DISSERTATION

Generative Speech Enhancement in Multimodal Applications

Kumulative Dissertation zur Erlangung des akademischen Grades

Dr. rer. nat.

an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Universität Hamburg

eingereicht von

Julius Richter

Hamburg 2025

This thesis reprints IEEE copyrighted publications with permission. The respective copyright notice and full reference for each article is displayed on the cover page that precedes each included publication. For each publication, the accepted version of the publication is reprinted.

Julius Richter: Generative Speech Enhancement in Multimodal Applications

PRÜFUNGSKOMMISSION:

Erster Gutachter: (noch zu genehmigen)

Zweiter Gutachter: (noch zu genehmigen)

Dritter Gutachter: (noch zu genehmigen)

Vorsitz: (noch zu genehmigen)

Stellvertretender Vorsitz: (noch zu genehmigen)

TAG DER EINREICHUNG:

12.05.2025

TAG DER DISPUTATION:

(noch zu genehmigen)

Abstract

This dissertation advances generative speech enhancement by investigating both unsupervised and supervised machine learning approaches, with a focus on integrating visual information to improve robustness. The work is organized into three main contributions:

The first contribution focuses on *unsupervised* generative speech enhancement. We explore a Bayesian framework combining variational autoencoders (VAEs) trained on clean speech with a non-negative matrix factorization (NMF) noise model. We propose to use stochastic temporal convolutional networks (STCNs) with temporal and hierarchical latent variables to capture the dynamic structure of speech. We employ a Monte Carlo expectation-maximization algorithm for joint optimization of speech and noise parameters. Replacing the VAE with an STCN in the VAE-NMF framework enables us to learn a more expressive generative model for speech, leading to improved performance in speech enhancement. To incorporate visual cues, we propose a disentanglement learning approach for the latent variables, which allows the VAE to be conditioned on voice activity labels inferred from an audio-visual classifier. Conditioning on visual features enables the model to learn a more robust speech representation, improving the quality of the enhanced speech.

The second contribution pertains to *supervised* generative speech enhancement. We investigate diffusion models for high-quality speech restoration, introducing score-based generative models for speech enhancement (SGMSE), a novel method that adapts the diffusion process to learn clean speech posteriors conditioned on corrupted inputs. Notably, SGMSE is not limited to addressing additive corruptions; it is also suitable for restoring general speech communication artifacts, effectively handling diverse distortions such as background noise, reverberation, bandwidth limitation, codec artifacts, and packet loss. We provide a comprehensive review of diffusion models for audio restoration, highlighting their data-driven nature while also discussing their potential for integration into model-based approaches. We extend SGMSE to audio-visual speech enhancement by conditioning on visual features, and propose causal processing by adapting the network architecture. Additionally, we explore alternative diffusion processes, including the Schrödinger bridge, to improve efficiency and perceptual quality.

The third contribution of this dissertation is an analysis of generative speech enhancement methods in comparison to predictive approaches. We conduct evaluations of generative methods against predictive methods using the Expressive Anechoic Recordings of Speech (EARS) dataset, a high-quality 48 kHz speech corpus we curated, which encompasses a variety of speaking styles, emotional prosody, and conversational speech. To facilitate this evaluation, we create two speech enhancement benchmarks based on the EARS dataset, with controlled background noise and reverberation, respectively. Our improved model *SGMSE+* consistently outperforms all baseline methods on these benchmarks across both objective measures and subjective listening evaluations.

By addressing critical research questions within unsupervised, supervised, and audio-visual frameworks, this work demonstrates the use of generative models as a powerful paradigm for speech enhancement, with significant implications for reliable communication and audio restoration.

Zusammenfassung

In dieser Dissertation wird die generative Sprachverbesserung weiterentwickelt, indem sowohl unüberwachte als auch überwachte Ansätze des maschinellen Lernens untersucht werden. Ein Schwerpunkt liegt dabei auf der Integration visueller Informationen zur Verbesserung der Robustheit. Die Arbeit gliedert sich in drei Hauptbeiträge:

Der erste Beitrag konzentriert sich auf die *unüberwachte* generative Sprachverbesserung. Wir untersuchen einen Bayes'schen Ansatz, der Variational Autoencoder (VAEs), die auf sauberer Sprache trainiert werden, mit einem nichtnegativen Matrixfaktorisierungs (NMF)-Rauschmodell kombiniert. Wir schlagen vor, Stochastic Temporal Convolutional Networks (STCNs) mit zeitlichen und hierarchischen latenten Variablen zu verwenden, um die dynamische Struktur von Sprache zu erfassen. Wir verwenden einen Monte-Carlo Erwartungs-Maximierungs-Algorithmus für die gemeinsame Optimierung von Sprach- und Rauschparametern. Das Ersetzen des VAEs durch ein STCN im VAE-NMF-Framework ermöglicht es uns, ein aussagekräftigeres generatives Modell für Sprache zu lernen, was zu einer verbesserten Leistung bei der Sprachverbesserung führt. Um visuelle Hinweise einzubeziehen, schlagen wir einen Disentanglement Learning-Ansatz für die latenten Variablen vor, der es ermöglicht, den VAE auf Sprachaktivitäts-Labels zu konditionieren, die von einem audiovisuellen Klassifikator abgeleitet werden. Die Konditionierung auf visuelle Merkmale ermöglicht es dem Modell, eine robustere Sprachrepräsentation zu erlernen und so die Qualität der Sprache zu verbessern.

Der zweite Beitrag bezieht sich auf die *überwachte* generative Sprachverbesserung. Wir untersuchen Diffusionsmodelle für eine qualitativ hochwertige Sprach-Restauriering und führen Score-basierte generative Modelle für die Sprachverbesserung (SGMSE) ein, eine neuartige Methode, die den Diffusionsprozess anpasst, um saubere Sprachposteriori zu erlernen, die von beschädigten Eingaben abhängen. Insbesondere ist SGMSE nicht auf die Bearbeitung additiver Störungen beschränkt, sondern eignet sich auch für die Restaurierung allgemeiner Artefakte in der Sprachkommunikation, wobei verschiedene Störungen wie Hintergrundgeräusche, Nachhall, Bandbreitenbeschränkung, Codec-Artefakte und Paketverluste effektiv behandelt werden. Wir geben einen umfassenden Überblick über Diffusionsmodelle für die Audiorestauration, wobei wir ihre datengesteuerte Natur hervorheben und gleichzeitig ihr Potenzial für die Integration in modellbasierte Ansätze diskutieren. Wir erweitern SGMSE zur audiovisuellen Sprachverbesserung durch Konditionierung auf visuelle Merkmale und schlagen eine kausale Verarbeitung durch Anpassung der Netzwerkarchitektur vor. Darüber hinaus untersuchen wir alternative Diffusionsprozesse, einschließlich der Schrödinger-Brücke, um die Effizienz und Wahrnehmungsqualität zu verbessern.

Der dritte Beitrag dieser Dissertation beinhaltet die Analyse generativer Sprachverbesserung im Vergleich zu prädiktiven Ansätzen. Wir evaluieren generative Methoden im Vergleich zu prädiktiven Methoden anhand des Expressive Anechoic Recordings of Speech (EARS)-Datensatzes, einem von uns kuratierten hochwertigen 48 kHz-Sprachkorpus, der eine Vielzahl von Sprechstilen, emotionaler Prosodie und Konversationssprache umfasst. Zur Evaluierung erstellen wir zwei Sprachverbesserungs-Benchmarks auf der Grundlage des EARS-Datensatzes,

jeweils mit kontrolliertem Hintergrundrauschen und Nachhall. Unser verbessertes Modell *SGMSE+* übertrifft alle Baseline Methoden bei diesen Benchmarks sowohl bei objektiven Messungen als auch bei subjektiven Hörbewertungen.

Durch die Auseinandersetzung mit kritischen Forschungsfragen in unüberwachten, überwachten und audiovisuellen Rahmenwerken demonstriert diese Arbeit die Verwendung generativer Modelle als ein überzeugendes Paradigma für die Sprachverbesserung, mit bedeutenden Auswirkungen auf zuverlässige Kommunikation und Audiorestauration.

Acknowledgement

This dissertation represents not merely the culmination of five years of research, but a testament to the collaborative nature of scientific progress. I am grateful for the many individuals who contributed to this intellectual journey.

First and foremost, I owe my deepest gratitude to my supervisor, Timo Gerkmann. Your thoughtful guidance, high standards, and unwavering trust have been instrumental in shaping this work and my development as a researcher. Over these years, your mentorship has taught me not only how to conduct rigorous research but also how to think critically and independently. The lessons I have learned under your supervision—about academic integrity, scientific rigor, and perseverance—will stay with me throughout my career. I am particularly grateful for the freedom you gave me to explore ideas while always providing insightful feedback to steer me in the right direction.

I would also like to thank Reinhard Zierke for his invaluable technical support and Stephanie Schulte for her organizational assistance. Your contributions were essential to my daily work and progress.

My sincere thanks go to the crossmodal learning (CML) project team, especially Jianwei Zhang and Xiaolin Hu for facilitating my research visit to Tsinghua University, which opened new perspectives for my work. Our cross-continental collaboration demonstrated the power of distributed scientific inquiry.

To my colleagues—Alina, Danilo, David, Guillaume, Hector, Huajian, Jakob, Jean-Marie, Kristina, Long, Navin, Simon, and Tal—your brilliant minds created an environment where ideas were stress-tested through spirited debate. The coffee breaks discussions, the late-night coding sessions, the peer reviews that sharpened our work—these interactions exemplified the collaborative spirit of science.

My family provided the foundation that sustained me through this process. Michael Flamm, your thoughtful advice was always timely. To my father-in-law, your gracious acceptance across cultural differences meant more than you realize. To my brother Lorenz, my first research debug partner in life—your quiet but unwavering support spoke volumes, and your rare but razor-sharp technical advice always came exactly when needed. And to my parents, your endless encouragement and sacrifices made this achievement possible.

To Xiaochen Huang, my beloved wife: your ability to pull me from the abstract world of algorithms back to life's simple joys kept me balanced. Your love, patience, and daily encouragement transformed this marathon into a joyous journey. Every page of this thesis carries your unwavering support.

Finally, I'm grateful for all the experiences—the struggles that taught me resilience, the breakthroughs that brought joy, and the privilege of pursuing knowledge. The halting problem of knowledge remains undecided—there will always be another question to explore, another bug to fix, another optimization to discover. While this particular process terminates here, new threads await spawning. The program continues.

Table of Contents

Abstract **i**

Zusammenfassung **iii**

Acknowledgement **v**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Traditional Speech Enhancement . . . 5
1.3 Learning-Based Speech Enhancement . . . 8
1.4 Predictive Methods for Speech Enhancement . . . 11
1.5 Generative Methods for Speech Enhancement . . . 16
1.5.1 Unsupervised Generative Speech Enhancement . . . 20
1.5.2 Supervised Generative Speech Enhancement . . . 23
1.6 Audio-Visual Speech Enhancement . . . 28
1.7 Outline and Contributions . . . 33
1.7.1 Chapter 3: *Unsupervised Generative Speech Enhancement* . . . 33
1.7.2 Chapter 4: *Supervised Generative Speech Enhancement* . . . 34
1.7.3 Chapter 5: *Analysis and Comparison* . . . 35

2 List of Publications **37**

3 Unsupervised Generative Speech Enhancement **41**
3.1 Variational Autoencoders for Speech Enhancement . . . 42
3.1.1 Publication: *Speech Enhancement with Stochastic Temporal Convolutional Networks* . . . 42
3.2 Extension to Audio-Visual Speech Enhancement . . . 48
3.2.1 Publication: *Disentanglement Learning for Variational Autoencoders Applied to Audio-Visual Speech Enhancement* . . . 48

4 Supervised Generative Speech Enhancement **55**
4.1 Diffusion Models for Speech Enhancement . . . 56
4.1.1 Publication: *Speech enhancement and dereverberation with diffusion-based generative models* . . . 56
4.1.2 Publication: *Diffusion Models for Audio Restoration: A review* . . . 72
4.2 Causal Processing for Restoration of Speech Communication Artifacts . . . 96
4.2.1 Publication: *Causal Diffusion Models for Generalized Speech Enhancement* . . . 96
4.3 Extension to Audio-Visual Speech Enhancement . . . 106
4.3.1 Publication: *Audio-visual speech enhancement with score-based generative models* . . . 106
4.4 Alternative Diffusion Processes . . . 112

4.4.1 Publication: *Investigating Training Objectives for Generative Speech Enhancement* . . . 112

5 Analysis and Comparison **119**
5.1 Analysis of Generative Speech Enhancement and Comparison to Predictive Approaches . . . 120
5.1.1 Publication: *EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation* . . . 120

6 Discussion and Conclusions **127**
6.1 Analysis of Contributions . . . 127
6.1.1 Chapter 3: *Unsupervised Generative Speech Enhancement* . . . 127
6.1.2 Chapter 4: *Supervised Generative Speech Enhancement* . . . 129
6.1.3 Chapter 5: *Analysis and Comparison* . . . 134
6.2 Discussion and Future Research . . . 135

References **141**

List of Acronyms **159**

Eidesstattliche Versicherung **161**

A Related Publications **165**
A.1 Variational Autoencoders for Speech Enhancement . . . 165
A.1.1 Publication: *Guided Variational Autoencoder for Speech Enhancement with a Supervised Classifier* . . . 165
A.2 Diffusion Models for Speech Enhancement . . . 172
A.2.1 Publication: *Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain* . . . 172
A.3 Causal processing for restoration of speech communication artifacts . . . 178
A.3.1 Publication: *Speech Signal Improvement Using Causal Generative Diffusion Models* . . . 178
A.4 Extension to audio-visual speech enhancement . . . 181
A.4.1 Publication: *LipDiffuser: Lip-to-Speech Generation with Conditional Diffusion Models* . . . 181
A.5 Alternative Diffusion Processes . . . 195
A.5.1 Publication: *Normalize Everything: A Preconditioned Magnitude-Preserving Architecture for Diffusion-Based Speech Enhancement* . . . 195
A.6 Analysis of Generative Speech Enhancement and Comparison to Predictive Approaches . . . 201
A.6.1 Publication: *Diffusion-based speech enhancement: Demonstration of performance and generalization* . . . 201
A.6.2 Publication: *StoRM: A Diffusion-based Stochastic Regeneration Model for Speech Enhancement and Dereverberation* . . . 205
A.6.3 Publication: *Analysing Diffusion-based Generative Approaches versus Discriminative Approaches for Speech Restoration* . . . 219

B Derivations **225**
B.1 Reverse SDE for Diffusion Processes . . . 225
B.2 Mean and Variance of the Perturbation Kernel . . . 226

B.3 Derivation for the EDM framework 228
B.4 Connection between Score Matching and Denoising 229

1 Introduction

1.1 Motivation

Most environments in our everyday life are noisy environments. Imagine, for a moment, your current surroundings—whether you are at home, in a cafe, or on a train ride. As you focus on your surroundings, you may notice a multitude of overlapping sounds, from conversations and the hum of engines to various ambient noises. At the same time, we are often surrounded by surfaces and structures that reflect sound waves from multiple sources, leading to reverberation. Together, background noise and reverberation establish a constant backdrop that significantly impacts our cognitive processes and communication abilities. Despite these disturbances, humans are remarkably adept at communicating by focusing on a specific auditory stimulus. This ability, known as the *cocktail party effect* [1], allows the brain to concentrate a person's attention on a particular sound while filtering out a multitude of other stimuli from conscious awareness. For instance, at a lively party, a guest can focus on a single conversation amid the background chatter and music. This selective auditory attention is crucial for effective communication in noisy environments, enabling us to interact and connect with others in our daily lives.

Nowadays, speech communication is no longer constrained by distance, as technology enables remote interactions through various means, including telecommunication and video conferencing. The rapid advancement of technology has also resulted in the widespread availability of smart devices and automated systems, such as smartphones, home appliances, self-driving cars, and interactive machines. Despite their prevalence, these speech communication devices face a common challenge: the target speech captured by one or more microphones is often distorted by interfering noise and reverberation. This issue highlights the need for robust speech enhancement algorithms, designed to extract the desired speech by effectively suppressing unwanted interference. Such algorithms are essential for ensuring clear and intelligible communication in both human-to-human and human-to-machine interactions, making them a vital component in today's technology-driven world.

Speech enhancement is a critical component in a wide range of applications, from improving the clarity of voice communications in teleconferencing and hearing aids to enhancing the performance of automatic speech recognition (ASR) systems and voice-controlled devices. While the terms speech enhancement, speech separation, and speech extraction are often used interchangeably, they have subtle distinctions. Speech enhancement generally refers to the process of improving the quality and intelligibility of a speech signal by reducing noise and reverberation. In contrast, speech separation involves isolating individual speech sources from a mixture of sounds, and speech extraction focuses on isolating a specific target speech signal, regardless of whether multiple speakers are present. In practice, these tasks are closely related, and the techniques used in their algorithms often exhibit significant similarities. In

Source of Sound	**Distance**	**SPL (dB)**
Firecracker [3]	0.5 m	∼170
Jet engine [4]	1 m	∼150
Threshold of pain [5]	At ear	120–140
Loudest human voice [6]	2.54 cm	∼135
Teacher talking in a classroom [7]	1 m	60–80
Traffic of electric vehicles [8]	10 m	65–75
Normal conversation at home [7]	1 m	45–70
Whispering [9]	1 m	20–40
Calm breathing [4]	Ambient	∼10
Auditory threshold at 1 kHz	At ear	0

Table 1.1: Typical sound pressure levels (SPLs) of various sound sources and human hearing thresholds, measured in decibels (dB) at a specific distance.

this thesis, we will mainly focus on speech enhancement methods but emphasize their ability to encompass these other related tasks. Additionally, while modern approaches often utilize multi-channel speech enhancement techniques that exploit spatial correlations captured by multiple microphones to improve performance in challenging conditions, this thesis focuses on single-channel speech enhancement using only one microphone.

Ensuring that speech enhancement algorithms remain robust across diverse acoustic conditions is both essential and challenging. A key factor in designing and evaluating speech enhancement algorithms is the *signal-to-noise ratio (SNR)*,

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} \right), \tag{1.1}$$

which measures the ratio of the speech signal power P_{signal} to the background noise power P_{noise}, often expressed in decibels (dB). Different areas of applications and environments present varying degrees of difficulty regarding the SNR level. For instance, a quiet office environment might exhibit a high SNR, such as 20 dB, indicating that the signal power is 100 times greater than the noise power. In contrast, a busy street could have a significantly lower SNR, for example, 0 dB, implying that the signal and noise powers are equal.

In addition to considering the clarity of the signal through SNR, it is also vital to understand the overall loudness of a sound source, which is where the *sound pressure level (SPL)* becomes relevant. The SPL is defined as

$$\mathrm{SPL} = 20 \log_{10} \left(\frac{p}{p_0} \right), \tag{1.2}$$

which is the effective pressure p of a sound relative to a reference pressure p_0, typically the threshold of human hearing (roughly the sound of a mosquito flying 3 m away) [2]. Table 1.1 presents typical SPLs of various sound sources and human hearing thresholds, measured in dB at a specific distance.

Besides evaluating clarity and loudness through SNR and SPL, other acoustic characteristics like reverberation significantly influence speech perception. Key reverberation parameters include the T_{60} *time* and the *direct-to-reverberant ratio (DRR)*. T_{60} is the time it takes for

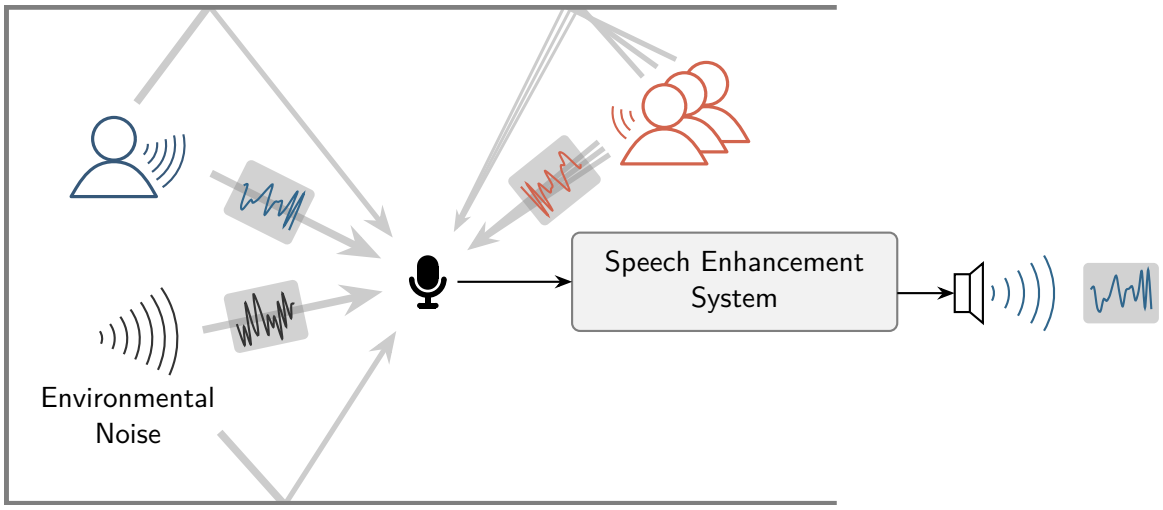


Figure 1.1: Speech enhancement scenario. This illustration depicts the interaction within a room enclosure among three sound sources: the target speech signal (blue), environmental noise from various sources (black), and babble noise from multiple other speakers (red). The sound waves are captured directly by the microphone and reflected off room surfaces, causing reverberation. The speech enhancement system processes this complex mixture to estimate and isolate the target speech signal (blue).

the reverberant sound to decay by 60 dB, indicating the reverberation time of a space, while the DRR measures the ratio between direct and reverberant sound energy, with higher values suggesting environments where the direct sound is predominant [10]. Understanding and integrating these factors into speech enhancement algorithms is vital for developing solutions that can adapt to varying noise and reverberation levels, ensuring clear communication across different environments.

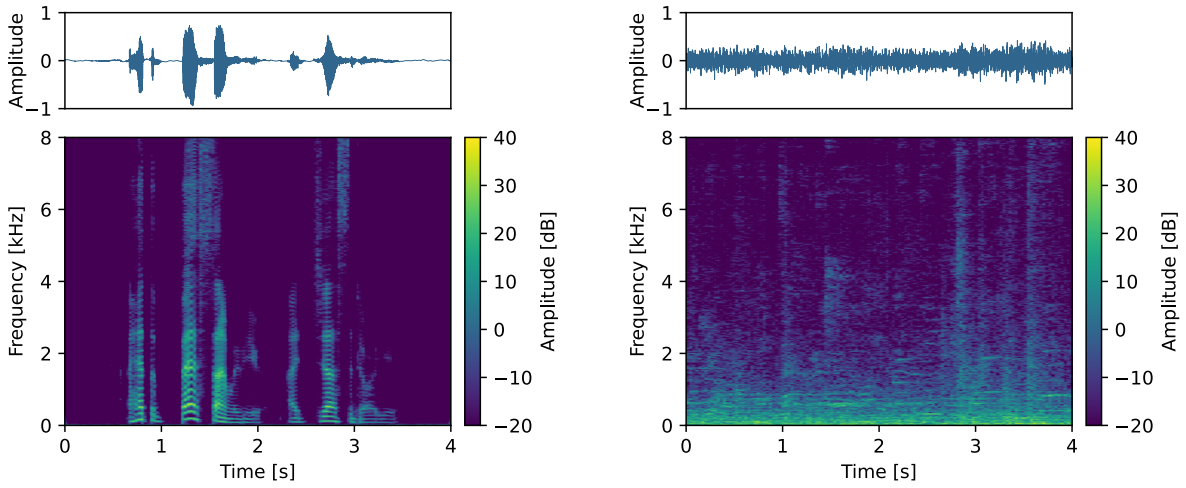
The challenges posed by noisy and reverberant environments can be mathematically described through a signal model that captures the effects of both additive background noise and reverberation. In this model, the observed microphone signal is typically represented as the sum of the target clean speech signal and a noise component originating from various environmental sources. This relationship is represented as

$$y(n) = x(n) + d(n), \tag{1.3}$$

where $y(n) \in \mathbb{R}^N$ denotes the observed noisy speech signal, $x(n) \in \mathbb{R}^N$ represents the clean speech signal, and $d(n) \in \mathbb{R}^N$ is the additive noise signal, all of which are discrete time-domain signals with N samples. Beyond the complications of additive background noise, reverberation introduces further challenges by convolving the clean speech with a room impulse response (RIR). Consequently, the signal model becomes

$$y(n) = x(n) * h(n) + d(n), \tag{1.4}$$

where $h(n) \in \mathbb{R}^{N_h}$ represents the RIR of the environment, with N_h indicating the number of samples in the RIR, and $*$ denotes the convolution operation. Figure 1.1 illustrates this scenario, showing the interaction between the target speech signal (blue), environmental



(a) Speech signal.

(b) Cafeteria background noise.

Figure 1.2: Comparison between a typical speech signal and cafeteria background noise: (a) depicts a speech signal, while (b) features cafeteria noise, which includes babble noise from distant speakers, dish clatter, and chair movement. Both are shown in the time domain (top) and the time-frequency domain (bottom) to illustrate their differences.

noise from various noise sources (black), and babble noise from multiple other speakers (red). Within a room enclosure, sound waves from these sources are not only directly captured by the microphone but also interact with the surfaces of the room, leading to reverberation. A speech enhancement system then processes the microphone signal, with the goal of estimating and isolating the target speech signal (blue) from this complex mixture.

A plethora of computational methods for speech enhancement have been proposed, with most of them operating in the time-frequency domain. These methods utilize the short-time Fourier transform (STFT) to analyze the temporal and spectral content of speech signals [11]. This is motivated by the fact that time-frequency representations of signals often reveal distinctive structures. For example, in the time-frequency domain, vowel sounds in speech exhibit harmonic structures characterized by frequency components that are integer multiples of the fundamental frequency. These harmonic structures provide valuable cues for distinguishing between speech and noise, as well as for isolating specific speakers in a mixture. Additionally, speech spectrograms typically appear sparse, with only a few time-frequency bins exhibiting significant amplitude values, while the majority remain minimal. This sparsity naturally results in less overlap of these prominent amplitude values [12]. Figure 1.2 compares a typical speech signal and cafeteria background noise, showcasing their representations in both the time domain and the time-frequency domain. It can be observed that the speech signal exhibits a more structured and sparse representation in the time-frequency domain, while the background noise appears more diffuse and stationary.

The next section, Section 1.2, will focus on traditional speech enhancement methods. This will be followed by Section 1.3, which discusses learning-based speech enhancement, with a specific emphasis on predictive methods in Section 1.4 and generative methods in Section 1.5. Additionally, Section 1.6 will cover audio-visual speech enhancement, leading into an outline of the contributions of this thesis in Section 1.7.

1.2 Traditional Speech Enhancement

Time-frequency domain techniques form the foundation for a variety of traditional speech enhancement methods [13]. Utilizing the STFT, the additive signal model introduced in (1.3) is applicable in the time-frequency domain, expressed as

$$Y(f,t) = X(f,t) + D(f,t), \tag{1.5}$$

where $Y(f,t) \in \mathbb{C}^{F\times T}$, $X(f,t) \in \mathbb{C}^{F\times T}$, and $D(f,t) \in \mathbb{C}^{F\times T}$ represent the noisy speech, clean speech, and noise spectrograms, respectively, Here, F represents the number of frequency bins and T represents the number of time frames in the spectrogram. Spectral masking and filtering techniques employ various models to derive masks based on signal statistics, with spectral subtraction and Wiener filtering being among the most fundamental approaches [14]. Since spectral masking is applied independently to each time frame, we will omit the index t for the remainder of this section.

Spectral subtractive algorithms are one of the earliest and most straightforward approaches for reducing noise in speech signals [15], [16]. In its fundamental form [15], spectral subtraction involves subtracting an estimated average noise magnitude spectrum $|\hat{D}(f)| \in \mathbb{R}^F$ from the magnitude spectral coefficients of the mixture to obtain an estimate of the clean speech amplitude spectrum $|\hat{X}_{\mathrm{SS}}(f)| \in \mathbb{R}^F$. This subtraction process can also be represented as a masking operation

$$|\hat{X}_{\mathrm{SS}}(f)| = |Y(f)| - |\hat{D}(f)| = \left(1 - \frac{|\hat{D}(f)|}{|Y(f)|}\right)|Y(f)|. \tag{1.6}$$

To reconstruct the clean speech spectrum, the estimated clean speech magnitude spectrum is combined with the phase of the noisy mixture. In online algorithms, the estimated noise spectrum $|\hat{D}(f)|$ is typically updated when the speech signal is absent, under the assumption that the noise is stationary or slowly varying. While these algorithms are computationally simple, they can introduce speech distortion and suffer from musical noise artifacts if the subtraction is not carefully managed. Numerous variations of the spectral subtraction algorithm have been proposed to mitigate such distortions [17], [18], [19], [20]. Nevertheless, it is crucial to recognize that spectral subtraction relies on a heuristic approach. While the complex spectral coefficients of speech and noise signals are additive, as shown in (1.5), their magnitudes, as well as the averages or expected values of these magnitudes, are not additive [14]. Consequently, from a mathematical perspective, the straightforward estimator in (1.6) is not optimal.

In contrast, Wiener filtering employs a more mathematically rigorous approach to optimize an error criterion, specifically the mean square error (MSE) between the target clean speech $X(f)$ and the estimate $\hat{X}_{\mathrm{WF}}(f) = H(f)Y(f)$. This is formulated as

$$H_{\mathrm{WF}}(f) = \arg\min_{H(f)} \mathbb{E}\left[|X(f) - H(f)Y(f)|^2\right], \tag{1.7}$$

where, the optimal filter $H(f)$ that minimizes the MSE is known as the Wiener filter $H_{\mathrm{WF}}(f)$. The filter is named after mathematician Norbert Wiener [21], who first formulated and solved this filtering problem in the continuous domain. Assuming both speech and noise are Gaussian

processes, the Wiener filter estimates the clean speech spectrum $\hat{X}_{\mathrm{WF}}(f) \in \mathbb{C}^F$ as

$$\hat{X}_{\mathrm{WF}}(f) = \underbrace{\frac{P_X(f)}{P_X(f) + P_D(f)}}_{H_{\mathrm{WF}}(f)} Y(f), \tag{1.8}$$

where $P_X(f) = \mathbb{E}[|X(f)|^2]$ and $P_D(f,t) = \mathbb{E}[|D(f)|^2]$ denote the power spectral densitys (PSDs) of the clean speech and noise, respectively [22]. The Wiener filter is a linear estimator of the complex clean spectrum and is optimal in the minimum mean square error (MMSE) sense when both the complex noise and speech discrete Fourier transform (DFT) coefficients are assumed to be independent Gaussian random variables [22]. However, from (1.8) it is evident that the Wiener filter requires accurate estimates of the PSDs of the clean speech and noise, since these quantities cannot be assumed known.

Other statistical model-based algorithms focus on the nonlinear estimation of DFT magnitudes instead of the complex coefficients and can be framed within the statistical estimation theory [23]. Consider a dataset $\mathcal{D} = \{Y_1(f), \ldots, Y_M(f)\}$ comprising M noisy speech spectra $Y_i(f) \in \mathbb{C}^F$, and assume the probability density function of Y, denoted as $p(Y(f); \theta)$, is parameterized by unknown parameters θ. The objective of *maximum likelihood (ML)* estimation is to determine the value of θ that most likely produced the observed dataset $\mathcal{D}$. Formally, the ML estimate $\hat{\theta}_{\mathrm{ML}}$ is defined as

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \sum_{i=1}^{M} p(Y_i(f); \theta), \tag{1.9}$$

where $p(Y(f); \theta)$ is referred to as the *likelihood function* in this context. McAulay and Malpass [24] were the first to apply the ML approach to speech enhancement. Under certain distributional assumptions and after eliminating the phase parameter, they derived the ML estimator for the clean speech amplitude spectra $|\hat{X}_{\mathrm{ML}}(f)|$ as

$$|\hat{X}_{\mathrm{ML}}(f)| = \left(\frac{1}{2} + \frac{1}{2} \sqrt{\frac{|Y(f)|^2 + \sigma_N^2(f)}{|Y(f)|^2}} \right) |Y(f)|, \tag{1.10}$$

where $|Y(f)| \in \mathbb{R}^F$ are the magnitudes of the noisy speech, and $\sigma_N^2 \in \mathbb{R}^F$ are the frequency-dependent noise variances. To reconstruct the clean speech signal, they combined the estimated magnitude spectrum with the noisy phase. Note that the derivation assumes the signal magnitude is unknown but deterministic, and the phase is assumed to be uniformly distributed.

In ML estimation, the searched quantity, denoted by θ, is assumed to be deterministic. The focus is on finding the parameter values that maximize the likelihood of the observed data, without regard for any prior information about θ. In contrast, when these parameters are modeled as random variables, it leads to a *Bayesian* approach. This framework allows for the integration of *a priori* knowledge about the parameters, facilitating the estimation of the *posterior distribution* $p(\theta|\mathcal{D})$, which represents the conditional distribution of θ given the observed dataset $\mathcal{D}$. In speech enhancement, the posterior distribution relates to the conditional distribution of clean speech given the noisy speech. Minimization of the Bayesian MSE with respect to the clean speech amplitudes $|\hat{X}|$ leads to the optimal MMSE estimator

given by,

$$|\hat{X}(f)|_{\text{MMSE}} = \mathbb{E}[|\hat{X}(f)|\,|Y(0),\ldots,Y(F)], \tag{1.11}$$

which is the mean of the posterior distribution of speech amplitudes after observing the noisy spectra. Note that there are two fundamental differences between the Wiener estimator in (1.7) and the MMSE estimator described in (1.11). First, during the derivation of the Wiener filter, it is assumed that there is a linear relationship between the observed data Y and the estimator $\hat{X}_{\text{WF}}$. Second, the Wiener filter is obtained by evaluating the mean of the posterior distribution of X rather than the posterior of the magnitudes $|X|$. This makes the Wiener filter the optimal estimator in the MMSE sense for complex spectra, not for magnitude spectra, under the assumed model. The posterior mean estimator, unlike the Wiener estimator, does not assume a linear relationship between the observed data and the estimator. However, it requires knowledge of the probability density functions of the speech and noise DFT coefficients. Obtaining the true probability density functions of the DFT coefficients is challenging because both the speech and the noise signals are neither stationary nor ergodic processes [11]. Attempts to measure these distributions often focus on the long-term behavior of the processes [25], [26], but it is uncertain whether large data histograms realistically measure the true probability densities of the DFT coefficients.

In the seminal paper by Ephraim and Malah [27], the authors circumvented the above-mentioned challenges by introducing a statistical model-based approach to speech enhancement that utilizes the asymptotic statistical properties of the DFT coefficients. They assume that the real and imaginary parts of the DFT coefficients are statistically independent and uncorrelated, following a Gaussian probability distribution with zero mean and time-varying variances to capture the non-stationary behavior of speech signals. Under these assumptions, the authors derive the MMSE estimator for the clean speech magnitude spectrum $|\hat{X}(f)|_{\text{MMSE}} \in \mathbb{R}^F$ as

$$|\hat{X}(f)|_{\text{MMSE}} = \frac{\sqrt{\nu(f)}}{\gamma(f)}\,\Gamma(1.5)\,\Phi(0.5,1;-\nu(f))\,|Y(f)| \tag{1.12}$$

where $\Gamma(z)$ denotes the gamma function [28, Eq. (8.310.1)], $\Phi(a,b;z)$ is the confluent hypergeometric function [28, Eq. (9.210.1)], ν is defined as

$$\nu(f) = \frac{\xi(f)}{1+\xi(f)}\gamma(f), \tag{1.13}$$

and the *a posteriori* SNR $\gamma(f)$ and the *a priori* SNR $\xi(f)$ are defined by

$$\gamma(f) = \frac{|Y(f)|}{\sigma_N^2(f)} \tag{1.14}$$

and

$$\xi(f) = \frac{\sigma_X^2(f)}{\sigma_N^2(f)}. \tag{1.15}$$

Furthermore, the authors show that the noisy phase is the optimal phase in the MMSE sense under the independence assumption of the DFT coefficients. Thus, the final enhanced speech signal is obtained by combining the estimated magnitude spectrum with the noisy phase.

The MMSE amplitude estimator in (1.12) was derived under the assumption that the *a priori* SNR $\xi(f)$ and the noise variance $\sigma_N^2(f)$ are known. In practice, however, these quantities are

unknown and need to be estimated from the noisy speech signal. Several methods have been proposed to estimate the *a priori* SNR, including [27], [29], [30]. The task of noise power estimation has been addressed in [31], [32], [33], [34].

Non-negative matrix factorization (NMF) is a widely used technique to model signals in the time-frequency domain, particularly for applications in speech enhancement and source separation [35]. NMF aims to factorize a non-negative matrix, such as an amplitude spectrogram, into the product of two lower-dimensional matrices while ensuring these matrices remain non-negative. Formally, given a non-negative matrix $\mathbf{D} \in \mathbb{R}_+^{F \times T}$, which represents for example the magnitude spectrogram $|D(f,t)|$ of a noise signal, NMF decomposes it into two matrices: a dictionary matrix $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and an activation matrix $\mathbf{H} \in \mathbb{R}_+^{K \times T}$, such that their product approximates the matrix $\widehat{\mathbf{D}} \approx \mathbf{WH}$. Here, F is the number of frequency bins, T is the number of time frames, and K is the number of basis components, typically smaller than both F and T. The matrix $\mathbf{W}$ serves as a set of basis spectra or templates, capturing the underlying patterns in the data, while $\mathbf{H}$ represents the temporal activity of these patterns. Thus, each column vector or frame $\mathbf{d}_t \in \mathbb{R}^F$ of the spectrogram can be interpreted as a linear combination of the basis spectra (columns of $\mathbf{W}$), weighted by the coefficients in the corresponding column $\mathbf{h}_t \in \mathbb{R}^K$ of the activation matrix. To optimize the decomposition, a cost function $D(\mathbf{D}|\mathbf{WH})$ is minimized, which measures the discrepancy between the original and approximated matrix. Common choices for the cost function include the Euclidean distance and Kullback-Leibler divergence. Févotte et al. [35] introduced an alternative cost function based on the Itakura-Saito divergence, which is specifically advantageous in audio applications because it treats frequency components equally regardless of energy distribution. Training schemes can vary based on application and data availability. For instance, dictionary matrices for known sources can be pre-learned, whereas unknown noise models are adapted dynamically at test time. Such a setup allows for flexible and robust integration of incomplete prior knowledge across various application scenarios, enhancing the generality and effectiveness of NMF-based speech enhancement systems.

Traditional signal processing techniques have played a crucial role in the development of successful speech enhancement algorithms. However, these methods often depend on certain distributional assumptions and signal model approximations that may not align with real-world scenarios. For example, the common assumption that noise changes more slowly than speech can limit the effectiveness when dealing with non-stationary noise. Learning-based approaches, on the other hand, do not require adherence to such simplifying assumptions. They offer the potential for significant performance improvements, particularly through deep learning-based methods, by learning spectro-temporal patterns directly from data. These methods have gained widespread recognition for their exceptional ability to learn intricate and meaningful representations from data. In the following section, we will introduce learning-based speech enhancement, with a focus on both predictive and generative deep learning techniques.

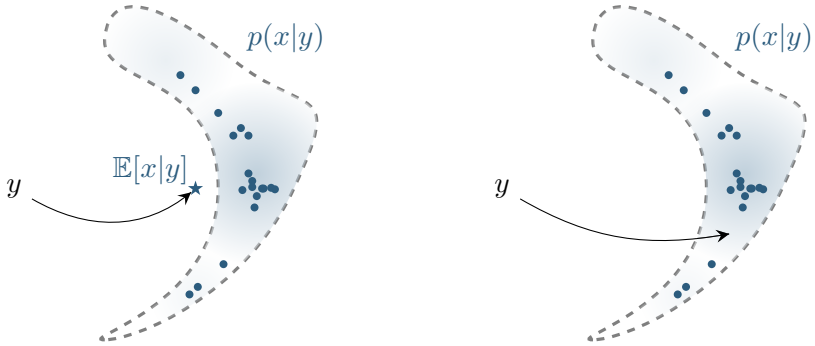
1.3 Learning-Based Speech Enhancement

Recent advances in machine learning, especially in deep learning, have enabled the development of data-driven signal processing algorithms that outperform traditional approaches by a large margin [36]. This transition to learning-based algorithms marks a paradigm change in the field of speech enhancement, bringing forward machine learning strategies that effectively

manage the diversity and complexity of real-world acoustic environments, such as varying noise levels, reverberation, and overlapping speech. Unlike traditional methods, such as those described in Section 1.2, which often struggle with challenging acoustic conditions, deep learning techniques can adapt more effectively due to their ability to learn and exploit temporal and spectral structures present in the data. These improvements are rooted in essential developments in machine learning, such as the introduction of backpropagation and the evolution of neural network architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Customized neural network architectures allow deep learning models to capture complex patterns and temporal dependencies in audio signals, providing a foundation for more robust and adaptive speech enhancement solutions. We will briefly explore these pivotal developments in the field of machine learning below.

In 1943, McCulloch and Pitts [37] introduced the first artificial neuron and considered RNNs. In 1960, Kelly [38] proposed a precursor of backpropagation in the field of control theory. In the same decade, Ivakhnenko and Lapa [39] trained neural networks layer by layer, and Amari [40] trained neural networks consisting of five layers by stochastic gradient descent. In 1972, Amari [41] published a paper about training RNNs. In 1980, Fukushima [42] introduced the *Neocognitron*, the first basic CNN. In 1982, Hopfield [43] introduced the *Hopfield network*, a type of RNN. This contribution was one of the achievements recognized when he was awarded the Nobel Prize in Physics in 2024. In 1986, Rumelhart et al. [44] showed that backpropagation yields useful internal representations. In 1987, Waibel [45] proposed the use of time-delay neural networks for phoneme recognition, employing techniques such as convolution, weight sharing, and backpropagation. This work was subsequently published as a journal paper two years later [46]. Time-delay neural networks have been employed in early methods of audio-visual speech recognition [47]. Since 1989, LeCun et al. [48] have made significant contributions to the development of CNNs, including the introduction of the *LeNet-5* architecture [49], which was one of the first successful applications of CNNs for handwritten digit recognition. LeCun's work also popularized the use of convolutional layers, pooling, and gradient-based learning for image recognition tasks, laying the foundation for modern deep learning [50], [51]. In 1997, Hochreiter and Schmidhuber [52] introduced the long short-term memory (LSTM) network, which is a type of RNN that can learn long-term dependencies. In 2006, Hinton and Salakhutdinov [53] demonstrated that autoencoder models could learn low-dimensional representations of data and described an effective method for initializing Boltzmann machines. This achievement was among those recognized when Hinton was awarded the Nobel Prize in Physics in 2024.

Recent developments in deep learning have been driven by the availability of large datasets and computational resources. In 2012, Krizhevsky et al. [54] introduced *AlexNet*, a deep CNN that achieved remarkable performance in image classification tasks and kickstarted the so-called "deep learning revolution" [55]. In 2014, Goodfellow et al. [56] introduced generative adversarial networks (GANs), a novel learning paradigm that can generate realistic images. A year later, Ronneberger et al. [57] introduced the *U-Net*, a deep learning architecture designed for medical image segmentation tasks. Since then, it has been widely adopted for various applications, including time-frequency domain speech processing. In 2016, He et al. [58] introduced *ResNet*, a deep residual network that helped train very deep neural networks. In the same year, van den Oord et al. [59] introduced *WaveNet*, a temporal convolutional network using dilated convolutions for generating audio waveforms. In 2017, Vaswani et al. [60] introduced the *Transformer*, a novel architecture based on self-attention mechanisms that achieved state-of-the-art performance in machine translation tasks. This has led to



(a) Predictive model. (b) Generative model.

Figure 1.3: Comparison of predictive and generative approaches for speech enhancement. Predictive models focus on learning a deterministic mapping from noisy to clean speech, aiming to approximate the mean of the posterior distribution $\mathbb{E}[x|y]$, under a least-squares objective. Generative models, in contrast, estimate and sample from the posterior distribution $p(x|y)$ of clean speech given the noisy mixture.

the development of large language models that are trained on vast amounts of text from the internet. For instance, *Llama 3* [61] is trained on over 15 T tokens, where tokens are the fundamental units of text, such as words or subwords, used by the model for processing language.

As we shift our focus back to speech enhancement, it is essential to recognize how the foundational developments in machine learning have influenced this field in recent decades. Deep neural networks (DNNs), with their ability to arbitrarily approximate any function, provide remarkable flexibility in incorporating complex relationships into algorithmic frameworks. In the field of speech enhancement, a distinction arises between *predictive methods* and *generative methods*. Figure 1.3 illustrates the difference between these two approaches. A predictive model learns a deterministic mapping from noisy to clean speech, aiming to minimize the average error over the training data according to a specific loss function. When trained with a least squares objective, the optimal model outputs the mean of the posterior distribution of clean speech given the noise input (see Figure 1.3a). Generative methods, on the other hand, aim at estimating and sampling from the data distribution. For speech enhancement, this relates to the posterior distribution of clean speech given the noisy mixture (see Figure 1.3b). This can be used to infer a measure of uncertainty for their predictions and to allow the generation of multiple valid estimates instead of a single best estimate as in predictive approaches [62]. Moreover, the landscape of machine learning techniques for speech enhancement is typically categorized into two primary learning paradigms: *supervised learning* and *unsupervised learning*. Supervised learning methodologies leverage paired datasets comprising clean and noisy speech. On the other hand, unsupervised learning in speech enhancement involves algorithms trained solely on clean speech data, without presenting the model with any noise examples during training.

In the following sections, we explore the details of predictive methods for speech enhancement in Section 1.4 and generative methods in Section 1.5, highlighting their unique characteristics and introducing selected methods. The section about generative speech enhancement is further divided into two subsections: Section 1.5.1 covers unsupervised generative speech

enhancement, while Section 1.5.2 focuses on supervised generative speech enhancement.

1.4 Predictive Methods for Speech Enhancement

The utilization of deep learning techniques for speech enhancement has emerged as an active area of research [63]. Recent algorithms leverage the advanced non-linear modeling capabilities of DNNs, resulting in impressive state-of-the-art performance. Among these approaches, predictive methods are particularly prevalent, focusing on learning a deterministic function $P_\theta : \mathbb{R}^N \to \mathbb{R}^N$, parameterized by θ, which maps noisy audio mixtures $\mathbf{y} \in \mathbb{R}^N$ to the corresponding clean speech estimation $\hat{\mathbf{x}} \in \mathbb{R}^N$. Predictive methods are typically trained using a dataset $\mathcal{D} = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(M)}, \mathbf{y}^{(M)})\}$ that comprises M pairs of clean and noisy speech samples. It is important to note that this dataset is often synthetically generated by artificially adding environmental noise to clean speech targets. This is a common practice in speech enhancement research due to the linear nature of the additive noise model in (1.3). The training objective is to minimize a point-wise distance between the predicted clean speech $\hat{\mathbf{x}} = P_\theta(\mathbf{y})$ and the true clean speech $\mathbf{x}$, thereby determining the optimal parameters θ^*_{pred} that minimize the expected loss over the training data,

$$\theta^*_{\text{pred}} = \arg\min_\theta \mathbb{E}_{(\mathbf{x},\,\mathbf{y})\sim\mathcal{D}} \left[\mathcal{L}(P_\theta(\mathbf{y}), \mathbf{x})\right], \tag{1.16}$$

where $\mathcal{L} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is called the *loss function*, which quantifies the discrepancy between the predicted clean speech and the ground-truth clean speech. The training dataset is typically designed to encompass diverse acoustic conditions, including multiple speakers, various types of background noise, and different input SNRs. This diversity helps ensure that the models can generalize effectively across real-world scenarios. However, it is nearly impossible to encompass all potential acoustic conditions within the training data, which poses a challenge for effective generalization. The performance gap becomes more pronounced as the distribution shift increases, particularly when the model encounters novel or extreme conditions that were not sufficiently represented during training. This limitation highlights the necessity of developing robust algorithms capable of adapting to unforeseen variations in real-world environments.

Neural networks have been explored for speech enhancement since the late 1980s. Tamura and Waibel [64] introduced a network that operates directly on the raw audio waveforms. Their approach employed feed-forward layers to process audio on a frame-by-frame basis using a speaker-dependent and isolated-word database. Parveen and Green [65] demonstrated the application of missing data techniques in speech enhancement through a two-stage system. The first stage involves an RNN that produces enhanced speech, while the second stage employs missing data techniques to further refine the speech quality. Lu et al. [66] developed a deep learning approach for speech enhancement using an autoencoder with layer-wise pretraining. Their model was trained exclusively on clean speech, with the assumption that it would reconstruct the clean speech when given noisy speech as input. In subsequent research, Lu et al. [67] enhanced their autoencoder by training with clean and noisy speech pairs. Huang et al. [68] explored deep learning for monaural speech separation. They explored a predictive training criterion for the neural networks to further enhance the separation performance. Xu et al. [69] proposed a supervised method for speech enhancement using a DNN trained on a large dataset composed of various combinations of speech and noise types. They introduced several techniques to optimize the DNN-based system, including global variance equalization

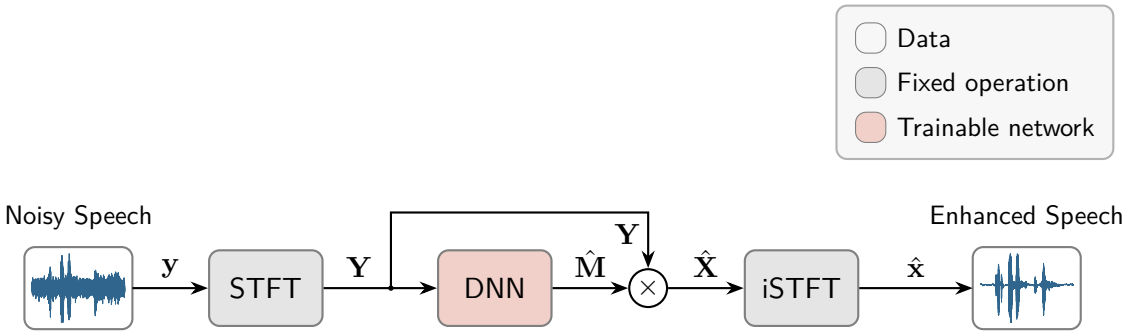


Figure 1.4: Masking-based speech enhancement. First, the noisy speech signal $\mathbf{y}$ is converted into the time-frequency domain using the short-time Fourier transform (STFT). A deep neural network (DNN) then predicts a mask $\hat{\mathbf{M}}$, which is applied to the noisy spectrogram $\mathbf{Y}$ via element-wise multiplication, resulting in the estimated clean speech spectrogram $\hat{\mathbf{X}}$. Finally, the enhanced speech signal $\hat{\mathbf{x}}$ is reconstructed using the inverse short-time Fourier transform (iSTFT).

to mitigate the over-smoothing problem inherent in regression models. Additionally, they applied dropout and noise-aware training strategies to enhance the DNN's ability to generalize to unseen noise conditions.

Other methods in predictive speech enhancement focus on integrating DNNs into traditional systems, yielding substantial improvements in speech quality and intelligibility. Zhang et al. [70] have proposed an MMSE-based noise PSD tracker that employs a temporal convolutional network as an *a priori* SNR estimator. Nicolsson and Paliwal [71] studied different learning approaches to clean speech magnitude spectrum estimation for speech enhancement. They found that training targets that estimate the *a priori* SNR for MMSE estimators produce the highest objective quality scores. Rehr and Gerkmann [72] addressed the challenge of generalizing DNN-based speech enhancement to unseen noise conditions, particularly when training data is limited in size and diversity. They proposed a SNR-based noise-aware training approach, which demonstrated robust and level-independent performance, even with simple architectures and limited training data. Tammen et al. [73] proposed to use an LSTM network to estimate the speech presence probability for each time-frequency bin. Their findings demonstrate that combining a multi-frame minimum power distortionless response filter with the proposed deep learning-based speech presence probability estimator results in enhanced speech quality compared to traditional model-based estimators of speech presence probability. In their follow-up work, Tammen and Doclo [74] proposed to embed the multi-frame minimum variance distortionless response (MVDR) filter within a deep learning framework. They trained a temporal convolutional network to map the noisy STFT coefficients to the required quantities by minimizing the scale-invariant source to distortion ratio (SI-SDR) loss function at the output of the multi-frame MVDR filter.

Recent deep learning-based approaches often utilize end-to-end speech enhancement frameworks, where the network directly maps noisy audio mixtures to clean speech signals. Various approaches have been proposed, including time-frequency masking [75], complex spectral mapping [76], or operating directly in the time domain [77]. Figure 1.4 presents a typical masking-based approach for speech enhancement in the time-frequency domain. The signal processing pipeline starts by converting the noisy speech signal $\mathbf{y}$ into a spectrogram $\mathbf{Y}$ using the STFT. Next, a DNN estimates a mask $\hat{\mathbf{M}}$, which is applied to the noisy spectrogram

$\mathbf{Y}$ through element-wise multiplication to produce the estimated clean speech spectrogram $\hat{\mathbf{X}}$. Finally, the enhanced speech signal $\hat{\mathbf{x}}$ is obtained by using the inverse short-time Fourier transform (iSTFT). Masking can be applied either to the complex spectrogram $\mathbf{Y} \in \mathbb{C}^{T \times F}$ or the magnitude spectrogram $|\mathbf{Y}| \in \mathbb{R}^{T \times F}$, using the noisy phase for reconstruction. The mask can be binary $\hat{\mathbf{M}} \in \{0,1\}^{T \times F}$, soft $\hat{\mathbf{M}} \in [0,1]^{T \times F}$, or complex $\hat{\mathbf{M}} \in \mathbb{C}^{T \times F}$. The loss function can be defined end-to-end in the time domain, with backpropagation through the STFT operation, in the time-frequency domain, or a combination of both.

The mask function can also be directly optimized by the DNN by minimizing the discrepancy between the estimated mask $\hat{\mathbf{M}}$ and an *ideal* or *oracle mask* $\mathbf{M}^*$. For overcomplete time-frequency transforms such as the STFT, calculating the oracle mask with respect to minimizing distortions in the time-domain is challenging due to the non-orthogonality of the transform [78]. In practice, however, this issue is often neglected, and the oracle mask is calculated independently for each time-frequency bin, which leads to the optimization problem

$$M^*(f,t) = \underset{M(f,t)}{\arg\min} \, |X(f,t) - M(f,t)Y(f,t)|^2. \tag{1.17}$$

Solving this optimization problem under the constraint $M(f,t) \in [0,1]$ results in the ideal soft mask also known as the ideal phase-sensitive mask (PSM) [79], which is defined as

$$\mathrm{PSM}(f,t) = \begin{cases} 0, & \text{if } r(f,t) < 0 \\ 1, & \text{if } r(f,t) > 1 \\ r(f,t), & \text{otherwise,} \end{cases} \tag{1.18}$$

with

$$r(f,t) = \mathrm{Re}\left(\frac{X(f,t)}{Y(f,t)}\right) = \frac{|X(f,t)|}{|Y(f,t)|}\cos(\angle X(f,t) - \angle Y(f,t)), \tag{1.19}$$

where $\angle X(f,t)$ and $\angle Y(f,t)$ are the phase angles of the clean and noisy speech coefficients, respectively.

Other works consider the ideal ratio mask (IRM) [80], defined as

$$\mathrm{IRM}(f,t) = \frac{|X(f,t)|}{|X(f,t)| + |D(f,t)|}, \tag{1.20}$$

or an ideal *"Wiener-like"* mask, defined as

$$\mathrm{IWM}(f,t) = \frac{|X(f,t)|^2}{|X(f,t)|^2 + |D(f,t)|^2}, \tag{1.21}$$

which is related to the Wiener filter in (1.7) that involves the expected values of the squared magnitudes of the signal components being estimated. It should be noted, however, that the coefficients obtained from adaptive Wiener filtering are optimal in a probabilistic sense under the assumption that the targets follow a zero-mean Gaussian distribution, but they do not necessarily ensure the smallest possible distortion [78].

Optimizing the problem in (1.17) under the constraint $M(f,t) \in \{0,1\}$ results in the ideal

binary mask (IBM) [14], which is defined as

$$\mathrm{IBM}(f,t) = \begin{cases} 0, & \text{if } r(f,t) < 0.5 \\ 1, & \text{otherwise.} \end{cases} \tag{1.22}$$

Other works defined the ideal binary mask as assigning a value of 1 when the target energy in a time-frequency bin is greater than the interference energy, and 0 otherwise [81], [82]. Looking at (1.22), it turns out that the IBM is a simple function of the ideal PSM in (1.18). In fact, it has been shown that soft masking improves speech intelligibility [83], [84], and increases the maximum achievable signal-to-distortion ratio (SDR) by 3 dB compared to binary masking [78].

Williamson et al. [75] proposed the complex ideal ratio mask (cIRM), defined as

$$\mathrm{cIRM}(f,t) = \underbrace{\frac{Y_r(f,t)X_r(f,t) + Y_i(f,t)X_i(f,t)}{Y_r(f,t)^2 + Y_i(f,t)^2}}_{M_r(f,t)} + j \underbrace{\frac{Y_r(f,t)X_i(f,t) - Y_i(f,t)X_r(f,t)}{Y_r(f,t)^2 + Y_i(f,t)^2}}_{M_i(f,t)}, \tag{1.23}$$

where the subscripts r and i indicate the real and imaginary parts, respectively, and j represents the imaginary unit. $M_i(f,t) \in \mathbb{R}$ and $M_r(f,t) \in \mathbb{R}$ are the real and imaginary parts of the cIRM, respectively, and a real-valued DNN can be trained to predict them separately [75]. Note that the complex mask may have large real and imaginary components with values in the range $(-\infty, \infty)$, which can be problematic for training. Therefore, the authors propose to compress the cIRM, such that

$$\mathrm{ccIRM}(f,t) = K\frac{1 - e^{-CM_r(f,t)}}{1 + e^{-CM_r(f,t)}} + jK\frac{1 - e^{-CM_i(f,t)}}{1 + e^{-CM_i(f,t)}}, \tag{1.24}$$

where this compression produces mask values within $[-K, K]$ and C controls its steepness. They found empirically that $K = 10$ and $C = 0.1$ work well for their experiments.

In addition to time-frequency methods that use the STFT, masking can also be applied in a learned representation space by employing an encoder and decoder in place of the STFT and iSTFT. Luo and Mesgarani [85] introduced *TasNet*, a model comprising a fully-connected encoder-decoder network designed to learn representations directly from raw waveforms. In addition, it includes a separator network that estimates a mask for source separation within this learned representation space. In principle, *TasNet* functions like a masking-based approach, as shown in Figure 1.4. However, instead of using the STFT and iSTFT, it employs a learned encoder and decoder network. Follow-up work by Luo and Mesgarani [86] introduced *Conv-TasNet*, which replaces the LSTM-based separator with a convolutional separator. This modification overcame the scalability and efficiency limitations of the original TasNet, resulting in improved performance in speech enhancement tasks. Luo et al. [87] further extended *Conv-TasNet* by introducing a dual-path architecture called *DPRNN* that processes the input waveform in parallel, allowing the model to capture both local and global information in the raw waveform. Hu et al. [88] introduced *DCCRN*, which performs complex-valued operations within both CNN and RNN structures to more effectively train on complex targets. This method achieved first place in the real-time track of the Interspeech 2020 deep noise suppression (DNS) challenge [89] with its competitive performance despite having only 3.7 M parameters.

Other end-to-end methods directly transform the noisy speech into clean speech samples.

Défossez et al. [90] developed *Demucs*, a causal speech enhancement model that operates directly on raw waveforms using an encoder-decoder architecture with skip-connections, optimized in both time and time-frequency domains. Li et al. [91] introduced *GaGNet*, a predictive speech enhancement method operating in the complex domain for monaural noise suppression. The proposed system comprises two modules: a spectral feature extraction module and so-called stacked glance-gaze modules. In the spectral feature extraction module, the authors employ *U-Net* blocks after each convolution layer, allowing for feature recalibration from multiple scales. They decompose the multi-target optimization in the complex spectrum into two sub-tasks: The glance path aims to suppress noise in the magnitude domain to achieve a coarse estimation, while the gaze path seeks to compensate for lost spectral details in the complex domain.

Several methods have been developed to optimize predictive speech enhancement systems for specific perceptual-based metrics. Koizumi et al. [92] proposed a method based on optimizing objective sound quality assessment scores, including the perceptual evaluation of speech quality (PESQ) and the short-time objective intelligibility (STOI) metric. In their approach, a DNN estimates the probability density function of output signals, aiming to maximize objective sound quality assessment scores by learning to produce output signals that yield high scores. Fu et al. [93] introduced *MetricGAN*, a predictive speech enhancement method that uses a generator network for masking-based prediction of clean speech and employs a discriminator network trained to approximate multiple evaluation metrics such as PESQ and STOI. Similar to traditional GANs, *MetricGAN* focuses on adversarial training where the generator aims to produce clean speech, while the discriminator learns the metric, thus ensuring minimal discrepancy between the enhanced and clean speech scores. However, contrary to what the method name suggests, *MetricGAN* is not a generative method, as it does not involve stochasticity or sampling from the posterior distribution. In a follow-up work, Fu et al. [94] developed *MetricGAN+*, which incorporates several improvements over their original approach. One improvement is the inclusion of noisy speech in the training of the discriminator, alongside enhanced and clean speech, to better align the discriminator with target objective metrics. Additionally, they reuse speech generated from previous epochs for training the discriminator, aiding in the prevention of catastrophic forgetting [95]. For the generator, they introduced a learnable sigmoid function for mask estimation, arguing that a conventional sigmoid is suboptimal due to its uniformity across frequency bands and its maximum value of 1. Instead, a per-frequency learnable sigmoid function offers greater flexibility, enhancing the performance of speech enhancement. De Oliveira et al. [22] presented the *PESQetarian*, which utilizes a differentiable version of PESQ [96] to optimize speech enhancement models. Despite achieving a "state-of-the-art" PESQ score of 3.82 on the VoiceBank-DEMAND dataset [97], their model performed poorly in listening tests, underscoring the dangers of overfitting to specific metrics. This highlights the importance of including diverse metrics in evaluations and confirming findings through listening experiments to ensure comprehensive performance assessments.

Recent predictive methods for speech enhancement leverage cutting-edge advances in deep learning, including *Transformer*-based architectures with self-attention mechanisms and state-space models such as *Mamba* [98]. Chen et al. [99] introduced *DPTNet*, a dual-path *Transformer* network that facilitates context-awareness in speech sequence modeling and efficiently manages long sequences without requiring positional encodings. Subakan et al. [100] proposed *SepFormer*, a *Transformer*-based neural network that uses a multi-scale approach to capture both short and long-term dependencies. Rouard [101] developed *HT Demucs*,

which integrates a cross-domain *Transformer* encoder within a hybrid temporal and spectral *U-Net* architecture, employing sparse attention to extend its receptive field. Chao et al. [102] presented *SEMamba*, a speech enhancement system based on the scalable state-space model *Mamba* [98], employing signal-level distances and metric-oriented loss functions, achieving an impressive PESQ score of 3.69 on the VoiceBank-DEMAND dataset [97].

While predictive methods for speech enhancement are popular due to their straightforward design and ease of training, they risk overfitting to the training data, particularly when the model lacks exposure to a diverse range of acoustic conditions. This generalization issue becomes especially pronounced when there is a mismatch between the training and test data [P4]. Moreover, in challenging scenarios, some masking-based approaches have been shown to introduce unpleasant speech distortions, potentially undermining the benefits of noise reduction [103]. Most predictive models are specifically designed for background noise removal and are not well-suited for non-additive corruptions such as dereverberation or bandwidth extension [P16]. Additionally, end-to-end methods are often regarded as black boxes, complicating their models' interpretability and explainability. To address these concerns, research in explainable artificial intelligence aims to provide greater insight into model decisions, thereby improving understanding, transparency, and trust. Furthermore, explainability forms the foundation for ensuring that such systems meet safety and ethical standards. Despite their unprecedented performance, predictive speech enhancement methods continue to face challenges in scenarios with very low SNRs (e.g., $< -5\,\mathrm{dB}$). They also struggle with edge cases, such as diverse speaking styles like whispering or yelling, children's speech, pathological speech [104], and enhancing languages and dialects, which have distinct phonemes and prosody.

1.5 Generative Methods for Speech Enhancement

Generative models for speech enhancement follow a different learning paradigm compared to their predictive counterparts. Instead of learning a deterministic mapping from noisy to clean speech by minimizing an objective function as in (1.16), they aim to explicitly or implicitly model the conditional distribution of clean speech given the noisy speech. This approach allows explicit generative models to capture the clean speech posterior under the model, while both explicit and implicit generative models can sample from this distribution. The goal of the generative modeling problem for speech enhancement is to learn some sort of mapping $\Phi_\theta : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ parameterized by θ that takes the noisy speech $\mathbf{y} \in \mathbb{R}^N$ and a random vector $\mathbf{z} \in \mathbb{R}^M$ from an easy-to-sample distribution $p(\mathbf{z})$ (e.g., a Gaussian), and outputs an estimated clean speech sample $\hat{\mathbf{x}} \in \mathbb{R}^N$ that resembles a sample from the clean speech data distribution $p_{\text{data}}(\mathbf{x})$. This can be expressed as

$$\hat{\mathbf{x}} = \Phi_\theta(\mathbf{y}, \mathbf{z}), \text{ with } \mathbf{y} \sim p_{\text{data}}(\mathbf{y}), \mathbf{z} \sim p(\mathbf{z}), \tag{1.25}$$

where $p_{\text{data}}(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})\, p_{\text{data}}(\mathbf{x})$ denotes the empirical noisy speech distribution, with $p(\mathbf{y}|\mathbf{x})$ representing the conditional probability that characterizes the corruption process. The goal is to determine the optimal parameters θ^*_{gen} that enable Φ_θ to accurately model the posterior distribution $p_\theta(\mathbf{x}|\mathbf{y})$. This is typically assessed by computing a divergence or distance metric between the distribution of the enhanced speech $p_{\text{enh},\theta}(\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{y})p(\mathbf{y}|\mathbf{x}')p_{\text{data}}(\mathbf{x}')$ and the empirical clean speech data distribution $p_{\text{data}}(\mathbf{x})$. Consequently, the generative modeling

objective can be expressed as,

$$\theta^*_{\text{gen}} = \arg\min_{\theta} D(p_{\text{enh},\theta}(\mathbf{x}),\, p_{\text{data}}(\mathbf{x})), \tag{1.26}$$

where D is a divergence measure between distributions, such as the Kullback-Leibler divergence, Wasserstein distance, or a custom loss function.

Unlike predictive models that focus on minimizing point-wise errors, generative models emphasize capturing the underlying variability and uncertainty in the data. This stochastic nature enables them to produce diverse and high-quality samples, making them particularly well-suited for tasks where both variability and naturalness are critical. In the context of speech enhancement, maintaining naturalness is crucial. Predictive methods, however, have been shown to result in unnatural speech and distortions that can outweigh the benefits of noise reduction [103]. Furthermore, predictive methods struggle with infilling tasks such as bandwidth extension and packet loss concealment [P16]. These pitfalls are addressed by generative methods, which allow sampling from the clean speech distribution and thus maintain high-quality and naturalness.

A distinction is made between *unsupervised generative speech enhancement*, which involves training the model without the need for paired clean and noisy speech data, often relying solely on clean speech data, and *supervised generative speech enhancement*, where the model is trained using paired clean and noisy speech data. In unsupervised generative speech enhancement using only clean speech data, the model learns the prior distribution of clean speech, denoted as $p_\theta(\mathbf{x})$, whereas in supervised generative speech enhancement, the model learns the conditional distribution of clean speech given the noisy mixture, denoted as $p_\theta(\mathbf{x}|\mathbf{y})$. With the learned distribution $p_\theta(\mathbf{x}|\mathbf{y})$, it becomes possible to directly sample multiple plausible clean speech signals given the noisy mixture. On the other hand, in unsupervised generative speech enhancement, $p_\theta(\mathbf{x})$ can serve as an explicit prior model for clean speech, such as in a Bayesian framework.

The term *generative model* originally stems from the field of statistical classification, where it refers to a statistical model that captures the joint probability distribution of the observed and target variables. In statistical classification, generative models represent one of the two main approaches to developing classifiers [105]. *Generative classifiers* learn a model of the joint probability $p(x, y)$ of the input x and the label y, and make their predictions by using Bayes rules to calculate $p(y|x)$, and then picking the most likely label y. In contrast, *discriminative classifiers* model the posterior $p(y|x)$ directly, or learn a direct mapping from input to the class label. Typical examples for generative classifiers include the naive Bayes classifier and linear discriminant analysis, while discriminative classifiers include logistic regression and support vector machines [106].

In generative modeling, on the other hand, the goal of a generative model is to implicitly or explicitly model the distribution of training data, enabling the generation of new samples that resemble the training data. This process often involves learning a joint distribution $p(\mathbf{x}, \mathbf{z})$ of observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$. The latent variables in a probabilistic model do not necessarily have a direct physical interpretation but may be introduced simply to allow a more complex joint distribution to be constructed from simpler components [106]. To generate new samples, the model can sample from the latent variables $\mathbf{z}$ and then generate the observed variables $\mathbf{x}$ given the latent variables. This process is known as *ancestral sampling* in graphical models, where the graphical model captures the causal process by which the

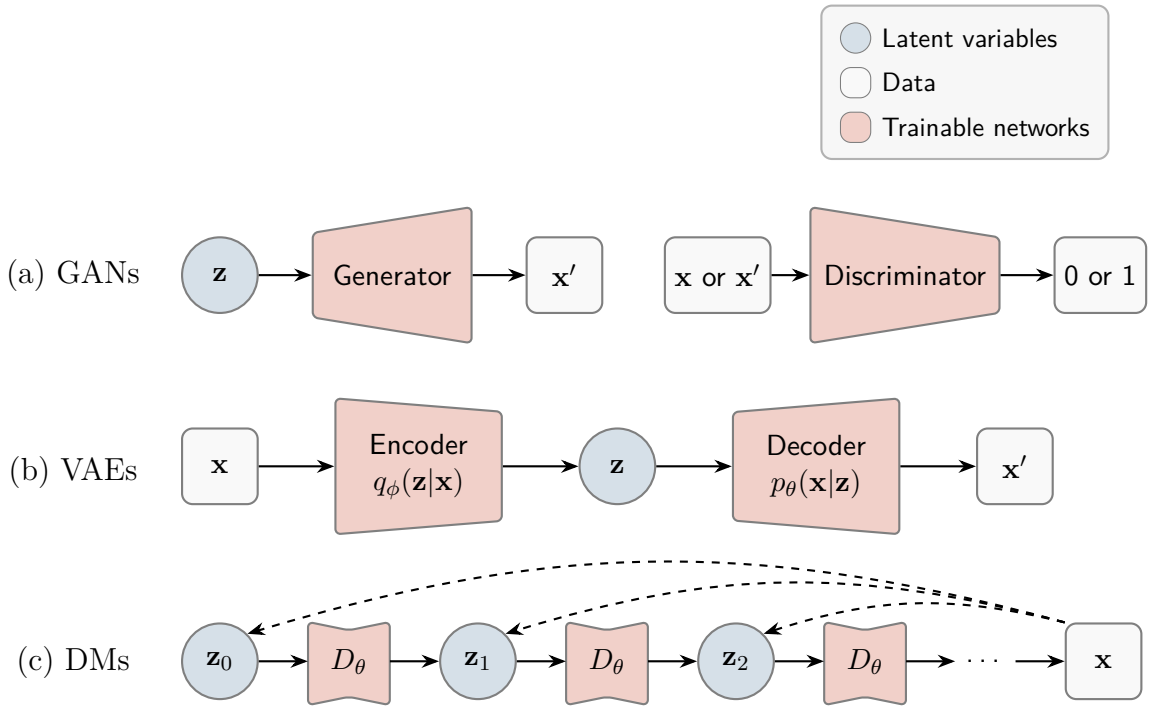


Figure 1.5: Schematic overview of deep generative models, showcasing (a) generative adversarial networks (GANs), which utilize an adversarial training approach involving a generator and a discriminator; (b) variational autoencoders (VAEs), which employ an encoder and a decoder modeling the posterior and likelihood, respectively; and (c) diffusion-based generative models (DMs), which are based on a Markov chain and utilize a denoiser model D_θ to progressively remove Gaussian noise from the data. During training, the perturbed data is directly sampled using the perturbation kernel, as indicated by dashed lines.

observed data was generated [107]. Probabilistic models like Bayesian networks and Markov random fields leverage these methods to generate samples that reflect the causal process by which the observed data was generated [107]. Other generative models include hidden Markov models [108], which are widely used for temporal data analysis; Gaussian mixture models [109], which model data as a mixture of multiple Gaussian distributions; and methods based on Markov chain Monte Carlo (MCMC) [110], which use stochastic simulations to approximate complex probability distributions.

Deep generative models extend the concept of generative modeling to deep learning, leveraging the representational power of DNNs [111], [112]. Early works in deep generative modeling include deep belief networks [53] and restricted Boltzmann machines [113], [114]. Deep belief networks have been used for phone recognition [115], and acoustic modeling [116]. More recent deep generative models include GANs [56], variational autoencoders (VAEs) [117], [118], and diffusion-based generative models [119], [120], [121]. Figure 1.5 presents a schematic overview of these three popular deep generative models. Although a variety of generative models are available, they often fall short in simultaneously meeting all three key criteria: high sample quality, mode coverage, and fast sampling. This limitation, known as the *"generative learning trilemma"* [122], means that most models typically excel in some areas while lacking in others. For example, denoising diffusion models achieve impressive sample quality and diversity, but their slow sampling can hinder their use in practical applications. GANs are known for their

high sample quality and fast sampling, but they often struggle with mode coverage, leading to the generation of samples that are not representative of the training data. VAEs are known for their mode coverage and fast sampling, but they often produce samples of lower quality compared to GANs and diffusion models. Below, we provide a brief overview of the three types of deep generative models: GANs, VAEs, and diffusion-based generative models.

GANs consist of two neural networks, a generator and a discriminator, that are trained in an adversarial manner. In a minimax-game, the generator takes a Gaussian vector $\mathbf{z}$ and learns to produce samples $\mathbf{x}'$ that mimic the training data, while the discriminator learns to distinguish between real and generated samples [56]. VAEs are based on variational inference and consist of two main components: an encoder network and a decoder network. The encoder network parameterized by ϕ maps the input data $\mathbf{x}$ to the first two moments of a latent variable, modeling the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$. The decoder network parameterized by θ maps the latent variable back to a reconstruction $\mathbf{x}'$ of the input data, thus modeling the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$. The model is trained by maximizing the evidence lower bound, which is a lower bound on the log-likelihood of the data [117], [118].

Diffusion-based generative models [119], [120], [121], [123] generate data through an incremental process

$$\mathbf{z}_{i+1} \leftarrow D_\theta(\mathbf{z}_i, \sigma_i), \quad \mathbf{z}_0 \sim \mathcal{N}(0, \sigma_0 \mathbf{I}), \quad i = 0, \ldots, K-1, \tag{1.27}$$

where a Gaussian vector $\mathbf{z}_0$ is progressively transformed into a data sample $\mathbf{x} := \mathbf{z}_K$ by estimating the latent variable at a lower noise level containing more information about the data. Here, K represents the total number of steps in the incremental process, where σ_i denotes a decreasing noise level, such that $\sigma_{i+1} < \sigma_i$. A *denoiser model*, denoted as $D_\theta : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$, is used to estimate the latent variable at a lower noise level given the current latent variable and noise level. The output from the neural network can either be an estimate of the data ($\mathbf{x}$-prediction) or an estimate of the Gaussian noise ($\boldsymbol{\epsilon}$-prediction), serving to predict the latent variable at a lower noise level. Unlike typical latent variable models, such as the VAE, diffusion-based generative models employ a fixed inference procedure $q(\mathbf{z}_0, \ldots, \mathbf{z}_{N-1}|\mathbf{x})$ for the latent variables instead of a trainable one, and they feature relatively high-dimensional latent variables. During training, a latent variable $\mathbf{z}_i$ is sampled from the *perturbation kernel* simulating the *forward process* (indicated by dashed lines in Figure 1.5), such that it is given by

$$\mathbf{z}_i = \alpha_i \mathbf{x} + \sigma_i \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1.28}$$

where α_i and σ_i define the *noise schedule*. Several approaches exist of how to define the noise schedule, including a *variance-preserving* schedule if $\alpha_i^2 + \sigma_i^2 = 1$, or a *variance exploding* schedule with exploding variance for $i \to 0$ such that $\sigma_0^2 \gg \alpha_0^2$ where $\alpha_i = 1$.

In the following sections, we explore generative speech enhancement methods based on deep generative models, exploring both unsupervised and supervised approaches. In Subsection 1.5.1, we focus on unsupervised generative speech enhancement where unsupervised models learn the prior distribution of clean speech. Subsequently, in Subsection 1.5.2, we discuss supervised generative speech enhancement, where models learn the conditional distribution of clean speech given the noisy mixture.

1.5.1 Unsupervised Generative Speech Enhancement

In the field of machine learning, *unsupervised learning* refers to a class of algorithms designed to identify patterns within datasets that do not have any labels [111]. Unlike *supervised learning*, which utilizes labeled or paired datasets to establish relationships between inputs and outputs, the goal of unsupervised learning is to reveal the underlying structure of the data and develop useful representations [124]. Unsupervised learning is typically applied to tasks such as clustering, which involves grouping data points into clusters based on similarity [106]; dimensionality reduction, which compresses data while preserving its essential structure [125]; and generative modeling, which aims to synthesize new data examples that are statistically similar to the training data [124]. As discussed in Section 1.5, some generative models explicitly define the probability distribution over the input domain, allowing the generation of new examples by sampling from this distribution. In contrast, other models, known as implicit generative models, focus on learning a mechanism to generate new examples without explicitly modeling the distribution. While unsupervised generative models can be used in their original form, they are more often adapted within Bayesian inference frameworks or fine-tuned for specific downstream applications.

A variety of unsupervised generative models have been developed to model the distribution of speech data. Blaauw and Bonada [126] introduced the use of a VAE for modeling frame-wise spectral envelopes of speech signals. They considered several downstream tasks, including formant manipulation through interpolation between latent representations of different vowels, and unconditional speech generation using a Gaussian diffusion process in latent space. Hsu et al. [127] utilized a convolutional VAE to model the distribution of magnitude spectra and filter bank representations of speech signals. They explored random sampling from the latent space and modifying phone and speaker attributes by transforming the latent representations of speech segments. In subsequent research, Hsu et al. [128] proposed a factorized hierarchical VAE, which learns disentangled representations for sequence-level and segment-level attributes. They demonstrated their ability to transform speaker characteristics or linguistic content by manipulating different sets of latent variables. Kong et al. [129] presented *DiffWave*, a diffusion-based generative model operating in the time domain, which is assessed on conditional and unconditional speech generation tasks. Iashchenko et al. [130] investigated various architectures for unsupervised speech enhancement and determined that a time-frequency domain architecture outperforms multiple time-domain architectures. De Oliveira et al. [131] introduced an unconditional diffusion model trained solely on clean speech for assessing speech quality. They demonstrated that the quality of a speech utterance can be evaluated by estimating the likelihood of a corresponding sample in the terminating Gaussian distribution.

In speech enhancement, unsupervised methods utilize models trained without requiring paired noisy-clean speech examples. These methods can involve using only clean speech, only noisy speech, or a combination of noisy speech files and noise samples. Unsupervised models trained on clean speech only can be employed to regularize the enhanced speech to make it more speech-like, such as by serving as a clean speech prior in Bayesian inference. They can be utilized to estimate parameters related to clean speech, such as the power spectral density, which in turn can inform statistical estimators like the Wiener filter in (1.7). This approach is based on the assumption that, while noise variability is difficult to model, clean speech signals exhibit a highly structured nature. Overall, unsupervised speech models provide two primary

advantages over supervised models. Firstly, they can be trained with large-scale clean speech datasets, as task-specific information like noisy speech is integrated only during inference or fine-tuning. Secondly, they tend to be more robust to various noise conditions because they do not depend on specific conditioning during training.

Smaragdis and Venkataramani [132] demonstrated that a linear autoencoder can serve as an analog to NMF and applied this concept to source separation, achieving significantly improved performance over standard NMF [35]. Bando et al. [133] proposed an unsupervised speech enhancement method based on a statistical framework for variance modeling in the time-frequency domain. They employed a VAE trained on clean speech power spectra to model the prior distribution of the frequency-dependent speech variance. This speech prior was combined with an untrained NMF model for estimating the time- and frequency-dependent noise variance to enhance noisy speech using a Bayesian inference algorithm. Specifically, they iteratively sampled the posterior distributions of the latent variables from the VAE and the NMF parameters using an MCMC algorithm, updating the signal variances necessary for forming a Wiener filter. Building on this approach, Leglaive et al. [134] introduced a time-dependent gain to improve robustness against the loudness variations of training examples. In addition, they proposed an iterative Monte Carlo expectation-maximization (MCEM) optimization method in which only the latent variables of the VAE are treated as random variables, while the NMF parameters remain unknowns. During each iteration, the Metropolis-Hastings algorithm [135] is used to evaluate the posterior distribution of the latent variables, and the unknown NMF parameters are updated with multiplicative update rules [35]. To improve the computational efficiency, Pariente et al. [136] proposed a variational expectation-maximization (EM) algorithm that factorizes the joint posterior distribution of the latent variables using mean-field variational inference principles.

In the original VAE-NMF speech enhancement framework [133], [134], the VAE is trained to reconstruct single spectrogram frames without modeling temporal dependencies. To alleviate this limitation, in [P1], we proposed to use a stochastic temporal convolution network that combines both hierarchical and temporal dependencies of latent variables (see Section 3.1.1). We show that our proposed approach outperforms the non-sequential VAE in terms of SI-SDR, as well as instrumental speech quality and intelligibility. Similar findings have been shown by Leglaive et al. [137], using a recurrent VAE to model the clean speech variance. Bie et al. [138] expanded on this work by introducing the use of a dynamical VAE in the unsupervised speech enhancement framework. During inference, the parameters are updated using a variational EM method, with the VAE encoder being fine-tuned via stochastic gradient descent. For a comprehensive review on dynamical VAEs, we refer the reader to [139]. Nortier et al. [140] proposed to replace the VAE with an unconditional diffusion model in the statistical framework. When evaluated under both matched and mismatched conditions, their method outperforms the dynamic VAE across nearly all standard instrumental evaluation metrics.

Despite being theoretically robust against generalization issues in various noisy environments, the performance of unsupervised speech enhancement methods often lags behind that of supervised methods trained with paired data. This is arguably due to the increased complexity in the conditioning design and training processes of supervised methods. To mitigate this, several works have been proposed that enhance the unsupervised learning approach with elements of supervised learning. Bando et al. [141] suggested training the VAE using noisy-clean speech pairs and stabilized the training by adding an encoder output branch to estimate a speech mask in a multi-task manner. Fang et al. [142] introduced a noise-aware VAE with a two-stage training process. In the first stage, an unsupervised VAE is trained on

clean speech only. In the second stage, a noise-aware encoder is optimized by minimizing the Kullback–Leibler divergence between the posterior distribution defined by the encoder of the unsupervised VAE using clean speech input and the posterior distribution modeled by the noise-aware encoder with noisy speech input. Another approach to incorporating supervision into the VAE-NMF framework is to guide the VAE with a supervised classifier trained separately on noisy speech. We have proposed this idea in [P2], where we condition the VAE on a high-level categorical variable describing the speech signal (see Section A.1.1). In the experiments, we considered voice activity and the ideal binary mask as labels. We demonstrated that when using the ideal binary mask as a label estimated by a simple DNN, the guided VAE outperformed both the standard VAE and a DNN-based Wiener filter, especially at low input SNRs. However, the label may not be completely disentangled from other latent variables. To address this, in [P3], we proposed using an adversarial training scheme for VAEs to disentangle the label from the other latent variables (see Section 3.2.1). This involves using a discriminator that competes with the VAE's encoder during training. Additionally, we employ an extra encoder to estimate the label for the VAE's decoder, which is critical for achieving disentanglement.

Other approaches employ unsupervised speech models as priors for clean speech within a statistical framework to explicitly model the posterior of clean speech given the noisy mixture using Bayes' rule. Qian et al. [143] proposed a Bayesian speech enhancement framework in the time domain, which predicts clean speech samples by estimating the prior distribution and the likelihood function of clean speech using *WaveNet*-like architectures. They decompose the posterior distribution into a causal prior model, which estimates the probability of a clean speech sample given the previous enhanced samples, and a non-causal likelihood model, which considers the past enhanced samples and the noisy samples within the temporal context of the current sample. Moliner et al. [144] introduced an unsupervised dereverberation method that integrates an unconditional diffusion model trained on anechoic speech with a model-based RIR estimator. The approach utilizes diffusion posterior sampling [145], which applies Bayes' rule to express the score of the posterior as the sum of the unconditional score and the likelihood score.

Another line of research in unsupervised speech enhancement follows the approach of training models using only noisy speech or adding noise to noisy speech. Xiang and Bao [146] proposed to use a cycle-consistent GAN called *CycleGAN* [147] trained on non-paired clean and noisy speech. Their method incorporates two separate encoders, one for clean speech and one for noisy speech. It also includes four generators: one to predict the ideal ratio mask of clean speech, another for the log-power spectrum of clean speech, a third for the magnitude spectrum of noisy speech, and a fourth for the log-power spectrum of noisy speech. Four discriminators are employed to differentiate between real and generated speech features. All models are trained simultaneously using multi-objective learning with distinct loss functions. Building upon their work, Yu et al. [148] improved performance by incorporating attention mechanisms into the *CycleGAN* architecture. Wisdom et al. [149] proposed mixture invariant training that requires only single-channel acoustic mixtures. It involves creating examples by combining existing mixtures. The model then separates these into a variable number of sources, enabling the separated sources to be remixed to closely approximate the original mixtures. Fujimura et al. [150] introduced a noisy-target training approach, where the training target is noisy speech, and the input signal is generated by combining this noisy speech with additional noise. Essentially, a DNN is trained to predict a noisy speech signal from an even more degraded signal.

Although unsupervised speech enhancement methods—both predictive and generative—are conceptually more robust to unknown noise conditions, they often underperform compared to supervised approaches. Nevertheless, the exploration of innovative unsupervised learning techniques for speech enhancement remains a vibrant area of research. Bayesian models, in particular, stand out for their interpretability and their ability to leverage well-established estimators with known optimal solutions. Additionally, unsupervised pretraining can serve as an effective initialization strategy for supervised models, offering a robust foundation for further learning in a supervised framework. On the other hand, supervised methods continue to advance by scaling up model sizes, leveraging larger datasets, and employing more sophisticated data augmentation strategies.

1.5.2 Supervised Generative Speech Enhancement

In comparison to unsupervised generative speech enhancement, supervised methods leverage paired noisy-clean speech data to model the conditional distribution of clean speech given the noisy mixture. By training on such paired data, supervised generative models learn to capture the conditional distribution, enabling them to generate multiple valid, clean speech estimates by sampling from this conditional distribution. This differs from predictive methods, which learn a deterministic mapping from noisy to clean speech and therefore produce a single output for each input. The stochastic nature of generative models, however, allows them to better represent the variability and uncertainty in speech signals, often resulting in improved naturalness and diversity in the generated samples. Another aspect of supervised generative approaches is their ability to incorporate explicit loss functions to improve perceptual quality, such as spectral distance metrics or adversarial losses, which guide the model to produce more natural-sounding speech. Additionally, supervised methods benefit from the growing availability of large-scale speech and noise datasets, enabling them to generalize well across diverse noise conditions and speech characteristics.

Sampling multiple valid estimates of clean speech from the posterior distribution of clean speech given the noisy mixture is a rather recent development in speech enhancement research [P4]. Traditional statistical model-based approaches for speech enhancement often considered only the mean of the distribution as an estimate [151], which is equivalent to the output of a deterministic model. This recent development has been driven by the success of generative models in other domains, such as image generation, where the ability to sample from the data distribution has led to significant improvements in the quality and diversity of generated images [56], [120]. In speech enhancement, this capability has the potential to improve the robustness and naturalness of the enhanced speech, while also providing an elegant solution to inpainting tasks such as bandwidth extension or packet loss concealment [P7].

Early works on supervised generative speech enhancement that model the conditional distribution of clean speech given the noisy mixture include concepts from deep generative models such as VAEs and GANs. Pascual et al. [152] introduced *SEGAN*, a GAN-based speech enhancement model that operates in the time domain. A generator model is designed as a fully convolutional autoencoder. The generator's encoder processes the noisy speech and compresses it into a compact representation. This representation is then combined with a latent vector sampled from a Gaussian distribution and fed into the decoder, which produces the enhanced speech output. The discriminator is a binary classifier trained to differentiate between clean and enhanced speech, considering the latter as "fake". By employing the stan-

dard GAN adversarial training objective, the generator is guided to produce enhanced speech that is indistinguishable from clean speech. In a follow-up work, Pascual et al. [153] proposed training their model on a broader range of speech enhancement tasks, including denoising, dereverberation, declipping, packet loss concealment, and bandwidth extension. Additionally, they introduced an adversarial acoustic regression loss to enhance feature extraction at the discriminator. They employed a two-step adversarial training schedule, serving as a warm-up and fine-tuning process. Li et al. [154] introduced a conditional GAN designed for speech enhancement in the time-frequency domain. They investigated various network architectures for the generator, such as a feedforward neural network, an RNN, and a CNN. Their findings indicated that the CNN generator surpassed the other architectures regarding speech quality and intelligibility. Donahue et al. [155] explored the use of GANs for speech enhancement and dereverberation operating on log-Mel filterbank spectra. They found that using log-Mel filterbank spectra as input representation improved the noise robustness of their model as a preprocessor for ASR compared to using raw waveforms.

Chien and Kuo [156] proposed a VAE-based model for speech separation that utilizes an RNN architecture and operates in the time-frequency domain. The encoder, responsible for recognizing hidden variables, and the decoder, tasked with generating regression outputs, were guided by a variational learning algorithm designed to maximize the variational lower bound of the logarithm of the conditional likelihood. In a supervised learning setting, the model estimates the separated speech signals from an input mixture of two speakers. Braithwaite and Kleijn [157] proposed using a variance-constrained VAE for speech enhancement in the time domain. Unlike the standard VAE, which learns both the mean and variance of a Gaussian distribution for the latent variables, the variance-constrained VAE learns only the mean, while restricting the variance of the latent variables to a predefined value. Strauss and Edler [158] introduced a time-domain generative speech enhancement model based on a normalizing flow model [159]. They showed that a straightforward nonlinear μ-law companding technique serves as an effective preprocessing tool to enhance the performance of the model.

Recently, diffusion-based generative models have gained popularity in speech enhancement [P5]. Lu et al. [160] proposed to use a diffusion probabilistic model for speech enhancement that operates in the time domain. Their model is based on *DiffWave* [129] and employs a modified reverse process, which adds noisy speech at each time step to the mean prediction. In subsequent research, Lu et al.[161] introduced *CDiffuSE*, a conditional diffusion model for speech enhancement that operates in the time domain. They proposed incorporating real noise recordings in the diffusion process by linearly interpolating between clean and noisy speech along the process. However, this choice of linear interpolation requires the denoiser network to explicitly estimate a portion of environmental noise at each step in the reverse process, giving the resulting objective function characteristics of a predictive learning task. Additionally, the authors implemented a post-processing step by blending the enhanced waveform with the original noisy signal in a 4:1 ratio. This was intended to recover the high-frequency components of speech in the final enhanced waveform but resulted in remaining background noise.

To address these limitations, in [P6], we introduced a diffusion-based speech enhancement model that operates in the time-frequency domain. Our approach involves the use of continuous-time diffusion models [121], which are based on stochastic differential equations (SDEs) to represent both the forward and reverse diffusion process. Specifically, we proposed modeling the corruption of clean speech by environmental background noise directly within the forward diffusion process by introducing a drift term, which causes the process mean to transition from

clean to noisy speech. In our follow-up work [P4], we introduced *SGMSE+*, a diffusion-based speech enhancement model with significantly improved performance by adopting the improved *noise conditional score network (NCSN)++* architecture [121]. We demonstrated that only 30 diffusion steps are needed to produce high-quality clean speech estimates, achieving state-of-the-art performance in generative speech enhancement based on both instrumental and subjective evaluation. Furthermore, we demonstrated that our proposed method is also effective for dereverberation, showing that it is not restricted to removing additive background noise.

Building on our pioneering work [P4], [P6], we have further explored the potential of diffusion-based speech enhancement in various directions. In [P7], we proposed a causal processing scheme for diffusion-based speech enhancement to address the restoration of speech communication artifacts, including denoising, dereverberation, declipping, packet loss concealment, bandwidth extension, and codec artifact removal. To ensure causal processing, we modified the *NCSN++* architecture from our previous work and eliminated non-causal normalization techniques. This work is based on [P8], our submission to the "International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2023 Speech Signal Improvement Challenge" [162], where our method achieved fifth place in the non-real-time track, despite competing against industry leaders with access to more data and resources. In our speech enhancement method called *StoRM* [P15], we proposed combining a predictive model with a generative model by using the estimate from the predictive model as the initial condition for the reverse process. We showed that the proposed approach alleviates generative vocalizing and breathing artifacts while producing very high-quality samples, even in adverse conditions. Furthermore, we showed that this approach enables the use of lighter sampling schemes with fewer diffusion steps without sacrificing quality. In [21], we introduced an alternative diffusion process utilizing a Brownian bridge [163]. We demonstrated that this approach reduces the mismatch between the terminal distribution of the forward process and the starting distribution for reverse sampling. As a result, it improves performance with only half the iteration steps and eliminates the need to tune one hyperparameter. In [P11], we explored the Schrödinger bridge [164] for speech enhancement and explained the difference to score-based generative models. Additionally, we proposed a perceptual loss function specifically designed for the Schrödinger bridge framework, showing that it enhances performance and improves the perceptual quality of the enhanced speech signals.

Our research has been a driving force in advancing diffusion-based speech enhancement, bridging the performance gap between predictive and generative methods. We demonstrated that diffusion-based speech enhancement can achieve state-of-the-art results in both instrumental and subjective evaluations, highlighting its potential for practical applications. This has been further confirmed by the preference for generative methods in contemporary challenge series, such as the "Speech Signal Improvement Challenge" [162] and the "URGENT Challenge" [165], where participants are tasked with developing a single speech enhancement system capable of adaptively handling input speech with various corruptions across different acoustic environments. The recent development of diffusion-based models has spurred dedicated workshops and special sessions on generative speech enhancement at major conferences, such as the Institute of Electrical and Electronics Engineers (IEEE) ICASSP and International Speech Communication Association (ISCA) Interspeech. These events have attracted a growing number of researchers and practitioners, fostering a vibrant research community around generative speech enhancement. In an attempt to find the underlying scheme of classes of current popular speech enhancement methods, we propose a taxonomy illustrated in Figure 1.6.

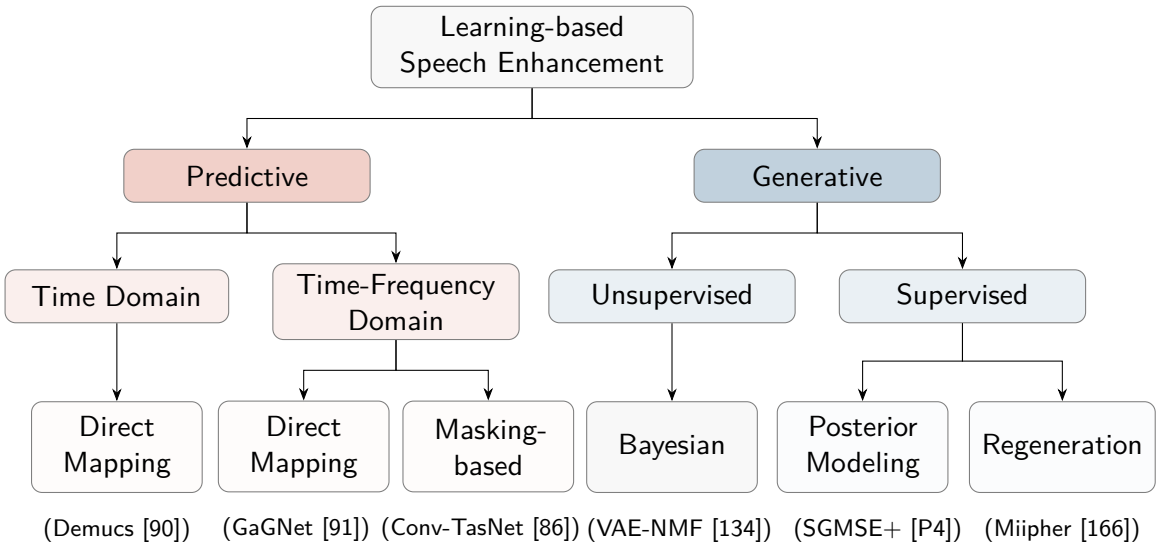


Figure 1.6: Taxonomy of learning-based speech enhancement. We begin by distinguishing between generative and predictive learning paradigms. Predictive methods are further divided into time domain and time-frequency domain approaches, while generative methods are categorized into unsupervised and supervised learning frameworks. Additional subdivisions highlight the specific training paradigms and inference strategies utilized. Representative methods from each class are included below each category.

In this taxonomy, we first distinguish between generative and predictive paradigms, further dividing predictive methods into time domain and time-frequency domain approaches, and generative methods into unsupervised and supervised learning frameworks. Further subdivisions are made based on the specific training paradigms and inference strategies employed by the methods. In addition, we denote one popular method from each class to provide a representative example.

Several works have adapted or extended the *SGMSE+* model [P4] based on our publicly available code repository[1]. Peer et al. [167] proposed adapting *SGMSE+* for STFT phase retrieval, demonstrating that the diffusion approach is well-suited for this task, outperforming both classical and modern methods. Scheibler et al. [168] introduced score-based generative models for single-channel source separation by initiating the forward process with the separated sources and letting it converge to a Gaussian distribution centered on their mixture. Kamo et al. [169] explored the task of target speech extraction by conditioning the extraction process on a clue specific to the target speaker. Additionally, they introduced ensemble inference to mitigate potential errors in the extraction process that may arise from the diffusion process. Nakatani et al. [170] proposed conditioning diffusion models on multiple estimates, including those derived from a predictive model and weighted prediction error (WPE) dereverberation. They demonstrated enhanced performance when tested on a reverberant and noisy speech dataset using a two-microphone setup. Kimura et al. [171] expanded *SGMSE+* into a multi-input multi-output (MIMO) approach named *multi-stream SGMSE* (*mSGMSE*). This method targets the estimation of clean speech signals from multiple microphones, preserving spatial cues, from noisy and reverberant speech captured by a distant microphone array. They

[1] `https://github.com/sp-uhh/sgmse`

introduced techniques to make *mSGMSE* computationally efficient and adaptable to various or unknown array geometries for general speech enhancement applications. Shi et al. [172] proposed adding a second decoder that shares the same encoder as the original decoder. This additional decoder is trained to predict the clean speech spectrogram using an MSE loss. They explored two inference strategies: In the first strategy, the second decoder is ignored during inference and serves solely as a regularizer during training. In the second strategy, the output of the second decoder is utilized during the initial and final reverse steps by linearly interpolating between the outputs of the first and second decoders. Their experiments on the VoiceBank-Demand dataset [97] indicate that incorporating the second decoder results in slightly higher PESQ values while significantly reducing the number of required diffusion steps. Lay et al. [23] tackled the issue of slow speech sampling through a two-stage training approach. In the first stage, they trained the diffusion model in the conventional manner. In the second stage, they fine-tuned the score model by computing the enhanced signal through solving the reverse process and comparing the resulting estimate to the clean speech target using a predictive loss.

Other approaches have explored various diffusion formalisms, input representations, and network architectures. Gonzales et al. [173] utilized the advanced *Elucidating the Design Space of Diffusion-Based Generative Models (EDM)* framework [174], using a change of variables to account for the SDE satisfied by the environmental noise, resulting in the necessary linear affine drift term. Scheibler et al. [175] developed *UNIVERSE++*, a speech enhancement method that utilizes score-based generative models and adversarial training in the time domain. This work builds upon the foundation laid by Serra et al. [176], one of the pioneering large-scale diffusion-based methods for speech enhancement. Trained on 55 different types of distortions, the model significantly outperformed state-of-the-art methods in a subjective test conducted by expert listeners. The authors advocated for a "universal approach" to speech enhancement, suggesting that the task could be reframed as a generative one. Liu et al. [177] introduced *SpeechFlow*, a foundational model tailored for various speech generation tasks, employing flow matching [178] for generative modeling along with masked audio prediction as a self-supervised pretraining strategy. Building on this, Vyas et al. [179] proposed *Audiobox*, a unified audio generation system guided by natural language prompts. Jukic et al. [180] explored the Schrödinger bridge approach for speech enhancement, a generative model that optimally transforms one probability distribution into another [164]. In contrast to *SGMSE+*, this method enables the generative process to begin directly from the noisy input and uses a data prediction loss.

Another area of research in generative speech enhancement focuses on speech regeneration using techniques from speech synthesis and language modeling. Polyak et al. [181] proposed a speech regeneration model that operates in the time domain. Their approach incorporates a speech synthesis model, an ASR model, a pitch extraction model, and a loudness feature. Koizumi et al. [166] introduced *Miipher*, a speech restoration model designed to produce high-quality clean speech from enhanced *w2v-BERT* [182] features while conditioning on speaker embeddings and phoneme features. Wang et al. [183] presented a speech enhancement paradigm that integrates discrete tokens and leverages language models, comprising three stages: encoding, modeling, and decoding. The encoder converts continuous waveform signals into discrete tokens using pre-trained self-supervised learning (SSL) models and a k-means tokenizer. Language models then capture comprehensive contextual information within these tokens. Finally, a de-tokenizer and *HiFi-GAN* [184], used as a neural vocoder, restores the tokens into enhanced speech. Yang et al. [185] proposed generating clean speech as

discrete codec tokens, conditioned on input speech features. They provide insights into the optimal generation scheme for enhancement among parallel prediction, autoregression, and masking, highlighting the benefits of conditioning on both pre-trained and jointly learned speech features. Xue et al. [186] introduced a low-latency speech enhancement method via speech token generation, which models clean speech through acoustic codes of a neural speech codec and generates speech codes conditionally on past noisy frames in an autoregressive manner. Additionally, they proposed an explicit alignment approach to synchronize noisy frames with the generated speech tokens, thereby enhancing robustness and scalability for various input lengths.

Recent advancements in generative speech enhancement are bringing a fresh perspective to the field, highlighting the advantages of generative methodologies compared to predictive approaches. Understanding the differences between these approaches is crucial, as they exhibit distinct behaviors and capabilities. In [P16], we compared *SGMSE+* to a predictively trained model with the identical network architecture, across three restoration tasks: speech denoising, dereverberation, and bandwidth extension. Our observations indicate that the generative approach outperforms the predictive one on all tasks, with significant advantages in handling non-additive distortions, such as those encountered in dereverberation and bandwidth extension. In [19], we assessed the performance of the same speech enhancement model under both predictive and generative paradigms across various metrics. Our findings reveal that intrusive and non-intrusive measures demonstrate differing correlations for each paradigm. To extend our evaluations further, we introduced the *EARS* (Expressive Anechoic Recordings of Speech) dataset in [P13], a high-quality 48 kHz speech dataset featuring 100 hours of clean, anechoic speech. We benchmarked several speech enhancement methods on this dataset, using a set of instrumental metrics and conducting a listening test with 20 participants. *SGMSE+* emerged as the best-performing algorithm among those tested.

Despite their successes, generative models encounter challenges similar to those of predictive models, such as managing extremely low SNRs and ensuring robustness against unfamiliar noise types. However, the nature of their distortions is distinct, as generative models often produce artifacts that sound more natural, like vocalizing and breathing, which present unique challenges for generative speech enhancement. Additionally, the iterative nature of diffusion models poses a drawback, particularly regarding the slow sampling speed during the reverse inference process. Nonetheless, supervised generative methods continue to set benchmarks in speech enhancement, driving progress in both research and practical applications.

1.6 Audio-Visual Speech Enhancement

Humans communicate and make sense of the world through various sensory channels, with audio and visual modalities playing a significant role in integrating information to understand their environment [187]. Despite the differences between audio and visual sensations, we can maintain a coherent and unified perception. This crossmodal ability provides significant behavioral advantages, particularly when one sense is impaired [188]. Combining sensory inputs across different modalities can greatly enhance the detection and discrimination of external stimuli and significantly improve response times [189], [190]. Several studies involving linguistic information have indicated the involvement of certain brain areas, including regions within the superior temporal sulcus, in the perception of audio-visual speech. Initial studies using magnetoencephalography (MEG) [191] and functional magnetic resonance imaging (fMRI) [192] demonstrated that visual speech, such as lip-reading, can activate areas of the

auditory cortex that were previously thought to be dedicated solely to processing sound-based signals. In a follow-up fMRI study, Calvert et al. [193] found that both seeing and hearing a speaker leads to increased activation in sensory-specific cortices, either auditory or visual, compared to the response to each modality individually. They suggested that the heightened activity in these modality-specific areas might represent the physiological basis for the perceived enhancements in hearing when observing a speaker's lip and mouth movements [194], as well as the improved visual localization of sound sources, such as distinguishing who is speaking in a crowded room, when both auditory and visual speech cues are available [195]. Behavioral studies suggest that similar principles might apply to the crossmodal perception of content information. For instance, speech comprehension is enhanced when the speaker is visible as well as audible [194], [196], yet it becomes impaired if the lip and mouth movements do not align semantically with the spoken words [197].

The McGurk effect [198] and the ventriloquist's illusion are compelling examples of crossmodal integration, where sensory information from different modalities is combined to form coherent perceptions [199]. The McGurk effect illustrates an interaction between auditory and visual components in speech perception, where mismatched stimuli can lead to the perception of an entirely different phoneme. For instance, when a speaker says "ba" while mouthing "ga," listeners may perceive "da". This phenomenon underscores the importance of visual cues in interpreting speech, highlighting how our brains integrate diverse sensory inputs related to stimulus identity. In contrast, the ventriloquist's illusion demonstrates the integration of spatial information across senses. During a ventriloquism performance, the audience perceives the voice as emanating from the puppet rather than the performer, as visual cues from the puppet override the actual auditory source. This reflects how crossmodal localization combines spatial coordinates across senses. Research suggests that different cortical networks may be involved in integrating inputs related to stimulus identity and location, as evidenced by the distinct cognitive constraints observed in these phenomena [200]. For instance, while desynchronization of auditory and visual inputs can disrupt the ventriloquist's illusion, the McGurk effect can persist even with a 180 ms lag between the onset of auditory and visual components [200]. This indicates a distinction in how sensory invariant information, such as temporal onset, duration, and correlated audio-visual frequency–amplitude information, is processed during audio-visual speech integration.

The above-mentioned characteristic of human crossmodal integration lays the motivation for developing computer systems capable of *multimodal learning*, especially in aligning and fusing audio and video inputs. The term *modality* typically refers to different types of data such as text, audio, images, or video, and a research problem is characterized as *multimodal* when it includes at least two such modalities [201]. Multimodal machine learning seeks to develop models capable of processing and integrating information from multimodal data, such as audio-visual speech. This objective is illustrated in Figure 1.7, where a speech utterance, the associated lip movements, a speaker portrait, and the written text are combined into a multimodal space. To effectively integrate and leverage information from these multiple modalities, robust fusion techniques are essential. Traditionally, fusion techniques are categorized into two types: early fusion and late fusion [202]. Early fusion consists of combining the different modalities into a joint representation at the feature level. Late fusion, on the other hand, consists of combining the modalities only at the decision level, after which the acoustic and visual information is processed separately with two different models. Each fusion technique has its advantages and disadvantages, and the choice between them depends on the specific task and the characteristics of the data. Early fusion is often preferred when the modalities are highly

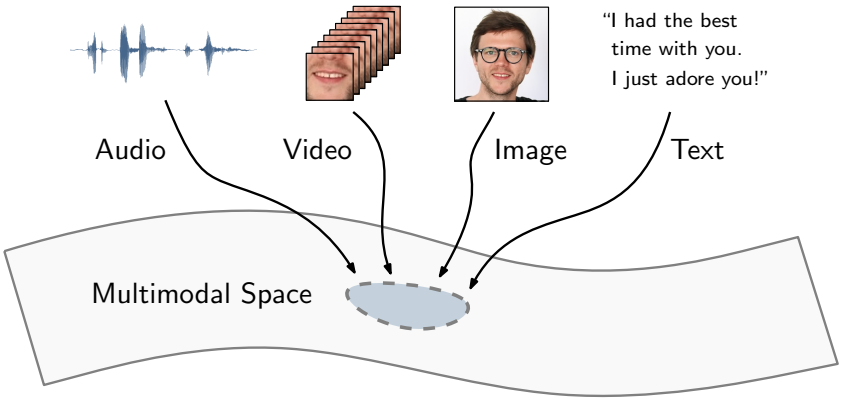


Figure 1.7: Illustration of multimodal machine learning, showcasing the integration of various data types in the analysis of audio-visual speech. The figure represents a speech utterance alongside its corresponding lip movements, a portrait of the speaker, and the written text, all combined into a unified multimodal space to enhance understanding and processing of interconnected information.

correlated, as it allows the model to learn a joint representation that captures the interactions between the modalities. In contrast, late fusion is more flexible and can be used when the modalities are less correlated or when the model needs to be more interpretable.

Early research using audio-visual speech data focused on automatic lipreading to enhance speech recognition. The pioneering work of Petajan et al. [203], [204] introduced the incorporation of visual information to enhance the performance of ASR systems. Their system relied on a codebook of images to translate incoming images into corresponding symbols. These symbol strings were then compared to stored sequences representing different words in the vocabulary. They utilized vector quantization, dynamic time warping, and a heuristic distance measure, resulting in significantly improved performance compared to relying solely on the acoustic recognition system. Yuhas et al. [205] proposed an alternative approach for processing visual speech signals, leveraging analog computation within a distributed network architecture. They demonstrated that these networks could extract speech information from visual images and that this information could be used to enhance automatic vowel recognition. Early works on audio-visual speech enhancement can be traced back to the 1990s. Girin et al. [206] proposed enhancing noisy speech using visual information from the speaker's face. Their approach aimed to estimate adaptive filters based on the speaker's lip pattern. While the results for stationary vowels were encouraging, the performance was less satisfactory for consonants and speech in noisy environments.

More recently, a number of algorithms have been developed in the context of deep learning, including predictive methods based on complex spectral mapping and masking-based approaches that are trained in a supervised fashion. These methods typically consist of two modality-specific encoders, one for processing noisy speech and another for processing video information. The resulting audio and visual features are combined and processed by a fusion model to obtain audio-visual embeddings that are fed into a decoder to predict the clean speech target. Hou et al. [207] proposed an end-to-end audio-visual speech enhancement model utilizing a CNN-based encoder-decoder architecture. The model is trained with a multi-task loss approach that aims to estimate the clean speech target while simultaneously

reconstructing the video features through two MSE terms. Afouras et al. [208] introduced a method designed to isolate individual speakers from simultaneous multi-talker speech in videos. They employed a deep audio-visual speech enhancement network that effectively separates a speaker's voice by leveraging the lip regions in the corresponding video, predicting both the magnitude and phase of the target signal. Ephrat et al. [209] introduced *Looking to Listen at the Cocktail Party* (*L2L*), a joint audio-visual model designed to isolate a single speech signal from a mixture of sounds, including other speakers and background noise. To train their model, they created *AVSpeech*, an extensive audio-visual dataset comprising thousands of hours of video segments sourced from the internet. Martel et al. [20] presented *AVLIT*, a lightweight neural network that employs progressive learning for audio-visual speech separation in noisy environments. To achieve this, they utilize the asynchronous fully recurrent convolutional neural network (AFRCNN) [210], which has demonstrated effective performance in audio-only speech separation. Several other end-to-end methods have investigated audio-visual speech enhancement within the framework of deep learning, making use of advanced network architectures and larger volumes of training data.

Sadeghi et al. [211] introduced a generative approach to audio-visual speech enhancement utilizing a conditional VAE. This approach serves as an audio-visual extension of the original VAE-NMF framework [133], [134], as discussed in Section 1.5.1. The encoder processes a single frame of the speech power spectrogram along with a visual feature vector extracted from the corresponding video frame, and it outputs the parameters of the posterior distribution of the latent variables. The decoder network uses a random vector sampled from this posterior distribution, together with a video feature as input, to model the variance of the speech. Using the noisy speech and the observed visual features, they employ an EM algorithm to iteratively estimate both the clean speech and the noise variance. Similar to the original framework, the estimated variances are used to construct a probabilistic Wiener filter for speech enhancement. In later research, Sadeghi and Alameda-Pineda [212], [213], [214] proposed processing audio and video modalities using separate encoders and addressed cases of unreliable visual inputs by switching between audio-only and audio-visual VAE-based priors. In [P3], we proposed a principled method for conditioning the VAE on visual information by disentangling the latent variable from high-level categorical information, which can be inferred from the video modality. To achieve a disentangled latent representation, we use a discriminator that competes with the encoder of the VAE and an additional encoder to estimate the label for the decoder. In the experiments, we demonstrated the advantages of the proposed disentanglement learning approach when a voice activity label, estimated from visual data, is utilized for speech enhancement. To explore the noise robustness of audio-visual classifiers, in [18], we developed a continuous phoneme recognizer using an end-to-end approach. Utilizing the NTCD-TIMIT dataset [215], we demonstrated an increase in phoneme accuracy from 64.1% to 73.4% by incorporating video information with our proposed audio-visual classifier, compared to an audio-only classifier built on the same architecture.

Recently, SSL frameworks have shown great promise for learning representations from large-scale unlabelled data [216], [217]. Baevski et al. [218] introduced *Wav2Vec*, a contrastive learning approach that predicts quantized representations of raw audio to achieve state-of-the-art performance in speech recognition tasks. Similarly, Hsu et al. [219] developed *HuBERT*, a SSL model for speech representation that uses masked prediction tasks over continuous speech inputs to learn robust acoustic features. Both frameworks demonstrate the effectiveness of SSL in extracting meaningful representations from unlabeled audio data. In multimodal SSL, the goal is to learn a joint representation that captures the interactions between different

modalities, utilizing large-scale datasets. Shi et al. [220] introduced *AV-HuBERT*, an SSL framework designed for audio-visual speech representation learning. The framework employs masked prediction and an iterative clustering mechanism to learn audio-visual representations from audio filterbank features and region of interest (ROI) video frames focused on the mouth region. Each modality is processed by a separate modality-specific encoder, and the resulting representations are concatenated and fed into a *Transformer*-based architecture. The hybrid encoder-*transformer* model encodes masked audio and image sequences into audio-visual features to predict a predetermined sequence of discrete cluster assignments. Initially, these target cluster assignments are developed from signal processing-based acoustic features, such as mel-frequency cepstral coefficients (MFCCs), and are iteratively refined using the features learned by the audio-visual encoder through k-means clustering. Haliassos et al. [221] introduced *RAVEn*, a self-supervised multi-modal approach to jointly learn visual and auditory speech representations. The model's objective involves encoding masked inputs and then predicting contextualized targets generated by slowly evolving momentum encoders. Due to the inherent differences between video and audio, the model design employed an asymmetric approach to the pretext tasks for the two modalities. The auditory stream is tasked with predicting both the visual and auditory targets, while the visual stream is only required to predict the auditory targets. Unlike *AV-HuBERT*, their approach learns robust speech representations directly from raw video and audio data, without relying on handcrafted features.

In [P9], we introduced *AV-Gen*, an audio-visual extension of the *SGMSE+* model [P4] for speech enhancement. We utilized audio-visual embeddings obtained from an *AV-HuBERT* model fine-tuned on lipreading. The layer-wise features from its *Transformer*-based encoder are aggregated, time-aligned, and integrated into the *NCSN++* architecture. Experimental evaluations demonstrated that our proposed audio-visual speech enhancement system improves speech quality and reduces generative artifacts, such as phonetic confusions, compared to the audio-only version. This is further supported by a noticeable decrease in the word error rate (WER) of a downstream ASR model, particularly at low SNRs. Lee et al. [222] proposed a method to isolate a target speaker's voice from a mixture of voices by utilizing visual cues. They employed a diffusion model with a cross-attention-based feature fusion mechanism to effectively integrate the two modalities. For visual feature extraction, they used an architecture based on *ResNet* layers combined with a temporal convolutional network. Jung et al. [223] proposed an efficient method to enhance the quality of corrupted speech signals by incorporating both acoustic and visual cues. They utilized a conditional flow matching algorithm that facilitates the generation of high-quality speech in a single sampling step.

Other approaches treat audio-visual speech enhancement as a speech regeneration task, aiming to produce an output that closely matches the original signal in terms of sound quality, prosody, and other acoustic characteristics [36]. Yang et al. [224] developed a speaker-dependent model that generates clean speech codes from a neural audio codec, conditioned on both noisy audio and visual inputs. These discrete codes are used to reconstruct a mel-spectrogram, which is then processed by a neural vocoder. Hsu et al. [225] decompose audio-visual speech enhancement into two stages: audio-visual speech recognition and text-to-speech synthesis, utilizing hidden units from a pre-trained SSL model. Mira et al. [226] proposed a two-stage approach that first predicts clean mel-spectrograms from audio-visual speech and then converts them into waveform audio using a neural vocoder. In general, audio-visual speech enhancement systems outperform their audio-only counterparts in severe acoustic conditions, but they tend to be more computationally complex due to the integration of an additional modality into

an audio-focused task. Moreover, these systems often rely on clear and reliable video inputs, which is not always feasible in real-world scenarios where video might be occluded, poorly lit, or captured from a suboptimal angle, such as the side of the face [227]. Thus, audio-visual speech processing remains an active research area, with ongoing efforts to develop more robust multimodal integration and efficient processing techniques to overcome these challenges and make such systems practical for real-world applications.

1.7 Outline and Contributions

This dissertation addresses both unsupervised and supervised generative speech enhancement methods. Building on audio-only models, we develop approaches for integrating the video modality to perform audio-visual speech enhancement. We organize the dissertation in three chapters:

- Chapter 3: *Unsupervised Generative Speech Enhancement*
- Chapter 4: *Supervised Generative Speech Enhancement*
- Chapter 5: *Analysis and Comparison*

Below, we outline the research questions addressed in each chapter and provide a brief overview of the contributions.

1.7.1 Chapter 3: Unsupervised Generative Speech Enhancement

Chapter 3 explores unsupervised generative speech enhancement by combining VAEs trained on clean speech with an untrained noise model based on NMF. We propose a method for incorporating the visual modality into this framework by conditioning a VAE on a high-level attribute of speech, which can be inferred by an audio-visual classifier. We address the research questions **RQ1** and **RQ2** listed below, with answers provided in Section 6.1.1 as **A1** and **A2**, respectively.

Research Questions

RQ1 Can an unsupervised generative model, such as the variational autoencoder (VAE), be used to capture the distribution of clean speech, including its temporal dependencies, to enhance the quality and intelligibility of noisy speech within a Bayesian speech enhancement framework?

RQ2 How can a Bayesian speech enhancement method, such as the framework consisting of a variational autoencoder (VAE) and non-negative matrix factorization (NMF), be extended to incorporate the video modality for improved performance in unsupervised settings?

In Section 3.1, we propose to use stochastic temporal convolutional networks (STCNs) with hierarchical and temporal latent variables to learn the distribution of clean speech in an unsupervised manner. At inference time, we combine the learned speech prior with an unlearned noise model based on NMF. Using an MCEM algorithm, we iteratively maximize the likelihood of the speech model and update the parameters of the linear noise model. After

convergence, we use the resulting speech and noise variances in a Wiener filter to enhance the noisy input.

In Section 3.2, we introduce an extension to audio-visual speech enhancement by conditioning the VAE on a voice activity label inferred from the visual modality. To achieve disentanglement of the latent variables, we incorporate a discriminator that competes with the VAE's encoder and an additional encoder to estimate the label for the decoder. We demonstrate the advantages of this disentanglement learning approach when the voice activity label is utilized for speech enhancement. Furthermore, we explore the robustness of audio-visual classifiers and develop an audio-visual phoneme recognizer using an end-to-end approach.

1.7.2 Chapter 4: Supervised Generative Speech Enhancement

Chapter 4 investigates supervised generative speech enhancement, with a particular focus on diffusion models for improving speech quality and restoring common speech communication artifacts. We develop a causal processing scheme with diffusion models and propose an extension to audio-visual speech enhancement by incorporating information from the visual modality. In addition, we explore alternative diffusion processes, including the Schrödinger bridge, to enhance computational speed and robustness. We address the research questions **RQ3**–**RQ6** listed below, with answers provided in Section 6.1.2 as **A3**–**A6**, respectively.

Research Questions

RQ3 How can generative models, such as diffusion models, be effectively employed in supervised settings to learn the posterior distribution of clean speech conditioned on noisy speech, and what are their benefits over predictive methods and other generative methods in enhancing speech quality and intelligibility, particularly under mismatched acoustic conditions and non-additive corruptions?

RQ4 How can causal processing be ensured when using diffusion models for restoring speech communication artifacts?

RQ5 How can visual information be utilized to improve the performance of generative speech enhancement in adverse acoustic conditions, particularly in alleviating phonetic errors in the generated speech?

RQ6 How can alternative diffusion processes influence the performance and computational speed of diffusion-based speech enhancement?

In Section 4.1, we present our method *SGMSE+*, a supervised generative method for speech enhancement that utilizes score-based generative models. We develop a task-adapted diffusion process by introducing a novel drift term to the forward SDE, causing the process to transition from clean speech to noisy speech. This enables sampling from the posterior distribution of clean speech conditioned on noisy speech using the corresponding reverse process. In addition, we present a tutorial on diffusion models for audio restoration with a focus on applications in speech and music [P5].

In Section 4.2, we propose a causal processing system using diffusion models designed to address common speech communication artifacts. The model is trained on a variety of speech distortions, including additive background noise, reverberation, bandwidth limitations, packet loss, and codec artifacts. We demonstrate that this generalized causal approach does not

experience a significant performance penalty compared to its non-causal counterpart, while offering the flexibility needed for real-world applications where various types of distortions may occur.

In Section 4.3, we extend the *SGMSE+* model to audio-visual speech enhancement by conditioning the diffusion process on visual information. We demonstrate that the model can effectively leverage the visual modality to enhance speech quality and intelligibility, particularly in challenging acoustic environments.

In Section 4.4, we explore different training objectives for diffusion-based generative models, including the *EDM* framework [174] and the Schrödinger bridge approach [180]. We propose a perceptual-based loss function for the Schrödinger bridge approach, which improves the quality of the generated speech and achieves leading performance in speech enhancement benchmarks.

1.7.3 Chapter 5: Analysis and Comparison

Chapter 5 provides an analysis of generative speech enhancement methods and investigates differences from predictive methods. We address the research question **RQ7** listed below, with the answer **A7** provided in Section 6.1.3.

Research Questions

RQ7 Is diffusion-based speech enhancement, particularly the *SGMSE+* model, capable of producing studio-quality speech when trained on a high-quality, fullband dataset, and how does it compare to other generative and predictive methods?

In Section 5.1, we present the EARS dataset, a high-quality 48 kHz speech corpus that encompasses a wide array of speaking styles. This includes emotional speech, various reading styles, non-verbal sounds, and conversational freeform speech. To evaluate speech enhancement techniques, we create the EARS-WHAM benchmark and test several predictive and generative speech enhancement methods using both instrumental metrics and human evaluations. Our generative method, *SGMSE+*, achieves the highest scores on most instrumental metrics and emerges as the preferred method in a listening test.

2 List of Publications

The peer-reviewed articles that form the core of this cumulative dissertation are highlighted in grey boxes. All other related publications and preprints are listed in Appendix A.

3. Unsupervised Generative Speech Enhancement

3.1 Variational Autoencoders for Speech Enhancement

[P1] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks", in *ISCA Interspeech*, 2020, pp. 4516–4520.

[P2] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021, pp. 681–685.

3.2 Extension to Audio-Visual Speech Enhancement

[P3] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021, pp. 126–130.

4. Supervised Generative Speech Enhancement

4.1 Diffusion Models for Speech Enhancement

[P4] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models", *IEEE/ACM Transactions Audio, Speech, Language Processing*, vol. 31, pp. 2351–2364, 2023.

[P5] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration: A review", *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2025.

[P6] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain", in *ISCA Interspeech*, 2022, pp. 2928–2932.

4.2 Causal processing for restoration of speech communication artifacts

[P7] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement", *IEEE Open Journal of Signal Processing*, pp. 780–789, 2024.

[P8] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Speech signal improvement using causal generative diffusion models", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–2.

4.3 Extension to audio-visual speech enhancement

[P9] J. Richter, S. Frintrop, and T. Gerkmann, "Audio-visual speech enhancement with score-based generative models", in *Proceedings of ITG Conference on Speech Communication*, 2023, pp. 275–279.

[P10] D. de Oliveira, J. Richter, T. Peer, and T. Gerkmann, "LipDiffuser: Lip-to-speech generation with conditional diffusion models", submitted to *Advances in Neural Information Processing Systems*, 2025.

4.4 Alternative Diffusion Processes

[P11] J. Richter, D. de Oliveira, and T. Gerkmann, "Investigating training objectives for generative speech enhancement", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2025.

[P12] J. Richter, D. de Oliveira, and T. Gerkmann, "Normalize everything: A preconditioned magnitude-preserving architecture for diffusion-based speech enhancement", submitted to *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2025.

5. Analysis and Comparison

5.1. Analysis of Generative Speech Enhancement and Comparison to Predictive Approaches

[P13] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation", in *Proceedings of Interspeech*, 2024, pp. 4873–4877.

[P14] J. Richter and T. Gerkmann, "Diffusion-based speech enhancement: Demonstration of performance and generalization", in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[P15] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation", *IEEE/ACM Transactions Audio, Speech, Language Processing*, vol. 31, pp. 2724–2737, 2023.

[P16] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

Further Publications

[17] Q. Nguyen, J. Richter, M. Lauri, T. Gerkmann, and S. Frintrop, "Improving mix-and-separate training in audio-visual sound source separation with an object prior", in *IEEE International Conference on Pattern Recognition*, 2021, pp. 5844–5851.

[18] J. Richter, J. Liebold, and T. Gerkmann, "Continuous phoneme recognition based on audio-visual modality fusion", in *International Joint Conference on Neural Networks*, 2022, pp. 1–8.

[19] D. de Oliveira, J. Richter, J.-M. Lemercier, T. Peer, and T. Gerkmann, "On the behavior of intrusive and non-intrusive speech enhancement metrics in predictive and generative settings", in *Speech Communication - 15th ITG Conference*, 2023, pp. 260–264.

[20] H. Martel, J. Richter, K. Li, X. Hu, and T. Gerkmann, "Audio-visual speech separation in noisy environments with a lightweight iterative model", in *Proceedings of Interspeech*, 2023, pp. 1673–1677.

[21] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement", in *Proceedings of Interspeech*, 2023, pp. 3809–3813.

[22] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The PESQetarian: On the relevance of goodhart's law for speech enhancement", in *Proceedings of Interspeech*, 2024, pp. 3854–3858.

[23] B. Lay, J.-M. Lermercier, J. Richter, and T. Gerkmann, "Single and few-step diffusion for generative speech enhancement", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024, pp. 626–630.

Unsupervised Generative Speech Enhancement

3

In this chapter, we explore a Bayesian approach for speech enhancement using unsupervised generative models trained on clean speech only. We explore a framework that utilizes VAEs to model the distribution of clean speech spectra, which are combined with an untrained NMF noise model during inference.

Section 3.1 contains publication [P1], which describes an approach to learn the distribution of speech spectrograms using STCNs. We combine this model with an untrained NMF noise model to perform speech enhancement in an unsupervised manner.

Section 3.2 starts with publication [P3], which extends the VAE-NMF framework to include visual data. In this approach, we condition VAEs on voice activity labels derived from video inputs, and propose disentangled learning for VAEs. Additionally, the section includes publication [18], in which we develop an audio-visual phoneme recognizer using an end-to-end approach.

3.1 Variational Autoencoders for Speech Enhancement

3.1.1 Publication: Speech Enhancement with Stochastic Temporal Convolutional Networks

Abstract

We consider the problem of speech modeling in speech enhancement. Recently, deep generative approaches based on variational autoencoders have been proposed to model speech spectrograms. However, these approaches are based either on hierarchical or temporal dependencies of stochastic latent variables. In this paper, we propose a generative approach to speech enhancement based on a stochastic temporal convolutional network, which combines both hierarchical and temporal dependencies of stochastic variables. We evaluate our method with real recordings of different noisy environments. The proposed speech enhancement method outperforms a previous non-sequential approach based on feed-forward fully-connected networks in terms of speech distortion, instrumental speech quality and intelligibility. At the same time, the computational cost of the proposed generative speech model remains feasible, due to inherent parallelism of the convolutional architecture.

Reference

Julius Richter, Guillaume Carbajal, and Timo Gerkmann, "Speech Enhancement with Stochastic Temporal Convolutional Networks", *Proceedings of Interspeech*, 2020, pp. 4516—4520. DOI: 10.21437/Interspeech.2020-2588

Copyright Notice

The following article is the accepted version of the article published with ISCA.
© 2020 ISCA. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Julius Richter is the lead author of this publication. He was responsible for implementing all the algorithms, training the neural networks used in the study, conducting the experimental validation, and writing the manuscript. Guillaume Carbajal provided valuable feedback on all the methods through discussions and assisted with the manuscript review. Timo Gerkmann contributed insights on the experimental validation and mathematical derivations, and also reviewed the manuscript.

Speech Enhancement with Stochastic Temporal Convolutional Networks

Julius Richter, Guillaume Carbajal, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

{julius.richter, guillaume.carbajal, timo.gerkmann}@uni-hamburg.de

Abstract

We consider the problem of speech modeling in speech enhancement. Recently, deep generative approaches based on variational autoencoders have been proposed to model speech spectrograms. However, these approaches are based either on hierarchical or temporal dependencies of stochastic latent variables. In this paper, we propose a generative approach to speech enhancement based on a stochastic temporal convolutional network, which combines both hierarchical and temporal dependencies of stochastic variables. We evaluate our method with real recordings of different noisy environments. The proposed speech enhancement method outperforms a previous non-sequential approach based on feed-forward fully-connected networks in terms of speech distortion, instrumental speech quality and intelligibility. At the same time, the computational cost of the proposed generative speech model remains feasible, due to inherent parallelism of the convolutional architecture.

Index Terms: speech enhancement, stochastic temporal convolutional networks, generative model, variational inference.

1. Introduction

Speech processing systems such as mobile phones, VoIP, teleconferencing systems, speech recognition, and hearing aids require improving speech quality and intelligibility [1]. Single-channel speech enhancement aims at recovering a clean speech signal from a mixture which can contain speech and additive noise [2]. In traditional speech enhancement, Bayesian estimators are often used to estimate speech coefficients [3], and speech parameters such as the speech power spectral density [4], the noise power spectral density [5], and the phase [6].

Besides Bayesian estimators, a Bayesian generative model of the observed signal can be built based on available knowledge about its production process [7]. The resulting probabilistic speech enhancement methods aim to mimic the hidden random process of speech and may be used to generate artificial data that resembles the properties of a given dataset.

Several speech enhancement methods combine concepts of Bayesian inference and deep learning, e.g. [8–10]. These approaches make use of generative models, in particular the variational autoencoder (VAE) [11, 12]. Similar to methods based on denoising autoencoders [13, 14], VAE-based methods are also capable of denoising speech spectrograms by using stochastic latent variables, which are hierarchically arranged in a top-down fashion. Although both achieve good performance compared against traditional speech enhancement methods, most of them only reconstruct single spectrogram frames without modeling temporal dependencies.

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and ahoi.digital. We would like to thank Rohde & Schwarz SwissQual AG for their support with POLQA.

Recently, a generative model called stochastic temporal convolutional network (STCN) was proposed which can simultaneously capture temporal dependencies of variable-length sequences and learn correlations between output variables [15]. The STCN uses a top-down hierarchy of stochastic latent variables which are conditioned on deterministic representations computed bottom-up. The deterministic representations correspond to dilated convolutions of a temporal convolutional network (TCN) [16]. Lateral shortcut connections between the deterministic and latent variables allow higher levels of the latent hierarchy to focus on more abstract invariant features [17]. The network architecture is capable of generating high-quality synthetic samples and achieves state-of-the-art log-likelihoods in speech synthesis. However, to our best knowledge, STCNs have not been applied in speech enhancement yet.

In this work, motivated by the aforementioned advantages, we propose a generative approach to speech enhancement using an STCN as a speaker-independent speech model to estimate the variance of clean speech. The estimation of the noise variance, on the other hand, is based on a non-negative matrix factorization (NMF) [18, 19]. We compare the proposed convolutional approach to a non-sequential approach based on a feed-forward fully-connected architecture using the same NMF parameter optimization algorithm [10].

The rest of this paper is organized as follows: in Section 2 we introduce a general speech enhancement method based on speech and noise variance estimation. Section 3 describes deep generative speech models based on variational inference. The novel STCN architecture is introduced in Section 4 followed by the evaluation in Section 5.

2. Speech Enhancement

The general goal of single-channel speech enhancement is defined as recovering a speech signal from an observed signal involving interfering sound sources or reverberation. For this purpose, a common approach consists in estimating the speech and noise variances in order to reconstruct the clean speech signal.

2.1. Model

In the time-frequency domain using the short time Fourier transform (STFT), the mixture signal $y_{ft} \in \mathbb{C}$ is the sum of the clean speech $s_{ft} \in \mathbb{C}$ and the noise $n_{ft} \in \mathbb{C}$, such that

$$y_{ft} = s_{ft} + n_{ft}, \tag{1}$$

for every frequency bin $f \in \{1, \dots, F\}$ and time frame $t \in \{1, \dots, T\}$, where F denotes the number of frequency bins and T the number of time frames of the utterance. The signals s_{ft} and n_{ft} are modeled as mutually independent circularly-symmetric complex Gaussian random variables,

$$s_{ft} \sim \mathcal{CN}(0, \sigma^2_{s,ft}), \quad n_{ft} \sim \mathcal{CN}(0, \sigma^2_{n,ft}), \tag{2}$$

where $\mathcal{CN}(\mu, \sigma^2)$ denotes a complex Gaussian distribution with mean μ and variance σ^2. Under a local stationary assumption, the variances $\sigma^2_{s,ft}$ and $\sigma^2_{n,ft}$ represent the short-time power spectral density of s_{ft} and n_{ft}, respectively [20]. Given the noisy mixture, clean speech coefficients can be estimated in the minimum mean square sense using the Wiener estimator

$$\hat{s}_{ft} = \frac{\hat{\sigma}^2_{s,ft}}{\hat{\sigma}^2_{s,ft} + \hat{\sigma}^2_{n,ft}} \, y_{ft} \tag{3}$$

where $\hat{\sigma}^2_{s,ft}$ and $\hat{\sigma}^2_{n,ft}$ are the estimated variances using the corresponding signal variance models as described next.

2.2. Signal variance models

NMF is a popular choice for modeling signal variances based on previously trained clean speech spectra [18]. However, it is limited in its modeling capacity due to the linear parametrization of the variances. Therefore, we use a non-linear deep generative model for speech variance estimation which depends on stochastic latent variables. The noise variance, on the other hand, is estimated based on an untrained NMF noise model.

2.3. Robust mixture model

In order to provide some robustness with respect to the time-varying loudness of different speech signals, the mixture model in Eq. (1) is extended with a frequency-independent but time-varying gain $g_t \in \mathbb{R}_+$ [10], such that

$$y_{ft} = \sqrt{g_t}\, s_{ft} + n_{ft}. \tag{4}$$

3. Generative Speech Models

3.1. Variational Autoencoders

With the original VAE framework [11, 12], speech power coefficients $x_{ft} = |s_{ft}|^2$ are created by a random process, involving an unobserved random variable $\mathbf{z}_t \in \mathbb{R}^D$. This process is shown in Fig. 1a and consists of two steps: first a value $\mathbf{z}_t^{(i)}$ is drawn from a prior probability distribution $p(\mathbf{z}_t)$ as the i-th sample, and second a power spectrogram frame $\mathbf{x}_t^{(i)} \in \mathbb{R}_+^F$ is generated from a conditional probability distribution $p_\theta(\mathbf{x}_t|\mathbf{z}_t)$ which is also called generative distribution.

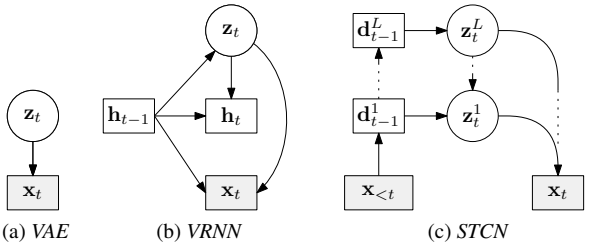
In variational inference, a recognition model $q_\phi(\mathbf{z}_t|\mathbf{x}_t)$ is introduced as an approximation to the intractable true posterior $p(\mathbf{z}_t|\mathbf{x}_t)$ [21]. All distributions are modeled as neural networks and its parameters θ and ϕ are jointly optimized maximizing the variational lower bound on the marginal log-likelihood for a given spectrogram frame

$$\begin{aligned} \log p(\mathbf{x}_t^{(i)}) \geq & - D_{\mathrm{KL}}(q_\phi(\mathbf{z}_t|\mathbf{x}_t^{(i)}) \,||\, p(\mathbf{z}_t)) \\ & + \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_t^{(i)})}[\log p_\theta(\mathbf{x}_t^{(i)}|\mathbf{z}_t)], \end{aligned} \tag{5}$$

where D_{KL} denotes the Kullback-Leibler divergence between the approximate posterior $q_\phi(\mathbf{z}_t|\mathbf{x}_t^{(i)})$ and the prior probability distribution $p(\mathbf{z}_t)$ which is commonly chosen to be a standard Gaussian distribution with zero mean and unit variance. The second term in Eq. (5) is an expected reconstruction error and requires estimation by sampling.

The statistical properties of clean speech, which are learned by the generative model during training, can be used to estimate the distribution of the noisy mixture coefficients y_{ft}. Since speech and noise signals are supposed to be mutually independent given the latent variable $\mathbf{z}_t$, we have

$$y_{ft}|\mathbf{z}_t \sim \mathcal{CN}(0, g_t \hat{\sigma}^2_{s,ft}(\mathbf{z}_t) + \hat{\sigma}^2_{n,ft}). \tag{6}$$



(a) *VAE* (b) *VRNN* (c) *STCN*

Figure 1: *Generative models: (a) the variational autoencoder (VAE), (b) the variational recurrent neural network (VRNN), and (c) the stochastic temporal convolutional network (STCN). Circles are stochastic variables, squares are deterministic variables, and gray boxes represent the observed sequence.*

Straightforward maximum likelihood estimation of Eq. (6) is intractable due to the non-linear relation between the speech variance and the latent variables. However, a Monte Carlo expectation maximization (MCEM) algorithm can be employed which iteratively optimizes the NMF noise parameters and the time-varying gain g_t using a block-coordinate approach [10]. Samples of the latent variable are drawn from the approximate posterior distribution using the random walk Metropolis-Hastings algorithm [22].

3.2. Temporal dependencies

In the VAE framework [11, 12], the generative model does not consider temporal dependencies of sequential data. Thus, we introduce the variational recurrent neural network (VRNN) as a common deep generative model capable of modeling the temporal dependencies by introducing a deterministic hidden state $\mathbf{h}_t \in \mathbb{R}^H$ which is updated in a recurrent fashion [23]. The generation of new samples and the recurrence of the hidden state is illustrated in Fig. 1b.

VRNNs have an infinite internal memory due to occurring feedback loops. As we are interested in modeling finite length phonemes, this infinite internal memory may be seen as a conceptual disadvantage. Furthermore, due to the necessary back-propagation through time, recurrent network architectures are rather slow to train and may also suffer from the vanishing gradient problem [24].

4. Stochastic Temporal Convolutional Networks

The STCN is another generative model capable of modeling the joint probability distribution of variable-length sequences. For this purpose, the architecture employs two main modules: 1) a TCN with deterministic representations computed bottom-up; and 2) a stochastic latent variable hierarchy with top-down dependencies, as illustrated in Fig. 1c.

4.1. Temporal convolutional network

In the TCN, dilated causal convolutions are applied over the input sequence $\mathbf{x}$ to compute a set of deterministic representations $\mathbf{d} = \{\mathbf{d}^1, \ldots, \mathbf{d}^L\}$ where L is the total number of stacked layers with corresponding sequences $\mathbf{d}^l = (\mathbf{d}^l_1, \ldots, \mathbf{d}^l_T)$ and $\mathbf{d}^l_t \in \mathbb{R}^{D_l}$. The computation of the deterministic representations is bottom-up and recursively defined as $\mathbf{d}^l_t = f(\mathbf{d}^{l-1}_t)$ starting with $\mathbf{d}^0 = \mathbf{x}$. The function f is a series of transformations containing residual connections, dilated convolutions, weight normalization [25], ReLUs, and spatial dropout [26].

Using larger dilation factors in the convolution enables a deterministic representation at higher levels to represent a wider range of inputs, thus effectively expanding the receptive field compared to a standard convolution. Therefore, contrary to VRNNs, the STCN has the advantage to accurately set a desired receptive field size, which may be useful for the task of speech enhancement. The overall receptive field size is given as $(k-1)(2^L-1)$ with filter size k and can be set in line with the typical length of phonemes, or in line with the temporal integration time of the human auditory system, which has an upper bound of a few hundreds of milliseconds [27].

Furthermore, residual and skip connections allow local information to propagate through the network while avoiding the vanishing gradient problem [28, 29].

4.2. Stochastic latent variable hierarchy

Stochastic latent variables are arranged in correspondence to the TCN layers. Thus, there also exists a set of random variables $\mathbf{z} = \{\mathbf{z}^1, \dots, \mathbf{z}^L\}$ with sequences $\mathbf{z}^l = (\mathbf{z}_1^l, \dots, \mathbf{z}_T^l)$ and $\mathbf{z}_t^l \in \mathbb{R}^{Z_l}$ which capture temporal dependencies at different time scales. The decoupling of deterministic and stochastic layers is shown in Fig. 2. The stochastic latent variable hierarchy can be seen as a modular add-on for any temporal convolutional network architecture.

4.3. Inference

The recognition model of the STCN relies on a top-down dependency of the latent variables, as illustrated in Fig. 2. As a result, the parameters of the approximate posterior for each latent layer l are computed by

$$[\hat{\boldsymbol{\mu}}_{t,q}^l, \hat{\mathbf{v}}_{t,q}^l] = \begin{cases} f_q^{(l)}(\mathbf{z}_t^{l+1}, \mathbf{d}_t^l), & \text{for } l \in [1, L-1] \\ f_q^{(L)}(\mathbf{d}_t^L), & \text{for } l = L, \end{cases} \tag{7}$$

where $\{f_q^{(l)}\}_{l=1,\dots,L}$ is a set of neural networks consisting of stacked layers of 1D convolutions with kernel size 1.

This is contrary to ordinary VAEs, where the inference is defined as a bottom-up process. Furthermore, the mean $\hat{\boldsymbol{\mu}}_{t,q}^l$ and the diagonal variance entries $\hat{\mathbf{v}}_{t,q}^l$ are subsequently corrected by precision-weighted addition [17], such that

$$\begin{aligned} \boldsymbol{\mu}_{t,q}^l &= \mathbf{v}_{t,q}^l(\hat{\boldsymbol{\mu}}_{t,q}^l(\hat{\mathbf{v}}_{t,q}^l)^{-2} + \boldsymbol{\mu}_{t,p}^l(\mathbf{v}_{t,p}^l)^{-2}), \\ \mathbf{v}_{t,q}^l &= \frac{1}{(\hat{\mathbf{v}}_{t,q}^l)^{-2} + (\mathbf{v}_{t,p}^l)^{-2}}, \end{aligned} \tag{8}$$

where $\boldsymbol{\mu}_{t,p}^l$ and $\mathbf{v}_{t,p}^l$ are the parameters of the prior distribution which is described in Sec. 4.4.

Finally, the approximate posterior distribution of the set of latent variables $\mathbf{z}_t = \{\mathbf{z}_t^1, \dots, \mathbf{z}_t^L\}$ conditioned on the set of deterministic representations $\mathbf{d}_t = \{\mathbf{d}_t^1, \dots, \mathbf{d}_t^L\}$ at time step t is given as

$$q(\mathbf{z}_t|\mathbf{d}_t) = q(\mathbf{z}_t^L|\mathbf{d}_t^L) \prod_{l=1}^{L-1} q(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_t^l), \tag{9}$$

where each distribution of every latent layer l is modeled as a multidimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}_{t,q}^l, \mathrm{diag}(\mathbf{v}_{t,q}^l))$.

4.4. Prior distribution

The prior distribution of the set of latent variables $\mathbf{z}_t$ at time step t depends on the set of deterministic representations $\mathbf{d}_{t-1}$ of the previous time step and is layer-wise factorized as

$$p(\mathbf{z}_t|\mathbf{d}_{t-1}) = p(\mathbf{z}_t^L|\mathbf{d}_{t-1}^L) \prod_{l=1}^{L-1} p(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_{t-1}^l), \tag{10}$$

where each distribution of every latent layer l is modeled as a multidimensional real-valued Gaussian $\mathcal{N}(\boldsymbol{\mu}_{t,p}^l, \mathrm{diag}(\mathbf{v}_{t,p}^l))$ with mean and diagonal variance entries

$$[\boldsymbol{\mu}_{t,p}^l, \mathbf{v}_{t,p}^l] = \begin{cases} f_p^{(l)}(\mathbf{z}_t^{l+1}, \mathbf{d}_{t-1}^l), & \text{for } l \in [1, L-1] \\ f_p^{(L)}(\mathbf{d}_{t-1}^L), & \text{for } l = L, \end{cases} \tag{11}$$

and the set of neural networks $\{f_p^{(l)}\}_{l=1,\dots,L}$ consisting of stacked layers of 1D convolutions with kernel size 1.

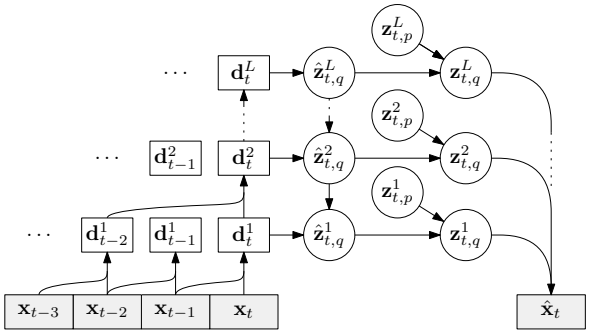


Figure 2: *Computational graph of the inference based on the prior and the approximate posterior distribution.*

4.5. Observation model

The final prediction of the clean speech coefficients $\hat{s}_{ft}$ given the latent variable $\mathbf{z}_t$ is modeled as

$$\hat{s}_{ft}|\mathbf{z}_t \sim \mathcal{CN}(0, \{\hat{\boldsymbol{\sigma}}_{s,t}^2(\mathbf{z}_t)\}_f)) \tag{12}$$

where $\hat{\boldsymbol{\sigma}}_{s,ft}^2 : \mathbb{R}^D \mapsto \mathbb{R}_+^F$ represents a multidimensional nonlinear function modeled by stacked TCN layers. In order to take all latent variables into account, we use the concatenation of samples from all latent layers $\mathbf{z}_t = (\mathbf{z}_t^1 \cdots \mathbf{z}_t^L)$, which takes inspiration from recent convolutional architectures [30].

4.6. Learning objective

The learning objective of the STCN is similar to Eq. (5), namely optimizing the variational lower bound on the log-likelihood at time step t, resulting in the loss function

$$\mathcal{L}(\mathbf{x}_t) = \mathcal{L}_t^{\mathrm{KL}}(\mathbf{x}_t) + \mathcal{L}_t^{\mathrm{Recon}}(\mathbf{x}_t). \tag{13}$$

Using the factorizations from Eq. (9) and (10), the regularization term becomes

$$\begin{aligned} \mathcal{L}_t^{\mathrm{KL}}(\mathbf{x}_t) = \; & D_{\mathrm{KL}}(q(\mathbf{z}_t^L|\mathbf{d}_t^L) \,||\, p(\mathbf{z}_t^L|\mathbf{d}_{t-1}^L)) \\ & + \sum_{l=1}^{L-1} \mathbb{E}_{q(\mathbf{z}_t^{l+1}|\cdot)}[D_{\mathrm{KL}}(q(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_t^l) \,||\, p(\mathbf{z}_t^l|\mathbf{z}_t^{l+1}, \mathbf{d}_{t-1}^l))], \end{aligned} \tag{14}$$

where the expectation over the latent variables $\mathbf{z}_t^{l+1}$ is approximated using the reparameterization trick [11]. The Kullback-Leibler divergences can be calculated analytically since all appearing distributions are diagonal multivariate Gaussians.

The output of $\hat{\boldsymbol{\sigma}}^2_{s,t}(\mathbf{z}_t)$ also represents the estimated short-time power spectral density $\hat{\mathbf{x}}_t$ at frame t, therefore the reconstruction loss is defined as

$$\mathcal{L}_t^{\text{Recon}}(\mathbf{x}_t) = \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x}_{\leq t})}\left[\left(\log \frac{\mathbf{x}_t + \epsilon}{\hat{\boldsymbol{\sigma}}^2_{s,t}(\mathbf{z}_t) + \epsilon}\right)^2\right] \quad (15)$$

where the expectation is approximated using the reparameterization trick and ϵ is numerically motivated in order not divide by zero.

5. Evaluation

In this section, we compare the performance of the proposed STCN approach with a non-sequential VAE approach [10]. We assess the performance in terms of speech distortion, speech intelligibility, and speech quality.

Very recently Leglaive et al. proposed a speech enhancement method based on VRNNs [31]. However, their approach is based on a different optimization algorithm than in the non-sequential VAE and the STCN considered here, which would make a direct comparison of the approaches difficult.

We also refrain from comparing with the other popular generative speech model which is based on generative adversarial networks [8], because its training needs clean and noisy speech pairs. Thus, the model is restricted to a limited set of noise types which is not the case in our approach. Furthermore, the number of learnable parameters differs by two orders of magnitude compared to the STCN approach.

5.1. Dataset

As training data we use approximately 25 hours of clean speech from the "si_tr_s" subset of the Wall Street Journal (WSJ0) dataset [32]. For testing we use 651 synthetic mixtures corresponding to approximately 1.5 hours of noisy speech. The clean speech signals are taken from the "si_et_05" subset of WSJ0 (8 unseen speakers), and the noise signals from the "verification" subset of the QUT-NOISE dataset [33]. Each mixture is created by uniformly sampling a noise type among "cafe", "home", "street", "car" and a signal-to-noise ratio (SNR) among -5, 0, 5 dB. All signals have a sampling rate of 16 kHz.

5.2. Hyperparameter settings

Following the baseline method, the STFT is computed using a 64 ms sine window with 75% overlap, resulting in a frame period of 16 ms and $F = 513$ unique frequency bins.

The re-implementation of the non-sequential VAE follows the same hyperparameter setting in [10]. Both the encoder and decoder consist of a 128-dimensional fully-connected feed-forward representation layer before mapping to a latent layer of dimension 16, resulting in 171,297 learnable parameters.

For the STCN we use $L = 4$ layers and a filter size $k = 2$ which results in a receptive field of 240 ms. The deterministic representations have dimensions [64, 32, 16, 8], whereas the latent layers have dimensions [32, 16, 8, 4]. The networks which calculate the parameters of the approximate posterior and the prior distribution in each layer, $f_q^{(l)}$ and $f_p^{(l)}$ respectively, are modeled as three stacked layers of 1D convolutions with kernel size 1, mapping from the input dimension $Z_{l+1} + D_l$ to the latent dimension Z_l. The observation model uses a TCN of two layers with dimensionality 256, mapping from input dimension $Z = \sum_l Z_l$ to the output dimension F. All variances in the latent layers are clamped between 0.001 and 5 and dropout is set to 0.2. The total number of learnable parameters is 325,497 which is less than factor two compared against the non-sequential VAE.

Method	SI-SDR (dB)	ESTOI	POLQA
VAE [10]	3.74 ± 0.21	0.58 ± 0.01	1.83 ± 0.03
STCN	$\mathbf{4.48 \pm 0.30}$	$\mathbf{0.66 \pm 0.01}$	$\mathbf{2.20 \pm 0.04}$
Mixture	0.54 ± 0.31	0.65 ± 0.01	2.26 ± 0.05

Table 1: *Average results and confidence intervals*

We use the Adam optimizer with standard configuration and a learning rate of 10^{-3} [34]. For the VAE, we set the batch size to 128, whereas for the STCN to 16. The training takes about 100 epochs until the loss converges. In the training of the STCN we gradually turn on the KL-term $\mathcal{L}_t^{\text{KL}}(\mathbf{x}_t)$ within the first 50 epochs, in order not to collapse into the prior [35].

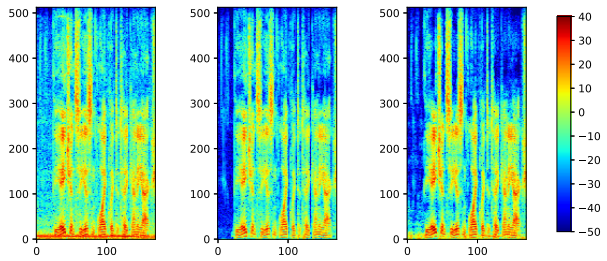
For the optimization of the NMF noise parameters and the time-varying gain, we follow the MCEM algorithm in [10] and set the rank of the NMF to 8. The Metropolis-Hastings algorithm takes the same parameters as in the baseline method.

5.3. Results

To measure performance we use the scale-invariant signal-to-distortion ratio (SI-SDR) [36], raw scores of the extended short-time objective intelligibility (ESTOI) with values between 0 and 1 [37], and the perceptual objective listening quality analysis (POLQA) score with values between 1 and 5 [38].

The average results are shown in Table 1. It may be seen that the proposed STCN outperforms the VAE by 0.8 dB in terms of SI-SDR. It is interesting to see that, at the same time, it also outperforms the VAE in terms of ESTOI and POLQA.

Fig. 3 shows spectrograms $\mathbf{S}_{\text{dB}} = 20\log_{10}(|\mathbf{S}|)$ for an example utterance for: a) the mixture signal, b) the corresponding clean speech, and c) the reconstructed clean speech. It may be seen that the proposed method covers speech while attenuating additive noise. Code and audio examples are available online[1].



(a) *Mixture* (b) *Clean* (c) *Reconstructed*

Figure 3: *Magnitude of frequency bins in dB over time frames of an example utterance for: (a) the mixture signal, (b) the corresponding clean speech, and (c) the reconstructed speech.*

6. Conclusion

In this work, we proposed a generative speech model based on an STCN for speech enhancement. This approach combines both hierarchical and temporal dependencies of stochastic variables. We evaluated our approach with real recordings of different noise environments. The proposed approach outperforms the non-sequential VAE in terms of signal-to-distortion ratio, and instrumental speech quality and intelligibility.

[1] https://uhh.de/inf-sp-stcn2020

7. References

[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2013.

[3] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement – A Survey of the State-of-the-art*. Morgan & Claypool, Jan. 2013.

[4] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4897–4900.

[5] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," in *Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4. IEEE, 2011, pp. 1383–1393.

[6] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," in *Transactions on Signal Processing*, vol. 62, no. 16. IEEE, 2014, pp. 4199–4208.

[7] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI global, 2011, pp. 162–185.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proceedings Interspeech*, 2017, pp. 3642–3646.

[9] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 716–720.

[10] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations ICLR*, Y. Bengio and Y. LeCun, Eds., 2014.

[12] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning*, 2014, pp. 1278–1286.

[13] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[14] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[15] E. Aksan and O. Hilliges, "STCN: Stochastic temporal convolutional networks," in *7th International Conference on Learning Representations (ICLR 2019)*, 2019.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[17] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Advances in neural information processing systems*, 2016, pp. 3738–3746.

[18] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[19] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 45–48.

[20] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[22] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

[23] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[24] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[25] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in neural information processing systems*, 2016, pp. 901–909.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] G. Van Den Brink, "Detection of tone pulse of various durations in noise of various bandwidths," *The Journal of the Acoustical Society of America*, vol. 36, no. 6, pp. 1206–1211, 1964.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[31] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020.

[32] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[33] D. B. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," in *Interspeech*, 2015, pp. 3456–3460.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[37] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[38] "P.863: Perceptual objective listening quality prediction," International Telecommunication Union, Mar. 2018, iTU-T recommendation. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

3.2 Extension to Audio-Visual Speech Enhancement

3.2.1 Publication: Disentanglement Learning for Variational Autoencoders Applied to Audio-Visual Speech Enhancement

Abstract

Recently, the standard variational autoencoder has been success-fully used to learn a probabilistic prior over speech signals, which is then used to perform speech enhancement. Variational autoencoders have then been conditioned on a label describing a high-level speech attribute (e.g. speech activity) that allows for a more explicit control of speech generation. However, the label is not guaranteed to be disentangled from the other latent variables, which results in limited performance improvements compared to the standard variational autoencoder. In this work, we propose to use an adversarial training scheme for variational autoencoders to disentangle the label from the other latent variables. At training, we use a discriminator that competes with the encoder of the variational autoencoder. Simultaneously, we also use an additional encoder that estimates the label for the decoder of the variational autoencoder, which proves to be crucial to learn disentanglement. We show the benefit of the proposed disentanglement learning when a voice activity label, estimated from visual data, is used for speech enhancement.

Reference

Guillaume Carbajal, Julius Richter, and Timo Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement", *IEEE Workshop on Applications Signal Processing to Audio and Acoustics*, 2021. DOI: 10.1109/WASPAA52581.2021.9632676

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2021 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Guillaume Carbajal and Julius Richter made equal contributions to this work. Timo Gerkmann provided insights into the experimental validation and mathematical derivations and participated in reviewing the manuscript.

DISENTANGLEMENT LEARNING FOR VARIATIONAL AUTOENCODERS APPLIED TO AUDIO-VISUAL SPEECH ENHANCEMENT

Guillaume Carbajal, Julius Richter, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany
{guillaume.carbajal, julius.richter, timo.gerkmann}@uni-hamburg.de

ABSTRACT

Recently, the standard variational autoencoder has been successfully used to learn a probabilistic prior over speech signals, which is then used to perform speech enhancement. Variational autoencoders have then been conditioned on a label describing a high-level speech attribute (e.g. speech activity) that allows for a more explicit control of speech generation. However, the label is not guaranteed to be disentangled from the other latent variables, which results in limited performance improvements compared to the standard variational autoencoder. In this work, we propose to use an adversarial training scheme for variational autoencoders to disentangle the label from the other latent variables. At training, we use a discriminator that competes with the encoder of the variational autoencoder. Simultaneously, we also use an additional encoder that estimates the label for the decoder of the variational autoencoder, which proves to be crucial to learn disentanglement. We show the benefit of the proposed disentanglement learning when a voice activity label, estimated from visual data, is used for speech enhancement.

Index Terms— Speech enhancement, conditional generative model, variational autoencoder, disentanglement learning, adversarial training, semi-supervised learning, audio-visual.

1. INTRODUCTION

Single-channel speech enhancement consists in recovering a speech signal from a mixture signal captured with one microphone in a noisy environment [1]. Common speech enhancement approaches estimate the speech signal using a filter in the time-frequency domain to reduce noise while avoiding speech artifacts [2]. Under the Gaussian assumption, the optimal filter in the minimum mean square error sense requires estimating the signal variances [3, 4, 5].

Recently, deep generative models based on the variational autoencoder (VAE) have gained attention for learning the probability distribution of complex data [6]. VAEs have been used to learn a prior distribution of clean speech, and have been combined with an untrained non-negative matrix factorization (NMF) noise model to estimate the signal variances using a Monte Carlo expectation maximization (MCEM) algorithm [7, 8]. However, since the VAE speech model is trained with clean speech only in an unsupervised manner, the speech prior generates a speech-like signal when only noise is present. As a result, the VAE often outputs speech-like noise when applied to noisy speech [9].

An increased robustness can be obtained by incorporating temporal dependencies [10], noise-aware training [11], or by conditioning the VAE on a label describing an attribute of the data that allows for a more explicit control of data generation [12]. For various speech-related tasks, VAEs have been conditioned on a label describing a speech attribute, such as speaker identity [13, 14], phoneme [14] or speech activity [15]. Ideally, the label should be independent from the other latent dimensions to obtain an explicit control of speech generation. However, the semi-supervised learning of conditional VAEs does not guarantee that it will promote independence between the label and the other latent dimensions [16]. As a result, the speech generation can only partially be controlled by the label.

Disentanglement learning aims at making all the dimensions of the latent distribution independent from each other [17, 18, 19], e.g. by holding different axes of variation fixed during training [20]. Semi-supervised disentanglement approaches, on the other hand, tackle the problem of making only some observed (often interpretable) variations in the data independent from the other latent dimensions which themselves remain entangled [12, 16, 21]. In particular, some of them rely on adversarial training in the latent space for image generation [22, 23].

In this work, we propose to use a semi-supervised disentanglement approach based on adversarial training in the latent space. At training, we use a discriminator that competes with the encoder of the VAE. The discriminator aims at identifying the label from the other latent dimensions, whereas the encoder aims at making it unable to estimate the label. Simultaneously, we also use an additional encoder that estimates the label for the decoder of the VAE, which proves to be crucial to learn disentanglement during training. We show the benefit of the proposed disentanglement learning when a voice activity label, estimated from visual data, is used for speech enhancement.

The rest of this paper is organized as follows. In Section 2 we summarize the background related to the VAE for speech enhancement. Section 3 describes our proposed approach. The experimental setup is described in Section 4 which is followed by the evaluation in Section 5.

2. BACKGROUND

2.1. Mixture model and filtering

In the time-frequency domain using the short time Fourier transform (STFT), the mixture signal $x_{nf} \in \mathbb{C}$ is the sum of the clean speech $s_{nf} \in \mathbb{C}$ and the noise $b_{nf} \in \mathbb{C}$:

$$x_{nf} = \sqrt{g_n}\, s_{nf} + b_{nf}, \tag{1}$$

at time frame index $n \in [1, N]$ and frequency bin $f \in [1, F]$, where N denotes the number of time frames and F the number of frequency bins of the utterance. The scalar $g_n \in \mathbb{R}_+$ represents a frequency-independent but time-varying gain providing

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and ahoi.digital.

some robustness with respect to the time-varying loudness of different speech signals [8].

Under the Gaussian assumption, the clean speech s_{nf} can be estimated in the minimum mean square error sense using the Wiener estimator:

$$\widehat{s}_{nf} = \frac{\widehat{g}_n \widehat{v}_{s,nf}}{\widehat{g}_n \widehat{v}_{s,nf} + \widehat{v}_{b,nf}} \, x_{nf}, \tag{2}$$

where $\widehat{v}_{s,nf}$ and $\widehat{v}_{b,nf}$ are the estimated variances of the clean speech s_{nf} and the noise b_{nf}, respectively. Under a local stationary assumption, short-time power spectra $|s_{nf}|^2$ and $|b_{nf}|^2$ are unbiased estimates of the signal variances [24].

2.2. VAE as speech prior

2.2.1. *Standard VAE*

The standard VAE is used to learn a prior over clean speech [7, 8]. Fig. 1a shows the training scheme of the standard VAE. At time frame n, the frequency bins of clean speech $\mathbf{s}_n \in \mathbb{C}^F$ are modeled as

$$p_\theta(\mathbf{s}_n|\mathbf{z}_n) = \mathcal{CN}(\mathbf{0}, \mathrm{diag}(\mathbf{D}_\theta(\mathbf{z}_n))), \quad \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $\mathbf{z}_n \in \mathbb{R}^L$ denotes a latent variable of dimension L and $\mathbf{D}_\theta(\mathbf{z}_n) \in \mathbb{R}_+^F$ is the output of the *decoder* network $\mathbf{D}_\theta(\cdot)$ parametrized by θ.

In variational inference, the posterior of $\mathbf{z}_n$ is approximated as

$$q_\phi(\mathbf{z}_n|\, \mathbf{s}_n) = \mathcal{N}(\boldsymbol{\mu}_\phi(|\mathbf{s}_n|^2), \mathrm{diag}(\mathbf{v}_\phi(|\mathbf{s}_n|^2))), \tag{4}$$

where $\boldsymbol{\mu}_\phi(|\mathbf{s}_n|^2) \in \mathbb{R}^D$ and $\mathbf{v}_\phi(|\mathbf{s}_n|^2) \in \mathbb{R}_+^D$ are the outputs of the *encoder* network $\mathbf{E}_{\phi,z}(\cdot)$ parametrized by ϕ.

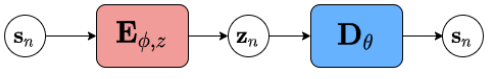
The encoder $\mathbf{E}_{\phi,z}(\cdot)$ and decoder $\mathbf{D}_\theta(\cdot)$ are jointly trained by maximizing a lower bound of the marginal per-frame log-likelihood, called the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{z}_n|\mathbf{s}_n)}[\log p_\theta(\mathbf{s}_n|\mathbf{z}_n)] - \mathcal{D}_{\text{KL}}(q_\phi(\mathbf{z}_n|\mathbf{s}_n)||p(\mathbf{z}_n)), \tag{5}$$

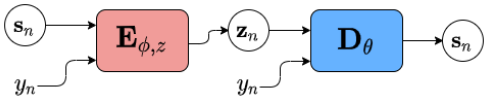
where the first term is the reconstruction loss and $\mathcal{D}_{\text{KL}}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence. However, since the training loss $\mathcal{L}_{\text{ELBO}}$ is unsupervised, there is no explicit control of speech generation.

2.2.2. *Conditional VAE*

The VAE can be conditioned on a label $y_n \in \mathcal{Y}$ describing a speech attribute (e.g. speech activity) that allows for a more explicit control of speech generation [15]. A common approach is to make use of the label y_n by directly inputting it in both the encoder $\mathbf{E}_{\phi,z}(|\mathbf{s}_n|^2, y_n)$ and the decoder $\mathbf{D}_\theta(\mathbf{z}_n, y_n)$ (see Fig 1b) [13, 14, 15]. The training loss remains the same as for the standard VAE, i.e. $\mathcal{L}_{\text{ELBO}}$.



(a) Standard VAE



(b) Conditional VAE [15]

Figure 1: Training scheme for (a) the standard VAE and (b) the conditional VAE.

2.3. Non-negative matrix factorization as noise model

As in Leglaive et al. [8], the noise variance is modeled with an untrained NMF as

$$v_{b,nf} = \{\mathbf{HW}\}_{nf}, \tag{6}$$

where $\mathbf{H} \in \mathbb{R}_+^{N \times K}$ and $\mathbf{W} \in \mathbb{R}_+^{K \times F}$ are two non-negative matrices representing the temporal activations and spectral patterns of the noise power spectrogram. K denotes the NMF rank.

2.4. Clean speech estimation

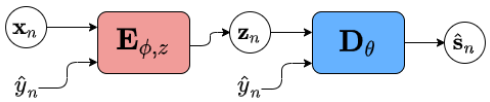


Figure 2: Estimation of the speech variance $v_{s,nf}$ at test time.

Speech variance estimation at test time is shown in Fig. 2. First, a pretrained noise-robust classifier which is trained separately from the VAE provides the label estimate $\widehat{y}_n$ from the mixture $\mathbf{x}_n$ as input for the VAE. The outputs of the encoder and decoder are then $\mathbf{E}_{\phi,z}(|\mathbf{x}_n|^2, \widehat{y}_n)$ and $\mathbf{D}_\theta(\mathbf{z}_n, \widehat{y}_n)$, respectively. Given the speech prior provided by the VAE and the noise model, the mixture signal x_{nf} is distributed as

$$x_{nf}|\, \mathbf{z}_n \sim \mathcal{CN}(0, g_n\{\mathbf{D}_\theta(\mathbf{z}_n, \widehat{y}_n)\}_f + \{\mathbf{HW}\}_{nf}) \tag{7}$$

where $\Theta_u = \{g_n, \mathbf{H}, \mathbf{W}\}$ are the unsupervised parameters to be estimated. Since the resulting optimization problem is intractable due to the non-linear relation between the speech variance $v_{s,nf}$ and the latent variable $\mathbf{z}_n$, an MCEM algorithm is employed to iteratively optimize the unsupervised parameters Θ_u [8]. At each iteration, the estimated terms $\widehat{g}_n\{\mathbf{D}_\theta(\mathbf{z}_n)\}_f$ and $\{\widehat{\mathbf{H}}\widehat{\mathbf{W}}\}_{nf}$ are supposed to get closer to the true variances $v_{s,nf}$ and $v_{b,nf}$, respectively.

Ideally, in order to obtain an explicit control of speech generation, the label y_n should be independent from the latent variable $\mathbf{z}_n$. However, the training loss $\mathcal{L}_{\text{ELBO}}$ does not guarantee that it will promote independence between the label y_n and the latent variable $\mathbf{z}_n$ [16]. As a result, speech generation can only partially be controlled by the label y_n.

3. DISENTANGLEMENT LEARNING

Inspired by recent work on semi-supervised disentanglement learning [22, 23], we propose to use an adversarial training scheme on the latent space to disentangle the label y_n from the latent variable $\mathbf{z}_n$. More particularly, we propose a different training loss for the encoder of the VAE.

3.1. Architecture

Fig. 3 shows the adversarial training scheme. At training, we use a discriminator $\boldsymbol{\delta}_\psi(\cdot)$ that competes with the encoder $\mathbf{E}_{\phi,z}(\cdot)$ of

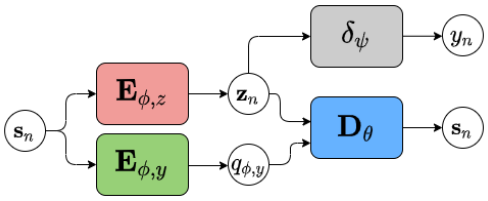


Figure 3: Adversarial training scheme for the conditional VAE

the VAE, which we refer to as the *adversarial-encoder*. The discriminator $\boldsymbol{\delta}_\psi(\cdot)$, parametrized by ψ, aims at identifying the label y_n from the latent variable $\mathbf{z}_n$ whereas the *adversarial-encoder* $\mathbf{E}_{\phi,z}(\cdot)$ aims at making it unable to estimate the label y_n. Simultaneously, we also estimate the label y_n using a classifier that we denote as the *classifier-encoder* $\mathbf{E}_{\phi,y}(\cdot)$, parametrized by ϕ.

Label y_n We consider voice activity detection (VAD) for the label, i.e. $y_n \in \{0, 1\}$, which is here obtained from visual data.

Discriminator $\boldsymbol{\delta}_\psi(\cdot)$ Similarly to Fader Networks for image generation [22], the discriminator $\boldsymbol{\delta}_\psi(\cdot)$ estimates the probability that the label $y_n = 1$, i.e. $p_\psi := P(y_n = 1|\mathbf{z}_n) = \boldsymbol{\delta}_\psi(\mathbf{z}_n)$. We use the binary cross entropy (BCE) as the learning objective:

$$\mathcal{L}_{\text{dis}} = y_n \log p_\psi + (1 - y_n) \log(1 - p_\psi). \tag{8}$$

Adversarial-Encoder $\mathbf{E}_{\phi,z}(\cdot)$ To force the discriminator $\boldsymbol{\delta}_\psi(\cdot)$ to make incorrect classification, adversarial approaches in image generation proposed to use $\mathcal{L}_{\text{adv-enc}} = -\mathcal{L}_{\text{dis}}$ as the training loss for the *adversarial-encoder* $\mathbf{E}_{\phi,z}(\cdot)$ [22, 23]. However, this loss does not make the discriminator $\boldsymbol{\delta}_\psi(\cdot)$ unable to predict the label y_n. Instead, we thus propose to minimize the negative binary entropy function with $q_{\phi,z} := P(y_n = 1|\mathbf{s}_s) = \boldsymbol{\delta}_\psi(\mathbf{E}_{\phi,z}(|\mathbf{s}_n|^2))$:

$$\mathcal{L}_{\text{adv-enc}} = q_{\phi,z} \log q_{\phi,z} + (1 - q_{\phi,z}) \log(1 - q_{\phi,z}). \tag{9}$$

Classifier-Encoder $\mathbf{E}_{\phi,y}(\cdot)$ Similarly to Creswell et al. for image generation [23], we estimate the label y_n for the decoder $\mathbf{D}_\theta(\cdot)$ using the *classifier-encoder* $\mathbf{E}_{\phi,y}(\cdot)$ as $q_{\phi,y} := P(y_n = 1|\mathbf{s}_n) = \mathbf{E}_{\phi,y}(|\mathbf{s}_n|^2)$. We use the BCE as the learning objective:

$$\mathcal{L}_{\text{clf-enc}} = y_n \log q_{\phi,y} + (1 - y_n) \log(1 - q_{\phi,y}). \tag{10}$$

Decoder $\mathbf{D}_\theta(\cdot)$ Instead of using the binary label y_n as an input to the decoder such that $\mathbf{D}_\theta(\mathbf{z}_n, y_n)$, we use the posterior probability $q_{\phi,y}$ as a soft value, such that $\mathbf{D}_\theta(\mathbf{z}_n, q_{\phi,y})$.

3.2. Adversarial training

The complete loss of the conditional VAE is:

$$\mathcal{L}_{\text{adv-VAE}} = \mathcal{L}_{\text{ELBO}} + \alpha\mathcal{L}_{\text{adv-enc}} + \beta\mathcal{L}_{\text{clf-enc}}. \tag{11}$$

At each step of the training, we update the networks as follows:

1. Update $\mathbf{E}_{\phi,z}(\cdot)$, $\mathbf{E}_{\phi,y}(\cdot)$ and $\mathbf{D}_\theta(\cdot)$ using $\mathcal{L}_{\text{adv-VAE}}$
2. Update $\boldsymbol{\delta}_\psi(\cdot)$ using $\alpha\mathcal{L}_{\text{dis}}$[1]

3.3. Clean speech estimation

Speech variance estimation at test time is shown in Fig. 4. At test time, we estimate the label y_n with a pretrained noise-robust classifier which is trained separately from the VAE on visual data. For estimating the unsupervised parameters Θ_u, we use the same MCEM configuration as in Section 2.4.

[1] α is related to $\mathcal{L}_{\text{adv-enc}}$ and proved to provide better results in practice.

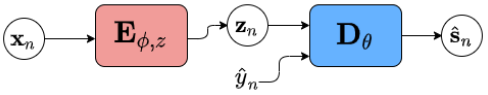


Figure 4: Estimation of the speech variance $v_{s,nf}$ at test time.

4. EXPERIMENTAL SETUP

Dataset For training, validation, and test we used the NTCD-TIMIT dataset [25], which consists of audio-visual recordings. The visual data consist of videos of the lip region at 30 frames-per-second (FPS). All audio signals have a sampling rate of 16 kHz. In addition to clean speech signals, we use the noisy versions, with 2 types of stationary noise {*Car*, *White*} and four types of nonstationary noises {*Living Room*, *Cafe*, *Babble*, *Street*}. We evaluate on 3 signal-to-noise ratios (SNRs): $\{0, +5, +10\}$ dB. Note that these levels were computed on speech segments only. For training, the clean-speech and visual sets used are of about 5 h. For the test, the noisy speech subset consists of 15,876 utterances of about 5 s each.

Baselines For the noise-robust classifier estimating the label y_n at test time, we use the visual-only variant of Ariav and Cohen's audio-visual VAD classifier [26]. We pretrain the classifier on the visual training set of the NTCD-TIMIT dataset. For the baselines, we use the standard and the conditional VAEs, which we denote as *VAE* and *CVAE*, respectively, both trained using $\mathcal{L}_{\text{ELBO}}$.

Hyperparameter settings The STFT is computed using a 64 ms Hann window with 75% overlap, resulting in $F = 513$ unique frequency bins and an audio frame rate of 62.5 FPS. To obtain the same visual frame rate as the audio frame rate, we use FFMPEG [27]. This guarantees that the audio and visual frames are synchronized. To obtain the ground truth for the VAD label y_n, we use simple thresholding in the time domain.

For a fair comparison between all the approaches, we consider a similar architecture for each subnetwork. Tab. 1 shows the configuration of the models. We set the dimension of the latent space to $L = 16$. For the loss, we empirically found that $\alpha = \beta = 10$ provides excellent performance. We use the Adam optimizer with standard configuration [28]. We set the batch size to 128. Early stopping with a patience of 10 epochs is performed using $\mathcal{L}_{\text{adv-enc}}$ on the validation set. For the MCEM we follow the settings of Leglaive et al. and set the NMF rank to $K = 10$ [8].

Metrics To evaluate speech enhancement performance, we use the scale-invariant signal-to-distortion ratio (SI-SDR) measured in dB [29], raw scores of the extended short-time objective intelligibility (ESTOI) with values between 0 and 1 [30], and the perceptual objective listening quality analysis (POLQA) score with values between 1 and 5 [31]. To evaluate the classification performance of the visual-only VAD classifier used at test time, we use the F1-score which combines the precision and recall rates.

	Hidden layers			Output layer
Subnetwork	**# layers**	**# units**	**act. fn**	**act. fn**
$\boldsymbol{\delta}_\psi(\cdot)$	2	128	ReLU	sigmoid
$\mathbf{E}_{\phi,z}(\cdot)$	2	128	tanh	identity
$\mathbf{E}_{\phi,y}(\cdot)$	2	128	ReLU	sigmoid
$\mathbf{D}_\theta(\cdot)$	2	128	tanh	exp

Table 1: Model configurations.

Model	SI-SDR (dB)	ESTOI	POLQA
Mixture	-1.3 ± 0.1	0.38 ± 0.00	1.40 ± 0.01
VAE + $\mathcal{L}_{\text{ELBO}}$	3.9 ± 0.1	0.36 ± 0.00	1.52 ± 0.01
CVAE + $\mathcal{L}_{\text{ELBO}}$	4.6 ± 0.1	0.38 ± 0.00	$\mathbf{1.57 \pm 0.01}$
CVAE + $\mathcal{L}_{\text{adv-VAE}}$	$\mathbf{5.3 \pm 0.1}$	$\mathbf{0.38 \pm 0.00}$	1.56 ± 0.01

Table 2: Average performance for nonstationary noises. ESTOI and POLQA are only evaluated during speech activity while the proposed disentanglement yields strong improvements in speech absence.

Model	SI-SDR (dB)	ESTOI	POLQA
Mixture	-6.8 ± 0.2	0.46 ± 0.00	1.65 ± 0.02
VAE + $\mathcal{L}_{\text{ELBO}}$	6.0 ± 0.2	0.42 ± 0.00	1.57 ± 0.01
CVAE + $\mathcal{L}_{\text{ELBO}}$	$\mathbf{7.7 \pm 0.1}$	$\mathbf{0.46 \pm 0.00}$	$\mathbf{1.67 \pm 0.01}$
CVAE + $\mathcal{L}_{\text{adv-VAE}}$	6.8 ± 0.1	0.43 ± 0.00	1.57 ± 0.01

Table 3: Average performance for stationary noises.

5. RESULTS

5.1. Average performance

The visual-only VAD classifier gives F1-score = 88% on the test set. Fig. 5 shows an example of the reconstructed spectrograms by the VAEs, i.e. without using the MCEM algorithm. The proposed *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ is the only approach with a meaningful output when speech absence is detected, i.e. $\widehat{y}_n = 0$. This illustrates that disentanglement works with the proposed approach.

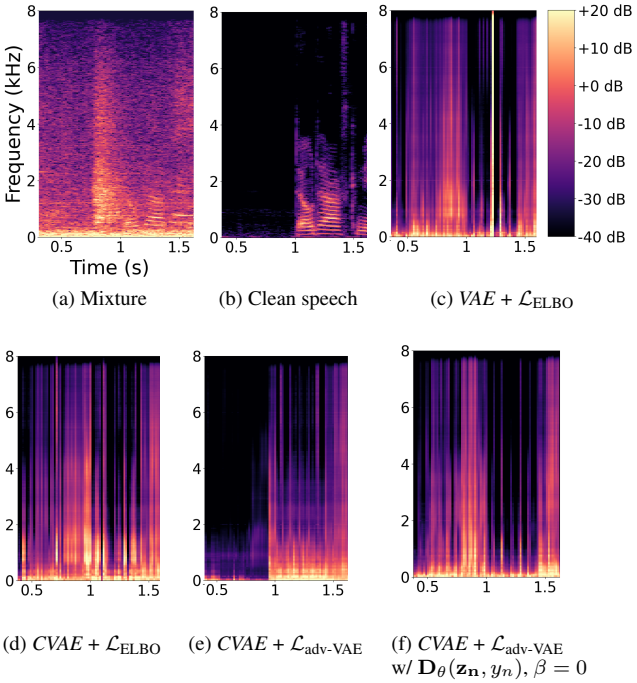
Tab. 2 shows the average results for nonstationary noises. *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ outperforms both *VAE* + $\mathcal{L}_{\text{ELBO}}$ and *CVAE* + $\mathcal{L}_{\text{ELBO}}$ in terms of SI-SDR. This is again explained by our successful disentanglement which prevents *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ from outputting a noise-like signal when speech absence is detected, i.e. $\widehat{y}_n = 0$. Conversely, *VAE* + $\mathcal{L}_{\text{ELBO}}$ and *CVAE* + $\mathcal{L}_{\text{ELBO}}$ often outputs speech-like noise when $\widehat{y}_n = 0$. Note that, unlike SI-SDR, the benefits of the proposed approach can not be visible with ESTOI and POLQA because these metrics are only computed in speech presence.

Tab. 3 shows the average results for stationary noises. *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ outperforms *VAE* + $\mathcal{L}_{\text{ELBO}}$ in terms of SI-SDR. However, *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ is outperfomed by *CVAE* + $\mathcal{L}_{\text{ELBO}}$. This is also explained by our successful disentanglement which forces *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ to output a signal when speech presence is detected, i.e. $\widehat{y}_n = 1$. Since stationary noise is present on the entire utterance, unlike *VAE* + $\mathcal{L}_{\text{ELBO}}$ and *CVAE* + $\mathcal{L}_{\text{ELBO}}$, *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ is forced to output both speech and noise when speech presence is detected which is not always beneficial, but can likely be addressed by more informative labels in future work. Code and audio examples are available online[2].

5.2. Analysis of the training scheme

While the above results show the performance of *CVAE* with the proposed training scheme using $\mathcal{L}_{\text{adv-VAE}}$, we need further analysis regarding what contributes to the performance of the proposed training scheme *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$. Since the proposed disentanglement yields strong improvements in speech absence while ESTOI and POLQA are only evaluated during speech presence, we only consider SI-SDR here.

[2] https://uhh.de/inf-sp-disentangled2021



(a) Mixture (b) Clean speech (c) *VAE* + $\mathcal{L}_{\text{ELBO}}$

(d) *CVAE* + $\mathcal{L}_{\text{ELBO}}$ (e) *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ (f) *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ w/ $\mathbf{D}_\theta(\mathbf{z_n}, y_n), \beta = 0$

Figure 5: Reconstructed spectrograms (without MCEM).

Training scheme	Average	Noise stationarity: Nonstationary	Noise stationarity: Stationary
$\mathcal{L}_{\text{adv-VAE}}$	$\mathbf{5.8 \pm 0.1}$	$\mathbf{5.3 \pm 0.1}$	$\mathbf{6.8 \pm 0.1}$
w/ $\mathbf{D}_\theta(\mathbf{z}_n, y_n), \beta = 0$	5.1 ± 0.1	4.3 ± 0.1	6.8 ± 0.1
w/ $\mathbf{D}_\theta(\mathbf{z}_n, y_n), \beta = 0$, $\mathcal{L}_{\text{adv-enc}} = -\mathcal{L}_{\text{dis}}$	5.1 ± 0.1	4.4 ± 0.1	6.4 ± 0.1

Table 4: Average SI-SDR (in dB) w.r.t. different training schemes.

Tab. 4 shows the results on average and per noise stationarity with variants of the proposed training scheme. The first line corresponds to the proposed training scheme using *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$. The second and third lines correspond to the case where the binary label y_n is used as an input to the decoder $\mathbf{D}_\theta(\mathbf{z}_n, y_n)$, instead of the soft value $q_{\phi,y}$. These training scheme variants are outperformed by the initially-proposed training scheme *CVAE* + $\mathcal{L}_{\text{adv-VAE}}$ in terms of SI-SDR. We can also observe this on Fig. 5f. Thus, we conclude that estimating $q_{\phi,y}$ using the *classifier-encoder* $\mathbf{E}_{\phi,y}(\cdot)$ as an input to the decoder $\mathbf{D}_\theta(\mathbf{z}_n, q_{\phi,y})$ is crucial to learn disentanglement.

6. CONCLUSION

In this work, we propose to use adversarial training to learn disentanglement between a label describing speech activity and the other latent variables of the VAE. We showed the beneficial effect of learning disentanglement when reconstructing clean speech from noisy speech. In the presence of nonstationary noise, the proposed approach outperforms the standard and conditional VAEs, both trained using the ELBO as loss function, in terms of SI-SDR. Our proposed approach is particularly interesting for audio-visual speech enhancement, where speech activity can be estimated from visual information, which is not affected by the noisy environment.

7. REFERENCES

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Hoboken, NJ: John Wiley & Sons, 2018.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art*, ser. Synthesis Lectures on Speech and Audio Processing. Williston, VT: Morgan & Claypool, 2013, no. 11.

[3] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement Without Musical Noise," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.

[4] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[5] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[6] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[7] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization," in *ICASSP*, Apr. 2018, pp. 716–720.

[8] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *MLSP*, Sept. 2018, pp. 1–6.

[9] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder," in *ISCA Interspeech*, 2020, pp. 2437–2441.

[10] J. Richter, G. Carbajal, and T. Gerkmann, "Speech Enhancement with Stochastic Temporal Convolutional Networks," in *ISCA Interspeech*, Oct. 2020, pp. 4516–4520.

[11] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *ICASSP*, 2021, pp. 676–680.

[12] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *NeurIPS*, vol. 27. Curran Associates, Inc., 2014, pp. 3581–3589.

[13] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, Sept. 2019.

[14] Y. Du, K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara, "Semi-supervised Multichannel Speech Separation Based on a Phone- and Speaker-Aware Deep Generative Model of Speech Spectrograms," in *EUSIPCO*, Jan. 2021, pp. 870–874.

[15] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *ICASSP*, June 2021, pp. 681–685.

[16] N. Siddharth, B. Paige, J.-W. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *NeurIPS*, vol. 30. Curran Associates, Inc., 2017.

[17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[18] H. Kim and A. Mnih, "Disentangling by factorising," in *Int. Conf. Machine Learning*, vol. 80. PMLR, July 2018, pp. 2649–2658.

[19] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *NeurIPS*, vol. 31. Curran Associates, Inc., 2018.

[20] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, "Disentangling Disentanglement in Variational Autoencoders," vol. 97. PMLR, June 2019, pp. 4402–4412.

[21] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem, "Disentangling Factors of Variation Using Few Labels," in *ICLR*, 2020.

[22] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader networks: Manipulating images by sliding attributes," in *NeurIPS*, vol. 30. Curran Associates, Inc., 2017, pp. 5969–5978.

[23] A. Creswell, Y. Mohamied, B. Sengupta, and A. A. Bharath, "Adversarial Information Factorization," *arXiv:1711.05175 [cs]*, Sept. 2018.

[24] A. Liutkus, R. Badeau, and G. Richard, "Gaussian Processes for Underdetermined Source Separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, July 2011.

[25] A. H. Abdelaziz, "NTCD-TIMIT: A New Database and Baseline for Noise-Robust Audio-Visual Speech Recognition," in *ISCA Interspeech*, Aug. 2017, pp. 3752–3756.

[26] I. Ariav and I. Cohen, "An End-to-End Multimodal Voice Activity Detection Using WaveNet Encoder and Residual Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 265–274, May 2019.

[27] FFmpeg Developers, "ffmpeg (version 4.3.2)," 2021. [Online]. Available: http://ffmpeg.org/

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, Dec. 2014.

[29] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *ICASSP*, May 2019, pp. 626–630.

[30] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[31] Int. Telecomm. Union (ITU-T) Rec., "Recommendation P.863: Perceptual objective listening quality assessment," 2011.

Supervised Generative Speech Enhancement

4

In this chapter, we explore supervised generative models for speech enhancement, focusing on diffusion models to improve speech quality and address common communication artifacts. We introduce a causal processing framework using these models and extend it to audio-visual speech enhancement by integrating visual information. We investigate alternative diffusion processes, such as the Schrödinger bridge, to enhance computational efficiency and robustness.

Section 4.1 introduces our *SGMSE+* method [P4], utilizing score-based generative models with a novel drift term in the forward SDE to transition from clean to noisy speech, enabling effective sampling of clean speech given noisy inputs. A tutorial on diffusion models for audio restoration, particularly in speech and music, is provided in [P5].

In Section 4.2, we propose a causal processing system using diffusion models trained on various distortions like background noise, reverberation, and codec artifacts [P7]. This approach maintains strong performance without significant loss compared to non-causal models, offering practical adaptability.

Section 4.3 presents *AV-Gen* [P9], an audio-visual extension of *SGMSE+* that incorporates visual data for audio-visual speech enhancement, showing improvements in speech quality and intelligibility in difficult acoustic conditions.

Finally, Section 4.4 investigates training objectives for diffusion-based speech enhancement models and explores the Schrödinger bridge as an alternative diffusion process. A perceptual-based loss function for the Schrödinger bridge is introduced in [P11], which enhances the quality of generated speech indicated by exceptional performance in instrumental quality metrics.

4.1 Diffusion Models for Speech Enhancement

4.1.1 Publication: Speech enhancement and dereverberation with diffusion-based generative models

Abstract

In this work, we build upon our previous publication and use diffusion-based generative models for speech enhancement. We present a detailed overview of the diffusion process that is based on a stochastic differential equation and delve into an extensive theoretical examination of its implications. Opposed to usual conditional generation tasks, we do not start the reverse process from pure Gaussian noise but from a mixture of noisy speech and Gaussian noise. This matches our forward process which moves from clean speech to noisy speech by including a drift term. We show that this procedure enables using only 30 diffusion steps to generate high-quality clean speech estimates. By adapting the network architecture, we are able to significantly improve the speech enhancement performance, indicating that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. We complement the results with an instrumental evaluation using real-world noisy recordings and a listening experiment, in which our proposed method is rated best. Examining different sampler configurations for solving the reverse process allows us to balance the performance and computational speed of the proposed method. Moreover, we show that the proposed method is also suitable for dereverberation and thus not limited to additive background noise removal.

Reference

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, Timo Gerkmann, "Speech Enhancement and Dereverberation with Diffusion-Based Generative Models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2351 - 2364, 2023. DOI: 10.1109/TASLP.2023.3285241

This publication received the **VDE ITG 2024 award**.

Copyright Notice

The following article is the accepted version of the article published with IEEE.
©2023 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Julius Richter and Simon Welker collaboratively designed the original SGMSE algorithm [P6]. Julius Richter, the lead author of this paper, extended it to develop the SGMSE+ algorithm presented here. He conducted the denoising experiments and wrote the majority of the manuscript. Simon Welker contributed to writing and revising the manuscript, prepared figures, engaged in discussions about the method, and conducted experiments. Jean-Marie Lemercier,

the third author of this publication, was responsible for the dereverberation experiments (Sections V-A-3, V-F-7, V-F-8, and VI-B) and assisted in reviewing the manuscript. Bunlong Lay contributed to the denoising experiments by training the ConvTasNet baseline (Section V-F-6) and also helped review the manuscript. Timo Gerkmann provided insights into the experimental validation, mathematical derivations, and reviewed the manuscript.

Speech Enhancement and Dereverberation with Diffusion-based Generative Models

Julius Richter, *Student Member, IEEE*, Simon Welker, *Student Member, IEEE*, Jean-Marie Lemercier, *Student Member, IEEE*, Bunlong Lay, Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—In this work, we build upon our previous publication and use diffusion-based generative models for speech enhancement. We present a detailed overview of the diffusion process that is based on a stochastic differential equation and delve into an extensive theoretical examination of its implications. Opposed to usual conditional generation tasks, we do not start the reverse process from pure Gaussian noise but from a mixture of noisy speech and Gaussian noise. This matches our forward process which moves from clean speech to noisy speech by including a drift term. We show that this procedure enables using only 30 diffusion steps to generate high-quality clean speech estimates. By adapting the network architecture, we are able to significantly improve the speech enhancement performance, indicating that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. We complement the results with an instrumental evaluation using real-world noisy recordings and a listening experiment, in which our proposed method is rated best. Examining different sampler configurations for solving the reverse process allows us to balance the performance and computational speed of the proposed method. Moreover, we show that the proposed method is also suitable for dereverberation and thus not limited to additive background noise removal. Code and audio examples are available online[1].**

***Index Terms*—speech enhancement, dereverberation, diffusion models, score-based generative models, score matching.**

I. Introduction

SPEECH enhancement aims to recover clean speech signals from audio recordings that are impacted by acoustic noise or reverberation [1]. To this end, computational approaches often exploit the different statistical properties of the target and interference signals [2]. Machine learning algorithms can be used to extract these statistical properties by learning useful representations from large datasets. A wide class of methods employed for speech enhancement are discriminative models that learn to directly map noisy speech to the corresponding clean speech target [3]. Common approaches include time-frequency (T-F) masking [4], complex spectral mapping [5], or operating directly in the time domain [6]. These supervised methods are trained with a variety of clean/noisy speech pairs containing multiple speakers, different noise types, and a large range of signal-to-noise ratios (SNRs). However, it is nearly impossible to cover all possible acoustic conditions in the training data to guarantee generalization. Furthermore, some discriminative approaches have been shown to result in unpleasant speech distortions that outweigh the benefits of noise reduction [7].

The use of generative models for speech enhancement, on the other hand, follows a different paradigm, namely to learn a prior distribution over clean speech data. Thus, they aim at learning the inherent properties of speech, such as its spectral and temporal structure. This prior knowledge can be used to make inferences about clean speech given noisy or reverberant input signals that are assumed to lie outside the learned distribution. Several approaches follow this idea and utilized deep generative models for speech enhancement [8]–[17]. Among them are methods that employ likelihood-based models for explicit density estimation such as the variational autoencoder (VAE) [18], or leverage generative adversarial networks (GANs) [19] for implicit density estimation. Bando et al. propose a statistical framework using a VAE trained in an unsupervised fashion to learn a prior distribution over clean speech [9]. At test time they combine the speech model with a low-rank noise model to infer the signal variances of speech and noise to build a Wiener filter for denoising. However, since the VAE is trained with clean speech only, the inference model (i.e. the encoder) that predicts the latent variable remains sensitive to noise. This has been shown to cause the generative speech enhancement method to produce speech-like sounds although only noise is present [9]. To mitigate this, it has been proposed to make the inference model robust to noisy speech by training on labeled data in a supervised manner [14], [15], or by disentangling the latent variable from high-level information such as speech activity which can be estimated by supervised classifiers [12], [13]. Nevertheless, VAE-based speech enhancement methods remain limited due to the dimensionality reduction in the latent layer and the combined use of a linear noise model based on non-negative matrix factorization [9]–[15].

More recently, a new class of generative models called diffusion-based generative models, has been introduced to the task of speech enhancement [20]–[23]. Diffusion-based generative models, or simply *diffusion models*, are inspired by non-equilibrium thermodynamics and exist in several variants

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169), DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380. We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

Simon Welker is with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany, and with the Center for Free-Electron Laser Science, DESY, 22607 Hamburg, Germany (e-mail: simon.welker@uni-hamburg.de). The other authors are all with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany (e-mail: {julius.richter; jeanmarie.lemercier; bunlong.lay; timo.gerkmann}@uni-hamburg.de).

[1]https://github.com/sp-uhh/sgmse

[24]–[26]. All of them share the idea of gradually turning data into noise, and training a neural network that learns to invert this process for different noise scales. More specifically, the inference model is a fixed Markov chain, that slowly transforms the data into a tractable prior, such as the standard normal distribution. The generative model is another Markov chain that is trained to revert this process iteratively [25]. Therefore, diffusion models can be considered as deep latent variable models and have similar properties to VAEs, with the crucial difference that the inference model is not trained and that the latent variables have the same dimensionality as the input. This has the advantage of not relying on surrogate objectives to approximate maximum likelihood training such as the evidence lower bound and enforces no strong restrictions on the model architecture. Recently, diffusion models have been connected with score matching [27] by looking at the stochastic differential equation (SDE) associated with the discrete-time Markov chain [28]. The forward process can be inverted, resulting in a corresponding reverse SDE which depends only on the score function of the perturbed data [29]. Using this continuous-time SDE formalism creates the opportunity to design novel diffusion processes that support the underlying generation task. In contrast to discrete Markov chains, it also allows the use of general-purpose SDE solvers to numerically integrate the reverse process for sampling.

Concerning the application of diffusion models for speech enhancement, there exist currently two approaches that differ conceptually in how the diffusion process is used. One approach is based on speech re-generation, i.e. a diffusion-based vocoder network is used to synthesize clean speech by sampling from an unconditional prior, while a conditioner network takes noisy speech as input and performs the core part of denoising by providing enhanced speech representations to the vocoder network [23], [30]. An auxiliary loss is introduced for the conditioner network to facilitate its ability to estimate clean speech representations [23]. The second approach, on the other hand, does not require any auxiliary loss and is not using two separate models for generation and denoising. Instead, it models the corruption of clean speech by environmental background noise or reverberation directly within the forward diffusion process, so that reversing this process would consequently result in generating clean speech. This has been proposed as a discrete diffusion process for time-domain speech signals [21], and as a continuous SDE-based diffusion process in the complex spectrogram domain [22]. Interestingly, the original denoising score matching objective [31], which is to estimate the white Gaussian noise in the perturbed data, is essentially reminiscent of the goal of speech enhancement, which is to remove interfering noise or reverberation from speech signals. However, under realistic conditions, the environmental noise or reverberation may not match the assumption of stationary white Gaussian noise. Therefore, it was proposed to include real noise recordings in the diffusion process, either by linearly interpolating between clean and noisy speech along the process [21], or by defining such a transformation within the drift term of an SDE [22]. The choice of linear interpolation in [21], however, implies that the trained deep neural network (DNN) must explicitly estimate a portion of environmental noise at each step in the reverse process. This can be seen in the resulting objective function [21, Eq. (21)] which exhibits characteristics of a discriminative learning task. In contrast, an SDE-based formulation results in a pure generative objective function [22, Eq. (9)] and avoids any prior assumptions on the noise distribution.

Nonetheless, note that diffusion-based speech enhancement methods, unlike the VAE-based method described above, are not counted as unsupervised methods, since labeled data (i.e. clean and noisy speech pairs) are used for training. However, the learning objective remains generative in nature which is to learn a prior for clean speech per se rather than a direct mapping from noisy to clean speech. In fact, supervision is only exploited to learn the conditional generation of clean speech when noisy speech is given. Thus, current diffusion-based models for speech enhancement, such as [21]–[23], can be considered as conditional generative models trained in a supervised manner.

In this work, we build upon our previous publication which defines the diffusion process in the complex short-time Fourier transform (STFT) domain [22]. We present a comprehensive theoretical review of the underlying score-based generative model and include an expanded discussion on the conditional generation process which is based on the continuous-time SDE formalism. By using a network architecture developed in the image processing community [28], in the work at hand we significantly improve performance in comparison to our previous model [22]. This indicates that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. To confirm the effectiveness of the proposed method on non-simulated data, we perform an instrumental evaluation with real-world noisy recordings using non-intrusive metrics. We complement the results with a listening experiment, in which our proposed method is rated best. Interestingly, using the improved network, we show that the proposed method is also suitable for dereverberation when an individual model is trained on simulated reverberant data. Thus, the method is not limited to the removal of additive background noise and can also be applied to non-additive corruptions such as reverberation or, as shown in [32], for bandwidth extension. Furthermore, we investigate different sampler configurations for solving the reverse process which reveals a trade-off between the performance and computational speed of the proposed method.

We summarize our major contributions as follows. Regarding the novelty with respect to Song et al. [28], we introduce a drift term to the SDE to achieve the required task adaptation for reconstruction problems and furthermore apply the diffusion process and score matching objective to a complex data representation. Also note that the approach in [28] is not explicitly trained on reconstruction tasks and the application is different from ours. Regarding the novelty with respect to our previous publication [22], we use an improved network architecture and increase the performance significantly. Moreover, we include

an extended theoretical discussion and investigate different sampler configurations. Finally, we expand the evaluation by means of a cross-dataset evaluation, an instrumental evaluation with real-world noisy recordings, and a listening experiment.

II. Method: Score-based Generative Model for Speech Enhancement (SGMSE)

In this section, we motivate and describe in detail the approach of using score-based generative models for speech enhancement, as proposed in our previous publication [22].

A. Data representation

We represent our data in the complex-valued STFT domain, as it has been observed that both real and imaginary parts of clean speech spectrograms exhibit clear structure and are therefore amenable to deep learning models [4]. Following the approach of complex spectral mapping [5], we use our conditional generative model to estimate the clean real and imaginary spectrograms from the noisy ones.

The use of complex coefficients as data representation allows the definition of the diffusion process in the complex spectral domain, in which additive Gaussian noise corresponds to the signal model used for the denoising task. This relates to traditional STFT-based methods, where spectral coefficients are usually assumed to be complex Gaussian distributed and mutually independent [1], [2]. Statistical approaches often consider an additive signal model assuming that the speech process and the noise process are realizations of stochastic processes that are statistically independent. Observing that the overall noise process is a sum of several independent sources, the central limit theorem ensures that the observed noise process tends to be Gaussian [1].

Although it would be theoretically possible to define the diffusion process in the magnitude domain, additive Gaussian noise would not relate to the signal model anymore. This becomes evident considering that in the magnitude domain, additive Gaussian noise could result in negative amplitudes which are physically not defined.

Thus, we operate on complex spectrograms that are elements of $\mathbb{C}^{K\times F}$, where K denotes the number of time frames dependent on the audio length, and F represents the number of frequency bins. To compensate for the typically heavy-tailed distribution of STFT speech amplitudes [33], we apply an amplitude transformation

$$\tilde{c} = \beta|c|^{\alpha}\mathrm{e}^{i\angle(c)} \tag{1}$$

to all complex STFT coefficients c, where $\angle(\cdot)$ represents the angle of a complex number, $\alpha \in (0, 1]$ is a compression exponent which brings out frequency components with lower energy (e.g. fricative sounds of unvoiced speech) [34], and $\beta \in \mathbb{R}_+$ is a simple scaling factor to normalize amplitudes roughly to within $[0, 1]$. Such a compression has been argued to be perceptually more meaningful in speech enhancement [35], [36], and the transformation ensures that the neural network operates on consistently scaled inputs with respect to the Gaussian diffusion noise [25].

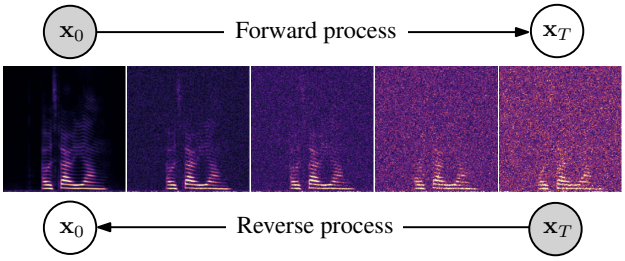


Fig. 1: Diffusion process on a spectrogram: In the forward process noise is gradually added to the clean speech spectrogram $\mathbf{x}_0$, while the reverse process learns to generate clean speech in an iterative fashion starting from the corrupted signal $\mathbf{x}_T$.

B. Stochastic Process

The tasks at hand, speech enhancement and dereverberation, can be considered as conditional generation tasks: Given the corrupted noisy/reverberant speech, generate clean speech by using a conditional generative model. Most previously published diffusion-based generative models are adapted to such conditional tasks either through explicit conditioning channels added to the DNN [37], [38], or through combining an unconditionally trained score model with a separate model (such as a classifier) that provides conditioning in the form of a gradient [28], [39]. With our method, we explore a third possibility, which is to incorporate the particular task directly into the forward and reverse processes of a diffusion-based generative model.

a) Forward Process: Following Song et al. [28], we design a stochastic diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ that is modeled as the solution to a linear SDE of the general form,

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{y})\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{2}$$

where $\mathbf{x}_t$ is the current process state, $t \in [0, T]$ a continuous time-step variable describing the progress of the process (not to be confused with the time index of any signal in the time or T-F domain), $\mathbf{y}$ the noisy or reverberant speech, and $\mathbf{w}$ denotes a standard Wiener process. The vector-valued function $\mathbf{f}(\mathbf{x}_t, \mathbf{y})$ is referred to as the *drift coefficient*, while $g(t)$ is called the *diffusion coefficient* and controls the amount of Gaussian white noise injected at each time-step. Note that different to Song et al. [28], our drift term is now a function of $\mathbf{y}$, by which we tailor the proposed SDE to reconstruction tasks. The process is defined for each T-F bin independently. Thus, the variables in bold are assumed to be vectors in $\mathbb{C}^d$ with $d = KF$ containing the coefficients of a flattened complex spectrogram.

The forward process in Eq. (2) turns a clean speech sample $\mathbf{x}_0$ into a corrupted sample $\mathbf{x}_T$ by gradually adding noise from the Wiener process, as illustrated in Fig. 1. To account for the intended task adaptation of speech enhancement or dereverberation, we propose a drift term that ensures the mean of the process moving from clean speech $\mathbf{x}_0$ to noisy/reverberant speech $\mathbf{y}$. In particular, we define the drift coefficient $\mathbf{f}$ and the diffusion coefficient g as

$$\mathbf{f}(\mathbf{x}_t, \mathbf{y}) := \gamma(\mathbf{y} - \mathbf{x}_t)\,, \tag{3}$$

$$g(t) := \sigma_{\text{min}} \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)^t \sqrt{2 \log \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)}, \tag{4}$$

where γ is a constant called *stiffness* controlling the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and σ_{min} and σ_{max} are parameters defining the noise schedule of the Wiener process. Note that we choose the diffusion coefficient identical to that of the so-called *Variance Exploding SDE* from Song et al. [28]. Our novel contribution lies in the modified drift term, by which the intended task adaptation is achieved.

b) Reverse Process: Following Anderson [29] and Song et al. [28], the SDE in Eq. (2) has an associated *reverse SDE*,

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})\right] \mathrm{d}t + g(t) \mathrm{d}\bar{\mathbf{w}}\,, \tag{5}$$

where the *score* $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is the term to be approximated by a DNN which is therefore called a *score model*. We denote the score model as $\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)$, which is parameterized by a set of parameters θ and receives the current process state $\mathbf{x}_t$, the noisy speech $\mathbf{y}$, and the current time-step t as an input. Finally, by substituting the score model into the reverse SDE in Eq. (5), we obtain the so-called *plug-in reverse SDE* [40],

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t + g(t) \mathrm{d}\bar{\mathbf{w}}\,, \tag{6}$$

which can be solved by various solver procedures, to be discussed in detail in Sec. III.

For inference, we assume that a trained score model $\mathbf{s}_\theta$ is given, which approximates the true score for all $t \in [0, T]$. We can then generate clean speech $\mathbf{x}_0$ conditioned on the noisy or reverberant speech $\mathbf{y}$ by solving the plug-in reverse SDE in Eq. (6). To determine the initial condition of the reverse process at $t = T$, we sample

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma(T)^2 \mathbf{I}), \tag{7}$$

which is a strongly corrupted version of the noisy speech $\mathbf{y}$, as illustrated in Fig. 1. The denoising process which solves the task of speech enhancement or dereverberation is then based on iterating through the reverse process starting at $t = T$ and ending at $t = 0$.

C. Training objective

Next, we derive the objective function used for training the score model $\mathbf{s}_\theta$. Since the SDE in Eq. (2) describes a Gaussian process, the mean and variance of the process state $\mathbf{x}_t$ can be derived when its initial conditions are known [41]. This allows for direct sampling of $\mathbf{x}_t$ at an arbitrary time step t given $\mathbf{x}_0$ and $\mathbf{y}$ by using the so-called *perturbation kernel*,

$$p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_t; \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t), \sigma(t)^2 \mathbf{I}\right), \tag{8}$$

where $\mathcal{N}_{\mathbb{C}}$ denotes the circularly-symmetric complex normal distribution and $\mathbf{I}$ denotes the identity matrix. We utilize Eqs. (5.50, 5.53) in Särkkä & Solin [41] to determine closed-form solutions for the mean

$$\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) = \mathrm{e}^{-\gamma t} \mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t}) \mathbf{y}\,, \tag{9}$$

and the variance

$$\sigma(t)^2 = \frac{\sigma_{\text{min}}^2 \left(\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)^{2t} - \mathrm{e}^{-2\gamma t}\right) \log\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)}{\gamma + \log\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)}\,. \tag{10}$$

Vincent [31] shows that fitting the score model $\mathbf{s}_\theta$ to the score of the perturbation kernel $\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})$ is equivalent to implicit and explicit score matching [27] under some regularity conditions. This technique is called *denoising score matching* and essentially results in estimating

$$\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \nabla_{\mathbf{x}_t} \log \left[|2\pi\sigma\mathbf{I}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{\|\mathbf{x}_t - \boldsymbol{\mu}\|_2^2}{2\sigma^2}}\right] \tag{11}$$

$$= \nabla_{\mathbf{x}_t} \log |2\pi\sigma(t)\mathbf{I}|^{-\frac{1}{2}} - \nabla_{\mathbf{x}_t} \frac{\|\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)\|_2^2}{2\sigma(t)^2} \tag{12}$$

$$= -\frac{\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)}{\sigma(t)^2}\,, \tag{13}$$

where for simplicity we derived the score for the real and imaginary part of the complex normal distribution in Eq. (8), assuming they are independently distributed and each follows a real-valued multivariate normal distribution. Note that Eq. (13) involves division by $\sigma(t)^2$, which has very small numerical values (including 0) around $t = 0$. To avoid undefined values and numerical instabilities, we thus introduce a small minimum process time t_ε, as done previously in the literature [28].

At each training step, the procedure can then be described as follows: 1) sample a random $t \sim \mathcal{U}[t_\varepsilon, T]$, 2) sample $(\mathbf{x}_0, \mathbf{y})$ from the dataset, 3) sample $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; 0, \mathbf{I})$, and 4) sample $\mathbf{x}_t$ from Eq. (8) by effectively computing

$$\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) + \sigma(t)\mathbf{z}. \tag{14}$$

After passing $(\mathbf{x}_t, \mathbf{y}, t)$ to the score model, the final loss is an unweighted L_2 loss between the model output and the score of the perturbation kernel. By substituting Eq. (14) into Eq. (13), the overall training objective becomes,

$$\arg\min_\theta \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})} \left[\left\| \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) + \frac{\mathbf{z}}{\sigma(t)} \right\|_2^2\right], \tag{15}$$

where the expectation is approximated by sampling all random variables at each training step as described above. Note that due to the cancellation of $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)$, the loss function does not explicitly involve $\mathbf{y}$, only as an input to the score model. This means that the score model is not tasked with estimating any portion of the environmental noise directly. Finally, the minimization is achieved by optimizing the parameters θ using stochastic gradient descent.

D. Interpretation and limitations

Let p_t be the distribution of the perturbed data $\mathbf{x}_t$ from the diffusion process for a given dataset. Then its time evolution can be thought of as a continuum of distributions $\{p_t\}_{t\in[0,T]}$ which is determined by the drift and the diffusion coefficient of the forward SDE. For fixed $\mathbf{x}_0$ and $\mathbf{y}$, this time evolution can be described in close form using Eq. (8), which we illustrate in Fig. 2 for a one-dimensional case. In the reverse process, the DNN has the task of learning this continuous family of distributions starting from $\tilde{p}_T$ as defined in Eq. (7). Due to the exponential increase of the diffusion coefficient in the forward process with initial condition $\sigma(0)^2 = 0$ the distribution p_0 essentially corresponds to the clean speech distribution, whereas the terminating distribution p_T is a strongly corrupted version

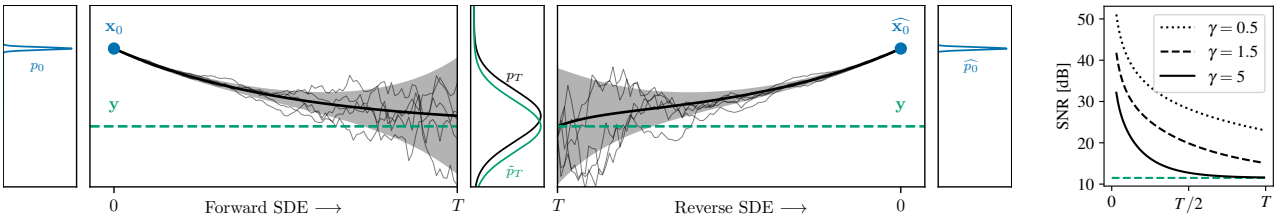


Fig. 2: (Left) The forward and reverse process illustrated with a single scalar variable. The mean $\boldsymbol{\mu}$ (thick black line) of the forward process exponentially decays from clean speech $\mathbf{x}_0$ (blue) towards noisy speech $\mathbf{y}$ (green), and the standard deviation (shaded gray region) increases exponentially. The reverse process moves back to $\mathbf{x}_0$, starting from a slightly mismatched distribution $\tilde{p}_T$ which is centered around $\mathbf{y}$ rather than $\mathbf{x}_T$. Sample paths from both processes are shown as thin black lines. (Right) Time evolution of the SNR of the mean $\boldsymbol{\mu}$ (black) with respect to the SNR of $\mathbf{y}$ (green) for three different values of γ.

of the noisy speech distribution. The particular characteristics in each noisy speech sample are strongly masked by the Gaussian white noise at $t = T$. Therefore, by learning the reverse process, the generative model learns a strong prior p_0 on clean speech, whereas the forward process terminates in a strongly corrupted distribution of the noisy speech, used as a weakly informative prior for generation. In Fig. 2, we simulate five sample paths from the diffusion process. All sample paths of the forward process start exactly at $\mathbf{x}_0$ but exhibit starkly different trajectories at large t. The reverse process should then turn a high-variance sample $\mathbf{x}_T$ back into a low-variance estimate of $\mathbf{x}_0$.

Eq. (9) indicates that the mean $\boldsymbol{\mu}$ of the forward process exponentially decays from $\mathbf{x}_0$ to $\mathbf{y}$, which can also be seen in Fig. 2 (thick black line). However, for finite t it does not fully reach the corrupted speech $\mathbf{y}$ (dashed green line), particularly we have $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, T) \neq \mathbf{y}$. Thus, the final distribution of the forward process p_T exhibits a slight mismatch to the initial distribution of the reverse process $\tilde{p}_T$. We can make this mismatch arbitrarily small by either choosing a high stiffness parameter γ or by increasing σ_{max} to further smooth the density functions of both distributions. However, increasing γ would bring the mean close to $\mathbf{y}$ within a short time of the forward process, which may lead to an unstable reverse process because only the last steps are concerned with removing environmental noise. This effect can be seen in Fig. 2 (right plot), where we plot the SNR of the process mean $\boldsymbol{\mu}$ averaged over 256 randomly selected files from the dataset for three values of γ. Note that we calculated the SNR in the time domain as the ratio of the power of clean speech to the power of environmental noise after inverting the non-linear amplitude transformation of Eq. (1). We see that while for $\gamma = 5$ the mismatch at $t = T$ becomes virtually zero, the change in SNR occurs mainly in the first half of the process. For $\gamma = 0.5$, on the other hand, the mismatch is already more than 10 dB. However, the slope in the SNR is still apparent at the end of the process. Therefore, there is a trade-off to consider when choosing γ which depends on the dataset to be used. Increasing σ_{max} would come at the cost of more reverse iterations since the more white Gaussian noise is added, the less high-level information about the structure of the speech is preserved to serve as a guide in the reverse process. In the experiments, we choose a set of parameters based on empirical hyperparameter optimization.

III. Numerical SDE Solvers

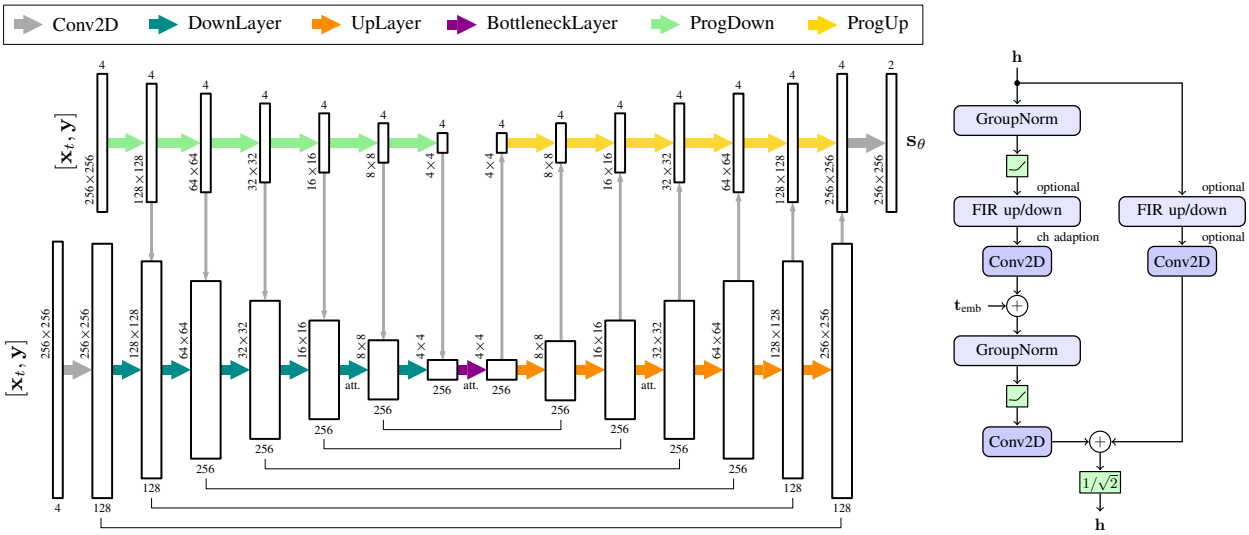
There exist several computational methods to find numerical solutions for SDEs, which are based on an approximation to discrete time steps. To this end, the interval $[0, T]$ is partitioned into N equal subintervals of width $\Delta t = T/N$, which approximates the continuous formulation into the discrete reverse process $\{\mathbf{x}_T, \mathbf{x}_{T-\Delta t}, \ldots, \mathbf{x}_0\}$. A common single-step method for solving this discretization is the Euler-Maruyama method. In each iteration step, the method refers to a previous state of the process and utilizes the drift and the Brownian motion to determine the current state.

In this work, we employ so-called predictor-corrector (PC) samplers proposed by Song et al. [28], which combine single-step methods for solving the reverse SDE with numerical optimization approaches such as annealed Langevin Dynamics [26]. PC samplers consist of two parts, a predictor and a corrector. The predictor can be any single-step method that aims to solve the reverse process by iterating through the reverse SDE. After each iteration step of the predictor, the current state of the process is refined by the corrector. The correction is based on Markov chain Monte Carlo sampling and can be understood as a stochastic gradient ascent optimizer that adds at each iteration step a small amount of noise after taking a step in the direction of the estimated score. One possible intuition about the use of stochastic correctors is that they allow the process state to escape local minima by the use of stochasticity. However, Karras et al. [42] have recently argued that in the reverse process, stochasticity is only necessary to correct for numerical truncation errors of the predictor, a need that could be effectively circumvented by further improving the quality of the score model and predictor.

Another numerical way of approximating the reverse process is by solving the corresponding *probability flow ordinary differential equation (ODE)*,

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t\,, \qquad (16)$$

which is the associated *deterministic process* of the stochastic reverse SDE in Eq. (6). It can be shown that for each diffusion process, there exists an ODE that describes the same marginal probability density $p_t(\mathbf{x}_t)$ [28]. Enhancing the noisy or reverberant mixture is then based on solving this ODE. In Sec. V-E, we also evaluate and compare this class of solvers



(a) Feature maps at multiple resolutions.

(b) Residual block.

Fig. 3: NCSN++ network architecture used as a score model $\mathbf{s}_\theta$: The architecture is based on a multi-resolution U-Net structure containing skip connections and an additional progressive growing path as shown in (a). Each up- and downsampling layer and the bottleneck layer consist of multiple residual blocks in series which are illustrated in (b).

for our task, specifically employing the Runge-Kutta method of fourth order with an error estimator of fifth order [43].

IV. Network Architecture

We utilize the *Noise Conditional Score Network* (NCSN++) architecture [28] for the score model $\mathbf{s}_\theta$ and adapt it for the use of complex spectrograms. For this purpose, we consider the real and imaginary parts of the complex input as separate channels, since the original network only works with real-valued numbers. Estimating both the real and imaginary parts of the score allows to generate complex spectrograms of clean speech.

The network is based on a multi-resolution U-Net structure, which has been experimentally shown to be powerful for tasks such as generation and segmentation [44]. In Fig. 3a, we illustrate the architecture by showing the feature maps at each resolution, indicating their spatial dimension and the corresponding number of channels. The transformations between the feature maps are represented by arrows, where the color of the arrow specifies the type of transformation (see the legend on top). We use Conv2D layers with a 3x3 kernel and stride 1 as input and output layers, and 1x1 Conv2D layers to aggregate information from the progressive growing path that we describe later. Up- and downsampling layers are based on residual network blocks which are taken from the BigGAN architecture [45], shown in Fig. 3b. A residual block consists of Conv2D layers with the same configuration as above, group normalization [46], up- or downsampling with finite impulse response (FIR) filters [47], and the Swish activation function [48]. Each upsampling layer consists of three residual blocks and each downsampling layer of two residual blocks in series with the last block performing the up- or downsampling. Global attention mechanisms [49] are added at a resolution of 16×16 and in the bottleneck layer to better learn global dependencies within the feature maps.

To make the model time-dependent, information about the current progression of the diffusion process is fed into the network architecture. A common practice is to use Fourier-embeddings [49], i.e., a learned projection that maps the scalar time coordinate t to an M-dimensional vector $\mathbf{t}_{\text{emb}}$ that is integrated into every residual block as can be seen in Fig. 3b.

In addition to the main feature extraction path of the multi-resolution U-Net structure, the network incorporates a so-called progressive growing of the input which is seen at the top of Fig. 3a. The idea is to provide a downsampled version of the input to every feature map in the contracting path, which has been successful in stabilizing high-resolution image generation [50]. Note that the downsampling operation in the progressive growing use shared weights for each resolution. The same procedure is also used in the expansive path where a progressive growing of the output is informed by the feature maps at each resolution, resulting in the final score estimate.

V. Experiments

In this section, we describe the experimental setup for our speech enhancement and speech dereverberation experiments using the proposed method.

A. Datasets

For the evaluation of the speech enhancement task we use two datasets, the WSJ0-CHiME3 dataset and the VB-DMD dataset, which are described below. The use of two datasets allows cross-dataset evaluation, i.e. the test is performed on the other dataset than the one used for training. This mismatched condition reveals information about how well the

method generalizes to unseen data with different characteristics such as distinct noise types or different recording conditions. Moreover, to train and evaluate our proposed method on the dereverberation task, we create the WSJ0-REVERB dataset, which is also described below.

a) WSJ0-CHiME3: We create the WSJ0-CHiME3 dataset using clean speech utterances from the Wall Street Journal (WSJ0) dataset [51] and noise signals from the CHiME3 dataset [52]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance. Each utterance is used only once, and the SNR is sampled uniformly between 0 and 20 dB for the training, validation, and test set.

b) VB-DMD: We use the publicly available VoiceBank-DEMAND dataset (VB-DMD) [53] which is often used as a benchmark for single-channel speech enhancement. The utterances are artificially contaminated with eight real-recorded noise samples from the DEMAND database [54] and two artificially generated noise samples (babble and speech shaped) at SNRs of 0, 5, 10, and 15 dB. The test utterances are mixed with different noise samples at SNR levels of 2.5, 7.5, 12.5, and 17.5 dB. We split the training data into a training and validation set using speakers "p226" and "p287" for validation.

c) WSJ0-REVERB: To create the WSJ0-REVERB dataset, we use clean speech data from the WSJ0 dataset [51] and convolve each utterance with a simulated room impulse response (RIR). We use the `pyroomacoustics` engine [55] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a T_{60} between 0.4 and 1.0 seconds. A dry version of the room is generated with the same geometric parameters but a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target. The resulting average direct-to-reverberant ratio (DRR) is around -9 dB.

B. Instrumental evaluation metrics

To evaluate the performance of the proposed method we use standard metrics which we will describe in detail below. Metrics (a)-(d) employ full reference algorithms that rate the processed signal in relation to the clean reference signal using conventional digital signal analysis. On the other hand, metrics (e)-(g) are non-intrusive metrics that can be used to evaluate real recordings when the clean reference is unavailable.

a) POLQA: The Perceptual Objective Listening Quality Analysis (POLQA) is an ITU-T standard that includes a perceptual model for predicting speech quality [56]. The POLQA score takes values from 1 (poor) to 5 (excellent) as usual for mean opinion scores (MOS).

b) PESQ: The Perceptual Evaluation of Speech Quality (PESQ) is used for objective speech quality testing and is standardized in ITU-T P.862 [57]. Although it is the predecessor of POLQA, it is still widely used in the research community. The PESQ score lies between 1 (poor) and 4.5 (excellent) and there exist two variants, namely wideband PESQ and narrowband PESQ denoted as $\mathrm{PESQ}_{\mathrm{nb}}$.

c) ESTOI: The Extended Short-Time Objective Intelligibility (ESTOI) is an instrumental measure for predicting the intelligibility of speech subjected to various kinds of degradation [58]. The metric is normalized and lies between 0 and 1, with higher values indicating better intelligibility.

d) SI-SDR, SI-SIR, SI-SAR: Scale-Invariant (SI-) Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) are standard evaluation metrics for single-channel speech enhancement and speech separation [59]. They are all measured in dB, with higher values indicating better performance.

e) DNSMOS: The Deep Noise Suppression MOS (DNS-MOS) is a reference-free metric to evaluate perceptual speech quality [60]. The evaluation method uses a DNN that is trained on human ratings obtained by using an online framework for listening experiments [61] based on ITU-T P.808 [62].

f) SIG, BAK, OVRL: The non-intrusive speech quality assessment model DNSMOS P.835 [63] is based on a listening experiment according to ITU-T P.835 [64] and provides three MOS scores: speech quality (SIG), background noise quality (BAK), and the overall quality (OVRL) of the audio.

g) WVMOS: Wav-to-Vec MOS (WVMOS) [65] is a MOS prediction method for speech quality evaluation using a fine-tuned wav2vec2.0 model [66].

C. Listening Experiment

Instrumental evaluation metrics do not always correlate to human perception because there are many aspects of perception that are very difficult to capture by computational means. Therefore, we conduct a MUSHRA listening experiment [67] with ten participants using the webMUSHRA framework [68]. The participants were asked to rate the overall quality of twelve randomly sampled examples from the WSJ0-CHiME3 test set as reconstructed by the compared algorithms. The results are reported on a quality scale from 0 to 100.

D. Hyperparameters and training configuration

a) Input representation: We convert each audio input with sampling rate 16 kHz into a complex-valued STFT representation using a window size of 510, resulting in $F = 256$, a hop length of 128 (i.e. approximately 75% overlap), and a periodic Hann window. To process multiple examples for batch training, the length of each spectrogram is trimmed to $K = 256$ STFT time frames, with start and end times selected randomly at each training step. For the spectrogram transformation in Eq. (1), we have chosen $\alpha = 0.5$ and $\beta = 0.15$ empirically.

b) Stochastic process: The SDE in Eq. (2) is parameterized with $\sigma_{\mathrm{min}} = 0.05, \sigma_{\mathrm{max}} = 0.5$, and $\gamma = 1.5$ based on hyperparameter optimization with grid search.

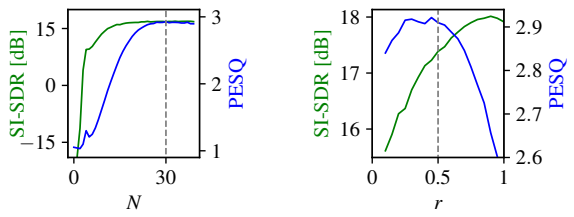
c) Training configuration: We train the DNN on four Quadro RTX 6000 (24 GB memory each) for 160 epochs using the distributed data-parallel (DDP) approach in PyTorch Lightning [69], which takes about one day. We use the Adam optimizer [70] with a learning rate of 10^{-4} and an effective batch size of $4 \times 8 = 32$. We track an exponential moving average of the DNN weights with a decay of 0.999, to be used for sampling [71]. We log the average PESQ value of 20 randomly chosen examples from the validation set during training and select the best-performing model for evaluation.

E. Sampler settings

To find optimal sampler settings for the reverse process, we run a hyperparameter search using the VB-DMD dataset.

TABLE I: Results for different sampler configurations tested on VB-DMD with the average number of function evaluations (NFE) and the respective average real-time factor (RTF)[2].

Type	Sampler settings	NFE	RTF	PESQ	SI-SDR [dB]
PC	0 corrector steps	30	0.89	2.80	15.38
PC	1 corrector steps	60	1.77	2.93	17.35
PC	2 corrector steps	90	2.65	2.92	17.52
ODE	atol=10^{-1}, rtol=10^{-1}	14	0.46	2.78	12.83
ODE	atol=10^{-6}, rtol=10^{-3}	49	1.55	2.71	12.76



(a) Varying N, fix $r = 0.33$ (b) Varying r, fix $N = 30$

Fig. 4: Model performance in PESQ and SI-SDR as a function of (a) the number of reverse steps N and (b) the step size parameter r in the annealed Langevin corrector.

a) Sampler type: We investigate which choice of sampler yields the best speech enhancement performance, comparing the PC sampler with different numbers of corrector steps and an ordinary ODE sampler as described in Sec. III. In Tab. I we can see that use of one correction step in the PC sampler seems to be advantageous, but the use of two steps does not lead to a further increase in performance. Thus, we decide to use the PC sampler with one corrector step for the evaluation. However, it should be noted that the use of one correction step doubles the number of function evaluations (NFE) of the sampler. The *function* being the expensive score model, this results in an average real-time factor (RTF) of 1.77, i.e., 1 sec of audio requires 1.77 sec of processing[2]. Comparing the PC sampler with the ODE sampler, we find that the PC sampler performs better in both metrics. However, with suitable settings, the ODE sampler requires only 14 NFE on average which results in an improved RTF of only 0.46.

b) Number of reverse steps N: The number of reverse steps N can be used to set a balance between the computational effort and the performance of the model. In Fig. 4a, we show the speech enhancement performance as a function of N. It can be seen that SI-SDR starts to stagnate earlier than PESQ. We opt for a value of $N = 30$, at which both metrics show no further increase in performance.

c) Step size in corrector: In Fig. 4b, we vary the step size r of the annealed Langevin dynamics in the corrector. Interestingly, this parameter represents a compromise between PESQ and SI-SDR. We choose $r = 0.5$ to achieve a maximum PESQ value while still obtaining a good value for SI-SDR.

[2]Average processing time for 10 audio files on an NVIDIA GeForce RTX 2080 Ti GPU, in a machine with an Intel Core i7-7800X CPU @ 3.50GHz.

F. Baselines

We compare the performance of our proposed method with four generative and four discriminative baselines which we describe in more detail below. All methods are re-trained by us, except for DVAE, MetricGAN+, and CDiffuSE on VB-DMD, for which we obtained the pre-trained model from the authors who used the exact same training data.

a) STCN [11]: A generative VAE-based speech enhancement method which uses a stochastic temporal convolutional network (STCN) [72] that allows the latent variables to have both hierarchical and temporal dependencies. The parameters of the noise model and the latent variables are estimated using a Monte Carlo expectation maximization (MCEM) algorithm.

b) DVAE [17]: Generative speech enhancement method based on an unsupervised dynamical VAE (DVAE) [73] which models temporal dependencies between successive observable and latent variables. Parameters are updated at test time using a variational expectation maximization (VEM) method where the encoder is fine-tuned using stochastic gradient ascent.

c) CDiffuSE [21]: Most related to our proposed method is CDiffuSE, a generative speech enhancement method based on a conditional diffusion process defined in the time domain.

d) SGMSE [22]: Score-based Generative Model for Speech Enhancement (SGMSE) is our previous publication on which the proposed method is based. The main difference is that it uses a deep complex U-Net [74] instead of the NCSN++ architecture as the score model.

e) MetricGAN+ [75]: A discriminative speech enhancement method that uses a generator network for mask-based prediction of clean speech and introduces a discriminator network trained to approximate the PESQ score.

f) Conv-TasNet [76]: An end-to-end neural network that estimates a mask that is used for filtering a learned representation of the noisy mixture. The filtered representation is transformed back to the time domain by a learned decoder.

g) GaGNet [77]: This neural network is trained on a hybrid complex-domain and magnitude-domain regression objective for single-channel dereverberation. It uses so-called "glance" and "gaze" (GaG) modules, which respectively perform a coarse estimation of the magnitude and refine it with phase estimation in the complex domain.

h) TCN+SA+S [78]: This single-channel dereverberation approach uses a self-attention module to extract features from the input magnitude. This representation is then used by a temporal convolutional network followed by a single-layer convolutional smoother that outputs a magnitude estimate, which is used as the training objective. Griffin-Lim iterations are used to reconstruct the phase.

VI. Results

A. Speech Enhancement

In Tab. II, we report the speech enhancement results on the WSJ0-CHiME3 test set for the matched and mismatched condition, i.e. when the training set was also WSJ0-CHiME3 or when the training set was VB-DMD. We compare our proposed method, which we call *SGMSE+*, with selected baseline methods and sort the results by the type of algorithm,

TABLE II: Speech enhancement results obtained for WSJ0-CHiME3 under matched and mismatched training conditions. Values indicate mean and standard deviation. Methods are sorted by the algorithm type, generative (G) or discriminative (D).

Method	Type	Training set	POLQA	PESQ	ESTOI	SI-SDR [dB]	SI-SIR [dB]	SI-SAR[dB]	DNSMOS
Mixture	-	-	2.64 ± 0.68	2.01 ± 0.55	0.81 ± 0.12	13.5 ± 4.7	18.7 ± 5.5	15.4 ± 4.7	3.34 ± 0.37
STCN [11]	G	WSJ0	2.64 ± 0.68	2.01 ± 0.55	0.81 ± 0.12	13.5 ± 4.7	18.7 ± 5.5	15.4 ± 4.7	3.34 ± 0.37
RVAE [17]	G	WSJ0	2.97 ± 0.63	2.31 ± 0.55	0.85 ± 0.11	15.8 ± 5.0	21.6 ± 6.1	17.6 ± 4.9	3.61 ± 0.29
CDiffuse [21]	G	WSJ0-C3	3.08 ± 0.58	2.27 ± 0.51	0.83 ± 0.09	9.2 ± 2.3	19.8 ± 5.9	10.0 ± 2.3	3.43 ± 0.32
SGMSE [22]	G	WSJ0-C3	2.98 ± 0.60	2.28 ± 0.57	0.86 ± 0.09	14.8 ± 4.3	25.4 ± 5.6	15.3 ± 4.2	3.70 ± 0.27
SGMSE+	G	WSJ0-C3	$\mathbf{3.73 \pm 0.53}$	2.96 ± 0.55	0.92 ± 0.06	18.3 ± 4.4	$\mathbf{31.1 \pm 4.6}$	18.6 ± 4.5	$\mathbf{3.99 \pm 0.19}$
MetricGAN+ [75]	D	WSJ0-C3	3.52 ± 0.61	$\mathbf{3.03 \pm 0.45}$	0.88 ± 0.08	10.5 ± 4.5	24.5 ± 5.1	10.7 ± 4.6	3.67 ± 0.30
Conv-TasNet [76]	D	WSJ0-C3	3.65 ± 0.54	2.99 ± 0.58	$\mathbf{0.93 \pm 0.05}$	$\mathbf{19.9 \pm 4.3}$	29.2 ± 4.6	$\mathbf{20.6 \pm 4.5}$	3.79 ± 0.27
STCN [11]	G	VB	2.53 ± 0.66	1.80 ± 0.45	0.79 ± 0.12	11.9 ± 4.5	17.3 ± 4.9	13.8 ± 4.6	3.40 ± 0.34
RVAE [17]	G	VB	2.84 ± 0.61	2.08 ± 0.49	0.82 ± 0.11	13.9 ± 4.8	19.5 ± 5.9	15.8 ± 4.7	3.52 ± 0.31
CDiffuse [21]	G	VB-DMD	2.15 ± 0.57	1.79 ± 0.42	0.71 ± 0.11	3.2 ± 3.2	21.8 ± 7.0	3.4 ± 3.2	3.17 ± 0.29
SGMSE [22]	G	VB-DMD	2.66 ± 0.58	1.94 ± 0.47	0.81 ± 0.11	13.3 ± 4.3	23.5 ± 6.0	13.8 ± 4.2	3.76 ± 0.25
SGMSE+	G	VB-DMD	$\mathbf{3.43 \pm 0.61}$	$\mathbf{2.48 \pm 0.58}$	$\mathbf{0.90 \pm 0.07}$	$\mathbf{16.2 \pm 4.1}$	$\mathbf{28.9 \pm 4.6}$	$\mathbf{16.4 \pm 4.1}$	$\mathbf{4.00 \pm 0.19}$
MetricGAN+ [75]	D	VB-DMD	2.47 ± 0.67	2.13 ± 0.53	0.76 ± 0.12	6.8 ± 3.1	22.9 ± 4.9	7.0 ± 3.1	3.51 ± 0.29
Conv-TasNet [76]	D	VB-DMD	3.13 ± 0.60	2.40 ± 0.53	0.88 ± 0.08	15.2 ± 3.9	26.5 ± 4.6	15.6 ± 4.0	3.68 ± 0.30

which is either generative or discriminative. Considering the matched condition in the upper half of Tab. II, we see that SGMSE+ outperforms all other generative methods in all metrics. Note that STCN and RVAE are both unsupervised speech enhancement methods, i.e. they are trained on clean speech only (WSJ0 or VB). RVAE shows competitive results for SI-SAR, however, its VEM optimization algorithm is very time-consuming due to the fine-tuning of the encoder at test time, resulting in a RTF of $>$10000. This is significant in contrast to STCN with a RTF of 0.64 and SGMSE+ with a RTF of 1.77[2]. Although both VAE-based methods model temporal dependencies, they are limited in their ability to produce high-quality speech, likely due to the dimensionality reduction of the latent variable and the encoder's sensitivity to noisy input, which causes the latent variable to be incorrectly initialized [15].

Comparing SGMSE+ to our previous model SGMSE, we find a significant improvement, especially for the perceptual metrics. We report improvements of 0.75 for POLQA and 0.68 for PESQ. This shows that the proposed generative diffusion process benefits significantly from the adapted network architecture. In our previous paper [22], we have already shown improvements over CDiffuSE for SGMSE in SI-SDR and SI-SAR, which we now back up with also reporting an improvement in ESTOI and DNSMOS and on par results in PESQ and POLQA. With SGMSE+, these improvements become even more significant, e.g. with 0.65 improvement in POLQA and 9.1 dB in SI-SDR compared to CDiffuSE. In qualitative analysis, we found that SGMSE+ is more accurate than CDiffuSE in preserving the high frequencies of fricatives after the completion of the reverse process. To compensate for that, CDiffuSE combines the enhanced files with the original noisy speech signal at a ratio of 0.2 for the final prediction [21]. This results in a trade-off between noise removal and the conservation of the signal. In our proposed approach, on the other hand, we found no significant suppression of high frequencies after completing the reverse process. Therefore, it is not necessary to mix back the noisy mixture to improve the signal quality, resulting in a significantly higher SI-SIR.

The comparison with Conv-TasNet and MetricGAN+ shows that SGMSE+ can keep up with the performance of discriminative methods and even surpasses them in terms of POLQA, SI-SIR, and DNSMOS. Discriminative methods are based on regression problems that optimize certain point-wise loss functions between the corrupted speech and a clean speech reference. For Conv-TasNet and MetricGAN+ these loss functions correspond to established intrusive metrics, namely SI-SDR for Conv-TasNet and PESQ for MetricGAN+. Note that both these discriminative methods shine in particular on the respective metric they used as a loss function. In contrast, generative methods like SGMSE+ are usually not trained to achieve the exact reconstruction of the reference clean speech but rather aim at generating a realization of speech that is on the manifold of clean speech. Thus, we suggest the use of non-intrusive metrics as a complementary measure since they allow an estimation of speech quality without relying on the exact reconstruction of a reference signal. In fact, for the non-intrusive metric DNSMOS our proposed method yields a significantly higher value than the discriminative baselines, indicating the strong ability of our generative model to generate high-quality clean speech.

Looking at the results for the mismatched condition in the bottom half of Tab. II, a general trend of decreasing metrics can be seen for all methods when compared to the corresponding values of the matched condition. This was to be expected since particular properties of the mismatched test set, such as distinct noise types or different recording characteristics of the clean speech have not been seen during training. However, generative methods generally show less degradation in the mismatched condition than discriminative methods. CDiffuSE is an exception, as this method shows significant degradation in the mismatched case. Informal listening reveals a problem with gain control, which is evident in strong volume fluctuations in the enhanced files. Furthermore, we see that SGMSE+ outperforms all other methods in all metrics under this condition, which shows the ability of our proposed method to generalize well.

Complementary to the average results above, we present in

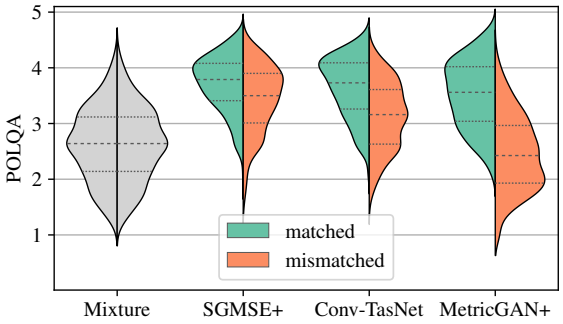


Fig. 5: Violin plots showing POLQA results for the matched and the mismatched condition with dashed and dotted lines representing median and quartiles, respectively.

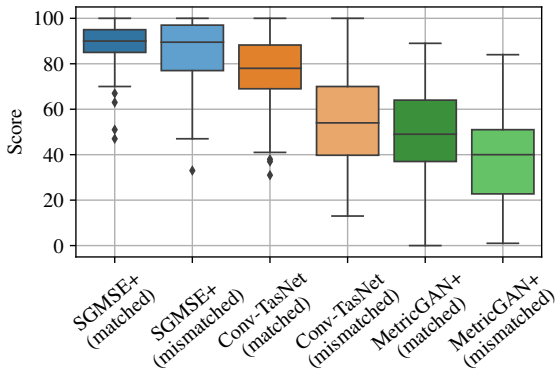


Fig. 6: Boxplot showing the results of the MUSHRA listening experiment with ten participants on twelve randomly selected examples.

Fig. 5 violin plots of the full distribution of the POLQA scores obtained for SGMSE+, Conv-TasNet, MetricGAN+, and the noisy mixture for reference. For each method, the distributions are plotted side by side for the matched and mismatched conditions, so that the ability to generalize can be inferred from the horizontal alignment between both distributions. It can be seen that both distributions for SGMSE+ are relatively similar, whereas they are skewed for Conv-TasNet and especially for MetricGAN+.

In Fig. 6, we report the results of the MUSHRA listening experiment in a boxplot. On average, the ten participants rated the overall quality of our proposed approach with the highest score. In addition, our method remains fairly robust when the model was trained on a different training set, while discriminative methods show much stronger degradation for the mismatched condition. This also corresponds with the results of the non-intrusive metric DNSMOS in Tab. II and thus supports the use of non-intrusive methods for instrumental evaluation. Interestingly, MetricGAN+ was only rated with a median score less than 50 for the matched condition, although the method performed best among all baselines for PESQ (see Tab. II). This reveals the discrepancy between the use of instrumental metrics for evaluation and people's actual perceptions. We suspect that MetricGAN+ has simply learned to utilize the internal operations of the PESQ algorithm to obtain a high value in this metric, neglecting the naturalness of the clean speech estimate. In fact, listening to the enhanced files, it can be recognized that the energy of the speech signal estimated by MetricGAN+ is mainly concentrated in the low- and mid-frequency area of the spectrogram, while high frequencies are strongly attenuated.

Listening to the enhanced files of our method, we notice that at very low input SNRs, some "vocalizing" artifacts with very poor articulation and no linguistic meaning are occasionally produced. In other examples, we find that breathing sounds or speech-like sounds were generated in noisy regions where no speech was originally present. These artifacts may also explain the outliers of our method in the listening experiment (see Fig. 6). For the matched condition, for example, the two lowest outliers come from the same utterance with clearly noticeable vocalizing artifacts. We hypothesize that these artifacts can be linked to the generative nature of the proposed approach.

TABLE III: Speech enhancement results obtained for VB-DMD. Models marked with an asterisk (*) are additional baselines with values taken from the corresponding papers.

Method	PESQ	PESQ_{nb}	ESTOI	SI-SDR	DNSMOS
Mixture	1.97	2.88	0.79	8.4	3.09
SEGAN* [8]	2.16	-	-	-	-
RVAE [17]	2.43	3.11	0.81	16.4	3.30
MetricGAN-U* [82]	2.45	-	0.77	8.2	-
CDiffuse [21]	2.52	3.31	0.79	12.4	3.09
SGMSE [22]	2.28	3.22	0.80	16.2	3.46
SGMSE+	2.93	**3.66**	**0.87**	17.3	**3.56**
UMX* [83]	2.35	-	0.83	14.0	-
Conv-TasNet [76]	2.63	3.42	0.85	**19.1**	3.37
MetricGAN+ [75]	**3.13**	3.63	0.83	8.5	3.37

Indeed, for very noisy inputs, the score model may erroneously identify noise energy in some T-F areas as corrupted speech. The reverse diffusion process then produces speech where it did not originally exist. We argue that this behavior could be mitigated if some conditioning with respect to speech activity and phoneme identity would be added to the score model.

Finally, Tab. III lists the results for the standardized VB-DMD dataset. This has the advantage that one can take values from other methods and copy them from the corresponding papers for a quick algorithmic comparison. It can be seen that SGMSE+ outperforms all other generative baselines, further narrowing the performance gap with discriminative methods that currently lead the benchmark based on PESQ[3], including recent approaches such as [79] and [80]. It should however be noted that PESQ formally requires a minimum file length of 3.2 sec according to P.862.3 [81], which is not the case for most files in VB-DMD [53].

Investigating whether phase estimation has actually been improved with the modeling of the complex coefficients, we use the noisy phase in place of the estimated phase which does not show a significant performance difference. This is also in line with a recent study on the role of phase enhancement, where it has been shown that the impact of phase enhancement is rather small for $\sim$32 ms spectral analysis frames but increasingly

[3]https://paperswithcode.com/sota/speech-enhancement-on-demand

TABLE IV: Single-channel dereverberation results obtained for WSJ0-REVERB test set. Values indicate mean and standard deviation. Methods are sorted by the algorithm type which is either generative (G) or discriminative (D).

Method	Type	POLQA	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR
Mixture	-	1.76 ± 0.29	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5	-7.5 ± 5.4	-
SGMSE [22]	G	1.79 ± 0.28	1.35 ± 0.15	0.57 ± 0.07	-7.4 ± 5.8	-1.1 ± 7.0	-6.2 ± 5.5
SGMSE+	G	**3.24 ± 0.46**	**2.66 ± 0.48**	**0.84 ± 0.07**	**1.6 ± 7.8**	9.4 ± 10.2	**2.3 ± 7.2**
Conv-TasNet [76]	D	2.41 ± 0.52	1.84 ± 0.42	0.73 ± 0.10	**1.6 ± 5.3**	**12.1 ± 5.1**	1.9 ± 5.4
TCN+SA+S [78]	D	2.92 ± 0.33	2.29 ± 0.36	0.79 ± 0.05	-4.4 ± 5.3	-2.3 ± 5.2	-0.6 ± 5.1
GaGNet [77]	D	2.62 ± 0.47	1.98 ± 0.46	0.73 ± 0.08	-0.6 ± 4.9	6.1 ± 3.9	0.4 ± 5.1

large with shorter frame lengths [84].

B. Dereverberation

We report in Tab. IV the performance of our approach when trained and tested on a single-channel dereverberation task. We compare with SGMSE [22] and three discriminative baselines, namely Conv-TasNet [76], GaGNet [77] and TCN+SA+S [78].

Our proposed SGMSE+ approach performs particularly well in terms of instrumental metrics compared to all other baseline models. The low average input DRR of -9 dB constitutes a real challenge for discriminative approaches, which do not manage to separate the reverberation from the target without distorting the target signal, resulting in low-quality scores. On the other hand, our approach benefits from generative modeling and is able to reconstruct speech with very high quality in most cases. When comparing our previous SGMSE model [22] with SGMSE+, we see that for speech dereverberation, the method benefits greatly from the improved network architecture. This effect is even more significant than for additive background noise removal in the speech enhancement task.

In particular, using the proposed approach SGMSE+ on a single-channel dereverberation task does not produce any of the vocalized artifacts observed in the speech enhancement experiments for low input SNRs. Although the reverberant signal is formally decorrelated in the time domain from the target by the randomness of reflections across the room, it still originates from the dry speech source. Therefore, we conjecture that the score model effectively detects whether the energy in a particular time-frequency area is associated with the clean speech nearby that needs to be reconstructed.

C. Evaluation on real data

Complementing the experiments using simulated data, we evaluate the speech enhancement performance on real-world noisy recordings. For real-world noisy recordings, there exists no clean speech reference. Thus, we can only non-intrusive metrics to evaluate the perceptual speech quality which we describe in Sec. V-B (e)-(g). For the evaluation, we use 300 files from the test set of the Deep Noise Suppression (DNS) Challenge 2020 [85]. In Tab. V, we report the results for models that were trained on VB-DMD. It turns out that our proposed method performs better than all other methods in all non-intrusive metrics, demonstrating its robustness to real-world noisy examples. Interestingly, a trend of degradation in speech quality (SIG) can be observed for the discriminative methods, whereas all generative models improve this metric with respect to the mixture. For the background noise quality (BAK) metric, on the other hand, discriminative models seem to perform well, yet our proposed method performs superior. It is important to note that non-intrusive metrics do not require a corresponding clean reference signal and only assess speech quality based on the method's estimate. We hypothesize that our generative model works well on these metrics, as it was trained to generate clean speech. However, "vocalizing" artifacts as mentioned above or phonetic confusions may not be captured with these metrics.

TABLE V: Speech enhancement results obtained for real-world noisy recordings from the DNS Challenge 2020 test set.

Method	DNSMOS	SIG	BAK	OVRL	WVMOS
Mixture	3.05	3.05	2.51	2.26	1.12
RVAE [17]	3.29	3.16	2.91	2.44	1.87
CDiffuse [21]	3.14	3.15	3.19	2.55	1.86
SGMSE [22]	3.38	3.22	3.02	2.52	1.80
SGMSE+	**3.64**	**3.42**	**3.82**	**3.04**	**2.54**
Conv-TasNet [76]	3.07	2.87	3.59	2.52	2.07
MetricGAN+ [75]	3.26	2.88	3.39	2.45	1.52

We provide on our project page[4] some listening examples for all evaluated tasks. Furthermore, we include real reverberant examples from the MC-WSJ-AV dataset [86].

VII. Conclusions

In this work, we built upon our existing work [22] that uses a novel stochastic diffusion process to design a generative model for speech enhancement in the complex STFT domain. We presented an extended theoretical analysis of the underlying score-based generative model and derived in detail the objective function used for training. In further explorations, we considered the time evolution of the conditional diffusion process which revealed a slight mismatch between the forward and reverse process, which can be adjusted with a careful parameterization of the forward SDE.

By using an adopted network architecture, we were able to significantly improve the performance compared to our previous model. In addition, we trained and evaluated the proposed method on the task of speech dereverberation and show significantly superior performance compared to discriminative baseline methods. Hence, we showed that with our proposed

[4]https://uhh.de/inf-sp-sgmse

method, a single framework can be used to train individual models for different distortion types. For the task of speech enhancement, we evaluated performance under matched and mismatched conditions, i.e. when the training and test data were taken from the same or different corpora. For the matched condition, the proposed generative speech enhancement method performs on par with competetive discriminative methods. For the mismatched condition, our method shows strong generalization capabilities and outperforms all baselines in all metrics, as confirmed by a listening experiment. In very adverse conditions, however, we observe that the proposed method sometimes introduces vocalizing and breathing artifacts. We argue that these could be mitigated in future work if some conditioning concerning speech activity and phoneme information would be added to the score model.

In addition, we explored different sampling strategies to solve the reverse process at test time which allows us to balance the performance and computational speed of the proposed method. Future work could include other sampling techniques to further reduce the number of diffusion steps [87] and thus the computational complexity.

REFERENCES

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state-of-the-art*. Morgan & Claypool, 2013.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. John Wiley & Sons, 2018.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 24, no. 3, pp. 483–492, 2015.

[5] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, pp. 1–6, 2017.

[6] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," *IEEE Asia-Pacific Signal and Inf. Proc. Assoc. Annual Summit and Conf. (APSIPA ASC)*, 2017.

[7] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 39–48, 2020.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *ISCA Interspeech*, pp. 3642–3646, 2017.

[9] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 716–720, 2018.

[10] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, pp. 1–6, 2018.

[11] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks." *ISCA Interspeech*, pp. 4516–4520, 2020.

[12] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 681–685, 2021.

[13] ——, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics (WASPAA)*, pp. 126–130, 2021.

[14] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder." *ISCA Interspeech*, pp. 2437–2441, 2020.

[15] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 676–680, 2021.

[16] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 1104–1117, 2020.

[17] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 30, pp. 2993–3007, 2022.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Int. Conf. on Learning Representations (ICLR)*, 2014.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 27, 2014.

[20] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *IEEE Asia-Pacific Signal and Inf. Proc. Assoc. Annual Summit and Conf. (APSIPA ASC)*, pp. 659–666, 2021.

[21] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 7402–7406, 2022.

[22] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," *ISCA Interspeech*, pp. 2928–2932, 2022.

[23] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[24] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *Int. Conf. on Machine Learning (ICML)*, pp. 2256–2265, 2015.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[26] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 32, 2019.

[27] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[28] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[29] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[30] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," *ISCA Interspeech*, pp. 803–807, 2022.

[31] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[32] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[33] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," *Int. Workshop on Acoustic Echo and Noise Control*, 2010.

[34] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," *Int. Conf. on Telecom. and Signal Proc. (TSP)*, pp. 72–76, 2021.

[35] C. H. You, S. N. Koh, and S. Rahardja, "/spl beta/-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 13, no. 4, pp. 475–486, 2005.

[36] C. Breithaupt and R. Martin, "Analysis of the decision-directed snr estimator for speech enhancement with respect to low-snr and transient conditions," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 19, no. 2, pp. 277–289, 2010.

[37] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[38] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," *arXiv preprint arXiv:2111.13606*, 2021.

[39] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 34, 2021.

[40] C.-W. Huang, J. H. Lim, and A. C. Courville, "A variational perspective on diffusion-based generative models and score matching," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 34, 2021.

[41] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

[42] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 35, 2022.

[43] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *Journal of Computational and Applied Mathematics*, vol. 6, pp. 19–26, 1980.

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Int. Conf. on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

[45] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *Int. Conf. on Learning Representations (ICLR)*, 2018.

[46] Y. Wu and K. He, "Group normalization," *Proc. of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[47] R. Zhang, "Making convolutional networks shift-invariant again," *Int. Conf. on Machine Learning (ICML)*, pp. 7324–7334, 2019.

[48] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: a self-gated activation function," *arXiv preprint arXiv:1710.05941*, 2017.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 30, 2017.

[50] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020.

[51] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete." [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A

[52] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, 2015.

[53] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *ISCA Speech Synthesis Workshop (SSW)*, pp. 146–152, 2016.

[54] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[55] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2018.

[56] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[57] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 2, pp. 749–752, 2001.

[58] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 24, no. 11, pp. 2009–2022, 2016.

[59] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 626–630, 2019.

[60] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 6493–6497, 2021.

[61] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," *ISCA Interspeech*, pp. 2862–2866, 2020.

[62] ITU-T Rec. P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," *Int. Telecom. Union (ITU)*, 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.808-202106-I/en

[63] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2022.

[64] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *Int. Telecom. Union (ITU)*, 2003. [Online]. Available: https://www.itu.int/rec/T-REC-P.835-200311-I/en

[65] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," *arXiv preprint arXiv:2203.13086*, 2022.

[66] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020.

[67] ITU-R Rec. BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," *Int. Telecom. Union (ITU)*, 2014. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534

[68] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[69] W. Falcon et al., "Pytorch lightning," *GitHub: https://github.com/PyTorchLightning/pytorch-lightning*, vol. 3, 2019.

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. on Learning Representations (ICLR)*, 2015.

[71] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 12 438–12 448, 2020.

[72] E. Aksan and O. Hilliges, "Stcn: Stochastic temporal convolutional networks," *Int. Conf. on Learning Representations (ICLR)*, 2018.

[73] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *Foundations and Trends in Machine Learning*, vol. 15, no. 1-2, pp. 1–175, 2021.

[74] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware Speech Enhancement with Deep Complex U-Net," *arXiv preprint arXiv:1903.03107*, 2019.

[75] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," *arXiv preprint arXiv:2104.03538*, 2021.

[76] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.

[77] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[78] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 1598–1607, 2020.

[79] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, "DBT-Net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 30, pp. 2629–2644, 2022.

[80] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *ISCA Interspeech*, 2022, pp. 936–940.

[81] ITU-T Rec. P.862.3, "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," *Int. Telecom. Union (ITU)*, 2007. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.3/en

[82] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 7412–7416, 2022.

[83] S. Uhlich and Y. Mitsufuji, "Open-unmix for speech enhancement (UMX SE)," 2020. [Online]. Available: https://github.com/sigsep/open-unmix-pytorch

[84] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *JASA Express Letters*, vol. 2, no. 10, p. 104802, 2022.

[85] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The Interspeech 2020 Deep Noise Suppression Challenge: Datasets, subjective testing framework, and challenge results," *ISCA Interspeech*, pp. 2492–2496, 2020.

[86] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 357–362, 2005.
[87] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," *Int. Conf. on Learning Representations (ICLR)*, 2021.



Julius Richter (S'22) received a B.Sc. and M.Sc in Electrical Engineering in 2017 and 2019 from the Technical University of Berlin, Germany. He is currently with the Signal Processing group at Universität Hamburg under the supervision of Prof. Dr.-Ing. Timo Gerkmann. His research interests include deep generative models and multi-modal learning with applications to audio-visual speech processing.



Simon Welker received a B.Sc. in Computing in Science (2019) and M.Sc. in Bioinformatics (2021) from Universität Hamburg, Germany. He is currently a second-year PhD student in the labs of Prof. Timo Gerkmann (Signal Processing, Universität Hamburg) and Prof. Henry N. Chapman (Center for Free-Electron Laser Science, DESY, Hamburg), researching machine learning techniques for solving inverse problems that arise in speech processing and X-ray imaging.



Jean-Marie Lemercier (S'21) received a M.Eng in Electrical Engineering in 2019 from Ecole Polytechnique, Paris, France. In 2020, he received a M.Sc. in Communications and Signal Processing from Imperial College London, London, United Kingdom. He is currently with the Signal Processing group at the Universität Hamburg under the supervision of Prof. Dr.-Ing. Timo Gerkmann. His research interests span machine learning-based speech enhancement and dereverberation for hearing devices applications. Recent works also include the design and analysis of diffusion-based generative models for various speech restoration tasks.



Bunlong Lay obtained a B.Sc. and M.Sc. in Mathematics in 2015 and 2017 from the University of Bonn, Germany. He subsequently joined the research institute Fraunhofer FKIE in Wachtberg Germany from 2018 until 2021, where he focused on research in the field of electronic warfare. In 2021 he started his Ph.D. at the University of Hamburg. Currently researching diffusion-based models for Speech Enhancement.



Timo Gerkmann (S'08–M'10–SM'15) is a professor for Signal Processing at the Universität Hamburg, Germany. He has previously held positions at Technicolor Research & Innovation in Germany, the University of Oldenburg in Germany, KTH Royal Institute of Technology in Sweden, Ruhr-Universität Bochum in Germany, and Siemens Corporate Research in Princeton, NJ, USA. His main research interests are on statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and a Senior Area Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing. He received the VDE ITG award 2022.

4.1.2 Publication: Diffusion Models for Audio Restoration: A review

Abstract

With the development of audio playback devices and fast data transmission, the demand for high sound quality is rising for both entertainment and communications. In this quest for better sound quality, challenges emerge from distortions and interferences originating at the recording side or caused by an imperfect transmission pipeline. To address this problem, audio restoration methods aim to recover clean sound signals from the corrupted input data. We present here audio restoration algorithms based on diffusion models, with a focus on speech enhancement and music restoration tasks. Traditional approaches, often grounded in handcrafted rules and statistical heuristics, have shaped our understanding of audio signals. In the past decades, there has been a notable shift toward data-driven methods that exploit the modeling capabilities of deep neural networks (DNNs). Deep generative models, and among them diffusion models, have emerged as powerful techniques for learning complex data distributions. However, relying solely on DNN-based learning approaches carries the risk of reducing interpretability, particularly when employing end-to-end models. Nonetheless, data-driven approaches allow more flexibility in comparison to statistical model-based frameworks, whose performance depends on distributional and statistical assumptions that can be difficult to guarantee. Here, we aim to show that diffusion models can combine the best of both worlds and offer the opportunity to design audio restoration algorithms with a good degree of interpretability and a remarkable performance in terms of sound quality. In this article, we review the use of diffusion models for audio restoration. We explain the diffusion formalism and its application to the conditional generation of clean audio signals. We believe that diffusion models open an exciting field of research with the potential to spawn new audio restoration algorithms that are natural-sounding and remain robust in difficult acoustic situations.

Reference

Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann, “Diffusion Models for Audio Restoration: A review ”, *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2025. DOI: 10.1109/MSP.2024.3445871

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2025 IEEE. Reprinted, with permission, from the reference displayed above.

Author’s Contributions

Jean-Marie Lemercier and Julius Richter are the lead authors of this paper. Jean-Marie Lemercier wrote the final section, “Practical Requirements for Diffusion-based Sampling for Audio Tasks,” as well as the paragraph “Model-based Processing with Diffusion Models” in the “Basics of Diffusion Models” section. Julius Richter wrote the abstract, the introduction, and the “Basics of Diffusion Models” section. Simon Welker authored the section “Conditional Generation with Diffusion Models,” while Eloi Moliner wrote the section “Diffusion Models for

Inverse Problems." Timo Gerkmann and Vesa Välimäki provided insights into the organization and structure of the manuscript. All authors contributed to reviewing and revising the manuscript.

Diffusion Models for Audio Restoration

Invited paper for the SPM Special entitled “Model-based and Data-Driven Audio Signal Processing”.

Jean-Marie Lemercier†, Julius Richter†, Simon Welker†, Eloi Moliner°, Vesa Välimäki°, Timo Gerkmann†

†Signal Processing (SP), Department of Informatics, Universität Hamburg, Germany

°Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland

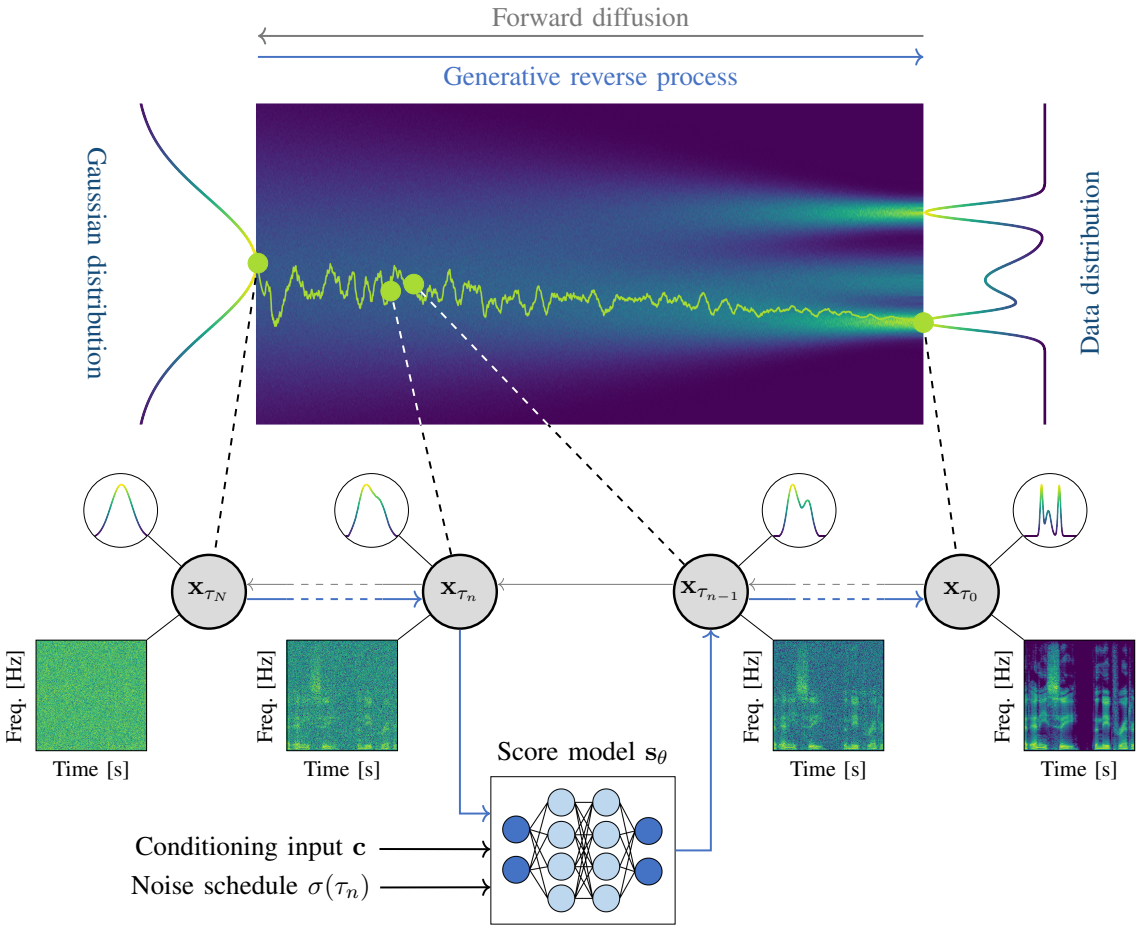


Fig. 1: A continuous-time diffusion model [8] transforms (left) a Gaussian distribution to (right) an intractable data distribution through a stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$ with marginal distributions $\{p_\tau(\mathbf{x}_\tau)\}_{\tau\in[0,T]}$. During training, the forward diffusion is simulated by adding Gaussian noise and rescaling the data, and a score model $\mathbf{s}_\theta$ with parameters θ learns the score function $\nabla_{\mathbf{x}_\tau}\log p_\tau(\mathbf{x}_\tau)$. During the generative reverse process, the process time τ is discretized to steps $\{\tau_0,\dots,\tau_N\}$ and followed in reverse from $\tau_N = T$ to $\tau_0 = 0$. (Bottom) The next state $\mathbf{x}_{\tau_{n-1}}$ is obtained based on the previous state $\mathbf{x}_{\tau_n}$ using an estimate given by the score model. The score model is conditioned by the noise scale at the current time step, $\sigma(\tau_n)$, and optional conditioning $\mathbf{c}$ to guide the generation such as, e.g., a text description.

With the development of audio playback devices and fast data transmission, the demand for high sound quality is rising for both entertainment and communications. In this quest for better sound quality, challenges emerge from distortions and interferences originating at the recording side or caused by an imperfect transmission pipeline. To address this problem, audio restoration methods aim to recover clean sound signals from the corrupted input data. We present here audio restoration algorithms based on diffusion models, with a focus on speech enhancement and music restoration tasks.

Traditional approaches, often grounded in handcrafted rules and statistical heuristics, have shaped our understanding of audio signals. In the past decades, there has been a notable shift towards data-driven methods that exploit the modeling capabilities of deep neural networks (DNNs). Deep generative models, and among them diffusion models, have emerged as powerful techniques for learning complex data distributions. However, relying solely on DNN-based learning approaches carries the risk of reducing interpretability, particularly when employing end-to-end models. Nonetheless, data-driven approaches allow more flexibility in comparison to statistical model-based frameworks, whose performance depends on distributional and statistical assumptions that can be difficult to guarantee. Here, we aim to show that diffusion models can combine the best of both worlds and offer the opportunity to design audio restoration algorithms with a good degree of interpretability and a remarkable performance in terms of sound quality.

In this article, we review the use of diffusion models for audio restoration. We explain the diffusion formalism and its application to the conditional generation of clean audio signals. We believe that diffusion models open an exciting field of research with the potential to spawn new audio restoration algorithms that are natural-sounding and remain robust in difficult acoustic situations.

INTRODUCTION

Traditional audio restoration methods exploit statistical properties of audio signals, such as auto-regressive modeling for click removal [1] or probabilistic modeling for speech enhancement and separation [2], by using various representations like time-domain waveforms, spectrograms, or cepstra. Although they are robust to many scenarios, such methods struggle with highly non-stationary sources or interferences that appear in real-life scenarios. In the past decade, audio signal processing algorithms have benefited greatly from the introduction of data-driven approaches based on DNNs [3]. Among these methods, a broad class leverages *predictive models* that learn to map a given input to a desired output. Note that the term *predictive models* covers both classification and regression tasks, unlike *discriminative models* [4]. In a typical supervised setting, a predictive model is trained on a labeled dataset to minimize a certain point-wise loss function between the processed input and the clean target. Following the principle of empirical risk minimization, the goal of predictive modeling is to find a model with minimal average error over the training data, where the generalization ability of the model is usually assessed on a validation set of unseen data. By employing ever-larger models and datasets—a current trend in deep learning—strong generalization

can be achieved. However, many purely data-driven approaches are considered black boxes and remain largely unexplainable and non-interpretable. Moreover, these models typically produce deterministic outputs, disregarding the inherent uncertainty in their results.

Generative models follow a different learning paradigm, namely estimating and sampling from an unknown data distribution. This can be used to infer a measure of uncertainty for their predictions and to allow the generation of multiple valid estimates instead of a single best estimate as in predictive approaches [4]. Furthermore, incorporating prior knowledge into generative models can guide the learning process and enforce desired properties about the learned distribution. Generative adversarial networks (GANs) [5] and variational autoencoders [6] have been instrumental in the early developments of generative models. Subsequent to these approaches, *diffusion models* [7], [8] have emerged as a distinct class of deep generative models that boast an impressive ability to learn complex data distributions such as that of natural images [7], [8], music [9], and speech [10]. Diffusion models generate data samples through iterative transformations, transitioning from a tractable prior distribution (e.g., Gaussian) to a target data distribution, as visualized in Figure 1. This iterative generation scheme is formalized as a stochastic process and is parameterized with a DNN that is trained to address a Gaussian denoising task.

From a practical point of view, diffusion models have become popular because they can generate high-quality samples while being simpler to train than GANs. Moreover, combining data-driven machine learning techniques with mathematical concepts, such as stochastic processes, opens up possibilities for modeling conditional data distributions and integrating Bayesian inference tools. In audio processing, this has spawned new types of algorithms that adopt diffusion models for restoration tasks such as speech enhancement [11], [12] or music restoration [9]. Here, we present a comprehensive overview and categorization of novel techniques for solving audio restoration problems using diffusion models in a data-driven, model-based fashion.

In the following, we first look at the basics of diffusion models and show how they can be used for model-based processing. We then examine conditional generation with diffusion models for audio restoration tasks, distinguishing between three different conditioning techniques. In particular, we look at diffusion models for audio inverse problems with a known degradation operator and its extension to blind inverse problems when the degradation operator is unknown. We conclude by discussing the practical requirements of diffusion models for audio restoration tasks, examining sampling speed and robustness to adverse conditions.

Basics of Diffusion Models

With the development of DNNs and the increase in computational power, deep generative modeling has become one of the leading directions in machine learning with a variety of applications. Deep generative models aim to design a generation process for data that resembles real-world examples, e.g., natural speech produced by a human speaker. This involves modeling the probability distribution of highly structured and complex data such that learning

and sampling are computationally tractable. One way to realize generative modeling is based on the assumption that the data is generated by some random process involving unobserved variables. Such *hidden variable models* map samples from a tractable distribution, such as the Gaussian distribution, to samples that are likely to represent target data points. From this perspective of hidden variable models, we discuss diffusion models as a distinct class of deep generative models whose hidden variables are parameterized via a stochastic process.

Diffusion models break down the problem of generating high-dimensional complex data into a series of easier *denoising* tasks. Training such a denoising model first requires defining a *forward diffusion process*, which gradually adds noise to the data points of a dataset. This corruption process progressively turns the data distribution into a Gaussian distribution, as shown in Figure 1 from right to left (gray arrows). In turn, data generation is accomplished by reversing the corruption process. First, a random sample is drawn from a Gaussian distribution, and then the model iteratively removes noise from this initial point, ultimately yielding a sample from the data distribution. This *reverse diffusion process* is illustrated in Figure 1 from left to right (blue arrows).

Formally, the forward diffusion process can be represented by the Markov chain $\mathbf{x}_0 \to \mathbf{x}_1 \to \cdots \to \mathbf{x}_N$ with $\mathbf{x}_0 \in \mathbb{R}^d$ sampled from the data distribution and fixed Gaussian transition probabilities $q(\mathbf{x}_n|\mathbf{x}_{n-1})$. The resulting directed graphical model is depicted with gray circles in Figure 1. The generative model is then described by a Markov chain in reverse order $\mathbf{x}_N \to \mathbf{x}_{N-1} \to \cdots \to \mathbf{x}_0$ with $\mathbf{x}_N$ sampled from the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. To accomplish the generation task, a DNN is trained to denoise the sample $\mathbf{x}_n$. Specifically, it learns to approximate the transition probabilities of the reverse Markov chain $p(\mathbf{x}_{n-1}|\mathbf{x}_n)$ [7].

This discrete-time Markov chain formulation of diffusion models can be generalized to continuous-time stochastic processes by letting the number of steps N grow infinitely, conversely making the distance between steps infinitely small. This facilitates the design of novel diffusion processes and allows the use of more flexible sampling schemes [8]. Specifically, the corresponding forward diffusion process is defined as a stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$, i.e., a collection of random variables indexed by a continuous process time $\tau \in [0, T]$ [13]. The process time τ in stochastic processes intuitively corresponds to the index n in Markov chains. It is important to note that the process time τ is completely unrelated to the time dimension of the audio signal. A single random realization of the stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$ is depicted by the green trajectory traversing Figure 1. The conditional distribution characterizing the forward diffusion model is the *transition kernel* $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ instinctively related to the probability $q(\mathbf{x}_n|\mathbf{x}_0) := \prod_{i=1}^{n} q(\mathbf{x}_i|\mathbf{x}_{i-1})$ in Markov chains. The transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ can be computed by solving a stochastic differential equation (SDE), which is a differential equation where some of the coefficients are random [13]. Specifically, we define the so-called *forward SDE* as

$$\mathrm{d}\mathbf{x}_\tau = \mathbf{f}(\mathbf{x}_\tau, \tau)\,\mathrm{d}\tau + g(\tau)\,\mathrm{d}\mathbf{w}_\tau\,, \tag{1}$$

where the function $\mathbf{f} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is referred to as the *drift coefficient* and relates to the deterministic part of the SDE. The function $g : \mathbb{R} \to \mathbb{R}$ is called the *diffusion coefficient* and controls the amount of randomness in the SDE. More precisely, the diffusion coefficient $g(\tau)$ scales the noise injected by the stochastic process $\mathbf{w}_\tau$. In most cases, $\mathbf{w}_\tau$ is chosen to be a Wiener process, which is a stochastic process with independent and normally distributed increments, i.e., $\mathbf{w}_{\tau+\mathrm{d}\tau} - \mathbf{w}_\tau \sim \mathcal{N}(\mathbf{0}, \mathrm{d}\tau\, \mathbf{I})$ [13]. If the drift coefficient $\mathbf{f}$ is an affine function of $\mathbf{x}_\tau$ and the diffusion coefficient g is independent of $\mathbf{x}_\tau$, then the transition kernel has a simple Gaussian form

$$q_\tau(\mathbf{x}_\tau|\mathbf{x}_0) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}_0, \tau), \sigma(\tau)^2\mathbf{I})\,, \tag{2}$$

where the mean $\boldsymbol{\mu}(\mathbf{x}_0, \tau)$ and standard deviation $\sigma(\tau)$ are obtained by analytically solving the SDE and computing the first and second moments of the solution [13].

The reverse diffusion process, i.e. the generation process, is also a stochastic process $\{\mathbf{x}_\tau\}_{\tau\in[0,T]}$ parameterized by the process time $\tau \in [0, T]$ flowing in the reverse direction, with $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma(T)^2\boldsymbol{I})$. Reversing the process time axis in (1) results in another SDE called the *reverse SDE* whose marginal distributions match those of the forward SDE [8]. Therefore, denoising the sample $\mathbf{x}_\tau$ boils down to solving the reverse SDE

$$\mathrm{d}\mathbf{x}_\tau = \left[\mathbf{f}(\mathbf{x}_\tau, \tau) - g(\tau)^2\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)\right] \mathrm{d}\tau + g(\tau)\, \mathrm{d}\bar{\mathbf{w}}_\tau\,, \tag{3}$$

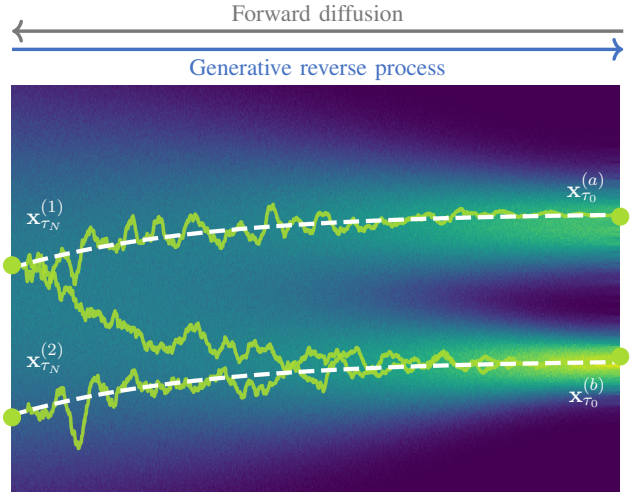
where $\mathrm{d}\tau < 0$ as the process axis is traveled in the reverse direction. The stochastic process $\bar{\mathbf{w}}_\tau$ is another Wiener process associated to this reverse process axis, i.e., $\bar{\mathbf{w}}_{\tau+\mathrm{d}\tau} - \bar{\mathbf{w}}_\tau \sim \mathcal{N}(\mathbf{0}, -\mathrm{d}\tau\, \mathbf{I})$. The quantity $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ (with $\nabla_{\mathbf{x}_\tau}$ representing the gradient operator with respect to $\mathbf{x}_\tau$) is called *score function* and is a vector field informative about the variations of the process state's logarithmic probability density. The score function $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is generally intractable and we need to approximate it with a DNN called *score model* $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ with parameters θ. Vincent et al. [14] have shown that the score model $\mathbf{s}_\theta(\mathbf{x}_\tau, \tau)$ can be optimized using *denoising score matching*, i.e., matching the score of the Gaussian transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ instead of the score of the unknown probability $p_\tau(\mathbf{x}_\tau)$. The score of the transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ can be obtained from (2) as

$$\nabla_{\mathbf{x}_\tau} \log q_\tau(\mathbf{x}_\tau|\mathbf{x}_0) = -\frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0, \tau)}{\sigma(\tau)^2}\,. \tag{4}$$

The score model $\mathbf{s}_\theta$ is therefore trained using the denoising score-matching objective [14]

$$\mathbb{E}_{\substack{\mathbf{x}_0\sim p(\mathbf{x}_0)\\ \tau\sim\mathcal{U}(0,T)\\ \mathbf{x}_\tau\sim q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)}} \left[\lambda(\tau) \left\| \mathbf{s}_\theta(\mathbf{x}_\tau, \tau) + \frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0, \tau)}{\sigma(\tau)^2} \right\|_2^2\right], \tag{5}$$

where a data point $\mathbf{x}_0$ is first sampled from the training set. Then, a process time τ is sampled uniformly between 0 and T, and the diffusion state $\mathbf{x}_\tau$ is obtained by sampling from the transition kernel (2). Here $\lambda(\tau)$ is a time-dependent



1: Sample initial state $\mathbf{x}_{\tau_N} \sim \mathcal{N}(\mathbf{0}, \sigma(\tau_N)^2\mathbf{I})$
2: **for** $n \in \{N, \dots, 1\}$ **do**
3: Get reverse step size $\Delta\tau_n = \tau_{n-1} - \tau_n < 0$
4: Obtain score estimate $\mathbf{s}_\theta(\mathbf{x}_{\tau_n}, \tau_n)$
5: **if** *Deterministic sampler* **then**
6: Take probability-flow ODE step
7: **else if** *Stochastic sampler* **then**
8: Take reverse SDE step
9: Obtain next state $\mathbf{x}_{\tau_{n-1}} = \mathbf{x}_{\tau_n} + \mathrm{d}\mathbf{x}_{\tau_n}$
10: Output: $\mathbf{x}_0$

Fig. 2: Stochastic (green solid lines) and deterministic (white dashed lines) sampling trajectories. The stochastic sampler discretizes the reverse SDE (3) where noise is added at each sampling step by the Wiener process $\bar{\mathbf{w}}_\tau$. The deterministic sampler uses the probability-flow ordinary differential equation (ODE), which does not re-introduce noise. Two different initial points $\mathbf{x}_{\tau_N}^{(1)}$ and $\mathbf{x}_{\tau_N}^{(2)}$ are sampled, and two realizations of the stochastic sampler are shown for the same initial state $\mathbf{x}_{\tau_N}^{(1)}$. The target distribution has two modes $\mathbf{x}_{\tau_0}^{(a)}$ and $\mathbf{x}_{\tau_0}^{(b)}$.

scaling factor, chosen empirically to stabilize training [8].

Once the score model $\mathbf{s}_\theta$ has been trained, it allows the generation of new samples from the learned data distribution by solving the reverse SDE (3). In practice, this is done by first discretizing the process time axis into N steps $\{\tau_N, \tau_{N-1}, \dots, \tau_0\}$ with a step size $\Delta\tau_n := \tau_{n-1} - \tau_n < 0$, often chosen uniformly. Then an initial condition $\mathbf{x}_{\tau_N}$ is sampled and the reverse SDE (3) is integrated between $\tau_N = T$ and $\tau_0 = 0$ using a numerical approximation method called *SDE solver* [15]. A differential equation solver approximates the trajectory between successive steps $\mathbf{x}_{\tau_n}$ and $\mathbf{x}_{\tau_{n-1}}$ as a piecewise polynomial function (linear if first-order solver, quadratic if second-order, etc.), whose coefficients depend on the terms in (3). An SDE solver, in particular, considers two polynomial functions (with potentially distinct degrees) to model the deterministic and stochastic terms, respectively. For instance, the widespread Euler-Maruyama method is an SDE solver with first-order polynomial approximation for both the deterministic and stochastic components [15]. The generation process is summarized in the algorithm in Figure 2.

Deterministic sampling can also be used in place of stochastic sampling by deactivating the randomness source, i.e., removing the Wiener process $\bar{\mathbf{w}}_\tau$ in (3) and scaling the diffusion coefficient g. This turns the reverse SDE into a so-called *probability flow ordinary differential equation (ODE)* [8]. A comparison between stochastic and deterministic sampling is presented in Figure 2. We display three realizations of the stochastic and deterministic sampler. Note that because of the stochastic noise injected at each step, two stochastic sampler trajectories starting at the same initial state may end up reaching two different modes of the target data distribution, whereas the

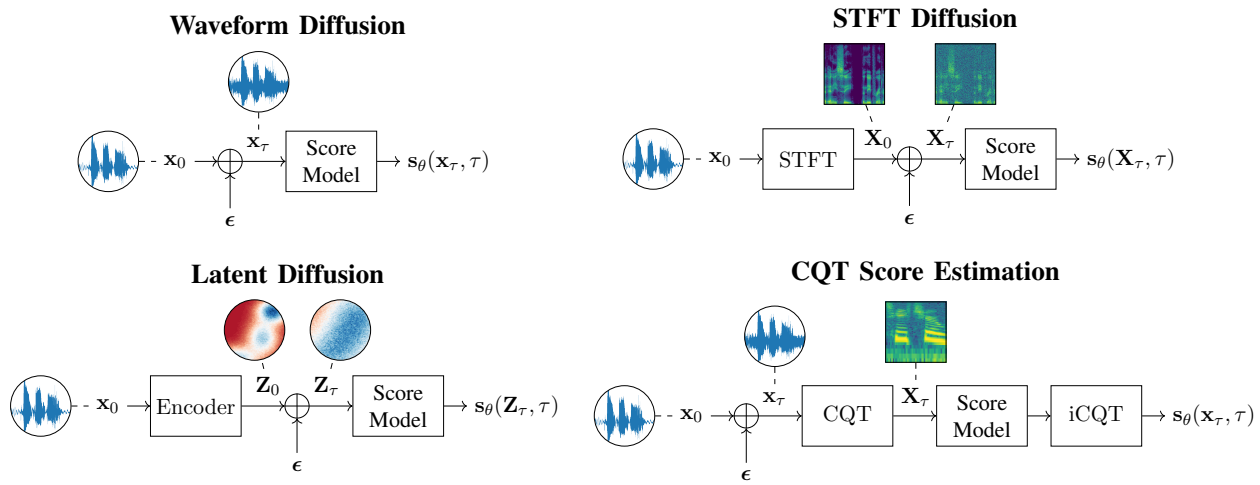


Fig. 3: A diffusion model may be trained in (top-left) waveform [16], (top-right) STFT [17], (bottom-left) latent [21], or (bottom-right) CQT [9] domains. Sampling from the transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ can be realized by rescaling the clean data sample $\mathbf{x}_0$ and adding Gaussian noise $\boldsymbol{\epsilon}$ with standard deviation $\sigma(\tau)$ (see (2)). In the top-left, top-right, and bottom-left figures, the noise and the score functions are in the same domain. In the bottom-right figure, the diffusion process is formulated in the time domain but the score model pipeline includes a CQT and its inverse.

corresponding mean trajectory systematically reaches the same mode. This suggests that a stochastic sampler could be used to obtain more diverse samples and also to improve mode coverage, i.e., better represent the modes of the data distribution $p(\mathbf{x}_0)$ regardless of the initial state.

Diffusion processes can be defined for various data representations, depending on the audio application considered. Early works such as [10], [11], [16] directly use the waveform representation, whereas some speech enhancement approaches employ the complex short-time Fourier transform (STFT) domain [12], [17], [18], and several music restoration works consider the Constant-Q Transform (CQT) domain, which is a natural space for harmonic music signals [9], [19], [20]. Learned domains like, e.g., autoencoder latent spaces, can also be exploited for diffusion to reduce the dimensionality of the original audio data or leverage auto-encoding properties, which gives birth to *latent diffusion models* [21]. Figure 3 offers a schematic overview of diffusion models defined in various domains.

It should be noted that a connection between score-based diffusion models and continuous normalizing flows has been drawn by Lipman et al., creating a new category of so-called *flow matching* models [22]. Flow matching methods generalize Gaussian diffusion models and allow to design more flexible probability paths (based on, e.g., optimal transport) between arbitrary terminal distributions. This approach has been used to train a foundational speech model, which can be finetuned to perform restoration tasks such as speech enhancement and separation [23].

Model-based processing with diffusion models

In statistical model-based speech enhancement, each time-frequency bin of the speech and noise spectrograms is often assumed to be mutually independent and to follow a zero-mean complex Gaussian prior distribution [2]. For

an additive mixture model, this yields a Gaussian likelihood model for the mixture and a Gaussian posterior model for the clean speech estimate using Bayes' rule. Under this Gaussian assumption, the posterior mean can be derived as the celebrated Wiener filter solution, providing the optimal speech estimate in the minimum mean square error (MMSE) sense. However, distributional and independence assumptions are merely approximations utilized out of convenience for the derivation of closed-form estimators, e.g., the mentioned Wiener filter. With diffusion models, there are no distributional and independence assumptions on the speech and noise signals themselves. Indeed, the very intent of deep generative modeling is to allow more flexibility by inferring the signal structure from data rather than the parameters of a fixed distribution.

In contrast to other deep learning approaches to audio restoration, two aspects make diffusion models well suited for the introduction of domain knowledge, showcasing them as model-based approaches. The first property is derived from the physical inspiration of diffusion models and their connection to Gaussian denoising [24], which makes them easier to interpret in comparison to other deep generative models such as GANs. In particular, the Gaussian parameterization of the transition kernel $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ enables the injection of knowledge in the form of specific schedules for the mean $\boldsymbol{\mu}$ and standard deviation σ [11], [17], [25]. Furthermore, domain knowledge can be leveraged to posit a distributional hypothesis for the noise process $\mathbf{w}_\tau$ used during forward and reverse diffusion. For instance, Nachmani et al. [26] propose a Gamma distribution instead of the usual Wiener process $\mathbf{w}_\tau$ with Gaussian increments, as it better fits the estimation error distribution. The authors consequently show improvements in speech generation quality compared to the Gaussian case.

The second powerful property of diffusion models is their natural integration within stochastic optimization and posterior sampling using Bayes' theorem, making them particularly suited for conditional generation. We consider the case of audio restoration under the scope of inverse problem solving, i.e., retrieval of clean audio $\mathbf{x}_0$ from a measurement $\mathbf{y}$. There, an approximation of the measurement likelihood $p(\mathbf{y}|\mathbf{x}_0)$ can be obtained via a closed-form model of the operation corrupting $\mathbf{x}_0$ into $\mathbf{y}$. Combining this likelihood model and the learned deep generative prior with Bayes' rule can provide sampling or stochastic optimization algorithms for the conditional generation of samples from the posterior distribution $p(\mathbf{x}_0|\mathbf{y})$ [9], [20], [27], [28].

In summary, first, we see that the data-driven nature of diffusion models allows a higher degree of versatility than traditional signal processing methods, which are often strictly based on simple closed-form distributions and independence assumptions. Secondly, it is important to note that diffusion models transcend the stereotype of being non-interpretable black-boxes. Instead, they benefit from strong integration within stochastics and enable significant potential for the injection of domain knowledge for model-based audio processing.

CONDITIONAL GENERATION WITH DIFFUSION MODELS

One of the most fundamental uses of diffusion models is to perform unsupervised learning from a finite collection of samples to learn an underlying complex data distribution. This provides the ability of *unconditional generation*, i.e., to generate new samples from the learned data distribution. To solve audio restoration tasks, a diffusion model must be adapted to generate audio that not only conforms to the learned clean audio distribution but, importantly, is also a plausible reconstruction of a given corrupted signal. This effectively requires the diffusion model to perform *conditional generation*. We distinguish between three families of approaches for diffusion-based generative audio restoration: *(i) input conditioning*, where the score model is provided with a task-specific conditioning signal as input, *(ii) task-adapted diffusion*, where the forward and reverse diffusion processes are modified to interpolate between clean and corrupted signals, and *(iii) external conditioning*, where the score model is trained purely on clean audio data and is later combined with an external conditioner during inference. Approaches that use input conditioning (i) or external conditioning (iii) often initialize the iterative generation process with pure Gaussian noise, and then generate a clean signal by iteratively filtering this noise while being guided by the conditioning signal. In contrast, in task-adapted diffusion (ii), the corrupted audio itself is used for initialization and iteratively filtered, making this approach conceptually closer to a denoising procedure.

(i) Input conditioning: Diffusion models that use input conditioning are provided with a task-specific conditioning signal $\mathbf{c}$ (usually some representation of the corrupted signal $\mathbf{y}$) as an additional input during training and inference. To this end, they employ DNNs as score models that are specifically designed to perform feature fusion between the inputs $\mathbf{x}_\tau$ and $\mathbf{c}$. It should be noted that, in most cases, input conditioning approaches require the use of paired data, as the conditioning signal $\mathbf{c}$ and the target data sample $\mathbf{x}_0$ should be representations of the same data instance, or at least share some semantics. The earliest works to follow this approach include DiffWave [16], which uses mel-spectrograms as conditioning signals for neural vocoding and text-to-speech tasks. While DiffWave focuses on audio generation rather than restoration, the authors of DiffWave also provide preliminary evidence that an unconditional speech diffusion model can perform speech enhancement by using the corrupted audio $\mathbf{y}$ as a starting point of the sampling process even though the diffusion model was only trained to remove Gaussian noise. DiffuSE [29] builds upon DiffWave to solve speech enhancement tasks, using noisy spectral features as conditioning $\mathbf{c}$.

In the worst case, the score model may not use conditioning $\mathbf{c}$ at all, thus inadvertently performing unconditional rather than conditional generation. One possible solution to this is *classifier-free guidance*, where the conditioning signal is randomly set to zero with a fixed probability during training. This results in a single model that can both provide an estimate for the conditional score $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau|\mathbf{c})$ and the unconditional score $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$. At inference, the two estimates can then be weighted at will to trade quality (more weight on conditional score) for variety (more weight on unconditional score). This idea has been used, for instance, by Liu et al. [21] to perform

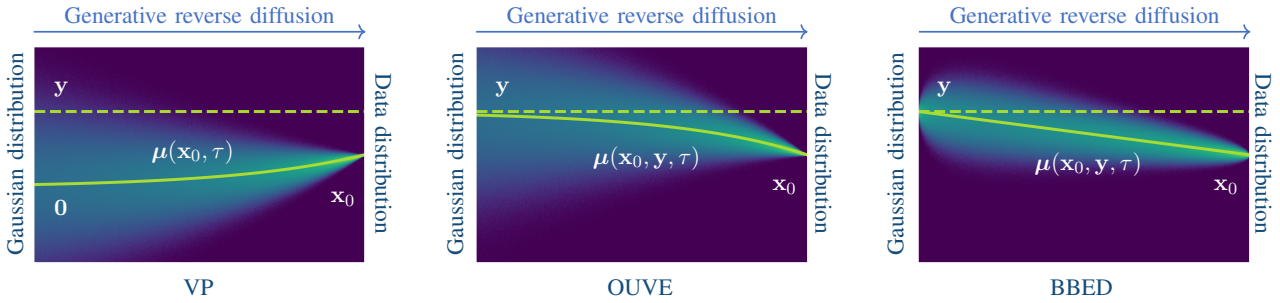


Fig. 4: Comparison of different diffusion processes. (left) Classical variance-preserving (VP) diffusion: mean exponentially interpolates between clean audio $\mathbf{x}_0$ and $\mathbf{0}$, irrespectively of the degraded audio $\mathbf{y}$ [8]. (middle) Task-adapted Ornstein–Uhlenbeck variance exploding (OUVE) diffusion: mean exponentially interpolates between clean audio $\mathbf{x}_0$ and degraded audio $\mathbf{y}$ [12], [17]. (right) Task-adapted brownian bridge with exponential diffusion (BBED): mean linearly interpolates between clean audio $\mathbf{x}_0$ and degraded audio $\mathbf{y}$ [25].

controllable full-band audio synthesis and can also be employed for various audio restoration tasks.

(ii) Task-adapted diffusion: In many restoration tasks such as denoising, dereverberation and separation, the corrupted signal $\mathbf{y}$ and the clean signal $\mathbf{x}_0$ have same dimensionality. This allows one to define what we denote as *task-adapted* diffusion processes, i.e., diffusion processes whose mean $\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, \tau)$ is $\mathbf{x}_0$ at $\tau = 0$ and $\mathbf{y}$ at $\tau = T$, and that interpolates between these terminal values for $\tau \in [0, T]$. This is a form of conditioning which is not introduced as an auxiliary variable to the score model $\mathbf{s}_\theta$ as for input conditioning, but rather directly injected in the parameters of the diffusion process itself. Examples of classical and task-adapted diffusion processes are visualized on Figure 4. CDiffuSE [11] is one of the earliest methods using task-adapted diffusion, formulating the process in discretized time steps. Score-based generative models for speech enhancement (SGMSE) [17] and SGMSE+ [12] extend this idea to the continuous SDE-based formalism of diffusion models to derive pairs of forward and backward processes. Subsequent works [18], [25] build upon this formalism to design alternative forward and backward processes which result in fewer sampling steps and/or higher reconstruction quality. In practice, these methods combine task-adapted diffusion processes with input conditioning, by also providing $\mathbf{y}$ as an auxiliary input to the score model.

The interpolation between $\mathbf{x}_0$ and $\mathbf{y}$ underlying these approaches assumes an additive signal model typical for denoising tasks, which can also be treated as natural for separation tasks [30] or for convolutive corruptions such as reverberation. The aforementioned methods also achieve excellent reconstruction quality for non-additive corruptions like in bandwidth extension [18] and STFT phase retrieval [31], which shows their ability to perform *blind* restoration, i.e., when the corruption operator is not perfectly known during inference.

(iii) External conditioning: External conditioning approaches combine an unconditional diffusion model with an external conditioner that provides a conditioning signal during inference. Since the diffusion model is unconditional, no knowledge of the restoration task is accessed during the training stage and no supervision nor paired data is

required. Instead, the task-specific information is injected only at inference by the external conditioner. Therefore, external condition methods can leverage diffusion-based foundation models pre-trained on large-scale data, and adapt them for inference without further re-training. One such type of external conditioner is a pre-trained classifier enabling the combined model to perform class-conditional data generation. For audio restoration, the external conditioner usually takes the form of a task-specific closed-form measurement model. This results in an overall model that combines a strong data-driven prior for clean audio (score model) with a model-based formulation of the specific restoration task (measurement model). This approach shows good results even when the observation $\mathbf{y}$ is affected by measurement noise [9], [28] and has the advantage of not requiring retraining of the diffusion model for new restoration tasks. These approaches can be applied to blind restoration tasks if a good parameterization of the measurement operator is found. The parameterization enables classical estimation algorithms to be utilized for joint inference of the measurement model and target audio sample estimation, as Moliner et al. [20] accomplished in the blind bandwidth extension of historical music recordings.

Diffusion Models for Inverse Problems

We have seen different strategies to condition diffusion models for audio restoration tasks. This section delves into the *external conditioning* approach, specifically focusing on the application of diffusion models for solving inverse problems in the audio domain. Several audio restoration tasks can be formulated as an inverse problem, wherein an observed audio signal $\mathbf{y}$ is the result of corrupting a clean signal $\mathbf{x}_0$ with a degradation model $\mathcal{A}(\cdot)$ and additive noise $\mathbf{n}$, which can be expressed as

$$\mathbf{y} = \mathcal{A}(\mathbf{x}_0) + \mathbf{n}. \tag{6}$$

This model covers an infinite set of possible degradations, depending on how the operator $\mathcal{A}(\cdot)$ is defined. Three cases of particular interest are showcased in Figure 5. Initially, we concentrate on scenarios in which both the degradation model $\mathcal{A}(\cdot)$ and the noise statistics $\mathbf{n}$ are known. The goal is to recover the original signal $\mathbf{x}_0$ from the corrupted observations $\mathbf{y}$. However, in many cases, the problem is ill-posed, lacking a unique solution and defying straightforward resolution.

Often, solving an inverse problem is approached with a maximum a posteriori (MAP) objective

$$\arg\max_{\mathbf{x}_0} p(\mathbf{x}_0|\mathbf{y}) \,, \tag{7}$$

where the posterior distribution factorizes into likelihood and prior $p(\mathbf{x}_0|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}_0)p(\mathbf{x}_0)$. Under a zero-mean Gaussian measurement noise assumption, denoted as $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2\mathbf{I})$, the MAP estimate takes the form

$$\arg\min_{\mathbf{x}_0} \frac{1}{\sigma_y^2}\|\mathbf{y} - \mathcal{A}(\mathbf{x}_0)\|_2^2 + \mathcal{R}(\mathbf{x}_0) \,, \tag{8}$$

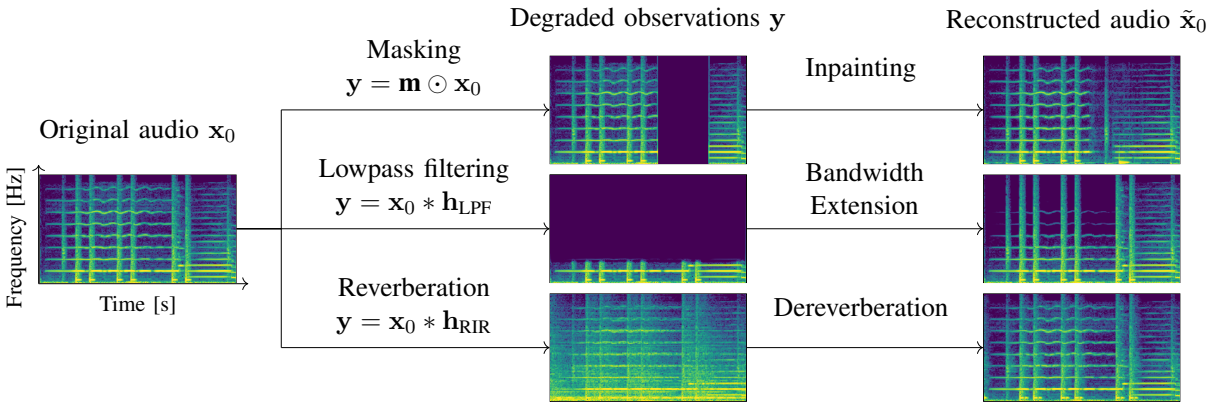


Fig. 5: Visual representation of several inverse problems in audio: (top to bottom) inpainting, bandwidth extension, and dereverberation. (Left) The spectrogram of the original audio signal, $\mathbf{x}_0$, undergoes various transformations via different measurement operators, (middle) the resulting degraded observations, $\mathbf{y}$, correspond to specific audio distortions, and (right) the reconstructed audio signal $\tilde{\mathbf{x}}_0$ is obtained by solving each inverse problem. Notably, the reconstructed example spectrograms (right) closely mirror the original (left), but minor differences appear because of the inherent ill-posed nature of these inverse problems.

where the first term is a reconstruction cost function, in this case an L^2-norm, designed to preserve fidelity with the observations $\mathbf{y}$. The second term, $\mathcal{R}(\mathbf{x}_0)$, functions as a regularizer, incorporating prior information or domain knowledge about the signal. Its purpose is to mitigate the under-determination of the problem by constraining the space of suitable solutions, thereby making the optimization feasible in practice. In audio processing, a frequently employed regularizer is the sparsity-promoting L^1-norm, which assumes that the true signal is sparse in a specified transform domain, such as time-frequency representations.

A diffusion model learns the statistical characteristics of the training data, in our case of clean audio signals. One can then expect diffusion models to have the potential to serve as strong data-driven priors for solving inverse problems. We will now elaborate on how to leverage these diffusion-based generative priors for solving (8).

To solve an inverse problem using a diffusion model, the score $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ in the reverse SDE (3) is replaced with the score of the posterior using Bayes' rule

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau)\,, \tag{9}$$

where the *prior score* $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is approximated with the unconditional score model $\mathbf{s}_\theta$ (see (5)). The term $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau)$ represents the *likelihood score*. However, it is important to note that the likelihood $p(\mathbf{y}|\mathbf{x}_\tau)$ is only analytically tractable for $\tau = 0$, as $\mathbf{x}_\tau$ refers to a noisy version of $\mathbf{x}_0$ and the true likelihood is defined through an intractable integral over all possible $\mathbf{x}_0$

$$p_\tau(\mathbf{y}|\mathbf{x}_\tau) = \int_{\mathbf{x}_0} p(\mathbf{y}|\mathbf{x}_0) p_\tau(\mathbf{x}_0|\mathbf{x}_\tau)\,\mathrm{d}\mathbf{x}_0\,. \tag{10}$$

Some works alleviate this issue by simply bypassing the likelihood term, and instead project the state $\mathbf{x}_\tau$ onto the set of the observations $\mathbf{y}$ at every step of the discretized inference process [8]. The objective of such *projection-based* method is to inject the reliable parts of the observations into the intermediate predictions. This ensures that at each step, the intermediate output of the algorithm is consistent with the algorithm input, i.e., the degraded audio, which is often refered to as *data consistency* and helps avoiding degenerate solutions. Projection-based methods offer the advantage of ensuring data consistency and simplicity in terms of algorithmic implementation. However, their applicability is limited to a reduced set of linear inverse problems, such as audio inpainting or bandwidth extension [9], [21], where closed-form expressions for the projection step are available.

Other works adopt more theoretically grounded approximations of the likelihood that allow a broader versatility by incorporating a model-based approach. In particular, Chung et al. [27] proposed Diffusion Posterior Sampling, and approximate the likelihood as $p_\tau(\mathbf{y}|\mathbf{x}_\tau) \approx p(\mathbf{y}|\hat{\mathbf{x}}_0(\mathbf{x}_\tau))$. There, $\hat{\mathbf{x}}_0(\mathbf{x}_\tau)$ is a coarse estimate of $\mathbf{x}_0$ obtained by denoising from state $\mathbf{x}_\tau$ in just one deterministic reverse diffusion step. When modeling the measurement noise $\mathbf{n}$ in (6) as a Gaussian $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2\mathbf{I})$, the resulting approximated likelihood is a Gaussian distribution $p\left(\mathbf{y}|\hat{\mathbf{x}}_0(\mathbf{x}_\tau)\right) = \mathcal{N}(\mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_\tau)), \sigma_y^2\mathbf{I})$. It follows that the likelihood score can be computed as:

$$\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{y}|\mathbf{x}_\tau) \approx -\frac{1}{\sigma_y^2} \nabla_{\mathbf{x}_\tau} \|\mathbf{y} - \mathcal{A}\left(\hat{\mathbf{x}}_0(\mathbf{x}_\tau)\right)\|_2^2 . \tag{11}$$

The L^2-norm in (11) can be replaced by any other objective function that better fits the statistics of the measurements [27]. For example in [32], the measurement noise $\mathbf{n}$ is modelled as a Gaussian in the compressed STFT domain, where the compression is a square-root power law on the STFT magnitude, leaving the phase unchanged. This helps accounting for the heavy-tailedness of speech distributions [33]. It is important to note that the gradient operator $\nabla_{\mathbf{x}_\tau}$ requires differentiating through the score model by backpropagation, which introduces a computational overhead. In practice, the unknown measurement noise variance σ_y^2 is estimated empirically using, e.g., the norm of the gradients in (11) [9]. Compared to projection-based methods, this approach is not limited to linear problems and can be applied to cases where $\mathcal{A}(\cdot)$ is nonlinear, as long as the operator $\mathcal{A}(\cdot)$ is differentiable. A geometrical perspective on the sampling process is displayed in Figure 6. This diagram illustrates the intuition behind conditional sampling with a diffusion model, in this case in the context of bandwidth extension. This strategy has been successfully applied in audio bandwidth extension [9], audio inpainting [19], and dereverberation [28].

Blind inverse problems

Until this point, our analysis has proceeded under the assumption that the degradation operator $\mathcal{A}(\cdot)$ is known. However, in practical applications, the degradation operator is often unknown. This lack of knowledge about the degradation operator renders the calculation of the posterior $p(\mathbf{x}_0|\mathbf{y})$ a *blind* inverse problem, substantially raising

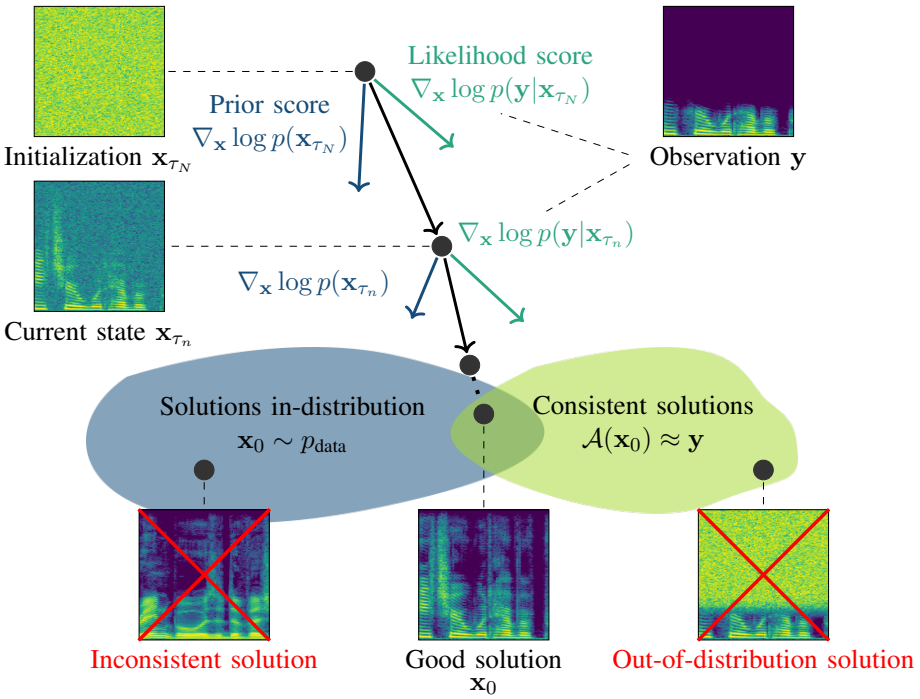


Fig. 6: Geometrical interpretation of posterior sampling with diffusion models (e.g., [27]). The prior score guides the trajectories towards solutions within the training data manifold, or in-distribution with the training data (gray space). Simultaneously, the role of the likelihood score is to steer the sampling trajectories toward a solution space consistent with the observed data (light green space). When properly weighted, the two components pull the sampling process to the intersection of these two manifolds. This intersection exists and contains the solutions to the inverse problem if the two score functions are properly estimated and if the solutions are contained in the manifold spanned by the training data, i.e., if the training dataset is properly adapted to the problem.

the difficulty of the task. The Diffusion Posterior Sampling approach [27], as previously explained, provides a valuable foundation that can be extended to tackle blind inverse problems. In scenarios where we possess at least some knowledge of the structure of the degradation operator, a viable strategy is to embrace a model-based approach. This involves designing a parametric model of the degradation operator, denoted as $\mathcal{A}_\phi(\cdot)$, and jointly optimizing its parameters ϕ alongside the restored audio signal throughout the sampling process.

An example of this approach is the Blind Audio Bandwidth Extension (BABE) [20], which addresses the problem of blind reconstruction of missing high frequencies in music from bandlimited observations without knowledge of the lowpass degradation, such as the cutoff frequency. This challenge is typical in restoring historical audio recordings. In BABE, the measurement model $\mathcal{A}_\phi(\cdot)$ is parameterized by a piecewise approximation of a low-pass filter in the frequency domain, where the parameters ϕ represent the cutoff frequencies and decay slopes of this filter [20]. The optimization process, as illustrated in Figure 7, alternates between sampling updates of the audio signal $\mathbf{x}_\tau$ and refining ϕ through stochastic gradient descent, using a maximum likelihood objective as the guiding principle. BUDDy [32] takes a similar approach and solves joint speech dereverberation and room acoustics estimation by

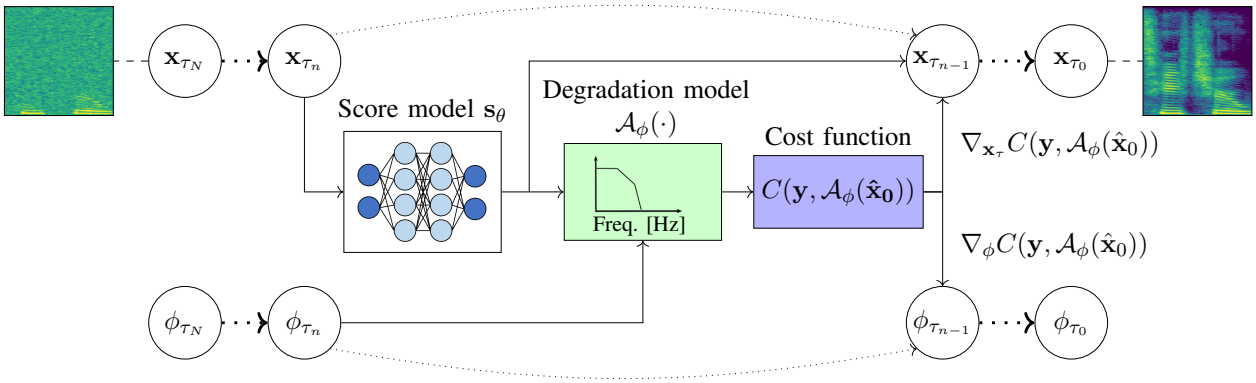


Fig. 7: BABE: posterior sampling algorithm for solving blind bandwidth extension using a prior score model $\mathbf{s}_\theta$ and a parameterized degradation operator $\mathcal{A}_\phi$ [20]. The optimization alternates between updating the reconstructed signal $\mathbf{x}$ (top) and the degradation parameters ϕ (bottom). For ease of reading, we write $\hat{\mathbf{x}}_0 := \hat{\mathbf{x}}_0(\mathbf{x}_{\tau_n})$.

combining Diffusion Posterior Sampling with a model-based subband filter approximating room impulse response. The resulting method largely outperforms other blind unsupervised dereverberation methods. Thanks to unsupervised learning, BUDDy seamlessly adapts to new acoustic scenarios, whereas supervised methods typically struggle when there is a mismatch between training and testing conditions.

Practical Requirements of Diffusion-based Sampling for Audio Tasks

While diffusion models provide powerful priors that can be employed for various audio restoration tasks, they require some improvements to be suitable for real-time acoustic communications. We divide these requirements into two categories: *(i) inference speed and causal processing*, which can be prohibitive for low-latency real-time applications, and *(ii) robustness to adverse conditions*, which must be assured for integration into reliable systems.

Inference speed and causal processing

One major drawback of diffusion models is their slow inference. As the score model is called at each step of the reverse process, the computational complexity is directly proportional to the number of steps used and the order of the solver, i.e., the number of score estimations used per time step. Using more diffusion steps naturally provides better sample reconstruction since the truncation error of the numerical solver is reduced when the step size is decreased. Similarly, increasing the solver order reduces the per-step truncation error. However, both these options lead to an increased computational cost. Furthermore, accumulating truncation errors over the diffusion trajectory can make the samples diverge from the distribution learned during training, and therefore make the score model produce unreliable estimates, which is referred to as the *drifting bias*. These two sources of error compound over the diffusion trajectory, therefore, without further optimization, high-quality reconstruction can only be obtained

at a high computational cost. This section presents several methods to reduce the computational complexity of diffusion-based methods in audio applications.

Reducing per-step inference time: A natural way to accelerate inference is to reduce the cost of each call to the score model. This can be obtained by minimizing the size of the neural network used for score inference through, e.g., knowledge distillation, or by reducing the size of the space itself where diffusion is performed, resulting in *latent diffusion models*. The latent space should be designed such that its reduced dimensionality has a limited impact on the reconstruction quality, and its structure allows for score estimation with a reasonably-sized neural network. Latent diffusion is popular in text-to-audio generation and has been recently applied to audio editing (including restoration) in AUDIT [34], which uses latents provided by a variational autoencoder.

Improving initialization: Another possibility to accelerate sampling is to find a better initial prediction to reduce the distance between the initial condition $\mathbf{x}_T$ and the target sample $\mathbf{x}_0$. This can be provided by a separate plug-in predictive network providing an estimate of the posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{y}]$ as proposed by Lemercier et al. in their Stochastic Regeneration Model (StoRM) [18] for speech enhancement (see Figure 8). The diffusion-based generative model can restore target cues potentially destroyed by the predictive stage while additionally removing residual corruption. The resulting approach requires significantly fewer function evaluations than the original diffusion-only model in [12], for a better-sounding result. Figure 9 shows the clean, degraded, and restored speech spectrograms produced with StoRM. As a simpler alternative, the corrupted utterance $\mathbf{y}$ can be directly used as the mean of the initial state $\mathbf{x}_T$. This latter strategy is sometimes referred to as *warm initialization* and has already been used in audio-related tasks such as speech enhancement [11] and bandwidth extension [20]. A good initial prediction can also be obtained by designing a more suitable diffusion trajectory to reduce the mismatch between training and inference, as suggested by Lay et al. [25] for speech enhancement. As shown in Figure 4, the BBED diffusion process proposed in [25] has a linear, constant speed mean interpolating between the clean and noisy speech, which effectively terminates at the clean speech in finite time, unlike the original OUVE diffusion process proposed in [17].

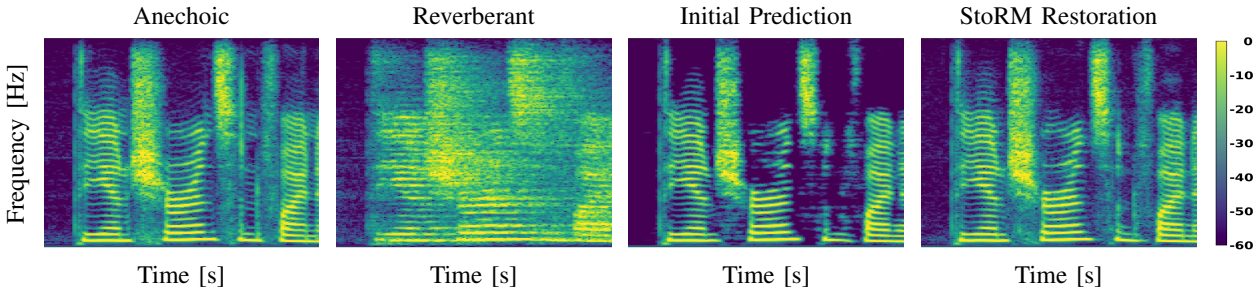


Fig. 9: Dereverberation results with StoRM [18]. Input T_{60} is 1.06 s. Three seconds of audio are shown, and the bandwidth is 8 kHz. Severe speech distortions are observed in the initial prediction because of the harsh reverberant conditions. StoRM corrects the distortions and restores the formant structure without residual reverberation.

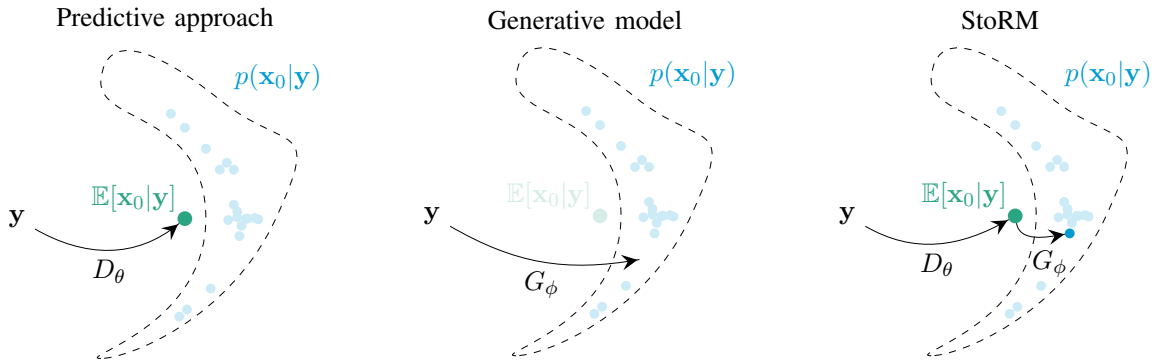


Fig. 8: Visualization of the inference process for the predictive, generative and StoRM [18] models for a complex posterior distribution. With the proposed two-stage inference, StoRM uses the predictive mapping to the posterior mean $\mathbb{E}[\mathbf{x}_0|\mathbf{y}]$ as an intermediate step for generation of a sample which is more likely to lie in high-density regions of the posterior $p(\mathbf{x}_0|\mathbf{y})$.

Reducing the number of steps: The remaining approaches investigate how to reduce the number of diffusion steps of the reverse process. As in most ODE/SDE integration problems, using off-the-shelf higher-order samplers can improve the per-step precision but here it comes at the cost of more calls to the neural network for each step, which leads to a non-trivial tradeoff between computational complexity and sample quality. In denoising diffusion implicit models (DDIM) [35] instead, the Markovian property of the transition kernel is deliberately removed by conditioning the next reverse diffusion estimate $\mathbf{x}_{\tau_{n-1}}$ on both the previous state $\mathbf{x}_{\tau_n}$ and $\hat{\mathbf{x}}_0$, a coarse estimate of the clean signal obtained via one-step denoising (see the section above on inverse problems). This allows an arbitrary number of steps to be skipped during reverse diffusion, which can significantly accelerate inference.

A progressive distillation method for reverse diffusion is used for text-to-speech generation in [36]. Leveraging DDIM sampling, a new student diffusion sampler learns at each iteration of the distillation process how to perform reverse diffusion using half as many steps as the current teacher. The resulting distilled sampler generates speech with similar quality as the original sampler using 64x more steps.

The noise variance schedule and time discretization used for reversed diffusion can also be optimized to reduce the number of steps, instead of being pre-defined. In [37], the schedule is learned by training an auxiliary hyper-network on top of existing denoising diffusion models. The resulting approach enables impressive speech generation results in as few as three reverse diffusion steps.

Finally, some auxiliary losses and training schemes are designed to ensure that the diffusion states remain as close as possible to the domain seen by the score network during training, thereby mitigating the so-called drifting bias. Lay et al. [38] propose a two-stage training method for diffusion-based speech enhancement following such a concept. The score network is first trained with denoising score matching and then fine-tuned to overfit a particular reverse diffusion sampler, by matching the final estimate of the solver to the clean speech target. High-quality speech enhancement is obtained with as few as one reverse diffusion step, reaching real-time computational complexity.

Causal processing: In real-time acoustic communications (e.g., hearing aids), future information can not be used to process the current signal which means processing must be causal. Diffusion models can be adapted for causal processing, as in [39], where the convolutional score network architecture and the audio level normalization procedure are modified to meet causality requirements.

Robustness to adverse conditions

Artifacts produced by diffusion models can differ in nature from those produced by statistical signal processing methods or predictive deep learning models. It was observed in [12] that speech enhancement diffusion models tend to hallucinate for negative input signal-to-noise ratios (SNRs), i.e., when noise dominates clean speech. This can lead to speech inpainting in noise-only regions, breathing and gasping artifacts, or the introduction of phonetic confusion, which may have a negative impact in real-world applications. This behavior can be mitigated by introducing external modalities such as video in [40], where lip movements are analyzed to determine the phoneme used as conditioning for score estimation guidance. Alternatively, as presented in StoRM [18] the input SNR can be first increased by using a predictive deep learning model to remove parts of the noise, at the potential cost of speech distortions. A generative diffusion model is then used to reconstruct the noisy and distorted speech, which was shown to help avoid hallucination effects and thus increase the robustness to challenging conditions.

Generative pre-training is another approach to increase robustness to outliers. It involves using a pretext task such as masked modeling to train the diffusion model in a self-supervised fashion. Masked modeling involves randomly masking some regions of audio and instructing the model to fill in those masked sections using the available context information, i.e., the non-masked regions. This pre-trained model can then be fine-tuned for a particular downstream task (e.g., speech enhancement, music restoration, etc.) using a supervised setting. Liu et al. [23] show that their diffusion model SpeechFlow benefits from generative pre-training, as it increases its robustness to adverse scenarios such as noise-dominated utterances in speech enhancement. They also notice that generative pre-training consistently increases performance for most speech restoration tasks.

Finally, running several realizations of the reverse diffusion process and measuring the empirical standard deviation of the obtained estimates can provide the user with a natural measure of uncertainty, which can help detect outliers and estimate the robustness of the approach on the given task.

CONCLUSION

This article discussed diffusion models as deep conditional generative models for audio restoration. We suggested that diffusion models can be considered as serious candidates for model-based audio processing, as we recalled that domain knowledge can be injected into various aspects of their design such as parameterization of diffusion trajectories, or modeling of a measurement likelihood for posterior sampling with diffusion priors. By categorizing

the various forms of conditioning proposed in diffusion approaches—namely input conditioning, task-adapted processes, and external conditioning—we highlight the structural flexibility of diffusion models and their resulting appreciable degree of interpretation. In particular, looking at audio restoration under the scope of solving inverse problems, we showed that we can combine diffusion models with Bayesian tools and stochastic optimization, thereby leveraging various parameterizations of degradation operators for informed and blind inverse problems. The quality of diffusion-based audio generation is remarkable, and although this can be originally outbalanced by disadvantages regarding practical requirements, e.g., robustness to adverse conditions or inference speed, we exposed several approaches and studies solving these drawbacks. We believe these solutions can be combined to yield robust, fast diffusion models for real-time acoustic communications.

ACKNOWLEDGMENTS

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169), DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and NordicSMC (Nordic Sound and Music Computing Network) with NordForsk project 86892.

AUTHORS

Jean-Marie Lemercier (jeanmarie.lemercier@uni-hamburg.de) received an M.Eng in Electrical Engineering in 2019 from Ecole Polytechnique, Paris, France. In 2020, he received a M.Sc. in Communications and Signal Processing from Imperial College London, London, United Kingdom. He is currently a PhD student in the Signal Processing group at Universität Hamburg, 20148 Hamburg, Germany. For his work on speech enhancement with diffusion-based generative models, he was the recipient of the Association for Electrical, Electronic, and Information Technologies Informationstechnische Gesellschaft (VDE ITG) Award 2024. He received the Best Student Paper Award at IWAENC 2024. His research interests span machine learning-based speech enhancement and dereverberation for hearing devices applications. Recent works also include the design and analysis of diffusion-based generative models for various speech restoration tasks. He is a Student Member of IEEE.

Julius Richter (julius.richter@uni-hamburg.de) received a B.Sc. and M.Sc. in Electrical Engineering in 2017 and 2019 from the Technical University of Berlin, Germany. He is currently a PhD student in the Signal Processing group at Universität Hamburg, 20148 Hamburg, Germany. For his work on speech enhancement with diffusion-based generative models, he was the recipient of the Association for Electrical, Electronic, and Information Technologies Informationstechnische Gesellschaft (VDE ITG) Award 2024. His research interests include deep generative models and multi-modal learning with applications to audio-visual speech processing. He is a Student Member of IEEE.

Simon Welker (simon.welker@uni-hamburg.de) received a B.Sc. in Computing in Science (2019) and an M.Sc. in Bioinformatics (2021) from Universität Hamburg, Germany. He is currently a Ph.D. student in the Signal Processing Group, Universität Hamburg, 20148 Hamburg, Germany, and the Coherent Imaging Division, Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany. For his work on speech enhancement with diffusion-based generative models, he was the recipient of the Association for Electrical, Electronic, and Information Technologies Informationstechnische Gesellschaft (VDE ITG) Award 2024. His research interests include machine learning techniques for solving inverse problems that arise in speech processing and X-ray imaging. He is a Student Member of IEEE.

Eloi Moliner (eloi.moliner@aalto.fi) received his B.Sc. degree in telecommunications technologies and services engineering in 2018 and his M.Sc. degree in telecommunications engineering in 2021 from the Polytechnic University of Catalonia, Spain. He is currently a Ph.D. candidate at the Acoustics Lab of Aalto University, 02150 Espoo, Finland. He is the winner of the Best Student Paper Award of the 2023 IEEE ICASSP conference. He received the Best Student Paper Award at IWAENC 2024. His research interests include digital audio restoration and audio applications of machine learning.

Vesa Välimäki (vesa.valimaki@aalto.fi) received his D.Sc. degree in electrical engineering from the Helsinki University of Technology in 1995. In 1996, he was a post-doctoral research fellow at the University of Westminster, London, U.K. In 2008–2009 he was a visiting scholar at Stanford University. He currently is a full professor of audio signal processing and vice dean for Research at Aalto University, 02150 Espoo, Finland. He is a fellow of the Audio Engineering Society and of the Asia-Pacific Artificial Intelligence Association, and is the editor-in-chief of the *Journal of the Audio Engineering Society*. His research interests include the application of machine learning and signal processing to audio technology. He is a Fellow of IEEE.

Timo Gerkmann (timo.gerkmann@uni-hamburg.de) received his Dr.-Ing. degree in electrical engineering and information sciences from Ruhr-Universität Bochum, Germany. He is a professor for signal processing with the Universität Hamburg, 20148 Hamburg, Germany. He was the recipient of the Association for Electrical, Electronic, and Information Technologies Informationstechnische Gesellschaft (VDE ITG) Award 2022 and 2024. He served in the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and is currently a senior area editor of the emphIEEE/ACM Transactions on Audio, Speech and Language Processing. His research interests include statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio–visual media, and human–machine interfaces. He is a Senior Member of IEEE.

REFERENCES

[1] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration—A Statistical Model Based Approach*. Springer, 1998.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement* (E. Vincent, T. Virtanen, and S. Gannot, eds.), John Wiley & Sons, 2018.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Neural Inf. Process. Syst.*, vol. 27, pp. 139–144, 2014.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Proc. Int. Conf. Learning Repr.*, pp. 1–14, 2014.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Neural Inf. Process. Syst.*, pp. 6840–6851, 2020.

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learning Repr.*, pp. 1–36, 2021.

[9] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–5, 2023.

[10] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Proc. Int. Conf. Learning Repr.*, pp. 1–15, 2021.

[11] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 7402–7406, 2022.

[12] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[13] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*. Springer, 2013.

[14] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd ed., 2007.

[16] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *Proc. Int. Conf. Learning Repr.*, pp. 1–17, 2021.

[17] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech*, 2022.

[18] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2724–2737, 2023.

[19] E. Moliner and V. Välimäki, "Diffusion-based audio inpainting," *J. Audio Eng. Soc.*, vol. 72, pp. 100–113, Mar. 2024.

[20] E. Moliner, F. Elvander, and V. Välimäki, "Blind audio bandwidth extension: A diffusion-based zero-shot approach," *arXiv:2306.01433*, 2024.

[21] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Machine Learning*, pp. 21450–21474, 2023.

[22] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proc. Int. Conf. Learning Repr.*, pp. 1–28, 2023.

[23] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching," in *Proc. Int. Conf. Learning Repr.*, pp. 1–20, 2024.

[24] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Neural Inf. Process. Syst.*, pp. 1–13, 2019.

[25] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," in *Proc. Interspeech*, pp. 1–5, 2023.

[26] E. Nachmani, R. S. Roman, and L. Wolf, "Denoising diffusion gamma models," in *Proc. Int. Conf. Learning Repr.*, pp. 1–13, 2022.

[27] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *Proc. Int. Conf. Learning Repr.*, pp. 1–30, 2023.

[28] J.-M. Lemercier, S. Welker, and T. Gerkmann, "Diffusion posterior sampling for informed single-channel dereverberation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 1–5, 2023.

[29] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, pp. 659–666, 2021.

[30] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023.

[31] T. Peer, S. Welker, and T. Gerkmann, "DiffPhase: Generative diffusion-based STFT phase retrieval," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–5, IEEE, 2023.

[32] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-channel blind unsupervised dereverberation with diffusion models," in *Proc. Int. Workshop Acoustic Signal Enhancement*, pp. 1–5, 2024.

[33] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop Acoustic Signal Enhancement*, pp. 1–4, 2010.

[34] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, and S. Zhao, "Audit: Audio editing by following instructions with latent diffusion models," in *Proc. Neural Inf. Process. Syst.*, pp. 71340–71357, 2023.

[35] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learning Repr.*, pp. 1–22, 2022.

[36] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "ProDiff: Progressive fast diffusion model for high-quality text-to-speech," in *ACM Multimedia*, 2022.

[37] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *Proc. Int. Conf. Learning Repr.*, pp. 1–22, 2022.

[38] B. Lay, J.-M. Lemercier, J. Richter, and T. Gerkmann, "Single and few-step diffusion for generative speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 636–630, 2024.

[39] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 780–789, 2024.

[40] J. Richter, S. Frintrop, and T. Gerkmann, "Audio-visual speech enhancement with score-based generative models," in *Proc. ITG Conf. Speech Communication*, pp. 275–279, 2023.

4.2 Causal Processing for Restoration of Speech Communication Artifacts

4.2.1 Publication: Causal Diffusion Models for Generalized Speech Enhancement

Abstract

In this work, we present a causal speech enhancement system that is designed to handle different types of corruptions. This paper is an extended version of our contribution to the "ICASSP 2023 Speech Signal Improvement Challenge". The method is based on a generative diffusion model which has been shown to work well in scenarios beyond speech-in-noise, such as missing data and non-additive corruptions. We guarantee causal processing with an algorithmic latency of 20 ms by modifying the network architecture and removing non-causal normalization techniques. To train and test our model, we generate a new corrupted speech dataset which includes additive background noise, reverberation, clipping, packet loss, bandwidth reduction, and codec artifacts. We compare the causal and non-causal versions of our method to investigate the impact of causal processing and we assess the gap between specialized models trained on a particular corruption type and the generalized model trained on all corruptions. Although specialized models and non-causal models have a small advantage, we show that the generalized causal approach does not suffer from a significant performance penalty, while it can be flexibly employed for real-world applications where different types of distortions may occur.

Reference

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, Tal Peer, and Timo Gerkmann, "Causal Diffusion Models for Generalized Speech Enhancement," *IEEE Open Journal of Signal Processing*, pp. 780–789, 2024. DOI: 10.1109/OJSP.2024.3379070

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2024 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

This is an extended paper based on the challenge submission [P8] to the *Signal Improvement Challenge* 2023. Julius Richter was taking the lead role in the collaborative challenge submission, while all other authors participated in developing the algorithm, training framework, and data preparation. Julius Richter is the lead author of this paper. He designed the new corrupted speech dataset, trained all models, conducted the experiments, and wrote the manuscript. Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Tal Peer helped with writing and revising the manuscript. Timo Gerkmann contributed through insights into experimental validation, mathematical derivations, and manuscript review.

Received 11 August, 2023; revised 16 January, 2024; accepted 24 January, 2024; Date of publication 19 March, 2024; date of current version 16 January, 2024.

Digital Object Identifier 10.1109/OJSP.2024.3379070

Causal Diffusion Models for Generalized Speech Enhancement

Julius Richter[1], Simon Welker[1, 2], Jean-Marie Lemercier[1], Bunlong Lay[1], Tal Peer[1], Timo Gerkmann[1]

[1]Signal Processing (SP), Universität Hamburg, Germany
[2]Center for Free-Electron Laser Science, DESY, Hamburg, Germany

Corresponding author: Julius Richter (email: julius.richter@uni-hamburg.de).

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169), DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380.

ABSTRACT In this work, we present a causal speech enhancement system that is designed to handle different types of corruptions. This paper is an extended version of our contribution to the "ICASSP 2023 Speech Signal Improvement Challenge". The method is based on a generative diffusion model which has been shown to work well in scenarios beyond speech-in-noise, such as missing data and non-additive corruptions. We guarantee causal processing with an algorithmic latency of 20 ms by modifying the network architecture and removing non-causal normalization techniques. To train and test our model, we generate a new corrupted speech dataset which includes additive background noise, reverberation, clipping, packet loss, bandwidth reduction, and codec artifacts. We compare the causal and non-causal versions of our method to investigate the impact of causal processing and we assess the gap between specialized models trained on a particular corruption type and the generalized model trained on all corruptions. Although specialized models and non-causal models have a small advantage, we show that the generalized causal approach does not suffer from a significant performance penalty, while it can be flexibly employed for real-world applications where different types of distortions may occur.

INDEX TERMS causal processing, diffusion models, generalized speech enhancement

I. INTRODUCTION

SPEECH enhancement algorithms typically address the problem of recovering clean speech signals from mixtures that contain speech and additive background noise [1]. Common approaches estimate the speech signal using a filter in the time-frequency domain to reduce noise while avoiding speech distortions [2]. In real-world speech communication, however, there are numerous other factors besides additive background noise that can degrade speech signals. These include reverberation, transmission errors, limited bandwidth, codec artifacts, and non-linearities of recording and playback equipment. To address this broad range of corruptions, several recent attempts have been made to find *generalized* (sometimes called *universal*) speech enhancement models [3]–[5]. This can be achieved by training a single model on a dataset comprising multiple corruptions [6], or employing a re-generation strategy, where a two-step approach is used to first enhance and then synthesize speech signals [7].

Classical speech processing tasks include denoising and dereverberation, which can be described as an (approximate) inversion of an additive or convolutive degradation model. For this purpose, deep learning-based methods often employ *predictive models* to either accomplish regression of clean waveform samples [8] and/or spectral components [9], or to estimate a multiplicative mask that is applied to the input spectrogram [10]. In contrast, tasks such as packet loss concealment or bandwidth extension are missing-data problems, where regression and masking-based approaches often reach their limitations [11]. These missing-data problems can be interpreted as audio inpainting tasks which require the generation of new signal content, i.e., filling the gaps of lost packets or missing frequency bands in a speech signal. Consequently, the use of *generative models* is a natural choice for these tasks. Per definition, generative models aim at learning the prior distribution of the target data. The learned prior distribution is then used to estimate

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

clean speech given a corrupted input that is assumed to lie outside the learned distribution.

Generative and predictive speech enhancement models have been shown to differ in the type of distortions they introduce into the clean speech estimates [12]. While predictive methods are prone to noise leakage and speech distortions, generative models tend to hallucinate when presented with challenging inputs, resulting in speech-like sounds with poor articulation and no semantic meaning [13]. However, since generative models leverage the learned speech prior, the enhanced speech will resemble the characteristics of the clean speech training data which is typically noise-free and of high quality. Moreover, due to their ability to deal with non-additive corruption types and missing data problems while still performing well on denoising and dereverberation tasks [11], generative speech enhancement models are good candidates for generalized speech enhancement.

Recently, diffusion-based generative models [14], or simply *diffusion models*, have been successfully applied to the task of speech enhancement [13], [15], [16]. Diffusion models define data recovery and generation problems as an iterative denoising task, in which a deep neural network (DNN), also called the *score model*, learns to remove the Gaussian noise that was progressively added in a forward diffusion process. To enable the conditional generation of clean speech given a corrupted signal, it was proposed to use a task-adapted diffusion process and condition the score model on the corrupted input [15], [16]. Currently, most diffusion models employ convolutional-based U-Net architectures as score models [13], [15]. However, standard U-Net-like architectures do not have a causal structure, meaning that the network considers information from the future to make predictions about the current time frame. This makes the resulting diffusion model unsuitable for online (and potentially real-time capable) signal processing where causality is a critical requirement.

In this paper, we present a causal speech enhancement method that is based on generative diffusion models and is designed to handle different types of corruptions. This is an extended version of our previous publication [17], which we contributed to the "ICASSP 2023 Speech Signal Improvement Challenge" ("SIG challenge") [18]. Compared to our original work on diffusion-based speech enhancement [13], we modify the network architecture to meet the causality requirement and now operate on super-wideband speech (32 kHz). By adjusting the short-time Fourier transform (STFT), we achieve an *algorithmic latency* of 20 ms, which we define as the latency introduced by frame-based processing, i.e., the latency that remains when using an infinitely fast processing device. Moreover, we address the more generalized speech enhancement problem by training the model on different corruption types. For this purpose, we generate a new dataset that contains additive background noise, reverberation, clipping, packet loss, bandwidth reduction, codec artifacts, as well as different combinations thereof. In the experiments, we evaluate on each individual corruption set and compare the model trained on all corruptions with specialized models only trained on the specific corruption type. Furthermore, we compare the causal approach against the non-causal counterpart. The results suggest that while specialized models and non-causal models have certain benefits, the causal model trained on all corruptions can keep up despite the causality constraint and provides a single model for generalized speech enhancement.

II. METHOD

In this section, we summarize our contributions to diffusion-based speech enhancement [13], [16] and highlight all modifications that were carried out to meet the causality requirement.

A. DATA REPRESENTATION

We represent all audio signals in the complex-valued STFT domain. Thus, comparable to complex spectral mapping [19], the diffusion model aims at estimating the clean real and imaginary spectrograms from the corrupted ones. Formally, we operate on complex spectrograms that are elements of $\mathbb{C}^{K\times F}$, where K denotes the number of time frames dependent on the audio length, and F is the number of frequency bins. To compensate for the typically heavy-tailed distribution of STFT speech amplitudes [20], we apply an amplitude transformation

$$\widetilde{c} = \beta|c|^{\alpha}\mathrm{e}^{i\angle(c)} \tag{1}$$

to all complex STFT coefficients c, where $\angle(\cdot)$ represents the angle of a complex number, $\alpha \in (0, 1]$ is a compression exponent which brings out frequency components with lower energy (e.g. fricative sounds of unvoiced speech) [21], and $\beta \in \mathbb{R}_+$ is a scaling factor to roughly normalize amplitudes within $[0, 1]$.

Hereafter, we denote clean speech as $\mathbf{x}_0 \in \mathbb{C}^d$ and its corrupted version as $\mathbf{y} \in \mathbb{C}^d$, both of which represent flattened and amplitude-transformed spectrograms consisting of $d = KF$ complex coefficients.

B. DIFFUSION MODEL

Following our original approach [13], we use a task-adapted diffusion process for the conditional generation of clean speech $\mathbf{x}_0$ given a corrupted input $\mathbf{y}$. To this end, we define a forward stochastic process whose mean interpolates linearly between $\mathbf{x}_0$ and $\mathbf{y}$. Using the continuous-time formulation for diffusion models [14], the forward process is modeled as the solution to the stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{2}$$

where $\mathbf{x}_t \in \mathbb{C}^d$ denotes the process state at time $t \in [0, T]$, $\gamma \in \mathbb{R}$ controls the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and $g(t) \in \mathbb{R}$ is the diffusion coefficient that controls the amount of Gaussian noise induced by a standard Wiener process $\mathbf{w}$. Note that t is only used to index the stochastic process and is completely unrelated to the time dimension of the audio signal.

The forward process can be time-inverted [14], resulting in a corresponding reverse process

$$\mathrm{d}\mathbf{x}_t = [-\gamma(\mathbf{y} - \mathbf{x}_t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{3}$$

where the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is intractable and therefore approximated by a *score model* $\mathbf{s}_\theta$, parameterized by θ.

To train the score model $\mathbf{s}_\theta$, we use the denoising score matching objective [22], which approximates the score function by estimating the Gaussian noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ injected by the Wiener process, resulting in the loss function

$$\arg\min_\theta \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})} \left[\left\| \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) + \frac{\mathbf{z}}{\sigma_t} \right\|_2^2 \right], \tag{4}$$

where the expectation is approximated by sampling all random variables at each training step. For a complete derivation of the loss function, we refer to our previous work [13].

Once the score model is trained, the reverse process (3) can be solved by replacing the score function with its approximation $\mathbf{s}_\theta$ and using a numerical SDE solver. For this purpose, we use the predictor-corrector sampler [14] with the configuration described in Section IV-B and initialize the reverse process with

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma_T^2 \mathbf{I})\,, \tag{5}$$

where σ_T^2 corresponds to the noise power at time step $t = T$. The enhancement process is then based on iterating through the reverse process starting at $t = T$ and ending at $t = 0$.

C. NETWORK ARCHITECTURE

As a score model, we use a modified version of the Noise Conditional Score Network (NCSN++) [14]. The network is a U-Net-like encoder-decoder architecture based on 2D convolutions, which takes complex spectrograms $\mathbf{x}_t$ and $\mathbf{y}$, and the process timestep t as input. Real and imaginary parts are considered as separate channels and the convolutions are performed over time and frequency. For the non-causal baseline, we use the same network configuration as in SGMSE+ [13], except that attention is only used in the bottleneck of the network, resulting in 64.8 M parameters.

To design a causal version of NCSN++, we apply the following modifications to the architecture. First, we modify the padding in the 2D convolutions and truncate the output such that the convolution along the time dimension is causal. An example of this procedure is illustrated in Figure 1 for a 1D convolution of kernel size 3. Next, we replace all group normalization layers with cumulative group normalization [23], aggregating statistics recursively. Downsampling in the time dimension is performed with strided convolutions and corresponding upsampling with transposed strided convolutions. Note that the strided and transposed strided convolutions remain causal operations with the padding described above. Up- and downsampling in the frequency dimension are realized with 1D finite impulse response filters [24]. Finally, all attention layers as well as the progressive growing path are removed from the network, resulting in 55.7M parameters in total. Considering all modifications, we would like to point out that the comparison between the causal and non-causal approach is also a comparison of different network architectures, respectively a comparison of the method in [13] with the proposed causal method.

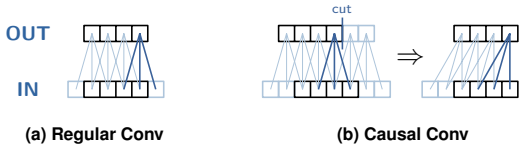


FIGURE 1. To obtain a causal convolution kernel, zero padding (light grey elements) is modified and the output is truncated.

Contrary to the SIG challenge submission [17], we choose not to apply an automatic gain control. Instead, we train the score model on various input gains. For the reverberation corruption set, however, we noted that the causal diffusion model had difficulties in estimating the correct output gain, resulting in enhanced files with monotonically increasing gain over the whole utterance. We addressed this issue by training the score model on normalized clean speech targets while maintaining the varying input level. Thus, the diffusion model learns to generate normalized clean speech conditioned on variable input levels which results in learning a simpler target distribution. This approach has shown improved robustness for reverberant data.

III. GENERALIZED SPEECH ENHANCEMENT

In this work, we extend the classical speech enhancement task of removing additive background noise and address a more generalized class of speech enhancement. More specifically, we consider the task of reconstructing clean speech that has been degraded by a set of different signal modifications, which we will describe in detail below. Besides training the model on different corruptions, we augment the data during training with variable input gains to make the model more robust against loudness variations.

A. BACKGROUND NOISE

To simulate speech recordings that have been degraded by additive background noise, we assume an additive mixture model

$$y = x + n \tag{6}$$

where the corrupted signal y is the sum of clean speech x and background noise n.

B. REVERBERATION

Reverberation is caused by room acoustics and is characterized by multiple reflections on the room enclosures [25]. Here, we assume a convolutive corruption model

$$y = x * h \tag{7}$$

where the clean speech x is convolved with a room impulse response (RIR) h representing the acoustic path between the source and the listener.

C. CLIPPING

Clipping distortion occurs when the amplitude of a signal exceeds the maximum range that can be faithfully represented by, e.g., a fixed precision digital system or an electrodynamic microphone with limited membrane displacement range. This is modeled by the corruption model

$$y = \min(\max(x, A_{\text{min}}), A_{\text{max}}) \tag{8}$$

where the speech signal x is clipped between a minimum amplitude A_{min} and a maximum amplitude A_{max}.

D. PACKET LOSS

In digital speech communication, such as voice-over-IP, data packets can get lost because of different transmission breakdown causes such as software corruption, interference, or jammed network nodes. Following the work in [26], we assume packets of length 20 ms and replace the signal with zeros if a packet is lost.

E. BANDWIDTH REDUCTION AND CODECS

Bandwidth reduction occurs when an audio signal at a high sampling rate is converted to a lower rate, e.g., to reduce the data transmission rate (bitrate) in a telecommunication system. The corruption process is linear and typically involves an anti-aliasing low-pass filter followed by a decimation operation:

$$y = \text{Resample}(x * a,\ f_s^{\text{h}},\ f_s^{\text{l}}) \tag{9}$$

where a is the impulse response of the anti-aliasing filter, f_s^{h} is the original high sampling rate and f_s^{l} is the lower target sampling rate.

Modern digital speech communication systems also employ source coding techniques to reduce the required bitrate. By considering specific signal characteristics (e.g. speech) and psychoacoustic models, lossy coding schemes offer a large reduction in bitrate but generally result in audible distortions and artifacts in the decoded signal, which become more severe as the compression ratio increases. In this work, we simulate the effect of coding by applying a speech or audio encoder, $\text{Enc}(\cdot)$, and the corresponding decoder, $\text{Dec}(\cdot)$, subsequently:

$$y = \text{Codec}(x) = \text{Dec}\left(\text{Enc}(x)\right) \tag{10}$$

F. COMBINATIONS

To simulate the occurrence of multiple corruptions at the same time, we use the aforementioned signal modifications and randomly select corruption chains among plausible candidates, e.g. $\{\text{Reverb} \rightarrow \text{BackgroundNoise} \rightarrow \text{PacketLoss}\}$.

TABLE 1. Clean speech files statistics for the MultiCorruption dataset.

		LibriVox [28]	VCTK [29]	Total
# Files	Train	26299	42830	69129
	Valid	414	780	1194
	Test	418	845	1263
# Speakers	Train	220	106	326
	Valid	4	4	8
	Test	4	2	6
Length [h]	Train	75.5	40.2	115.7
	Valid	1.2	0.7	1.9
	Test	1.2	0.6	1.8

IV. EXPERIMENTAL SETUP

A. DATA

To train and test our model, we generate a new dataset which we call *MultiCorruption* consisting of pairs of clean and corrupted speech files. We utilize clean speech data from the DNS challenge [27]. More specifically, we use a subset of the clean speech files from the LibriVox recordings [28] ("*read_speech*" directory) and the entire VCTK speech data [29]. To select the subset of the LibriVox recordings, we use speakers who have between 100 and 140 utterances and make the dataset gender-balanced by using an open-source gender recognition method[1]. This results in 228 speakers (114 male / 114 female), of which 220 are used for training and 4 each for validation and testing. The VCTK corpus contains 110 speakers with approximately equal numbers of male and female speakers. Following previous works [13], [15], [30], we are using speakers "*p226*" and "*p287*" for validation, speakers "*p232*" and "*p257*" for the test, and the remaining speakers for training. Table 1 contains further details regarding the clean speech data.

To generate the corrupted speech files, we define six corruption datasets (*Noise*, *Reverb*, *Clipping*, *PacketLoss*, *BWR-Codecs*, and *Combinations*), corresponding to the signal modifications listed in Section III. Note that we group bandwidth reduction and codecs into one corruption set as codecs also often include limited bandwidth. Each corruption set is generated using all clean speech files. Thus the total length of training data for the model trained on all corruptions results in 694.2 h. For additive background noise, we use noise files from the DNS challenge [27] and mix clean speech and noise at a segmental signal-to-noise ratio (SNR)[2] uniformly sampled in $[-10, 15]$ dB. Reverberant data is generated using RIRs also taken from the DNS challenge [27], but cropped such that the direct path occurs at the very beginning. This helps guide the training process because no delay is introduced between the reverberant and clean speech file. For clipping, we define A_{min} and A_{max} as percentile thresholds,

[1] https://github.com/x4nth055/gender-recognition-by-voice

[2] For the computation of the segmental SNR, we use the script from the DNS challenge [27].

such that A_{min} correspond to the threshold of the k-th percentile, where k is uniformly sampled between $[0, 30]$, and A_{max} corresponds to the threshold of the $(100 - k)$-th percentile. For packet loss, we randomly sample the amounts of consecutive packets lost per drop between $[1, 5]$, and sample the number of drops per second between $[3, 6]$. For bandwidth reduction, we pick an anti-aliasing filter type among Chebyshev, Butterworth, Elliptic, and Bessel and a filter order among $\{2,4,8\}$. Decimating is then realized with a down-scaling factor sampled in $\{2,4,8\}$. The utterance is then resampled at the original 32 kHz with polyphase filtering. For the codec corruption, we use the AMR-NB, AMR-WB, and EVS speech communication codecs, as well as MP3 coding with bitrates varying between 6.6 kb/s and 64 kb/s. Note that the AMR-NB and AMR-WB codecs implicitly include bandwidth reduction to 8 kHz and 16 kHz, respectively. In addition to the single corruptions, we also simulate various plausible combinations, using the same random corruption parameters as mentioned above.

Besides using the *MultiCorruption* dataset, we train and test our model on a 32 kHz version of the publically available *Voicebank-Demand* dataset [30]. This dataset is often used as a benchmark for single-channel speech enhancement. The utterances are artificially contaminated with eight real-recorded noise samples from the DEMAND database [31] and two artificially generated noise samples (babble and speech shaped) at SNRs of 0, 5, 10, and 15 dB. The test utterances are mixed with different noise samples at SNR levels of 2.5, 7.5, 12.5, and 17.5 dB.

B. HYPERPARAMETER SETTINGS

All processing is performed at a sampling frequency of 32 kHz. We use an STFT with a 638-point Hann window and 160-point hop, which results in an algorithmic latency of 20 ms, as we use purely causal processing. The diffusion process hyperparameters are identical to those in [13]. We use 30 diffusion steps for the reverse diffusion and adopt the predictor-corrector-scheme [14] with one step of annealed Langevin dynamics correction and a step size of 0.5. Please note that we have not optimized the sampler configuration for the individual corruption types, which could potentially improve the performance, as it was the case for dereverberation in [13]. We train the score model $\mathbf{s}_\theta$ with the denoising score matching objective [14] using the Adam optimizer with a learning rate of 10^{-4}. We train each model on two NVIDIA RTX A6000 GPUs using an effective batch size of $2 \times 8 = 16$ and an exponential moving average over the parameters with a factor of 0.999. To allow a fair comparison between the specialized and generalized models, we train all models for 300k training steps which takes around three days. Different from all other conducted experiments, the training for the Voicebank-Demand dataset is trained on four NVIDIA RTX A5000 GPUs with an effective batch size of $4 \times 4 = 16$ lasting for about two days.

C. EVALUATION METRICS

We describe here the standard speech enhancement metrics we used to assess the performance of the proposed method. A distinction is made between *intrusive* metrics computed by algorithms rating the processed signal in relation to the clean reference signal, and *non-intrusive* metrics which can be employed to evaluate real recordings when no clean reference is available.

Intrusive metrics include POLQA [32] for predicting speech quality which takes values from 1 (poor) to 5 (excellent) as usual for mean opinion scores (MOS). We also report PESQ [33], which is the predecessor of POLQA and is still widely used in the research community. The PESQ score lies between 1 (poor) and 4.5 (excellent). We further use ESTOI [34] as an intrusive measure of speech intelligibility. This metric yields values between 0 and 1, with higher values indicating better intelligibility. We report the speech mode of ViSQOL [35], a full-reference metric for perceived audio quality that is based on a model trained with data from subjective tests. Moreover, we calculate standard energy-based metrics including the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [36], and the log spectral distance (LSD). Both are measured in dB, with higher values for SI-SDR and lower values for LSD indicating better performance.

For non-intrusive speech quality assessment, we use the model DNSMOS P.835 [37] which is based on a listening experiment according to ITU-T P.835 [38] and provides three MOS scores: speech quality (SIG), background noise quality (BAK), and the overall quality (OVRL) of the audio. Furthermore, we use Wav-to-Vec MOS (WVMOS) [39] which is a non-intrusive MOS prediction method for speech quality evaluation using a fine-tuned wav2vec2.0 model [40].

V. RESULTS

In this section, we present the results of the proposed causal speech enhancement method and compare its performance against the non-causal counterpart. First, we present results for background noise removal using the standardized Voicebank-Demand dataset. Then, we investigate the method's generalization capability on different corruption types using the *MultiCorruption* dataset. Finally, we report the results evaluating on the SIG challenge blind set.

A. RESULTS ON VOICEBANK-DEMAND

In Table 2, we report speech enhancement results on the standardized Voicebank-Demand dataset. First, we discuss ablations using different normalization techniques. The non-causal model does not use any normalization before processing, whereas the method *Non-causal (norm. noisy)* is normalized based on the noisy mixture. We observe that these two methods differ only in a small amount in all reported metrics with a slight advantage for the model without normalization. Second, we show that changing the up- and downsampling in the NCSN++ architecture to non-

TABLE 2. Results obtained for the Voicebank-Demand. Values indicate mean and standard deviation.

Method	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL
Unprocessed	3.00 ± 0.88	2.01 ± 0.73	0.68 ± 0.17	8.4 ± 5.6	3.59 ± 0.68	2.66 ± 0.82	2.66 ± 0.88	2.16 ± 0.66
Causal	3.89 ± 0.65	2.74 ± 0.71	0.75 ± 0.14	16.8 ± 3.2	4.12 ± 0.47	3.13 ± 0.50	3.67 ± 0.45	2.76 ± 0.50
Non-causal	3.98 ± 0.63	2.63 ± 0.74	0.75 ± 0.14	18.3 ± 3.5	4.12 ± 0.48	3.16 ± 0.49	3.81 ± 0.35	2.83 ± 0.49
Non-causal (norm. noisy)	3.94 ± 0.60	2.62 ± 0.66	0.75 ± 0.13	17.8 ± 3.5	4.11 ± 0.46	3.16 ± 0.50	3.74 ± 0.41	2.81 ± 0.50
Non-causal (strided convs)	3.88 ± 0.60	2.69 ± 0.68	0.74 ± 0.14	16.3 ± 3.4	4.08 ± 0.48	3.13 ± 0.51	3.62 ± 0.47	2.74 ± 0.50

causal strided and transposed strided convolutions results in a performance reduction mostly audible in the amount of background noise reduction. This is also reflected in the metrics as *Non-causal (strided convs)* shows a 0.19 lower performance in BAK compared to the non-causal method using the original FIR filter as in [13]. Last, we see that enforcing the strided convolution to be causal as shown in Figure 1 has no negative impact on performance (see last row in Table 2). In summary, when accumulating the ablations we observe that the proposed causal configuration performs on par with the original non-causal NCSN++.

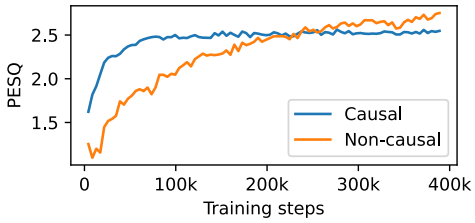


FIGURE 2. Training convergence on the Noise corruption set for the causal and the non-causal model.

B. RESULTS ON MULTIPLE CORRUPTIONS

To systematically analyze how well the proposed method performs on different corruption types, we test the model on the individual corruption sets described in Section III. For both causal and non-causal settings, we compare the generalized model trained on the entire dataset (denoted as *All*) with specialized models only trained on the respective corruption set. Listening examples can be found online[3].

We start by looking at the evaluation on the *Noise* corruption set which contains additive background noise (see Table 3). It can be seen that the non-causal method performs generally better than the causal method. This behavior is expected: due to speech being more temporally correlated than noise, the non-causal model can effectively use future information to disentangle speech and noise. However, the performance difference between the causal and non-causal variants is not significant for the generalized models trained on all data, where the non-causal model leads by a margin of only 0.08 in POLQA, 0.04 in PESQ, and 0.6 dB in SI-SDR. Listening to the enhanced files confirms that both methods sound comparable. For the specialized models, however, this difference is somewhat more pronounced, which can also be confirmed by informal listening, with fewer speech distortions for the non-causal model for challenging input.

Interestingly, we have noted different convergence behavior during training for the causal and the non-causal models. Figure 2 shows the speech enhancement performance in PESQ evaluated on 20 randomly selected utterances from the validation set after each epoch. While the causal model converges faster, the non-causal model eventually surpasses the causal model. This different behavior is potentially due to the modified network architecture and may result from different up- and downsampling with strided and transposed strided convolutions as opposed to regular convolutions with FIR filters (see Section II-C).

Next, in Table 4 we compare the results on the *Reverb* corruption set, which contains noiseless reverberant speech data. We note that the specialized models have an advantage over the generalized models for the intrusive metrics (except for ViSQOL, which seems to be insensitive to reverberation). This reveals that task-specific training on dereverberation helps to improve the performance compared to multi-task training. Furthermore, we observe that non-causal models have a consistent improvement over the causal models, which shows that future information can be effectively leveraged for dereverberation. Nevertheless, the performance of the causal models remains competitive and shows significant improvements with respect to the unprocessed files.

In Table 5, we report the results on the *Clipping* corruption set. For the intrusive metrics, we observe a rather large performance difference between the specialized models and the generalized models trained on all data. In the causal setup, the generalized model sees a drop of 0.57 in POLQA and PESQ, 0.06 in ESTOI, and 3.2 dB in SI-SDR. A possible cause for this is the inclusion of combinations such as clipping followed by bandwidth reduction in the *All* dataset. This leads to utterances where the resulting corruption is relatively different from that of pure clipping since clipping mainly introduces high frequencies that are subsequently removed by bandwidth reduction. Interestingly, we note that the SIG metric for the unprocessed data (i.e. the clipped signals) shows a high value which does not reflect our listening experience, indicating that this metric is not particularly suitable for assessing clipping artifacts.

[3]https://uhh.de/inf-sp-causal-diffusion

TABLE 3. Results obtained for the Noise test set. Values indicate mean and standard deviation.

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	1.76 ± 0.83	1.38 ± 0.50	0.57 ± 0.21	3.1 ± 10.0	2.7 ± 0.9	2.8 ± 1.0	2.2 ± 0.8	2.1 ± 0.7	1.1 ± 2.1
Causal	Noise	2.76 ± 1.10	2.30 ± 0.85	0.74 ± 0.19	14.0 ± 9.6	3.7 ± 0.8	3.5 ± 0.2	3.9 ± 0.3	3.2 ± 0.3	3.6 ± 0.9
Causal	All	2.60 ± 1.14	2.14 ± 0.84	0.71 ± 0.21	11.8 ± 11.4	3.6 ± 0.8	3.5 ± 0.3	3.9 ± 0.4	3.1 ± 0.3	3.4 ± 1.0
Non-causal	Noise	2.99 ± 1.08	2.46 ± 0.88	0.76 ± 0.18	15.0 ± 9.3	3.9 ± 0.7	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.8 ± 0.6
Non-causal	All	2.68 ± 1.08	2.18 ± 0.86	0.71 ± 0.21	12.4 ± 11.1	3.7 ± 0.8	3.4 ± 0.3	4.0 ± 0.2	3.1 ± 0.3	3.6 ± 0.9

TABLE 4. Results obtained for the Reverb test set. Values indicate mean and standard deviation.

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unprocessed	-	2.18 ± 0.57	1.70 ± 0.49	0.70 ± 0.13	-2.4 ± 6.6	4.1 ± 0.5	2.8 ± 0.7	3.2 ± 0.9	2.4 ± 0.6	3.0 ± 1.1
Causal	Reverb	3.03 ± 0.77	2.75 ± 0.66	0.86 ± 0.07	7.2 ± 5.4	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.3 ± 0.2	3.9 ± 0.5
Causal	All	2.83 ± 0.74	2.45 ± 0.66	0.82 ± 0.09	3.2 ± 6.5	4.1 ± 0.3	3.5 ± 0.2	4.0 ± 0.1	3.2 ± 0.2	3.7 ± 0.5
Non-causal	Reverb	3.19 ± 0.81	2.83 ± 0.71	0.87 ± 0.08	7.0 ± 6.8	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.3 ± 0.2	3.9 ± 0.5
Non-causal	All	3.05 ± 0.75	2.58 ± 0.66	0.84 ± 0.08	4.4 ± 6.5	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.2	3.9 ± 0.4

TABLE 5. Results obtained for the Clipping test set. Values indicate mean and standard deviation.

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	1.76 ± 0.80	1.63 ± 0.74	0.73 ± 0.13	3.3 ± 4.9	3.4 ± 0.8	3.4 ± 0.4	3.8 ± 0.4	3.0 ± 0.5	2.4 ± 1.4
Causal	Clipping	3.31 ± 0.79	3.15 ± 0.69	0.88 ± 0.07	12.7 ± 5.4	4.4 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.2	4.1 ± 0.4
Causal	All	2.74 ± 0.90	2.58 ± 0.78	0.82 ± 0.09	9.5 ± 5.9	4.1 ± 0.5	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.9 ± 0.5
Non-causal	Clipping	3.32 ± 0.73	2.99 ± 0.65	0.87 ± 0.08	12.6 ± 5.1	4.3 ± 0.4	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.2	4.1 ± 0.4
Non-causal	All	2.83 ± 0.83	2.49 ± 0.69	0.83 ± 0.09	9.9 ± 5.5	3.9 ± 0.5	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.9 ± 0.5

TABLE 6. Results obtained for the PacketLoss test set. Values indicate mean and standard deviation.

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	1.21 ± 0.27	1.31 ± 0.23	0.75 ± 0.07	5.8 ± 3.9	4.0 ± 0.4	2.9 ± 0.5	3.8 ± 0.3	2.6 ± 0.4	3.4 ± 0.6
Causal	PacketLoss	1.79 ± 0.64	2.02 ± 0.53	0.84 ± 0.06	7.4 ± 4.5	4.4 ± 0.2	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	4.1 ± 0.4
Causal	All	1.48 ± 0.55	1.86 ± 0.50	0.82 ± 0.07	6.7 ± 4.3	4.3 ± 0.3	3.4 ± 0.2	4.0 ± 0.1	3.2 ± 0.3	4.0 ± 0.4
Non-causal	PacketLoss	2.03 ± 0.64	1.95 ± 0.51	0.83 ± 0.06	7.7 ± 4.5	4.3 ± 0.3	3.4 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	4.0 ± 0.4
Non-causal	All	1.81 ± 0.62	1.87 ± 0.47	0.83 ± 0.07	7.0 ± 4.3	4.2 ± 0.3	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	4.0 ± 0.4

Table 6 reports the results using the *PacketLoss* corruption set. We can observe a clear improvement in all metrics with respect to the unprocessed files with benefits for the specialized models. This advantage of the specialized model seems to be more pronounced for the causal case, which is also confirmed by informal listening. However, for this typical inpainting task, we actually assumed that the non-causal model would perform significantly better than the causal model since future information should be heavily used by the non-causal model. This is, however, only the case for POLQA and is not reflected in the other metrics. We argue that the causal model still performs comparably well, because the gaps are relatively short (20-100 ms), and the past temporal context given by the receptive field of the causal model still covers enough speech parts to draw inferences to fill the gaps.

In Table 7, we show the results for the *BWR-Codecs* corruption set, which contains bandwidth-reduced audio and codec artifacts. We notice that the unprocessed files already yield good values for speech quality and intelligibility measures. While the specialized models further improve these metrics, the generalized models slightly reduce them, although they still turn out to be quite high (PESQ and POLQA > 3.5) in absolute terms. It should, however, be noted that PESQ and POLQA have been reported to correlate poorly with listening experiments for bandwidth reduction [11]. We thus also include the log spectral distance (LSD) as a common metric used for bandwidth extension and observe that the causal models perform slightly better in it than the non-causal model. Note that for brevity, we omit SIG, BAK, and OVRL as they have not shown any difference on this corruption set.

TABLE 7. Results obtained for the BWR-Codecs test set. Values indicate mean and standard deviation.

Method	Training	POLQA ↑	PESQ ↑	ESTOI ↑	SI-SDR [dB] ↑	LSD [dB] ↓	ViSQOL ↑	WVMOS ↑
Unproc.	-	3.88 ± 0.74	3.76 ± 0.78	0.91 ± 0.11	11.2 ± 18.6	1.90 ± 0.41	4.4 ± 0.5	3.86 ± 0.60
Causal	BWR+Codecs	4.06 ± 0.65	3.89 ± 0.70	0.94 ± 0.07	17.9 ± 7.7	1.46 ± 0.19	4.6 ± 0.4	4.12 ± 0.39
Causal	All	3.66 ± 0.93	3.59 ± 0.95	0.92 ± 0.10	13.1 ± 14.2	1.45 ± 0.24	4.5 ± 0.5	4.01 ± 0.48
Non-causal	BWR+Codecs	4.13 ± 0.57	3.94 ± 0.60	0.94 ± 0.07	18.4 ± 7.1	1.57 ± 0.26	4.5 ± 0.4	4.13 ± 0.39
Non-causal	All	3.88 ± 0.74	3.68 ± 0.73	0.91 ± 0.10	12.2 ± 17.6	1.46 ± 0.21	4.4 ± 0.4	4.05 ± 0.48

TABLE 8. Results obtained for the All test set. Values indicate mean and standard deviation.

Method	Training	POLQA	PESQ	ESTOI	SI-SDR [dB]	ViSQOL	SIG	BAK	OVRL	WVMOS
Unproc.	-	2.03 ± 1.09	1.83 ± 1.03	0.68 ± 0.21	2.2 ± 11.9	3.5 ± 1.0	2.9 ± 0.8	3.2 ± 1.1	2.5 ± 0.8	2.3 ± 2.0
Causal	All	2.49 ± 1.13	2.35 ± 0.98	0.77 ± 0.18	7.1 ± 10.7	3.9 ± 0.8	3.5 ± 0.2	4.0 ± 0.3	3.1 ± 0.3	3.6 ± 0.9
Non-causal	All	2.67 ± 1.09	2.39 ± 0.95	0.78 ± 0.19	7.2 ± 11.8	3.9 ± 0.7	3.5 ± 0.2	4.1 ± 0.1	3.2 ± 0.3	3.8 ± 0.6

TABLE 9. Results obtained for the blind set of the SIG challenge.

Method	SIG	BAK	OVRL	WVMOS
Unprocessed	2.9 ± 0.6	3.4 ± 0.7	2.5 ± 0.6	2.2 ± 1.2
Causal	3.2 ± 0.4	3.8 ± 0.4	2.9 ± 0.4	2.4 ± 0.8
Non-causal	3.3 ± 0.4	4.0 ± 0.2	3.0 ± 0.4	2.7 ± 0.8

Finally, Table 8 shows the average performance using all corruption sets, including different combinations of corruptions. It can be seen, that the non-causal model has a slight advantage over the causal approach but does not suffer from a significant performance penalty except a drop of 0.22 in POLQA.

In summary, we find that task-specific models trained on a specific corruption type typically outperform generalized models using multi-task training. This follows the intuitive explanation that a more complex task leads to worse overall performance for the same network capacity. Similar behavior has also been reported for image restoration tasks [41]. In addition, we observe that the modification to causal processing slightly degrades the performance. This is consistent with the intuition that the causal model can only use information from the past, while the non-causal model can also leverage future information for the estimation.

C. RESULTS ON THE SIG CHALLENGE BLIND SET

In Table 9, we present the results when evaluating the causal and non-causal model on the 500 files from the 2023 SIG challenge's blind test set [18]. Since there is no reference signal available, we can only report non-intrusive metrics. While both models improve all metrics, there is a small advantage for the non-causal model.

D. COMPUTATIONAL COMPLEXITY

With the given reverse sampler configuration, the proposed method requires calling the score model 60 times to process an utterance. While designed to run on a GPU, this currently poses a practical limitation in real-time applications on a CPU. Processing a file of 10 s requires 236 TMACs and takes 30.5 s on a GPU and 30 minutes on a CPU[4]. However, since the method performs strictly causal signal processing, real-time operation is still theoretically possible, assuming that the computational complexity can be decreased in the future. One possible avenue towards decreased complexity is reducing the number of diffusion steps, e.g., as in [42]. This is an active area of research which we plan to focus on in future work.

VI. CONCLUSION

In this work, we have presented a causal speech enhancement method with an algorithmic latency of 20 ms that is based on generative diffusion models. To guarantee causal processing, we have modified the network architecture and removed all non-causal normalization techniques. We addressed a more generalized speech enhancement scenario beyond speech-in-noise by training and testing the model on multiple corruption types. For this purpose, we generated a new corrupted speech dataset which includes additive background noise, reverberation, clipping, packet loss, bandwidth reduction, and codec artifacts. In the experiments, we have conducted evaluations on the individual corruption sets and compared the generalized model trained on all data with specialized models trained solely on each specific corruption type. Moreover, we compared the causal diffusion model against the non-causal baseline. Our findings revealed that, although specialized models and non-causal models have a small advantage, the generalized causal approach does not suffer from a significant performance penalty, while being more practicable for real-world applications.

[4] Average processing time for 10 audio files on an NVIDIA GeForce RTX 2080 Ti GPU, in a machine with an Intel Core i7-7800X CPU @ 3.50GHz.

REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state-of-the-art*. Morgan & Claypool, 2013.

[3] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," in *ISCA Interspeech*, 2019, pp. 1791–1795.

[4] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv preprint arXiv:2109.13731*, 2021.

[5] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[6] J. Su, A. Finkelstein, and Z. Jin, "Perceptually-motivated environment-specific speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7015–7019.

[7] J. Chen, Y. Shi, W. Liu, W. Rao, S. He, A. Li, Y. Wang, Z. Wu, S. Shang, and C. Zheng, "Gesper: A unified framework for general speech restoration," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[8] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *ISCA Interspeech*, 2020, pp. 3291–3295.

[9] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2014.

[10] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 7, pp. 1492–1501, 2017.

[11] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[12] D. de Oliveira, J. Richter, J.-M. Lemercier, T. Peer, and T. Gerkmann, "On the behavior of intrusive and non-intrusive speech enhancement metrics in predictive and generative settings," in *ITG Conference on Speech Communication*, 2023.

[13] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 31, pp. 2351–2364, 2023.

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *International Conference on Learning Representations (ICLR)*, 2021.

[15] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7402–7406, 2022.

[16] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," *ISCA Interspeech*, pp. 2928–2932, 2022.

[17] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Speech signal improvement using causal generative diffusion models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[18] R. Cutler, A. Saabas, B. Naderi, N.-C. Ristea, S. Braun, and S. Branets, "ICASSP 2023 speech signal improvement challenge," *arXiv preprint arXiv:2303.06566*, 2023.

[19] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," *IEEE International Workshop on Machine Learning for Signal Proc. (MLSP)*, 2017.

[20] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," *International Workshop on Acoustic Echo and Noise Control*, 2010.

[21] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," *International Conference on Telecom. and Signal Processing*, pp. 72–76, 2021.

[22] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[23] Y. Wu and K. He, "Group normalization," *Proc. of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[24] R. Zhang, "Making convolutional networks shift-invariant again," *International Conference on Machine Learning (ICML)*, pp. 7324–7334, 2019.

[25] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2011, vol. 59.

[26] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "Interspeech 2022 audio deep packet loss concealment challenge," *arXiv preprint arXiv:2204.05222*, 2022.

[27] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 deep noise suppression challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[28] "LibriVox: Free public domain audiobooks," https://librivox.org/, accessed: 2023-07-24.

[29] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *ISCA Speech Synthesis Workshop (SSW)*, pp. 146–152, 2016.

[31] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[32] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[33] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, 2001.

[34] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 11, pp. 2009–2022, 2016.

[35] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *2020 twelfth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.

[37] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[38] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *Int. Telecom. Union (ITU)*, 2003. [Online]. Available: https://www.itu.int/rec/T-REC-P.835-200311-I/en

[39] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[40] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[41] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[42] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," in *ISCA Interspeech*, 2023.

4.3 Extension to Audio-Visual Speech Enhancement

4.3.1 Publication: Audio-visual speech enhancement with score-based generative models

Abstract

This paper introduces an audio-visual speech enhancement system that leverages score-based generative models, also known as diffusion models, conditioned on visual information. In particular, we exploit audio-visual embeddings obtained from a self-supervised learning model that has been fine-tuned on lipreading. The layer-wise features of its transformer-based encoder are aggregated, time-aligned, and incorporated into the noise conditional score network. Experimental evaluations show that the proposed audiovisual speech enhancement system yields improved speech quality and reduces generative artifacts such as phonetic confusions with respect to the audio-only equivalent. The latter is supported by the word error rate of a downstream automatic speech recognition model, which decreases noticeably, especially at low input signal-to-noise ratios.

Reference

Julius Richter, Simone Frintrop, and Timo Gerkmann, "Audio-visual speech enhancement with score-based generative model", *Proceedings of ITG Conference on Speech Communication*, pp. 275–279., 2023. DOI: 10.30420/456164054

Copyright Notice

The following article is the accepted version of the article published with VDE Verlag GmbH. © 2023 VDE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Julius Richter is the lead author of this paper. He developed the algorithm, trained all models, conducted the experiments, and wrote the manuscript. Simone Frintrop and Timo Gerkmann contributed through insights into experimental validation, mathematical derivations, and manuscript review.

Audio-Visual Speech Enhancement with Score-Based Generative Models

Julius Richter[1], *Simone Frintrop*[2], *Timo Gerkmann*[1]

[1]Signal Processing Group, [2]Computer Vision Group, Department of Informatics, Universität Hamburg, Germany
Email: {julius.richter,simone.frintrop,timo.gerkmann}@uni-hamburg.de

Abstract

This paper introduces an audio-visual speech enhancement system that leverages score-based generative models, also known as diffusion models, conditioned on visual information. In particular, we exploit audio-visual embeddings obtained from a self-supervised learning model that has been fine-tuned on lipreading. The layer-wise features of its transformer-based encoder are aggregated, time-aligned, and incorporated into the noise conditional score network. Experimental evaluations show that the proposed audio-visual speech enhancement system yields improved speech quality and reduces generative artifacts such as phonetic confusions with respect to the audio-only equivalent. The latter is supported by the word error rate of a downstream automatic speech recognition model, which decreases noticeably, especially at low input signal-to-noise ratios.

Index Terms— audio-visual speech enhancement, diffusion models, self-supervised learning.

1 Introduction

Robust speech processing, e.g. in telecommunications or for automatic speech recognition (ASR), often requires improved speech quality and intelligibility. For this purpose, speech enhancement aims at estimating clean speech signals from noisy audio recordings [1]. However, in various real-world scenarios, such as in negative signal-to-noise ratios (SNRs), speech enhancement systems may face limitations, typically resulting in speech distortions and impaired speech recognition.

To mitigate these limitations, promising approaches have been proposed that leverage both audio and video information which is referred to as audio-visual speech enhancement (AVSE) [2]. In fact, incorporating visual information into audio-based systems becomes an emerging research direction and improves both the robustness and the overall performance [3]. The principle is in general twofold: first, the visual information (e.g. lip movements) is usually not affected by the acoustic environment, and second, it has been proven that visual information is able to provide additional speech and speaker-related cues [4, 5].

Prevailing deep learning approches proposed for AVSE typically involve models with modality-specific encoders, feature fusion, and a decoder, all trained end-to-end in a supervised fashion [6, 7]. Their goal is to learn audio-visual representations suitable for speech enhancement which is usually defined as a predictive regression task. Another approach to speech enhancement is to use generative models which follow a different learning paradigm. They aim to model the underlying statistical distribution of clean speech and do not directly estimate the mapping from noisy input to clean output.

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169). We would like to thank J. Berger and Rohde&Schwarz SwissQual AG for their support with POLQA.

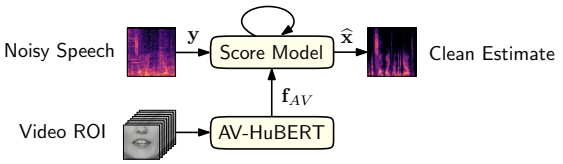


Figure 1: Proposed audio-visual speech enhancement system. The score model iteratively refines noisy speech $\mathbf{y}$ to obtain a clean estimate $\hat{\mathbf{x}}$ and is conditioned on audio-visual embeddings $\mathbf{f}_{AV}$ obtained from AV-HuBERT.

Recently, score-based generative models have been proposed for speech enhancement [8, 9]. The idea is to corrupt clean speech with slowly increasing levels of noise and train a deep neural network (DNN) to reverse this corruption iteratively. This exciting new research direction provides an alternative to other generative approaches and differs from predictive approaches in its methodology and results. Since the models are trained to generate clean speech conditioned on the noisy input, they tend to hallucinate for signals with low SNRs [9]. This can lead to speech-like sounds with poor articulation and no linguistic meaning or cause phonetic confusions, which may be problematic in many practical applications.

In this paper, we build upon our previous work on score-based generative models for speech enhancement [9] and develop an audio-visual extension of the generative approach. We condition the noise conditional score network on audio-visual embeddings, that are obtained from AV-HuBERT [10], a self-supervised learning (SSL) model fine-tuned on lipreading (see Fig. 1). This follows a recent trend in deep learning-based speech processing to leverage powerful representations from pre-trained models, which have been trained on large-scale datasets using SSL techniques.

Experimental evaluations show that the proposed AVSE system provides better speech quality and reduces generative artifacts, such as phonetic confusions, compared to its audio-only counterpart. This observation is supported by the reduction of the word error rate (WER) of a downstream ASR model, which is particularly evident at low input SNRs.

2 Related Work

2.1 Score-based Generative Models

Score-based generative models, also known as diffusion models, have recently gained attention due to their ability to generate high-quality samples and provide flexible and efficient training procedures. Originally inspired by non-equilibrium thermodynamics [11], they exist in several variants [12, 13]. All of them share the idea of gradually turning data into noise, and training a DNN that learns to invert this diffusion process for different noise scales. In an iterative refinement procedure, the DNN is evaluated multiple times to generate data, achieving a trade-off between computational complexity and sample quality.

Song et al. [14] showed that the diffusion process can be expressed as a solution of a stochastic differential equation (SDE), establishing a continuous-time formulation for diffusion models. The forward SDE can be inverted in time, resulting in a corresponding reverse SDE which depends only on the score function of the perturbed data [15].

2.2 Audio-Visual Speech Enhancement

Early works on AVSE can be traced back to the 1990s [16]. More recently, a number of interesting algorithms have been developed in the context of deep learning, including methods based on spectral mapping/masking that are trained in a supervised fashion [6, 7]. These methods typically consist of two modality-specific encoders, one for processing noisy speech and another for processing video information. The resulting audio and visual features are then concatenated and processed to obtain an audio-visual embedding that is fed into a decoder to predict the clean speech.

Alternatively, Sadeghi et al. [17] proposed a generative approach to AVSE using a conditional variational autoencoder trained on clean speech only. The model learns a clean speech prior that is conditioned on visual information. At inference, this prior is combined with a noise model based on nonnegative matrix factorization to estimate a probabilistic Wiener filter for denoising.

Other approaches consider AVSE as a speech resynthesis task, i.e., to create an output that closely resembles the original signal in terms of sound quality, prosody, and other acoustic characteristics [18]. Yang et al. [19] developed a speaker-dependent model that generates clean speech codes of a neural audio codec conditioned on noisy audio and visual inputs. The discrete codes obtained are used to reconstruct a mel-spectrogram which is then fed into a neural vocoder. Hsu et al. [20] break AVSE into audio-visual speech recognition and text-to-speech exploiting hidden units from a pre-trained SSL-model. Mira et al. [21] proposed a two-stage approach that predicts clean mel-spectrograms from audio-visual speech and converts them into waveform audio using a neural vocoder.

2.3 Self-Supervised Learning

Recently, SSL has gained significant attention in the speech community, with well-known models such as wav2vec 2.0 [22] and HuBERT [23] for audio-only and AV-HuBERT [10] for audio-visual speech representation learning. In contrast to supervised learning, which involves direct optimization of models for a specific task, SSL adopts a two-step approach. First, the model undergoes pre-training on large-scale unlabeled data to extract task-agnostic representations. Then, these learned representations are either used as input to a supervised model, or the model itself is fine-tuned for a particular downstream task. The expected outcome is to either improve the downstream task's performance or reduce the amount of labeled data required for training.

Pasad et al. [24] have analyzed the layer-wise features of wav2vec 2.0 and found that the transformer layers follow an autoencoder-like behavior in which the representation deviates from the input speech features as the depth of the model increases, followed by a reverse trend in which even deeper layers become more similar to the input. Furthermore, the evolution of the representations follows an acoustic-linguistic hierarchy. The shallowest layers encode acoustic features, followed by phonetic, word identity, and word meaning information [24]. Huang et al. [25] have investigated SSL model for speech enhancement and separation and concluded although they are not designed for waveform generation tasks like enhancement and separation, some of them achieve remarkable improvements over spectrogram magnitudes. Interestingly, instead of extracting representations from the last hidden layer, they proposed to aggregate embeddings from all layers with a learnable weighted sum.

3 Method

In this section, we describe the proposed AVSE system, which we name *AV-Gen*. Fig. 1 shows an overview of AV-Gen which consists of a score model and a pre-trained AV-HuBERT model as an upstream video processor.

3.1 Score Model

Following the work in [9], we cast speech enhancement as a conditional generation task using score-based generative models. For a pair of clean and noisy spectrograms $\mathbf{x}_0, \mathbf{y} \in \mathbb{C}^d$ with d denoting the number of time-frequency bins, we define a stochastic diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$, called *forward process*, as the solution to the SDE,

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{1}$$

where $\mathbf{x}_t$ denotes the process state at time $t \in [0, T]$, $\gamma \in \mathbb{R}$ controls the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and $g(t) \in \mathbb{R}$ is the diffusion coefficient that controls the amount of Gaussian white noise induced by a standard Wiener process $\mathbf{w}$.

The forward process has an associated *reverse process* which is given by the reverse SDE [14, 15],

$$\mathrm{d}\mathbf{x}_t = [-\gamma(\mathbf{y} - \mathbf{x}_t) + g(t)^2 \underbrace{\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})}_{\approx \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, \mathbf{f}_{AV}, t)}]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}\,, \tag{2}$$

where the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is intractable and therefore approximated by the score model $\mathbf{s}_\theta$ with learnable parameters θ. As in [9], we use the Noise Conditional Score Network (NCSN++) [14] as the score model. The training objective is denoising score matching, as defined in [9, Eq. (15)]. Note, however, that in contrast to [9], the score model is additionally conditioned on the video modality by taking conditional features $\mathbf{f}_{AV}$ as input.

At inference, we can generate a clean speech estimate $\widehat{\mathbf{x}}$ by solving the reverse SDE in Eq. (2). For this purpose, a numerical SDE solver is employed which solves the task of speech enhancement iterating through the reverse process starting at $t = T$ and ending at $t = t_\epsilon$, where t_ϵ is a small minimum process time to avoid numerical instabilities.

3.2 AV-HuBERT as Conditional Model

As an upstream video processor for the score model, we use a pre-trained AV-HuBERT model consisting of a hybrid ResNet-transformer architecture for which we freeze the parameters at training. In particular, we exploit audio-visual embeddings from each layer of its transformer encoder, as visualized in Fig. 2. For each encoder layer k of the score model, we learn a conditional feature $\mathbf{f}_k$ by aggregating the layer-wise embeddings $\mathbf{e}_l$ of AV-HuBERT's transformer encoder, as

$$\mathbf{f}_k = \mathrm{TempAlign}_k\left(\sum_{l=1}^{L} w_{kl}\,\mathbf{e}_l\right), \quad k = 1, \ldots, K\,, \tag{3}$$

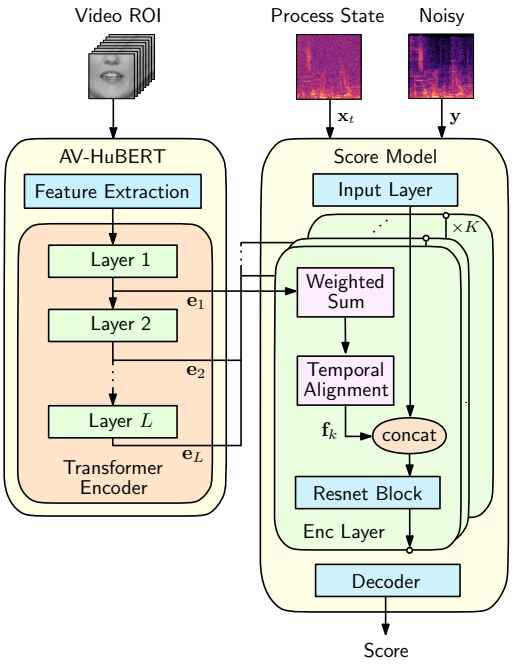


Figure 2: Network architecture: Each encoder layer k of the score model is conditioned on a conditional feature $\mathbf{f}_k$ which is a weighted sum of the layerwise embedding $\mathbf{e}_l$ of the AV-HuBERT's transformer encoder.

with trainable weights w_{kl} with the properties $w_{kl} \geq 0$ and $\sum_l w_{kl} = 1$ for all k. The operation $\text{TempAlign}_k(\cdot)$ ensures that the features are temporally aligned to the feature rates of the score model's embeddings for different resolutions. In particular, we use linear interpolation for upsampling and trainable Conv1D layers with kernel size and stride d for downsampling with factor d. The features are broadcasted over the frequency dimension to match the feature maps of the score model and the number of channels is adopted in the temporal alignment operation. Finally, the conditional features $\mathbf{f}_{AV} = \{\mathbf{f}_1, \ldots, \mathbf{f}_K\}$ have the same shapes as the embeddings of the score model before concatenation.

4 Experimental Setup

4.1 Dataset

For audio-visual speech data, we use the LRS3 dataset [26], a large-scale dataset primarily designed for the task of audio-visual speech recognition. Besides the audio tracks sampled at 16 kHz, it includes face tracks sampled at 25 Hz from over 400 hours of TED and TEDx videos. Note that recording conditions vary among the files and the data sometimes contain reverberant speech and is not as clean as data recorded in controlled environments such as a recording studio. For practical reasons, we limit the size of training data and use the low-resource 30 hours subset introduced in [10]. To meet the requirements for evaluating speech quality [27], we only consider files from the test set with a minimum length of 3.2 s resulting in 244 test files in total.

We generate pairs of clean and noisy speech by mixing the audio tracks of LRS3 with noise signals from the CHiME3 dataset [28] at SNRs sampled uniformly between -6 and 12 dB for the training, validation, and test. Consequently, we name the resulting data set LRS3-CHiME3.

4.2 Baselines

The primary objective of this paper is to investigate the impact of incorporating visual information into SGMSE+ [9], our previous work on score-based generative models for speech enhancement. To accomplish this, we train SGMSE+ using identical settings as those employed for AV-Gen. This encompasses the input representation, SDE parameters, and matching training configurations, including batch size and learning rate. Considering that both methods employ the same DNN as the score model, with the only distinction being the visual conditioning, it is reasonable to regard SGMSE+ as the audio-only counterpart to AV-Gen.

Furthermore, we consider VisualVoice [29] as an audio-visual baseline, a method originally proposed for audio-visual speech separation. This method learns a direct mapping from noisy to clean speech conditioned on visual information. To this end, it utilizes an encoder-decoder architecture (54,7M parameters) for complex mask prediction, in which the audio features are combined with visual features extracted from a lip motion analysis network. Here, we only use the mask prediction loss and omit the facial analysis network, as we adopt the method for AVSE.

4.3 Hyperparameter Setting

4.3.1 Input Representation

The complex spectrogram is computed with the short-time Fourier transform (STFT) using a Hann window of length 510 and a hop size of length 160 resulting in 256 unique frequency bins and an audio frame rate of 100 Hz. Consequently, the audio frame rate is a multiple of the video frame rate, which provides simple temporal alignment for feature fusion to ensure audio-visual synchronization. We use the same spectrogram transformation as in [9]. Unlike the original work, we do not normalize the signal in the time domain, which should make the score model more robust to varying input conditions.

For video pre-processing, we follow the protocol in [10] which stems from prior works in lipreading [30, 31]. However, instead of using `dlib` for facial landmark detection, we use the Face Alignment Network [32] because we found better robustness when inspecting the processed data. Using the 68 facial landmarks, each video clip is aligned to a reference face frame via affine transformation and cropped to a 88×88 region-of-interest (ROI) centered on the mouth.

4.3.2 Score model and training configuration

We parameterize the SDE in Eq. (1) and the NCSN++ architecture with the same configuration as in [9], resulting in 76M learnable parameters. We train the score model with variable length sequences (between 2 and 12 seconds), using the `MaxTokenBucketizer` from TorchData, which forms batches of variable batch size. We utilize four Quadro RTX 6000 (24 GB memory each) and train for 60 epochs using the distributed data-parallel (DDP) approach in PyTorch Lightning, which takes about 2.5 days. We use the Adam optimizer [33] with a learning rate of $5 \cdot 10^{-4}$. At inference, we use 30 diffusion steps and adopt the predictor-corrector-scheme [14] with one step of annealed Langevin dynamics correction with step size 0.5 and set $t_\epsilon = 0.03$.

4.3.3 AV-HuBERT

We use AV-HuBERT Base with $L = 12$ transformer layers. We use the fine-tuned model for visual speech recognition

Method	AV	POLQA	PESQ	ESTOI	SI-SDR [dB]	WER	S	D	I
Mixture		1.68 ± 0.55	1.20 ± 0.20	0.77 ± 0.13	2.6 ± 5.1	0.28 ± 0.30	2.2	1.5	**0.1**
SGMSE+ [9]		2.73 ± 0.61	2.08 ± 0.56	0.89 ± 0.08	10.6 ± 4.8	0.22 ± 0.25	2.3	0.6	0.2
VisualVoice [29]	✓	2.78 ± 0.60	2.11 ± 0.57	0.89 ± 0.07	11.0 ± 4.0	0.15 ± 0.18	1.5	**0.3**	0.2
AV-Gen (ours)	✓	$\mathbf{2.87 \pm 0.55}$	$\mathbf{2.23 \pm 0.53}$	$\mathbf{0.90 \pm 0.06}$	$\mathbf{11.4 \pm 3.9}$	$\mathbf{0.11 \pm 0.14}$	**1.1**	**0.3**	0.2

Table 1: Average speech enhancement performance obtained for the LRS3-CHiME3 test set. Values indicate the mean and standard deviation for the 244 test files.

that used the complete LRS3 dataset (433 h) as pretraining and finetuning data. This model takes only the video modality as input. We also experimented with the pre-trained model trained on the complete LRS3 dataset using both modalities as input, which resulted in inferior performance.

4.4 Metrics

For an instrumental evaluation of speech quality and intelligibility, we use standard metrics, including POLQA [34], PESQ [27], ESTOI [35], and SI-SDR [36].

To evaluate the effect of speech enhancement on ASR, we use QuartzNet (QuartzNet15x5Base-En) from the NeMo toolkit [37] as a downstream ASR system. We measure the WER, which is defined as WER $= (S + D + I)/N$, where S is the number of substitutions, D is the number of deletions, I is the number of insertions needed to get from the predicted to the target sequence, and N is the number of words in the target sequence.

5 Results

In Tab. 1, we report the speech enhancement performance obtained for the LRS3-CHiME3. Comparing AV-Gen to its audio-only counterpart SGMSE+, we see that the inclusion of video information improves the speech quality, indicated by an improvement of 0.14 in POLQA and 0.15 in PESQ. ESTOI is only marginally better, suggesting that both methods are fairly similar in terms of intelligibility. However, the evaluation on ASR shows an improved WER of 0.11 on average. We argue that this is mainly because SGMSE+ tends to produce generative artifacts like phonetic confusions at low input SNRs [9], which is alleviated by using video information. In fact, the reduced number of substitutions and deletions on average reveals a stronger robustness with respect to generative hallucinations, which is also verified by informal listening. Moreover, AV-Gen seems to be competitive with the audio-visual baseline VisualVoice, outperforming it in all metrics.

In informal listening, we find that AV-Gen occasionally generates a more reverberant output compared to the noisy input. This could be due to the generative behavior of the proposed model, as the data used for training was recorded in the wild and also contains reverberant speech. To prevent this effect, the training data could be cleaned beforehand.

To gain more insights into the effect of exploiting video information in AV-Gen as opposed to SGMSE+, we evaluate PESQ and WER at different input SNRs. Fig. 3 shows the mean performance in PESQ, and Fig. 4 for WER, when evaluated at specific ranges of input SNRs. As expected, both metrics degrade for decreasing input SNRs. However, the proposed audio-visual method shows stronger robustness, indicated by larger relative improvements. This is particularly evident for WER, with drastic degradations for SGMSE+ at negative input SNRs, at which AV-Gen remains fairly robust. Please find some listening examples on our project page[1].

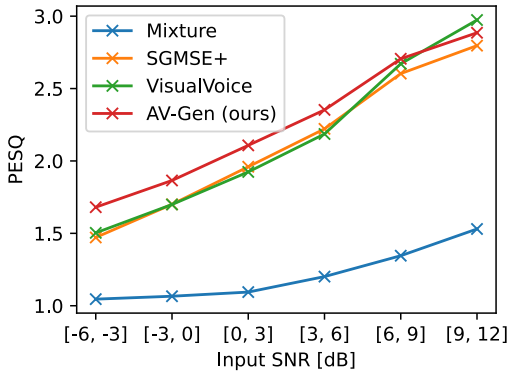


Figure 3: Mean performance in PESQ evaluated for different ranges of input SNRs.

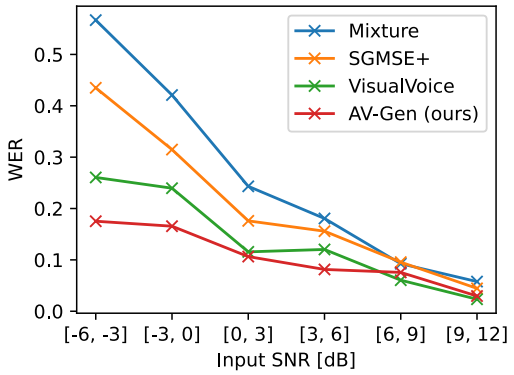


Figure 4: Mean performance in WER evaluated for different ranges of input SNRs

6 Conclusion

In this paper, we presented an AVSE system that utilizes score-based generative models conditioned on visual information. To this end, we leverage audio-visual embeddings obtained from AV-HuBERT, a self-supervised learning model that was fine-tuned on lipreading. The experimental evaluations demonstrated that the proposed method improves speech quality and noticeably reduces generative artifacts, such as phonetic confusions, in comparison to its audio-only counterpart. The effectiveness of the system is further supported by the significant reduction in WER, as observed in a downstream ASR model, particularly for low input SNRs.

[1] https://uhh.de/inf-sp-av-gen

References

[1] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement* (E. Vincent, T. Virtanen, and S. Gannot, eds.), John Wiley & Sons, 2018.

[2] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 29, pp. 1368–1396, 2021.

[3] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. Journal of Automation and Computing*, vol. 18, pp. 351–376, 2021.

[4] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[6] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[7] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *Proc. Interspeech 2018*, pp. 3244–3248, 2018.

[8] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," *ISCA Interspeech*, pp. 2928–2932, 2022.

[9] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 31, pp. 2351–2364, 2023.

[10] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *Int. Conf. on Learning Repr. (ICLR)*, 2022.

[11] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *Int. Conf. on Machine Learning (ICML)*, pp. 2256–2265, 2015.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[13] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 32, 2019.

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. on Learning Repr. (ICLR)*, 2021.

[15] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[16] L. Girin, G. Feng, and J.-L. Schwartz, "Noisy speech enhancement with filters estimated from the speaker's lips.," in *Eurospeech*, 1995.

[17] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 1788–1800, 2020.

[18] A. Mehrish, N. Majumder, R. Bhardwaj, and S. Poria, "A review of deep learning techniques for speech processing," *arXiv preprint arXiv:2305.00359*, 2023.

[19] K. Yang, D. Marković, S. Krenn, V. Agrawal, and A. Richard, "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8227–8237, 2022.

[20] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "ReVISE: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18795–18805, 2023.

[21] R. Mira, B. Xu, J. Donley, A. Kumar, S. Petridis, V. K. Ithapu, and M. Pantic, "LA-VocE: Low-SNR audio-visual speech enhancement using neural vocoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 29, pp. 3451–3460, 2021.

[24] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, IEEE, 2021.

[25] Z. Huang, S. Watanabe, S. Yang, P. García, and S. Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 6837–6841, 2022.

[26] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[27] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 2, pp. 749–752, 2001.

[28] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 504–511, 2015.

[29] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15490–15500, IEEE, 2021.

[30] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6319–6323, 2020.

[31] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7613–7617, 2021.

[32] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *International Conference on Computer Vision*, 2017.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. on Learning Repr. (ICLR)*, 2015.

[34] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018.

[35] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 24, no. 11, pp. 2009–2022, 2016.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 626–630, 2019.

[37] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, *et al.*, "NeMo: a toolkit for building AI applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

4.4 Alternative Diffusion Processes

4.4.1 Publication: Investigating Training Objectives for Generative Speech Enhancement

Abstract

Generative speech enhancement has recently shown promising advancements in improving speech quality in noisy environments. Multiple diffusion-based frameworks exist, each employing distinct training objectives and learning techniques. This paper aims to explain the differences between these frameworks by focusing our investigation on score-based generative models and the Schrödinger bridge. We conduct a series of comprehensive experiments to compare their performance and highlight differing training behaviors. Furthermore, we propose a novel perceptual loss function tailored for the Schrödinger bridge framework, demonstrating enhanced performance and improved perceptual quality of the enhanced speech signals. All experimental code and pre-trained models are publicly available to facilitate further research and development in this domain.

Reference

Julius Richter, Danilo de Oliveira, and Timo Gerkmann, "Investigating Training Objectives for Generative Speech Enhancement", *IEEE Conference on Acoustics, Speech and Signal Processing*, 2025. DOI: 10.1109/ICASSP49660.2025.10887784

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2025 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Julius Richter is the lead author of this paper. He developed the algorithm, trained all models, conducted the experiments, and wrote the manuscript. Danilo de Oliveira and Timo Gerkmann contributed through insights into experimental validation, mathematical derivations, and manuscript review.

Investigating Training Objectives for Generative Speech Enhancement

Julius Richter
Signal Processing Group
University of Hamburg
Hamburg, Germany

Danilo de Oliveira
Signal Processing Group
University of Hamburg
Hamburg, Germany

Timo Gerkmann
Signal Processing Group
University of Hamburg
Hamburg, Germany

***Abstract*—Generative speech enhancement has recently shown promising advancements in improving speech quality in noisy environments. Multiple diffusion-based frameworks exist, each employing distinct training objectives and learning techniques. This paper aims to explain the differences between these frameworks by focusing our investigation on score-based generative models and the Schrödinger bridge. We conduct a series of comprehensive experiments to compare their performance and highlight differing training behaviors. Furthermore, we propose a novel perceptual loss function tailored for the Schrödinger bridge framework, demonstrating enhanced performance and improved perceptual quality of the enhanced speech signals. All experimental code and pre-trained models are publicly available to facilitate further research and development in this domain[1].**

I. Introduction

Diffusion-based generative models have been successfully employed in various audio restoration tasks, most notably in speech enhancement [1]. Generative methods in this task aim at estimating and sampling from the clean speech distribution conditioned on noisy speech. Unlike predictive models, generative models enable the generation of multiple valid estimates for a given input and can be utilized for generalized (or universal) speech enhancement, effectively addressing various corruption types [2].

Numerous diffusion-based generative approaches exist, all centered around the idea of defining a transformation between the data distribution and a tractable prior distribution (e.g., Gaussian). Popular frameworks include continuous-time diffusion models [3], EDM [4], flow matching [5], and the Schrödinger bridge (SB) [6]. Each of these approaches has been applied to speech enhancement.

Score-based generative models for speech enhancement (SGMSE) [7], [8] employs continuous-time diffusion models based on stochastic differential equations (SDEs). Follow-up work utilizes the EDM framework; it has been proposed to use a change of variable to consider the SDE satisfied by the environmental noise [9], which results in the required linear affine drift term. Flow matching has been used in SpeechFlow [10], where the authors apply masked audio prediction as a self-supervised pretraining technique.

We acknowledge the support by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).

[1]https://github.com/sp-uhh/sgmse

More recently, the SB has been proposed for speech enhancement [11]. The SB is a generative model that seeks an optimal way to transport one probability distribution to another distribution [6]. This approach enables starting the generative process directly from the noisy input and allows for using a data prediction loss [12].

This paper builds upon the above-mentioned advances and explores multiple training objectives and learning techniques for generative speech enhancement. We begin by exploring score-based generative models and connecting various loss functions used to learn the score function. For further details, we refer to [13], showing how various diffusion-based generative model objectives can be understood as special cases of a weighted loss. Second, we examine the SB approach for speech enhancement [11] and establish a connection to SGMSE [8]. Moreover, we propose a novel perceptual loss function for the SB framework and perform ablation studies to evaluate its effect. Perceptually motivated loss functions have been extensively integrated into speech enhancement frameworks [14], [15]. However, their use in diffusion-based generative models, particularly within the Schrödinger bridge framework, is a novel approach. We specifically choose the PESQ metric because it is a widely used instrumental measure in speech enhancement for assessing speech quality and, unlike POLQA [16], a differentiable version is available [17].

Our experiments demonstrate that score-based generative models trained with different objective functions exhibit varying training behaviors, although they theoretically model the same underlying concepts. We hypothesize this is due to the neural network's different training tasks. Additionally, we show that our novel perceptual loss for the SB achieves state-of-the-art performance in PESQ on the Voiceband-Demand (VB-DMD) benchmark [18].

Contemporaneously to our work, Wang et al. [19] explore the SB and set a symmetric noise scheduling, where the diffusion shrinks at both boundaries. Furthermore, they combine the SB concept with a two-stage approach inspired by StoRM [20] by aiding the generative model with a magnitude ratio mask.

The paper is structured as follows: Sec. II explores SGMSE and the SB framework. This is followed by describing the experimental setup in Sec. III. Then, we present the results of the experiments in Sec. IV. Finally, in Sec. V, we summarize our key findings and provide an outlook on future work.

II. METHODS

This section discusses two existing generative approaches for speech enhancement and explores their connection. First, we introduce SGMSE [8]. Second, we examine the SB for speech enhancement [11]. Both approaches are diffusion-based stochastic processes aiming to model and manipulate probability distributions. Score-based models emphasize learning score functions, whereas the SB approach can be considered as an optimal transport problem.

A. Score-based generative models for speech enhancement

Following [8], the diffusion forward process is described by the solution to the *forward SDE*

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{C}^d$ is the process state at time $t \in [0, 1]$ and $\mathbf{y} \in \mathbb{C}^d$ is the noisy speech. The diffusion coefficient $g(t) = \sqrt{c}k^t$ controls the Gaussian noise introduced by the Wiener process $\mathbf{w} \in \mathbb{C}^d$. Moreover, γ, c, and k are positive scalar constants that are set as hyperparameters. The forward process is also called Ornstein-Uhlenbeck process with variance exploding (OUVE) SDE [21].

The marginals of the time-reversed forward process can be represented as marginals of another stochastic process (see Theorem A.1 in [22]). This resulting process is described by the solution to the so-called *reverse SDE*

$$\mathrm{d}\mathbf{x}_t = [-\gamma(\mathbf{y}-\mathbf{x}_t)+g(t)^2\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{y})]\mathrm{d}t+g(t)\mathrm{d}\bar{\mathbf{w}}\,, \tag{2}$$

where $\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t|\mathbf{y})$ is the conditional score function, and $\bar{\mathbf{w}}$ is the Wiener process backward in time.

It can be shown that the OUVE SDE results in an interpolation between clean speech $\mathbf{x}_0$ and noisy speech $\mathbf{y}$ with exponentially increasing variance [8]. The evolution of the marginals is described by the time-dependent mean

$$\boldsymbol{\mu}_t(\mathbf{x}_0, \mathbf{y}) = e^{-\gamma t}\mathbf{x}_0 + (1 - e^{-\gamma t})\mathbf{y} \tag{3}$$

and the time-dependent variance

$$\sigma_t^2 = \frac{c(k^{2t} - e^{-2\gamma t})}{2(\gamma + \log k)}\,, \tag{4}$$

that allow for direct sampling of the process state $\mathbf{x}_t$ at time t using the perturbation kernel

$$p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}(\mathbf{x}_t; \boldsymbol{\mu}_t(\mathbf{x}_0, \mathbf{y}), \sigma_t^2\mathbf{I})\,. \tag{5}$$

The score function is typically intractable and approximated by a score model $\mathbf{s}_\theta$ with parameters θ. To train the score model, we use the denoising score-matching objective [23]

$$\mathcal{L}_{\text{DSM}} = \lambda(t)\|\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) - \nabla_{\mathbf{x}_t}\log p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})\|_2^2 \tag{6}$$

where $\lambda(t)$ is a weighting function, and the other variables are sampled according to $t \sim \mathcal{U}[0, 1]$, $(\mathbf{x}_0, \mathbf{y}) \sim p_{\text{data}}(\mathbf{x}_0, \mathbf{y})$ from the dataset, and $\mathbf{x}_t \sim p_{0t}(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})$. The score matching loss is essentially equivalent to a noise prediction loss

$$\mathcal{L}_{\text{score}} = \|\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\,\sigma_t + \mathbf{z}\|_2^2\,, \tag{7}$$

when $\lambda(t) = \sigma_t^2$ in Eq. (6), and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To improve numerical stability, the output of an employed neural network F_θ is often scaled by a factor of $1/\sigma_t$ such that $\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) = F_\theta(\mathbf{x}_t, \mathbf{y}, t)/\sigma_t$.

Following the derivations in [4], it can also be shown that denoising score matching for SGMSE is equivalent to training a denoiser model D_θ with the denoising loss

$$\mathcal{L}_{\text{denoise}} = \lambda(t)\|D_\theta(\mathbf{x}_t, \mathbf{y}, t) - \boldsymbol{\mu}_t(\mathbf{x}_0, \mathbf{y})\|_2^2\,. \tag{8}$$

Furthermore, it was argued that it is beneficial to precondition a neural network F_θ to obtain the denoiser

$$D_\theta(\mathbf{x}_t, \mathbf{y}, t) = c_{\text{skip}}(t)\mathbf{x}_t + c_{\text{out}}(t)F_\theta(c_{\text{in}}(t)\mathbf{x}_t,\, c_{\text{in}}(t)\mathbf{y},\, t) \tag{9}$$

where $c_{\text{skip}}(t)$, $c_{\text{out}}(t)$, $c_{\text{in}}(t)$, and $c_{\text{in}}(t)$ are time-dependent functions that can be derived from first principles (see Appendix B.6 in [4]). Then, the score is then given by

$$\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) = \frac{D_\theta(\mathbf{x}_t, \mathbf{y}, t) - \mathbf{x}_t}{\sigma_t^2}\,. \tag{10}$$

B. Schrödinger bridge for speech enhancement

The SB [24] is defined as the minimization of the Kullback-Leibler divergence D_{KL} between a path measure p and a reference path measure p_{ref}, subject to boundary conditions

$$\min_{p \in \mathcal{P}_{[0,1]}} D_{\text{KL}}\left(p, p_{\text{ref}}\right) \quad \text{s. t.} \quad p_0 = p_x,\ p_1 = p_y, \tag{11}$$

where $\mathcal{P}_{[0,1]}$ is the space of path measures on $[0, 1]$ [6]. An optimal transport solution is given by a pair of symmetric forward and reverse SDEs, with the forward SDE being

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t) + g(t)^2\nabla_{\mathbf{x}_t}\log\Psi_t(\mathbf{x}_t)\right]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t,\ \mathbf{x}_0 \sim p_x, \tag{12}$$

and the reverse SDE being

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t) - g(t)^2\nabla_{\mathbf{x}_t}\log\bar{\Psi}_t(\mathbf{x}_t)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t,\ \mathbf{x}_1 \sim p_y, \tag{13}$$

where the functions $\Psi_t, \bar{\Psi}_t$ are described by coupled partial differential equations (see Theorem 1 in [6])

$$\begin{cases} \frac{\partial\Psi_t}{\partial t} = -\nabla_{\mathbf{x}_t}\Psi_t(\mathbf{x}_t)^{\mathrm{T}}\mathbf{f}(\mathbf{x}_t) - \frac{1}{2}\operatorname{Tr}(g(t)^2\nabla_{\mathbf{x}_t}^2\Psi_t(\mathbf{x}_t)) \\ \frac{\partial\bar{\Psi}_t}{\partial t} = -\nabla_{\mathbf{x}_t}\cdot(\bar{\Psi}_t(\mathbf{x}_t)\mathbf{f}(\mathbf{x}_t)) + \frac{1}{2}\operatorname{Tr}(g(t)^2\nabla_{\mathbf{x}_t}^2\bar{\Psi}_t(\mathbf{x}_t)) \\ \qquad\qquad \textit{s.t. } \Psi_0\bar{\Psi}_0 = p_x, \Psi_1\bar{\Psi}_1 = p_y. \end{cases} \tag{14}$$

However, for a system of symmetric forward and reverse SDEs in Eqs. (12, 13), and arbitrary Ψ_t and $\bar{\Psi}_t$, there are infinitely many solutions bridging the prior to the target [25]. According to Nelson's identity [26]

$$\nabla_{\mathbf{x}_t}\log\Psi_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\log\bar{\Psi}_t(\mathbf{x}_t) = \nabla_{\mathbf{x_t}}\log p_t(\mathbf{x}_t), \tag{15}$$

which is a necessary condition for time-reversal [25], we note that in score-based generative models, this corresponds to setting $\nabla_{\mathbf{x}_t}\log\Psi_t$ to zero. This implies that in score-based generative models, the drift of the forward process, $\mathbf{f}_{\text{SGMSE}}(\mathbf{x}_t) = \gamma(\mathbf{y} - \mathbf{x}_t)$ in Eq. (1), has to be chosen such that the perturbation kernel is known analytically. However, this is not required for the general SB formulation.

Although solving the general SB is typically intractable, closed-form solutions are available for specific cases, such as those involving Gaussian boundary conditions [27]. Assume a drift $\mathbf{f}(\mathbf{x}_t) = f(t)\,\mathbf{x}_t$ and conditional Gaussian boundary conditions $p_0(\mathbf{x}|\mathbf{x}_0) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}; \mathbf{x}_0, \epsilon_0^2\mathbf{I}\right)$ and $p_1(\mathbf{x}|\mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}; \mathbf{y}, \epsilon_1^2\mathbf{I}\right)$ where $\epsilon_1 = \mathrm{e}^{\int_0^1 f(\tau)\mathrm{d}\tau}\epsilon_0$. For $\epsilon_0 \to 0$, the SB solution between clean speech $\mathbf{x}_0$ and noisy speech $\mathbf{y}$ can be expressed as

$$\bar{\Psi}_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}_{\mathbb{C}}(\alpha_t\mathbf{x}_0, \alpha_t^2\sigma_t^2\mathbf{I}),\ \Psi_t(\mathbf{x}_t|\mathbf{y}) = \mathcal{N}_{\mathbb{C}}(\bar{\alpha}_t\mathbf{y}, \alpha_t^2\bar{\sigma}_t^2\mathbf{I}) \tag{16}$$

with parameters $\alpha_t = \mathrm{e}^{\int_0^t f(\tau)\mathrm{d}\tau}$, $\sigma_t^2 = \int_0^t \frac{g^2(\tau)}{\alpha_\tau^2}\mathrm{d}\tau$, $\bar{\alpha}_t = \alpha_t\alpha_1^{-1}$ and $\bar{\sigma}_t^2 = \sigma_1^2 - \sigma_t^2$ [12]. Therefore, the marginal distribution is the Gaussian distribution

$$p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_t; \boldsymbol{\mu}_t(\mathbf{x}_0, \mathbf{y}), \sigma_{\mathbf{x}_t}^2\mathbf{I}\right) \tag{17}$$

whose mean and variance are defined as

$$\boldsymbol{\mu}_t(\mathbf{x}_0, \mathbf{y}) = w_x(t)\mathbf{x}_0 + w_y(t)\mathbf{y}, \quad \sigma_{\mathbf{x}_t}^2 = \frac{\alpha_t^2\bar{\sigma}_t^2\sigma_t^2}{\sigma_1^2}, \tag{18}$$

with $w_x(t) = \alpha_t\bar{\sigma}_t^2/\sigma_1^2$, and $w_y(t) = \bar{\alpha}_t\sigma_t^2/\sigma_1^2$ [12].

In this paper, we adopt the same variance exploding (VE) diffusion coefficient $g(t) = \sqrt{c}k^t$ as used in Eq. (1), and set $f(t) = 0$. This SB configuration has shown strong robustness for both denoising and dereverberation [11]. Consequently, we get $\alpha_t = 1$ and $\sigma_t^2 = c(k^{2t} - 1)/2\log k$. Due to the optimal transport characteristics of the SB, the mean exactly interpolates between the clean speech $\mathbf{x}_0$ at $t = 0$ and the noisy speech y at $t = 1$.

A key advantage of the SB compared to SGMSE is that the neural network F_θ can be trained to directly predict the data $\mathbf{x}_0$ [12]. This is in contrast to SGMSE, where the Gaussian noise $\mathbf{z}$ is predicted as shown in Eq. (7). Using the data prediction loss, it has been proposed to include a time-domain auxiliary loss term based on the ℓ_1-norm [11]. Additionally, we propose incorporating a perceptual loss term such that

$$\mathcal{L}_{\text{SB}} = \|F_\theta(\mathbf{x}_t, \mathbf{y}, t) - \mathbf{x}_0\|_2^2 + \alpha\|\hat{\underline{\mathbf{x}}}_\theta - \underline{\mathbf{x}}_0\|_1 - \alpha_{\text{P}}\,\text{PESQ}(\hat{\underline{\mathbf{x}}}_\theta, \underline{\mathbf{x}}_0)\,, \tag{19}$$

where $\hat{\underline{\mathbf{x}}}_\theta = \text{iSTFT}(F_\theta(\mathbf{x}_t, \mathbf{y}, t))$ and $\underline{\mathbf{x}}_0 = \text{iSTFT}(\mathbf{x}_0)$ represent the corresponding time-domain signals using the inverse short-time Fourier transform (iSTFT). Moreover, $\text{PESQ}(\cdot,\cdot)$ denotes a differentiable version of the PESQ metric[2], and α and α_{P} are hyperparameters to weight the different loss terms.

At inference, the reverse SDE in Eq. (13) can be solved with an ordinary differential equation (ODE) sampler or an SDE sampler [12]. Here, we make use of the ODE sampler because it has shown better performance for the speech enhancement task [11]. For a given discretization schedule $(t_N = 1, t_{N-1}, \ldots, t_0 = 0)$ with N steps, the ODE sampler is recursively defined as

$$\mathbf{x}_{t_{n-1}} = a_n\mathbf{x}_{t_n} + b_nF_\theta(\mathbf{x}_{t_n}, \mathbf{y}, t_n) + c_n\mathbf{y}, \quad \mathbf{x}_{t_N} = \mathbf{y}\,, \tag{20}$$

[2]https://github.com/audiolabs/torch-pesq

Model	# params	GMACs	proc/s [s]
Conv-TasNet [28]	8.7 M	28	0.015
MetricGAN+ [29]	1.9 M	106	0.016
SGMSE+ [8]	65.6 M	15,995	1.155
SEMamba [30]	2.3 M	131	0.075

TABLE I: Baseline methods. Number of parameters, GMACs for an input of 4 seconds, and average processing time.

$$a_n = \frac{\alpha_{t_{n-1}}\sigma_{t_{n-1}}\bar{\sigma}_{t_{n-1}}}{\alpha_{t_n}\sigma_{t_n}\bar{\sigma}_{t_n}}, \tag{21}$$

$$b_n = \frac{\alpha_{t_{n-1}}}{\sigma_1^2}\left(\bar{\sigma}_{t_{n-1}}^2 - \frac{\bar{\sigma}_{t_n}\sigma_{t_{n-1}}\bar{\sigma}_{t_{n-1}}}{\sigma_{t_n}}\right), \tag{22}$$

$$c_n = \frac{\alpha_{t_{n-1}}}{\alpha_1\sigma_1^2}\left(\sigma_{t_{n-1}}^2 - \frac{\sigma_{t_n}\sigma_{t_{n-1}}\bar{\sigma}_{t_{n-1}}}{\bar{\sigma}_{t_n}}\right). \tag{23}$$

III. Experimental Setup

A. Models settings

We train eight models (M1-M8), each utilizing complex spectrograms as the input representation by computing the short-time Fourier transform (STFT) with a periodic Hann window of size of 510 and a hop length of 128. We use the identical amplitude compression as in [8]. We train the models with a batch size of 16 using two NVIDIA RTX A6000 graphics processing units (GPUs) with 48 GB memory.

Models M1-M4 employ the OUVE SDE, utilizing the recommended hyperparameters from [8]. The models vary in the loss type and the preconditioning, as shown in Table II.

Models M5-M8 use the Schrödinger bridge with variance exploding diffusion coefficient (SB-VE) with the recommended hyperparameters from [11]. The models differ in the hyperparameter α_{P}, as indicated in Table II.

B. Network architecture

In all experiments except for model M4, we employ the NCSN++ architecture [3] using the same parameterization described in [8].

For M4, we employ the EDM2 network architecture [31]. The core idea in EDM2 is to restructure the network layers to ensure that the expected magnitudes of activations, weights, and updates maintain unit variance. Additionally, all additive biases are removed, and an extra channel of constant one is concatenated to the network's input instead. We use the same number of layers and channels as for the NCSN++. Furthermore, the authors propose to use a power function exponential moving average (EMA) that automatically scales according to training time and has zero contribution at the initial training step.

C. Metrics

As intrusive speech enhancement metrics, we include POLQA [16] and PESQ [33] for predicting speech quality. Moreover, we employ ESTOI [34] as an instrumental measure of speech intelligibility and calculate the scale invariant signal-to-distortion ratio (SI-SDR) [35] measured in dB. As a non-intrusive metric, we use DNSMOS [36], which employs a

Model	SDE	Loss	α_P	Precon	POLQA	PESQ	SI-SDR	ESTOI	DNSMOS
Noisy					3.11 ± 0.79	1.97 ± 0.75	8.4 ± 5.6	0.79 ± 0.15	3.09 ± 0.39
Conv-TasNet+ [28]	-	-	-	-	3.56 ± 0.57	2.63 ± 0.60	19.1 ± 3.5	0.85 ± 0.10	3.37 ± 0.32
MetricGAN+ [29]	-	-	-	-	3.72 ± 0.68	3.13 ± 0.55	8.5 ± 3.8	0.83 ± 0.11	3.37 ± 0.30
SEMamba [30]	-	-	-	-	$\mathbf{4.33 \pm 0.40}$	3.56 ± 0.60	$\mathbf{19.7 \pm 3.2}$	$\mathbf{0.89 \pm 0.08}$	3.58 ± 0.29
PESQetarian [32]	-	-	-	-	1.46 ± 0.48	$\mathbf{3.82 \pm 0.57}$	-19.8 ± 3.3	0.84 ± 0.09	2.39 ± 0.22
SGMSE+ [8]	OUVE	score	-	✗	3.95 ± 0.52	2.93 ± 0.62	17.3 ± 3.3	0.87 ± 0.10	3.56 ± 0.28
M1	OUVE	score	-	✗	3.93 ± 0.51	2.84 ± 0.61	17.7 ± 3.6	0.86 ± 0.10	3.54 ± 0.28
M2	OUVE	denoise	-	✗	3.96 ± 0.53	2.90 ± 0.67	18.0 ± 3.3	0.86 ± 0.10	3.55 ± 0.28
M3	OUVE	denoise	-	✓	3.86 ± 0.50	2.77 ± 0.59	17.8 ± 3.2	0.86 ± 0.10	3.51 ± 0.27
M4 (EDM2)	OUVE	denoise	-	✓	3.87 ± 0.54	2.87 ± 0.65	18.0 ± 3.2	0.86 ± 0.10	3.54 ± 0.27
M5	SB-VE	predict	0	✗	4.15 ± 0.54	2.91 ± 0.76	19.4 ± 3.5	0.88 ± 0.10	$\mathbf{3.59 \pm 0.30}$
M6	SB-VE	predict	1e-3	✗	4.15 ± 0.53	3.70 ± 0.58	8.3 ± 2.8	0.86 ± 0.09	3.44 ± 0.34
M7	SB-VE	predict	5e-4	✗	4.25 ± 0.50	3.50 ± 0.66	14.1 ± 2.9	0.87 ± 0.09	3.55 ± 0.29
M8	SB-VE	predict	2.5e-4	✗	4.20 ± 0.51	3.44 ± 0.73	15.3 ± 2.8	0.87 ± 0.09	3.58 ± 0.29

TABLE II: Speech enhancement performance on VB-DMD. Values indicate mean and standard deviation.

neural network trained on human ratings. For all metrics it holds, the higher, the better.

D. Baselines and Data

Table I shows all baseline methods, the number of parameters, multiply–accumulate operations (MACs) for an input of 4 seconds, and the processing time per input second on a GPU. We use the provided checkpoints and the official implementations. As a dataset, we use the standardized VB-DMD [18], commonly employed as a benchmark for speech enhancement.

IV. Results

In Table II, we present the results for the speech enhancement task using the VB-DMD dataset. We begin by comparing different training objectives for the OUVE SDE. Model M1 employs the standard SGMSE training objective, replicating the results reported in the original paper [8]. Minor discrepancies in these outcomes may be attributed to a different batch size. Model M2 utilizes the denoising loss in Eq. (8) and achieves slightly higher scores than M1. However, its training process is more unstable, as depicted by the orange lines in Fig. 1. In contrast, the preconditioned model M3 displays faster training times but produces results comparable to those of M1. Model M4 explores the EDM2 network, exhibiting performance that is competitive with M1. Yet, similar to M2, it experiences instability during training, indicated by the red lines in Fig. 1. It is worth noting that the results for M4 are preliminary, as we have not experimented with various post-hoc EMA configurations and fine-tuned learning rate schedulers. Among models M1-M4, differences in ESTOI and DNSMOS are negligible.

Turning the attention to the SB approach, model M5 shows strong performance in SI-SDR; an expected result due to the ℓ_1-loss applied in the time domain. Moreover, M5 demonstrates relatively stable training behavior, as illustrated by the purple lines in Fig. 1, and achieves the best score for DNSMOS among all methods. Model M6 introduces the PESQ loss and achieves state-of-the-art results in PESQ with a score of 3.70. However, similar to the extreme behavior of the PESQetarian [32], this improvement comes at the expense of SI-SDR. Models M7 and M8 investigate different weightings

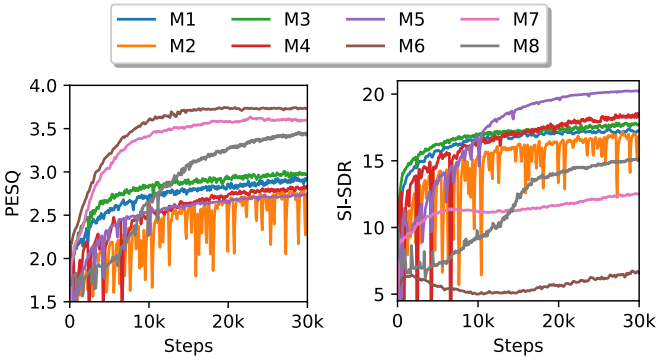


Fig. 1: PESQ and SI-SDR performance over the training steps.

of the PESQ term, achieving a better balance between SI-SDR and PESQ, while also obtaining higher ESTOI scores compared to M5 and M6. Specifically, M7 performs best in POLQA among all models based on the SB.

Lastly, we examine a comparison with the given baseline methods. With the SB approach, we surpass the performance of most baselines, including SGMSE while remaining on par with SEMamba [30]. For results obtained with 48 kHz speech data using the EARS-WHAM dataset [37], we refer to [38]. Audio files are available online[3]

V. Conclusion

This paper explored the distinctions among various diffusion-based frameworks for generative speech enhancement, explicitly focusing on score-based generative models and the Schrödinger bridge (SB). Through comprehensive experimental analysis, we highlighted the variations in training behaviors and performance across these frameworks. We proposed adding ad-hoc cost functions to the SB framework, significantly enhancing the performance and perceptual quality of the processed speech signals. To support ongoing research and innovation in this field, we made all experimental code and pre-trained models publicly accessible.

[3]https://sp-uhh.github.io/gen-se/

REFERENCES

[1] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration: A review," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2025.

[2] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open Journal of Signal Processing*, 2024.

[3] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *International Conference on Learning Representations*, 2021.

[4] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, 2022.

[5] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations*, 2023.

[6] T. Chen, G.-H. Liu, and E. Theodorou, "Likelihood training of Schrödinger bridge using forward-backward SDEs theory," in *International Conference on Learning Representations*, 2021.

[7] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proceedings of Interspeech*, 2022, pp. 2928–2932.

[8] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[9] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Investigating the design space of diffusion models for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 4486–4500, 2024.

[10] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching," in *International Conference on Learning Representations*, 2024.

[11] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement," in *Proceedings of Interspeech*, 2024, pp. 1175–1179.

[12] Z. Chen, G. He, K. Zheng, X. Tan, and J. Zhu, "Schrödinger bridges beat diffusion models on text-to-speech synthesis," *arXiv preprint arXiv:2312.03491*, 2023.

[13] D. P. Kingma and R. Gao, "Understanding diffusion objectives as the ELBO with simple data augmentation," in *Conference on Neural Information Processing Systems*, 2023.

[14] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[15] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*, 2019, pp. 2031–2041.

[16] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[17] J. Kim, M. El-Khamy, and J. Lee, "End-to-end multi-task denoising for joint SDR and PESQ optimization," *arXiv preprint arXiv:1901.09146*, 2019.

[18] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *ISCA Speech Synthesis Workshop*, pp. 146–152, 2016.

[19] S. Wang, S. Liu, A. Harper, P. Kendrick, M. Salzmann, and M. Cernak, "Diffusion-based speech enhancement with Schrödinger bridge and symmetric noise schedule," *arXiv preprint arXiv:2409.05116*, 2024.

[20] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2724–2737, 2023.

[21] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," in *Proceedings of Interspeech*, 2023, pp. 3809–3813.

[22] J. Berner, L. Richter, and K. Ullrich, "An optimal control perspective on diffusion-based generative modeling," *Transactions on Machine Learning Research*, 2024.

[23] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[24] E. Schrödinger, "Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique," in *Annales de l'institut Henri Poincaré*, vol. 2, no. 4, 1932, pp. 269–310.

[25] L. Richter and J. Berner, "Improved sampling via learned diffusions," in *International Conference on Learning Representations*, 2024.

[26] E. Nelson, *Dynamical Theories of Brownian Motion*. Princeton University Press, 1967, vol. 3.

[27] C. Bunne, Y.-P. Hsieh, M. Cuturi, and A. Krause, "The Schrödinger bridge between gaussian measures has a closed form," in *International Conference on Artificial Intelligence and Statistics*, 2023, pp. 5802–5833.

[28] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[29] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proceedings of Interspeech*, 2021, pp. 201–205.

[30] R. Chao, W.-H. Cheng, M. La Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating Mamba for speech enhancement," *IEEE Spoken Language Technology Workshop*, 2024.

[31] T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine, "Analyzing and improving the training dynamics of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 174–24 184.

[32] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The PESQetarian: On the relevance of Goodhart's law for speech enhancement," in *Proceedings of Interspeech*, 2024, pp. 3854–3858.

[33] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001, pp. 749–752.

[34] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[35] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 626–630.

[36] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6493–6497, 2021.

[37] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proceedings of Interspeech*, 2024.

[38] J. Richter and T. Gerkmann, "Diffusion-based speech enhancement: Demonstration of performance and generalization," *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

5 Analysis and Comparison

In this chapter, we presents a comprehensive analysis of generative speech enhancement methods, highlighting their differences from predictive approaches.

Section 5.1.1 introduces the EARS dataset, a newly curated 48 kHz speech corpus that captures a broad range of speaking styles, encompassing emotional speech, various reading styles, non-verbal vocalizations, and conversational freeform speech. To enable rigorous evaluation of speech enhancement methods, the EARS-WHAM benchmark is established. Several predictive and generative speech enhancement models are systematically compared using both objective instrumental metrics and human listening tests. Notably, our generative model *SGMSE+* demonstrates superior performance, earning the highest scores across most instrumental evaluations and being the top choice among listeners in testing test.

5.1 Analysis of Generative Speech Enhancement and Comparison to Predictive Approaches

5.1.1 Publication: EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation

Abstract

We release the EARS (Expressive Anechoic Recordings of Speech) dataset, a high-quality speech dataset comprising 107 speakers from diverse backgrounds, totaling in 100 hours of clean, anechoic speech data. The dataset covers a large range of different speaking styles, including emotional speech, different reading styles, non-verbal sounds, and conversational freeform speech. We benchmark various methods for speech enhancement and dereverberation on the dataset and evaluate their performance through a set of instrumental metrics. In addition, we conduct a listening test with 20 participants for the speech enhancement task, where a generative method is preferred. We introduce a blind test set that allows for automatic online evaluation of uploaded data. Dataset download links and automatic evaluation server can be found online.

Reference

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, Timo Gerkmann, "EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation," *Proceedings of Interspeech*, Kos, Greece, Sep. 2024, pp. 4873–4877. DOI: 10.21437/Interspeech.2024-153

Copyright Notice

The following article is the accepted version of the article published with ISCA.
© 2024 ISCA. Reprinted, with permission, from the reference displayed above.

Author's Contributions

This paper is a collaboration among the Signal Processing Group at the University of Hamburg, the Codec Avatars Lab at Meta, and the Language Technologies Institute at Carnegie Mellon University. Each lab handled distinct tasks.

At the University of Hamburg, the Signal Processing Group was tasked with setting up the benchmark and adapting all baseline models for 48 kHz. Julius Richter led the efforts for the University of Hamburg, serving as the main author of this paper. He was responsible for designing the benchmark, adapting the baseline models for 48 kHz, conducting experiments, and writing the manuscript. Simon Welker managed the listening experiment and contributed to the manuscript review. Bunlong Lay assisted in training the baseline models and also reviewed the manuscript. Timo Gerkmann provided insights into the experimental validation and participated in the manuscript review.

In the Codec Avatars Lab at Meta, the team was responsible for recording the speech data. Alexander Richard took the lead on the Meta side, initiated the collaboration, managed the project, oversaw data collection, and co-wrote the manuscript. Yi-Chiao Wu shared insights into the benchmark design and conducted a listening experiment at Meta. Steven Krenn was responsible for setting up the recording and managing data collection.

At the Language Technologies Institute, Shinji Watanabe organized the evaluation server and contributed insights into the experimental validation and manuscript review.

EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation

Julius Richter[1], *Yi-Chiao Wu*[2], *Steven Krenn*[2], *Simon Welker*[1], *Bunlong Lay*[1], *Shinji Watanabe*[3], *Alexander Richard*[2], *Timo Gerkmann*[1]

[1]Signal Processing (SP), Universität Hamburg, Germany
[2]Codec Avatars Lab, Meta, Pittsburgh, PA, USA
[3]Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA
julius.richter@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

We release the EARS (**E**xpressive **A**nechoic **R**ecordings of **S**peech) dataset, a high-quality speech dataset comprising 107 speakers from diverse backgrounds, totaling in 100 hours of clean, anechoic speech data. The dataset covers a large range of different speaking styles, including emotional speech, different reading styles, non-verbal sounds, and conversational freeform speech. We benchmark various methods for speech enhancement and dereverberation on the dataset and evaluate their performance through a set of instrumental metrics. In addition, we conduct a listening test with 20 participants for the speech enhancement task, where a generative method is preferred. We introduce a blind test set that allows for automatic online evaluation of uploaded data. Dataset download links and automatic evaluation server can be found online[1].

Index Terms: speech dataset, speech enhancement, dereverberation, benchmark

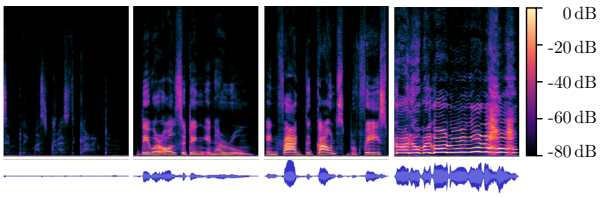
1. Introduction

Learning-based speech processing has seen huge leaps forward in recent years, with the impact of deep learning spanning essentially all areas from speech representation learning [1] over text-to-speech [2] to speech enhancement [3]. Publicly available datasets such as LibriSpeech [4] or VCTK [5] have undoubtedly been a key driver of open and reproducible research in our field and have enabled steady progress. However, these datasets typically come with multiple shortcomings and are either too small, of low recording quality or do not span a large enough variety of different speakers and speaking styles.

To overcome these shortcomings, we release the **E**xpressive **A**nechoic **R**ecordings of **S**peech (EARS) dataset. EARS contains 100 h of anechoic speech recordings at 48 kHz from over 100 English speakers with high demographic diversity. The dataset spans the full range of human speech, including reading tasks in seven different reading styles, emotional reading and freeform speech in 22 different emotions, conversational speech, and non-verbal sounds like laughter or coughing.

In addition, we set up a speech enhancement and speech dereverberation benchmark on EARS, comparing several predictive [6, 7] and generative [8, 9] speech enhancement methods. The benchmarks are intended to provide valuable insights into models' strengths, limitations, and comparability, thus promoting the development of robust and efficient speech enhancement systems.

[1]https://sp-uhh.github.io/ears_dataset/



(a) *whisper* (b) *regular* (c) *loud* (d) *yelling*

Figure 1: ***High Dynamic Range.*** *The EARS dataset spans the complete dynamic range of human speech, from whispering to yelling and screaming.*

2. EARS dataset

A good speech dataset is characterized by its scale, diversity, and high recording quality. However, most existing datasets fall short in one or more of these characteristics; see Table 1. Most notably, a dataset that is of high recording quality (clean 48 kHz audio), has a sufficient scale and covers the full range of human speech as opposed to only reading or neutral speech does not exist to the best of our knowledge. Yet, such a dataset is strongly required to advance research ranging from speech synthesis over voice and style conversion to speech enhancement.

We overcome these limitations with the EARS dataset, which provides high speaker and speech diversity paired with the highest recording quality.

High Recording Quality. All speech is recorded in an anechoic chamber as 32-bit audio at 48 kHz. We simultaneously record with a low-noise GRAS 40HH and a GRAS 48BL microphone, which are both mounted about 1 m in front of the participant. The first microphone has low self-noise and high sensitivity to capture subtle and nuanced speech signals, while the second has lower sensitivity to capture high-energy speech like yelling without clipping, allowing us to capture the full dynamic range of human speech, see Figure 1. We use the high-sensitivity recording for our dataset whenever possible. In the few (5% of the dataset) cases, like yelling, where the high-sensitivity microphone clips, we replace it with the lower-sensitivity microphone. To maintain the same audio characteristics between both microphones, we measure the transfer function between them using a sine-sweep and deconvolution and equalize the low-sensitivity microphone accordingly. See the project page for examples[1].

High Speaker Diversity. We recorded 107 speakers from diverse demographic backgrounds, each for close to one hour, resulting in a dataset with 100 h of clean speech. Our speakers

	hours	speakers	sample rate	anechoic	reading styles	freeform speech	emotional speech	speaker metadata
DNS (LibriVox) [10]	556	1948	48 kHz†	✗	n/a	✗	✗	✗
MSP-Conversation (v1.0) [11]	14	197	16 kHz	✗	n/a	✓	✓	✗
MSP-Podcast (v1.10) [12]	166	1458	16 kHz	✗	n/a	✓	✓	✗
LibriSpeech [4]	982	2484	16 kHz	✗	n/a	✗	✗	✗
LJSpeech [13]	24	1	22.05 kHz	✗	neutral	✗	✗	✗
TIMIT [14]	5	632	16 kHz	✗	neutral	✗	✗	✓
VCTK [5]	44	110	48 kHz	✗	neutral	✗	✗	✗
WSJ0 [15]	29	119	16 kHz	✗	neutral	✗	✗	✗
EARS (ours)	100	107	48 kHz	✓	7 styles	✓	22 emotions	✓

Table 1: ***Speech datasets.*** *In contrast to existing datasets, the EARS dataset is of higher recording quality, large, and more diverse. Reading tasks feature seven styles (regular, loud, whisper, fast, slow, high pitch, and low pitch). Additionally, the dataset features unconstrained freeform speech and speech in 22 different emotional styles. We provide transcriptions of the reading portion and metadata of the speakers (gender, age, race, first language).* †*contains files with limited bandwidth*

	# files	rooms
ACE-Challenge [16]	84	building lobby, lecture room, meeting room, office
AIR [17]	344	auditorium, corridor, lecture room, meeting room, stairway
ARNI [18]	1000	variable acoustics laboratory
BRUDEX [19]	144	variable acoustics laboratory
dEchorate [20]	648	variable acoustics laboratory
DetmoldSRIR [21]	49	concert hall, music chamber, theater
Palimpsest [22]	44	air raid shelter, dockyard, submarine

Table 2: ***RIR datasets***. *To construct EARS-Reverb, we use 2313 RIR files with different room characteristics.*

	pub. date	# params	GMACs	proc/s [s]
Conv-TasNet [6]	May 2019	8.7 M	28	0.015
CDiffuSE [8]	May 2022	18.1 M	18,382	42.268
Demucs [7]	June 2023	83.6 M	60	0.027
SGMSE+ [9]	June 2023	64.8 M	47,984	2.575

Table 3: ***Baseline methods***. *Date of publication, number of parameters, MACs for an input of four seconds, and average processing time per one-second input length.*

range from age 18 to 75 and span various ethnicities, including African American, Caucasian, Hispanic, and Asian. Participants are 44% male, 53% female, and 3% non-binary.

High Content Diversity. Each speaker follows a script that covers a wide variety of speech styles. The script contains a large portion of phonetically balanced sentence reading in seven different styles (regular, loud, whisper, fast, slow, low pitch, and high pitch). Additionally, it contains 18 minutes of conversational freeform speech, where participants freely reply to open-ended questions asked by an operator or talk about vacations, hobbies, or professions. To cover the wide range of emotional speech, we ask participants to read three sentences and describe an image with a specific emotional tone for each of 22 different emotions, including base emotions like ecstasy, fear, anger, or sadness, and nuanced emotions like serenity or adoration. To cover the full variety of human sounds, we additionally include short sections with non-speech sounds like laughter, yelling, or crying, vegetative sounds like coughing or yawning, interjection words, and melodic sounds. A trained operator monitors the participant during the recordings and ensures that speaking styles and prompts are followed as requested and re-record faulty segments.

3. Benchmarks

The EARS dataset enables various speech processing tasks to be evaluated in a controlled and comparable way. Here, we present benchmarks for speech enhancement and dereverberation tasks. According to the typical convention, we divide the data into training, validation, and test splits. We select participants *p001* to *p099* for training, *p100* and *p101* as validation speakers, and *p102* to *p107* as test speakers. We use all speech files except utterances containing interjection, melodic, nonverbal, or vegetative sounds. We cut longer files in the validation and training splits every 10 s to be at least 4 s long. For the test set, we provide cutting times and exclude files that are longer than 29 s. This results in 32,485 files (86.8 h) for training, 632 files (1.7 h) for validation, and 886 files (3.7 h) for the test. Data generation scripts can be found online[1].

3.1. EARS-WHAM

For the speech enhancement task, we construct the EARS-WHAM dataset, which mixes speech from the EARS dataset with real noise recordings from the WHAM! dataset [23] (CC BY-NC 4.0 license). We mix speech and noise files at signal-to-noise ratios (SNRs) randomly sampled in a range of $[-2.5, 17.5]$ dB, where we compute the SNR using loudness K-weighted relative to full scale (LKFS) standardized in ITU-R BS.1770 [24] to obtain a more perceptually meaningful scaling and also to remove silent regions from the SNR computation [25]. We additionally create a **blind test set** for which we only publish the noisy audio files but not the clean ground truth. It contains 743 files (2 h) from six speakers (3 male, 3 female) that are not part of the EARS dataset and noise especially recorded for this test set. We set up an evaluation server for blind evaluation on this test set, which can be found online[1].

3.2. EARS-Reverb

For the task of dereverberation, we use real recorded room impulse responses (RIRs) from multiple public datasets [16, 17, 18, 19, 20, 21, 22] (CC BY 4.0, MIT license). Table 2 shows statistics on the RIR datasets used. All RIRs are fullband, and we use a randomly selected channel for multi-channel recordings. We generate reverberant speech by convolving the clean speech with the RIR. To avoid a time delay between the reverberant and clean speech signal caused by the direct path of the RIR, we cut off the beginning of the RIR up to the index with the highest amplitude. We only use RIRs with an RT_{60} rever-

	POLQA (14 kHz)	**SI-SDR** [dB] (24 kHz)	**PESQ** (7 kHz)	**ESTOI** (5 kHz)	**SIGMOS** (24 kHz)	**DNSMOS** (8 kHz)	**WER** [%] (8 kHz)
Noisy	1.71 ± 0.56	5.98 ± 6.10	1.24 ± 0.22	0.49 ± 0.15	1.95 ± 0.39	2.74 ± 0.29	33 ± 29
Conv-TasNet [6]	2.73 ± 0.78	$\mathbf{16.93 \pm 4.36}$	2.31 ± 0.59	0.70 ± 0.14	2.69 ± 0.42	3.47 ± 0.31	20 ± 20
CDiffuSE [8]	1.81 ± 0.50	8.35 ± 3.13	1.60 ± 0.40	0.53 ± 0.15	2.08 ± 0.31	2.87 ± 0.26	32 ± 27
Demucs [7]	2.97 ± 0.75	16.92 ± 4.35	2.37 ± 0.58	0.71 ± 0.14	2.87 ± 0.43	3.66 ± 0.30	17 ± 18
SGMSE+ [9]	$\mathbf{3.40 \pm 0.73}$	16.78 ± 4.47	$\mathbf{2.50 \pm 0.62}$	$\mathbf{0.73 \pm 0.13}$	$\mathbf{3.41 \pm 0.41}$	$\mathbf{3.88 \pm 0.26}$	$\mathbf{16 \pm 18}$

Table 4: ***Results on EARS-WHAM****. Column groups are the method name, intrusive metrics, non-intrusive metrics, and WER. Below each metric is the maximum frequency taken into account for the assessment. Values indicate mean and standard deviation.*

	POLQA	**SI-SDR** [dB]	**PESQ**	**ESTOI**	**SIGMOS**	**DNSMOS**	**WER** [%]
Noisy	1.81 ± 0.60	6.48 ± 6.76	1.28 ± 0.32	0.57 ± 0.18	1.97 ± 0.44	2.79 ± 0.37	28 ± 25
Conv-TasNet [6]	2.68 ± 0.75	16.56 ± 5.80	2.41 ± 0.63	0.75 ± 0.14	2.70 ± 0.38	3.43 ± 0.35	23 ± 22
CDiffuSE [8]	1.93 ± 0.61	8.22 ± 3.97	1.64 ± 0.46	0.59 ± 0.17	2.09 ± 0.34	2.92 ± 0.29	31 ± 25
Demucs [7]	3.03 ± 0.79	$\mathbf{16.81 \pm 5.94}$	2.50 ± 0.63	0.76 ± 0.14	2.82 ± 0.43	3.62 ± 0.34	19 ± 20
SGMSE+ [9]	$\mathbf{3.35 \pm 0.82}$	16.43 ± 6.12	$\mathbf{2.59 \pm 0.70}$	$\mathbf{0.78 \pm 0.13}$	$\mathbf{3.30 \pm 0.40}$	$\mathbf{3.79 \pm 0.30}$	$\mathbf{19 \pm 19}$

Table 5: ***Results for the blind test****. Column groups are the method name, intrusive metrics, non-intrusive metrics, and WER. Values indicate mean and standard deviation.*

beration time that does not exceed 2 s. Finally, we normalize the loudness of the reverberant speech to the loudness of the clean speech using LKFS.

4. Baselines and Evaluation

4.1. Baselines

Table 3 shows all baseline methods with the date of publication, number of parameters, multiply–accumulate operations (MACs) for an input of 4 s using the ptflops package[1], and the processing time per input second. We calculate the processing time per second averaged over 20 utterances from the test set using an NVIDIA RTX A6000 graphics processing unit (GPU).

Conv-TasNet [6] is a predictive method initially proposed for speech separation that operates in the time domain. Identical to the original approach, we learn 2 ms filters, which correspond to kernels of size 120 and stride of 60 at a sampling rate of 48 kHz. We train with a batch size of 4 using one GPU.

CDiffuSE [8] is a generative speech enhancement method based on a conditional diffusion process defined in the time domain. We adapt the method for 48 kHz by using a 3072-point short-time Fourier transform (STFT), resulting in 1537 frequency bins for the conditioner. We train the large model with a batch size of 16 using two GPUs.

Demucs v4 [7] is a predictive model originally proposed for music separation. We train with batch size 8 using one GPU.

SGMSE+ [9] is a generative speech enhancement method based on a conditional diffusion process defined in the complex STFT domain. We adapt the method for 48 kHz by using 1534-point STFT with hop size 384. We use $\alpha = 0.667$ and $\beta = 0.065$ for the STFT amplitude compression and $\sigma_{\text{min}} = 0.1$, $\sigma_{\text{max}} = 1$, and $\gamma = 2$ for the stochastic differential equation. We train with a batch size of 4 using four GPUs.

4.2. Metrics

We employ *intrusive* metrics that rate the processed signal in relation to the clean reference signal and *non-intrusive* metrics, which assess the performance only using the processed signal.

Intrusive metrics include the perceptual objective listening quality analysis (POLQA) [26] for predicting speech quality, which takes values from 1 (poor) to 5 (excellent) as usual for mean opinion scores (MOS). We report the perceptual evaluation of speech quality (PESQ) [27], which is the predecessor of POLQA and is still widely used in the research community. The PESQ score lies between 1 (poor) and 4.5 (excellent). We further use extended short-time objective intelligibility (ESTOI) [28] as an intrusive measure of speech intelligibility. This metric yields values between 0 and 1, with higher values indicating better intelligibility. Moreover, we calculate the scale-invariant signal-to-distortion ratio (SI-SDR) [29] measured in dB, with higher values indicating better performance.

Non-intrusive metrics include the SIGMOS estimator [30], which is a speech quality assessment model based on a multi-dimensional listening test [31]. We report the overall quality (SIGMOS) and the reverberation assessment (MOS Reverb, only in Table 7). In addition, we use the speech quality assessment model DNSMOS [32] that is trained on human ratings obtained from listening experiments based on ITU-T P.808 [33].

To evaluate the effect of speech enhancement on automatic speech recognition (ASR), we use QuartzNet15x5Base-En from the NeMo toolkit [34] as a downstream ASR system and report the word error rate (WER). We obtain the reference transcriptions by performing ASR on the clean speech utterances.

4.3. Evaluation

We provide an empirical evaluation of the speech enhancement and dereverberation benchmarks. Listening examples for both tasks can be found online[1].

Speech Enhancement. In Table 4 and Table 5, we show speech enhancement results on the EARS-WHAM test set and the blind test set, respectively. Among the methods, the generative SGMSE+ [9] performs the best across most metrics, with particularly high scores in POLQA and SIGMOS. Demucs [7], as a representative of predictive methods, convinces with strong results, too, although falling slightly behind SGMSE+.

Listening Test. We conduct a MUSHRA-like (*Multiple Stimuli with Hidden Reference and Anchor*) listening test on EARS-WHAM with 20 participants. We randomly sample 10 distinct utterances from the test set in a gender-balanced way (5 male, 5 female). We use the clean audio as the hidden refer-

[1] `https://pypi.org/project/ptflops/`

	(a) POLQA				(b) SI-SDR [dB]			
	0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Noisy	1.2 ± 0.2	1.4 ± 0.3	1.9 ± 0.4	2.4 ± 0.4	-1.6 ± 2.4	3.5 ± 2.2	8.6 ± 2.6	13.5 ± 2.2
Conv-TasNet [6]	1.9 ± 0.5	2.5 ± 0.5	3.1 ± 0.5	3.5 ± 0.5	11.7 ± 2.6	15.2 ± 2.0	$\mathbf{19.0 \pm 2.0}$	$\mathbf{21.8 \pm 1.7}$
CDiffuSE [8]	1.3 ± 0.2	1.6 ± 0.3	2.0 ± 0.4	2.3 ± 0.5	4.5 ± 1.9	8.1 ± 1.9	10.1 ± 2.0	10.7 ± 2.1
Demucs [7]	2.1 ± 0.5	2.7 ± 0.5	3.4 ± 0.4	3.7 ± 0.3	$\mathbf{11.9 \pm 2.6}$	$\mathbf{15.3 \pm 2.3}$	18.9 ± 2.3	21.7 ± 2.1
SGMSE+ [9]	$\mathbf{2.6 \pm 0.6}$	$\mathbf{3.3 \pm 0.6}$	$\mathbf{3.8 \pm 0.3}$	$\mathbf{4.0 \pm 0.3}$	11.6 ± 2.7	15.1 ± 2.2	18.8 ± 2.2	21.8 ± 2.0

	(c) ESTOI				(d) WER [%]			
	0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Noisy	0.32 ± 0.08	0.44 ± 0.09	0.56 ± 0.10	0.65 ± 0.11	63 ± 25	39 ± 24	18 ± 17	12 ± 16
Conv-TasNet [6]	0.58 ± 0.14	0.67 ± 0.12	0.75 ± 0.11	0.79 ± 0.10	39 ± 22	21 ± 18	11 ± 12	8 ± 12
CDiffuSE [8]	0.37 ± 0.10	0.50 ± 0.11	0.60 ± 0.11	0.65 ± 0.11	61 ± 24	36 ± 22	18 ± 16	14 ± 18
Demucs [7]	0.60 ± 0.13	0.69 ± 0.11	0.76 ± 0.10	0.80 ± 0.10	32 ± 21	$\mathbf{17 \pm 16}$	9 ± 11	7 ± 11
SGMSE+ [9]	$\mathbf{0.63 \pm 0.13}$	$\mathbf{0.72 \pm 0.11}$	$\mathbf{0.78 \pm 0.10}$	$\mathbf{0.81 \pm 0.10}$	$\mathbf{30 \pm 21}$	$\mathbf{17 \pm 16}$	$\mathbf{8 \pm 9}$	$\mathbf{6 \pm 10}$

Table 6: ***Results per input SNR***. *Mean and standard deviation for (a) POLQA, (b) SI-SDR, (c) ESTOI, and (d) WER on EARS-WHAM.*

	POLQA	**SI-SDR** [dB]	**PESQ**	**ESTOI**	**SIGMOS**	**MOS Reverb**	**WER** [%]
Reverberant	1.75 ± 0.48	-16.17 ± 9.77	1.48 ± 0.37	0.52 ± 0.17	2.77 ± 0.43	2.99 ± 0.74	25 ± 25
SGMSE+ [9]	$\mathbf{3.61 \pm 0.63}$	$\mathbf{5.79 \pm 7.97}$	$\mathbf{3.03 \pm 0.67}$	$\mathbf{0.85 \pm 0.09}$	$\mathbf{3.49 \pm 0.43}$	$\mathbf{4.73 \pm 0.21}$	$\mathbf{9 \pm 12}$

Table 7: ***Results on EARS-Reverb***. *Column groups are the method name, intrusive metrics, non-intrusive metrics, and WER in percent. Values indicate mean and standard deviation.*

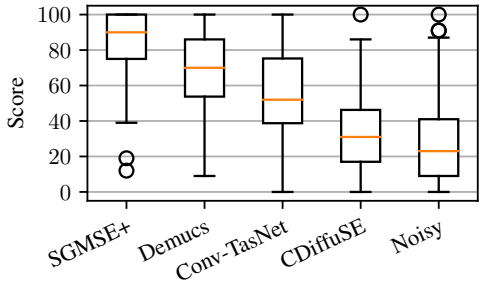


Figure 2: ***Results of the listening test***. *Subjective scores based on 20 participants visualized in a standard box plot.*

ence and the noisy audio as the hidden anchor. As stimuli, we use enhanced files of each noisy utterance from the four methods. We present participants with six audio files (four stimuli, hidden reference, hidden anchor) per utterance and ask them to "rate the overall quality considering artifacts and residual noise" of each on a scale of 0–100. The trends support the quantitative evaluation, demonstrating that SGMSE+ [9] is the preferred approach, closely followed by Demucs [7], see Figure 2.

Effect of Input SNR. Table 6 shows POLQA, SI-SDR, ESTOI, and WER scores segmented by input SNR, where 0 dB denotes the range $[-2.5, 2.5]$ dB, and each subsequent 5 dB increment representing the next range. As expected, there is a trend for better performance at higher input SNR, as well as smaller standard deviations than on the full test set.

Effect of Speaking Style and Emotion. We compare the performance of all baseline methods with respect to speaking style and selected core emotions in Table 8 and 9. We observe worse performance for whispered speech, which is expected since such voiceless speech is particularly difficult to recover after contamination with noise. Furthermore, it can be seen that all considered approaches trained on EARS-WHAM generalize well to emotional speech.

	regular	whisper	loud	slow	fast
Noisy	1.74	1.85	1.75	1.68	1.72
Conv-TasNet [6]	2.82	2.49	2.79	2.75	3.17
CDiffuSE [8]	1.78	1.68	1.93	1.72	2.02
Demucs [7]	2.95	2.82	3.10	2.86	3.27
SGMSE+ [9]	3.39	2.89	3.70	3.33	3.64

Table 8: ***POLQA for different speaking styles***. *Mean values.*

	neutral	anger	desire	pain	relief
Noisy	1.77	1.61	1.65	1.86	1.74
Conv-TasNet [6]	2.78	2.53	2.71	2.72	2.76
CDiffuSE [8]	1.65	1.94	1.76	1.94	1.78
Demucs [7]	2.93	2.91	2.82	3.02	2.91
SGMSE+ [9]	3.18	3.45	3.15	3.40	3.28

Table 9: ***POLQA for different emotions***. *Mean values.*

Dereverberation. Blind dereverberation with only a single microphone is known to be challenging, and recent results suggest that generative approaches are particularly well suited for this task [35]. In Table 7, we show dereverberation results on the EARS-Reverb test set, using the diffusion-based generative model SGMSE+ [9].

5. Conclusion

We released EARS, a dataset with high speaker and speaking style diversity spanning the full range of human speech. We hope this dataset will serve the community as a useful source to tackle new frontiers in speech processing. We additionally provided a speech enhancement and dereverberation benchmark on this new large-scale dataset and compared predictive and generative baselines to set a standard for future speech enhancement work on EARS.

6. Acknowledgements

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and DASHH (Data Science in Hamburg – Helmholtz Graduate School for the Structure of Matter) with Grant-No. HIDSS-0002. We would like to thank J. Berger and Rohde & Schwarz SwissQual AG for their support with POLQA.

7. References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[5] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[8] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7402–7406.

[9] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[10] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner, "ICASSP 2023 deep noise suppression challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[11] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *ISCA Interspeech*, 2020, pp. 1823–1827.

[12] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[13] K. Ito and L. Johnson, "The LJ Speech Dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[14] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[15] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete - Linguistic Data Consortium," 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93s6a

[16] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[17] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *IEEE Int. Conference on Digital Signal Processing*, 2009.

[18] K. Prawda, S. J. Schlecht, and V. Välimäki, "Robust selection of clean swept-sine measurements in non-stationary noise," *The Journal of the Acoustical Society of America*, vol. 151, no. 3, pp. 2117–2126, 2022.

[19] D. Fejgin, W. Middelberg, and S. Doclo, "BRUDEX database: Binaural room impulse responses with uniformly distributed external microphones," in *Proc. ITG Conference on Speech Communication*, 2023, pp. 126–130.

[20] D. D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dEchorate: a calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[21] S. V. Amengual Gari, B. Sahin, D. Eddy, and M. Kob, "Open database of spatial room impulse responses at Detmold university of music," in *Audio Engineering Society Convention 149*, 2020.

[22] "A sonic Palimpsest: Revisiting Chatham historic dockyards." [Online]. Available: https://research.kent.ac.uk/sonic-palimpsest/impulse-responses/

[23] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending speech separation to noisy environments," in *ISCA Interspeech*, 2019, pp. 1368–1372.

[24] Recommendation ITU-R BS.1770-5, "Algorithms to measure audio programme loudness and true-peak audio level," *International Telecommunication Union (ITU)*, 2023. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1770-5-202311-I/en

[25] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

[26] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[27] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001, pp. 749–752.

[28] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 626–630.

[30] N. C. Ristea, A. Saabas, R. Cutler, B. Naderi, S. Braun, and S. Branets, "ICASSP 2024 speech signal improvement challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[31] B. Naderi, R. Cutler, and N.-C. Ristea, "Multi-dimensional speech quality assessment in crowdsourcing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[32] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6493–6497.

[33] ITU-T Rec. P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," *International Telecommunication Union*, 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.808-202106-I/en

[34] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "NeMo: a toolkit for building AI applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

[35] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

6 Discussion and Conclusions

6.1 Analysis of Contributions

In this section, we summarize and analyze the key contributions of this dissertation. We adhere to the structure outlined in Section 1.7, beginning with unsupervised generative methods for speech enhancement, which includes a discussion on the extension to audio-visual speech enhancement. Next, we proceed with supervised generative methods, with a particular focus on diffusion models and causal processing for the restoration of speech communication artifacts. Additionally, we address the extension to audio-visual speech enhancement and alternative diffusion processes. Finally, we present our analysis of generative speech enhancement methods and compare them with predictive approaches.

6.1.1 Chapter 3: Unsupervised Generative Speech Enhancement

Our first contribution is the use of STCNs to model the distribution of clean speech power spectrograms, as proposed in [P1]. STCNs incorporate both hierarchical and temporal latent variables [228], enabling the modeling of temporal dependencies within the time-frequency domain. We train the model in an unsupervised manner on a clean speech dataset ("si_tr_s" subset of the Wall Street Journal (WJS0) dataset [229]) by using a variational lower bound on the log-likelihood as the loss function, similar to the approach taken in VAEs [117]. We substitute the VAE with the STCN within the VAE-NMF speech enhancement framework, as presented in [134]. With fewer than twice the number of parameters compared to the VAE (325 K versus 171 K parameters), the STCN is designed to balance complexity with computational efficiency. For testing, we use 651 synthetic mixtures generated from a clean speech dataset ("si_et_05" from the WJS0 dataset) combined with noise samples from the QUT-NOISE dataset [230]. The mixtures are created at SNRs uniformly chosen from the values -5, 0, and 5 dB. In our experiments, we compare our proposed method to the VAE-NMF framework presented in [134] using SI-SDR, extended short-time objective intelligibility (ESTOI), and perceptual objective listening quality assessment (POLQA) as evaluation metrics. Our results show that our approach significantly outperforms the baseline in all metrics, demonstrating the effectiveness of modeling the speech prior with STCNs. We argue that the hierarchical and temporal latent variables capture the temporal dependencies within the time-frequency domain, which is crucial for speech enhancement tasks.

Our second contribution extends the VAE-NMF framework for audio-visual speech enhancement. We adopt a late fusion approach, estimating high-level speech attributes, such as speech activity, from the video modality to guide the speech enhancement process. In [P3], we propose an adversarial training scheme for VAEs to separate a binary random variable from other continuous latent variables. For this purpose, we employ a discriminator network that

is trained with a binary cross-entropy loss to predict voice activity as a binary variable from the continuous latent variables. Concurrently, the encoder network is trained to maximize the prediction error of the discriminator by minimizing the negative binary entropy function of the binary random variable. Additionally, we use a classifier network, also trained with a binary cross-entropy loss, to predict the voice activity from clean speech. The decoder network then reconstructs the clean speech by using the sampled continuous latent variables along with the estimated probability of voice activity. During each training step, the parameters of the encoder, decoder, and classifier are first updated using a loss function that combines the evidence lower bound of the VAE, the binary cross-entropy loss from the classifier, and the negative binary entropy function. Subsequently, the parameters of the discriminator are updated using its own binary cross-entropy loss. For training, validation, and testing, we used the NTCD-TIMIT dataset [215], which contains audio-visual recordings from the TCD-TIMIT dataset [231] mixed with noise from the NOISEX-92 dataset [232] and CHiME challenges [233], [234]. During inference, we utilize a pretrained video-only classifier to estimate voice activity using the visual input. In our experiments, we compare our proposed method to two baseline frameworks: one using an unconditional VAE as described in [134], and another using a conditional VAE without disentanglement learning. We demonstrate that our method outperforms both baselines in terms of SI-SDR in the presence of non-stationary noise, highlighting the effectiveness of disentanglement learning for audio-visual speech enhancement.

Building on the above findings, we address the research questions **RQ1** and **RQ2** posed in Section 1.7.1 with answers **A1** and **A2** provided below.

Answers to Research Questions

RQ1 *Can an unsupervised generative model, such as the variational autoencoder (VAE), be used to capture the distribution of clean speech, including its temporal dependencies, to enhance the quality and intelligibility of noisy speech within a Bayesian speech enhancement framework?*

A1 Unsupervised generative models trained exclusively on clean speech can serve as a speech prior in Bayesian speech enhancement frameworks, such as the VAE-non-negative matrix factorization (NMF)-framework [133], [134]. This framework has shown strong generalization capabilities in unseen noisy conditions; however, its performance typically lags behind that of supervised methods. In [P1], we increased the modeling capacity of the speech prior by using stochastic temporal convolutional networks (STCNs), which capture temporal dependencies within the time-frequency domain. We show increased performance in enhancing the quality and intelligibility of noisy speech compared to the original VAE-NMF framework, indicated by significant improvements in SI-SDR, ESTOI, and POLQA metrics. Nevertheless, VAE-based speech enhancement methods remain limited due to the dimensionality reduction in the latent layer and the combined use of a linear noise model based on NMF.

RQ2 *How can a Bayesian speech enhancement method, such as the framework consisting of a variational autoencoder (VAE) and non-negative matrix factorization (NMF), be extended to incorporate the video modality for improved performance in unsupervised settings?*

A2 Audio-visual speech enhancement can be achieved by various fusion strategies that combine audio and visual information. For unsupervised speech enhancement, where no clean-noisy speech pairs are available, we propose a late fusion approach that uses voice activity as a high-level speech attribute estimated from the video modality to guide the speech enhancement process in the VAE-NMF framework [134]. In [P3], we propose a disentanglement learning scheme for VAEs that separates a binary random variable from other continuous latent variables. Inferring voice activity from a pretrained video-only classifier, we show that our proposed method outperforms both the original VAE-NMF framework and a conditional VAE without disentanglement learning in terms of SI-SDR in the presence of non-stationary noise.

6.1.2 Chapter 4: Supervised Generative Speech Enhancement

One of our most significant contributions, highlighted not only by the number of citations but also by the impact on the field, is the introduction and development of score-based generative models for speech enhancement. Our research has been pivotal in advancing diffusion-based speech enhancement, effectively bridging the performance gap between predictive and generative methods. We have demonstrated that diffusion-based generative approaches can achieve state-of-the-art results in both instrumental and subjective evaluations, underscoring their potential for practical applications. This significance is further endorsed by the preference for generative methods in contemporary challenge series, such as the "Speech Signal Improvement Challenge" [162] and the "URGENT Challenge" [235], where participants develop adaptable speech enhancement systems for various corruptions across different acoustic environments. Additionally, the rise of dedicated workshops and special sessions on generative speech enhancement at major conferences, including the IEEE ICASSP and the ISCA Interspeech conference, highlights the emerging interest and growing research community around these innovative models.

As our first contribution in Chapter 4, we present *SGMSE+* [P4], a supervised generative model that enables sampling from the posterior distribution of clean speech given the noisy observations. We propose modifying the drift term of the forward SDE to incorporate the noisy speech observations, causing the process mean to transition from clean speech to noisy speech throughout the forward process. This allows the reverse process to begin with the noisy observations combined with Gaussian noise, facilitating the generation of high-quality clean speech samples in just 30 diffusion steps. As an input representation, we propose applying exponential amplitude compression and multiplicative scaling to each complex coefficient of the spectrogram while keeping the phase unchanged. We adopt the *NCSN++* architecture [121] as the score model, using the process state, the noisy speech spectrogram, and the process time as inputs. For our experimental setup, we utilize two distinct paired datasets: the *WJS0+CHiME3* dataset, which combines the WJS0 dataset [229] with the *CHiME-3* dataset [234] at SNR levels ranging from 0 to 20 dB, and the *VoiceBank-Demand* dataset [97], a standardized dataset for speech enhancement tasks. This allows for cross-dataset evaluation, which is crucial for assessing the generalization capabilities of speech enhancement models. We examine a matched case in which both the training and test subsets are taken from the *WJS0+CHiME3* dataset, and a mismatched case where the training subset comes from the *WJS0+CHiME3* dataset while the test subset is taken from the VoiceBank-Demand dataset. Note that the matched case still consists of unseen noise and speakers, as the training

and test subsets are specifically designed to have no overlap in these characteristics. However, the recording conditions, such as the recording equipment and environment, remain the same in the matched case. As baselines for comparison, we consider predictive methods such as *Conv-TasNet* [86] and *MetricGAN+* [94], as well as generative methods including *RVAE* [138] and *CDiffuSE* [161]. We consider a set of instrumental evaluation metrics, including POLQA, PESQ, ESTOI, SI-SDR, and deep noise suppression mean opinion score (DNSMOS), to assess the performance of our proposed method. Our results indicate that *SGMSE+* outperforms all generative baselines and exhibits competitive performance against predictive methods in the matched case. In the mismatched case, *SGMSE+* significantly outperforms all baselines, showcasing the strong generalization capabilities of score-based diffusion models for speech enhancement tasks. In a listening experiment involving 10 participants, we further demonstrate that *SGMSE+* achieves the highest perceived quality scores among all baselines. In another experiment, we train the model on *WSJ0-REVERB*, a dataset that combines the *WSJ0* dataset with simulated RIRs. The results demonstrate that *SGMSE+* significantly outperforms all baselines in terms of speech quality and intelligibility, as indicated by instrumental metrics. This underscores the capability of diffusion-based speech enhancement models to handle not only additive corruption types but also different distortions, such as reverberation.

In our second contribution to supervised generative speech enhancement, we present a review article on diffusion models for audio restoration in [P5]. We begin by introducing the basics of diffusion models in a tutorial format. In particular, we highlight two aspects that make diffusion models well-suited for incorporating domain knowledge, showcasing them as model-based approaches. The first property stems from the physical inspiration behind diffusion models and their connection to Gaussian denoising, which makes them easier to interpret compared to other deep generative models, such as GANs. The second property of diffusion models is their natural integration within stochastic optimization and posterior sampling using Bayes' theorem, which makes them particularly well-suited for conditional generation. To classify the conditional generation techniques using diffusion models, we propose three categories: (1) *input conditioning*, where the score model incorporates the conditional variable as an input during both training and inference; (2) *task-adapted diffusion*, in which the diffusion process interpolates between the condition and the target variable; and (3) *external conditioning*, which combines an unconditional diffusion model with an external conditioner. Additionally, we provide a comprehensive review of diffusion models for inverse problems in audio restoration applications. A particular focus is on diffusion posterior sampling [145], which combines an unconditional diffusion model as a speech prior with a likelihood model based on a degradation model to approximate the posterior using Bayes' rule. We conclude by discussing the practical requirements of diffusion models for audio restoration tasks, focusing on sampling speed and robustness under adverse conditions.

In [P7], we adopt *SGMSE+* for the restoration of speech communication artifacts in a causal processing framework. Our contribution involves adapting the network architecture to create a causal version of the *NCSN++* architecture. We adjust the padding in the 2D convolutions and truncate the output to ensure that the convolution along the time dimension remains causal. We replace all group normalization layers with cumulative group normalization layers aggregating statistics recursively. Downsampling in the time dimension is performed with strided convolutions and corresponding upsampling with transposed strided convolutions. Finally, all attention layers as well as the progressive growing path are removed from the network. To address various types of speech communication artifacts, we construct the *MultiCorruption* dataset, which utilizes clean speech files from the *LibriVox* recordings [236]

and the entire *VCTK* speech data [237]. Using clean speech, we create six corrupted datasets, each featuring different types of corruption: additive noise, reverberation, clipping, packet loss, and bandwidth limitation combined with codec artifacts. To introduce additive background noise, we use noise files from the DNS challenge [238] and mix them with clean speech at a segmental SNR, which is uniformly sampled between −10 and 15 dB. Reverberant speech is created using RIRs from the DNS challenge [238], which are cropped so that the direct path occurs at the very beginning. In the experiments, we compare the proposed causal speech enhancement method with its non-causal counterpart. Additionally, we evaluate specialized models trained on individual corruption types against a generalized model trained on all types of corruption. Our findings indicate that while specialized models and non-causal models have a slight advantage, the generalized causal approach does not experience a significant performance penalty and is more practical for real-world applications.

Another contribution to supervised generative speech enhancement is the extension of *SGMSE+* to audio-visual speech enhancement. In [P9], we introduce *AV-Gen*, an audio-visual generative speech enhancement model that leverages visual information to improve speech quality and intelligibility. We propose to utilize audio-visual embeddings obtained from an SSL model fine-tuned on lipreading. The layer-wise features from its transformer-based encoder are aggregated, time-aligned, and integrated into the *NCSN++* architecture. We utilize the *LRS3* dataset [239] for training, validation, and testing, which includes audio-visual recordings from over 400 hours of TED and TEDx videos. To generate pairs of clean and noisy speech, we mix the audio tracks of *LRS3* with noise signals from the *CHiME3* dataset [234], using SNRs uniformly sampled between −6 and 12 dB. In the experiments, we compare *AV-Gen* to the audio-only *SGMSE+* model and a predictive baseline audio-visual model. The results show that the proposed method enhances speech quality, as evidenced by an increase of +.14 POLQA and +.15 PESQ compared to the audio-only model. An informal listening test indicates a noticeable reduction in generative artifacts, such as phonetic confusions. The effectiveness of the system is further validated by a significant reduction in WER in a downstream ASR model, particularly at low input SNRs.

In [P11], we explore alternative diffusion processes and investigate various training objectives for diffusion-based speech enhancement. We formulate an alternative learning objective similar to the *EDM* framework [174]. Instead of adopting a noise prediction loss, we train a denoiser model to minimize the MSE between the process mean and the denoiser's output. We investigate preconditioning to the input and output signals of the denoiser model based on the statistical properties of the data to improve the training convergence and performance of the generative model. Furthermore, we investigate the connection between *SGMSE+* and the Schrödinger bridge for speech enhancement [180]. The Schrödinger bridge is a diffusion bridge that aims to find the optimal transport between two distributions, which, in our case, are the clean and noisy speech distributions. We propose incorporating ad-hoc cost functions into the Schrödinger bridge framework, with a specific focus on adding a differentiable PESQ term to the loss function. In our experiments, we compare the various diffusion processes and training objectives on the *VoiceBank-Demand* dataset [97]. The results indicate variations in training behaviors and performance across these frameworks. The model based on the denoising loss achieves slightly higher scores than *SGMSE+*, but its training process is more unstable. Preconditioning the model results in faster training times, with performance comparable to that of *SGMSE+*. The Schrödinger bridge outperforms *SGMSE+* in most metrics, highlighting the potential of this framework for speech enhancement tasks. Furthermore, the Schrödinger bridge model can enhance speech using only a single diffusion step, which is significantly faster

compared to the 30 steps required by *SGMSE+*. By incorporating our proposed differentiable PESQ term, we observe significantly improved PESQ performance, albeit at the expense of SI-SDR. Nevertheless, the model trained with the PESQ term achieves the highest POLQA score among all models. Additionally, it leads to a leading result with a score of 3.70 in PESQ on the *VoiceBank-Demand* benchmark.

Building on the above findings, we address the research questions **RQ3**–**RQ6** posed in Section 1.7.2 with answers **A3**–**A6** provided below.

Answers to Research Questions

RQ3 *How can generative models, such as diffusion models, be effectively employed in supervised settings to learn the posterior distribution of clean speech conditioned on noisy speech, and what are their benefits over predictive methods and other generative methods in enhancing speech quality and intelligibility, particularly under mismatched acoustic conditions and non-additive corruptions?*

A3 Diffusion models, such as the proposed *SGMSE+* [P4], can be effectively employed in supervised settings to learn the posterior distribution of clean speech conditioned on noisy speech. By incorporating noisy speech into the drift term of the forward SDE in score-based generative models, *SGMSE+* allows for efficient sampling from the posterior using only 30 diffusion steps, resulting in high-quality clean speech generation. Compared to other generative methods like *RVAE* and *CDiffuSE*, *SGMSE+* demonstrates significantly superior performance across various instrumental metrics. While it achieves comparable results to strong predictive baselines such as *Conv-TasNet* and *MetricGAN+*, *SGMSE+* shows enhanced robustness in mismatched scenarios, where it significantly outperforms these predictive models and excels in addressing non-additive artifacts, such as reverberation. Furthermore, the model achieves higher sound quality and naturalness, as indicated by listening tests. These capabilities highlight the strong potential of using supervised generative models for diverse speech enhancement tasks.

RQ4 *How can causal processing be ensured when using diffusion models for restoring speech communication artifacts?*

A4 In [P7], we ensured causal processing in diffusion models by adapting the network architecture to process only past and current information, without relying on future data. This involved modifying the network architecture to include causal elements such as adjusting the padding in 2D convolutions and truncating the output to maintain causality along the time dimension. To address various types of speech communication artifacts, we trained specialized models on individual corruption types, including additive noise, reverberation, clipping, packet loss, bandwidth limitation, and codec artifacts. We also developed a generalized or universal model trained on all types of corruption. Our results indicated that while specialized models and their non-causal counterparts exhibited a slight performance advantage, the generalized causal approach did not experience a significant performance penalty. This makes the generalized causal model more practical and applicable for real-world scenarios where multiple types of corruption

may be present simultaneously.

RQ5 *How can visual information be utilized to improve the performance of generative speech enhancement in adverse acoustic conditions, particularly in alleviating phonetic errors in the generated speech?*

A5 Our research in [P9] introduces *AV-Gen*, an audio-visual generative speech enhancement model that incorporates visual cues to boost both the quality and intelligibility of the enhanced speech. This is achieved by integrating audio-visual embeddings from *AV-HuBERT* [220], an SSL model, specifically fine-tuned for lipreading tasks, into the *NCSN++* architecture. The process involves aggregating and time-aligning features derived from *AV-HuBERT*'s transformer-based encoder. In our experiments, which use the *LRS3* dataset [239] combined with noise signals from the *CHiME3* dataset [234] to simulate various adverse conditions, *AV-Gen* demonstrates superior performance compared to the audio-only *SGMSE+* model and a baseline audio-visual model. The results highlight improvements in objective speech quality metrics, with increases of +.14 in POLQA and +.15 in PESQ. Moreover, informal listening tests suggest a reduction in phonetic confusions, and a downstream ASR evaluation indicates a substantial decrease in word error rates, especially at low input SNRs.

RQ6 *How can alternative diffusion processes influence the performance and computational speed of diffusion-based speech enhancement?*

A6 Alternative diffusion processes can significantly impact both the performance and computational efficiency of speech enhancement models. In [P11], we investigate alternative diffusion processes and training objectives for improving diffusion-based speech enhancement. By formulating a learning objective akin to the *EDM* framework [174], we move away from noise prediction losses and instead train a denoiser model to minimize the MSE between the process mean and its output. This approach, coupled with input and output preconditioning based on data statistics, accelerates training convergence and enhances model performance. Moreover, in [P11] we explore the Schrödinger bridge for speech enhancement [180], a diffusion process that aims for an optimal transport between clean and noisy speech distributions. By introducing ad-hoc cost functions, particularly a differentiable PESQ term into the Schrödinger bridge framework, we observe a significant improvement in PESQ scores. Our experiments on the *VoiceBank-Demand* [97] dataset highlight that the Schrödinger bridge model not only surpasses *SGMSE+* [P4] in most metrics but also allows speech enhancement with a single diffusion step, significantly reducing computation time relative to the 60 function evaluations needed by *SGMSE+*. Despite a trade-off with SI-SDR, the inclusion of the differentiable PESQ term yields the highest POLQA score among evaluated models and achieves a remarkable PESQ score of 3.70 on the benchmark.

6.1.3 Chapter 5: Analysis and Comparison

In [P13], we release a new high-quality 48 kHz speech dataset called *EARS*, which contains 100 hours of anechoic speech recordings from 107 speakers, including emotional speech, different reading styles, and conversational freeform speech. The data is publicly available and can be used for various speech processing tasks[1]. We introduce two benchmarks using the *EARS* dataset: *EARS-WHAM* for speech enhancement, which mixes clean speech with noise from the WHAM dataset [240] at SNRs uniformly sampled between −2.5 and 17.5 dB; and *EARS-Reverb* for speech dereverberation, which combines clean speech with real-recorded RIRs from various RIR datasets. In the experiments, we compare two predictive methods, *Conv-TasNet* [86] and *Demucs* [101] with two generative methods, *CDiffuSE* [161], and *SGMSE+* [P4] on the *EARS-WHAM* dataset. The results show that *SGMSE+* significantly outperforms all baselines in all metrics, except for SI-SDR where it is on par with *Demucs* and *Conv-TasNet*. A MUSHRA-like (multiple stimuli with hidden reference and anchor) listening test on *EARS-WHAM* with 20 participants further confirms the superior performance of *SGMSE+* over the baselines. Additionally, we assess the performance of *SGMSE+* on the *EARS-Reverb* dataset as an initial step for future research on speech dereverberation tasks in 48 kHz.

Building on the above findings, we address the research question **RQ7** posed in Section 1.7.3 with answer **A7** provided below.

Answers to Research Questions

RQ7 Is diffusion-based speech enhancement, particularly the *SGMSE+* model, capable of producing studio-quality speech when trained on a high-quality, fullband dataset, and how does it compare to other generative and predictive methods?

A7 By adapting the STFT configuration and the network settings, *SGMSE+* [P4] is capable of generating studio-quality speech at 48 kHz, benefiting significantly from the high-quality, clean training data provided by the large-scale *EARS* dataset [P13]. The *SGMSE+* model demonstrates superior performance compared to the generative method *CDiffuSE* and predictive methods *Conv-TasNet* and *Demucs* in most metrics on the *EARS-WHAM* benchmark [P13]. A listening experiment with 20 participants further confirms the superior perceived quality of *SGMSE+*, indicating its potential to produce studio-quality speech. Additionally, evaluation of *SGMSE+* on the *EARS-Reverb* benchmark suggests strong capabilities in speech dereverberation tasks at 48 kHz. We demonstrate that diffusion-based generative models, such as *SGMSE+*, substantially narrowed the performance gap between predictive and generative methods for speech enhancement, encouraging the use of generative techniques as the preferred starting point for tackling speech restoration tasks.

[1] `https://sp-uhh.github.io/ears_dataset/`

6.2 Discussion and Future Research

Machine learning and speech processing continue to be dynamic and rapidly evolving fields, characterized by a growing number of peer-reviewed publications each year. In Figure 6.1, we show the number of accepted papers at leading conferences in speech processing and machine learning over the last six years. This growth is driven by an increasing number of researchers entering the field, greater accessibility to computational resources, and improving computational capabilities through constantly better hardware. An important observation related to the advancement of computational hardware is Huang's law [241], which highlights the rapid progression of graphics processing units (GPUs). In contrast to Moore's law [242], which predicts the doubling of transistors in a dense integrated circuit approximately every two years, Huang's law suggests that the performance of GPUs will more than double every two years. This accelerated advancement in GPU technology underpins many of the computational gains seen in machine learning, although the hypothesis does face scrutiny regarding its continued validity. Despite these advancements, the challenges of power consumption and memory constraints limit the deployment of large and computationally intensive models on resource-constrained hardware, such as hearing aids and embedded devices. Therefore, future research should focus not only on enhancing the robustness and performance of speech enhancement algorithms but also on improving their computational efficiency and memory footprint. This will enable broader application and integration of these technologies into practical, everyday systems.

In this thesis, we have demonstrated the effectiveness of generative models for speech enhancement tasks, particularly using score-based diffusion models [121]. Our flagship model *SGMSE+* [P4] has shown strong performance in enhancing speech quality and intelligibility at moderate SNRs, even in unseen acoustic conditions. The model can generate studio-quality clean speech samples when trained on our large-scale *EARS* dataset [P13]. Furthermore, we have demonstrated causal processing for speech communication artifacts in [P7], and the integration of visual information for audio-visual speech enhancement in [P9]. However, there are several avenues for future research that can enhance the performance and efficiency of generative speech enhancement models. Below, we will explore directions to improve their robustness and performance, as well as strategies to optimize model capacity and sampling speed.

In challenging acoustic conditions, such as low input SNRs and non-stationary noise like impulsive sounds, generative models for speech enhancement have been shown to introduce generative artifacts [19], including breathing sounds, phonetic confusions, and speech-like noise. In our demonstration [P14], we also experimented with varying speech levels and microphone distances, both of which can influence the quality of the generated speech. To alleviate these limitations, one straightforward approach is to increase the model capacity of the employed DNN. Additionally, one can expand the amount of training data or apply data augmentation strategies to incorporate more diverse acoustic conditions. Although this approach has proven successful in practice, in research, we are more interested in identifying novel methodological improvements that can significantly enhance the robustness of generative models, using publicly available datasets that allow for fair algorithmic comparisons. One such methodological improvement we proposed is in our method *StoRM* [P15], where we demonstrated that combining predictive models with generative models can enhance the robustness of speech enhancement models under adverse conditions. While using a predictive model for preprocessing and employing *SGMSE+* to learn the posterior distribution of clean

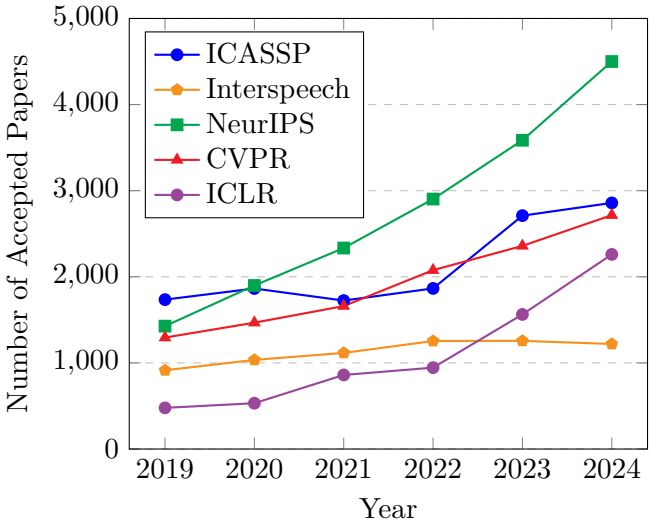


Figure 6.1: Number of accepted papers at leading conferences in speech processing and machine learning over the last six years.

speech given the processed speech seems like a straightforward approach, future research could explore more sophisticated methods for integrating predictive and generative models. One approach is proposed by Shi et al. [172], which involves adding a second decoder that shares the encoder with the original decoder. This additional decoder is trained to predict the clean speech spectrogram using a predictive loss. During inference, the output of the second decoder is utilized in the initial and final reverse steps by linearly interpolating between the outputs of the first and second decoders. In their experiments on the VoiceBank-Demand dataset, they demonstrated that incorporating the second decoder results in slightly higher PESQ values while significantly reducing the number of required diffusion steps compared to *StoRM*.

Another direction is to explore alternative diffusion processes that can improve the performance and robustness of generative speech enhancement models. In [21], we have proposed to use a Brownian bridge as a continuous-time diffusion process that interpolates linearly between the noisy and clean speech distributions. Compared to the Ornstein-Uhlenbeck process used in *SGMSE+* [P4], the Brownian bridge reduces the prior mismatch, which is the difference between the terminal distribution of the forward diffusion process and the prior distribution of the reverse diffusion process. As a result, the Brownian bridge improves performance with only half the iteration steps and removes the necessity of tuning the stiffness parameter as in the Ornstein-Uhlenbeck process. In [P11], we have explored the Schrödinger bridge as an alternative diffusion process for speech enhancement. The Schrödinger bridge is defined as the minimization of the Kullback-Leibler divergence between a path measure and a reference path measure, subject to boundary conditions, and can be regarded as an optimal transport problem. In contrast to the Ornstein-Uhlenbeck process in *SGMSE+*, the Schrödinger bridge enables the generative process to begin directly from the noisy input and uses a data prediction loss. This leads to improved performance and enables speech enhancement using just a single diffusion step. Other works have explored the use of Gaussian flow matching [178] as an alternative generative modeling framework. Liu et al. [177] have applied flow matching to speech restoration tasks, including speech enhancement. Using the *VoiceBank-Demand* dataset [97], they found that their flow matching-based generative model performs comparably to *SGMSE+* on most metrics but is inferior in ESTOI. However, this comparison is not

entirely fair, as different network architectures and training setups were used. Future research could ensure fair comparisons between flow matching and diffusion models by employing the same network architecture and training setup.

Flow matching has recently gained popularity due to its straightforward formulation and the "straightness" of its induced sampling trajectories [178]. Although rooted in a distinct theoretical perspective compared to diffusion models, flow matching in practice results in a similar inference algorithm. In fact, Gao et al. [243] demonstrate that diffusion models and Gaussian flow matching are essentially the same. However, variations in model specifications can result in different network outputs and sampling schedules, which can lead to performance differences. This relates to our work in [P11], where we specifically examined various training objectives for diffusion models and identified distinct differences in training behaviors and performance across these frameworks. Despite sharing many similarities with diffusion models, flow matching allows for the definition of arbitrary sampling trajectories, which can be advantageous for specific applications. Thus, future work could investigate the benefits of using flow matching for speech enhancement to define specific sampling trajectories and develop new samplers. Particularly, the amount of Gaussian noise introduced during inference is a crucial hyperparameter that significantly influences the characteristics of the generated samples. Lay et al. [244] realize that the scale of Gaussian noise is a crucial parameter in diffusion-based speech enhancement and demonstrate that it governs the balance between noise reduction and speech distortion. Specifically, they show that a larger variance increases noise attenuation and reduces the computational load by requiring fewer function evaluations to generate the estimate. Future research could explore the use of ordinary differential equation (ODE) samplers, as they have demonstrated high performance for speech enhancement using the Schrödinger bridge [180]. In ODE samplers for continuous-time diffusion models [121] and in the denoising diffusion implicit model (DDIM) sampler [123], randomness originates solely from the initial Gaussian sample, while the subsequent reverse process remains deterministic. In contrast, stochastic samplers add Gaussian noise at each step, arguably resulting in more diverse samples, as demonstrated in Figure 2 in Section 4.1.2. Interestingly, the ODE sampler for the Schrödinger bridge operates in a completely deterministic manner, as the reverse process begins with the noisy mixture and deterministically converges to the clean speech estimate without introducing any Gaussian noise. This deterministic behavior raises the question of whether the Schrödinger bridge, when using the ODE sampler, can still be classified as a generative model, since it does not generate samples according to our definition. In contrast, when using the SDE sampler for the Schrödinger bridge, the reverse process starts from the noisy mixture and adds Gaussian noise at each step, resulting in diverse samples. Future research could investigate the implications of using deterministic samplers for generative models and explore the potential benefits of using stochastic samplers for speech restoration tasks such as bandwidth extension and packet loss concealment.

In addition to exploring alternative diffusion processes, future research could investigate novel network architectures for generative speech enhancement. In SGMSE+ [P4], we employed the *NCSN++* architecture [121] as the score model, which is a noise conditional score network based on a multi-resolution U-Net structure. The architecture utilizes learned transformations including 2D convolutions, group normalization, and self-attention layers. Up- and downsampling is performed with finite impulse response filters. Another popular architecture for diffusion models is the *ablated diffusion model (ADM)* architecture [245], which is similar to the *NCSN++* architecture but uses a slightly different set of transformations. Despite the use of expressive architectures such as *NCSN++* and *ADM*, the training dynamics

of diffusion models remain challenging due to the highly stochastic nature of the loss function. The quality of the generated data is determined by the subtle details predicted throughout the sampling chain, where minor errors at intermediate steps can accumulate and lead to significant effects in subsequent iterations. The network needs to accurately estimate the average clean data across a wide range of noise levels, Gaussian noise realizations, and conditioning inputs. Learning to do so is challenging due to the chaotic training signal that is randomized across all these factors. To learn effectively in such a noisy training environment, the network should ideally respond predictably and consistently to parameter updates. Karras et al. [246], however, contend that current designs fall short of this ideal, which adversely affects model quality and complicates improvements. This is due to the intricate interactions between hyperparameters, network design, and training setups. Specifically, they have identified inconsistencies and inefficiencies in training behavior with the *ADM* architecture. They observed uncontrolled fluctuations and imbalances in the magnitudes of both network activations and weights throughout the training process. To address these issues, they redesigned the network layers to preserve activation, weight, and update magnitudes on expectation. With their magnitudes-preserving design, they achieved more stable training behavior, resulting in considerably better networks at equal computational complexity. In our preprint [P12], we investigate the magnitude preserving *ADM* for the Schrödinger bridge framework, and show promising results. We investigate different exponential moving average (EMA) profiles and their effects on the speech enhancement performance. In contrast to image generation tasks, where longer EMA lengths often enhance mode coverage, we observe that shorter EMA lengths consistently lead to better performance on standard speech enhancement metrics. Future research could explore the benefits of using magnitudes-preserving network architectures for generative speech enhancement, particularly for diffusion-based approaches.

Alongside exploring different diffusion methods and network architectures, future research could expand its focus to operate the diffusion process in alternative input spaces. Latent diffusion models [247] have gained popularity in recent years because they enable the application of diffusion processes in latent spaces rather than the original data space. Latent spaces can be learned using powerful autoencoders, which can significantly reduce the dimensionality of the data and allow for more efficient processing. Furthermore, model distillation techniques are employed to transfer knowledge from large models to smaller ones while keeping the impact on performance low. Knowledge distillation is a technique used to transfer information from a pretrained model, known as the *teacher*, to a smaller target model, known as the *student*. This approach reduces the computational demands of the model, as the student model is typically more compact and efficient than the teacher model. The teacher model, which is heavily parameterized, is initially trained on a large dataset but tends to be slow in inference and requires substantial memory. To address these issues, Hinton et al. [248] proposed training the student model to emulate the outputs of the pretrained teacher. This method successfully achieved the speech recognition performance of an ensemble of 10 models using just a single distilled model.

To enhance sampling speed, future research could consider employing supervised fine-tuning and adversarial training in generative models. One method involves initially pretraining a score model and then fine-tuning it using the discretized simulated reverse process. The fine-tuning approach integrates a predictive regression loss to refine the model's predictive abilities and accelerate the sampling process. Lay et al. [244] have demonstrated this two-stage training method by first pretraining a diffusion model with a score matching loss and subsequently fine-tuning it using a regression loss to predict clean speech signals. Another promising

strategy is adversarial training, which introduces GAN objectives into the diffusion model training process [122]. Xu et al. [249] proposed a method called *UFOGen*, which integrates diffusion models with GAN objectives to boost the quality of generated samples. *UFOGen* can utilize pretrained diffusion models, fine-tuning them with GAN objectives to improve sample quality. This approach is especially effective for generating high-quality images in a single step and can be adapted for speech enhancement tasks.

Diffusion models have seen an increasing application in generating temporal data, such as audio and video [250]. In the field of time-frequency domain speech processing, CNN architectures often use the same treatment for temporal and frequency axes, typically employing square kernels like 3×3, such as in *SGMSE+* [P4]. In [P7], we have adopted the convolution operations in the *NCSN++* architecture to be causal by adjusting the padding in the 2D convolutions and truncating the output to ensure that the convolution along the time dimension remains causal. However, the treatment of time and frequency remains the same. This approach could be refined by using handcrafted strides and pooling with learned kernels across the time dimension. Alternatively, RNNs could be used to handle the temporal dimension specifically in temporal data. Another idea would be to use the *Audio Spectrogram Transformer* [251], which leverages the *Transformer's* sequence dimension to natively capture temporal dynamics, replacing handcrafted CNN strides and pooling with learned attention patterns across time-frequency patches. Nevertheless, these models apply a uniform noise level across all time frames during the diffusion process and do not connect the physical time to the diffusion process time. Recent advancements, however, have suggested more nuanced approaches to generating temporal data with diffusion models. Ruhe et al. [252] have introduced rolling diffusion models that utilize a sliding window denoising process to generate temporal data. This method varies the noise levels across different frames within a sequence, progressively increasing noise for frames that appear later. This reflects the increasing uncertainty of future frames as the sequence progresses, offering a dynamic and adaptive approach to temporal data generation. Similarly, Chen et al. [253] have proposed a diffusion model that applies independent noise levels to tokens, which facilitates sequence generation through causal next-token prediction. This model effectively combines the flexibility of next-token models, known for handling variable-length generation, with the comprehensive capacity of full-sequence diffusion models, which are proficient in guided sampling. These developments present exciting opportunities for real-time speech enhancement applications, such as those required in telecommunications or hearing aids. By leveraging these advanced methods in diffusion-based speech enhancement, it is possible to develop an online system operating in a sliding-window fashion. Such a system would be able to process audio in real-time by utilizing past context, thus delivering enhanced speech quality and intelligibility in practical applications.

Besides exploring novel high-level design choices and methodological improvements, future research could investigate low-level computation and hardware design for generative speech enhancement. Optimizing inference and training pipelines through hardware-aware implementations could significantly improve efficiency, particularly for real-time applications. For instance, developing specialized causal *Compute Unified Device Architecture (CUDA)* kernels for diffusion models could reduce redundant computations in autoregressive or sliding-window settings, akin to the optimizations in *FlashAttention* [254] for *Transformers*. Similarly, leveraging hardware-specific features, such as tensor cores for mixed-precision training [255] or custom neural processing units like Google's TPUs [256], could accelerate diffusion sampling. Recent work on neural audio codecs (e.g., SoundStream [257]) demonstrates the benefits of kernel-level optimizations for low-latency audio generation, suggesting similar gains could be achievable

for speech enhancement. Further efficiency improvements could come from quantization techniques, e.g., *SmoothQuant* [258], or sparsity-aware architectures [259], which reduce memory bandwidth and computational overhead. Integrating these low-level optimizations with high-level algorithmic advances may enable deployment in resource-constrained environments, such as edge devices [260] or hearing aids, where real-time processing is critical.

As research in generative speech enhancement progresses, driven by more powerful computation, larger datasets, and improved algorithms, a fundamental question arises: "*When predictive methods become more robust and natural-sounding, and generative models adopt strong predictive behaviors, which paradigm do users ultimately prefer?*" Hybrid approaches, such as those combining predictive and generative objectives [P15], appear particularly promising as they aim to utilize the complementary strengths of both paradigms—the stability and efficiency of predictive methods with the naturalness and flexibility of generative approaches. Reinforcement learning from human feedback [261] could help align model behavior with human preferences, ensuring enhanced speech meets subjective quality standards. Additionally, generative speech enhancement can leverage spatial information from multi-microphone arrays [262], improving robustness in challenging acoustic environments. Ultimately, the success of these algorithms in real-world applications, from hearing aids to teleconferencing, will depend on their efficiency and reliability across diverse conditions. The path forward lies in balancing these priorities while adapting to user needs and hardware constraints.

References

[1] B. Arons, "A review of the cocktail party effect", *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.

[2] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*. John Wiley & Sons, 2000.

[3] G. A. Flamme, K. Liebe, and A. Wong, "Estimates of the auditory risk from outdoor impulse noise. I: Firecrackers", *Noise and Health*, vol. 11, no. 45, pp. 223–230, 2009.

[4] E. Winer, *The Audio Expert: Everything You Need to Know About Audio*. Routledge, 2017.

[5] J. R. Franks, *Preventing occupational hearing loss: A practical guide*. US Department of Health and Human Services, Public Health Service, Centers, 1996.

[6] Shure Inc., *Can a dynamic microphone handle really loud sounds? (Maximum SPL)*, `https://service.shure.com/s/article/can-a-dynamic-microphone-handle-really-loud-sounds-maximum-spl?language=en_US`, Accessed: 2025-02-22, 2021.

[7] K. S. Pearsons, R. L. Bennett, and S. A. Fidell, *Speech Levels in Various Noise Environments*. Office of Health, Ecological Effects, Office of Research, and Development, 1977.

[8] N. Misdariis and L.-F. Pardo, "The sound of silence of electric vehicles–issues and answers", in *InterNoise*, 2017.

[9] J. Cirillo, "Communication by unvoiced speech: The role of whispering", *Anais da Academia Brasileira de Ciências*, vol. 76, pp. 413–423, 2004.

[10] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.

[11] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC Press, 2007.

[12] T. Virtanen, E. Vincent, and S. Gannot, "Time-frequency processing: Spectral properties", *Audio source separation and speech enhancement*, pp. 15–29, 2018.

[13] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain Based Single-microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-art*. Morgan & Claypool Publishers, 2013, vol. 11.

[14] T. Gerkmann and E. Vincent, "Spectral masking and filtering", in *Audio Source Separation and Speech Enhancement*, Wiley, 2018, pp. 65–85.

[15] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 1979, pp. 208–211.

[17] P. Handel, "Low-distortion spectral subtraction for speech enhancement", in *European Conference on Speech Communication and Technology*, 1995, pp. 1549–1552.

[18] Z. Goh, K.-C. Tan, and T. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction", *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287–292, 1998.

[19] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, 2001.

[20] J. Beh and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: Spectral subtraction using spectral harmonics of speech", in *International Conference on Multimedia and Expo*, vol. 3, 2003, pp. 648–651.

[21] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949.

[22] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[23] S. M. Kay, *Fundamentals of statistical signal processing: Estimation theory*. Prentice Hall, 1993.

[24] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[25] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. 253–256.

[26] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model", *EURASIP Journal on Advances in Signal Processing*, pp. 1–17, 2005.

[27] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech, and Signal Srocessing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[28] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

[29] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[30] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4897–4900.

[31] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics”, *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[32] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging”, *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[33] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments”, *Speech communication*, vol. 48, no. 2, pp. 220–231, 2006.

[34] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay”, *IEEE Transactions Audio, Speech, Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[35] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis”, *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[36] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, “A review of deep learning techniques for speech processing”, *Information Fusion*, vol. 99, p. 101 869, 2023.

[37] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity”, *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[38] H. J. Kelley, “Gradient theory of optimal flight paths”, *ARS Journal*, vol. 30, no. 10, pp. 947–954, 1960.

[39] A. G. Ivakhnenko and V. G. Lapa, “Cybernetic predicting devices”, *CCM Information Corporation*, 1966.

[40] S. Amari, “A theory of adaptive pattern classifiers”, *IEEE Transactions on Electronic Computers*, no. 3, pp. 299–307, 1967.

[41] S.-I. Amari, “Characteristics of random nets of analog neuron-like elements”, *IEEE Transactions on systems, man, and cybernetics*, no. 5, pp. 643–657, 1972.

[42] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”, *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[43] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities.”, *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[44] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors”, *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[45] A. Waibel, “Phoneme recognition using time-delay neural networks”, in *SP87-100, Meeting of the Institute of Electrical, Information and Communication Engineers (IEICE), December, 1987, Tokyo, Japan*, 1987, pp. 19–24.

[46] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, “Phoneme recognition using time-delay neural networks”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[47] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1993, pp. 557–560.

[48] Y LeCun, "Generalization and network design strategies", *Connections in Perspective*, 1989.

[49] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[50] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series", *The handbook of brain theory and neural networks*, vol. 3361, no. 10, 1995.

[51] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation MIT-Press*, 1997.

[53] G. E. Hinton, "Deep belief networks", *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[55] C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts*. Springer Nature, 2023.

[56] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[57] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", in *Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[59] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio", in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, p. 125.

[60] A Vaswani, "Attention is all you need", *Advances in Neural Information Processing Systems*, 2017.

[61] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., "The LLaMA 3 herd of models", *arXiv preprint arXiv:2407.21783*, 2024.

[62] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT press, 2023.

[63] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview", *IEEE/ACM transactions on audio, speech, and language processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[64] S. Tamura and A. Waibel, "Noise reduction using connectionist models", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1988, pp. 553–556.

[65] S. Parveen and P. Green, "Speech enhancement with missing data techniques using recurrent neural networks", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–733.

[66] X. Lu, S. Matsuda, C. Hori, and H. Kashioka, "Speech restoration based on deep learning autoencoder with layer-wised pretraining.", in *Interspeech*, vol. 2012, 2012, pp. 1504–1507.

[67] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder", in *ISCA Interspeech*, 2013, pp. 436–440.

[68] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 1562–1566.

[69] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[70] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, 2020.

[71] A. Nicolson and K. K. Paliwal, "On training targets for deep learning approaches to clean speech magnitude spectrum estimation", *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3273–3293, 2021.

[72] R. Rehr and T. Gerkmann, "SNR-based features and diverse training data for robust DNN-based speech enhancement", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1937–1949, 2021.

[73] M. Tammen, D. Fischer, B. T. Meyer, and S. Doclo, "DNN-based speech presence probability estimation for multi-frame single-microphone speech enhancement", in *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing*, 2020, pp. 191–195.

[74] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 8443–8447.

[75] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[76] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning", *IEEE Internaional Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2017.

[77] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks", *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017.

[78] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms", *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[79] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[80] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7092–7096.

[81] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis", in *Speech separation by humans and machines*, Springer, 2005, pp. 181–197.

[82] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks", *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[83] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 92–102, 2011.

[84] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 63–72, 2012.

[85] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation", in *International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 696–700.

[86] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation", *IEEE/ACM Transactions Audio, Speech, Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[87] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020, pp. 46–50.

[88] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement", in *ISCA Interspeech*, 2020, pp. 2472–2476.

[89] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, subjective testing framework, and challenge results", in *ISCA Interspeech*, 2020, pp. 2492–2496.

[90] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain", in *ISCA Interspeech*, 2020, pp. 3291–3295.

[91] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement", *Applied Acoustics*, vol. 187, p. 108 499, 2022.

[92] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[93] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement", in *International Conference Machine Learning*, 2019.

[94] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of metricgan for speech enhancement", in *ISCA Interspeech*, 2021, pp. 201–205.

[95] I. J. Goodfellow, M. Mirza, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks", in *International Conference Learning Representations*, 2014.

[96] International Audio Laboratories Erlangen, *torch-pesq: Loss function inspired by the PESQ score*, `https://github.com/audiolabs/torch-pesq`, 2022.

[97] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech", *ISCA Speech Synthesis Workshop*, pp. 146–152, 2016.

[98] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces", in *Conference on Language Modeling*, 2024.

[99] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation", in *ISCA Interspeech*, 2020, pp. 2642–2646.

[100] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 21–25.

[101] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[102] R. Chao, W.-H. Cheng, M. La Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating Mamba for speech enhancement", *IEEE Spoken Language Technology Workshop*, 2024.

[103] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2020.

[104] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 6470–6474.

[105] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes", *Advances in Neural Information Processing Systems*, vol. 14, 2001.

[106] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[107] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, 1988.

[108] L. Rabiner and B. Juang, "An introduction to hidden markov models", *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[109] D. A. Reynolds et al., "Gaussian mixture models", *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[110] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods", 1993.

[111] I. Goodfellow, *Deep learning*. MIT press, 2016.

[112] J. M. Tomczak, *Deep Generative Modelling*. Springer, 2024.

[113] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[114] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines", in *Artificial intelligence and statistics*, PMLR, 2009, pp. 448–455.

[115] A.-r. Mohamed, G. Dahl, G. Hinton, et al., "Deep belief networks for phone recognition", in *Nips workshop on deep learning for speech recognition and related applications*, Vancouver, Canada, vol. 1, 2009, p. 39.

[116] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2011.

[117] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes", *International Conference Learning Representations*, 2014.

[118] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models", in *International Conference Machine Learning*, PMLR, 2014, pp. 1278–1286.

[119] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics", *International Conference Machine Learning*, pp. 2256–2265, 2015.

[120] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[121] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations", *International Conference Learning Representations*, 2021.

[122] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs", in *International Conference on Learning Representations*, 2022.

[123] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models", in *International Conference Learning Representations*, 2021.

[124] S. J. Prince, *Understanding deep learning*. MIT press, 2023.

[125] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: A review", *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, 2022.

[126] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders", in *ISCA Interspeech*, 2016, pp. 1770–1774.

[127] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation", in *ISCA Interspeech*, 2017, pp. 1273–1277.

[128] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data", *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[129] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis", in *International Conference Learning Representations*, 2021.

[130] A. Iashchenko, P. Andreev, I. Shchekotov, N. Babaev, and D. Vetrov, "Undiff: Unsupervised voice restoration with unconditional diffusion model", in *ISCA Interspeech*, 2023, pp. 4294–4298.

[131] D. de Oliveira, J. Richter, J.-M. Lemercier, S. Welker, and T. Gerkmann, "Non-intrusive speech quality assessment with diffusion models trained on clean speech", *arXiv preprint arXiv:2410.17834*, 2024.

[132] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017, pp. 86–90.

[133] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 716–720.

[134] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement", in *IEEE International Workshop on Machine Learning for Signal Processing*, 2018, pp. 1–6.

[135] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines", *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[136] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders", in *ISCA Interspeech*, 2019, pp. 3158–3162.

[137] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020, pp. 371–375.

[138] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2993–3007, 2022.

[139] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review", *Foundations and Trends in Machine Learning*, vol. 15, no. 1-2, pp. 1–175, 2021.

[140] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised speech enhancement with diffusion-based generative models", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024, pp. 12 481–12 485.

[141] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder", in *ISCA Interspeech*, 2020, pp. 2437–2441.

[142] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021, pp. 676–680.

[143] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian WaveNet", in *ISCA Interspeech*, 2017, pp. 2013–2017.

[144] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-channel blind unsupervised dereverberation with diffusion models", in *International Workshop on Acoustic Signal Enhancement*, 2024, pp. 120–124.

[145] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems", in *International Conference on Learning Representations*, 2023.

[146] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1826–1838, 2020.

[147] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[148] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, "CycleGAN-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism", in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2021, pp. 523–529.

[149] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training", *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[150] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech", in *European Signal Processing Conference*, 2021, pp. 436–440.

[151] Y. Ephraim, "Statistical-model-based speech enhancement systems", *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[152] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network", in *ISCA Interspeech*, 2017, pp. 3642–3646.

[153] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks", in *ISCA Interspeech*, 2019, pp. 1791–1795.

[154] Z.-X. Li, L.-R. Dai, Y. Song, and I. McLoughlin, "A conditional generative model for speech enhancement", *Circuits, Systems, and Signal Processing*, vol. 37, pp. 5005–5022, 2018.

[155] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 5024–5028.

[156] J.-T. Chien and K.-T. Kuo, "Variational recurrent neural networks for speech separation", in *ISCA Interspeech*, 2017, pp. 1193–1197.

[157] D. Braithwaite and W. B. Kleijn, "Speech enhancement with variance constrained autoencoders", in *ISCA Interspeech*, 2019, pp. 1831–1835.

[158] M. Strauss and B. Edler, "A flow-based neural network for time domain speech enhancement", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021, pp. 5754–5758.

[159] D. Rezende and S. Mohamed, "Variational inference with normalizing flows", in *International conference on machine learning*, PMLR, 2015, pp. 1530–1538.

[160] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model", in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2021, pp. 659–666.

[161] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022, pp. 7402–7406.

[162] R. Cutler, A. Saabas, B. Naderi, N.-C. Ristea, S. Braun, and S. Branets, "ICASSP 2023 speech signal improvement challenge", *IEEE Open Journal of Signal Processing*, 2024.

[163] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*. New York: Springer, 2019.

[164] T. Chen, G.-H. Liu, and E. Theodorou, "Likelihood training of Schrödinger bridge using forward-backward SDEs theory", in *International Conference on Learning Representations*, 2021.

[165] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, et al., "URGENT challenge: Universality, robustness, and generalizability for speech enhancement", in *ISCA Interspeech*, 2024, pp. 4868–4872.

[166] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, and M. Bacchiani, "Miipher: A robust speech restoration model integrating self-supervised speech and text representations", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2023, pp. 1–5.

[167] T. Peer, S. Welker, and T. Gerkmann, "Diffphase: Generative diffusion-based stft phase retrieval", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023, pp. 1–5.

[168] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023, pp. 1–5.

[169] N. Kamo, M. Delcroix, and T. Nakatani, "Target speech extraction with conditional diffusion model", in *ISCA Interspeech*, 2023, pp. 176–180.

[170] T. Nakatani, N. Kamo, M. Delcroix, and S. Araki, "Multi-stream diffusion model for probabilistic integration of model-based and data-driven speech enhancement", in *18th International Workshop on Acoustic Signal Enhancement*, 2024, pp. 65–69.

[171] R. Kimura, T. Nakatani, N. Kamo, D. Marc, S. Araki, T. Ueda, and S. Makino, "Diffusion model-based mimo speech denoising and dereverberation", in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, 2024, pp. 455–459.

[172] H. Shi, K. Shimada, M. Hirano, T. Shibuya, Y. Koyama, Z. Zhong, S. Takahashi, T. Kawahara, and Y. Mitsufuji, "Diffusion-based speech enhancement with joint generative and predictive decoders", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024, pp. 12 951–12 955.

[173] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Investigating the design space of diffusion models for speech enhancement", *IEEE Transactions on Audio, Speech, and Language Processing*, 2024.

[174] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models", *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 565–26 577, 2022.

[175] R. Scheibler, Y. Fujita, Y. Shirahata, and T. Komatsu, "Universal score-based speech enhancement with high content preservation", in *ISCA Interspeech*, 2024, pp. 1165–1169.

[176] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion", *arXiv preprint arXiv:2206.03065*, 2022.

[177] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching", in *International Conference on Learning Representations*, 2024.

[178] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling", in *International Conference on Learning Representations*, 2023.

[179] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, et al., "Audiobox: Unified audio generation with natural language prompts", *arXiv preprint arXiv:2312.15821*, 2023.

[180] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement", in *ISCA Interspeech*, 2024, pp. 1175–1179.

[181] A. Polyak, L. Wolf, Y. Adi, O. Kabeli, and Y. Taigman, "High fidelity speech regeneration with application to speech enhancement", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021, pp. 7143–7147.

[182] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training", in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 244–250.

[183] Z. Wang, X. Zhu, Z. Zhang, Y. Lv, N. Jiang, G. Zhao, and L. Xie, "Selm: Speech enhancement using discrete tokens and language models", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024, pp. 11 561–11 565.

[184] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis", *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[185] H. Yang, J. Su, M. Kim, and Z. Jin, "Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens", in *ISCA Interspeech*, 2024, pp. 1170–1174.

[186] H. Xue, X. Peng, and Y. Lu, "Low-latency speech enhancement via speech token generation", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024, pp. 661–665.

[187] B. Stein, *The Merging of the Senses*. MIT Press, 1993.

[188] L. Shams and A. R. Seitz, "Benefits of multisensory learning", *Trends in cognitive sciences*, vol. 12, no. 11, pp. 411–417, 2008.

[189] J. R. Zahn, L. Abel, and L. Dell'Osso, "Audio-ocular response characteristics.", *Sensory processes*, vol. 2, no. 1, pp. 32–37, 1978.

[190] B. E. Stein, M. A. Meredith, W. S. Huneycutt, and L. McDade, "Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli", *Journal of Cognitive Neuroscience*, vol. 1, no. 1, pp. 12–24, 1989.

[191] M. Sams, R. Aulanko, M. Hämäläinen, R. Hari, O. V. Lounasmaa, S.-T. Lu, and J. Simola, "Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex", *Neuroscience letters*, vol. 127, no. 1, pp. 141–145, 1991.

[192] G. A. Calvert, E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. Williams, P. K. McGuire, P. W. Woodruff, S. D. Iversen, and A. S. David, "Activation of auditory cortex during silent lipreading", *Science*, vol. 276, no. 5312, pp. 593–596, 1997.

[193] G. A. Calvert, M. J. Brammer, E. T. Bullmore, R. Campbell, S. D. Iversen, and A. S. David, "Response amplification in sensory-specific cortices during crossmodal binding", *Neuroreport*, vol. 10, no. 12, pp. 2619–2623, 1999.

[194] B. E. Dodd and R. E. Campbell, *Hearing by eye: The psychology of lip-reading.* Lawrence Erlbaum Associates, Inc, 1987.

[195] J. J. McDonald, W. A. Teder-Sälejärvi, and S. A. Hillyard, "Involuntary orienting to sound improves visual perception", *Nature*, vol. 407, no. 6806, pp. 906–908, 2000.

[196] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise", *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[197] B. Dodd, "The role of vision in the perception of speech", *Perception*, vol. 6, no. 1, pp. 31–40, 1977.

[198] H. McGurk and J. MacDonald, "Hearing lips and seeing voices", *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[199] G. A. Calvert, M. J. Brammer, and S. D. Iversen, "Crossmodal identification", *Trends in cognitive sciences*, vol. 2, no. 7, pp. 247–253, 1998.

[200] P Bertelson, *The cognitive architecture behind auditory-visual interaction in scene analysis and speech identification*, 1994.

[201] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy", *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[202] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[203] E. D. Petajan, "Automatic lipreading to enhance speech recognition", in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 1985, pp. 40–47.

[204] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition", in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1988, pp. 19–25.

[205] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks", *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.

[206] L. Girin, G. Feng, and J.-L. Schwartz, "Noisy speech enhancement with filters estimated from the speaker's lips.", in *European Conference on Speech Communication and Technology*, 1995.

[207] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks", *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[208] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement", *ISCA Interspeech*, pp. 3244–3248, 2018.

[209] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation", *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[210] X. Hu, K. Li, W. Zhang, Y. Luo, J.-M. Lemercier, and T. Gerkmann, "Speech separation using an asynchronous fully recurrent convolutional neural network", *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 509–22 522, 2021.

[211] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.

[212] M. Sadeghi and X. Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7534–7538.

[213] M. Sadeghi and X. Alameda-Pineda, "Mixture of inference networks for vae-based audio-visual speech enhancement", *IEEE Transactions on Signal Processing*, vol. 69, pp. 1899–1909, 2021.

[214] M. Sadeghi and X. Alameda-Pineda, "Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6663–6667.

[215] A. H. Abdelaziz et al., "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition.", in *ISCA Interspeech*, 2017, pp. 3752–3756.

[216] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey", *Patterns*, vol. 3, no. 12, 2022.

[217] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised learning: A succinct review", *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2761–2775, 2023.

[218] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations", *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[219] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[220] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction", in *International Conference on Learning Representations*, 2022.

[221] A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic, "Jointly learning visual and auditory speech representations from raw data", in *International Conference Learning Representations*, 2023.

[222] S. Lee, C. Jung, Y. Jang, J. Kim, and J. S. Chung, "Seeing through the conversation: Audio-visual speech separation based on diffusion model", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024, pp. 12 632–12 636.

[223] C. Jung, S. Lee, J.-H. Kim, and J. S. Chung, "FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching", in *ISCA Interspeech*, 2024, pp. 7412–7416.

[224] K. Yang, D. Marković, S. Krenn, V. Agrawal, and A. Richard, "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis", in *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8227–8237.

[225] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, "ReVISE: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 795–18 805.

[226] R. Mira, B. Xu, J. Donley, A. Kumar, S. Petridis, V. K. Ithapu, and M. Pantic, "LA-VocE: Low-SNR audio-visual speech enhancement using neural vocoders", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023, pp. 1–5.

[227] H. Fang, S. Frintrop, and T. Gerkmann, "Uncertainty-driven hybrid fusion for audio-visual phoneme recognition", in *Speech Communication; 15th ITG Conference*, VDE, 2023, pp. 255–259.

[228] E. Aksan and O. Hilliges, "STCN: Stochastic temporal convolutional networks", in *International Conference Learning Representations*, 2019.

[229] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, *CSR-I (WSJ0) Complete*, May 2007.

[230] D. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition", in *ISCA Interspeech*, 2015, pp. 3456–3460.

[231] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech", *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[232] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems", *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[233] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines", in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 126–130.

[234] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines", in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[235] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, "URGENT challenge: Universality, robustness, and generalizability for speech enhancement", in *ISCA Interspeech*, 2024, pp. 4868–4872.

[236] *LibriVox: Free public domain audiobooks*, `https://librivox.org/`, Accessed: 2023-07-24.

[237] J. Yamagishi, C. Veaux, K. MacDonald, et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)", *University of Edinburgh: The Centre for Speech Technology Research*, 2019.

[238] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 deep noise suppression challenge", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022.

[239] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition", *arXiv preprint arXiv:1809.00496*, 2018.

[240] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending speech separation to noisy environments", in *ISCA Interspeech*, 2019, pp. 1368–1372.

[241] T. Perry, "Move over, Moore's law: Make way for Huang's law", *IEEE Spectrum*, vol. 22, no. 08, 2021.

[242] G. Moore, "Cramming more components onto integrated circuits", *Electronics Magazine*, 1965.

[243] R. Gao, E. Hoogeboom, J. Heek, V. D. Bortoli, K. P. Murphy, and T. Salimans, "Diffusion meets flow matching: Two sides of the same coin", 2024.

[244] B. Lay and T. Gerkmann, "An analysis of the variance of diffusion-based speech enhancement", in *ISCA Interspeech*, 2024, pp. 2205–2209.

[245] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis", *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[246] T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine, "Analyzing and improving the training dynamics of diffusion models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 174–24 184.

[247] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[248] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network", *Stat*, vol. 1050, p. 9, 2015.

[249] Y. Xu, Y. Zhao, Z. Xiao, and T. Hou, "Ufogen: You forward once large scale text-to-image generation via diffusion gans", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8196–8206.

[250] Y. Yang, M. Jin, H. Wen, C. Zhang, Y. Liang, L. Ma, Y. Wang, C. Liu, B. Yang, Z. Xu, et al., "A survey on diffusion models for time series and spatio-temporal data", *CoRR*, 2024.

[251] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer", in *Interspeech 2021*, 2021, pp. 571–575.

[252] D. Ruhe, J. Heek, T. Salimans, and E. Hoogeboom, "Rolling diffusion models", in *International Conference Machine Learning*, 2024.

[253] B. Chen, D. M. Monsó, Y. Du, M. Simchowitz, R. Tedrake, and V. Sitzmann, "Diffusion forcing: Next-token prediction meets full-sequence diffusion", in *Conference on Neural Information Processing Systems*, 2024.

[254] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness", *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 344–16 359, 2022.

[255] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al., "Mixed precision training", in *International Conference on Learning Representations*, 2018.

[256] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, et al., "In-datacenter performance analysis of a tensor processing unit", in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[257] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[258] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "Smoothquant: Accurate and efficient post-training quantization for large language models", in *International Conference on Machine Learning*, PMLR, 2023, pp. 38 087–38 099.

[259] P. Wang, Y. Ji, C. Hong, Y. Lyu, D. Wang, and Y. Xie, "SNrram: An efficient sparse neural network computation architecture based on resistive random-access memory", in *Proceedings of the 55th Annual Design Automation Conference*, 2018, pp. 1–6.

[260] B. Liu, Z. Luo, H. Chen, and C. Li, "A survey of state-of-the-art on edge computing: Theoretical models, technologies, directions, and development paths", *IEEE Access*, vol. 10, pp. 54 038–54 063, 2022.

[261] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences", *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[262] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2022.

List of Acronyms

ADM	ablated diffusion model 137, 138
ASR	automatic speech recognition 1, 24, 27, 30, 32, 131
cIRM	complex ideal ratio mask 14
CNN	convolutional neural network 9, 14, 24, 30, 139
CUDA	Compute Unified Device Architecture 139
dB	decibels 2, 3, 14, 16, 129
DDIM	denoising diffusion implicit model 137
DFT	discrete Fourier transform 6, 7
DNN	deep neural network 10–15, 18, 22, 135
DNS	deep noise suppression 14, 131
DNSMOS	deep noise suppression mean opinion score 130
DRR	direct-to-reverberant ratio 2, 3
EDM	Elucidating the Design Space of Diffusion-Based Generative Models 27, 35, 131, 133
EM	expectation-maximization 21, 31
EMA	exponential moving average 138
ESTOI	extended short-time objective intelligibility 127, 128, 130, 136
fMRI	functional magnetic resonance imaging 28
FPE	Fokker-Planck equation 225
GAN	generative adversarial network 9, 15, 18, 19, 22–24, 130, 139
GPU	graphics processing unit 135
IBM	ideal binary mask 13, 14
ICASSP	International Conference on Acoustics, Speech, and Signal Processing 25, 129
IEEE	Institute of Electrical and Electronics Engineers 25, 129
IRM	ideal ratio mask 13
ISCA	International Speech Communication Association 25, 129
iSTFT	inverse short-time Fourier transform 13, 14
LSTM	long short-term memory 9, 12, 14
MCEM	Monte Carlo expectation-maximization 21, 33
MCMC	Markov chain Monte Carlo 18, 21

MEG magnetoencephalography 28
MFCC mel-frequency cepstral coefficient 32
MIMO multi-input multi-output 26
ML maximum likelihood 6
MMSE minimum mean square error 6, 7, 12
MSE mean square error 5, 6, 27, 31, 131, 133, 229
MVDR minimum variance distortionless response 12

NCSN noise conditional score network 25, 129–131, 137, 139
NMF non-negative matrix factorization 8, 21, 22, 31, 33, 41, 127–129

ODE ordinary differential equation 137

PESQ perceptual evaluation of speech quality 15, 16, 130–133, 136
POLQA perceptual objective listening quality assessment 127, 128, 130–133
PSD power spectral density 6, 12
PSM phase-sensitive mask 13, 14

RIR room impulse response 3, 22, 130, 131, 134
RNN recurrent neural network 9, 11, 14, 24, 139
ROI region of interest 32

SDE stochastic differential equation 24, 27, 34, 55, 129, 132, 137, 225
SDR signal-to-distortion ratio 14
SI-SDR scale-invariant source to distortion ratio 12, 21, 127–130, 132–134
SNR signal-to-noise ratio 2, 7, 8, 11, 12, 16, 22, 28, 32, 127, 129, 131, 133–135
SPL sound pressure level 2
SSL self-supervised learning 27, 31, 32, 131, 133
STCN stochastic temporal convolutional network 33, 41, 127
STFT short-time Fourier transform 4, 5, 12–14, 26, 134
STOI short-time objective intelligibility 15

VAE variational autoencoder 18–24, 31, 33, 34, 41, 127–129

WER word error rate 32, 131
WJS0 Wall Street Journal 127, 129
WPE weighted prediction error 26

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, 12.05.2025



Julius Richter

Appendices

Related Publications

A

A.1 Variational Autoencoders for Speech Enhancement

A.1.1 Publication: Guided Variational Autoencoder for Speech Enhancement with a Supervised Classifier

Abstract

Recently, variational autoencoders have been successfully used to learn a probabilistic prior over speech signals, which is then used to perform speech enhancement. However, variational autoencoders are trained on clean speech only, which results in a limited ability of extracting the speech signal from noisy speech compared to supervised approaches. In this paper, we propose to guide the variational autoencoder with a supervised classifier separately trained on noisy speech. The estimated label is a high-level categorical variable describing the speech signal (e.g. speech activity) allowing for a more informed latent distribution compared to the standard variational autoencoder. We evaluate our method with different types of labels on real recordings of different noisy environments. Provided that the label better informs the latent distribution and that the classifier achieves good performance, the proposed approach outperforms the standard variational autoencoder and a conventional neural network-based supervised approach.

Reference

Guillaume Carbajal, Julius Richter, and Timo Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2021. DOI: 10.1109/ICASSP39728.2021.9414363

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2021 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Guillaume Carbajal and Julius Richter made equal contributions to this work. Timo Gerkmann provided insights into the experimental validation and mathematical derivations and participated in reviewing the manuscript.

GUIDED VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT WITH A SUPERVISED CLASSIFIER

Guillaume Carbajal, Julius Richter, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany
{guillaume.carbajal, julius.richter, timo.gerkmann}@uni-hamburg.de

ABSTRACT

Recently, variational autoencoders have been successfully used to learn a probabilistic prior over speech signals, which is then used to perform speech enhancement. However, variational autoencoders are trained on clean speech only, which results in a limited ability of extracting the speech signal from noisy speech compared to supervised approaches. In this paper, we propose to guide the variational autoencoder with a supervised classifier separately trained on noisy speech. The estimated label is a high-level categorical variable describing the speech signal (e.g. speech activity) allowing for a more informed latent distribution compared to the standard variational autoencoder. We evaluate our method with different types of labels on real recordings of different noisy environments. Provided that the label better informs the latent distribution and that the classifier achieves good performance, the proposed approach outperforms the standard variational autoencoder and a conventional neural network-based supervised approach.

***Index Terms*—** Speech enhancement, deep generative model, variational autoencoder, semi-supervised learning.

1. INTRODUCTION

The task of single-channel speech enhancement consists in recovering a speech signal from a mixture signal captured with one microphone in a noisy environment [1]. Common speech enhancement approaches estimate the speech signal using a filter in the time-frequency domain to reduce the noise signal while avoiding speech artifacts [2]. Under the Gaussian assumption, the optimal filter in the minimum mean square error sense requires estimating the signal variances [3–5].

Supervised deep neural networks (DNNs) have demonstrated excellent performance in estimating the speech signal [6–9]. However, supervised approaches require labeled data which originates from pairs of noisy and clean speech. These pairs can be created synthetically. However, since supervised approaches may not generalize well to unseen situations, a large amount of pairs is needed to cover various acoustic conditions, e.g. different noise types, reverberation and different signal-to-noise ratios (SNRs).

Recently, deep generative models based on the variational autoencoder (VAE) have gained attention for learning the probability distribution of complex data [10]. VAEs have been used to learn a prior distribution of clean speech, and have been combined with an untrained non-negative matrix factorization (NMF) noise model to estimate the signal variances using a Monte Carlo expectation maximization (MCEM) algorithm [11, 12]. However, since the VAE speech model is trained in an unsupervised manner on clean speech only, its ability of extracting speech characteristics from noisy speech is limited in low SNRs. This results in limited speech enhancement performance compared to supervised approaches in already-seen noisy environments [11].

To overcome this limitation, the VAE can be conditioned on an auxiliary variable that allows for a more informed probabilistic latent distribution [13]. Kameoka et al. used a VAE conditioned on the speaker identity to inform the speech prior for multichannel speech separation [14]. However, their approach can only separate speakers which are included in the training set. As a result, their approach aims at speaker-dependent speech separation and not at speaker-independent speech enhancement.

In this work, we propose to guide the VAE with a classifier fully decoupled from the VAE. The classifier is trained separately in a supervised manner with pairs of noisy and clean speech. The estimated label is a high-level categorical variable describing the speech signal (e.g. speech activity). We show that the choice of label is crucial for the performance of the proposed guided VAE. In addition, we show that a noise-robust classifier is also required to outperform the standard VAE and a conventional supervised DNN-based approach.

The rest of this paper is organized as follows. In Section 2 we summarize the background related to the VAE for speech enhancement. Section 3 describes our proposed approach. The experimental setup is described in Section 4 which is followed by the evaluation in Section 5.

2. BACKGROUND

2.1. Mixture model and filtering

In the time-frequency domain using the short time Fourier transform (STFT), the mixture signal $x_{nf} \in \mathbb{C}$ is the sum of the clean speech $s_{nf} \in \mathbb{C}$ and the noise $b_{nf} \in \mathbb{C}$:

$$x_{nf} = \sqrt{g_n}\, s_{nf} + b_{nf}, \tag{1}$$

at time frame index $n \in [1, N]$ and frequency bin $f \in [1, F]$, where N denotes the number of time frames and F the number of frequency bins of the utterance. The scalar $g_n \in \mathbb{R}_+$ represents a frequency-independent but time-varying gain providing some robustness with respect to the time-varying loudness of different speech signals [12].

Under the Gaussian assumption, the clean speech s_{nf} can be estimated in the minimum mean square error sense using the Wiener estimator:

$$\widehat{s}_{nf} = \frac{\widehat{g}_n \widehat{v}_{s,nf}}{\widehat{g}_n \widehat{v}_{s,nf} + \widehat{v}_{b,nf}}\, x_{nf}, \tag{2}$$

where $\widehat{v}_{s,nf}$ and $\widehat{v}_{b,nf}$ are the estimated variances of the clean speech s_{nf} and the noise b_{nf}, respectively. Under a local stationary

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and ahoi.digital.

assumption, short-time power spectra $|s_{nf}|^2$ and $|b_{nf}|^2$ are unbiased estimates of the signal variances [15].

2.2. Model M1: standard VAE as a speech prior

The standard VAE which we refer to as model M1 is used to learn a prior over clean speech [11, 12]. At time frame n, the frequency bins of clean speech $\mathbf{s}_n \in \mathbb{C}^F$ are modeled as

$$\mathbf{s}_n|\mathbf{z}_{1,n} \sim \mathcal{CN}(\mathbf{0}, \mathrm{diag}(\mathbf{v}_\theta(\mathbf{z}_{1,n}))), \quad \mathbf{z}_{1,n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3)$$

where $\mathbf{z}_{1,n} \in \mathbb{R}^D$ denotes a latent variable of dimension D and $\mathbf{v}_\theta : \mathbb{R}^D \mapsto \mathbb{R}_+^F$ represents a trainable feedforward DNN called the generative model or *decoder* parametrized by θ. (see Fig 1a).

In variational inference, the posterior of $\mathbf{z}_{1,n}$ is approximated as

$$\mathbf{z}_{1,n}|\,\mathbf{s}_n \sim \mathcal{N}(\boldsymbol{\mu}_\phi(|\mathbf{s}_n|^2), \mathrm{diag}(\mathbf{v}_\phi(|\mathbf{s}_n|^2))), \quad (4)$$

where $\boldsymbol{\mu}_\phi : \mathbb{R}_+^F \mapsto \mathbb{R}^D$ and $\mathbf{v}_\phi : \mathbb{R}_+^F \mapsto \mathbb{R}_+^D$ represent feedforward DNNs sharing the same input and hidden layers called the recognition model or *encoder* which are parametrized by ϕ (see Fig 1b). Note that the absolute value and squaring in (4) are performed element-wise.

The generative model and recognition model are simultaneously trained by maximizing the evidence lower bound (ELBO) on the per-frame log-likelihood

$$\begin{aligned}\log p_\theta(\mathbf{s}_n) \geq\ & \mathbb{E}_{q_\phi(\mathbf{z}_{1,n}|\mathbf{s}_n)}[\log p_\theta(\mathbf{s}_n|\mathbf{z}_{1,n})] \\ & - \mathcal{D}_{\mathrm{KL}}(q_\phi(\mathbf{z}_{1,n}|\mathbf{s}_n)||p(\mathbf{z}_{1,n})),\end{aligned} \quad (5)$$

where the first term is the reconstruction loss and $\mathcal{D}_{\mathrm{KL}}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

2.3. Non-negative matrix factorization as noise model

The noise variance is modeled with an untrained NMF as

$$v_{b,nf} = \{\mathbf{HW}\}_{nf}, \quad (6)$$

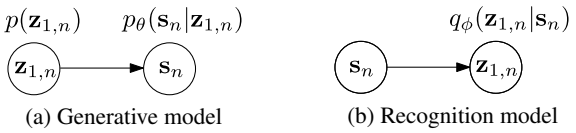
where $\mathbf{H} \in \mathbb{R}_+^{N\times K}$ and $\mathbf{W} \in \mathbb{R}_+^{K\times F}$ are two non-negative matrices representing the temporal activations and spectral patterns of the noise power spectrogram. K denotes the NMF rank.

2.4. Clean speech estimation

Given the speech prior provided by model M1 and the noise model, the mixture signal x_{nf} is distributed as

$$x_{nf}|\,\mathbf{z}_{1,n} \sim \mathcal{CN}(0, g_n\{\mathbf{v}_\theta(\mathbf{z}_{1,n})\}_f + \{\mathbf{HW}\}_{nf}) \quad (7)$$

where $\Theta_u = \{g_n, \mathbf{H}, \mathbf{W}\}$ are the unsupervised parameters to be estimated. Since the resulting optimization problem is intractable due to the non-linear relation between the speech variance and the latent variable, an MCEM algorithm is employed to iteratively optimize the unsupervised parameters Θ_u [12]. At each iteration, the estimated terms $\widehat{g}_n\{\mathbf{v}_\theta(\mathbf{z}_{1,n})\}_f$ and $\{\widehat{\mathbf{HW}}\}_{nf}$ are supposed to get closer to the true variances $v_{s,nf}$ and $v_{b,nf}$, respectively, reaching a local optimum. Note that while the VAE operates on a frame-by-frame basis, the MCEM algorithm is offline, resulting in an offline estimation of the parameters.



(a) Generative model (b) Recognition model

Fig. 1. Model M1 consisting of (a) a generative model $p_\theta(\mathbf{s}_n|\mathbf{z}_{1,n})p(\mathbf{z}_{1,n})$ and (b) a recognition model $q_\phi(\mathbf{z}_{1,n}|\mathbf{s}_n)$.

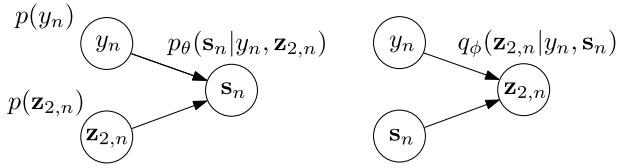
At test time, the recognition model takes the mixture signal $\mathbf{x}_{nf}$ as input instead of clean speech $\mathbf{s}_{nf}$. However, since model M1 is trained on clean speech only, its ability of extracting speech characteristics from the mixture $\mathbf{x}_{nf}$ is limited. As a result, the speech enhancement performance of the MCEM using M1 may be lower compared to supervised approaches trained on already-seen noisy environments [11].

3. GUIDED VARIATIONAL AUTOENCODER

In this section, we propose a guided VAE which consists in an extension of model M1 combined with a supervised classifier.

3.1. Model M2: labeled VAE as a speech prior

Inspired by Kingma et al.'s deep generative model for semi-supervised learning [13], we extend model M1 with a categorical variable $y_n \in \mathcal{Y}$ that characterizes a high-level feature of the speech signal (e.g. speech activity). Hereafter, we denote y_n as the *label*. The label is supposed to allow for a more informed probabilistic speech prior learned by the VAE. We describe our choice for y_n in Section 3.3.



(a) Generative model (b) Recognition model

Fig. 2. Model M2 consisting of (a) the guided generative model $p_\theta(\mathbf{s}_n|y_n, \mathbf{z}_{2,n})p(\mathbf{z}_{2,n})p(y_n)$ and (b) the guided recognition model $q_\phi(\mathbf{z}_{2,n}|\mathbf{s}_n, y_n)$.

At time frame n, the frequency bins of clean speech $\mathbf{s}_n$ are generated as

$$\mathbf{s}_n|y_n, \mathbf{z}_{2,n} \sim \mathcal{CN}(\mathbf{0}, \mathrm{diag}(\mathbf{v}_\theta(y_n, \mathbf{z}_{2,n}))), \quad (8)$$

where $\mathbf{z}_{2,n}$ has the same prior as $\mathbf{z}_{1,n}$ and $\mathbf{v}_\theta : \mathcal{Y} \circ \mathbb{R}^D \mapsto \mathbb{R}_+^F$ is a feedforward DNN resulting in the guided generative model (see Fig 2a). The posterior of $\mathbf{z}_{2,n}$ is approximated as

$$\mathbf{z}_{2,n}|y_n, \mathbf{s}_n \sim \mathcal{N}(\boldsymbol{\mu}_\phi(y_n, |\mathbf{s}_n|^2), \mathrm{diag}(\mathbf{v}_\phi(y_n, |\mathbf{s}_n|^2))), \quad (9)$$

where $\boldsymbol{\mu}_\phi : \mathcal{Y} \circ \mathbb{R}_+^F \mapsto \mathbb{R}^D$ and $\mathbf{v}_\phi : \mathcal{Y} \circ \mathbb{R}_+^F \mapsto \mathbb{R}_+^D$ are feedforward DNNs sharing the same input and hidden layers resulting in the guided recognition model (see Fig 2b).

The guided generative model and recognition model are simultaneously trained by maximizing the ELBO on the per-frame joint log-likelihood

$$\begin{aligned}\log p_\theta(y_n, \mathbf{s}_n) \geq\ & \mathbb{E}_{q_\phi(\mathbf{z}_{2,n}|y_n,\mathbf{s}_n)}[\log p_\theta(\mathbf{s}_n|y_n, \mathbf{z}_{2,n})] \\ & - \mathcal{D}_{\mathrm{KL}}(q_\phi(\mathbf{z}_{2,n}|y_n, \mathbf{s}_n)||p(\mathbf{z}_{2,n})) + \log p(y_n),\end{aligned} \quad (10)$$

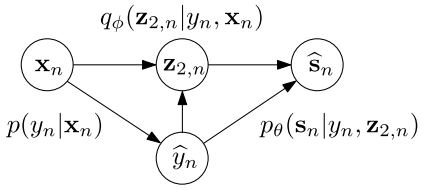


Fig. 3. Combined model M2 and classifier $p(y_n|\mathbf{x}_n)$ at test time.

where the first term is the reconstruction loss and $p(y_n)$ is the prior distribution of y_n.

3.2. Clean speech estimation

The estimation path at test time is shown in Fig. 3. First, we use a classifier to estimate $\widehat{y}_n$ from the mixture $\mathbf{x}_n$. Then, we use the mixture signal $\mathbf{x}_n$ and the estimated label $\widehat{y}_n$ as inputs for the guided recognition model. Given $\widehat{y}_n$ and the latent variable $\mathbf{z}_{2,n}$, the mixture signal x_{nf} is distributed as

$$x_{nf}|\widehat{y}_n, \mathbf{z}_{2,n} \sim \mathcal{CN}(0, g_n\{\mathbf{v}_\theta(\widehat{y}_n, \mathbf{z}_{2,n})\}_f + \{\mathbf{HW}\}_{nf}) \quad (11)$$

For estimating the unsupervised parameters we use the same MCEM configuration as for model M1. Provided that 1) the classifier is noise-robust and that 2) the label y_n better informs the speech prior, model M2 is supposed to better extract speech characteristics from the mixture $\mathbf{x}_n$ than model M1.

3.3. Classifier for label estimation

Since the classifier is fully decoupled from model M2, it can be trained separately. This fact can be used to construct the best classifier possible. In order to obtain a noise-robust classifier, we train a feedforward DNN in a supervised manner using the mixture power spectra $|\mathbf{x}_n|^2$ as inputs and corresponding labels y_n as targets. The classifier outputs the posterior probability $p(y_n|\mathbf{x}_n)$ and the estimated label $\widehat{y}_{nf}$ is subsequently determined by selecting the class c corresponding to the highest posterior probability $p(y_n = c|\mathbf{x}_n)$.

We consider two types of labels related to speech activity. First, we use a classifier to perform voice activity detection (VAD), i.e. $y_n^{\text{VAD}} \in \{0, 1\}$. We use the binary cross entropy (BCE) as the learning objective for this classifier. The prior of y_n^{VAD} is a symmetric Bernoulli distribution. Second, we consider a classifier to perform ideal binary mask (IBM) estimation, i.e. $\mathbf{y}_n^{\text{IBM}} \in \{0, 1\}^F$, which is equivalent to perform VAD per time-frequency bin. Thus, we use the BCE averaged over all frequency bins f and the prior for each frequency bin is a symmetric Bernoulli distribution.

4. EXPERIMENTAL SETUP

4.1. Dataset

For training, we use the "si_tr_s" subset of the Wall Street Journal (WSJ0) dataset which consists of approximately 25 h of clean speech [16], and the noise signals DWASHING, NRIVER, OOFICE and TMETRO of the DEMAND dataset [17]. For validation, we use the "si_dt_05" subset of WSJ0 and the noise signals NFIELD, OHALLWAY, PSTATION and TBUS of the DEMAND dataset. All signals have a sampling rate of 16 kHz. For the test, we use the "si_et_05" subset of WSJ0 consisting of 651 utterances, resulting in 1.5 h and the noise signals from the "verification" subset of the QUT-NOISE dataset [18], which we downsample to 16 kHz. Note that both speakers and noise types in the test set are different than in the training set. Each mixture signal is created by uniformly sampling a noise type and mixing speech and noise signals at SNRs of -5, 0 and $+5$ dB.

4.2. Baselines

Hereafter, we denote model M2 with VAD labels as *M2+VAD* and model M2 with IBM labels as *M2+IBM*. We use the DNN-based classifier described in Section 3.3 for the estimation of the VAD and IBM labels. To compare with our DNN-based IBM classifier, we also use a non-learned classifier which consists of the IBM estimator used inside the algorithm of Gerkmann and Hendriks [5], originally employed for noise PSD estimation.

For the baselines, we use model M1 and a feedforward DNN estimating a Wiener-like mask trained with the magnitude spectrum approximation loss [19], which we denote as *Supervised*.

4.3. Hyperparameter settings

The STFT is computed using a 64 ms Hann window with 75% overlap, resulting in a frame period of 16 ms and $F = 513$ unique frequency bins. To obtain the ground truth for the VAD and IBM labels, we use the method of Heymann et al. related to clean speech [20].

For a fair comparison between all the approaches, we consider a similar architecture for each model. Tab. 1 shows the configuration of the models. In particular, we consider 5 hidden layers for *Supervised* to match the same number of layers as models M1 and M2 (*encoder* + *decoder*). Model M1 has 171,297 learnable parameters whereas *M2+VAD* has 177,729 and *M2+IBM* has 302,625. The VAD classifier has 82,433 learnable parameters whereas the IBM classifier has 148,993 and *Supervised* has 198,017.

We use the Adam optimizer with standard configuration and a learning rate of 10^{-3} [21]. We set the batch size to 128. Note that because the learning objective of the classifier is scale-dependent, the DNN input $|\mathbf{x}_n|^2$ needs to be normalized at training time. This is not the case for models M1 and M2 since the reconstruction loss (i.e. the Ikatura-Saito distance) is scale-independent. Early stopping with a patience of 20 epochs is performed using the validation set. For the MCEM we follow the settings of Leglaive et al. and set the NMF rank to $K = 10$ [12]. For the non-learned IBM classifier, we use the standard configuration of the IBM estimator as in Gerkmann and Hendriks [5].

	Hidden layers			Output layer
Model	**# layers**	**# units**	**act. fn**	**act. fn**
Encoder	2	128	tanh	identity
Decoder	2	128	tanh	exp
DNN classifer	2	128	ReLU	sigmoid
Supervised	5	128	ReLU	sigmoid

Table 1. Model configurations.

4.4. Metrics

To evaluate the classification performance of the different classifiers, we use the F1-score which combines the precision and recall rates. To evaluate the speech enhancement performance of the approaches, we use the scale-invariant signal-to-distortion ratio (SI-SDR) measured in dB [22].

Model	Classifier	F1-score	SI-SDR	Input SNR		
				−5 dB	0 dB	5 dB
Mixture	–	–	0.2 ± 0.3	-5.0 ± 0.0	0.0 ± 0.0	5.0 ± 0.0
Supervised	–	–	4.7 ± 0.4	-0.9 ± 0.5	4.5 ± 0.4	9.7 ± 0.3
M1	–	–	6.4 ± 0.4	1.4 ± 0.6	6.3 ± 0.5	10.8 ± 0.4
M2+VAD	DNN	0.82	6.3 ± 0.6	1.3 ± 0.6	6.3 ± 0.5	10.7 ± 0.4
	oracle	1.00	6.8 ± 0.4	2.0 ± 0.6	6.8 ± 0.5	11.2 ± 0.4
M2+IBM	[5]	0.34	-0.5 ± 0.3	-2.2 ± 0.6	-0.2 ± 0.5	0.8 ± 0.6
	DNN	0.62	$\mathbf{7.3 \pm 0.4}$	$\mathbf{2.8 \pm 0.6}$	$\mathbf{7.1 \pm 0.5}$	$\mathbf{11.5 \pm 0.4}$
	oracle	1.00	9.5 ± 0.4	5.7 ± 0.5	9.3 ± 0.5	13.3 ± 0.4

Table 2. Average SI-SDR (in dB) and 95% confidence intervals on the test set on average and for different input SNRs.

5. RESULTS

Tab. 2 shows the results on average and per input SNR. Regarding the results on average, *M2+VAD* with DNN classifier outperforms *Supervised* by 1.7 dB but is outperformed by *M1* by 0.1 dB. *M2+VAD* with the oracle classifier outperforms *M1* by 0.4 dB but the difference is not statistically significant. We conclude that, even with the best classifier, VAD does not inform the speech prior learned by the VAE statistically better.

M2+IBM with DNN classifier outperforms *Supervised* by 2.6 dB and *M1* by 0.9 dB on average. The performance of *M2+IBM* is also statistically significant compared to these two models. However, the performance of *M2+IBM* with the non-learned classifier dramatically drops compared to the DNN classifier. Since the classification performance of the non-learned classifier is worse than the DNN classifier, we conclude that the performance of *M2+IBM* crucially depends on the classifier. Finally, the performance of *M2+IBM* with the oracle classifier shows that IBM informs the speech prior significantly better. Thus, the performance of *M2+IBM* could be improved with a better classifier.

M2+IBM with DNN classifier also outperforms *M1* for all input SNRs. The difference between the two models gets larger as the input SNR decreases. Therefore, *M2+IBM* with the DNN classifier is particularly more robust to noise than *M1* in low SNRs.

From informal listening tests, we can state that *M2+IBM* with DNN classifier typically reduces the noise better than *M1*, which is particularly obvious in the presence of nonstationary noise and transient interferences such as bursts. Code and audio examples are available online[1].

6. CONCLUSION

We proposed to guide a VAE for speech enhancement with a supervised classifier separately trained on noisy speech. We evaluated our method with labels corresponding to VAD and IBM on real recordings of different noisy environments. Using the IBM as label and a feedforward DNN classifier, the guided VAE outperforms the standard VAE and a feedforward DNN-based Wiener filter, particularly in low SNRs. Improving the classifier by taking time dependencies and/or visual information into account could further improve the guided VAE. The model could then be compared to other deep generative models taking temporal dependencies into account [23].

7. REFERENCES

[1] E. Vincent, T. Virtanen, and S. Gannot, eds., *Audio Source Separation and Speech Enhancement*. Hoboken, NJ: John Wiley & Sons, 2018.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art*. No. 11 in Synthesis Lectures on Speech and Audio Processing, Williston, VT: Morgan & Claypool, 2013.

[3] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement Without Musical Noise," *IEEE Signal Processing Letters*, vol. 14, pp. 1036–1039, Dec. 2007.

[4] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, pp. 793–830, Mar. 2009.

[5] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, May 2012.

[6] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, pp. 7092–7096, May 2013.

[7] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 2136–2147, Dec. 2015.

[8] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 483–492, Mar. 2016.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, Aug. 2019.

[10] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[11] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization," in *ICASSP*, pp. 716–720, Apr. 2018.

[12] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *MLSP*, pp. 1–6, Sept. 2018.

[1] https://uhh.de/inf-sp-guided2021

[13] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *NeurIPS*, pp. 3581–3589, Curran Associates, Inc., 2014.

[14] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, pp. 1891–1914, Sept. 2019.

[15] A. Liutkus, R. Badeau, and G. Richard, "Gaussian Processes for Underdetermined Source Separation," *IEEE Transactions on Signal Processing*, vol. 59, pp. 3155–3167, July 2011.

[16] J. S. Garofolo, D. Graff, D. Paul, and D. S. Pallett, *CSR-I (WSJ0) Sennheiser.* 1993.

[17] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: A Collection Of Multi-Channel Recordings Of Acoustic Noise In Diverse Environments," June 2013.

[18] D. Dean, A. Kanagasundaram, H. Ghaemmaghami, H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE Protocol for the Evaluation of Noisy Speaker Recognition," in *Interspeech*, pp. 3456–3460, 2015.

[19] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *GlobalSIP*, pp. 577–581, Dec. 2014.

[20] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, pp. 196–200, Mar. 2016.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, Dec. 2014.

[22] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?," in *ICASSP*, pp. 626–630, May 2019.

[23] J. Richter, G. Carbajal, and T. Gerkmann, "Speech Enhancement with Stochastic Temporal Convolutional Networks," in *Interspeech*, Oct. 2020.

A.2 Diffusion Models for Speech Enhancement

A.2.1 Publication: Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain

Abstract

We consider the problem of speech modeling in speech enhancement. Recently, deep generative approaches based on variational autoencoders have been proposed to model speech spectrograms. However, these approaches are based either on hierarchical or temporal dependencies of stochastic latent variables. In this paper, we propose a generative approach to speech enhancement based on a stochastic temporal convolutional network, which combines both hierarchical and temporal dependencies of stochastic variables. We evaluate our method with real recordings of different noisy environments. The proposed speech enhancement method outperforms a previous non-sequential approach based on feed-forward fully-connected networks in terms of speech distortion, instrumental speech quality and intelligibility. At the same time, the computational cost ofthe proposed generative speech model remains feasible, due to inherent parallelism of the convolutional architecture.

Reference

Simon Welker, Julius Richter, Timo Gerkmann, "Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain," Proceedings of Interspeech, Incheon, Korea, Sep. 2022, pp. 2928–2932. DOI: 10.21437/Interspeech.2022-10653

Copyright Notice

The following article is the accepted version of the article published with ISCA.
© 2022 ISCA. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Simon Welker and Julius Richter collaboratively designed the SGMSE algorithm, ran all training and evaluation experiments, and jointly wrote the manuscript. Simon Welker proposed the idea of interpolating stochastic differential equations (OUVE/OUVP SDE) for speech enhancement and wrote most of the code. Julius Richter conducted the data generation, contributed significant portions of the code, and performed baseline experiments. Timo Gerkmann provided insights into the experimental validation and mathematical derivations and participated in reviewing the manuscript.

Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain

Simon Welker[1,2,†], *Julius Richter*[1,†], *Timo Gerkmann*[1]

[1]Signal Processing (SP), Universität Hamburg, Germany
[2]Center for Free-Electron Laser Science, DESY, Hamburg, Germany
[†]Authors contributed equally to this work.
simon.welker@uni-hamburg.de, julius.richter@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

Score-based generative models (SGMs) have recently shown impressive results for difficult generative tasks such as the unconditional and conditional generation of natural images and audio signals. In this work, we extend these models to the complex short-time Fourier transform (STFT) domain, proposing a novel training task for speech enhancement using a complex-valued deep neural network. We derive this training task within the formalism of stochastic differential equations (SDEs), thereby enabling the use of predictor-corrector samplers. We provide alternative formulations inspired by previous publications on using generative diffusion models for speech enhancement, avoiding the need for any prior assumptions on the noise distribution and making the training task purely generative which, as we show, results in improved enhancement performance.

Index Terms: speech enhancement, generative modeling, score-based generative models, deep learning

1. Introduction

Speech enhancement aims at estimating clean speech signals from audio recordings that are impacted by acoustic noise [1]. The task is well studied in the signal processing literature, and conventional approaches often make assumptions regarding the statistical properties of speech signals [2]. With the advent of deep learning, approaches to speech enhancement have made significant progress in the last decade. Most methods are based on a discriminative learning task that aims to minimize a certain distance between clean and noisy speech. However, since supervised methods are inevitably trained on a finite set of training data with limited model capacity for practical reasons, they may not generalize to unseen situations, e.g., different noise types, reverberation, and different signal-to-noise ratios (SNRs). In addition, some discriminative approaches have been shown to result in unpleasant speech distortions that outweigh the benefits of noise reduction [3].

Instead of learning a direct mapping from noisy to clean speech, generative models aim to learn the distribution of clean speech as a prior for speech enhancement. Several approaches have utilized deep generative models for speech enhancement using generative adversarial networks (GANs) [4,5], variational autoencoders (VAEs) [6–11], flow-based models [12], and more recently generative diffusion models [13, 14]. The main principle of these approaches is to learn the inherent properties of clean speech, such as its spectral and temporal structure, which are then used as prior knowledge for making inferences about clean speech given a noisy input. Thus, they are trained solely to generate clean speech and are therefore considered more robust to different acoustic environments compared to their discriminative counterparts. In fact, generative approaches have shown to perform better under mismatched training and test conditions [10, 11, 14, 15]. However, they are currently less studied and still lag behind discriminative approaches, which is a strong incentive to conduct more research to realize their full potential.

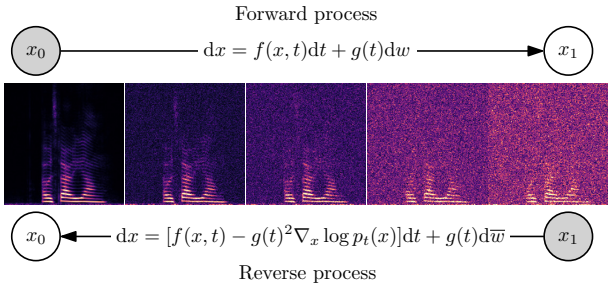


Figure 1: *The forward and reverse process on a spectrogram as a solution to an SDE. The reverse process gradually converts the corrupted signal x_1 into a clean speech spectrogram x_0.*

Although GANs and VAEs have become a popular choice for deep generative modeling, generative diffusion models [16–18] provide a powerful alternative which has recently shown to be state-of-the-art in natural image generation [19]. The idea is to gradually turn data into noise, and to train a neural network that learns to invert this procedure for different noise scales (see Fig. 1). Other works have adopted this scheme for generating speech in the time domain using clean Mel spectrograms as a conditioner [20, 21]. Interestingly, the model in [20] investigated its zero-shot speech denoising capabilities for different noise types. Although it was only trained to remove white noise added in the diffusion process, it already shows preliminary results for performing speech enhancement. Lu et al. built upon this idea and designed a supportive reverse process using the same architecture but with noisy spectrograms as the conditioner [13]. In their follow-up paper they devised a conditional diffusion process with an adopted forward and reverse process incorporating the noisy data into the diffusion process [14]. However, the derivation of their objective function is based on the assumption that the global distribution of the additive noise follows a standard white Gaussian, which is normally not the case for environmental noise. Moreover, the neural network used in their reverse process is trained to explicitly estimate the difference between clean and noisy speech, which is usually considered a discriminative learning task.

We acknowledge the support by DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002 and the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).

In this work, we propose an alternative diffusion process which is based on stochastic differential equations (SDEs) for score-based generative models (SGMs) [18] which is a specific class of generative diffusion models (see Section 2). Using this formulation, we are able to avoid the need for any prior assumptions on the noise distribution and make the training task purely generative. Unlike [13, 14] which work in the time domain, we perform generative speech enhancement in the complex STFT domain, i.e., working directly on the Fourier coefficients with amplitude and phase. This choice should allow deep-learning models to exploit the rich structure of speech in the time-frequency domain. In contrast to approaches that modify only the amplitudes, this choice avoids the need for phase retrieval techniques, since the phase information is enhanced directly as well. In our experiments, we demonstrate good speech enhancement performance and show that our method introduces less speech distortions compared to the baseline method.

2. Background

2.1. Score-based generative models

SGMs [16, 18] are generative diffusion models that center around the idea of corrupting training data with slowly increasing levels of noise (forward process), and training a *score model* to reverse this corruption (reverse process) by estimating the *score* $\boldsymbol{\nabla}_x \log p_{\text{data}}(x)$, i.e., the gradient of the log probability density with respect to the data. Once a score model is trained, an iterative procedure called Langevin dynamics can be used to draw samples from it [22].

2.2. SGMs through stochastic differential equations

Song et al. [18] introduced the formalism of using SDEs for score-based generative modeling. Following the Itô interpretation [23, ch. 4], the process $\{x_t\}_t$ can be defined by the SDE

$$\mathrm{d}x_t = f(x_t, t)\mathrm{d}t + g(t)\mathrm{d}w, \tag{1}$$

where f is the *drift*, g is the *diffusion*, t is a coordinate which denotes how far the forward process has run, $\mathrm{d}t$ is an infinitesimal time step, and w is the standard Wiener process. We stress that the "time" t mentioned in this context is a conceptual idea related to the progression of the process, and is not related to the time axis of an audio signal or its STFT. According to Anderson [24], every such SDE has a corresponding *reverse SDE*

$$\mathrm{d}x_t = \left[f(x_t, t) - g(t)^2 \boldsymbol{\nabla}_{x_t} \log p_t(x_t)\right] \mathrm{d}t + g(t)\mathrm{d}\overline{w}, \tag{2}$$

where $\mathrm{d}t$ is an infinitesimal negative time step and $\overline{w}$ is a standard Wiener process for time flowing in reverse. The score term $\boldsymbol{\nabla}_{x_t} \log p_t(x_t)$ is with respect to the log density of the process at time t and can be approximated by a learned time-dependent score model $s_\theta(x_t, t)$. The reverse SDE is then solved by means of some solver procedure (see Section 3.4), providing the basis for score-based generative modeling with SDEs.

3. Proposed method

3.1. A stochastic process for speech enhancement

Lu et al. [14] proposed the idea of linearly interpolating between clean speech x_0 (at $t = 0$) and noisy speech $y = x_0 + n$ (at $t = 1$) along a discrete-time forward process, so that the reverse process should lead to clean speech at $t = 0$. Note, that discrete-time diffusion processes using Markov chains [17] are different from the continuous SDE formulation used in the context of SGMs [18] and in this work. Furthermore, as can be seen from their subsequent derivations, the choice of linear interpolation implies that the trained deep neural network (DNN) must explicitly estimate a portion of environmental noise n at each step in the reverse process. From an intuitive standpoint, this occurs because it is necessary to estimate the slope between every value of x_0 and y to be able to formulate the reverse process. As a consequence, the resulting evidence lower bound (ELBO) loss [14, eq. (21)] exhibits characteristics of a discriminative learning task, making it difficult to assess whether the network is learning a prior for clean speech per se or the mapping from noisy to clean speech. To avoid this, we propose the following novel diffusion process based on an SDE, which leads to a *pure* generative training objective without the model estimating any portion of n:

$$\mathrm{d}x_t = \gamma(y - x_t)\mathrm{d}t + g(t)\mathrm{d}w \tag{3}$$

$$g(t) = \sigma_{\text{min}} \left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)^t \sqrt{2 \log\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)} \tag{4}$$

where σ_{min} and σ_{max} parameterize the variance schedule of the added Gaussian noise, and γ is a constant that can be interpreted as a stiffness parameter for x_0 being pulled towards y as t becomes larger. This SDE is inspired by the concept of stochastic processes that exhibit *mean reversion* [25], i.e., a convergence of the mean to a particular value as $t \to \infty$. Since the drift $f(x_t, t) = \gamma(y - x_t)$ is affine with respect to x_t, the SDE (3) describes a Gaussian process [23]. Let $\mathbf{I}$ be the identity matrix. The state distribution of this process, called perturbation kernel,

$$p_{0t}(x_t|x_0, y) = \mathcal{N}_{\mathbb{C}}\left(x_t; \mu_{x_0,y}(t), \sigma(t)^2\, \mathbf{I}\right), \tag{5}$$

admits efficient sampling x_t at an arbitrary timestep t because it is fully characterized by its mean $\mu_{x_0,y}(t)$ and variance $\sigma(t)^2$, which can be derived in closed form via eqns. (5.50, 5.53) in Särkkä & Solin [23] as:

$$\mu_{x_0,y}(t) = e^{-\gamma t} x_0 + (1 - e^{-\gamma t}) y \tag{6}$$

$$\sigma(t)^2 = \frac{\sigma_{\text{min}}^2 \left(\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)^{2t} - e^{-2\gamma t}\right) \log\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)}{\gamma + \log\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right)} \tag{7}$$

Since $\mu_{x_0,y}(t)$ does not exactly reach y for finite t, we empirically choose a stiffness γ so that $\mathbb{E}\left[|\mu_{x_0,y}(1) - y|^2\right] < 10^{-3}$, where the expectation is calculated over all complex bins in a random sample of 256 spectrogram pairs from the chosen dataset. Our SDE is, in essence, a synthesis of an Ornstein-Uhlenbeck SDE [26] and the Variance Exploding SDE by Song et al. [18]. The affine drift term (as in an Ornstein-Uhlenbeck SDE) leads to exponential decay of the mean from x_0 towards y, reflected in (6), and the exponentially increasing diffusion term (as in a Variance Exploding SDE) leads to exponentially increasing corruption of features by Gaussian noise.

We note that in (5), we assume circularity and a scaled identity matrix as the covariance matrix for the process. This would be a strong assumption to make if the process should model real-world additive noise. We argue that the assumption is acceptable in this case, as the noise added by the forward process is entirely artificial, intended to mask the particular characteristics of the environmental noise at $t = 1$, and thus represents only a means to an end for the generative task.

3.2. Data representation

In this work, we represent speech signals in the complex-valued one-sided STFT domain. We therefore treat clean speech x_0

and noisy speech y as elements of $\mathbb{C}^{(\frac{F}{2}+1)\times T}$, where F is the discrete Fourier transform (DFT) length and T is the number of time frames, dependent on the audio length. Since the global distribution of STFT speech amplitudes is typically heavy-tailed [27], the information visible in untransformed spectrograms is dominated by only a small portion of bins. Furthermore, the naturally encountered amplitudes often lie far outside the interval $[0, 1]$ often used in SGMs [17, 18]. As an engineering trick, we thus apply an amplitude transformation to all complex STFT coefficients c [28], in an effort to bring out frequency components with lower energy (e.g. fricative sounds of unvoiced speech) and to normalize amplitudes roughly to within $[0, 1]$ to better fit the usual SGM assumptions. The transformation and its associated inverse transformation is defined as:

$$\tilde{c} = \frac{|c|^\alpha}{\beta} e^{i\angle(c)} \quad \Leftrightarrow \quad c = \beta |\tilde{c}|^{1/\alpha} e^{i\angle(\tilde{c})} \tag{8}$$

where we have chosen $\alpha = 0.5$ and $\beta = 3$ empirically. The forward and backward Gaussian processes, as well as the input and output of the score-based DNN model, are then formulated and applied within this transformed domain.

3.3. Training task

We adopt the training strategy of *denoising score matching* [29], which trains a model to approximate the score $\nabla_{x_t} \log p_t(x_t)$ by estimating the Gaussian noise added by the forward process at time t. This training strategy can be efficiently realized by sampling t uniformly from $[0, 1]$ and then sampling x_t according to (5) for each data point in the training batch. The training task is to learn parameters θ that minimize the following term, where $x_0 \sim p_0(x), x_t \sim p_{0t}(x_t|x_0, y)$:

$$\mathbb{E}_{t,x_0,x_t|x_0} \left[\| s_\theta(x_t, t, y) - \nabla_{x_t} \log p_{0t}(x_t|x_0, y) \|_2^2 \right] \tag{9}$$

We note that in the baseline work, CDiffuSE [14, eq. (21)], the loss function includes a term $(y - x_0)$, i.e., the network is in part trained to remove the environmental noise directly in each step. In contrast, our objective function in (9) does not task the DNN with estimating a portion of the environmental noise n, and $y = x_0 + n$ enters (9) only as an input to the score-based model, functioning as a conditioning signal. Our model is thus trained to only estimate the Gaussian noise that is artificially added during the forward process. We therefore argue that the CDiffuSE training task has significant discriminative characteristics, whereas ours remains purely generative in nature.

3.4. Speech enhancement procedure

To proceed with speech enhancement, an initial complex spectrogram x_1 is determined by sampling from the corresponding prior distribution, with mean y and standard deviation $\sigma(1)$:

$$x_1 \sim \mathcal{N}_{\mathbb{C}}(y, \sigma(1)^2 \mathbf{I}) \tag{10}$$

We then choose a reverse sampling method to start from x_1 and solve the reverse SDE, from $t = 1$ up until a small $t_\varepsilon \approx 0$ to avoid numerical issues close to $t = 0$, see [18]. There are various possible choices of reverse sampling methods. Song et al. [18] show impressive performance of so-called *Predictor-Corrector samplers*, which we therefore adopt as well. We use the specific combination of *reverse diffusion sampling* as the predictor and *annealed Langevin dynamics* as the corrector, as in [18, Algorithm 2]. The corrector requires a so-called SNR parameter r, which we empirically set to $r = 0.33$. We use one corrector step per iteration, at 50 total iterations. The enhanced audio is finally retrieved by applying the backward transformation (8) and a subsequent inverse STFT.

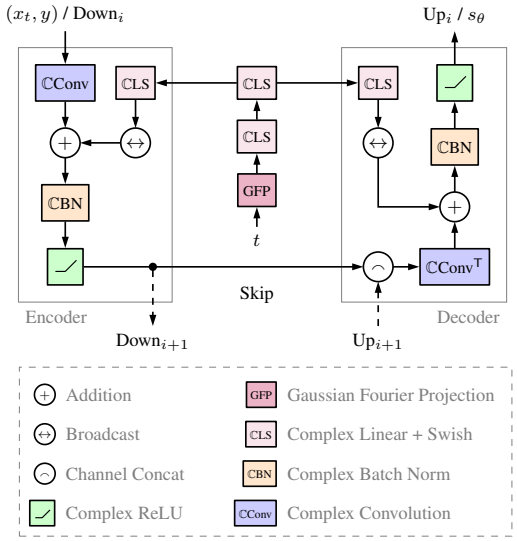


Figure 2: *Architecture of each encoder/decoder pair in our U-Net DNN. Activation functions and norms are applied to the real and imaginary parts separately, while the complex convolution and complex linear layers use the natural complex algebra.*

4. Experimental setup

4.1. Neural network architecture

For our experiments in this paper, we construct a complex-valued U-Net architecture based on blocks shown schematically in Fig. 2 and parameterized as listed in Table 1, resulting in 3.56M parameters overall. We adapt the *Deep Complex U-Net* (DCUNet) architecture [30] in several ways for our task: **(1)** We insert *time-embedding layers* into all encoder and decoder blocks, providing the DNN with information about the time-step t. We encode t via 128 random Fourier feature embeddings [31], which are passed through complex-valued affine layers and activations in each block. These embeddings are broadcasted over and added to each respective channel of the layer output, as in previous SGM-based work [17,18,20]. **(2)** We add multiple non-strided encoder/decoder pairs at the top of the U-Net to increase the DNN's capacity for finer, non-downsampled features. **(3)** We add exponentially increasing dilations in the frequency axis to the lower layers. This enlarges the receptive field in the frequency direction, to help the DNN differentiate between spectral characteristics encountered in speech without affecting the amount of required computation. **(4)** We change the network to have one complex-valued output channel estimating the score, and two complex-valued input channels (x_t, y), so that it has access to the information contained in the noisy speech y. **(5)** We replace all 3×3 kernels by 4×4 kernels, to avoid checkerboard patterns arising from the kernel size not being divisible by the stride in the transposed convolutions [32].

4.2. Dataset

We use the standardized VoiceBank-DEMAND [33, 34] dataset for training and testing, as was done in the baseline work (DiffuSE, [13]). We normalize the pairs of clean and noisy audio (x_0, y) by the maximum absolute value of x_0. We then convert

Table 1: *Encoder and decoder parameters of our modified DCUNet architecture. C_i / C_o are input / output channels, K is the kernel size, S is the stride, and D is the dilation. Tuples indicate the axes (frequency, time).*

Depth	**1**	**2**	**3**	**4**	**5**	**6**
C_i	2	32	32	32	64	128
C_o	32	32	32	64	128	256
K	(4,4)	(4,4)	(4,4)	(4,4)	(4,4)	(4,4)
S	(1,1)	(1,1)	(1,1)	(2,1)	(2,2)	(2,2)
D	(1,1)	(1,1)	(1,1)	(2,1)	(4,1)	(8,1)

each input into the complex-valued one-sided STFT representation, using an DFT length of $F = 512$, a hop length of 128 (i.e., 75% overlap) and a periodic Hann window. We randomly crop each spectrogram to 256 STFT time frames during each epoch.

4.3. Training procedure and hyperparameters

We train our DNN for 325 epochs, using the Adam optimizer [35] with a learning rate of 10^{-4} and a batch size of 32. We parameterize our SDE (3) as follows: $\gamma = 1.5, \sigma_{\text{min}} = 0.05, \sigma_{\text{max}} = 0.5, t_\varepsilon = 0.03$. We track an exponential moving average of the DNN weights with a decay of 0.999, to be used for sampling [36]. We also train two baseline models [13, 14] to compare against, using publicly available code.

4.4. Evaluation

For evaluation, we follow the speech enhancement procedure described in Section 3.4. We report the scale-invariant signal-to-distortion ratio (SI-SDR), signal-to-interference ratio (SI-SIR) and signal-to-artifacts ratio (SI-SAR) [37], comparing against the noisy speech y and the baseline method. We avoid the use of the Perceptual Evaluation of Speech Quality (PESQ) measure, since the P.862.3 standard [38] states that "there should be a minimum of 3.2s active speech in the reference", which does not hold for the majority of audio files in the VoiceBank-DEMAND [33] dataset used by us and (C)DiffuSE [13, 14]. We provide listening examples in order to assess the perceptual quality of the compared methods[1].

5. Results

In Table 2, we compare the performance of our proposed method (SGMSE) with DiffuSE [13], CDiffuSE [14] and the noisy mixture. The results from CDiffuSE, for which executable code was not available at the time of submission, were added to the table after acceptance. We report the mean results and their 95% confidence interval on the test set. We also report values for $m = (0.8\hat{x} + 0.2y)$, as suggested and used in [13, 14].

Our proposed method shows an improvement in SI-SDR of 4.6 dB over DiffuSE and 3.0 dB over CDiffuSE when comparing the raw model output. We can see that the three compared methods follow a trend of increasing SI-SAR and decreasing SI-SIR, with our method achieving the most favorable balance between the two metrics, resulting in the best SI-SDR. Arguably, our method also achieves more natural-sounding speech due to the lower amount of artifacts, which we confirmed by informal listening. Audio examples and code are available online[1].

This qualitative behavior is corroborated by Fig. 3, where we show the power spectrograms of an example utterance (clean and noisy) and the corresponding estimates of our method and CDiffuSE. Both methods are able to remove the environmental noise, which is clearly visible in the specified dynamic range of 50 dB. Furthermore, it can be seen that our method preserves the high frequencies of the fricatives better than CDiffuSE. Interestingly, CDiffuSE removes too much energy around the formants, whereas our method maintains the natural structure.

[1] https://uhh.de/inf-sp-sgmse

Table 2: *Average performance of our method (SGMSE) and DiffuSE [13], comparing their raw output to the noisy speech mixture. Best values in each column are bold. Values for $m = (0.8\hat{x} + 0.2y)$, as used in [13, 14], are listed in gray.*

Model	**SI-SDR**	**SI-SIR**	**SI-SAR**
Mixture	$8.4 \pm .38$	$8.4 \pm .38$	∞
DiffuSE [13]	$10.5 \pm .14$	$\mathbf{30.0 \pm .71}$	$10.8 \pm .11$
CDiffuSE [14]	$12.1 \pm .10$	$28.2 \pm .36$	$12.3 \pm .09$
SGMSE	$\mathbf{15.1 \pm .27}$	$24.9 \pm .42$	$\mathbf{15.7 \pm .25}$
DiffuSE$_m$ [13]	$11.4 \pm .18$	$19.1 \pm .46$	$13.0 \pm .11$
CDiffuSE$_m$ [14]	$12.6 \pm .15$	$18.8 \pm .34$	$14.4 \pm .09$
SGMSE$_m$	$14.5 \pm .29$	$17.9 \pm .36$	$17.7 \pm .25$

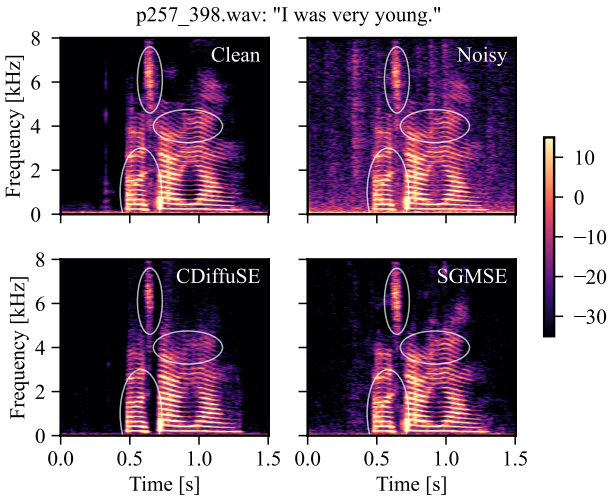


Figure 3: *Power spectrograms for an example audio file from VoiceBank-DEMAND. As can be seen from the highlighted regions, our approach (SGMSE) exhibits fewer voice distortions than CDiffuSE [14] at some expense of noise reduction.*

6. Conclusion

In this work, we have designed a novel stochastic process for a score-based generative modeling approach to speech enhancement in the complex STFT domain, based on the formalism of SDEs. To our knowledge, this is the first work to apply SGMs in the complex time-frequency domain and, after CDiffuSE [14], the second theoretically principled approach to use SGMs specifically for speech enhancement. Our approach exhibits more natural sounding reconstructions with fewer artifacts than previous methods at some expense of noise removal, which together leads to an SI-SDR improvement of about 5 and 3 dB over DiffuSE [13] and CDiffuSE [14], respectively. Further investigation into other SDEs, data representations and DNN architectures may prove to be a fruitful research avenue for speech enhancement using generative models. An extended journal article with additional analysis and improved performance is currently in preparation [39].

7. References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art*. Morgan & Claypool, 2013.

[2] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. John Wiley & Sons, 2018.

[3] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 39–48, 2020.

[4] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *ISCA Interspeech*, pp. 3642–3646, 2017.

[5] D. Baby and S. Verhulst, "SERGAN: Speech enhancement using relativistic generative adversarial networks with gradient penalty," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 106–110, 2019.

[6] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 716–720, 2018.

[7] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, pp. 1–6, 2018.

[8] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks." *ISCA Interspeech*, pp. 4516–4520, 2020.

[9] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," *IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics (WASPAA)*, pp. 126–130, 2021.

[10] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 676–680, 2021.

[11] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational auto-encoders," *arXiv preprint arXiv:2106.12271*, 2021.

[12] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE Trans. on Audio, Speech, and Language Proc. (TASLP)*, vol. 28, pp. 1104–1117, 2020.

[13] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," *Asia-Pacific Signal and Inf. Proc. Association Annual'Summit and Conf. (APSIPA ASC)*, pp. 659–666, 2021.

[14] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional Diffusion Probabilistic Model for Speech Enhancement," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2022.

[15] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder." *ISCA Interspeech*, pp. 2437–2441, 2020.

[16] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 32, 2019.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[19] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 34, 2021.

[20] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[21] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[22] G. Parisi, "Correlation functions and computer simulations," *Nuclear Physics B*, vol. 180, no. 3, pp. 378–384, 1981.

[23] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

[24] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[25] T. Chakraborty and M. Kearns, "Market making and mean reversion," *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 307–314, 2011.

[26] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," *Physical review*, vol. 36, no. 5, p. 823, 1930.

[27] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," *Int. Workshop on Acoustic Echo and Noise Control*, 2010.

[28] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," *Int. Conf. on Telecom. and Signal Proc. (TSP)*, pp. 72–76, 2021.

[29] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[30] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware Speech Enhancement with Deep Complex U-Net," *arXiv preprint arXiv:1903.03107*, 2019.

[31] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 7537–7547, 2020.

[32] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[33] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech." *ISCA Speech Synthesis Workshop (SSW)*, pp. 146–152, 2016.

[34] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. on Learning Representations (ICLR)*, 2015.

[36] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, pp. 12 438–12 448, 2020.

[37] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, pp. 626–630, 2019.

[38] ITU-T Rec. P.862.3, "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," *Int. Telecom. Union (ITU)*, 2007.

[39] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *(in preparation)*, 2022.

A.3 Causal processing for restoration of speech communication artifacts

A.3.1 Publication: Speech Signal Improvement Using Causal Generative Diffusion Models

Abstract

In this paper, we present a causal speech signal improvement system that is designed to handle different types of distortions. The method is based on a generative diffusion model which has been shown to work well in scenarios with missing data and non-linear corruptions. To guarantee causal processing, we modify the network architecture of our previous work and replace global normalization with causal adaptive gain control. We generate diverse training data containing a broad range of distortions. This work was performed in the context of an "ICASSP Signal Processing Grand Challenge" and submitted to the non-real-time track of the "Speech Signal Improvement Challenge 2023", where it was ranked fifth.

Reference

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, Tal Peer, Timo Gerkmann, "Speech Signal Improvement Using Causal Generative Diffusion Models," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Rhodes, Greece, Jun. 2023. DOI: 10.1109/ICASSP49357.2023.10095126

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2023 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

This paper presents the challenge submission for the *Signal Improvement Challenge* 2023. Julius Richter led the collaborative effort, while Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Tal Peer contributed to developing the algorithm, training framework, and data preparation. Timo Gerkmann provided valuable insights into experimental validation, mathematical derivations, and manuscript review.

SPEECH SIGNAL IMPROVEMENT USING CAUSAL GENERATIVE DIFFUSION MODELS

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, Tal Peer, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

ABSTRACT

In this paper, we present a causal speech signal improvement system that is designed to handle different types of distortions. The method is based on a generative diffusion model which has been shown to work well in scenarios with missing data and non-linear corruptions. To guarantee causal processing, we modify the network architecture of our previous work and replace global normalization with causal adaptive gain control. We generate diverse training data containing a broad range of distortions. This work was performed in the context of an "ICASSP Signal Processing Grand Challenge" and submitted to the non-real-time track of the "Speech Signal Improvement Challenge 2023", where it was ranked fifth.

Index Terms— Speech signal improvement, universal speech enhancement, diffusion models, causal processing

1. INTRODUCTION

High-quality voice communication requires clear audio and natural-sounding speech. However, there are numerous factors that can degrade speech signals including background noise, room acoustics, transmission errors, limited bandwidth, and codec artifacts. Prior works on improving speech signals have typically studied each type of distortion separately. However, there has been some recent interest in developing universal approaches that address a broader range of distortions. These approaches typically use generative modeling, which works particularly well in scenarios with missing data and non-linear corruptions [1].

In this paper, we present our diffusion-based speech enhancement method submitted to the non-real-time track of the "Speech Signal Improvement Challenge 2023" [2] as part of the "ICASSP Signal Processing Grand Challenges". The proposed model extends our previous work [3], incorporating significant modifications in the network architecture to meet the causality requirement and to output super wideband speech. Furthermore, we devise a data corruption approach to generate diverse training data resembling distortions observed in the blind data. Strong variations in loudness are compensated by using causal adaptive gain control.

In the challenge's subjective test, our proposed method yields a final score of 0.445 using ITU-T P.863.2 and P.804 with mean opinion scores (MOSs) of 2.570 (Overall), 2.998 (Signal), 3.765 (Noise), 3.330 (Coloration), 3.674 (Discontinuity), 3.435 (Loudness), and 4.241 (Reverberation). In the subjective test based on ITU-T P.835 our method yields a final score of 0.495 with MOSs of 2.842 (Overall), 3.119 (Signal), and 3.682 (Background). It is interesting to note that, as our approach is generative in nature, it is capable of achieving excellent performance for moderately distorted inputs, while it may generate phonetic confusions and insertions if the input distortions are too strong.

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169), DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380.

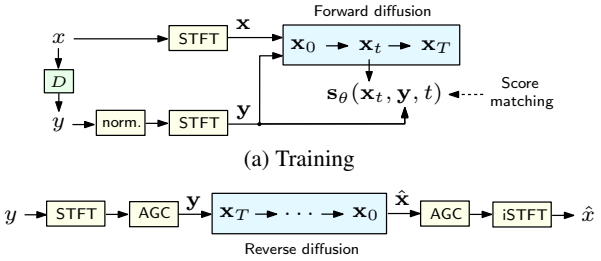


Fig. 1. Proposed system: (a) At training, a corruption model D generates y from clean speech x. The forward diffusion moves from the clean spectrogram $\mathbf{x}_0$ to the corrupted $\mathbf{y}$, while Gaussian noise is gradually added. The score model $\mathbf{s}_\theta$ is learned using score matching. (b) At inference, the data is normalized with causal adaptive gain control (AGC) and the reverse diffusion uses the trained score model.

2. PROPOSED SYSTEM

Fig. 1 shows an overview of our proposed system for speech signal improvement at training and inference time. The diffusion process (Sec. 2.1) is the core of the method and is accompanied by other processing blocks including the short-time Fourier transform (STFT), causal automatic gain control (AGC) (Sec. 2.3), and a data corruption model D used to simulate various distortion types (Sec. 3.1).

2.1. Diffusion process

The diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ for speech enhancement is essentially the same as in our previous works [3], [4], where $\mathbf{x}_t$ is the state of the process at time step t. During training, the forward process moves from clean speech $\mathbf{x}_0 := \mathbf{x}$ to corrupted speech $\mathbf{y}$, while increasing amounts of Gaussian noise are gradually added. At inference, the corresponding reverse process [5] is used to progressively remove the corruption and therefore generate an estimate of $\mathbf{x}$ starting from $\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma_t^2)$. This reverse process involves the *score function* of $\mathbf{x}_t$, i.e., the gradient of its log-probability. Functioning as a prior for clean speech, it is unavailable at inference time and is thus approximated by a trained deep neural network (DNN) $\mathbf{s}_\theta$ called the *score model*. We make no further adaptations to this process as it is defined for every STFT bin independently, and is thus inherently causal as long as $\mathbf{s}_\theta$ is implemented with a causal network architecture.

2.2. Network architecture

We use a modified version of NCSN++ [5] for score estimation. The network is an encoder-decoder architecture based on 2D convolutions,

taking complex spectrograms $\mathbf{x}_t$ and $\mathbf{y}$ and the process time t as input. Real and imaginary parts are considered as separate channels and the convolutions are performed over time and frequency.

We apply the following modifications to the architecture to meet the causality constraint: **(1)** Padding in the 2D convolutions is modified so that the convolution along the time-dimension is causal; **(2)** Batch normalization is replaced with cumulative group normalization, aggregating statistics recursively; **(3)** Downsampling in the time dimension is performed with strided convolutions and corresponding upsampling with transposed strided convolutions. Up- and downsampling in the frequency dimension are realized with finite impulse response filters, as in [3]. **(4)** All attention layers as well as the progressive growing path are removed.

2.3. Automatic gain control

To match the unit-scale training condition of the score model, i.e., normalization of corrupted speech y (see Fig. 1a), we use a causal AGC system before feeding the mixture to the diffusion process, and again after enhancement to maximize loudness, as part of the signal improvement task (see Fig. 1b). To this end, we recursively track the maximum value per magnitude frame averaged over the frequency bins to normalize the spectrogram in a causal manner. We start tracking when speech activity is first detected, using a voice activity detection method based on a causal speech presence probability estimator [6]. The speech probability is fed through an ideal low-pass filter to avoid having high-frequency noise bursts produce false positives. Voice activity is then assumed if the speech presence probability is higher than a threshold $\tau = 0.8$, for a duration of 100ms. When discovering a larger maximum than the previous one, we smooth the normalization with an exponential ramp going from the old value to the new one. Finally, to avoid clipping, we use a causal compressor from the `pedalboard` library[1].

3. EXPERIMENTAL SETUP

3.1. Dataset

We use the VCTK corpus [7] as the clean speech dataset and resample all utterances from 48kHz to our processing sampling frequency 32kHz. Using the `audiomentations` library[2], we simulate several corruptions observed in the blind data, namely stationary and non-stationary noise, reverberation, clipping, gain reduction, packet loss and lossy speech coding (the last two implemented by us). For each clean utterance, a random corruption chain is chosen among plausible candidates, e.g. $\{\text{Reverb} \rightarrow \text{Noise} \rightarrow \text{PacketLoss}\}$, with the parameters of each corruption chosen randomly as well. We use the QUT corpus [8] as the environmental noise dataset and take room impulse responses from the DNS challenge [9] for reverberation.

3.2. Hyperparameters and training configuration

All processing is performed at $f_s = 32$kHz and we upsample the processed files to the original 48kHz frequency. We use an STFT with a 638-point Hann window and 160-point hop, which guarantees the global latency to be below 20ms, as we use purely causal processing. We set the lowest two frequency bins of the spectrogram to zero, to remove the DC offset and low frequency noise. The diffusion process hyperparameters are identical to those in [3]. We train the score model $\mathbf{s}_\theta$ with the denoising score matching objective [5] using the Adam optimizer with mini-batch size of 8 and exponential moving average over the parameters with a factor of 0.999. Training takes around two days on two NVIDIA RTX A6000 GPUs.

[1] `https://github.com/spotify/pedalboard`

[2] `https://github.com/iver56/audiomentations`

4. RESULTS

In addition to the challenge's subjective evaluation [2], we use DNSMOS P.835 [10] to evaluate our method on the 500 files in the blind set. Fig. 2 depicts the histograms of DNSMOS scores: (a), (b) speech quality (SIG) and overall quality (OVRL) of the improved files compared to the corrupted files; (c) SIG with and without AGC.

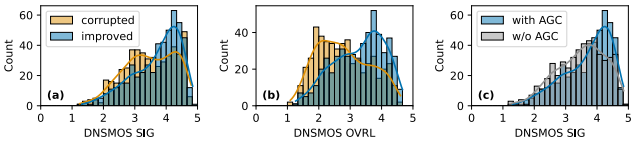


Fig. 2. Speech signal improvement results obtained with DNSMOS.

Computational complexity: The network has about 55.7M parameters and the time to infer a frame on a CPU (Intel Core i7-7800X @ 3.50GHz) takes 0.89s. Please note that the model is designed to run on a GPU, on which the inference time is orders of magnitude faster (0.02 s / frame with an NVIDIA GeForce RTX 2080 Ti).

5. CONCLUSION

In this work, we have built upon our previous work on diffusion-based speech enhancement, with the novel contribution in making the system causal and training the model on different distortion types. The proposed method was ranked fifth in the non-real-time track of the "Speech Signal Improvement Challenge 2023".

6. REFERENCES

[1] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," *ICASSP*, 2023.

[2] R. Cutler, A. Saabas, B. Naderi, N.-C. Ristea, S. Braun, and S. Branets, "ICASSP 2023 speech signal improvement challenge," 2023. [Online]. Available: `https://arxiv.org/abs/2303.06566`.

[3] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *arXiv preprint*, 2022.

[4] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," *Interspeech*, 2022.

[5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021.

[6] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," *WASPAA*, 2011.

[7] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2019.

[8] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms," in *Interspeech*, 2010.

[9] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 deep noise suppression challenge," in *ICASSP*, 2022.

[10] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2022.

A.4 Extension to audio-visual speech enhancement

A.4.1 Publication: LipDiffuser: Lip-to-Speech Generation with Conditional Diffusion Models

Abstract

We present *LipDiffuser*, a conditional diffusion model for lip-to-speech generation synthesizing natural and intelligible speech directly from silent video recordings. Our approach leverages the magnitude-preserving ablated diffusion model (MP-ADM) architecture as a denoiser model. To effectively condition the model, we incorporate visual features using magnitude-preserving feature-wise linear modulation (MP-FiLM) alongside speaker embeddings. A neural vocoder then reconstructs the speech waveform from the generated mel spectrograms. Evaluations on LRS3-CHiME3 and NTCD-TIMIT demonstrate that *LipDiffuser* outperforms existing lip-to-speech baselines in perceptual speech quality and speaker similarity, while remaining competitive in downstream automatic speech recognition (ASR) and audio-visual synchrony. Extensive ablation studies and cross-dataset evaluation confirm the effectiveness and generalization capabilities of our approach.

Reference

Reference

Danilo de Oliveira, Julius Richter, Tal Peer, Timo Gerkmann, "LipDiffuser: Lip-to-Speech Generation with Conditional Diffusion Models", submitted to *Advances in Neural Information Processing Systems*, 2025.

An up-to-date version is available on arXiv.

Author's Contributions

Danilo de Oliveira and Julius Richter collaboratively designed the LipDiffuser algorithm, conducted all training and evaluation experiments, and jointly wrote the manuscript. Tal Peer conducted the listening experiment, contributed to the evaluation code, and wrote significant portions of the manuscript. Timo Gerkmann provided insights into the experimental validation and mathematical derivations and participated in reviewing the manuscript.

LipDiffuser: Lip-to-Speech Generation with Conditional Diffusion Models

Danilo de Oliveira* **Julius Richter*** **Tal Peer** **Timo Gerkmann**
Signal Processing Group
University of Hamburg
{danilo.oliveira, julius.richter, tal.peer, timo.gerkmann}@uni-hamburg.de

Abstract

We present *LipDiffuser*, a conditional diffusion model for lip-to-speech generation synthesizing natural and intelligible speech directly from silent video recordings. Our approach leverages the magnitude-preserving ablated diffusion model (MP-ADM) architecture as a denoiser model. To effectively condition the model, we incorporate visual features using magnitude-preserving feature-wise linear modulation (MP-FiLM) alongside speaker embeddings. A neural vocoder then reconstructs the speech waveform from the generated mel spectrograms. Evaluations on LRS3-CHiME3 and NTCD-TIMIT demonstrate that *LipDiffuser* outperforms existing lip-to-speech baselines in perceptual speech quality and speaker similarity, while remaining competitive in downstream automatic speech recognition (ASR) and audio-visual synchrony. Extensive ablation studies and cross-dataset evaluation confirm the effectiveness and generalization capabilities of our approach.

1 Introduction

The task of conditional temporal data generation consists of creating new data sequences whose evolution over time is guided by corresponding conditioning data (Yang et al., 2024), where the generated data and conditioning information are often of different modalities. Generative models, such as conditional diffusion models (Rombach et al., 2022), are particularly well suited for this task as they possess a remarkable capability to learn complex posterior distributions over possible outputs given observed conditioning data. This enables the synthesis of realistic and diverse outcomes that reflect the variability present in many real-world nonstationary data streams. Such approaches are especially relevant in multimodal applications—for example, generating synchronized video frames from an audio signal (Zhang et al., 2023), or synthesizing realistic audio from corresponding visual cues, such as lip movements or gestural data (Viertola et al., 2025). Additionally, these models have shown promise in medical applications, where physiological temporal signals such as electroencephalography (EEG) can be generated or imputed based on other clinical time series (Torma and Szegletes, 2025). The main challenges in conditional temporal data generation include aligning the temporal data of different modalities and possibly different feature rates, as well as learning the intricate relationships between modalities in a robust and generalizable manner to ensure seamless integration and synchronization of the end result.

Diffusion models generate data by gradually transforming a noise sample into a sample from a given data distribution (Sohl-Dickstein et al., 2015; Ho et al., 2020). This diffusion process is often formulated as a stochastic differential equation (SDE) (Song et al., 2021), which defines the forward process describing the transition from the data distribution to a Gaussian noise distribution. While unconditional diffusion models generate a sample from the data distribution, as captured during training, conditional diffusion models add the ability to generate a sample from a conditional

*Authors contributed equally to this work.

Preprint. Under review.

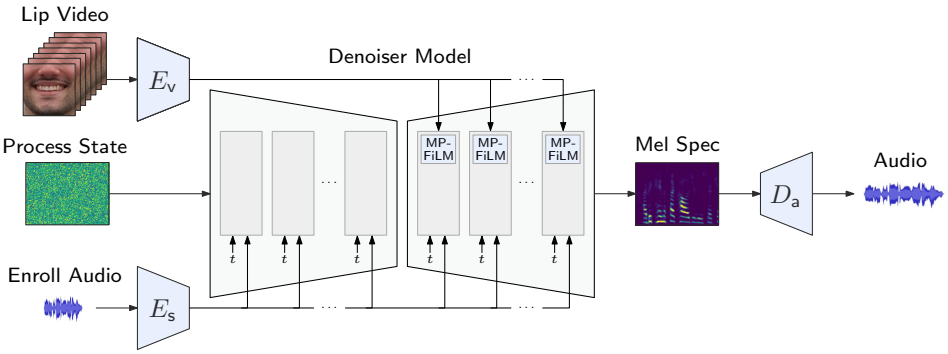


Figure 1: *LipDiffuser*: The model comprises a video encoder E_v, a speaker encoder E_s, a denoiser model, and a neural vocoder D_a. The denoiser model takes as input the process state, video features extracted by the video encoder, the current process time t, and the speaker embedding from the speaker encoder, and predicts a mel spectrogram. The video features are fused within the denoiser's decoder using magnitude-preserving feature-wise linear modulation (MP-FiLM) layers. After reverse diffusion, the neural vocoder synthesizes an audio signal from the generated mel spectrogram.

distribution, where the conditioning variable instructs the model to generate a sample that matches some other information. For example, an unconditional speech model can generate audio that resembles natural speech in its acoustic properties, but the resulting signal typically lacks any coherent semantic content (Kong et al., 2021). By conditioning the model on additional information, such as a text transcript, the generated speech can be guided to retain natural-sounding characteristics and accurately convey the intended meaning specified by the conditioning input (Chen et al., 2021; Kong et al., 2021).

In this work, we develop a conditional diffusion model and apply it to the lip-to-speech task, which involves synthesizing accurate speech signals based on lip movement data captured in video. Real-world video recordings often suffer from a very low-quality audio track. This can result from a noisy recording environment, inadequate equipment, or storage media corruption. In severe cases, the audio track may be completely unusable or even missing. In the context of human speech, visual and auditory information both play a role in the ability to understand what a person is saying (McGurk and MacDonald, 1976). However, these two modalities are not equally important; while missing visual information can lead to reduced intelligibility, the lack of audio is practically destructive to speech communication (Dodd, 1977), except for individuals highly trained in lip reading (Dodd and Campbell, 1987).

Lip-to-speech is the process of generating a natural-sounding speech signal corresponding to a given silent video of a person speaking. Lip-to-speech systems thus restore the ability to understand human speech even in scenarios where the audio track is missing or highly distorted. A high-quality lip-to-speech system must ensure that the generated speech signal is intelligible, synchronized with the speaker's lip movements, and sounds natural. Ideally, the speech signal must also match the speaker's characteristics, including age, gender, and accent. A major challenge for lip-to-speech techniques stems from the one-to-many mapping between visemes and phonemes (Richter et al., 2022, Table 1); different phonemes can correspond to the same sequence of lip movements, leading to ambiguity. This issue can be alleviated by including a longer temporal context in the generation process (Martinez et al., 2020). Another difficulty is alignment and synchronization of the visual cues from the lip movements with the generated audio, considering the large differences in feature rates between the visual and audio signals (e.g., 25 Hz for video compared to 48 kHz for audio).

We introduce *LipDiffuser* (Figure 1), a lip-to-speech method based on a conditional diffusion model. The proposed method comprises a video encoder, a speaker encoder, a denoiser model, and a neural vocoder. For the denoiser model, we utilize the magnitude-preserving ablated diffusion model (MP-ADM) architecture (Karras et al., 2024), which builds upon the ablated diffusion model (ADM) architecture (Dhariwal and Nichol, 2021) by introducing a series of modifications that significantly enhance output quality while retaining the overall structure. During diffusion sampling, the denoiser model takes as input the current process state, video features extracted by the video encoder, the current diffusion timestep, and a speaker embedding from the speaker encoder, and predicts a mel

spectrogram. To effectively incorporate visual cues, we propose integrating video features into the denoising network's decoder using magnitude-preserving feature-wise linear modulation (MP-FiLM). After the reverse diffusion process, the neural vocoder synthesizes the final audio waveform from the generated mel spectrogram.

We conduct our experiments on two audio-visual speech datasets: *LRS3* (Afouras et al., 2018b) and *TCD-TIMIT* (Harte and Gillen, 2015). To enable comparison with audio-visual speech enhancement methods, we create the *LRS3-CHiME3* dataset by mixing speech from *LRS3* with noise files from *CHiME3* (Barker et al., 2015) and use the *NTCD-TIMIT* benchmark (Abdelaziz et al., 2017). Our findings suggest that lip-to-speech approaches can offer advantages over audio-visual speech enhancement, particularly when the input audio is severely degraded (input signal-to-noise ratio (SNR) < -5 dB). Specifically, we show that our proposed *LipDiffuser* model outperforms all lip-to-speech baselines in terms of speech quality and speaker similarity, while remaining competitive in automatic speech recognition (ASR) performance and synchrony. Furthermore, cross-dataset evaluations demonstrate the strong generalization capabilities of our method in both speech quality and speaker similarity. Finally, we conduct extensive ablation studies on data selection, input representation, and visual features to validate various design choices. Audio-visual examples, code, and pretrained checkpoints are available online[1].

2 Related work

2.1 Lip-to-speech

A range of learning-based lip-to-speech methods has been introduced in recent years. These approaches vary regarding the underlying model architectures and learning paradigms, the types of visual and additional information leveraged, and the strategies used to guide the process during training and inference.

Kim et al. (2023) present *Lip2Speech*, a lip-to-speech method that includes a visual front-end, a *Conformer* model (Gulati et al., 2020) for capturing temporal relationships, and a Mel-spectrogram generator conditioned on speaker embeddings. They propose a multi-task learning approach that consists of: (a) a connectionist temporal classification (CTC) (Graves et al., 2006) loss between the visual representations and the text targets; (b) text prediction utilizing a pretrained ASR model; and (c) a Mel spectrogram reconstruction loss. Finally, they use the Griffin-Lim algorithm (Griffin and Lim, 1984) to convert the Mel spectrogram into a waveform.

Choi et al. (2023b) propose a lip-to-speech model that employs speech representations from an self-supervised learning (SSL) model. In a multi-task learning framework, their model predicts both mel-spectrograms and speech representations derived from a pretrained SSL model. In addition, they develop a vocoder that generates the waveform with the mel-spectrogram and speech units. During training, they augment the input mel spectrograms with blur and noise to help the vocoder learn to generate waveforms from the generated mel spectrograms by referencing the speech units. In a follow-up work, Choi et al. (2023a) present *DiffV2S*, a diffusion-based lip-to-speech method based on a speaker embedding derived from the visual input. This approach addresses the challenge of inferring the speaker's characteristics without audio.

Yemini et al. (2024) introduce *LipVoicer*, a lip-to-speech method that uses a diffusion model conditioned on lip video and incorporates a classifier-guidance mechanism (Dhariwal and Nichol, 2021) based on text. A pretrained ASR model (Burchi and Timofte, 2023) acts as the classifier, whereas a pretrained lip-reading network (Ma et al., 2023) predicts the spoken text from the silent video, offering guidance for the score model. They train the diffusion model to generate Mel spectrograms and utilize *DiffWave* (Kong et al., 2021) as a neural vocoder to produce the raw audio.

In contrast to the above-mentioned methods, our *LipDiffuser* model achieves aligned conditioning by interpolating the video features to match the temporal resolution of the audio features within the denoiser model, and fusing them through MP-FiLM layers.

[1]To be released at the time of acceptance.

2.2 Audio-only and audio-visual speech enhancement

Speech enhancement refers to the process of improving the quality and intelligibility of a speech signal by reducing noise and reverberation (Vincent et al., 2018). Numerous computational methods have been developed for this task, most operating in the time-frequency domain using the short-time Fourier transform (STFT). Machine learning-based speech enhancement methods are commonly divided into predictive and generative learning paradigms (Lemercier et al., 2025), and they vary in the types of information leveraged to estimate the clean speech signal. While practical approaches often exploit multi-channel input from microphone arrays to utilize spatial information for improved performance in adverse acoustic environments (Tesch and Gerkmann, 2022), the present work focuses on single-channel speech enhancement, relying solely on audio captured from a single microphone.

Human perception naturally integrates auditory and visual information to enhance speech understanding, particularly in challenging listening environments (Stein, 1993; Shams and Seitz, 2008). Neuroimaging and behavioral studies have shown that visual cues, such as a speaker's lip movements, can significantly improve both speech intelligibility and localization (Calvert et al., 1997; Sumby and Pollack, 1954; McDonald et al., 2000). This inherent crossmodal integration motivates the development of audio-visual speech enhancement methods, aiming to replicate these perceptual advantages in machine learning systems (Girin et al., 1995). By leveraging both audio and visual modalities, modern systems seek to enhance speech quality and intelligibility more robustly than is possible with audio alone (Ephrat et al., 2018; Afouras et al., 2018a; Richter et al., 2023a).

2.3 Diffusion models

Diffusion models break down data generation into more manageable denoising tasks by progressively removing noise and refining the data (Sohl-Dickstein et al., 2015; Ho et al., 2020). The forward process used for training can be described by a SDE, which transforms the target distribution into a terminating Gaussian distribution, which serves as the initial distribution for sampling (Song et al., 2021). Reversing this process in time yields the reverse process, which is characterized by a different SDE that is numerically integrated during inference using a denoiser or score model, depending on the specifics of the sampling procedure.

The denoiser model is a neural network trained to distinguish the target data from Gaussian noise introduced during the forward process. Expressive denoiser models are often implemented using the U-Net architecture (Ronneberger et al., 2015), augmented with self-attention layers (Vaswani et al., 2017). To keep network operations within an appropriate range and prevent large variations in gradient magnitudes, Karras et al. (2022) proposed using time-dependent scaling of the network's inputs and outputs, combined with a skip connection of the process state, known as preconditioning. In a subsequent study, Karras et al. Karras et al. (2024) observed that uncontrolled changes in the magnitude and imbalances in both network activations and weights persisted throughout the training process, even when preconditioning was applied. To address this issue, they redesigned the network layers of the ADM U-Net architecture (Dhariwal and Nichol, 2021) to better preserve the expected magnitudes of activations, weights, and parameter updates. This adjustment led to improved performance in image generation tasks.

Conditional diffusion models extend the capabilities of standard diffusion models by incorporating additional information, such as class labels, textual descriptions, or other context, into the data generation process. This conditioning enables the model to generate data that adheres to specific requirements or mimics particular patterns, significantly enhancing its practical applicability (Dhariwal and Nichol, 2021). Architecturally, conditioning is often achieved through cross-attention layers (Rombach et al., 2022), or by simply concatenating or adding conditional features (Richter et al., 2023a). These approaches enable the model to focus on relevant aspects of the conditioning input throughout the denoising process.

3 Method

Figure 1 shows an overview of our model called *LipDiffuser*. The model consists of a video encoder E_v, a speaker encoder E_s, a denoiser model D_θ parameterized by θ, and a neural vocoder D_a. The denoiser model receives as inputs the process state $\mathbf{x}_t \in \mathbb{R}^{n_a}$, a speaker embedding $\mathbf{s} \in \mathbb{R}^{n_s}$ from the

speaker encoder, video features $\mathbf{v} \in \mathbb{R}^{n_v}$ from the video encoder, and the process time t, and predicts the mel spectrogram $\hat{\mathbf{x}} \in \mathbb{R}^{n_a}$.

3.1 Training objective

We use denoising score matching (Vincent, 2011) as our training objective. For a process time t, the training objective is defined as minimizing

$$\mathcal{J}(D_\theta, t) = \mathbb{E}_{\mathbf{x},\mathbf{n},\mathbf{s},\mathbf{v}} \left[\|D_\theta(\mathbf{x} + \mathbf{n}, \mathbf{s}, \mathbf{v}, t) - \mathbf{x}\|_2^2 \right], \tag{1}$$

where $(\mathbf{x}, \mathbf{s}, \mathbf{v}) \sim p_{\text{data}}(\mathbf{x}, \mathbf{s}, \mathbf{v})$ is sampled from the dataset, and $\mathbf{n} \sim \mathcal{N}(0, \sigma(t)\mathbf{I})$ is a random Gaussian vector with time-dependent standard deviation $\sigma(t)$. The overall training objective is defined as a weighted expectation of $\mathcal{J}(D_\theta, t)$ over the process time,

$$\mathcal{J}(\theta') = \mathbb{E}_t \left[\frac{\lambda(t)}{\exp u_{\tilde{\theta}}(t)} \mathcal{J}(D_\theta; t) + u_{\tilde{\theta}}(t) \right], \tag{2}$$

where $t \sim U(0, 1)$ is uniformly sampled, and $\lambda : \mathbb{R} \to \mathbb{R}$ is a time-dependent weight. Following uncertainty-based multi-task learning (Kendall et al., 2018), the loss is further weighted by $u_{\tilde{\theta}}(t)$, a learnable linear projection of the process time parameterized by $\tilde{\theta}$, representing the uncertainty of the model. At the same time, the model is penalized for this uncertainty, encouraging $u_{\tilde{\theta}}(t)$ to be as low as possible. The full parameter set $\theta' = \{\theta, \tilde{\theta}\}$ thus includes both the denoiser model parameters and those of the learnable linear projection.

Following Karras et al. (2022), we define the denoiser model D_θ as

$$D_\theta(\mathbf{x}_t, \mathbf{s}, \mathbf{v}, t) = c_{\text{skip}}(t)\mathbf{x}_t + c_{\text{out}}(t)F_\theta(c_{\text{in}}(t)\mathbf{x}_t, \mathbf{s}, \mathbf{v}, t), \tag{3}$$

where $c_{\text{skip}} : \mathbb{R} \to \mathbb{R}$ is a skip scaling controlling the skip connection of $\mathbf{x}_t$, $c_{\text{out}} : \mathbb{R} \to \mathbb{R}$ is an output scaling, and $c_{\text{in}} : \mathbb{R} \to \mathbb{R}$ is an input scaling. The function $F_\theta : \mathbb{R}^{n_a} \times \mathbb{R}^{n_s} \times \mathbb{R}^{n_v} \times \mathbb{R} \to \mathbb{R}^{n_a}$ is a neural network parameterized by θ.

3.2 Network architecture

We utilize the MP-ADM architecture for our neural network (Karras et al., 2024). This architecture builds on the ADM architecture (Dhariwal and Nichol, 2021), introducing a series of modifications that retain the overall structure while significantly enhancing output quality. In particular, the network layers are carefully designed to preserve the expected magnitudes of activations and weight matrices during training. To condition the network on video features, we integrate feature-wise linear modulation (FiLM) (Perez et al., 2018) into the MP-ADM architecture. FiLM has been successfully applied in conditioning audio source separation (Liu et al., 2022) and speech enhancement (de Oliveira et al., 2024) systems on language prompts, and conditioning audio on video features for audio-visual sound event recognition (Brousmiche et al., 2022). Here, we introduce MP-FiLM, which we motivate as follows.

Magnitude-Preserving FiLM: For an input $\mathbf{x} \in \mathbb{R}^{d_x}$ and conditioning variable $\mathbf{c} \in \mathbb{R}^{d_c}$, FiLM layers are channel-wise defined as

$$\text{FiLM}(\mathbf{x}, \mathbf{c}|\theta, \phi) = \gamma_\theta(\mathbf{c}) \odot \mathbf{x} + \beta_\phi(\mathbf{c}), \tag{4}$$

where $\gamma_\theta : \mathbb{R}^{d_c} \to \mathbb{R}^{d_x}$ and $\beta_\phi : \mathbb{R}^{d_c} \to \mathbb{R}^{d_x}$ are neural networks parameterized by θ and ϕ and $\odot$ represents the Hadamard product (Perez et al., 2018). To ensure magnitude preservation of the addition in (4), we define MP-FiLM as

$$\text{MP-FiLM}(\mathbf{x}, \mathbf{c}|\theta, \phi) = \frac{(1 - \gamma_\theta(\mathbf{c}))\beta_\phi(\mathbf{c}) + \gamma_\theta(\mathbf{c}) \odot \mathbf{x}}{\sqrt{(1 - \gamma_\theta(\mathbf{c}))^2 + \gamma_\theta(\mathbf{c})^2}}. \tag{5}$$

Figure 2 shows our modified MP-ADM architecture, including MP-FiLM, which is integrated in each decoder block. Unlike the original FiLM, here the scaling and shifting of the network's activations are conducted per frame and not globally, so that each lip movement conditions the corresponding audio frame. With that intent, video features are initially interpolated in the time dimension to match the audio frame rate. We implement β_ϕ and γ_θ with two-layered convolutional blocks, containing a convolutional layer of kernel size 5, to account for potential misalignments between audio and lip movement, followed by a pointwise convolution. γ_θ blocks contain an additional learned gain, followed by a clamping operation to bound it between 0 and 1, defining the relative contribution of the video features at each fusion stage.

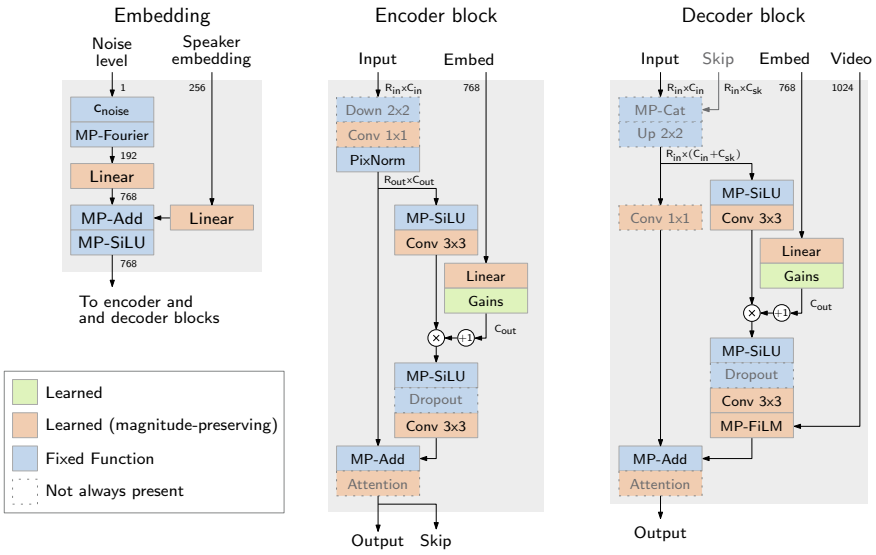


Figure 2: Modified MP-ADM network architecture with MP-FiLM layer to integrate video features.

3.3 Inference

For generating speech from video, we first extract video features $\mathbf{v} = E_v(\tilde{\mathbf{v}})$ from preprocessed video frames $\tilde{\mathbf{v}} \in \mathbb{R}^{N \times H \times W}$ using the video encoder E_v, where $\tilde{\mathbf{v}}$ is a grayscale video with N frames, height H, and width W depicting the speaker's lip movements. We extract a speaker embedding $\mathbf{s} = E_s(\tilde{\mathbf{s}})$ using the speaker encoder E_s, where $\tilde{\mathbf{s}} \in \mathbb{R}^{n_{\tilde{\mathbf{s}}}}$ is enrollment data of the target speaker's voice containing $n_{\tilde{\mathbf{s}}}$ samples.

Given the precomputed video features and the speaker embedding, we perform reverse diffusion using the trained denoiser model D_θ. We use the second-order deterministic sampler from Karras et al. (2022) with M sampling steps to generate a mel spectrogram $\hat{\mathbf{x}} \in \mathbb{R}^{n_a}$.

Finally, we synthesize the audio signal $\tilde{\mathbf{x}} = D_a(\hat{\mathbf{x}})$ using the neural vocoder D_a, where $\tilde{\mathbf{x}} \in \mathbb{R}^{n_{\tilde{x}}}$ denotes the estimated time-domain speech signal containing $n_{\tilde{x}}$ samples.

4 Experimental setup

We train our models for 32M samples using batch size 256, reference (peak) learning rate of 0.005, with linear ramp up followed by inverse square root decay, following Karras et al. (2024). Training is conducted on NVIDIA H100 graphics processing units (GPUs), typically requiring 60 GPU hours.

4.1 Data

We utilize the *LRS3* dataset (Afouras et al., 2018b) for training, validation, and the test. *LRS3* is an "in-the-wild" dataset comprising audio-visual recordings from TED Talks, in which the audio track often includes background noise and reverberation. Therefore, we use a pretrained generative speech enhancement model, specifically the Schrödinger bridge (Jukić et al., 2024; Richter et al., 2025), to enhance the audio tracks of the *LRS3* dataset. The preprocessing model has been trained on the *EARS-WHAM* (Richter et al., 2024) and VB-DMD (Valentini-Botinhao et al., 2016) datasets, for which a checkpoint is available online[2].

For the method to work, we have found that an audio-only pretraining stage is required, without video conditioning features. To increase the variety and quantity of training data, we pretrain the model on the 960 hours of the LibriSpeech dataset (Panayotov et al., 2015), also preprocessed by the speech enhancement model.

[2] `https://github.com/sp-uhh/sgmse`

We view lip-to-speech as an extreme case of audio-visual speech enhancement, where the audio is either completely absent or heavily corrupted. Consequently, we create an audio-visual speech enhancement benchmark called *LRS3-CHiME3*. This benchmark includes six subsets, each with a distinct SNR setting (5, 0, -5, -10, -15, and -20 dB). To create these subsets, we mixed clean speech from *LRS3* with noise files randomly sampled from the *CHiME3* noise database (Barker et al., 2015), adjusting the noise levels to achieve the specified SNR for each set.

For cross-dataset evaluation, we utilize all files with an SNR of -5 dB of the *NTCD-TIMIT* test set (Abdelaziz et al., 2017), which contains audio-visual recordings from the TCD-TIMIT dataset (Harte and Gillen, 2015) mixed with noise from the NOISEX-92 dataset (Varga and Steeneken, 1993) and CHiME challenges (Vincent et al., 2013; Barker et al., 2015).

4.2 Input representation

We use audio samples at a sampling rate of 16 kHz. We encode the waveforms into mel spectrograms with fast Fourier transform (FFT) length of 1024, and hop length of 256 samples, thus obtaining a feature rate of 62.5 Hz. The number of mel-frequency bins per frame is $n_a = 80$. The spectrograms are scaled and shifted to be standardized to zero-mean and a variance of 0.5, based on statistics computed on the training set. The time-frequency outputs of the model are converted into a time-domain signal using the neural vocoder D_a, for which we use the HiFi-GAN decoder (Kong et al., 2020), pretrained at 16 kHz. Speaker embeddings are obtained via the Wespeaker encoder D_s (Wang et al., 2023), which encode speaker characteristics into a feature vector of size $n_s = 256$. Video conditioning inputs are at 25 frames per second. They are first cropped into ROIs of 88×88 pixels, then fed into the self-supervised model BRAVEn (Haliassos et al., 2024), which results in features of dimension $n_v = 1024$ per video frame.

4.3 Metrics

We evaluate the proposed *LipDiffuser* method and compare it to other methods using a variety of metrics in order to capture different dimensions of performance. To evaluate speech quality, we employ two non-intrusive, deep neural network (DNN)-based speech quality assessment models, *DNS-MOS* (Reddy et al., 2021) and *NISQA*. Moreover, we also evaluate the speaker similarity (SpkSim), measured as the cosine similarity between speaker embeddings generated from the enrollment and synthesized speech signals. To quantify the effect on ASR tasks, we use the `QuartzNet15x5Base-En` model from the *NeMo* toolkit (Kuchaiev et al., 2019) as a downstream ASR system and report the word error rate (WER). We also measure the Levenshtein phoneme similarity (LPS) using the `w2v LV-60K` *wav2vec*-based phoneme predictor (Pirklbauer et al., 2023). To assess the synchronization between the generated speech and the input video, we use the lip sync error confidence (LSE-C) and the lip sync error distance (LSE-D) (Chung and Zisserman, 2017).

4.4 Baselines

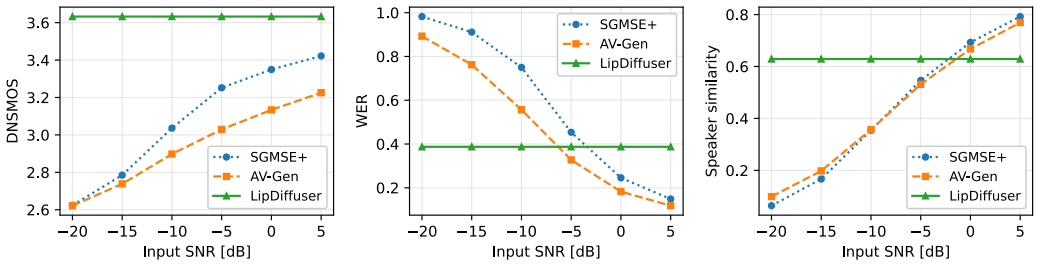
For lip-to-speech methods, we compare with *Lip2Speech* (Kim et al., 2023), *DiffV2* (Choi et al., 2023a), *IL2S* (Choi et al., 2023b), and *LipVoicer* (Yemini et al., 2024), all of which were trained on LRS3. *IL2S* comes in two configurations: the default with visual encoder trained from scratch, and a version which exploits visual features from the pretrained SSL model *AV-HuBERT* (Shi et al., 2022). As an audio-only speech enhancement method, we utilize *SGMSE+* (Richter et al., 2023b), which is a diffusion-based speech enhancement method using an Ornstein-Uhlenbeck process that interpolates between clean and noisy speech. We retrain the model with the original hyperparameters on the *LRS3-CHiME3* dataset. As an audio-visual speech enhancement method, we employ *AV-Gen* (Richter et al., 2023a). This method integrates audio-visual features extracted from *AV-HuBERT* (Shi et al., 2022) and inputs them into *SGMSE+* (Richter et al., 2023b). We retrain the model with the original hyperparameters using the *LRS3-CHiME3* dataset.

5 Results

We begin with in-domain results using the LRS3-CHiME3 test set comprising 1321 test files. Table 1 shows the mean performance of speech enhancement and lip-to-speech models at an SNR of -10 dB. Among the speech enhancement methods, the audio-visual *AV-Gen* model outperforms the audio-only

		Input	DNSMOS	NISQA	LPS	WER ↓	SpkSim	LSE-C	LSE-D ↓
	Clean	-	3.56	4.65	1.00	0.09	1.00	7.49	7.10
	Noisy	-	2.16	0.55	0.11	0.89	0.26	2.03	11.15
SE	*SGMSE+*	A	3.04	2.73	0.35	0.75	0.35	4.79	9.40
SE	*AV-Gen*	A+V	2.90	2.24	0.50	0.56	0.36	6.40	7.76
Lip-to-Speech	*Lip2Speech*	V+E	2.37	1.42	0.35	0.72	0.11	4.29	9.04
Lip-to-Speech	*DiffV2*	V	3.17	3.48	0.45	0.51	0.16	7.28	7.27
Lip-to-Speech	*IL2S*	V+E	2.84	2.53	0.52	0.59	0.34	**8.08**	**6.51**
Lip-to-Speech	*IL2S* (SSL)	V+E	2.86	2.62	**0.67**	0.36	0.37	8.03	6.54
Lip-to-Speech	*LipVoicer*	V	3.17	3.53	0.57	**0.36**	0.15	6.40	8.12
Lip-to-Speech	*LipDiffuser* (ours)	V+E	**3.64**	**4.57**	0.64	0.38	**0.63**	6.84	7.78

Table 1: Results on LRS3-CHiME3 showing mean values. Comparison of speech enhancement (SE) and lip-to-speech models at an SNR of −10 dB. For all metrics except LSE-D, higher values indicate better performance. Input modalities include audio (A), video (V), enrollment audio (E), or their combinations. Enrollment audio is derived from the clean reference.



(a) Speech quality (b) ASR performance (c) Speaker similariy

Figure 3: Performance across different dimensions—including speech quality (DNSMOS), ASR performance (WER), and speaker similarity—as a function of input SNR. Results are shown for audio-only speech enhancement (*SGMSE+*), audio-visual speech enhancement (*AV-Gen*), and lip-to-speech synthesis (*LipDiffuser*) on the LRS3-CHiME3 dataset.

SGMSE+ baseline regarding WER, LPS, and synchronization metrics, highlighting the benefit of visual information. For lip-to-speech approaches, our proposed *LipDiffuser* achieves the highest speech quality scores and the strongest speaker similarity, approaching the clean reference. In terms of ASR performance, *LipDiffuser*, *LipVoicer*, and *IL2S (SSL)* achieve comparable WER and LPS, all substantially outperforming the speech enhancement methods. Synchronization metrics (LSE-C/D) are generally highest for the *IL2S* models, but *LipDiffuser* remains competitive.

Figure 3 shows DNSMOS, WER, and speaker similarity for *SGMSE+*, *AV-Gen*, and *LipDiffuser* as a function of input SNR. While the performance of both speech enhancement models declines with decreasing SNR, *LipDiffuser's* performance remains constant since it does not rely on input audio. We observe that lip-to-speech consistently outperforms speech enhancement on all metrics below certain SNR thresholds, demonstrating its superiority in severely noisy environments.

Table 2 shows performance on the NTCD-TIMIT test set at an SNR of −5 dB, illustrating model generalization to out-of-domain data. All models show reduced performance compared to in-domain results, with SE systems (*SGMSE+*, *AV-Gen*) offering only minor improvements over noisy audio, especially in WER and LPS. Among lip-to-speech models, *LipDiffuser* achieves the best DNSMOS, NISQA, and speaker similarity, while *LipVoicer* provides the lowest WER and highest, Audio-visual synchronization is highest for *Lip2Speech* on LSE-C and LSE-D. Overall, *LipDiffuser* demonstrates strong robustness, maintaining high speech quality and speaker preservation even in mismatched, noisy conditions.

Table 3 reports the effects of various design choices on *LipDiffuser*'s performance. Training on the original LRS3 audio data instead of pre-enhancing the data, causes a clear drop in all metrics, including speech quality, ASR performance, and speaker similarity. Using Mel spectrograms as the target consistently outperforms STFT representations, especially for quality and intelligibility-related

		Input	DNSMOS	NISQA	LPS	WER ↓	SpkSim	LSE-C	LSE-D ↓
	Clean	-	3.59	3.61	1.00	0.14	1.00	2.62	9.01
	Noisy	-	2.24	1.02	0.10	0.93	0.19	0.87	8.11
SE	*SGMSE+*	A	2.74	1.99	0.11	0.97	0.23	1.16	9.81
SE	*AV-Gen*	A+V	2.75	1.98	0.24	0.96	0.21	1.72	8.46
L2S	*Lip2Speech*	V+E	2.33	1.72	0.31	0.94	0.00	**2.44**	**8.14**
L2S	*LipVoicer*	V	3.41	3.10	**0.61**	**0.49**	0.07	2.00	10.67
L2S	*LipDiffuser* (ours)	V+E	**3.73**	**4.38**	0.52	0.72	**0.52**	2.29	9.77

Table 2: Mismatched cross-dataset evaluation using the NTCD-TIMIT test set at an SNR of -5 dB. Values indicate the mean. For all metrics except LSE-D, higher values indicate better performance. Input modalities include audio (A), video (V), enrollment audio (E), or their combinations. Enrollment audio is derived from the clean reference.

	DNSMOS	NISQA	LPS	WER ↓	SpkSim	LSE-C	LSE-D ↓
LipDiffuser (Mel, BRAVEn)	3.64	4.57	0.64	0.38	0.63	6.84	7.78
– pre-enhanced training data	3.24	2.98	0.34	0.60	0.31	5.74	8.43
+ LRS3 pretraining (– LibriSpeech)	3.63	4.54	0.63	0.38	0.63	6.79	7.82
+ FiLM also in encoder	3.64	4.56	0.64	0.38	0.64	6.84	7.77
+ STFT (– Mel)	3.36	4.29	0.55	0.48	0.51	6.38	8.06
+ USR (– BRAVEn)	3.66	4.58	0.43	0.71	0.61	6.68	7.96
Same utterance	3.64	**4.58**	**0.65**	**0.43**	**0.64**	**6.93**	**7.75**
Other utterance	**3.65**	4.53	0.64	0.45	0.50	6.90	7.78
Enhanced utterance (-10 dB)	3.50	4.05	0.54	0.48	0.29	6.30	8.09
Noisy utterance (-10 dB)	3.09	3.27	0.27	0.78	0.18	4.12	9.98

Table 3: Design choice ablations on the LRS3-CHiME3 test set (top), and speaker embedding ablations limited to speakers with at least two utterances (bottom). The second block contains only speakers with more than one utterance in the test set

metrics (DNSMOS, WER, LPS). Pretraining on LRS3 instead of LibriSpeech, or adding FiLM layers to the encoder, yields minimal impact, indicating some flexibility in these components. Replacing the BRAVEn speaker encoder with the USR encoder yields similar quality and speaker similarity scores, but results in higher WER and lower LPS. This may be because the USR encoder was fine-tuned for visual speech recognition, whereas BRAVEn was used only in its pretrained configuration

Table 3 also investigates the effect of different enrollment utterances for speaker embedding on *LipDiffuser*'s performance. Using the very same, clean utterance as enrollment yields the best results across most metrics, including speaker similarity, intelligibility, and audio-visual synchronization. Using another clean utterance from the same speaker gives nearly identical quality but slightly reduced speaker similarity. Using either enhanced or noisy utterances for enrollment leads to noticeable drops, especially in speaker similarity and intelligibility. This highlights the benefit of clean enrollment audio for best performance, but also shows that speaker identity can still be maintained with clean utterances from other segments.

6 Conclusion

In this work, we proposed *LipDiffuser* a lip-to-speech diffusion model that synthesizes high-quality and intelligible speech from silent video recordings and an enrollment utterance. We proposed magnitude-preserving feature-wise linear modulation (MP-FiLM) layers for feature fusion within the magnitude-preserving ablated diffusion model (MP-ADM) network architecture. Experiments showed that below the threshold of -5 dB input SNR, *LipDiffuser* outperforms audio-visual and audio-only speech enhancement baselines. We conducted extensive experiments and ablations, with same- and cross-dataset evaluations. *LipDiffuser* consistently outperforms lip-to-speech baselines in audio quality and speaker similarity, with some robustness w.r.t different enrollment utterance scenarios. *LipDiffuser* demonstrates the capabilities of conditional diffusion models for generating high-quality speech in the lip-to-speech task.

References

Ahmed Hussen Abdelaziz et al. NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition. In *ISCA Interspeech*, pages 3752–3756, 2017.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *ISCA Interspeech*, pages 3244–3248, 2018a.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018b.

Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 504–511, 2015.

Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. Multimodal attentive fusion network for audio-visual event recognition. *Information Fusion*, 85:52–59, 2022.

Maxime Burchi and Radu Timofte. Audio-visual efficient conformer for robust speech recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2258–2267, 2023.

Gemma A Calvert, Edward T Bullmore, Michael J Brammer, Ruth Campbell, Steven CR Williams, Philip K McGuire, Peter WR Woodruff, Susan D Iversen, and Anthony S David. Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596, 1997.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021.

Jeongsoo Choi, Joanna Hong, and Yong Man Ro. Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7812–7821, 2023a.

Jeongsoo Choi, Minsu Kim, and Yong Man Ro. Intelligible lip-to-speech synthesis with speech units. In *ISCA Interspeech*, pages 4349–4353, 2023b.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision Workshops*, pages 251–263. Springer, 2017.

Danilo de Oliveira, Eric Grinstein, Patrick A. Naylor, and Timo Gerkmann. Laser: Language-queried speech enhancer. In *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 90–94, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Barbara Dodd. The role of vision in the perception of speech. *Perception*, 6(1):31–40, 1977.

Barbara Ed Dodd and Ruth Ed Campbell. *Hearing by eye: The psychology of lip-reading.* Lawrence Erlbaum Associates, Inc, 1987.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

Laurent Girin, Gang Feng, and Jean-Luc Schwartz. Noisy speech enhancement with filters estimated from the speaker's lips. In *European Conference on Speech Communication and Technology*, 1995.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, pages 369–376, 2006.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *ISCA Interspeech*, pages 5036–5040, 2020.

Alexandros Haliassos, Andreas Zinonos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Braven: Improving self-supervised pre-training for visual and auditory speech recognition. In *ICASSP*, pages 11431–11435, 2024.

Naomi Harte and Eoin Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ante Jukić, Roman Korostik, Jagadeesh Balam, and Boris Ginsburg. Schrödinger bridge for generative speech enhancement. In *ISCA Interspeech*, pages 1175–1179, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1–5, 2023.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, pages 17022–17033. Curran Associates, Inc., 2020.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference Learning Representations*, 2021.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. NeMo: a toolkit for building AI applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. Diffusion models for audio restoration: A review. *IEEE Signal Processing Magazine*, 41(6):72–84, 2025.

Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Separate what you describe: Language-queried audio source separation. In *Interspeech 2022*, pages 1801–1805, 2022.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6319–6323. IEEE, 2020.

John J McDonald, Wolfgang A Teder-SaÈlejaÈrvi, and Steven A Hillyard. Involuntary orienting to sound improves visual perception. *Nature*, 407(6806):906–908, 2000.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, pages 5206–5210, 2015.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Jan Pirklbauer, Marvin Sach, Kristoff Fluyt, Wouter Tirry, Wafaa Wardah, Sebastian Moeller, and Tim Fingscheidt. Evaluation metrics for generative speech enhancement methods: Issues and perspectives. In *ITG Speech Communication Conference*, pages 265–269. VDE, 2023.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6493–6497, 2021.

Julius Richter, Jeanine Liebold, and Timo Gerkamnn. Continuous phoneme recognition based on audio-visual modality fusion. In *International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2022.

Julius Richter, Simone Frintrop, and Timo Gerkmann. Audio-visual speech enhancement with score-based generative models. In *ITG Conference on Speech Communication*, pages 275–279, 2023a.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023b.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinjii Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *ISCA Interspeech*, pages 4873–4877, 2024.

Julius Richter, Danilo de Oliveira, and Timo Gerkmann. Investigating training objectives for generative speech enhancement. *ICASSP*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Ladan Shams and Aaron R Seitz. Benefits of multisensory learning. *Trends in cognitive sciences*, 12(11): 411–417, 2008.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference Machine Learning*, pages 2256–2265, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference Learning Representations*, 2021.

BE Stein. *The Merging of the Senses*. MIT Press, 1993.

William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

Kristina Tesch and Timo Gerkmann. Insights into deep non-linear filters for improved multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:563–575, 2022.

Szabolcs Torma and Luca Szegletes. Generative modeling and augmentation of eeg signals using improved diffusion probabilistic models. *Journal of Neural Engineering*, 22(1):016001, 2025.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. *ISCA Speech Synthesis Workshop*, pages 146–152, 2016.

Andrew Varga and Herman JM Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251, 1993.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2025.

Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 126–130, 2013.

Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7): 1661–1674, 2011.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP*, pages 1–5. IEEE, 2023.

Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. A survey on diffusion models for time series and spatio-temporal data. *CoRR*, 2024.

Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. LipVoicer: Generating speech from silent videos guided by lip reading. In *International Conference on Learning Representations*, 2024.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023.

A.5 Alternative Diffusion Processes

A.5.1 Publication: Normalize Everything: A Preconditioned Magnitude-Preserving Architecture for Diffusion-Based Speech Enhancement

Abstract

This paper presents a new framework for diffusion-based speech enhancement. Our method employs a Schrödinger bridge to transform the noisy speech distribution into the clean speech distribution. To stabilize and improve training, we employ time-dependent scalings of the inputs and outputs of the network, known as preconditioning. We consider two skip connection configurations, which either include or omit the current process state in the denoiser's output, enabling the network to predict either environmental noise or clean speech. Each approach leads to improved performance on different speech enhancement metrics. To maintain stable magnitude levels and balance during training, we use a magnitude-preserving network architecture that normalizes all activations and network weights to unit length. Additionally, we propose learning the contribution of the noisy input within each network block for effective input conditioning. After training, we apply a method to approximate different exponential moving average (EMA) profiles and investigate their effects on the speech enhancement performance. In contrast to image generation tasks, where longer EMA lengths often enhance mode coverage, we observe that shorter EMA lengths consistently lead to better performance on standard speech enhancement metrics. Code, audio examples, and checkpoints are available online.

Reference

Julius Richter, Danilo de Oliveira, and Timo Gerkmann, "Normalize Everything: A Preconditioned Magnitude-Preserving Architecture for Diffusion-Based Speech Enhancement", submitted to *IEEE Workshop on Applications Signal Processing to Audio and Acoustics*, 2025.

An up-to-date version is available on arXiv (`http://arxiv.org/abs/2505.05216`).

Author's Contributions

Julius Richter is the lead author of this paper. He developed the algorithm, trained all models, conducted the experiments, and wrote the manuscript. Danilo de Oliveira and Timo Gerkmann contributed through insights into experimental validation, mathematical derivations, and manuscript review.

Normalize Everything: A Preconditioned Magnitude-Preserving Architecture for Diffusion-Based Speech Enhancement

Julius Richter[1], *Danilo de Oliveira*[1], *Timo Gerkmann*[1]

[1]Signal Processing Group, University of Hamburg, Germany

***Abstract*—This paper presents a new framework for diffusion-based speech enhancement. Our method employs a Schrödinger bridge to transform the noisy speech distribution into the clean speech distribution. To stabilize and improve training, we employ time-dependent scalings of the inputs and outputs of the network, known as preconditioning. We consider two skip connection configurations, which either include or omit the current process state in the denoiser's output, enabling the network to predict either environmental noise or clean speech. Each approach leads to improved performance on different speech enhancement metrics. To maintain stable magnitude levels and balance during training, we use a magnitude-preserving network architecture that normalizes all activations and network weights to unit length. Additionally, we propose learning the contribution of the noisy input within each network block for effective input conditioning. After training, we apply a method to approximate different exponential moving average (EMA) profiles and investigate their effects on the speech enhancement performance. In contrast to image generation tasks, where longer EMA lengths often enhance mode coverage, we observe that shorter EMA lengths consistently lead to better performance on standard speech enhancement metrics. Code, audio examples, and checkpoints are available online[1].**

1. INTRODUCTION

Diffusion-based generative models for speech enhancement learn clean speech posteriors conditioned on noisy inputs [1]. Due to their data-driven nature, they are highly expressive and well-suited for general audio restoration, while also offering the flexibility to be integrated into model-based approaches [2]. Generative models for speech restoration are particularly effective at handling various speech communication artifacts, including background noise, reverberation, bandwidth limitation, codec artifacts, and packet loss [3].

Diffusion models break down data generation into more manageable denoising tasks by progressively removing noise and refining the data [4]. The forward process used for training can be described by a stochastic process, which transforms the target distribution into a Gaussian noise prior [5]. Reversing this process in time results in the reverse process, which is numerically integrated for inference. Various stochastic processes have been utilized for speech enhancement, including the Ornstein-Uhlenbeck process [1], Brownian bridge [6], and Schrödinger bridge [7]. All these methods require training a denoiser or score model, depending on the specifics of the sampling process.

The denoiser model is a neural network trained to distinguish clean speech from environmental noise and Gaussian noise introduced during the forward process. Expressive denoiser models are often implemented using the U-Net architecture [8], augmented with self-attention layers [9]. To maintain network operations within a suitable range and prevent large variations in gradient magnitudes, Karras et al. [10] proposed preconditioning the denoiser, which refers to inputs, targets, or updates being scaled or modified to improve robustness and accelerate training or sampling. Gonzales et al. [11] have explored preconditioning in diffusion-based speech enhancement using a cro:VEvariance exploding (VE) diffusion process and applying a change of variables to handle the stochastic process of environmental noise. However, preconditioning has not been explored for diffusion bridges like the Schrödinger bridge [12], [13] applied to speech. In a subsequent study, Karras et al [14] observed uncontrolled magnitude changes and imbalances in both network activations and weights throughout the training process, despite the application of preconditioning. To mitigate this effect, they redesigned the network layers of the cro:ADMablated diffusion model (ADM) U-Net architecture [15] to preserve the expected magnitudes of activations, weights, and updates, leading to improved performance for image generation tasks. De Oliveira et al. [16] used the magnitude-preserving ADM architecture for training an unconditional diffusion model on clean speech for non-intrusive speech quality assessment. Richter et al. [17] presented preliminary results employing the magnitude-preserving ADM architecture for diffusion-based speech enhancement with an Ornstein-Uhlenbeck process. However, they did not investigate its application to diffusion bridges or conduct thorough ablations, such as analyzing the effects of varying the cro:EMAexponential moving average (EMA) length.

In this work, we build upon the findings of Karras et al. [10], [14] by investigating the application of preconditioning and the magnitude-preserving ADM architecture to the Schrödinger bridge framework for speech enhancement [7], [17]. We derive time-dependent scalings for the network's inputs and outputs (a.k.a. preconditioning) and propose two versions for scaling a skip connection, which results in the network's training target being either environmental noise or clean speech. We show experimentally that while one configuration leads to a higher cro:SDRsignal-to-distortion ratio (SDR), another yields better perceptual scores. Within the magnitude-preserving network architecture, we propose learning the contribution of the noisy input within each network block to facilitate effective input conditioning. Additionally, we implement a method to approximate different EMA profiles post-training [14], finding that, in contrast to outcomes from image generation, shorter EMA lengths consistently yield better speech enhancement performance. This underscores the importance of carefully tuning EMA for speech restoration tasks, an area that has so far received little attention. Finally, we perform experiments using two speech enhancement benchmarks in 16 kHz: VoiceBank-DEMAND [18] and EARS-WHAM_v2 [19]. Compared to strong baselines, our approach remains competitive while introducing new concepts that extend beyond speech enhancement tasks.

2. RELATED WORK

2.1. Schrödinger Bridge for Speech Enhancement

The Schrödinger bridge problem was first introduced in the context of quantum mechanics [20]. It has since attracted wider interest due to its connections with optimal transport theory [21]. The dynamic Schrödinger bridge [22] is typically defined as

$$\min_{\mathbb{Q}\in\Pi(p_{\mathcal{A}},p_{\mathcal{B}})} D_{\mathrm{KL}}\left(\mathbb{Q},\mathbb{P}\right), \tag{1}$$

where $\Pi(p_{\mathcal{A}}, p_{\mathcal{B}})$ denotes the set of path measures with marginal densities $p_{\mathcal{A}}$ and $p_{\mathcal{B}}$ at the boundaries. By associating the path

[1]To be released at the time of acceptance.

measures $\mathbb{Q}$ and $\mathbb{P}$ with controlled and uncontrolled diffusion processes, respectively, we obtain the stochastic optimal control formulation

$$\begin{aligned} &\min_{\mathbf{u}(\mathbf{x}_t,t)} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|\mathbf{u}(\mathbf{x}_t,t)\|^2 \,\mathrm{d}t\right] \\ &\text{s.t. } \mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t,t)+\mathbf{u}(\mathbf{x}_t,t)]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \\ &\mathbf{x}_0 \sim p_{\mathcal{A}}, \quad \mathbf{x}_1 \sim p_{\mathcal{B}}, \end{aligned} \tag{2}$$

where $\mathbf{x}_t \in \mathbb{R}^N$ is a stochastic process governed by the cro:SDEstochastic differential equation (SDE) with drift coefficient $\mathbf{f} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$, diffusion coefficient $g : \mathbb{R} \to \mathbb{R}$, and $\mathbf{w}_t$ denoting a standard Wiener process. The optimization problem (2) aims to find an optimal control process $\mathbf{u}(\mathbf{x}_t, t)$ that minimizes the accumulated energy cost over the time horizon $[0, 1]$, while satisfying the distributional boundary conditions. Applying the Hopf-Cole transformation [23], [24] to the necessary conditions for (2) leads to a system of coupled partial differential equations [12],

$$\begin{aligned} &\begin{cases} \frac{\partial \Psi_t}{\partial t} = -\nabla_{\mathbf{x}_t}\Psi_t(\mathbf{x}_t)^{\mathrm{T}}\mathbf{f}(\mathbf{x}_t) - \frac{1}{2}\,\mathrm{Tr}(g(t)^2\nabla^2_{\mathbf{x}_t}\Psi_t(\mathbf{x}_t)) \\ \frac{\partial \bar{\Psi}_t}{\partial t} = -\nabla_{\mathbf{x}_t}\cdot(\bar{\Psi}_t(\mathbf{x}_t)\mathbf{f}(\mathbf{x}_t)) + \frac{1}{2}\,\mathrm{Tr}(g(t)^2\nabla^2_{\mathbf{x}_t}\bar{\Psi}_t(\mathbf{x}_t)) \end{cases} \\ &\textit{s.t. } \Psi_0\bar{\Psi}_0 = p_x, \Psi_1\bar{\Psi}_1 = p_y, \end{aligned} \tag{3}$$

where Ψ_t and $\bar{\Psi}_t$ are time-varying energy potentials satisfying Nelson's identity $\Psi_t\bar{\Psi}_t = p_t$ [25]. This leads to the optimal control law $\mathbf{u}^*(\mathbf{x}_t, t) = \beta_t\nabla \log \Psi(\mathbf{x}_t, t)$ and consequently, to the forward SDE being

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t) + g(t)^2\nabla_{\mathbf{x}_t}\log\Psi_t(\mathbf{x}_t)\right]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \tag{4}$$

Analogously, the optimal control process to the stochastic optimal control formulation in (2) running backward in time leads to the reverse SDE

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t) - g(t)^2\nabla_{\mathbf{x}_t}\log\bar{\Psi}_t(\mathbf{x}_t)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t. \tag{5}$$

For a system of symmetric forward and reverse SDEs in (4) and (5), and arbitrary Ψ_t and $\bar{\Psi}_t$, there are infinitely many solutions bridging the initial distribution to the target [26]. However, closed-form solutions are available for specific cases, such as those involving Gaussian boundary conditions [27]. Assume a drift $\mathbf{f}(\mathbf{x}_t) = f(t)\,\mathbf{x}_t$ and Gaussian boundary conditions $p_0(\mathbf{x}|\mathbf{x}_0) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}; \mathbf{x}_0, \epsilon_0^2\mathbf{I}\right)$ and $p_1(\mathbf{x}|\mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}; \mathbf{y}, \epsilon_1^2\mathbf{I}\right)$ where $\epsilon_1 = \mathrm{e}^{\int_0^1 f(\tau)\mathrm{d}\tau}\epsilon_0$. For $\epsilon_0 \to 0$, the Schrödinger bridge solution between clean speech $\mathbf{x}_0$ and noisy speech $\mathbf{y}$ can be expressed as

$$\bar{\Psi}_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}_{\mathbb{C}}(\alpha_t\mathbf{x}_0, \alpha_t^2\sigma_t^2\mathbf{I}), \tag{6}$$

$$\Psi_t(\mathbf{x}_t|\mathbf{y}) = \mathcal{N}_{\mathbb{C}}(\bar{\alpha}_t\mathbf{y}, \alpha_t^2\bar{\sigma}_t^2\mathbf{I}) \tag{7}$$

with parameters $\alpha_t = \mathrm{e}^{\int_0^t f(\tau)\mathrm{d}\tau}$, $\sigma_t^2 = \int_0^t \frac{g^2(\tau)}{\alpha_\tau^2}\mathrm{d}\tau$, $\bar{\alpha}_t = \alpha_t\alpha_1^{-1}$ and $\bar{\sigma}_t^2 = \sigma_1^2 - \sigma_t^2$ [28]. Therefore, the marginal distribution is the Gaussian distribution

$$p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_t; \boldsymbol{\mu}_t(\mathbf{x}_0, \mathbf{y}), \sigma^2_{\mathbf{x}_t}\mathbf{I}\right) \tag{8}$$

with mean

$$\boldsymbol{\mu}_t(\mathbf{x}_0, \mathbf{y}) = w_x(t)\mathbf{x}_0 + w_y(t)\mathbf{y}, \tag{9}$$

and variance

$$\sigma^2_{\mathbf{x}_t} = \frac{\alpha_t^2\bar{\sigma}_t^2\sigma_t^2}{\sigma_1^2}, \tag{10}$$

where $w_x(t) = \alpha_t\bar{\sigma}_t^2/\sigma_1^2$, and $w_y(t) = \bar{\alpha}_t\sigma_t^2/\sigma_1^2$ [28].

To train the Schrödinger bridge, a denoiser model D_θ parameterized by θ is trained using a data prediction loss. Jukić et al. [7] proposed to include a time-domain auxiliary loss term based on the ℓ_1-norm [7], resulting in the training objective

$$\mathcal{J}(\theta) = \mathbb{E}_{t,\mathbf{x}_t,\mathbf{x}_0,\mathbf{y}}\left[\|D_\theta(\mathbf{x}_t, \mathbf{y}, t) - \mathbf{x}_0\|_2^2 + \alpha\|\underline{\hat{\mathbf{x}}}_\theta - \underline{\mathbf{x}}_0\|_1\right] \tag{11}$$

where $t \sim \mathcal{U}(t_{\text{eps}}, 1)$ is uniformly sampled, $(\mathbf{x}_0, \mathbf{y}) \sim p_{\text{data}}(\mathbf{x}, \mathbf{y})$ are drawn from the empirical data distribution, and $\mathbf{x}_t \sim p_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y})$ is the noisy sample at process time t. The time-domain signals $\underline{\hat{\mathbf{x}}}_\theta$ and $\underline{\mathbf{x}}_0$ are obtained via the cro:iSTFTinverse short-time Fourier transform (iSTFT), enabling backpropagation through the iSTFT operation.

At inference, the reverse SDE in Eq. (5) can be solved with an cro:ODEordinary differential equation (ODE) or SDE sampler [28]. Here, we use the ODE sampler because it has shown better performance for the speech enhancement task [7].

2.2. Magnitude-Preserving Learned Layers

The operation of a fully-connected layer with input activations $\mathbf{a} \in \mathbb{R}^n$ and output activations $\mathbf{b} \in \mathbb{R}^m$, excluding a bias term, is defined as $\mathbf{b} = \mathbf{W}\,\mathbf{a}$, where $\mathbf{W} = [\mathbf{w}_i] \in \mathbb{R}^{m\times n}$ is a trainable weight matrix with row vectors $\mathbf{w}_i \in \mathbb{R}^n$. Equivalently, for a single element b_i in $\mathbf{b}$, we have $b_i = \mathbf{w}_i \cdot \mathbf{a}$. The same definition extends to convolutional layers by applying this formulation independently to each output element. In this case, the elements of $\mathbf{a}$ correspond to the activations of all input elements within the receptive field of the convolution kernel. Thus, the dimension of $\mathbf{a}$ is $\dim(\mathbf{a}) = N_j = N_c\,k^2$ where N_c is the number of input channels, and k is the spatial size of the convolution kernel.

For magnitude-preserving learned layers [14], the output activations are required to have the same variance as the input activations. This can be achieved by rescaling each output activation as $\tilde{b}_i = (\sigma_{\mathbf{a}}/\sigma_{b_i})b_i$, where $\sigma_{\mathbf{a}} = \sqrt{\mathrm{Var}(\mathbf{a})}$ and $\sigma_{b_i} = \sqrt{\mathrm{Var}(b_i)}$ are the standard deviations of the input and output activations, respectively. This can be achieved by normalizing the weight vector $\mathbf{w}_i$ by its ℓ_2-norm [14, Appendix B.4], resulting in

$$\tilde{b}_i = \frac{\mathbf{w}_i}{||\mathbf{w}_i||_2 + \epsilon} \cdot \mathbf{a}\,, \tag{12}$$

with a small constant ϵ added for numerical stability to avoid division by zero. Since $\hat{b}_i$ is now scale-invariant with respect to w_i, all weights are initialized as $w_{i,j} \sim \mathcal{N}(0, 1)$ to ensure uniform magnitude across all layers.

With the normalization in (12), Karras et al. [14] demonstrate that gradient descent optimization of the loss function preserves the norm of $\mathbf{w}_i$. However, in practical implementations with finite step sizes, discretization errors can still lead to changes in $\|\mathbf{w}_i\|_2$. To address this, magnitude-preserving learned layers strictly enforce the constraint $\|\mathbf{w}_i\|_2 = \sqrt{N_j}$ on each weight vector after every training step, where N_j denotes the dimensionality of $\mathbf{w}_i$. Accordingly, under standard gradient descent with learning rate α, the update rule becomes

$$\mathbf{w}_i \leftarrow \sqrt{N_j}\frac{\mathbf{w}_i'}{\|\mathbf{w}_i'\|_2}, \quad \text{with} \quad \mathbf{w}_i' = \mathbf{w}_i - \alpha\nabla_{\mathbf{w}_i}\mathcal{L}, \tag{13}$$

2.3. Post-training Exponential Moving Average (EMA)

The EMA maintains a running average $\hat{\theta}_\beta$ of the network parameters θ during training. At each training step n, the average is updated as

$$\hat{\theta}_\beta^{(n)} = \beta\hat{\theta}_\beta^{(n-1)} + (1-\beta)\theta^{(n)}, \tag{14}$$

in which β (typically close to 1) controls an exponential decay of contributions from previous steps.

Karras et al. [14] propose a slightly altered averaging profile based on power functions instead of exponential decay, which has the effect that the averaging profile automatically scales with training time (β_γ

is dependent on n). The update rule for the power function EMA is defined as

$$\hat{\theta}_\gamma^{(n)} = \beta_\gamma^{(n)}\hat{\theta}_\gamma^{(n-1)} + (1-\beta_\gamma^{(n)})\theta^{(n)} \text{ with } \beta_\gamma^{(n)} = (1-1/n)^{\gamma+1}, \quad (15)$$

where the constant γ controls the sharpness of the profile. In contrast to (14), the formulation in (15) has zero weight for the random weight initialization $\theta^{(0)}$ at the beginning of the training. While the parameter γ is mathematically well-defined, its impact on the averaging profile is non-intuitive. Therefore, the profile is characterized by its relative standard deviation σ_{rel}, representing the peak width relative to training duration. For example, when specifying an EMA length of 10%, this corresponds to $\sigma_{\text{rel}} = 0.10$ (equivalent to $\gamma \approx 6.94$).

Additionally, Karras et al. [14] propose a method for approximating the EMA profile post-training. This allows sampling the length of EMA densely and plotting its effect on the model performance. To achieve this, two EMA parameter vectors $\hat{\theta}_{\gamma_1}$ and $\hat{\theta}_{\gamma_2}$ with $\gamma_1 = 16.97$ ($\sigma_{\text{rel}} = 0.05$) and $\gamma_2 = 6.94$ ($\sigma_{\text{rel}} = 0.10$) are periodically saved as training snapshots throughout the training. To reconstruct $\hat{\theta}$ for an arbitrary EMA profile, the least-squares optimal linear combination of stored snapshots $\hat{\theta}_{\gamma_i}$ is computed that best matches the desired averaging profile.

3. METHOD

In this section, we present our diffusion-based speech enhancement framework, referred to as *cro:NESENormalize Everything Speech Enhancement (NESE)*. The diffusion process is formulated as a Schrödinger bridge, following the principles outlined in Section 2.1. We further introduce our novel contributions, incorporating a preconditioned, magnitude-preserving network architecture.

3.1. Preconditioning

Preconditioning refers to the technique of rescaling or transforming the network's inputs and outputs at each time step to enhance numerical stability and improve the efficiency of training the denoiser model.

Given the signal model

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \quad (16)$$

where $\mathbf{x} \in \mathbb{R}^n$ is the clean speech, $\mathbf{n} \in \mathbb{R}^n$ the environmental noise, and $\mathbf{y} \in \mathbb{R}^n$ the noisy mixture. Assume that $\mathbf{x} \sim p_x(\mathbf{x})$ and $\mathbf{n} \sim p_n(\mathbf{n})$ are independent and have variance $\sigma_{\mathbf{x}}^2$ and $\sigma_{\mathbf{n}}^2$, respectively. The learning objective on time step t is given by

$$\mathcal{J}(\theta; t) = \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{z}}\Big[||D_\theta(\underbrace{\boldsymbol{\mu}_t(\mathbf{x},\mathbf{y}) + \sigma_{\mathbf{x}_t}\mathbf{z}}_{\mathbf{x}_t}; t) - \mathbf{x}||_2^2\Big], \quad (17)$$

where $\mathbf{z} \sim \mathcal{N}(0, I)$ is a random Gaussian vector. We obtain the overall objective by taking a weighted expectation of $\mathcal{J}(\theta; t)$ over the time steps

$$\mathcal{J}(\theta) = \mathbb{E}_{t,\mathbf{x},\mathbf{y},\mathbf{z}}\big[\lambda(t)||D_\theta(\mathbf{x}_t + \sigma(t)\mathbf{z}; t) - \mathbf{x}||_2^2\big], \quad (18)$$

where $t \sim \mathcal{U}(t_{\text{eps}}, 1)$ is uniformly distributed, and $\lambda(t) : \mathbb{R} \to \mathbb{R}$ is a time-dependent weight.

We define the denoiser model D_θ as

$$D_\theta(\mathbf{x}_t, \mathbf{y}; t) = c_{\text{skip}}(t)\mathbf{x}_t + c_{\text{out}}(t)F_\theta(c_{\text{in}}(t)\mathbf{x}_t, c_{\text{in}}(1)\mathbf{y}, t), \quad (19)$$

where $c_{\text{skip}} : \mathbb{R} \to \mathbb{R}$ is a skip scaling controlling the skip connection of $\mathbf{x}_t$, $c_{\text{out}} : \mathbb{R} \to \mathbb{R}$ is an output scaling, and $c_{\text{in}} : \mathbb{R} \to \mathbb{R}$ is an input scaling. The input scaling for the conditioner is always set to $c_{\text{in}}(1)$, as the variance is expected to match that of $\mathbf{x}_t$ at $t = 1$. The function $F_\theta : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is a neural network parameterized by θ. Using the definition in (19), we can rewrite the objective in (18) as

$$\mathcal{J}(\theta) = \mathbb{E}_{t,\mathbf{x},\mathbf{y},\mathbf{z}}\big[\underbrace{\lambda(t)c_{\text{out}}(t)^2}_{:=w(t)} ||F_\theta\left(c_{\text{in}}(t)\mathbf{x}_t, c_{\text{in}}(1)\mathbf{y}, t\right) - \underbrace{\frac{1}{c_{\text{out}}(t)}(\mathbf{x} - c_{\text{skip}}(t)\mathbf{x}_t)}_{:=F_{\text{target}}(\mathbf{x}_z; t)}||_2^2\big], \quad (20)$$

where $w(t)$ is the effective time-dependent weight and $F_{\text{target}}(\mathbf{x}, \mathbf{y}, \mathbf{z}; t)$ is the target of the neural network

3.1.1. Input Scaling: We require the training inputs of the neural network F_θ to have unit variance, i.e., we want to have:

$$\text{Var}_{\mathbf{x},\mathbf{y},\mathbf{z}}[c_{\text{in}}(t)(\boldsymbol{\mu}(x, y; t) + \sigma(t)\mathbf{z})] = 1 \quad (21)$$

$$\Leftrightarrow c_{\text{in}}(t)^2 \,\text{Var}_{\mathbf{x},\mathbf{n},\mathbf{z}}[(w_{\mathbf{x}}(t) + w_{\mathbf{y}}(t))\mathbf{x} + w_{\mathbf{y}}(t)\mathbf{n} + \sigma(t)\mathbf{z}] = 1 \quad (22)$$

Since $\mathbf{x}$, $\mathbf{n}$, and $\mathbf{z}$ are independent, we obtain

$$c_{\text{in}}(t) = \frac{1}{(w_{\mathbf{x}}(t) + w_{\mathbf{y}}(t))^2\sigma_{\mathbf{x}}^2 + w_{\mathbf{y}}(t)^2\sigma_{\mathbf{n}}^2 + \sigma(t)^2}. \quad (23)$$

3.1.2. Output Scaling: We require the target of the neural network F_θ to have unit variance, i.e., we want to have:

$$\text{Var}_{\mathbf{x},\mathbf{y},\mathbf{z}}[F_{\text{target}}(\mathbf{x}, \mathbf{y}, \mathbf{z}; t)] = 1 \quad (24)$$

$$\Leftrightarrow\ c_{\text{out}}(t)^2 = \text{Var}_{\mathbf{x},\mathbf{y},\mathbf{z}}[(1 - c_{\text{skip}}(t)w_{\mathbf{x}}(t) - c_{\text{skip}}(t)w_{\mathbf{y}}(t))\mathbf{x} + c_{\text{skip}}(t)w_{\mathbf{y}}(t)\mathbf{n} + c_{\text{skip}}(t)\sigma(t)\mathbf{z}] \quad (25)$$

Thus, we get

$$c_{\text{out}}(t)^2 = (1 - c_{\text{skip}}(t)w_{\mathbf{x}}(t) - c_{\text{skip}}(t)w_{\mathbf{y}}(t))^2\sigma_{\mathbf{x}}^2 + c_{\text{skip}}(t)^2w_{\mathbf{y}}(t)^2\sigma_{\mathbf{n}}^2 + c_{\text{skip}}(t)^2\sigma(t)^2, \quad (26)$$

3.1.3. Skip Scaling: We explore selecting $c_{\text{skip}}(t)$ to be 1 or 0. For $c_{\text{skip}}(t) = 1$, the neural network learns to predict a scaled version of the environmental noise, and for $c_{\text{skip}}(t) = 0$, the neural network learns to predict a scaled version of the clean speech.

Selecting $c_{\text{skip}}(t) = 1$ results in the output scaling

$$c_{\text{out},1}(t) = \sqrt{(1 - w_{\mathbf{x}}(t) - w_{\mathbf{y}}(t))^2\sigma_{\mathbf{x}}^2 + w_{\mathbf{y}}(t)^2\sigma_{\mathbf{n}}^2 + \sigma(t)^2}, \quad (27)$$

whereas $c_{\text{skip}}(t) = 0$ leads to

$$c_{\text{out},0}(t) = \sigma_{\mathbf{x}}^2 \quad (28)$$

3.1.4. Loss Weighting: We require the effective weight to satisfy $w(t) = 1$. If we choose $c_{\text{skip}}(t) = 1$, the weighting function becomes

$$\lambda_1(t) = \frac{1}{(1 - w_{\mathbf{x}}(t) - w_{\mathbf{y}}(t))^2\sigma_{\mathbf{x}}^2 + w_{\mathbf{y}}(t)^2\sigma_{\mathbf{n}}^2 + \sigma(t)^2}, \quad (29)$$

whereas selecting $c_{\text{skip}}(t) = 0$ results in

$$\lambda_0(t) = 1/\sigma_{\mathbf{x}}^2. \quad (30)$$

3.2. Network architecture

We use the magnitude-preserving ADM architecture, with magnitude-preserving learned layers as described in Section 2.2. We propose incorporating the conditional features—specifically, the noisy speech or downsampled version of the noisy speech, depending on the resolution—into the network by adding them to the feature representations immediately after the merging of the residual branch within each network block. We use the magnitude-preserving addition for the fusion

$$\text{MP-Sum}(\mathbf{a}, \mathbf{b}, \tau) = \frac{(1 - \tau)\mathbf{a} + \tau\mathbf{b}}{\sqrt{(1 - \tau)^2 + \tau^2}}, \quad (31)$$

and propose to learn the interpolation coefficient τ.

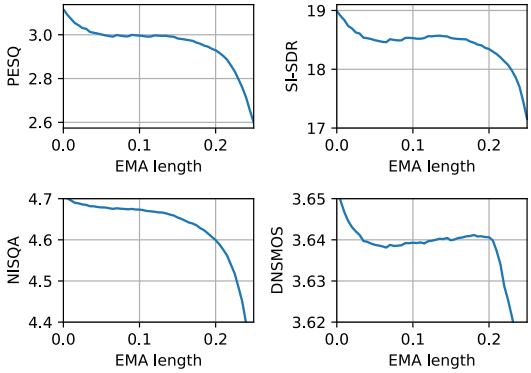


Fig. 1: Speech enhancement performance as a function of EMA length on the validation set of VoiceBank-DEMAND.

Table 1: Speech enhancement performance with different skip scaling, auxiliary loss configurations, and training datasets tested using the VoiceBank-DEMAND test set. Values indicate the mean.

		SI-SDR	PESQ	DNSMOS	NISQA
Matched	$c_{\text{skip}} = 1$, $\alpha = 0.001$	17.50	**2.97**	3.50	4.70
	$c_{\text{skip}} = 1$, $\alpha = 0.0$	17.58	2.91	3.52	4.71
	$c_{\text{skip}} = 0$, $\alpha = 0.001$	**18.07**	2.90	**3.55**	**4.76**
	$c_{\text{skip}} = 0$, $\alpha = 0.0$	18.04	2.89	**3.55**	4.75
Mismatched	$c_{\text{skip}} = 1$, $\alpha = 0.001$	14.79	2.69	**3.55**	4.42
	$c_{\text{skip}} = 1$, $\alpha = 0.0$	15.71	2.81	3.54	4.45
	$c_{\text{skip}} = 0$, $\alpha = 0.001$	14.23	2.64	3.54	4.34
	$c_{\text{skip}} = 0$, $\alpha = 0.0$	15.18	2.71	**3.55**	4.48

4. EXPERIMENTAL SETUP

Following [7], [17], we utilize the identical amplitude-compressed time-frequency representation as input. The forward and reverse SDEs in (4) and (5) are parameterized with $\mathbf{f}(\mathbf{x}, t) = \mathbf{0}$, and $g(t) = \sqrt{c}\, k^t$ with $c = 0.4$ and $k = 2.6$. For numerical stability, we select $t_{\text{eps}} = 0.02$, which corresponds to the step size of our uniform discretization scheme with 50 sampling steps. In all experiments, we train on 2 GPUs with a batch size of 16, stopping after processing exactly 6.291M training samples. We store a snapshot once every 1,024k training samples, or once every 64k training steps with a batch size of 16. We employ an inverse square root learning rate decay schedule [29], with an initial rate of 2.5×10^{-3} that decreases after processing 3×10^4 training samples. We train a total of eight models, covering all combinations of:

- Training dataset: {VoiceBank-DEMAND, EARS-WHAM}
- Skip connection: $\{c_{\text{skip}}(t) = 1,\ c_{\text{skip}}(t) = 0\}$
- Auxiliary loss weighting: $\{\alpha = 0,\ \alpha = 0.001\}$

5. RESULTS

We begin by investigating the effect of EMA length through an ablation study, using the model with $c_{\text{skip}}(t) = 1$ and $\alpha = 0.001$, trained on VoiceBank-DEMAND. Figure 1 presents the speech enhancement results on the validation set as a function of the EMA length. We observe that shorter EMA lengths consistently yield better speech enhancement performance across all metrics. After this initial peak, the performance plateaus until a sharp decline is observed when the EMA length exceeds $\sigma_{\text{rel}} = 0.2$. This result contrasts with findings in image generation, where larger EMA lengths are often preferred [14]. While EMA can improve system-level metrics such as cro:FIDFréchet inception distance (FID) in image generation, our results suggest that longer EMA lengths may be detrimental to instance-based metrics, as seen in the speech enhancement task. Based on these findings, we use $\sigma_{\text{rel}} = 0.001$ for the EMA length in all subsequent experiments.

Table 2: Speech enhancement performance on VoiceBank-DEMAND. Values indicate mean and standard deviation.

		SI-SDR	PESQ	DNSMOS	NISQA
	Clean	∞	4.64 ± 0.00	3.55 ± 0.28	4.50 ± 0.30
	Noisy	8.44 ± 5.61	1.97 ± 0.75	3.09 ± 0.39	3.03 ± 0.82
Match.	SGMSE+ [1]	17.35 ± 3.33	2.93 ± 0.62	3.56 ± 0.28	4.51 ± 0.38
	SB-VE [7]	$\mathbf{19.41 \pm 3.48}$	2.91 ± 0.76	$\mathbf{3.59 \pm 0.30}$	4.70 ± 0.39
	NESE (ours)	17.50 ± 2.63	$\mathbf{2.97 \pm 0.71}$	3.50 ± 0.31	$\mathbf{4.70 \pm 0.34}$
Mism.	SGMSE+ [1]	10.13 ± 5.68	2.62 ± 0.60	3.51 ± 0.29	4.52 ± 0.33
	SB-VE [7]	17.71 ± 4.05	2.00 ± 0.61	3.56 ± 0.29	4.32 ± 0.56
	NESE (ours)	14.79 ± 3.05	2.69 ± 0.63	3.55 ± 0.31	4.42 ± 0.47

Table 3: Speech enhancement performance on EARS-WHAM_v2. Models marked with * were trained at 48 kHz and require upsampling of test files before enhancement and downsampling after processing. Values indicate mean and standard deviation.

	SI-SDR	PESQ	DNSMOS	NISQA
Clean	∞	4.64 ± 0.00	3.89 ± 0.28	4.09 ± 0.83
Noisy	5.36 ± 5.90	1.24 ± 0.21	2.73 ± 0.31	1.95 ± 0.71
SGMSE+* [1]	14.52 ± 5.07	$\mathbf{2.19 \pm 0.59}$	$\mathbf{3.79 \pm 0.29}$	$\mathbf{4.08 \pm 0.80}$
SB-VE* [7]	12.40 ± 5.57	1.49 ± 0.35	3.54 ± 0.36	3.37 ± 0.83
NESE (ours)	$\mathbf{14.77 \pm 3.69}$	2.14 ± 0.61	3.74 ± 0.32	3.94 ± 0.86

Next, we compare the two skip configurations, examine different auxiliary loss weightings, and evaluate both matched and mismatched training-testing scenarios. Table 1 presents the speech enhancement results for all eight models, evaluated on the VoiceBank-DEMAND test set. The results show that different skip scalings improve different speech enhancement metrics. Specifically, setting $c_{\text{skip}}(t) = 1$ benefits PESQ optimization, while $c_{\text{skip}}(t) = 0$ yields slight improvements in SI-SDR and the non-intrusive metrics. Using the ℓ_1-loss in the time domain provides minor improvements in the matched case, but leads to deterioration in mismatched test scenarios. We select the $c_{\text{skip}} = 1$, $\alpha = 0.001$ as our default configuration for NESE.

Table 2 compares our method with other diffusion-based speech enhancement models, including SGMSE+ [1] and SB-VE [7]. The results show that our method performs competitively, with no statistically significant differences indicating that one method is better than the others. In the mismatched conditions, however, NESE demonstrates superior robustness, highlighting the strong generalization capabilities of our proposed generative speech enhancement method.

Table 3 presents results obtained using the EARS-WHAM_v2 dataset. It should be noted that we use pretrained checkpoints from [30], which were trained at 48 kHz. Accordingly, test files are upsampled to this sampling rate before enhancement and downsampled back to the original frequency after processing. NESE and SGMSE achieve comparable performance, whereas SB-VE struggles on this dataset, possibly due to the effects of the resampling procedure.

6. CONCLUSION

In this work, we introduced a magnitude-preserving network architecture for generative speech enhancement with diffusion bridges. We developed a preconditioning scheme for the denoiser model within a Schrödinger bridge framework and introduced two skip scaling strategies. These allow the network to use either environmental noise or clean speech as the training target, leading to different behaviors across various speech enhancement metrics. For the first time, we investigated the impact of EMA in diffusion-based speech enhancement by employing a method to approximate different EMA profiles post-training. While EMA is known to improve system-level metrics such as FID in image generation, our results show that longer EMA lengths can adversely affect instance-based metrics in speech enhancement.

REFERENCES

[1] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[2] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion models for audio restoration: A review," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2025.

[3] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Causal diffusion models for generalized speech enhancement," *IEEE Open Journal of Signal Processing*, 2024.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[5] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *International Conference on Learning Representations*, 2021.

[6] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," *arXiv preprint arXiv:2302.14748*, 2023.

[7] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement," in *Proceedings of Interspeech*, 2024, pp. 1175–1179.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, 2022.

[11] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Investigating the design space of diffusion models for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 4486–4500, 2024.

[12] T. Chen, G.-H. Liu, and E. Theodorou, "Likelihood training of Schrödinger bridge using forward-backward SDEs theory," in *International Conference on Learning Representations*, 2021.

[13] G.-H. Liu, A. Vahdat, D.-A. Huang, E. Theodorou, W. Nie, and A. Anandkumar, "I^2SB: Image-to-image Schrödinger bridge," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 042–22 062.

[14] T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine, "Analyzing and improving the training dynamics of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 174–24 184.

[15] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[16] D. de Oliveira, J. Richter, J.-M. Lemercier, S. Welker, and T. Gerkmann, "Non-intrusive speech quality assessment with diffusion models trained on clean speech," *arXiv preprint arXiv:2410.17834*, 2024.

[17] J. Richter, D. de Oliveira, and T. Gerkmann, "Investigating training objectives for generative speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.

[18] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *ISCA Speech Synthesis Workshop*, pp. 146–152, 2016.

[19] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proceedings of Interspeech*, 2024.

[20] E. Schrödinger, "Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique," in *Annales de l'institut Henri Poincaré*, vol. 2, no. 4, 1932, pp. 269–310.

[21] C. Léonard, "A survey of the Schrödinger problem and some of its connections with optimal transport," *Discrete and Continuous Dynamical Systems-Series A*, vol. 34, no. 4, pp. 1533–1574, 2014.

[22] M. Pavon and A. Wakolbinger, "On free energy, stochastic control, and Schrödinger processes," in *Modeling, Estimation and Control of Systems with Uncertainty: Proceedings of a Conference held in Sopron, Hungary, September 1990*. Springer, 1991, pp. 334–348.

[23] E. Hopf, "The partial differential equation $u_t + uu_x = \mu_{xx}$," *Communications on Pure and Applied Mathematics*, vol. 3, no. 3, pp. 201–230, 1950.

[24] J. D. Cole, "On a quasi-linear parabolic equation occurring in aerodynamics," *Quarterly of Applied Mathematics*, vol. 9, no. 3, pp. 225–236, 1951.

[25] E. Nelson, *Dynamical Theories of Brownian Motion*. Princeton University Press, 1967, vol. 3.

[26] L. Richter and J. Berner, "Improved sampling via learned diffusions," in *International Conference on Learning Representations*, 2024.

[27] C. Bunne, Y.-P. Hsieh, M. Cuturi, and A. Krause, "The Schrödinger bridge between gaussian measures has a closed form," in *International Conference on Artificial Intelligence and Statistics*, 2023, pp. 5802–5833.

[28] Z. Chen, G. He, K. Zheng, X. Tan, and J. Zhu, "Schrödinger bridges beat diffusion models on text-to-speech synthesis," *arXiv preprint arXiv:2312.03491*, 2023.

[29] D. Kingma, L. Ba *et al.*, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[30] J. Richter and T. Gerkmann, "Diffusion-based speech enhancement: Demonstration of performance and generalization," *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

A.6 Analysis of Generative Speech Enhancement and Comparison to Predictive Approaches

A.6.1 Publication: Diffusion-based speech enhancement: Demonstration of performance and generalization

Abstract

This demo presents advanced techniques in speech enhancement using deep generative models. It highlights the generalization capabilities of score-based generative models for speech enhancement and compares directly with Schrödinger bridge approaches. The presented methods focus on generating high-quality super-wideband speech at a sampling rate of 48 kHz. Participants will record speech using a single microphone in a noisy environment, such as a conference venue. These recordings will then be enhanced and played back through headphones, demonstrating the model's effectiveness in improving speech quality and intelligibility.

Reference

Julius Richter, Timo Gerkmann, "Diffusion-based Speech Enhancement: Demonstration of Performance and Generalization," *Audio Imagination NeurIPS 2024 Workshop*, Vancouver, Canada, Dec. 2024.

Author's Contributions

Julius Richter is the main author of the paper and was responsible for the implementation of the presented methods. Timo Gerkmann contributed through insights into experimental validation, mathematical derivations, and manuscript review.

Diffusion-based Speech Enhancement: Demonstration of Performance and Generalization

Julius Richter*
Signal Processing Group
University of Hamburg
Germany
julius.richter@uni-hamburg.de

Timo Gerkmann
Signal Processing Group
University of Hamburg
Germany
timo.gerkmann@uni-hamburg.de

Abstract

This demo presents advanced techniques in speech enhancement using deep generative models. It highlights the generalization capabilities of score-based generative models for speech enhancement and compares directly with Schrödinger bridge approaches. The presented methods focus on generating high-quality super-wideband speech at a sampling rate of 48 kHz. Participants will record speech using a single microphone in a noisy environment, such as a conference venue. These recordings will then be enhanced and played back through headphones, demonstrating the model's effectiveness in improving speech quality and intelligibility.

1 Introduction

Generative speech enhancement has recently shown promising advancements in improving speech quality in noisy environments [1]. Several diffusion-based frameworks exist, such as score-based generative models for speech enhancement (SGMSE) [2] and the Schrödinger bridge [3, 4].

This demo showcases the generalization capabilities of diffusion-based models for speech enhancement. A comparative evaluation is presented for SGMSE and the Schrödinger bridge approach. All presented methods generate high-quality super-wideband speech enhancement at a sampling rate of 48 kHz. Participants are asked to record their speech using a single microphone in a noisy environment, such as a conference venue. Once the speech is recorded, it undergoes enhancement and is subsequently played back through headphones.

2 Methods

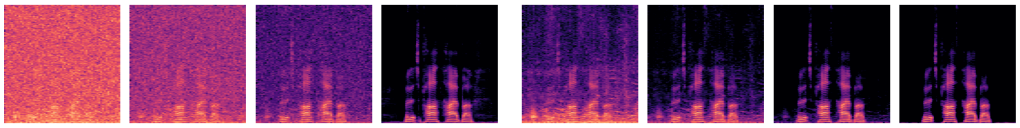
We represent audio signals in the time-frequency domain using the short-time Fourier transform (STFT). Thus, we have complex spectrograms $\bar{\mathbf{x}} \in \mathbb{C}^{T \times F}$ where T is the number of time frames, and F is the number of frequency bins. We apply an amplitude transformation $\widetilde{c} = \beta |c|^{\alpha} \mathrm{e}^{i \angle(c)}$ to all STFT coefficients c, where $\angle(\cdot)$ represents the angle of a complex number, $\alpha \in (0, 1]$, and $\beta \in \mathbb{R}$ [2]. Finally, we flatten $\bar{\mathbf{x}}$ into a vector of dimension $d = TF$, resulting in the input representation $\mathbf{x} \in \mathbb{C}^d$.

2.1 Score-based generative models for speech enhancement (SGMSE)

In SGMSE [2], we use a task-adapted diffusion process for the conditional generation of clean speech $\mathbf{x}_0 \in \mathbb{C}^d$ given a noisy input $\mathbf{y} \in \mathbb{C}^d$. Using the continuous-time formulation for diffusion models [5],

*We acknowledge the support by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) SDE solver for SGMSE. (b) ODE solver for the Schrödinger bridge.

Figure 1: Comparison between SGMSE and Schrödinger bridge.

	POLQA	**SI-SDR** [dB]	**PESQ**	**ESTOI**	**SIGMOS**	**DNSMOS**
Noisy	1.71 ± 0.56	5.98 ± 6.10	1.24 ± 0.22	0.49 ± 0.15	1.95 ± 0.39	2.74 ± 0.29
SGMSE+ [2]	3.40 ± 0.73	16.78 ± 4.47	2.50 ± 0.62	$\mathbf{0.73 \pm 0.13}$	3.41 ± 0.41	$\mathbf{3.88 \pm 0.26}$
SB [3]	3.46 ± 0.84	$\mathbf{17.85 \pm 4.37}$	2.33 ± 0.67	$\mathbf{0.73 \pm 0.13}$	$\mathbf{3.44 \pm 0.42}$	3.83 ± 0.27
SB-PESQ [4]	$\mathbf{3.71 \pm 0.80}$	16.29 ± 4.16	$\mathbf{3.09 \pm 0.63}$	$\mathbf{0.73 \pm 0.13}$	3.18 ± 0.44	3.72 ± 0.28

Table 1: Speech enhancement performance on EARS-WHAM. Mean and standard deviation.

the forward process is modeled as the solution to the stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}\,, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{C}^d$ denotes the process state at time $t \in [0, 1]$, $\gamma \in \mathbb{R}$ controls the transition from $\mathbf{x}_0$ to $\mathbf{y}$, and $g(t) \in \mathbb{R}$ is the diffusion coefficient that controls the amount of Gaussian noise induced by a standard Wiener process $\mathbf{w}$.

The forward process can be time-inverted [5], resulting in a corresponding reverse process

$$\mathrm{d}\mathbf{x}_t = [-\gamma(\mathbf{y} - \mathbf{x}_t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}\,, \tag{2}$$

where $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y})$ is the conditional score, and $\bar{\mathbf{w}}$ is the Wiener process backward in time.

The score function is typically intractable and approximated by a score model $\mathbf{s}_\theta$ with parameters θ. Once the score model is trained, the reverse process in Eq. (2) can be solved by replacing the score function with its approximation $\mathbf{s}_\theta$ and using a numerical SDE solver. Fig. 1a illustrates an example of reverse sampling utilizing an SDE solver. We use the same hyperparameters as in [6].

2.2 Schrödinger bridge for speech enhancement

The Schrödinger bridge (SB) for speech enhancement involves a pair of symmetric forward and reverse SDEs [3]. The general SB formulation is typically intractable, but closed-form solutions exist for certain cases, such as Gaussian boundary conditions [7].

Training the SB involves directly predicting the clean data $\mathbf{x}_0$, which differs from SGMSE, where Gaussian noise is predicted instead. This allows the incorporation of perceptual loss terms [4]. The reverse SDE is solved during inference using an ODE sampler [7]. The SB approach benefits from its optimal transport characteristics, interpolating efficiently between clean and noisy signals, as shown in Fig. 1b. We use the same hyperparameters as in [4], the same STFT as in [6], and $\alpha_{\mathrm{P}} = 10^{-4}$.

3 Experimental Setup and Results

We train all models using the EARS-WHAM dataset [6], which comprises 86.8 hours of audio. We report results on the EARS-WHAM test set in Table 1. In the demo, speech is recorded using a microphone connected to a sound card, which is in turn connected to a laptop. The laptop is equipped with a GPU to facilitate efficient processing. After the audio enhancement process is complete, we play back the enhanced audio through headphones. An example of enhanced audio from a noisy YouTube video is available at the following link: `https://www.youtube.com/watch?v=H5FiO0JxPK4`.

4 Conclusion

This demo illustrates the advanced capabilities of generative speech enhancement, showcasing its ability to improve speech quality and intelligibility in real-world, noisy environments. We highlight the generation of high-quality super-wideband speech, thereby emphasizing the potential of these techniques for practical applications.

References

[1] Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. Diffusion models for audio restoration. *Signal Processing Magazin, accepted paper*, 2025.

[2] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

[3] Ante Jukić, Roman Korostik, Jagadeesh Balam, and Boris Ginsburg. Schrödinger bridge for generative speech enhancement. In *ISCA Interspeech*, pages 1175–1179, 2024.

[4] Julius Richter, Danilo de Oliveira, and Timo Gerkmann. Investigating training objectives for generative speech enhancement. *arXiv preprint arXiv:2409.10753*, 2024.

[5] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

[6] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinjii Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *ISCA Interspeech*, pages 4873–4877, 2024.

[7] Zehua Chen, Guande He, Kaiwen Zheng, Xu Tan, and Jun Zhu. Schrödinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*, 2023.

A.6.2 Publication: StoRM: A Diffusion-based Stochastic Regeneration Model for Speech Enhancement and Dereverberation

Abstract

In this work, we build upon our previous publication and use diffusion-based generative models for speech enhancement. We present a detailed overview of the diffusion process that is based on a stochastic differential equation and delve into an extensive theoretical examination of its implications. Opposed to usual conditional generation tasks, we do not start the reverse process from pure Gaussian noise but from a mixture of noisy speech and Gaussian noise. This matches our forward process which moves from clean speech to noisy speech by including a drift term. We show that this procedure enables using only 30 diffusion steps to generate high-quality clean speech estimates. By adapting the network architecture, we are able to significantly improve the speech enhancement performance, indicating that the network, rather than the formalism, was the main limitation of our original approach. In an extensive cross-dataset evaluation, we show that the improved method can compete with recent discriminative models and achieves better generalization when evaluating on a different corpus than used for training. We complement the results with an instrumental evaluation using real-world noisy recordings and a listening experiment, in which our proposed method is rated best. Examining different sampler configurations for solving the reverse process allows us to balance the performance and computational speed of the proposed method. Moreover, we show that the proposed method is also suitable for dereverberation and thus not limited to additive background noise removal.

Reference

Jean-Marie Lemercier, Julius Richter, Simon Welker, Timo Gerkmann, "StoRM: A Diffusion-based Stochastic Regeneration Model for Speech Enhancement and Dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2724 -2737, 2023. DOI: 10.1109/TASLP.2023.3294692

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2023 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Jean-Marie Lemercier is the lead author of this publication, having implemented all the algorithms, trained the neural networks utilized in the study, conducted the experimental validation, and authored the manuscript. Julius Richter and Simon Welker contributed by providing feedback on all methodologies through discussions and assisted in reviewing the manuscript. Timo Gerkmann offered insights on experimental validation and mathematical derivations, in addition to reviewing the manuscript.

StoRM: A Diffusion-based Stochastic Regeneration Model for Speech Enhancement and Dereverberation

Jean-Marie Lemercier, *Student Member, IEEE*, Julius Richter, *Student Member, IEEE*, Simon Welker, *Student Member, IEEE*, Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—Diffusion models have shown a great ability at bridging the performance gap between predictive and generative approaches for speech enhancement. We have shown that they may even outperform their predictive counterparts for non-additive corruption types or when they are evaluated on mismatched conditions. However, diffusion models suffer from a high computational burden, mainly as they require to run a neural network for each reverse diffusion step, whereas predictive approaches only require one pass. As diffusion models are generative approaches they may also produce vocalizing and breathing artifacts in adverse conditions. In comparison, in such difficult scenarios, predictive models typically do not produce such artifacts but tend to distort the target speech instead, thereby degrading the speech quality. In this work, we present a stochastic regeneration approach where an estimate given by a predictive model is provided as a guide for further diffusion. We show that the proposed approach uses the predictive model to remove the vocalizing and breathing artifacts while producing very high quality samples thanks to the diffusion model, even in adverse conditions. We further show that this approach enables to use lighter sampling schemes with fewer diffusion steps without sacrificing quality, thus lifting the computational burden by an order of magnitude. Source code and audio examples are available online[1].**

***Index Terms*—score-based generative models, diffusion models, speech enhancement, speech dereverberation, predictive learning.**

I. Introduction

In real-life scenarios and modern communication devices, clean speech sources are often polluted by background noise, interfering speakers, room acoustics and codec degradation [1], [2]. We refer to this phenomenon as *speech corruption*, and denote by *speech restoration* the art of recovering clean speech from the corrupted signal [3]. On the one hand, traditional speech restoration methods leverage the statistical properties of the target and interference signals in various domains e.g. time, spectrum, cepstrum or spatial distribution [4]. On the other hand, machine learning techniques try to learn these statistical properties and how to exploit them from data [5]. Machine learning algortihms can be categorized into predictive (also called discriminative) approaches and generative approaches. We will choose the term *predictive* over *discriminative* as it fits both classification and regression tasks [6]. The field of speech restoration is dominated by predictive approaches that use supervised learning to learn a single best deterministic mapping between corrupted speech $\mathbf{y}$ and the corresponding clean speech target $\mathbf{x}$ [5]. These methods include for instance time-frequency (T-F) masking [7], time domain methods [8], [9] or direct spectro-temporal mapping [10]. They have contributed to drastically increasing the quality of speech restoration algorithms. However, they can distort target speech and suffer from generalizability issues [11], [12].

In contrast, generative models implicitly or explicitly learn the target distribution and allow to generate multiple valid estimates instead of a single best estimate as for predictive approaches [6]. Generative approaches include variational auto-encoders (VAEs) learning explicit density estimations [13]–[16], normalizing flows adding invertible transforms to obtain tractable marginal likelihoods [17], [18], generative adversarial networks (GANs) estimating implicit distributions [19], [20] and diffusion approaches [21]–[23]. We talk of *conditional* generative models when a covariate $\mathbf{c}$ is used to guide the generation, leading to the conditional distribution $p(\mathbf{x}|\mathbf{c})$ [6]. This conditioning can either be another modality describing the data (e.g. $\mathbf{c}$ could be video when $\mathbf{x}$ is speech), or a modified version of the data, an obvious example being corrupted speech $\mathbf{y}$ when the underlying task is speech restoration. By integrating stochasticity in their latent structure, generative models can capture the inherent uncertainty of the data distribution and produce realistic samples belonging to that distribution rather than a mean of optimal candidates [6]. In doing so, they may obtain better perceptual metrics at the cost of higher point-wise distortion [24]. In the imaging domain, it was observed that predictive approaches tend to brush over the fine-grained details of the considered domain [24], [25]. Furthermore, predictive models may result in limited generalization abilities towards unseen noise types or speakers as compared to generative models, which is demonstrated for diffusion-based generative speech enhancement in [12].

We focus in this work on such diffusion-based generative models, or simply *diffusion models*, which have met great

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380, the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002.

Simon Welker is with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany, and with the Center for Free-Electron Laser Science, DESY, 22607 Hamburg, Germany (e-mail: simon.welker@uni-hamburg.de).

The other authors are with the Signal Processing Group, Department of Informatics, Universität Hamburg, 22527 Hamburg Germany (e-mail: {jeanmarie.lemercier; julius.richter; timo.gerkmann}@uni-hamburg.de).

[1]https://uhh.de/inf-sp-storm

success in generating high-quality samples of natural images [21]–[23], [26]. Diffusion models use a *forward process* to slowly turn data into a tractable prior, usually a standard normal distribution, and train a neural network to solve the *reverse process* to generate clean data from this prior [27]. These diffusion models can also be used for conditional generation in restoration tasks, which has recently been proposed for speech processing tasks such as enhancement and dereverberation [12], [28]–[30] as well as bandwidth extension [11], [31].

One limiting aspect of diffusion models is their heavy computational burden. Several steps are needed for reverse diffusion, each of them calling the neural network used for score estimation. Much effort has been recently put into reducing this number of steps, either by optimization of the reverse noise schedule [32], modifications in the formulation of the diffusion processes [33], [34], or projection into a latent space [35] or a reduced subspace [36]. We also observed in past experiments that our previously proposed diffusion model is prone to confuse phonemes and generate vocalizing artifacts when facing very adverse conditions. This is due to the generative behaviour of the model under high uncertainty over the presence or nature of speech, and this naturally leads to a degradation, e.g. in automatic speech recognition (ASR).

In this work, we propose a *stochastic regeneration* scheme combining predictive and generative models to produce high quality samples while reducing the computational burden of diffusion models and their tendency to generate unwanted artifacts. We propose to first use a predictive approach to estimate a restored version of the corrupted speech. This estimate is then used as a guide by a diffusion model, which requires only a few diffusion steps to output a final clean speech estimation where the distortions introduced by the predictive stage are corrected while vocalizing artifacts and phonetic confusions are avoided. Both listening experiments and instrumental metrics confirm an impressive state-of-the-art perceptual quality of our proposed approach. Other refinement approaches using diffusion models were recently proposed. The *stochastic refinement* approach [24], [37] subtracts the output of the predictive model from the corrupted speech, and this residual is used for further estimation by a diffusion model. We argue hereafter that learning the residual is however a hard task and demonstrate that our approach outperforms this stochastic refinement in terms of instrumentally measured speech quality. Another refinement approach using diffusion models is *denoising diffusion restoration models* [38]–[40], where the corruption operator is assumed to be known (or at least its singular value decomposition) and is used to modify the reverse diffusion process at inference time.

We evaluate our proposed approach for speech enhancement with low input signal-to-noise ratios (SNRs) and speech dereverberation, using clean speech from the WSJ0 corpus [41]. We also show ASR results on the TIMIT dataset [42], and report results on the standardized Voicebank/DEMAND dataset [43]. Ablation studies are performed on sampling efficiency, intial predictor mismatch and training strategy.

II. Score-based Diffusion Models

Diffusion models originally use discrete-time diffusion processes modeled by Markov chains [22]. They have been recently extended to continuous-time diffusion processes formulated by stochastic differential equations (SDEs) in [44], allowing for new training paradigms such as score matching [45], [46]. This class model is subsequently denoted as *score-based diffusion models*. Score-based diffusion models are defined by three components: a forward diffusion process, a score function estimator, and a sampling method for inference.

A. Forward and reverse processes

The stochastic forward process $\{\mathbf{x}_\tau\}_{\tau=0}^{T}$ used in score-based diffusion models is defined as an Itô SDE [44], [47]:

$$\mathrm{d}\mathbf{x}_\tau = \mathbf{f}(\mathbf{x}_\tau, \tau)\mathrm{d}\tau + g(\tau)\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{x}_\tau$ is the current state of the process indexed by $\tau \in [0, T]$ with the initial condition representing clean speech $\mathbf{x}_0 = \mathbf{x}$. The continuous *diffusion time* variable τ relates to the progress of the stochastic process and should not be mistaken for our usual notion of *signal time*. As our process is defined in the complex spectrogram domain, independently for each T-F bin, the variables in bold are vectors in $\mathbb{C}^d$ containing the coefficients of a flattened complex spectrogram— with d the product of the time and frequency dimensions— whereas variables in regular font represent real scalar values. The set $\{\mathbf{x}_\tau\}_{\tau\in]0,T[}$ can be seen as latent variables used to parameterize the conditional distribution $p(\mathbf{x}_\tau|\mathbf{x}_0, \mathbf{y})$. The stochastic process $\mathbf{w}$ denotes a standard d-dimensional Brownian motion, that is, $\mathrm{d}\mathbf{w}$ is a zero-mean Gaussian variable with standard deviation $\mathrm{d}\tau$ for each T-F bin.

The *drift* function $\mathbf{f}$ and *diffusion* coefficient g as well as the initial condition $\mathbf{x}_0$ and the final diffusion time T uniquely define the Itô process $\{\mathbf{x}_\tau\}_{\tau=0}^{T}$ [47]. Under some regularity conditions on $\mathbf{f}, g$ allowing a unique and smooth solution to the Kolmogorov equations associated to (1), the reverse process $\{\mathbf{x}_\tau\}_{\tau=T}^{0}$ is another diffusion process defined as the solution of the following SDE [44], [48]:

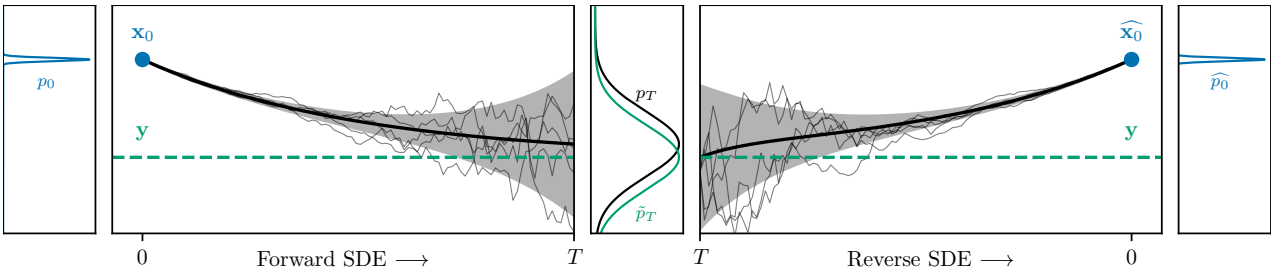
$$\mathrm{d}\mathbf{x}_\tau = \left[-\mathbf{f}(\mathbf{x}_\tau, \tau) + g(\tau)^2 \nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)\right] \mathrm{d}\tau + g(\tau)\mathrm{d}\bar{\mathbf{w}}, \tag{2}$$

where $\mathrm{d}\bar{\mathbf{w}}$ is a d-dimensional Brownian motion for the time flowing in reverse and $\nabla_{\mathbf{x}_\tau} \log p_\tau(\mathbf{x}_\tau)$ is the *score function*, i.e. the gradient of the logarithm data distribution for the current process state $\mathbf{x}_\tau$.

Speech restoration tasks can be regarded either as one-to-one mapping tasks between corrupted speech $\mathbf{y}$ and $\mathbf{x}_0$, which leads to predictive modelling; or as conditional generation tasks, i.e. generation of $\mathbf{x}_0$ conditioned on $\mathbf{y}$. Previous diffusion-based approaches proposed to condition the process explicitly within the neural network [49] or through guided classification [26]. In [28], the conditioning is directly incorporated into the diffusion process by defining the forward process as the solution to the following SDE:

$$\mathrm{d}\mathbf{x}_\tau = \underbrace{\gamma(\mathbf{y} - \mathbf{x}_\tau)}_{:=\mathbf{f}(\mathbf{x}_\tau, \mathbf{y})} \mathrm{d}\tau + \underbrace{\left[\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\tau} \sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}\right]}_{:=g(\tau)} \mathrm{d}\mathbf{w}. \tag{3}$$

Fig. 1: *Visualization of the forward and backward processes in* (3)*. Mean curve* (5) *is in solid black and variance* (6) *is represented by the greyed area. Several realizations of the diffusion process are represented by thin black lines. The mismatch between p_τ centered on $\mathbf{x}_\tau$ and $\tilde{p}_\tau$ centered on $\mathbf{y}$ comes from the fact that the mean in* (5) *can not reach $\mathbf{y}$ in finite time. This mismatch causes unavoidable bias in the reverse process, even were the score perfectly known.*



This equation belongs to the class of Ornstein-Uhlenbeck SDEs [47], a subclass of Itô SDEs in which the drift function $\mathbf{f}$ is affine in $\mathbf{x}_\tau$ and does not depend on τ, and the diffusion coefficient g only depends on τ. The equation introduces a *stiffness* hyperparameter γ controlling the slope of the decay from $\mathbf{y}$ to $\mathbf{x}_0$, and $\sigma_{\min}$ and $\sigma_{\max}$ are two hyperparameters controlling the *noise scheduling*, that is, the amount of Gaussian white noise injected at each timestep of the process.

The interpretation of our forward process in Eq. (3), visualized on Fig. 1, is as follows: at each time step and for each T-F bin independently, an infinitesimal amount of corruption is added to the current process state $\mathbf{x}_\tau$, along with Gaussian noise with standard deviation $g(\tau)\mathrm{d}\tau$. Therefore, the mean of the current process decays exponentially towards $\mathbf{y}$ while the variance increases as in the variance-exploding scheme of Song et al. [44], leading to a final distribution $\mathbf{x}_\tau$ which is the corrupted signal $\mathbf{y}$ with some additional Gaussian noise. Given an initial condition $\mathbf{x}_0$ and the covariate $\mathbf{y}$, the solution to (3) admits the following complex Gaussian distribution for the process state $\mathbf{x}_\tau$ called *perturbation kernel*:

$$p_{0,\tau}(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y}) = \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_\tau; \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau), \sigma(\tau)^2\mathbf{I}\right), \quad (4)$$

Following [50], we determine closed-form solutions for the mean $\boldsymbol{\mu}$ and variance $\sigma(\tau)^2$:

$$\boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau) = \mathrm{e}^{-\gamma\tau}\mathbf{x}_0 + (1-\mathrm{e}^{-\gamma\tau})\mathbf{y}\,, \quad (5)$$

$$\sigma(\tau)^2 = \frac{\sigma_{\min}^2\left(\left(\sigma_{\max}/\sigma_{\min}\right)^{2\tau} - \mathrm{e}^{-2\gamma\tau}\right)\log\left(\sigma_{\max}/\sigma_{\min}\right)}{\gamma + \log\left(\sigma_{\max}/\sigma_{\min}\right)}. \quad (6)$$

B. Score function estimator

When performing inference, one tries to solve the reverse SDE in Eq. (2). In the general case, the score function $\nabla_{\mathbf{x}_\tau}\log p_\tau(\mathbf{x}_\tau)$ is not readily available, it can however be estimated by a deep neural network (DNN) $\mathbf{s}_\phi$ called the *score model*. Given the simple Gaussian form of the perturbation kernel $p_{0,\tau}(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y})$ (4) and the regularity conditions exhibited by the mean and variance, a *denoising score matching* objective can be used to train the score model $\mathbf{s}_\phi$ [45], [46].

The score function of the perturbation kernel is:

$$\nabla_{\mathbf{x}_\tau}\log p_{0,\tau}(\mathbf{x}_\tau|\mathbf{x}_0,\mathbf{y}) = -\frac{\mathbf{x}_\tau - \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau)}{\sigma(\tau)^2}. \quad (7)$$

Once a clean utterance $\mathbf{x}_0$ and noisy utterance $\mathbf{y}$ are picked in the training set, the current process state is obtained as $\mathbf{x}_\tau = \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},\tau) + \sigma(\tau)\mathbf{z}$, with $\mathbf{z}\sim\mathcal{N}_{\mathbb{C}}(\mathbf{z};\mathbf{0},\mathbf{I})$. We can therefore write the denoising score matching objective as follows [44]:

$$\mathcal{J}^{(\mathrm{DSM})}(\phi) = \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),\mathbf{z},\mathbf{x}_\tau}\left[\left\|\mathbf{s}_\phi(\mathbf{x}_\tau,\mathbf{y},\tau) + \frac{\mathbf{z}}{\sigma(\tau)}\right\|_2^2\right]. \quad (8)$$

Here, we sample τ sampled uniformly in $[\tau_\epsilon, T]$ where τ_ϵ is a minimal diffusion time used to avoid numerical instabilities. This approach is analogous to the denoising objective used in the discrete-time formulation by [22], where one estimates the noise added at each step to learn the reverse process.

C. Inference through reverse sampling

At inference time, we first sample an initial condition of the reverse process, corresponding to $\mathbf{x}_T$, with:

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T;\mathbf{y},\sigma^2(T)\mathbf{I}), \quad (9)$$

This sample only approximates the training condition, as the final process distribution $p_T(\mathbf{x}_T)$ does not perfectly match $p(\mathbf{y})$ (see Fig. 1).

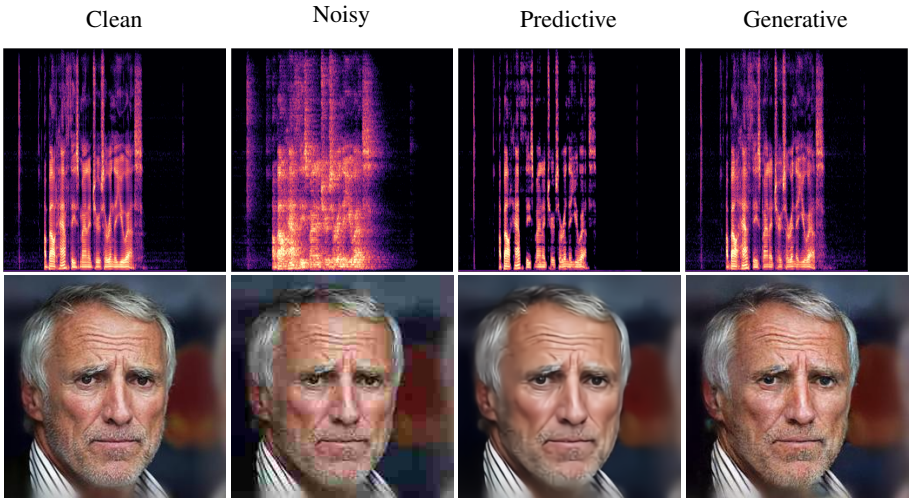
Conditional generation is then performed by solving the so-called *plug-in reverse SDE* from $\tau = T$ to $\tau = 0$, where the score function is replaced by its estimator $\mathbf{s}_\phi$, assuming the latter was trained e.g. according to Section II-B:

$$\mathrm{d}\mathbf{x}_\tau = \left[-\mathbf{f}(\mathbf{x}_\tau,\mathbf{y}) + g(\tau)^2\mathbf{s}_\phi(\mathbf{x}_\tau,\mathbf{y},\tau)\right]\mathrm{d}\tau + g(\tau)\mathrm{d}\bar{\mathbf{w}} \quad (10)$$

We use classical numerical solvers based on a discretization of (10) according to a uniform grid of N points on the interval $[0,T]$ (no minimal diffusion time is needed here). Classical solvers include the Euler-Maruyama method, higher-order single-step methods, and predictor-corrector sampler schemes [44]. In the latter, at each reverse step τ, the predictor uses a single-step method like Euler-Maruyama to generate $\mathbf{x}_\tau$, and the corrector uses the output of the score network to ensure consistency of the resulting sample with a marginal distribution consistent with the score estimate.

For notational convenience, we will denote by G_ϕ the generative model corresponding to the reverse diffusion process solver parameterized by the plug-in SDE (10) and the score network $\mathbf{s}_\phi$, such that the final estimate is $\widehat{\mathbf{x}} = G_\phi(\mathbf{y})$.

Fig. 2: *Visualization of samples obtained with predictive approach (NCSN++M, see Section IV) and generative model (SGMSE+M, see [12] and Section IV) for two ill-posed problems, namely speech dereverberation (top, from [11]) and JPEG artifact removal (bottom, from [25]). Spectrograms horizontal and vertical axes represent time and frequency respectively.*



III. Stochastic Regeneration with Diffusion Models

A. Predictive artifacts for images and spectrograms

Most problems in speech restoration (e.g. denoising without knowledge of the environmental noise signal, dereverberation or bandwidth extension) are ill-posed inverse problems. This means that either (i) it is not possible to exactly retrieve $\mathbf{x}$ given $\mathbf{y}$ (e.g. dereverberation with a known non-minimum-phase room impulse response); or (ii) many versions of clean speech $\mathbf{x}^{(i)}$ can correspond to the same corrupted speech $\mathbf{y}$ (e.g. bandwidth extension). Consider a predictive model D_θ trained with a L^2 regression objective $\mathbb{E}_{(\mathbf{x},\mathbf{y})}||\mathbf{x} - D_\theta(\mathbf{y})||_2^2$. Because of the expectation over all training examples, an optimal predictive model learns the mapping to the posterior mean $\mathbf{y} \to \mathbb{E}[\mathbf{x}|\mathbf{y}]$, thereby minimizing the *average* distortion over all training examples. This phenomenon is known as *regression to the mean* [6], [24]. This can be problematic if the posterior distribution $p(\mathbf{x}|\mathbf{y})$ has an intricate structure and is thus not well represented by its mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ (see Figure 5 here, or Figure 1 in [51]).

In image processing problems, this translates to predictive approaches being incapable of reproducing fine-grained details like e.g. edges and hair structure in human portraits [24], [25]. Our interpretation is that, given that these regions have the highest variability across natural image data [52], mapping directly to the posterior mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ will smooth out these details by mistaking them for noise. Therefore the predictive model will output a sample which does not necessarily lie on the posterior data manifold. When training predictive models for spectrogram estimation, we also observe that the models tend to introduce distortions in the target speech when the corruption level is high, leading to *overdenoising* effects and loss of output resolution [11], [53].

The link between these observations in the imaging and speech domains is the following: performing speech restoration in the spectrogram domain can be seen as image processing (with two pixel dimensions representing the real and imaginary parts instead of three for RGB image processing). The distortions observed in the predictive model output are a smoothing effect *in the spectrogram seen as an image*. This results in removal of fine-grained detail corresponding to quiet regions (i.e. low luminosity detail in images) and onsets or offsets (i.e. edges in images) in the spectrogram. This is visualized in the third column of Fig. 2, where we directly compare spectrograms from our previous study [11] with images in Welker et al. [25, Fig. 1] reproduced here.

Several paradigms using generative modelling can be envisaged to correct this bias of the predictive model without having to resort to a full-fletched computationally heavy diffusion-based generative model. Next, we present two of these approaches, namely *stochastic refinement* by Whang et al. [24] and *stochastic regeneration* which we propose here.

B. Stochastic refinement

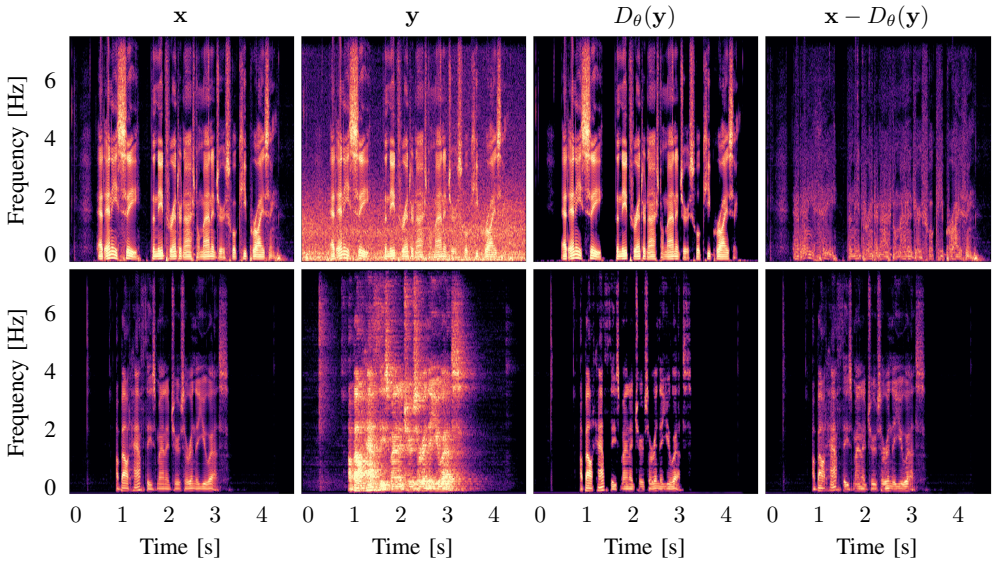
Instead of solving the reverse diffusion process from noisy speech to a clean speech estimate, the *stochastic refinement* approach by Whang et al. [24] uses both a predictive approach and a generative diffusion model for efficient inference.

A predictive model D_θ serves as an *initial predictor* producing an estimate $D_\theta(\mathbf{y})$. This estimate often lacks fine-grained detail and has significant target speech distortions, especially for corruption models like reverberation [11]. Let us write the predictive model output as:

$$D_\theta(\mathbf{y}) = \mathbf{x} - \mathbf{x}^{(\mathrm{dis})} + \tilde{\mathbf{n}}. \tag{11}$$

The *target distortion* $\mathbf{x}^{(\mathrm{dis})}$ is the artifact introduced by the predictive model: it contains target cues that were mistaken for corruption by the model, and consequently distorted. The residual corruption $\tilde{\mathbf{n}}$ is what remains of the interference (e.g. noise or reverberation) after being processed by the

Fig. 3: *Log-energy spectrograms of clean, noisy, processed and residual utterances for denoising (top) and dereverberation (bottom). The predictor used is NCSN++M .*



model. There is behind this decomposition an underlying orthogonality assumption between $\mathbf{x}$ and $\mathbf{n}$, which implies orthogonality between $\mathbf{x}^{(\text{dis})}$ and $\tilde{\mathbf{n}}$ [54].

A diffusion-based generative model G_ϕ is then used to learn the distribution of the *ideal residue* $\mathbf{r_x} = \mathbf{x} - D_\theta(\mathbf{y})$, starting from the *noisy residue* $\mathbf{r_y} = \mathbf{y} - D_\theta(\mathbf{y})$. Finally, the ideal residue estimate is added to the predictor estimate:

$$\widehat{\mathbf{x}} = D_\theta(\mathbf{y}) + \widehat{\mathbf{r_x}} \tag{12}$$
$$= D_\theta(\mathbf{y}) + G_\phi(\mathbf{y} - D_\theta(\mathbf{y})) \tag{13}$$

Results in [24], [37] seem to indicate that this stochastic refinement approach performs as expected, outperforming the initial predictor on perceptual metrics and the pure generative approach with fewer diffusion steps. However, we argue that learning the residual is suboptimal as the residual data distribution $p(\mathbf{r_x})$ does not have a structure like the target data distribution $p(\mathbf{x})$. Indeed, using (11), one can rewrite $\mathbf{r_x}$ as:

$$\mathbf{r_x} = \mathbf{x} - D_\theta(\mathbf{y})$$
$$= \mathbf{x}^{(\text{dis})} - \tilde{\mathbf{n}}, \tag{14}$$

and notice that the distribution of $\mathbf{r_x}$ highly depends on the choice of the predictive model as well as on the task, which does not assure a structured distribution in the general case. We show examples in Fig. 3 of residuals generated by predictive models for speech enhancement and dereverberation, which confirm this observation. For dereverberation (similarly for deblurring, shown in [24]), the residue has an overall structure somewhat similar to the target, because of the convolutional corruption model. However, the formants structure is severely degraded. For denoising, the residue has no clear structure (compared to e.g. clean speech) and we thus argue that it cannot be easily estimated by a generative process. In Whang et al. [24], it is shown that the residual distribution has lower entropy per pixel than the original distribution, which makes learning the residual easier. This is also true for our denoising and dereverberation tasks. However the pointwise entropy of a distribution relates to the quantity of information that needs to be learnt by the model and does not capture global structures in the data which can actually help facilitate training. Most importantly, when rewriting the noisy residue $\mathbf{r_y}$ as:

$$\mathbf{r_y} = \mathbf{y} - D_\theta(\mathbf{y})$$
$$= \mathbf{x}^{(\text{dis})} + \mathbf{n} - \tilde{\mathbf{n}}, \tag{15}$$

one notices that the resulting *a priori* SNR of the starting point of the reverse process (without accounting for the added Gaussian noise) is very low, as $||\mathbf{x}^{(\text{dis})}||, ||\tilde{\mathbf{n}}|| \ll ||\mathbf{n}||$ for low-enough SNRs and good-enough initial predictor. This makes learning difficult, and we therefore propose to use a different refinement process that we denote as *stochastic regeneration*.

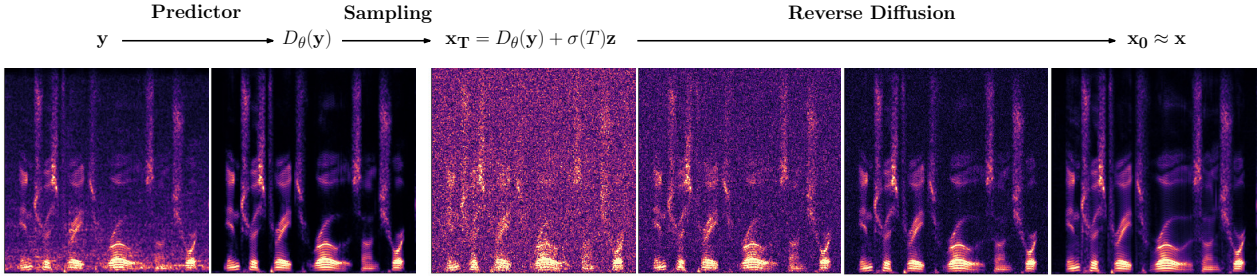
C. Stochastic regeneration

For *stochastic regeneration* we propose to cascade the predictive model D_θ and the generative diffusion model G_ϕ. The generative model learns to *regenerate* the clean speech based on the distorted version provided by the predictive approach. This is conceptually different from the stochastic refinement approach, where the target cues exist in the residual (but are hard to access given the noise present in the noisy residue $\mathbf{r}$) and need to be *refined* by the diffusion model.

The task of the diffusion model is then to guide generation of the clean speech $\mathbf{x}_0$ given the first estimate $D_\theta(\mathbf{y})$. If we look at the decomposition in (11), we simply have to remove the residual noise $\tilde{\mathbf{n}}$ and restore the distorted target cues $\mathbf{x}^{(\mathbf{dis})}$. The resulting *a priori* SNR in the starting point (again without considering the added Gaussian noise) is very high, as for a reasonable predictor $||\mathbf{x}^{(\text{dis})}||, ||\tilde{\mathbf{n}}|| \ll ||\mathbf{n}||$. The estimate is then obtained as:

$$\hat{\mathbf{x}} = G_\phi(D_\theta(\mathbf{y})) \tag{16}$$

Fig. 4: *Proposed stochastic regeneration inference process. The predictive network is first used to generate a denoised version $D_\theta(\mathbf{y})$. Diffusion-based generation G_ϕ is then performed by adding Gaussian noise $\sigma(T)\mathbf{z}$ to obtain the start sample $\mathbf{x}_T$ and solving the reverse diffusion SDE* (10)*, yielding a sample from the estimated posterior $\mathbf{x}_0 \sim p(\mathbf{x}|D_\theta(\mathbf{y}))$.*



The inference process is shown in Fig. 4. We name the resulting **Sto**chastic **R**egeneration **M**odel *StoRM*.

For training, we use a criterion $\mathcal{J}^{(\mathrm{StoRM})}$ combining denoising score matching and a supervised regularization term —e.g. mean square error—matching the output of the initial predictor to the target speech:

$$\begin{aligned}
\mathcal{J}^{(\mathrm{DSMS})}(\phi) &= \mathbb{E}_{\tau,(\mathbf{x},\mathbf{y}),\mathbf{z}} \left\| \mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau) + \frac{\mathbf{z}}{\sigma(\tau)} \right\|_2^2, \\
\mathcal{J}^{(\mathrm{Sup})}(\theta) &= \mathbb{E}_{(\mathbf{x},\mathbf{y})} \|\mathbf{x} - D_\theta(\mathbf{y})\|_2^2, \\
\mathcal{J}^{(\mathrm{StoRM})}(\phi,\theta) &= \mathcal{J}^{(\mathrm{DSMS})}(\phi) + \alpha \mathcal{J}^{(\mathrm{Sup})}(\theta),
\end{aligned} \tag{17}$$

where α is a balance term that we empirically set to 1.

One may object that the estimate $D_\theta(\mathbf{y})$ is not a sufficient statistic for the model to reconstruct target cues. However, by their very learning principle, generative models are able to create data based on the clean examples seen during training, hence our choice of the terminology *stochastic regeneration*. Stochastic regeneration is still a generative model with respect to the definition given in introduction, as it is able to output realistic samples belonging to a posterior distribution. We visually summarize the concepts of predictive, generative and our model in Fig. 5. We describe in algorithms 1 and 2 the training and inference stages of StoRM, respectively.

Algorithm 1 StoRM Training

Input: Training set of pairs $(\mathbf{x}, \mathbf{y})$
Output: Trained parameters $\{\phi, \theta\}$

Sample diffusion time $t \sim \mathcal{U}(t_\epsilon, T)$
Sample noise signal $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
Infer initial prediction $D_\theta(\mathbf{y})$
Generate perturbed state $\mathbf{x}_\tau \leftarrow \boldsymbol{\mu}(\mathbf{x}, D_\theta(\mathbf{y}), \tau) + \sigma(\tau)\mathbf{z}$
Estimate score $\mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)$
Compute loss $\mathcal{J}^{(\mathrm{StoRM})}(\phi, \theta)$ (eq. (17))
Backpropagate loss $\mathcal{J}^{(\mathrm{StoRM})}(\phi, \theta)$ to update $\{\phi, \theta\}$

Algorithm 2 StoRM Inference (based on PC sampling by [44])

Input: Corrupted speech $\mathbf{y}$, step size $\Delta\tau = \frac{T}{N}$
Output: Clean speech estimate $\widehat{\mathbf{x}}$

Infer initial prediction $D_\theta(\mathbf{y})$
Sample noise signal $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
Generate initial reverse state $\mathbf{x}_T = D_\theta(\mathbf{y}) + \sigma(T)\mathbf{z}$
for $n \in \{N, \ldots, 1\}$ **do**
    Get diffusion time $\tau = n\Delta\tau = \frac{n}{N}T$
    if using corrector **then**
        Estimate score $\mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)$
        Sample correction noise signal $\mathbf{w}_c \sim \mathcal{N}(0, \mathbf{I})$
        Correct estimate (Annealed Langevin Dynamics):
        $\mathbf{x}_\tau \leftarrow \mathbf{x}_\tau + 2r^2\sigma(\tau)^2 \mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau) + 2r\sigma(\tau)\mathbf{w}_c$
    Estimate score $\mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)$
    Sample prediction noise signal $\mathbf{w}_p \sim \mathcal{N}(0, \mathbf{I})$
    Predict next Euler-Maruyama step:
    $\mathbf{x}_{\tau-\Delta\tau} \leftarrow \mathbf{x}_\tau - \mathbf{s}_\phi(\mathbf{x}_\tau, [\mathbf{y}, D_\theta(\mathbf{y})], \tau)\Delta\tau + \gamma(D_\theta(\mathbf{y}) - \mathbf{x}_\tau)\Delta\tau + g(\tau)\mathbf{w}_p\sqrt{\Delta\tau}$
Output estimate: $\widehat{\mathbf{x}} = \mathbf{x}_0$

IV. Experimental Setup

A. Data

a) Speech Enhancement:

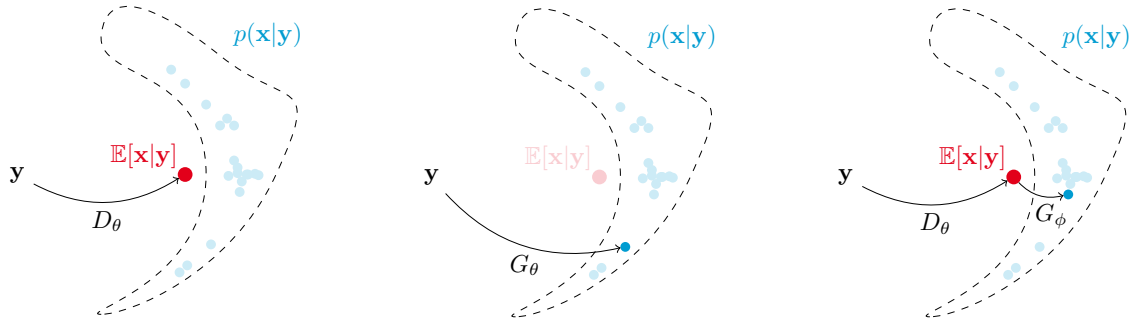
The WSJ0+Chime dataset is generated using clean speech from the WSJ0 corpus [41] and noise signals from the CHiME3 dataset [55]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance with a SNR sampled uniformly between -6 and 14 dB.

The TIMIT+Chime dataset is similarly generated as WSJ0+Chime, using TIMIT as the clean speech corpus [42]. We use this dataset for ASR as oracle annotations are available for word error rate (WER) evaluation.

The VoiceBank/DEMAND dataset is a classical benchmark dataset for speech enhancement using clean speech from the VCTK corpus [43] excluding two speakers. The utterances are corrupted by recorded noise from the DEMAND database [56] and two artificial noise types (babble and speech shaped) at SNRs of 0, 5, 10, and 15 dB for training and validation. The SNR levels of the test set are 2.5, 7.5, 12.5, and 17.5 dB.

b) Speech Dereverberation: The WSJ0+Reverb dataset is generated using clean speech data from the WSJ0 dataset and convolving each utterance with a simulated room impulse

Fig. 5: *Visualization of the inference process for the predictive, generative and proposed StoRM models for a complex posterior distribution (see also Figure 1 in [51]. With the proposed two-stage inference, StoRM uses the predictive mapping to the posterior mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ as an intermediate step for easier generative inference of a posterior sample $\mathbf{x}$ which is more likely to lie in high-density regions of the posterior $p(\mathbf{x}|\mathbf{y})$*



response (RIR). We use the `pyroomacoustics` engine [57] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a target T_{60} between 0.4 and 1.0 seconds and a room length, width and height in [5,15]×[5,15]×[2,6] m. This results in an average direct to reverberant ratio (DRR) of around -9 dB and *measured* T_{60} of 0.91 s. A dry version of the room is generated using the same geometric parameters with a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target.

B. Hyperparameters and training configuration

a) Data representation: Utterances are transformed using a short-time Fourier transform (STFT) with a window size of 510, a hop length of 128 and a square-root Hann window, at a sampling rate of 16kHz. A square-root magnitude warping is used to compress the dynamical range of the input spectrograms [12]. For training, sequences of 256 STFT frames (≈2s) are randomly extracted from the full-length utterances and normalized by the maximum absolute value of the noisy utterance before being fed to the network.

b) Forward and reverse diffusion: For all diffusion models, similar values are chosen to parameterize the forward and reverse stochastic processes. The stiffness parameter is fixed to γ =1.5, the extremal noise levels to $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 0.5$, and the extremal diffusion times to $T = 1$ and $\tau_\epsilon = 0.03$ as in [12]. Unless stated otherwise (that is, for all results except those in Figure 6), $N = 50$ time steps are used for reverse diffusion and we adopt the predictor-corrector scheme [44] with one step of annealed Langevin dynamics correction and a step size of $r = 0.5$.

c) Network architecture: The backbone architecture we use is a lighter configuration of the NCSN++ architecture variant proposed in [44], which was used in our previous study [11] and denoted as *NCSN++M*. The following modifications are carried on the up/down-sampling paths of the network: the attention layers are removed (we keep attention in the bottleneck), the number of layers in each encoder-decoder path is decreased from 7 to 4, and only one ResNet block is used per layer instead of two. This results in a network capacity of roughly 27.8M parameters instead of 65M, without significant degradation of the speech enhancement performance, be it for predictive or generative modelling.

When this NCSN++M configuration is used for score estimation in SGMSE+ [12], we call the resulting approach *SGMSE+M*. There, the noisy speech spectrogram $\mathbf{y}$ and the current diffusion process estimate $\mathbf{x}_\tau$ real and imaginary channels are stacked and fed to the network as input, and the current noise level $\sigma(\tau)$ is provided as a conditioner. For our proposed approach StoRM, the initial prediction $D_\theta(\mathbf{y})$ is also stacked together with $\mathbf{y}$ and $\mathbf{x}_\tau$: the influence of this double conditioning is examined in an ablation study in Section V-G. For the predictive approach, denoted directly as *NCSN++M*, the noise-conditioning layers are removed and only the noisy speech spectrogram real and imaginary channels are used. This ablation removes only 1.8% of the original number of parameters, which hardly modifies the network capacity.

We also use ConvTasNet [8] and GaGNet [58] (see next subsection) as alternative initial predictors for StoRM. We train using NCSN++M as the initial predictor and swap it during inference with one of the two networks mentioned above, in order to test the robustness of our proposed stochastic regeneration approach towards unseen predictors (see Section V-D.

d) Baselines: For comparison on WSJ0-based datasets, we compare StoRM to the purely generative SGMSE+M and purely predictive NCSN++M. We also report results using the non-causal version of GaGNet [58], a predictive denoiser using parallel magnitude- and complex-domain processing in the T-F domain. We complement the benchmark on the Voicebank/DEMAND dataset with the non-causal predictive ConvTasNet [8], MetricGAN+ [59], MANNER [60], VoiceFixer [61] as well as the generative unsupervised dynamical VAE (DVAE) [62], conditional time-domain diffusion model CDiffuse [29], stochastic refinement time-domain enhancement scheme SRTNet [37] and original SGMSE [28]. For all these, we use publicly available code provided by the authors.

e) Training configuration: We use the Adam optimizer [63] with a learning rate of 10^{-4} and an effective batch size of 16. We track an exponential moving average of the

DNN weights with a decay of 0.999 to be used for sampling, as it showed to be very effective [64]. We train DNNs for a maximum of 1000 epochs using early stopping based on the validation loss with a patience of 10 epochs. All models converged before reaching the maximum number of epochs. The generative approach is trained with the denoising score matching criterion (8), and the predictive methods use a simple mean-square error loss on the complex spectrogram. The stochastic regeneration approach uses the combined criterion in (17). The default training strategy is that we pre-train the initial predictor with a simple mean-square error loss, then jointly train the predictor and score networks with (17). Different training strategies are examined in an ablation study in Section V-G.

C. Evaluation metrics

For instrumental evaluation of the speech enhancement and dereverberation performance with clean test data available, we use intrusive measures such as Perceptual Evaluation of Speech Quality (PESQ) [65] to assess speech quality, extended short-term objective intelligibility (ESTOI) [66] for intelligibility and scale-invariant signal to distortion ratio (SI-SDR), scale-invariant signal to interference ratio (SI-SIR) and scale-invariant signal to artifacts ratio (SI-SAR) [67] for noise removal. As in [11], we complement our metrics benchmark with WV-MOS [68] , which is a DNN-based mean opinion score (MOS) estimation, and was used by the authors for reference-free assessment of bandwidth extension or speech enhancement performance.

We also evaluate our proposed approach on ASR, using NVidia's temporal convolutional network QuartzNet [69] as the speech recognition model, and classical WER dynamic programming evaluation with the `jiwer` Python library[2]. We use the pretrained `Base-en` 18.9M parameters version of QuartzNet for specialized English speech recognition.

Finally, we organize a medium-scale MUSHRA listening test with 9 participants. We ask the participants to rate 10 samples with a single number representing overall quality, including speech distortion, residual distortions and potential artifacts. We use the `webMUSHRA`[3] tool with `pymushra`[4] server management. The samples are randomly extracted from the WSJ0+Chime and WSJ0+Reverb test sets, ensuring gender and task balance as well as speaker exclusivity (within a given task, a speaker is used once at most). The approaches evaluated are the predictive NCSN++M, score-based generative model SGMSE+ and our proposed approach StoRM. The noisy mixture is given as a low anchor, and a supplementary anchor is created by increasing the input SNR by 10dB in comparison to the noisy mixture.

V. Experimental Results and Discussion

A. Comparison to baselines

a) WSJ0+Chime and WSJ0+Reverb: In tables I and II, we show results of the proposed stochastic regeneration StoRM approach as compared to purely predictive GaGNet and NCSN++M and purely generative SGMSE+M, for denoising on WSJ0+Chime and dereverberation on WSJ0+Reverb. Based on preliminary experiments we confirm that the approach can be successfully trained for joint dereverberation and denoising, and we refer the reader to our web page where audio samples are presented for this joint task. In this work, however, we separate the two tasks to get more insights into the individual performance.

We confirm the results from [11], which is that predictive NCSN++M and GaGNet provide samples with good interference removal (high SI-SDR) and intelligibility (high ESTOI) but lower quality (lower PESQ and WV-MOS) compared to diffusion-based generative SGMSE+M. This gap is stronger for dereverberation than for denoising as already observed, since the average input SNR for dereverberation is much lower than for denoising. Also, the reverberation interference being a filtered version of the target speech, the predictive method cannot suppress reverberation without introducing significant distortion, which is particularly audible in NCSN++M and GaGNet results. The generative SGMSE+M, however, is able to extract the speech cues and directly reconstructs with hardly any reverberation left.

It is generally observed that point-wise measures like SI-SDR, SI-SIR and SI-SAR provide generally worse results for generative models than for predictive models [6], [24]. This is because generative models try to estimate the posterior distribution, providing better perceptual metrics, whereas predictive models are implicitly trained to recover the posterior mean and average out the distortions for each point, thus yielding higher point-wise fidelity [6], [11].

We observe that our proposed StoRM associates the best of both the predictive and generative worlds, by producing samples with very high quality like generative SGMSE+M, while being approximately as good with interference removal as the predictive NCSN++M. Again, the observed gap is more significant for dereverberation, where the proposed StoRM outperforms both SGMSE+M and NCSN++M on all metrics. Example spectrograms are displayed on Figure 7 and 8, for denoising and dereverberation respectively.

b) VoiceBank/DEMAND: We report in Table III results of our StoRM configuration against various state-of-the-art speech enhancement baselines on the VoiceBank/DEMAND benchmark. The SNRs in Voicebank/DEMAND are always positive and distributed around 10dB, which is not very challenging compared to the conditions in our WSJ0+Chime dataset. Consequently, the gap between SGMSE+ and StoRM on Voicebank/DEMAND is not as large as on WSJ0+Chime, which shows that using the initial predictor is particularly useful in difficult conditions. In easier environments such as that simulated in Voicebank/DEMAND, diffusion-based generative modelling can take the noisy mixture as the initial condition for reverse diffusion without being further guided. Still, our proposed method StoRM still slightly outperforms the other generative models on ESTOI, WV-MOS and SI-SDR, setting a new state-of-the-art record for generative models on this benchmark.

[2]https://github.com/jitsi/jiwer
[3]https://github.com/audiolabs/webMUSHRA
[4]https://github.com/nils-werner/pymushra

TABLE I: *Denoising results obtained on WSJ0+Chime. Values indicate mean and standard deviation. All approaches (except GaGNet) use the NCSN++M architecture. Diffusion models (SGMSE+M and StoRM) use $N = 50$ steps for reverse diffusion.*

Method	WV-MOS	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR
Mixture	1.43 ± 0.66	1.38 ± 0.32	0.65 ± 0.18	4.3 ± 5.8	4.3 ± 5.8	-
SGMSE+M	3.63 ± 0.38	2.33 ± 0.61	0.86 ± 0.10	13.3 ± 5.0	27.4 ± 6.3	13.5 ± 4.9
NCSN++M	3.47 ± 0.53	2.21 ± 0.65	**0.89 ± 0.09**	**16.4 ± 4.4**	31.1 ± 5.0	**16.6 ± 4.4**
GaGNet	3.34 ± 0.54	2.19 ± 0.61	0.87 ± 0.09	15.7 ± 4.3	27.6 ± 4.7	16.0 ± 4.4
StoRM	**3.72 ± 0.40**	**2.58 ± 0.61**	0.88 ± 0.08	15.1 ± 4.2	**31.6 ± 5.0**	15.3 ± 4.2

TABLE II: *Dereverberation results on Reverb-WSJ0. Values indicate mean and standard deviation. All approaches (except GaGNet) use the NCSN++M architecture. Diffusion models (SGMSE+M and StoRM) use $N = 50$ steps for reverse diffusion.*

Method	WV-MOS	PESQ	ESTOI	SI-SDR	SI-SIR	SI-SAR
Mixture	1.78 ± 0.99	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5	-7.5 ± 5.4	-
SGMSE+M	3.49 ± 0.39	2.66 ± 0.45	0.85 ± 0.06	2.4 ± 7.2	11.6 ± 9.9	2.8 ± 6.8
NCSN++M	2.99 ± 0.38	2.08 ± 0.47	0.85 ± 0.06	6.1 ± 3.8	21.4 ± 7.0	6.1 ± 3.7
GaGNet	2.40 ± 0.52	1.59 ± 0.37	0.68 ± 0.09	-0.5 ± 4.8	7.7 ± 4.0	0.2 ± 5.1
StoRM	**3.73 ± 0.32**	**2.83 ± 0.42**	**0.88 ± 0.04**	**6.5 ± 4.0**	**22.9 ± 8.2**	**6.5 ± 3.9**

TABLE III: *Denoising results obtained on VoiceBank/DEMAND.* P *means predictive and* G *generative. All approaches were evaluated on the test set using publicly available code attached to the respective papers.*

Method	Type	PESQ	ESTOI	SI-SDR	WV-MOS
Mixture		1.97	0.79	8.4	2.99
NCSN++	P	2.83	**0.88**	**20.1**	4.07
NCSN++M	P	2.82	**0.88**	19.9	4.06
Conv-TasNet [8]	P	2.84	0.85	19.1	4.28
MetricGAN+ [59]	P	3.13	0.83	8.5	3.90
MANNER [60]	P	**3.21**	0.87	18.9	**4.38**
GaGNet [58]	P	2.94	0.86	18.1	4.23
VoiceFixer [61]	P	2.11	0.73	-4.3	4.00
DVAE [62]	G	2.43	0.81	16.4	3.73
CDiffuSE [29]	G	2.46	0.79	12.6	3.64
SRTNet [37]	G	2.11	0.81	8.5	3.58
SGMSE [28]	G	2.28	0.80	16.2	3.90
SGMSE+ [12]	G	2.93	0.87	17.3	4.24
SGMSE+M	G	**2.96**	0.87	17.3	4.26
StoRM (proposed)	G	2.93	**0.88**	**18.8**	**4.30**

TABLE IV: *Denoising results on WSJ0+Chime* (with input SNRs between 0dB and 20dB) in matched and mismatched settings. In the mismatch setting, the approaches are trained on VoiceBank/DEMAND. All methods use the NCSN++M architecture. SGMSE+M and StoRM use* 50 *diffusion steps.*

Method	Match	WV-MOS	PESQ	ESTOI	SI-SDR
NCSN++M	✓	3.78	2.73	**0.94**	**20.0**
	✗	3.33 (-0.35)	2.14 (-0.59)	**0.90** (-0.04)	**17.6** (-2.4)
SGMSE+M	✓	3.84	2.96	0.92	17.4
	✗	3.61 (-0.23)	**2.48** (**-0.48**)	**0.90** (**-0.02**)	15.7 (**-1.7**)
StoRM	✓	**3.93**	**3.01**	0.93	18.5
	✗	**3.71** (**-0.22**)	2.47 (-0.54)	**0.90** (-0.03)	16.8 (**-1.7**)

B. Efficient sampling

We report in Figure 6 the performance of the SGMSE+M and StoRM schemes as a function of the number of steps used for reverse diffusion. We additionally provide an estimation of the number of multiply-accumulate (MAC) operations per second as measured by the `python-papi` package.

We observe that StoRM is able to maintain performance at a near-optimal level even using only 10 steps, using the initial predictive estimate as a reasonable guess for further diffusion. In comparison, SGMSE+M performance degrades rapidly as the number of steps decreases. Furthermore, StoRM is able to produce very high-quality samples without even needing the Annealed Langevin Dynamics corrector during sampling, whereas SGMSE+M performance dramatically degrades without this corrector. Since each corrector step makes an additional call to the score network, avoiding its use further relaxes the computational complexity. StoRM therefore highlights a strong compromise between inference speed and sample quality. Using StoRM with 20 steps and no corrector produces near-optimal sample quality at a cost of 4.5 $\cdot 10^{11}$ MAC$\cdot$s^{-1}, versus 2.1 $\cdot 10^{12}$ MAC$\cdot$s^{-1} for the optimal SGMSE+M setting (50 steps and Annealed Langevin Dynamics correction). StoRM even outperforms the optimal SGMSE+M setting using 10 steps and no corrector, thus reducing computational complexity by a full order of magnitude. In our recent work [70], we proposed to make our diffusion generative model causal to demonstrate its application to real-world scenarios.

C. Generalization to unseen data

In Table IV, we examine robustness to mismatched training and test data. The mismatched condition is generated by training on Voicebank/DEMAND and testing on WSJ0+Chime*. WSJ0+Chime* is created in the same fashion as WSJ0+Chime,

Fig. 6: *Results for denoising on WSJ0+Chime as a function of the number of reverse diffusion steps N. All approaches use the same NCSN++M architecture. The corrector uses one step of Annealed Langevin Dynamics with $r = 0.5$.*

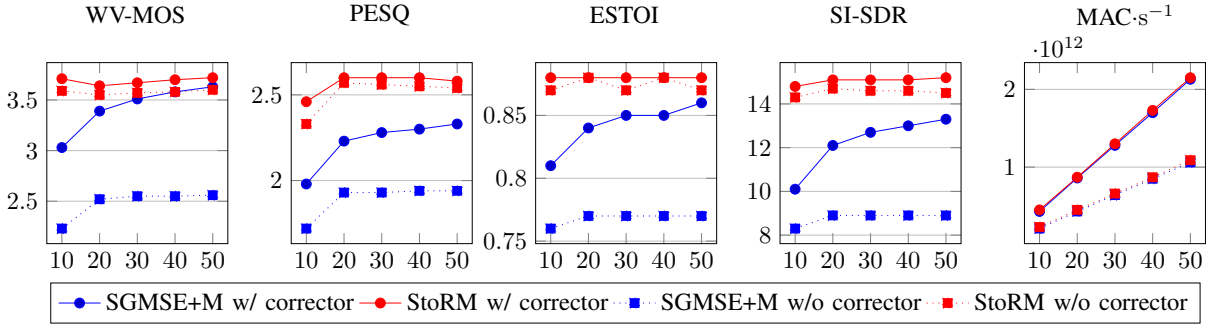


Fig. 7: *Clean, noisy and processed utterances from WS0+Chime. Input SNR is -0.9* dB. *Vocalizing artifacts are visible at the beginning of the SGMSE+M utterance. Speech distortions are observed in the NCSN++M sample. StoRM corrects these distortions without introducing vocalizing artifacts and yield high quality and intelligibility.*

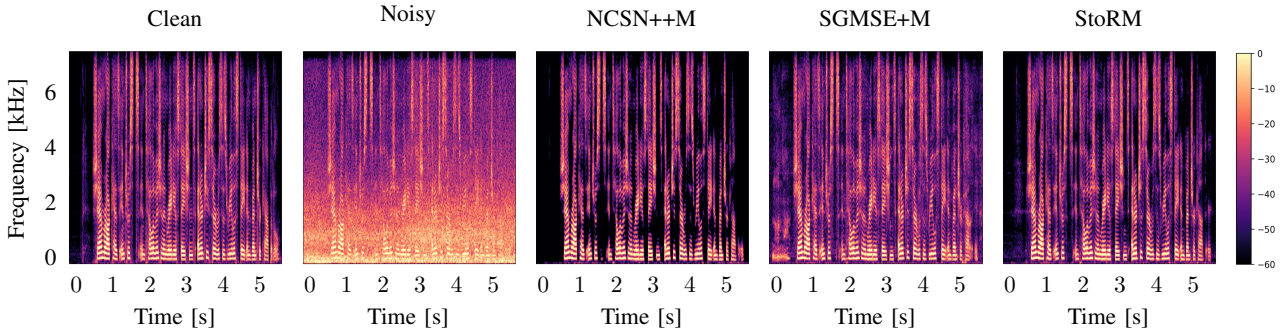
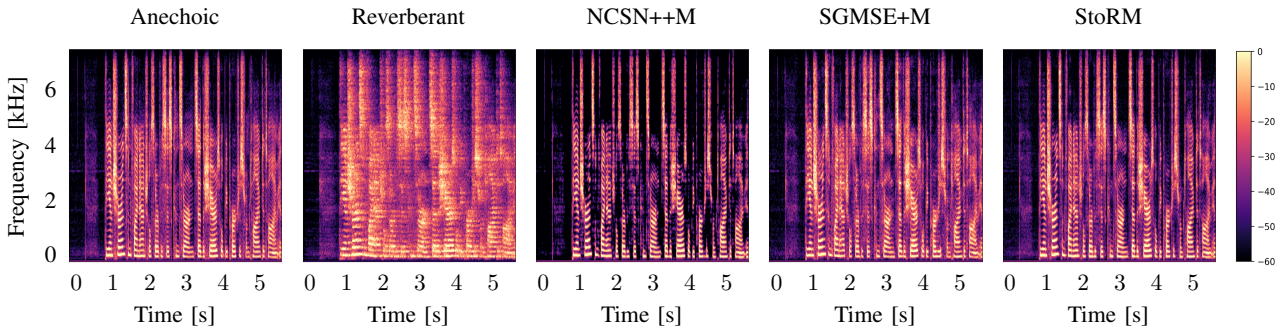


Fig. 8: *Anechoic, reverberant and processed utterances from WS0+Reverb. Input T_{60} is 1.06* s. *Formant structure is partly destroyed by SGMSE+M and severe speech distortions are observed in the NCSN++M sample. StoRM corrects the distortions and reproduces the formant structure without residual reverberation.*

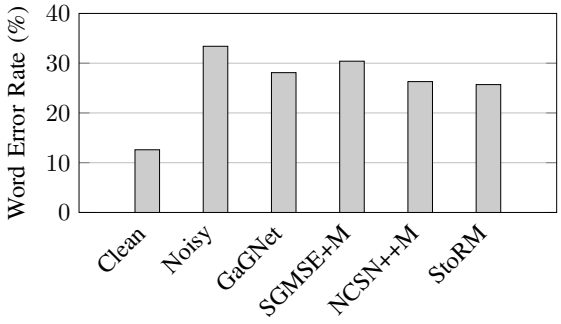


but the input SNRs range between 0 and 20dB to match the SNR range in Voicebank+DEMAND, and thus constitutes the same benchmark as in [12]. We observe that NCSN++M shows a reasonable generalization ability because of its sophisticated network architecture. However, SGMSE+M and StoRM's ability to maintain performance as compared to the matched case is even superior. This shows that StoRM can leverage generative modelling to correct the relative lack of robustness of the first predictive stage to this mismatched condition.

TABLE V: *Denoising results on WSJ0+Chime for StoRM using matched and mismatched initial predictors. The predictor architecture used for training is NCSN++M. All approaches use the NCSN++M as score network and $N = 50$ steps . Values indicate mean and standard deviation.*

Initial Predictor	Matched	PESQ	ESTOI	SI-SDR
Mixture	-	1.38 ± 0.32	0.65 ± 0.18	4.3 ± 5.8
NCSN++M	✓	**2.53** ± **0.63**	**0.88** ± **0.09**	**14.7** ± **4.3**
GaGNet [58]	✗	2.52 ± 0.62	0.87 ± 0.09	**14.7** ± **4.1**
ConvTasNet [8]	✗	2.36 ± 0.60	0.86 ± 0.09	9.9 ± 1.7

Fig. 9: *ASR results for speech enhancement on TIMIT+Chime. Diffusion models use $N = 50$ steps.*



D. Generalization to mismatched predictors

In Table V, we report results for StoRM using different initial predictors than the one used during training. The approach is trained using the NCSN++M as initial predictor as before, and we test using ConvTasNet [8] and GaGNet [58] as alternative initial predictors by exchanging this predictor during inference. We observe that the artifacts in GaGNet estimates are of a similar nature than those of NCSN++M, as both approaches process speech in the T-F domain. StoRM is entirely robust to such a slight mismatch, as indicated by the equivalent performance of using NCSN++M and GaGNet as the initial predictor. ConvTasNet is a time-domain method using a fully learnt encoder: the speech distortions are then different than those of NCSN++M or GaGNet. Additionally, ConvTasNet's original performance is slightly worse than its two counterparts. Consequently, we observe that the performance of StoRM using ConvTasNet as the initial predictor is poorer but close to that of using NCSN++M as the predictor. This demonstrates relative robustness to unseen conditions provided by the generative modelling stage.

E. ASR results

In Figure 9, we compare the predictive, generative and stochastic regeneration approaches on speech enhancement for ASR using the TIMIT+Chime dataset. We observe that SGMSE+M results in poorer speech recognition abilities than its predictive GaGNet and NCSN++M counterparts, and hardly improves the ASR performance over the noisy mixture. This can be explained by the previously mentioned undesired vocalizing artifacts and phonetic confusions, which are created by the generative approach under uncertainty over the presence and phonetic nature of speech respectively and are heavily punished by WER evaluation. Using the predictive estimate as a guide for generation, StoRM improves the WER performance by a relative factor of 19% as compared to SGMSE+M, even slightly outperforming NCSN++M, which shows that most of the artifacts and confusions are corrected.

F. Listening experiment

We show in the boxplot on Figure 10 the results of our MUSHRA listening test. On average, the participants clearly rated the proposed StoRM higher than the purely predictive NCSN++M and purely generative SGMSE+M. This confirms the results provided by the intrusive and non-intrusive metrics provided in tables I and II. Participants rated NCSN++M slightly better than SGMSE+M on average, which is linked to the rating criterion described in Section IV-C. This seems to indicate that participants put more weight on "residual distortions" and "potential artifacts" (vocalizing/breathing/confusions) than on "speech distortion".

Fig. 10: *Listening test results. CQS is the "continuous quality scale" on which participants are asked to rate. Inner line represents the median. 9 participants rated 10 samples randomly selected from WSJ0+Chime and WSJ0+Reverb.*

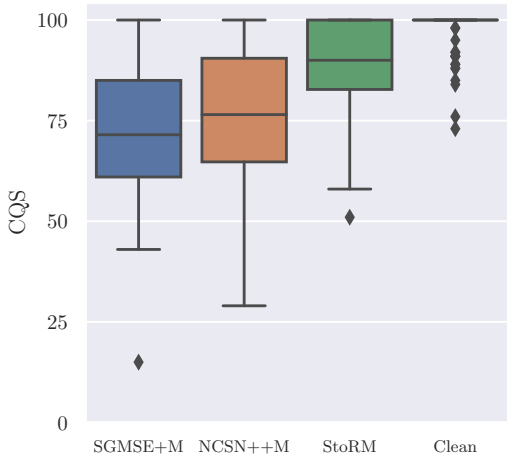


TABLE VI: *Denoising results on WSJ0+Chime for StoRM using different conditioning inputs for the score network. Values indicate mean and standard devation. All approaches use NCSN++M as backbone architecture and $N = 50$ steps.*

Conditioning	PESQ	ESTOI	SI-SDR
Noisy	2.30 ± 0.60	0.84 ± 0.10	11.5 ± 5.2
PostDenoiser	2.50 ± 0.62	0.87 ± 0.09	14.7 ± 4.3
Both	**2.53 ± 0.63**	**0.88 ± 0.08**	**15.1 ± 4.2**

G. Ablation studies

We conduct ablation studies on the WSJ0+Chime dataset, to observe the respective influence of score network conditioning on the one hand and the training strategy on the other hand.

a) Conditioning of the score network: In Table VI, we report instrumental results when using different conditioning inputs for the score network used in the proposed StoRM. We input either the noisy speech $\mathbf{y}$ ("Noisy"), the denoised estimate $D_\theta(\mathbf{y})$ ("PostDenoiser"), or both ("Both", which is the default setting for StoRM). Using only the noisy speech ("Noisy") is detrimental to the performance. It seems that the score network does need the information from the original distortions in $D_\theta(\mathbf{y})$ at time step τ=T, to properly learn the score at time step τ<T. This mismatch at the first denoising steps is detrimental to performance. We also observe that instrumental

TABLE VII: *Denoising results on WSJ0+Chime for StoRM using different training strategies for the score network. All approaches use NCSN++M as backbone architecture and $N = 50$ steps for reverse diffusion.*

Pre-train D_θ	Fine-tune D_θ	Use $\mathcal{J}^{(\mathrm{Sup})}$	PESQ	ESTOI	SI-SDR
✗	✓	✓	**2.58**	**0.88**	**15.1**
✓	✗	✗	2.53	**0.88**	14.7
✗	✓	✗	1.11	0.62	-0.3
✓	✓	✓	**2.58**	**0.88**	**15.1**

metrics tend to slightly favor the "Both" conditioning over the "PostDenoiser" conditioning.

b) Training strategies: We show in Table VII the results of StoRM using different training strategies. We see that jointly training the initial predictor and the score network slightly improves results for denoising. However, training the initial predictor from scratch or having it pre-trained first does not seem to make a difference, as long as one regularizes the training criterion with the supervised criterion $\mathcal{J}^{(\mathrm{Sup})}$ which matches the output of the initial predictor to the target. Indeed, as shown in the third line of Table VII, if we use a randomly initialized predictor and train both the predictor and score networks only with the score matching criterion $\mathcal{J}^{(\mathrm{DSM})}$—i.e. setting α in (17) to 0—the performance dramatically drops. This is to be expected since the learning task then becomes much more complicated given the size of the search space and the lack of regularization. The proposed combination of joint training and regularization with $\mathcal{J}^{(\mathrm{Sup})}$ performs most favorably. This indicates that it is best to train the predictor to output something resembling clean speech rather than arbitrary learned encoder features, while still leaving some room for the predictor to adapt its output to the score model. Our experiments also indicate that, once pre-trained, StoRM converges 25% faster than when training from scratch. However, the cumulated time of pre-training the predictor and fine-tuning is 50% larger than the cost of training from scratch.

VI. Conclusion

We presented a generative stochastic regeneration scheme combining a predictive model as initial predictor and a diffusion-based generative approach regenerating the target cues distorted by the first stage. On the one hand, the approach improves sample quality compared to pure predictive approaches as it leverages generative modelling to output samples that have high probability on the target posterior distribution manifold, rather than regressing to their mean. On the other hand, it uses predictive power to provide a good initial prediction of the target sample, which avoids typical generative artifacts such as vocalizing and breathing effects, and increases the interference removal performance, especially in difficult environments. Intrusive and reference-free instrumental metrics as well as formal listening tests confirmed the superiority of the stochastic regeneration approach over the baselines. The resulting approach allows efficient sampling, requiring fewer steps and avoiding the use of Annealed Langevin Dynamics correction during reverse diffusion, thus reducing computational complexity by an order of magnitude without sacrificing quality, compared to the original diffusion model.

References

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Noise Control Engineering Journal, 2011, vol. 59.

[2] S. J. Godsill, P. J. W. Rayner, and O. Cappé, *Digital audio restoration*. Springer, 1998.

[3] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state-of-the-art*. Morgan & Claypool, 2013.

[4] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio Source Separation and Speech Enhancement*, E. Vincent, Ed. John Wiley & Sons, 2018.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech, Language Proc.*, 2018.

[6] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[7] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[8] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *ISCA Interspeech*, 2018.

[9] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A two-stage approach to speech bandwidth extension," in *ISCA Interspeech*, 2021.

[10] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 23, no. 6, pp. 982–992, 2015.

[11] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2023.

[12] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE Trans. Audio, Speech, Language Proc.*, 2023.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Int. Conf. Learning Repr. (ICLR)*, 2014.

[14] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2021.

[15] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," *IEEE Int. Workshop on Machine Learning for Signal Proc. (MLSP)*, pp. 1–6, 2018.

[16] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks." *ISCA Interspeech*, 2020.

[17] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Int. Conf. Machine Learning (ICML)*, 2015.

[18] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, nov 2021.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Neural Information Processing Systems (NIPS)*, vol. 27, 2014.

[20] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Neural Information Processing Systems (NIPS)*, 2020.

[21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *Int. Conf. Machine Learning (ICML)*, 2015.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Neural Information Processing Systems (NIPS)*, 2020.

[23] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Neural Information Processing Systems (NIPS)*, vol. 32, 2019.

[24] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] S. Welker, H. N. Chapman, and T. Gerkmann, "DriftRec: Adapting diffusion models to blind image restoration tasks," *arXiv preprint*, 2022.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Neural Information Processing Systems (NIPS)*, vol. 34, 2021.

[27] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint*, 2022.

[28] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *ISCA Interspeech*, 2022.

[29] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2022.

[30] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint*, 2022.

[31] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *ISCA Interspeech*, 2022.

[32] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Neural Information Processing Systems (NIPS)*, 2021.

[33] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Int. Conf. Learning Repr. (ICLR)*, 2021.

[34] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *Int. Conf. Learning Repr. (ICLR)*, 2022.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] B. Jing, G. Corso, R. Berlinghieri, and T. Jaakkola, "Subspace diffusion generative models," in *European Conf. on Computer Vision (ECVA)*, 2022.

[37] Z. Qiu, M. Fu, Y. Yu, L. Yin, F. Sun, and H. Huang, "SRTNet: Time domain speech enhancement via stochastic refinement," *arXiv preprint*, 2022.

[38] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Neural Information Processing Systems (NIPS)*, 2022.

[39] K. Saito, N. Murata, T. Uesaka, C.-H. Lai, Y. Takida, T. Fukui, and Y. Mitsufuji, "Unsupervised vocal dereverberation with diffusion-based generative models," *arXiv preprint*, 2022.

[40] R. Sawata, N. Murata, Y. Takida, T. Uesaka, T. Shibuya, S. Takahashi, and Y. Mitsufuji, "A versatile diffusion-based generative refiner for speech enhancement," *arXiv preprint*, 2022.

[41] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete." [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A

[42] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[43] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," *9th ISCA Speech Synthesis Workshop*, 2016.

[44] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *Int. Conf. Learning Repr. (ICLR)*, 2021.

[45] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, 2005.

[46] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[47] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*. Journal of the American Statistical Association, 2000, vol. 82.

[48] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, 1982.

[49] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," *Int. Conf. Learning Repr. (ICLR)*, 2021.

[50] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

[51] M. Delbracio and P. Milanfar, "Inversion by direct iteration: An alternative to denoising diffusion for image restoration," *arXiv preprint*, 2023.

[52] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Neural Information Processing Systems (NIPS)*, 2017.

[53] A. Avila, A. Alam, D. O'Shaughnessy, and T. Falk, "Investigating speech enhancement and perceptual quality for speech emotion recognition," in *ISCA Interspeech*, 2018.

[54] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[55] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[56] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[57] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018.

[58] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, 2022.

[59] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of metricgan for speech enhancement," in *ISCA Interspeech*, 2022, pp. 7412–7416.

[60] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "Manner: Multi-view attention network for noise erasure," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2022.

[61] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv preprint*, 2021.

[62] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2993–3007, 2022.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learning Repr. (ICLR)*, 2015.

[64] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Neural Information Processing Systems (NIPS)*, 2020.

[65] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2001.

[66] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, 2016.

[67] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr - half-baked or well done?" in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2019.

[68] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," *arXiv preprint*, 2022.

[69] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building AI applications using neural modules," *arXiv preprint*, 2019.

[70] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Speech signal improvement using causal generative diffusion models," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, 2023.

A.6.3 Publication: Analysing Diffusion-based Generative Approaches versus Discriminative Approaches for Speech Restoration

Abstract

Diffusion-based generative models have had a high impact on the computer vision and speech processing communities these past years. Besides data generation tasks, they have also been employed for data restoration tasks like speech enhancement and dereverberation. While discriminative models have traditionally been argued to be more powerful e.g. for speech enhancement, generative diffusion approaches have recently been shown to narrow this performance gap considerably. In this paper, we systematically compare the performance of generative diffusion models and discriminative approaches on different speech restoration tasks. For this, we extend our prior contributions on diffusion-based speech enhancement in the complex time-frequency domain to the task of bandwith extension. We then compare it to a discriminatively trained neural network with the same network architecture on three restoration tasks, namely speech denoising, dereverberation and bandwidth extension. We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Code and audio examples can be found online.

Reference

Jean-Marie Lemercier, Julius Richter, Simon Welker, Timo Gerkmann, "Analysing Diffusion-based Generative Approaches versus Discriminative Approaches for Speech Restoration," *IEEE International Conference Acoustics, Speech, and Signal Processing*, Rhodes, Greece, Jun. 2023. DOI: 10.1109/ICASSP39728.2021.9414363

Copyright Notice

The following article is the accepted version of the article published with IEEE.
© 2023 IEEE. Reprinted, with permission, from the reference displayed above.

Author's Contributions

Jean-Marie Lemercier is the lead author of this publication, having implemented all the algorithms, trained the neural networks utilized in the study, conducted the experimental validation, and authored the manuscript. Julius Richter and Simon Welker contributed by providing feedback on all methodologies through discussions and assisted in reviewing the manuscript. Timo Gerkmann offered insights on experimental validation and mathematical derivations, in addition to reviewing the manuscript.

ANALYSING DIFFUSION-BASED GENERATIVE APPROACHES VERSUS DISCRIMINATIVE APPROACHES FOR SPEECH RESTORATION

Jean-Marie Lemercier[⋆]*, Julius Richter*[⋆]*, Simon Welker*[⋆×]*, Timo Gerkmann*[⋆]

[⋆]Signal Processing (SP), Universität Hamburg, Germany
[×] Center for Free-Electron Laser Science, DESY, Germany
{firstname.lastname}@uni-hamburg.de

ABSTRACT

Diffusion-based generative models have had a high impact on the computer vision and speech processing communities these past years. Besides data generation tasks, they have also been employed for data restoration tasks like speech enhancement and dereverberation. While discriminative models have traditionally been argued to be more powerful e.g. for speech enhancement, generative diffusion approaches have recently been shown to narrow this performance gap considerably. In this paper, we systematically compare the performance of generative diffusion models and discriminative approaches on different speech restoration tasks. For this, we extend our prior contributions on diffusion-based speech enhancement in the complex time-frequency domain to the task of bandwith extension. We then compare it to a discriminatively trained neural network with the same network architecture on three restoration tasks, namely speech denoising, dereverberation and bandwidth extension. We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Code and audio examples can be found online[1].

Index Terms— generative modelling, diffusion models, speech enhancement, dereverberation, bandwidth extension

1. INTRODUCTION

Speech corruptions arise in real-life scenarios and modern communication devices, when clean speech sources are impacted by background noise, interfering speakers, room acoustics and channel degradation. Speech restoration therefore aims at recovering clean speech from the corrupted signal. Traditional speech restoration methods leverage the different statistical properties of the target and interference signals [1]. Data-driven approaches based on machine learning predominately employ discriminative models that learn a single best deterministic mapping between corrupted speech and the corresponding clean speech target [2].

In contrast, generative models implicitly or explicitly learn the target distribution and allow to generate multiple valid estimates instead of a single best estimate as in discriminative approaches [3]. For example, diffusion-based generative models, or simply *diffusion models*, have shown great success in learning the data distribution of natural images [4, 5, 6]. This class of models uses a *forward process* to slowly turn data into a tractable prior, such as a standard normal distribution, and train a neural network to solve the *reverse process* to generate clean data from this prior. These diffusion models can also be used for conditional generation in restoration tasks, which has recently been proposed for speech enhancement and dereverberation [7, 8, 9, 10]. They can in that regard be functionally seen as a mean of generating clean speech based on noisy speech, and can be thus compared to discriminative approaches. However, to make a fair comparison of these two conceptually different approaches, similar network architectures and same training data should be used.

In this work, we present an analysis of a generative diffusion model as compared to its discriminative counterpart sharing the same deep neural network (DNN) architecture, for various speech restoration tasks. We use our previous method which defines the diffusion process in the complex spectrogram domain [7, 8]. We show that the performance gap between the generative and discriminative models varies with respect to the corruption at hand. We evaluate our proposed approaches on the WSJ0 corpus, using various simulated corruptions and recorded background noise. Finally we compare our bandwidth extension model with state-of-the-art bandwidth extension methods on the VCTK corpus.

The remainder of the paper is organized as follows. We first present the three speech restoration problems benchmarked, along with popular solutions for solving them. Then, we introduce diffusion-based generative models using the stochastic differential equation (SDE) formalism. We continue by explaining our experimental setup including data generation and training methods. Finally, we present and discuss our results.

2. SPEECH RESTORATION TASKS AND RELATED WORK

2.1. Speech enhancement

Speech enhancement consists in removing an additive interference n (e.g. background noise or interfering speakers) from the corrupted mixture y to extract the clean speech target s:

$$y = s + n \tag{1}$$

Popular enhancement methods include Wiener-inspired spectral filtering [1], discriminative machine learning methods [2] or generative approaches like denoising variational auto-encoders (VAEs) [11]. Recently, diffusion models were proposed to tackle speech enhancement either in the time domain [12] or in the complex time-frequency (T-F) domain [7, 10, 8].

2.2. Speech dereverberation

Reverberation is caused by room acoustics, and is characterized by multiple reflections on the room enclosures. Late reflections particularly degrade the speech signal and may result in reduced intelligibility [13].

This work has been funded by the Federal Ministry for Economic Affairs and Climate Action, project 01MK20012S, AP380, DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the Grant-No. HIDSS-0002, and the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169).

[1]https://uhh.de/inf-sp-sgmsemultitask

The corruption model is then convolutive, as the clean speech s is convolved with a room impulse response (RIR) h representing the acoustic path between the source and the listener:

$$y = s * h \tag{2}$$

Single-channel dereverberation methods range from spectral enhancement [14], inverse filtering [15], and cepstral processing [16] to machine learning algorithms using DNNs in the complex T-F domain [17] or in the time-domain [18].

2.3. Bandwidth extension

Audio super-resolution, or bandwidth extension, aims at converting a low-sampling rate signal back to a version sampled at a higher rate, regenerating time resolution, high-frequency content and audio quality. The corruption process is linear and involves an anti-aliasing low-pass filter followed by a decimation operation:

$$y = \mathrm{Resample}(s * a,\ f_s^{\mathrm{up}},\ f_s^{\mathrm{low}}) \tag{3}$$

where a is the anti-aliasing filter impulse response, f_s^{up} the original high sampling rate and f_s^{low} the low sampling rate.

Several discriminative methods were proposed to tackle bandwidth extension for speech signals [19, 20]. Generative approaches based on neural vocoders using generative adversarial networks (GANs) were also proposed [21, 22, 23]. A continuous-time diffusion model in the time-domain was proposed in [24].

3. SCORE-BASED DIFFUSION MODELS FOR SPEECH RESTORATION

Score-based diffusion models are defined by three components: a forward diffusion process, a score estimator and a sampling method for inference.

3.1. Forward and reverse processes

The stochastic forward process $\{\mathbf{x}_t\}_{t=0}^{T}$ is modeled as the solution to a SDE, in the Itô sense [25, 26]:

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w} \tag{4}$$

where $\mathbf{x}_t$ is the current state of the process indexed by a continuous time variable $t \in [0,T]$ with the initial condition $\mathbf{x}_0$ representing clean speech. As our process is defined in the T-F domain, the variables in bold are assumed to be one-dimensional vectors in $\mathbb{C}^d$ containing the coefficients of a flattened complex spectrogram, whereas variables in regular font represent real scalar values. The stochastic process $\mathbf{w}$ is a standard d-dimensional Brownian motion, which implies that $\mathrm{d}\mathbf{w}$ is a zero-mean Gaussian random variable with standard deviation $\sqrt{\mathrm{d}t}$ for each T-F bin.

The *drift* function $\mathbf{f}$ and *diffusion* coefficient g as well as the initial condition $\mathbf{x_0}$ and the final diffusion time T define uniquely the Itô process $\{\mathbf{x}_t\}_{t=0}^{T}$. Under some regularity conditions on $\mathbf{f}$,g allowing a unique and smooth solution to the Kolmogorov equations associated to (4), the reverse process $\{\mathbf{x}_t\}_{t=T}^{0}$ is another diffusion process defined as the solution of a SDE, with the following form [27, 26]:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\,\mathrm{d}\bar{\mathbf{w}}, \tag{5}$$

where $\mathrm{d}\bar{\mathbf{w}}$ is a d-dimensional Brownian motion for the time flowing in reverse and $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the *score function*, i.e. the gradient of the logarithm data distribution for the current process state $\mathbf{x}_t$.

Speech restoration tasks can be considered as conditional generation tasks, i.e. generation of clean speech $\mathbf{x_0}$ conditioned by the corrupted speech $\mathbf{y}$. In [7, 8] we proposed to incorporate the conditioning directly into the diffusion process by defining the forward process as the solution to the following Ornstein-Uhlenbeck SDE [25]:

$$\mathrm{d}\mathbf{x}_t = \underbrace{\gamma\,(\mathbf{y} - \mathbf{x}_t)}_{:=\,\mathbf{f}(\mathbf{x}_t,\mathbf{y})}\mathrm{d}t + \underbrace{\left[\sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{t}\sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}\right]}_{:=\,g(t)}\mathrm{d}\mathbf{w}, \tag{6}$$

with γ a *stiffness* hyperparameter, and $\sigma_{\min}$ and $\sigma_{\max}$ two hyperparameters controlling the *noise scheduling*, that is, the amount of Gaussian white noise injected at each timestep of the process.

The interpretation of our forward process in Eq. (6), visualized on Fig. 1, is as follows: at each time step and for each T-F bin independently, an infinitesimal amount of corruption is added to the current process state $\mathbf{x}_t$, along with Gaussian noise with standard deviation $g(t)\sqrt{\mathrm{d}t}$. Given an initial state $\mathbf{x}_0$ and $\mathbf{y}$, the Itô forward process corresponding to the solution of (6) admits a Gaussian distribution for the process state $\mathbf{x}_t$ called *perturbation kernel*:

$$p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y}) = \mathcal{N}_{\mathbb{C}}\big(\mathbf{x}_t; \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t), \sigma(t)^2\mathbf{I}\big), \tag{7}$$

where $\mathcal{N}_{\mathbb{C}}$ denotes the circularly-symmetric complex normal distribution and $\mathbf{I}$ the identity matrix. Given the simple Gaussian kernel, closed-form solutions for the mean $\boldsymbol{\mu}$ and variance $\sigma(t)^2$ can be determined [25]:

$$\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) = \mathrm{e}^{-\gamma t}\mathbf{x}_0 + (1 - \mathrm{e}^{-\gamma t})\mathbf{y}, \tag{8}$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2\big((\sigma_{\max}/\sigma_{\min})^{2t} - \mathrm{e}^{-2\gamma t}\big)\log(\sigma_{\max}/\sigma_{\min})}{\gamma + \log(\sigma_{\max}/\sigma_{\min})}. \tag{9}$$

3.2. Score function estimator

When performing inference by sampling through the reverse SDE in Eq. (5), the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is not readily available. Thus, it is approximated by a DNN $\mathbf{s}_\theta$, called the *score model*. In particular, given the Gaussian form of the perturbation kernel $p_{0t}(\mathbf{x}_t|\mathbf{x}_0,\mathbf{y})$ and the regularity conditions exhibited by the mean and variance, a *denoising score matching* objective can be used to train the score model $\mathbf{s}_\theta$ [28].

The score function of the perturbation kernel is:

$$\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y}) = -\frac{\mathbf{x}_t - \boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t)}{\sigma(t)^2}. \tag{10}$$

Therefore we can reparameterize the denoising score matching objective as follows [26]:

$$\begin{aligned}\mathcal{J}(\theta) &= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{y},\{\mathbf{x}_t|\mathbf{x}_0,\mathbf{y}\}}\left[\left\|\mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t) - \nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\,\mathbf{x}_0, \mathbf{y})\right\|_2^2\right] \\ &= \mathbb{E}_{t,\mathbf{x}_0,\mathbf{y},\mathbf{z}}\left[\left\|\mathbf{s}_\theta\big(\big[\boldsymbol{\mu}(\mathbf{x}_0, \mathbf{y}, t) + \sigma(t)\mathbf{z}\big], \mathbf{y}, t\big) + \frac{\mathbf{z}}{\sigma(t)}\right\|_2^2\right],\end{aligned} \tag{11}$$

using $\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_0,\mathbf{y},t) + \sigma(t)\mathbf{z}$, with $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. t is sampled uniformly in $[t_\epsilon, T]$ where t_ϵ is a minimal diffusion time used to avoid numerical instabilities.

3.3. Inference through reverse sampling

At inference time, we first sample an initial condition of the reverse process, corresponding to $\mathbf{x}_T$, with:

$$\mathbf{x}_T \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_T; \mathbf{y}, \sigma^2(T)\mathbf{I}), \tag{12}$$

This sample corresponds to corrupted speech $\mathbf{y}$ to which we add Gaussian noise with variance $\sigma(t)^2$, which approximates the training condition.

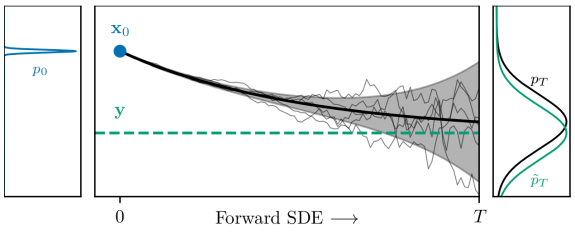


Fig. 1. *Visualization of the forward process* (6). *Mean curve is in solid black and variance is represented by the greyed area. Several realizations of the diffusion process are represented by thin black lines.*

Conditional generation is then performed by solving the following *plug-in reverse SDE* from $t = T$ to $t = 0$, where the score function is replaced by its estimator $\mathbf{s}_\theta$, assuming the latter was trained e.g. according to Section 3.2:

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(\mathbf{x}_t, \mathbf{y}) + g(t)^2\, \mathbf{s}_\theta(\mathbf{x}_t, \mathbf{y}, t)\right] \mathrm{d}t + g(t)\, \mathrm{d}\bar{\mathbf{w}}, \qquad (13)$$

where $\mathbf{f}$ and g are the drift and diffusion terms defined in (6).

We use classical numerical solvers based on discretization of (13) according to a N points grid of the interval $[0,T]$. Since each reverse diffusion step calls the score network, the inference time of diffusion models is higher than their discriminative counterparts, by two orders of magnitude in our case. Fast inference schemes are discussed in the literature and are outside of the scope of this paper.

4. EXPERIMENTAL SETUP

4.1. Data

We use the WSJ0 corpus [29] for most experiments to ensure easier comparison between tasks. For comparison to bandwidth extension baselines, we use the VCTK corpus [30]. All data generation methods are accessible via our web page[2].

Speech Enhancement: The WSJ0+Chime dataset is generated using clean speech extracts from the Wall Street Journal corpus and noise signals from the CHiME3 dataset [31]. The mixture signal is created by randomly selecting a noise file and adding it to a clean utterance with a signal-to-noise ratio (SNR) sampled uniformly between 0 and 20dB.

Speech Dereverberation: The WSJ0+Reverb dataset is generated using clean speech data from the WSJ0 dataset and convolving each utterance with a simulated RIR. We use the PyRoomAcoustics engine [32] to simulate the RIRs. The reverberant room is modeled by sampling uniformly a target T_{60} between 0.4 and 1.0 seconds and room length, width and height in [5,15]×[5,15]×[2,6] m. A dry version of the room is created with the same geometric parameters with a fixed absorption coefficient of 0.99, to generate the corresponding anechoic target.

Bandwidth Extension: The WSJ0+BWR dataset is built with clean speech extracted from the WSJ0 corpus and a similar bandwidth reduction recipe as in [21, 23]. We pick an anti-aliasing filter type among Chebyshev, Butterworth, Elliptic and Bessel and a filter order among$\{2,4,8\}$. Decimating is then realized with a down-scaling factor sampled in$\{2,4,8\}$. The utterance is then resampled at the original 16 kHz with polyphase filtering. To compare against other baselines, we generate VCTK+BWR by replacing WSJ0 with VCTK as the base speech corpus, which we first resample to 16kHz, and use the same process as explained above.

4.2. Hyperparameters and training configuration

Data representation: Utterances are transformed using a short-time Fourier transform (STFT) with a window size of 510, a hop length of 128 and a Hann window. Square-root magnitude compression is carried on the spectrogram. For training, sequences of 256 STFT frames (i.e. 2s) are randomly extracted from the full-length utterances and normalized with respect to the corrupted mixture before being fed to the network.

Forward diffusion: Defined in (6), the stiffness parameter is fixed to γ=1.5, the extremal noise levels to $\sigma_{\min} = 0.05$ and $\sigma_{\max} = 0.5$. The minimal diffusion time defined in (11) is set to $t_\epsilon = 0.03$ as in [8].

Network architecture: The original architecture used for score estimation in [8] is the NCSN++ network proposed in [26]. NCSN++ is a multiresolution U-Net structure which includes in each layer a series of ResNet blocks using 2D convolutions, group normalization and fixed down/upsampling. Attention mechanism is used in the bottleneck, and the network leverages a parallel progressive growing path in addition to the skip connections. The noisy speech spectrogram $\mathbf{y}$ and the current diffusion process estimate $\mathbf{x}_t$ real and imaginary channels are stacked and fed to the network as input. The model is made noise-conditional by feeding each ResNet block with an encoded version of the current noise level $\sigma(t)$. More details about the architecture can be found in [8, 26]. For the generative model proposed in this paper, denoted as *SGMSE+M*, we use a lighter configuration of the NCSN++ architecture called NCSN++M . Ablation studies were designed to halve the number of parameters with almost no degradation, resulting in a network capacity of roughly 27.8M parameters. For the discriminative approach, denoted simply as *NCSN++M* in the following, the noise-conditioning layers are removed This ablation removes only 1.8% of the original number of parameters, which hardly modifies the network capacity.

Training configuration: We train the DNN for a maximum of 300 epochs using early stopping with a patience of 10 epochs. The generative approach SGMSE+M is trained with the denoising score matching criterion (11), and discriminative NCSN++M uses a simple mean-square error loss on the complex spectrogram. We use the Adam optimizer with a learning rate of 10^{-4} and an effective batch size of 16. We track an exponential moving average of the DNN weights with a decay of 0.999.

Inference: 50 time steps are used for reverse inference, adopting the predictor-corrector scheme [26] with one step of annealed Langevin dynamics correction.

4.3. Evaluation metrics

For instrumental evaluation of the speech enhancement and dereverberation performance, we use Perceptual Evaluation of Speech Quality (PESQ) [33] , extended short-term objective intelligibility (ESTOI) [34] and scale-invariant signal to distortion ratio (SI-SDR) [35] . For bandwidth extension we also include log spectral distance (LSD) as a common metric used in the literature. However, it must be stated that the aforementioned instrumental metrics may relate poorly with listening experiments, especially for bandwidth extension. We therefore complement our metrics benchmark with WV-MOS [23][3], which is a DNN-based mean opinion score (MOS) estimation, and was used by the authors for assessment of bandwidth extension performance. For comparability purposes to baselines on the VCTK corpus, we use regular STOI [36] instead of its extended version .

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1. Speech enhancement

In Table 1, we report speech enhancement performance on the WSJ0+Chime dataset. We notice that the generative SGMSE+M produces higher quality samples as measured by WV-MOS and PESQ.

[2]https://uhh.de/inf-sp-sgmsemultitask

[3]https://github.com/AndreevP/wvmos

Table 1. Results for denoising on WSJ0+Chime data.

Method	Type	WV-MOS	PESQ	ESTOI	SI-SDR
Mixture		1.44 ± 1.62	1.70 ± 0.49	0.78 ± 0.14	10.0 ± 5.7
NCSN++M	D	3.65 ± 0.48	2.67 ± 0.69	$\mathbf{0.93 \pm 0.06}$	$\mathbf{19.5 \pm 4.4}$
SGMSE+M	G	$\mathbf{3.77 \pm 0.32}$	$\mathbf{2.94 \pm 0.60}$	0.92 ± 0.06	18.0 ± 5.1

Table 2. Results for dereverberation on WSJ0+Reverb data.

Method	Type	WV-MOS	PESQ	ESTOI	SI-SDR
Mixture		1.78 ± 0.99	1.36 ± 0.19	0.46 ± 0.12	-7.3 ± 5.5
NCSN++M	D	2.96 ± 0.38	2.19 ± 0.48	$\mathbf{0.87 \pm 0.05}$	$\mathbf{7.2 \pm 3.7}$
SGMSE+M	G	$\mathbf{3.43 \pm 0.33}$	$\mathbf{2.64 \pm 0.42}$	$\mathbf{0.87 \pm 0.05}$	6.4 ± 4.2

Table 3. Results for bandwidth extension on WSJ0+BWR data.

Method	Type	WV-MOS ↑	ESTOI ↑	LSD ↓
Mixture		2.45 ± 1.01	0.72 ± 0.21	2.31 ± 0.32
AE-NCSN++M	D	2.17 ± 0.93	0.71 ± 0.19	1.81 ± 0.21
NCSN++M	D	2.25 ± 0.87	0.73 ± 0.16	2.21 ± 0.30
SGMSE+M	G	$\mathbf{3.43 \pm 0.48}$	$\mathbf{0.83 \pm 0.13}$	$\mathbf{1.44 \pm 0.17}$

It is however slightly outperformed by discriminative NCSN++M on intelligibility and noise removal. Indeed, in a denoising task, the interference does not share any information with the target speech, making it relatively easy for a discriminative approach to remove the interference without distorting the target. However, we show in the uploaded listening examples that the discriminative approach tends to destroy low-energy speech regions for low SNRs, whereas the generative model does not. A larger benefit of the generative approach is observed when training and testing data have a stronger mismatch [8].

5.2. Speech dereverberation

In Table 2, we report dereverberation results on the WSJ0+Reverb dataset. Here, generative SGMSE+M clearly outperforms discriminative NCSN++M in terms of quality by a large margin on WV-MOS and PESQ, and performs on par on ESTOI and SI-SDR. For dereverberation, in contrast to denoising, the interference model is completely dependent on the target as it is a filtered version of the latter (Eq. (2)). The generative model is able to extract the speech cues and directly reconstructs it with very little reverberation. The discriminative method, however, cannot do so without introducing significant distortions.

5.3. Bandwidth extension

Results on WSJ0+BWR: In Table 3 we report bandwidth extension performance on the WSJ0+BWR dataset. Interestingly, using a STFT representation did not allow the discriminative approach to recreate the lost high-frequency content. The approach simply learnt an identity mapping, and similar results were observed when experimenting with other STFT-based DNN backbones and data. For this discriminative case, we modified the NCSN++M architecture to use a learnt encoder and decoder, as in e.g. [19] . The resulting approach, denoted as AE-NCSN++M in the following, uses a single 1D convolutional layer with 256 filters of length 510 and stride 128, so that the learnt representation is equivalent to the chosen STFT filterbank. As opposed to NCSN++M, AE-NCSN++M is able to generate high-frequency components, however the reconstruction quality is overall poor, which is to be expected given the generative nature of the bandwidth extension task. By learning an approximate identity mapping, NCSN++M performs better than AE-NCSN++M on instrumental metrics although it does not actually perform bandwidth extension.

Table 4. Results for bandwidth extension on VCTK data. * means that the results were taken from [23]. † means that the method was trained on each bandwidth reduction factor separately.

Bandwidth	Type	1kHz		2kHz		4kHz	
		WV-MOS	STOI	WV-MOS	STOI	WV-MOS	STOI
Mixture		1.36	0.79	2.34	0.89	3.52	0.99
TUNet† [20]	D	-	-	-	-	3.86	0.98
TFiLM*† [19]	D	1.65	0.81	2.27	0.91	3.49	**1.00**
HiFi++*† [23]	G	**3.71**	**0.86**	**3.95**	**0.94**	4.16	**1.00**
VoiceFixer [21]	G	2.50	0.73	3.35	0.78	3.81	0.83
NuWave2 [24]	G	-	-	-	-	3.76	0.97
SGMSE+M	G	3.25	0.83	3.70	0.93	**4.20**	**1.00**

In contrast, generative SGMSE+M performs much better in all metrics, generating plausible content even when the Nyquist frequency is down to 2kHz. When the Nyquist frequency is down to 1kHz, the approach can struggle with generating the right consonants in some cases. Typically the generation process may mistake $[ch]$ for $[s]$ as the information needed to differentiate those sounds is available only at frequencies way above 1kHz. Integrating a linguistic or visual model could be here envisaged to make the approach robust to this lack of acoustic cues.

Results on VCTK+BWR: In Table 4, we compare the proposed generative SGMSE+M on the VCTK+BWR test set against HiFI++ [23], VoiceFixer[4] [21], TFiLM [19], TUNet[5] [20] and NuWave2[6] [24]. Please note that HiFi++, TFiLM and TUNet are trained on each input bandwidth separately, while our generative model SGMSE+M as well as VoiceFixer and NuWave2 are bandwidth-agnostic. We use the official implementations for all approaches without retraining, except HiFi++ as no code is available, and TFiLM as no multi-speaker model is provided. When a method is not trained to restore speech at 16kHz, we use it at the nominal sampling rate then downsample its output to 16kHz. SGMSE+M achieves on par results with HiFi++ on 4kHz bandwidth and worse on 1kHz and 2kHz bandwidths, which is partially due to the fact that HiFi++ is trained separately for each input bandwidth. Using neural vocoders incorporating speech knowledge, as is the case for HiFi++, also probably helps improve robustness to very low input bandwidths. Against all other approaches than HiFi++, SGMSE+M performs significantly better on almost all metrics and conditions.

6. CONCLUSION

The goal of this work is to analyse the potential benefit of recent diffusion-based generative approaches against discriminative approaches on various speech restoration tasks. For this, we apply our recently proposed diffusion generative model to speech enhancement, dereverberation and bandwidth extension, and compare against a discriminative approach using the same DNN architecture . We observe that the generative approach performs globally better than its discriminative counterpart on all tasks, with the strongest benefit for non-additive distortion models, like in dereverberation and bandwidth extension. Furthermore, we show that the proposed bandwidth-agnostic method performs slightly worse or on par in comparison with a recent bandwidth-dependent approach, and largely outperforms other discriminative and bandwidth-agnostic generative approaches.

[4] https://github.com/haoheliu/voicefixer

[5] https://github.com/NXTProduct/TUNet

[6] https://github.com/mindslab-ai/nuwave2

7. REFERENCES

[1] T. Gerkmann and E. Vincent, *Spectral Masking and Filtering*. John Wiley & Sons, 2018.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Int. Conf. Machine Learning (ICML)*, Apr. 2015.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Neural Information Proc. Systems (NIPS)*, Dec. 2020.

[6] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Neural Information Proc. Systems (NIPS)*, Dec. 2019.

[7] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Interspeech*, Sept. 2022.

[8] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *arXiv*, 2022.

[9] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2022.

[10] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv*, 2022.

[11] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," June 2021.

[12] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Asia-Pacific Signal and Information Processing Association (APSIPA)*, Dec. 2021.

[13] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, vol. 59. Springer, 2011.

[14] E. Habets, *Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement*. PhD thesis, 2007.

[15] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2014.

[16] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Sept. 2011.

[17] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 25, no. 7, pp. 1492–1501, 2017.

[18] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. Euro. Signal Proc. Conf. (EUSIPCO)*, Sept. 2019.

[19] S. Birnbaum, V. Kuleshov, Z. Enam, P. W. Koh, and S. Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations," in *Neural Information Proc. Systems (NIPS)*, Dec. 2019.

[20] V.-A. Nguyen, A. H. T. Nguyen, and A. W. H. Khong, "TUNet: A block-online bandwidth extension model based on transformers and self-supervised pretraining," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, June 2022.

[21] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv*, 2021.

[22] H. Liu, W. Y. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural vocoder is all you need for speech super-resolution," in *Interspeech*, Sept. 2022.

[23] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," *arXiv*, 2022.

[24] S. Han and J. Lee, "Nu-wave 2: A general neural audio upsampling model for various sampling rates," in *Interspeech*, Sept. 2022.

[25] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, vol. 82. Journal of the American Statistical Association, 2000.

[26] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Int. Conf. Learning Repr. (ICLR)*, May 2021.

[27] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[28] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching.," *Journal of Machine Learning Research*, vol. 6, p. 695–709, 2005.

[29] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," May 2007.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Speech Synthesis Workshop (SSW)*, Sept. 2016.

[31] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015.

[32] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, Apr. 2018.

[33] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2001.

[34] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Proc.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[35] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - Half-baked or well done?," in *IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, May 2019.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.

B Derivations

B.1 Reverse SDE for Diffusion Processes

Let $\mathbf{x} = \{\mathbf{x}_t\}_{t\in[0,T]}$ be a diffusion process defined by the SDE,

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{B.1}$$

where $\mathbf{f} : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$, $g : [0,T] \to \mathbb{R}$, and $\mathbf{w}_t$ is a standard Wiener process. The density $p(\mathbf{x}, t)$ satisfies the Fokker-Planck equation (FPE),

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = -\nabla \cdot [\mathbf{f}(\mathbf{x}, t)p(\mathbf{x}, t)] + \frac{1}{2}g(t)^2\nabla^2 p(\mathbf{x}, t). \tag{B.2}$$

Define the time-reversed process $\tilde{\mathbf{x}}_t = \mathbf{x}_{T-t}$ with density $\tilde{p}(\mathbf{x}, t) = p(\mathbf{x}, T-t)$. Using the chain rule with $\tau = T - t$, we have

$$\frac{\partial \tilde{p}(\mathbf{x}, t)}{\partial t} = -\frac{\partial p(\mathbf{x}, \tau)}{\partial \tau}. \tag{B.3}$$

Substituting the FPE (B.2) gives the time-reversed FPE

$$\frac{\partial \tilde{p}(\mathbf{x}, t)}{\partial t} = -\nabla \cdot \left[-\tilde{\mathbf{f}}(\mathbf{x}, t)\tilde{p}(\mathbf{x}, t)\right] - \frac{1}{2}\tilde{g}(t)^2\nabla^2\tilde{p}(\mathbf{x}, t), \tag{B.4}$$

where $\tilde{\mathbf{f}}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, T-t)$ and $\tilde{g}(t) = g(T-t)$. To interpret (B.4) as a standard FPE, we rewrite the divergence term

$$\frac{\partial \tilde{p}(\mathbf{x}, t)}{\partial t} = -\nabla \cdot \left[\tilde{g}(t)^2\nabla \log \tilde{p}(\mathbf{x}, t) - \tilde{\mathbf{f}}(\mathbf{x}, t)\right]\tilde{p}(\mathbf{x}, t) + \frac{1}{2}\tilde{g}(t)^2\nabla^2\tilde{p}(\mathbf{x}, t).$$

This corresponds to the SDE

$$\mathrm{d}\mathbf{y}_t = \left[\tilde{g}(t)^2\nabla \log \tilde{p}(\mathbf{y}_t, t) - \tilde{\mathbf{f}}(\mathbf{y}_t, t)\right]\mathrm{d}t + \tilde{g}(t)\mathrm{d}\mathbf{w}_t. \tag{B.5}$$

The process $\mathbf{y}_t$ defined by (B.5) has the same marginal distributions as the time-reversed process $\tilde{\mathbf{x}}_t$. The key term $\nabla \log \tilde{p}(\mathbf{y}_t, t)$ is the score function, which accounts for the time-reversal adjustment.

B.2 Mean and Variance of the Perturbation Kernel

Consider the multi-dimensional SDE in the complex time-frequency domain

$$\mathrm{d}\mathbf{x}_t = \gamma(\mathbf{y} - \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{B.6}$$

where $\mathbf{y}, \mathbf{x}_t \in \mathbb{C}^{F\times N}$ represent the mixture and process state, respectively, and $\mathbf{w}_t \in \mathbb{C}^{F\times N}$ is a complex Wiener process with independent components. Since $\gamma \in \mathbb{R}$ and $g(t) \in \mathbb{R}$, the system decouples into independent scalar SDEs for each entry (i, j)

$$\mathrm{d}x_t(i,j) = \gamma(y(i,j) - x_t(i,j))\mathrm{d}t + g(t)\mathrm{d}w_t(i,j). \tag{B.7}$$

For simplicity, we drop the indices (i, j) and analyze the scalar case. The mean $m(t) = \mathbb{E}[x_t|x_0]$ satisfies the deterministic ODE

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \gamma(y - m(t)), \quad m(0) = x_0. \tag{B.8}$$

First consider the homogeneous equation

$$\frac{\mathrm{d}m_h(t)}{\mathrm{d}t} + \gamma m_h(t) = 0, \tag{B.9}$$

which has the solution

$$m_h(t) = Ce^{-\gamma t}. \tag{B.10}$$

For the non-homogeneous equation, we find a particular solution $m_p(t)$ using the constant ansatz method. Assume

$$m_p(t) = A \quad \text{(constant)}. \tag{B.11}$$

Substituting into the ODE

$$0 + \gamma A = \gamma y \implies A = y. \tag{B.12}$$

Thus, the particular solution is

$$m_p(t) = y. \tag{B.13}$$

Combining both solutions

$$m(t) = m_h(t) + m_p(t) = Ce^{-\gamma t} + y. \tag{B.14}$$

Applying the initial condition $m(0) = x_0$

$$x_0 = C + y \implies C = x_0 - y. \tag{B.15}$$

Therefore, the exact solution is

$$m(t) = e^{-\gamma t}x_0 + (1 - e^{-\gamma t})y. \tag{B.16}$$

The variance $\sigma^2(t) = \mathbb{E}[|x_t - m(t)|^2]$ satisfies

$$\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t} = -2\gamma\sigma^2(t) + g^2(t), \quad \sigma^2(0) = 0. \tag{B.17}$$

Using the integrating factor $e^{2\gamma t}$

$$\sigma^2(t) = e^{-2\gamma t} \int_0^t e^{2\gamma s} g^2(s)\, \mathrm{d}s.$$

When $g(t) = ck^t$ $(c, k > 0)$, the variance becomes

$$\begin{aligned}
\sigma^2(t) &= c^2 e^{-2\gamma t} \int_0^t e^{2\gamma s} k^{2s}\, \mathrm{d}s \\
&= c^2 e^{-2\gamma t} \int_0^t e^{2s(\gamma + \log k)}\, \mathrm{d}s \\
&= \frac{c^2}{2(\gamma + \log k)} \left(e^{2t(\gamma + \log k)} - 1 \right) e^{-2\gamma t} \\
&= \frac{c^2 (k^{2t} - e^{-2\gamma t})}{2(\gamma + \log k)}.
\end{aligned}$$

The perturbation kernel is given by

$$p_{0t}(x_t | x_0, y) = \mathcal{N}_{\mathbb{C}} \left(x_t; e^{-\gamma t} x_0 + (1 - e^{-\gamma t}) y, \frac{c^2 (k^{2t} - e^{-2\gamma t})}{2(\gamma + \log k)} \right). \tag{B.18}$$

B.3 Derivation for the EDM framework

Consider the SDE,

$$\mathrm{d}x(t) = f(t)\,x(t)\,\mathrm{d}t + g(t)\,\mathrm{d}w(t). \tag{B.19}$$

The perturbation kernel of the SDE (B.19) has the general form

$$p_{0t}(x(t)\,|\,x(0)) = \mathcal{N}\left(x(t); \alpha(t)x(0), \alpha(t)^2\sigma(t)^2\right), \tag{B.20}$$

where

$$\alpha(t) = \exp\left(\int_0^t f(s)\,\mathrm{d}s\right), \quad \sigma(t) = \sqrt{\int_0^t \frac{g(s)^2}{\alpha(s)^2}\,\mathrm{d}s}. \tag{B.21}$$

Proof: Using the integration factor $I(t) = \exp\left(-\int_0^t f(s)\mathrm{d}s\right)$, we multiply both sides of (B.19) to obtain:

$$I(t)\mathrm{d}x = I(t)f(t)x\mathrm{d}t + I(t)g(t)\mathrm{d}w. \tag{B.22}$$

Applying Itô's lemma to $y = I(t)x(t)$ gives:

$$\mathrm{d}y = I(t)\mathrm{d}x + x(t)\mathrm{d}I(t), \tag{B.23}$$

since $I(t)$ is deterministic. Substituting $\mathrm{d}I(t) = -f(t)I(t)\mathrm{d}t$ yields:

$$\mathrm{d}y = I(t)g(t)\mathrm{d}w. \tag{B.24}$$

Integrating from 0 to t:

$$I(t)x(t) - x(0) = \int_0^t I(s)g(s)\mathrm{d}w(s), \tag{B.25}$$

which gives the solution:

$$x(t) = \alpha(t)\left(x(0) + \int_0^t \frac{g(s)}{\alpha(s)}\mathrm{d}w(s)\right). \tag{B.26}$$

The mean and variance are:

$$\mathbb{E}[x(t)] = \alpha(t)x(0), \tag{B.27}$$

$$\mathrm{Var}[x(t)] = \alpha(t)^2 \int_0^t \frac{g(s)^2}{\alpha(s)^2}\mathrm{d}s. \tag{B.28}$$

Thus, $x(t)$ follows a normal distribution as given in (B.20).

B.4 Connection between Score Matching and Denoising

Consider a dataset $\mathcal{D} = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$ of N pairs of clean and noisy speech $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^d$. The time-dependent mean is given by

$$\boldsymbol{\mu}_t^{(i)} = e^{-\gamma t}\mathbf{x}^{(i)} + (1 - e^{-\gamma t})\mathbf{y}^{(i)} \tag{B.29}$$

where γ is the stiffness parameter.

The time-dependent data distribution is given by:

$$p_{\text{data}}(\mathbf{x}; t) = \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x} - \boldsymbol{\mu}_t^{(i)}) \tag{B.30}$$

where δ is the Dirac delta function.

The time-dependent perturbed data distribution is obtained by convolving with Gaussian noise:

$$p(\mathbf{x}; t) = p_{\text{data}}(\mathbf{x}; t) * \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma(t)^2\mathbf{I}) \tag{B.31}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_t^{(i)}, \sigma(t)^2\mathbf{I}) \tag{B.32}$$

Consider the denoising objective $\mathcal{L}$ based on MSE between a denoiser D_θ and target $\boldsymbol{\mu}_t$:

$$\mathcal{L}(D_\theta; t) = \mathbb{E}_{\boldsymbol{\mu}_t \sim p_{\text{data}}} \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_t, \sigma(t)^2\mathbf{I})} \|D_\theta(\mathbf{x}; t) - \boldsymbol{\mu}_t\|_2^2 \tag{B.33}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \int \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_t^{(i)}, \sigma(t)^2\mathbf{I}) \|D_\theta(\mathbf{x}; t) - \boldsymbol{\mu}_t^{(i)}\|_2^2 \mathrm{d}\mathbf{x} \tag{B.34}$$

The optimal denoiser is obtained by solving:

$$D_\theta(\mathbf{x}; t) = \frac{\sum_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_t^{(i)}, \sigma(t)^2\mathbf{I})\boldsymbol{\mu}_t^{(i)}}{\sum_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_t^{(i)}, \sigma(t)^2\mathbf{I})} \tag{B.35}$$

The score of the perturbed distribution is:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}; t) = \frac{\sum_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_t^{(i)}, \sigma(t)^2\mathbf{I}) \frac{\boldsymbol{\mu}_t^{(i)} - \mathbf{x}}{\sigma(t)^2}}{\sum_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_t^{(i)}, \sigma(t)^2\mathbf{I})} \tag{B.36}$$

$$= \frac{D_\theta(\mathbf{x}; t) - \mathbf{x}}{\sigma(t)^2} \tag{B.37}$$

The connection between score matching and denoising is given by:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}; t) = \frac{D_\theta(\mathbf{x}; t) - \mathbf{x}}{\sigma(t)^2} \tag{B.38}$$

This shows that the score function can be estimated using a denoising autoencoder.

List of Figures

1.1 Speech enhancement scenario 3
1.2 Typical speech signal and cafeteria background noise signal 4
1.3 Comparison of predictive and generative approaches for speech enhancement 10
1.4 Masking-based speech enhancement 12
1.5 Schematic overview of deep generative models 18
1.6 Taxonomy of learning-based speech enhancement 26
1.7 Illustration of multimodal machine learning 30

6.1 Number of accepted papers at selected conferences 136