LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading

Leyuan Qu, Cornelius Weber, and Stefan Wermter, *Member, IEEE*

***Abstract*— The aim of this work is to investigate the impact of crossmodal self-supervised pre-training for speech reconstruction (video-to-audio) by leveraging the natural co-occurrence of audio and visual streams in videos. We propose LipSound2 that consists of an encoder–decoder architecture and location-aware attention mechanism to map face image sequences to mel-scale spectrograms directly without requiring any human annotations. The proposed LipSound2 model is first pre-trained on ∼2400-h multilingual (e.g., English and German) audio-visual data (VoxCeleb2). To verify the generalizability of the proposed method, we then fine-tune the pre-trained model on domain-specific datasets (GRID and TCD-TIMIT) for English speech reconstruction and achieve a significant improvement on speech quality and intelligibility compared to previous approaches in speaker-dependent and speaker-independent settings. In addition to English, we conduct Chinese speech reconstruction on the Chinese Mandarin Lip Reading (CMLR) dataset to verify the impact on transferability. Finally, we train the cascaded lip reading (video-to-text) system by fine-tuning the generated audios on a pre-trained speech recognition system and achieve the state-of-the-art performance on both English and Chinese benchmark datasets.**

***Index Terms*— Lip reading, self-supervised pre-training, speech recognition, speech reconstruction.**

I. Introduction

INSPIRED by human bimodal perception [1] in which both sight and sound are used to improve the comprehension of speech, a lot of effort has been spent on speech processing tasks by leveraging visual information, for example, integrating simultaneous lip movement sequences into speech recognition [2], [3], guiding neural networks in isolating target speech signals with a static face image for speech separation [4], [5], and grounding speech recognition with visual objects and scene information [6], [7]. Multimodal audio-visual methods achieve significant improvement over single modality models since the visual signals are invariant to acoustic noise and complementary to auditory representations [8]. Moreover, the visual contribution becomes more important as the acoustic signal-to-noise ratio is decreased [9].

In most approaches, the visual information is mainly used as an auxiliary input to complement audio signals. However, in some circumstances, the auditory information may be absent or extremely noisy, which motivates speech reconstruction. Speech reconstruction aims to generate both intelligible and qualified speech by only conditioning on image sequences of talking mouths or faces. Generating intelligible speech from silent videos enables many applications, e.g., a silent visual input method on mobile phones for privacy protection in public areas [10]; communication assistance for patients suffering laryngectomy [11]; surveillance video understanding when only visual signals are available [12]; enhancement of video conferences or far-field human–robot interaction scenarios in a noisy environment [13]; and nondisruptive user intervention for autonomous vehicles [14].

It is challenging to reconstruct qualified and intelligible speech from only mouth or face movements since human speech is produced by not only externally observable organs such as lips and tongue but also internally invisible ones that are difficult to capture in most cases [15], for instance, vocal cords and pharynx. Consequently, it is hard to infer fundamental frequency or voicing information controlled by these organs. Moreover, some phonemes are acoustically discriminative but not easy to distinguish visually since the phonemes share the same places of articulation but with different manners of articulation [16], for example, /v/ and /f/ in English are both fricatives and look the same on lip and teeth movements but are different on the vibration of vocal cords (voiced versus unvoiced) and the attribute of aspirate (unaspirated versus aspirated) that are not visible in most video recordings. Hence, predicting human voices from appearance is still a challenging task [17].

In recent years, there has been a growing interest in speech reconstruction and variant methods have been proposed. A possible technique is to run lip reading (video-to-text) and text-to-speech (TTS) systems in cascade, but the lip reading performance is still unsatisfactory and the error is being propagated to TTS. Alternatively, other researchers directly estimate speech representations, for example, linear predictive coding (LPC) [18], bottleneck features [19], and mel-scale spectrograms [20], from videos, followed by a vocoder used to

Manuscript received 10 December 2020; revised 28 July 2021 and 16 February 2022; accepted 6 July 2022. Date of publication 22 July 2022; date of current version 6 February 2024. This work was supported in part by the China Scholarship Council (CSC) and in part by the German Research Foundation DFG under project CML TRR 169. *(Corresponding author: Leyuan Qu.)*

The authors are with the Knowledge Technology Institute, Department of Informatics, University of Hamburg, 22527 Hamburg, Germany (e-mail: quleyuan9826@gmail.com; cornelius.weber@uni-hamburg.de; stefan.wermter@uni-hamburg.de).

Digital Object Identifier 10.1109/TNNLS.2022.3191677

transform intermediate representations to audio, for instance, STRAIGHT [21] and WORLD vocoder [22]. In contrast, the information of speaker identity and speaking styles can be relatively preserved. However, most existing works only focus on speaker-dependent settings with a small vocabulary or artificial grammar dataset or even builds one model for each individual speaker, which does not meet the requirements in realistic scenarios.

In our previous work, we proposed LipSound [20] to directly map visual sequences to low-level speech representation, i.e., mel spectrogram, which is inspired by audio-visual self-supervised representation learning. By leveraging the natural co-occurrence of audio and visual streams in videos without requiring any human annotations or treating one modality as the supervision of the other, self-supervised representation learning has received substantial interest, for example, learning representations by matching the temporal synchronization [23] or spatial alignment [24] of audio and video clips for action recognition.

In comparison to our previous work, LipSound that only focuses on speaker-dependent settings for the GRID artificial grammar dataset, in this article, we further explore to what extent the large-scale crossmodal self-supervised pre-training can benefit speech reconstruction in generalizability (speak-independent) and transferability (non-Chinese to Chinese) on a large vocabulary continuous speech corpus TCD-TIMIT. In addition, we also changed the LipSound architecture substantially by replacing the 1-D convolutional neural network (CNN) with 3-D CNN blocks (Conv 3D + Batch Norm + ReLU + Max Pooling + Dropout). This should enable the model to directly learn stable representations from raw pixels and using a location-aware attention mechanism to make the alignments between encoder and decoder more robust to nonverbal areas. Moreover, we replace the Griffin–Lim algorithm [25] with a neural vocoder to smoothly generate waveforms and voices.

As shown in Fig. 1(a), our approach is first pre-training the Lipsound2 model on a large-scale multilingual audio-visual corpus (VoxCeleb2) to map silent videos to mel spectrogram and then fine-tuning the pre-trained model on specific domain datasets [GRID, TCD-TIMIT, and Chinese Mandarin Lip Reading (CMLR)], followed by a neural vocoder (WaveGlow [26]) to reconstruct estimated mel spectrogram to waveforms. Lip reading (video-to-text) experiments are performed by fine-tuning the generated audios on a pre-trained acoustic model (Jasper [27]) in Fig. 1(b).

The main contributions of this article are given as follows.

1) We propose an autoregressive encoder–decoder with attention architecture, LipSound2, to directly map silent facial movement sequences to mel-scale spectrograms for speech reconstruction, which does not require any human annotations.
2) We explore the model generalizability on speaker-independent and large-scale vocabulary datasets which few studies have focused on, and we achieve better performance on speech quality and intelligibility in the speech reconstruction task.

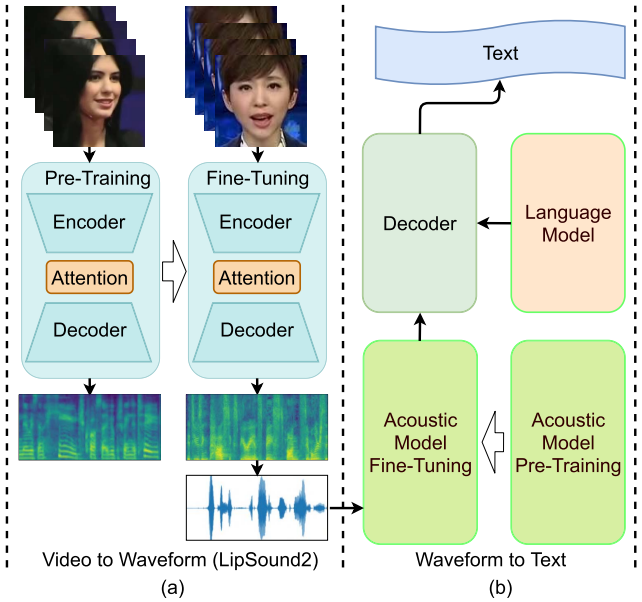

Fig. 1. Process of (a) video-to-waveform generation and (b) waveform-to-text transformation.

3) To the best of our knowledge, no previous research has investigated Chinese speech reconstruction in speaker-dependent and speaker-independent cases.
4) By leveraging the large-scale self-supervised pre-training on LipSound2 and the advanced Jasper speech recognition model, our cascaded lip reading system outperforms existing models by a margin on both English and Chinese corpora.

This article is organized as follows. Section II reviews related work on lip-to-speech reconstruction, lip reading, and self-supervised learning. Section III provides the model details, followed by the description of datasets and evaluation metrics in Section IV. Experimental results and discussion are presented in Sections V and VI, respectively. We conclude this article in Section VII.

II. Related Work

A. Lip-to-Speech Reconstruction

In recent years, researchers have investigated a variety of approaches to speech reconstruction from silent videos. We only review the neural network methods in this article.

Cornu and Milner [28] proposed to use fully connected (FC) neural networks to estimate spectral envelope representations, for instance, LPC coefficients and mel filter bank amplitudes, from visual feature inputs, such as 2-D discrete cosine transform, followed by a STRAIGHT vocoder [21], which is used to synthesize time-domain speech signals from the estimated representations. Follow-up work [29] predicts speech-related codebook entries with a classification framework to get further improvement on speech intelligibility. Instead of using handcrafted visual features, Ephrat and Peleg [18] utilized CNNs to automatically learn optimal features from raw pixels and show promising results on out-of-vocabulary experiments. Subsequently, improved results are reported by

Ephrat *et al.* [30] via combining a RestNet backbone and a postprocessing network on a large-scale vocabulary dataset, TCD-TIMIT [31]. Akbari *et al.* [19] treated the intermediate bottleneck features learned by a speech autoencoder as training targets by conditioning on lip reading network outputs. Kumar *et al.* [32] validated the effectiveness of using multiple views of faces on both speaker-dependent and speaker-independent speech reconstruction. Vougioukas *et al.* [33] utilized generative adversarial networks (GANs) to directly predict raw waveforms from visual inputs in an end-to-end fashion without generating an intermediate representation of audios. Inspired by the speech synthesis model, Tacotron2 [34], Qu *et al.* [20] proposed to directly map video inputs to low-level speech representations, mel spectrogram, with an encoder–decoder architecture and achieve better results on lip reading experiments. Afterward, Prajwal *et al.* [35] improved the model performance with 3-D CNN and skip connections. Recently, Michelsanti *et al.* [36] presented a multitask architecture to learn spectral envelope, aperiodic parameters, and fundamental frequency separately, which are then fed into a vocoder for waveform synthesis. They integrate a connectionist temporal classification (CTC) [37] loss to jointly perform lip reading, which is capable of further enhancing and constraining the video encoder.

In addition to sequences of lip or face images, further signals can be used for temporal self-supervision. For instance, Gonzalez *et al.* [38] generated speech from articulatory sensor data and Akbari *et al.* [39] reconstructed speech from invasive electrocorticography. However, most existing works only focus on a speaker-dependent setting and small vocabulary or artificial grammar datasets. In this article, we evaluate our method not only on speaker-dependent experiments but also pay attention to speaker-independent and large-scale vocabulary setups.

B. Lip Reading

Lip reading, also known as visual speech recognition, is the task to predict text transcriptions from silent videos, such as mouth or face movement sequences. Research on lip reading has a long tradition. Approaches to lip reading generally fall into two categories on feature level: 1) handcrafted visual feature extraction, such as discrete cosine transform [40], discrete wavelet transform [41], or active appearance models [42] and 2) representations learned by neural networks, which has become the dominant technique for this task, for example, using convolutional autoencoders [43], spatiotemporal CNNs [44], long short-term memory [45], or residual networks [46].

Alternatively, methods on modeling units for lip reading can be divided into word and character levels.

1) In the case of word-level units, lip reading is simplified as a classification task. Word-level lip reading datasets and benchmarks are built, for instance, LRW [47] for English and LRW-1000 [48] for Chinese. Stafylakis and Tzimiropoulos [46] adopted spatiotemporal convolutional networks and 2-D ResNet as front end to extract visual features and bidirectional long short-term memory networks as the backend to capture temporal information and attain significant improvement. Weng and Kitani [49] presented two separated deep 3-D CNN front ends to learn features from grayscale video and optical flow inputs. Martinez *et al.* [50] replaced recurrent neural networks widely used in past work with temporal convolutional networks to simplify the training procedure. The word-level methods are usually able to achieve high accuracy, and however, the models disregard the interaction or co-articulation phenomenon between phonemes or words. A predefined lexicon with closed-set vocabulary is used and words are usually treated as isolated units in speech. Thereby, long-term context information and assimilation or dissimilation effects are completely neglected. Moreover, it is hard to recognize out-of-vocabulary words.
2) Lip reading models with character or phoneme levels mainly use methods proposed in speech recognition. Assael *et al.* [44] conducted end-to-end lip reading experiments on sentence level with CTC loss. Subsequently, sequence discriminative training [51] and domain-adversarial training [52] are introduced to lip reading. Chung *et al.* [2] collected the dataset, "lip reading sentence" (LRS), which consists of hundreds of thousands of videos from BBC television, and significantly promoted the research on sentence-level lip reading. Shillingford *et al.* [53] verified the effectiveness of large-scale data (3886 h of video) for training continuous visual speech recognition. Afouras *et al.* [54] compared the performance of recurrent neural networks, fully CNNs, and transformer on lip reading character recognition.

Different from the mainstream methods which directly transform videos to text, we perform lip reading experiments in a cascaded manner, in which the silent videos are first mapped to audios with our LipSound2 model and, then, text transcriptions are predicted by fine-tuning on a pretrained speech recognition system.

C. Self-Supervised Learning

As a form of unsupervised learning, self-supervised learning leverages massive unlabeled data and aims to learn effective intermediate representations with the supervision of self-generated labels. Training unlabeled data in a supervised manner rely on the pretext tasks that determine what labels and loss functions to be used. In computer vision, the pretext tasks can be predicting angles of rotated images [55], learning the relative position of segmented regions in an image [56], placing shuffled patches back [57], or colorizing grayscale input images [58]. The video-based pretext tasks can be tracking moving objects in videos [59], validating temporal frame orders [60], video colorization [61], and so on.

Self-supervised learning is also widely used in natural language processing. Substantial progress has been made recently, where diverse pretext tasks are proposed, for instance, predicting center words using surrounding ones or vice versa [62], generating the next word by conditioning on previous words in an autoregressive fashion [63], completing masked tokens

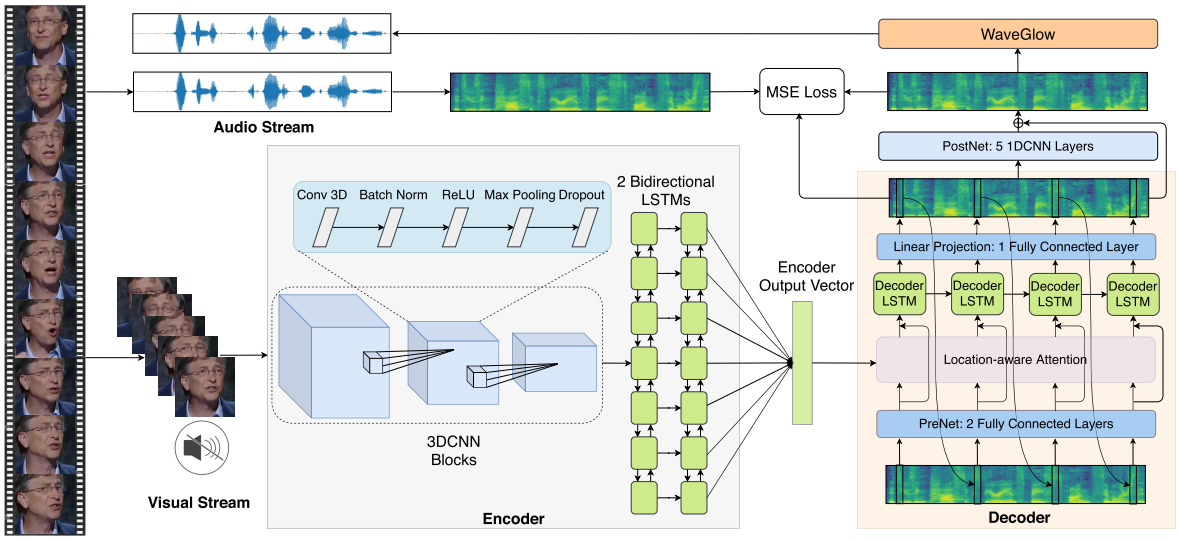

Fig. 2. Architecture of LipSound2. The video is split into visual and acoustic streams. The face region, which is cropped from the silent visual stream, is used as the model input. The acoustic spectrogram features extracted from the counterpart audio stream are used as the training target. During training, the ground-truth spectrogram frames are utilized to accelerate convergence, while, during inference, the outputs from previous steps are used.

or consecutive utterances [64], recovering the order of shuffled words [65], or the permutation of rotated sentence [66].

Inspired by the strong correlation between different modalities where, for example, the audio and visual modalities are semantically consistent or temporally synchronous, more and more researchers work on multimodal or cross-modal self-supervised learning. Multimodal self-supervised learning aims at learning joint or shared latent spaces or representations, while cross-modal self-supervised learning lets one modality supervise another. Here, we only review the audio-visual modalities since this is the main focus of this article. Different pretext tasks are designed according to the correspondence and synchronization of audio and visual modalities, for instance, predicting whether image and audio clips correspond, to enable neural networks to classify sounds [67], learn cross-modal retrieval [68], or locate the sound source in an image [69]. Besides, multimodal self-supervised representation learning can also be performed by matching the temporal synchronization [23] or spatial alignment [24] of audio and video clips in the context of action recognition, where a contrastive loss and a clustering loss are combined to learn high-level semantic representations for visual event and concept understanding [70]. In this article, we focus on cross-modal self-supervised learning where the corresponding audio signals provide the supervision for face sequence inputs.

III. Model Architecture

Fig. 2 shows the LipSound2 model architecture. We split the video clips into an audio stream used as training target and a visual stream used as model input. The system consumes the visual part to predict the audio counterpart in a self-supervised fashion. The proposed architecture is composed of an encoder–decoder and an attention model to map the soundless visual sequences to the low-level acoustic representation, mel-scale spectrograms. Advantages are that, in contrast to directly predicting raw waveform, working with mel spectrogram not only reduces computational complexity but also easily learns long-distance dependence. Model details are listed in Table I. Then, a pre-trained neural vocoder, WaveGlow, follows to reconstruct the raw waveform from the generated mel spectrogram.

TABLE I

Configuration of LipSound2 Encoder, Decoder, Attention, and Postnet

Layer	Kernel	Stride	Padding	Channels/Nodes
Encoder				
Conv3D 1	$5 \times 3 \times 3$	$[1, 2, 2]$	$[2, 0, 0]$	32
MaxPool3D	$1 \times 2 \times 2$	$[1, 2, 2]$	$[0, 0, 0]$	-
Conv3D 2	$5 \times 3 \times 3$	$[1, 2, 2]$	$[2, 0, 0]$	64
MaxPool3D	$1 \times 2 \times 2$	$[1, 2, 2]$	$[0, 0, 0]$	-
Conv3D 3	$5 \times 3 \times 3$	$[1, 1, 1]$	$[2, 0, 0]$	128
MaxPool3D	$1 \times 2 \times 2$	$[1, 2, 2]$	$[0, 0, 0]$	-
BiLSTM1	-	-	-	128
BiLSTM2	-	-	-	128
Attention				
Attention LSTM	-	-	-	1024
Query FC	-	-	-	128
Memory FC	-	-	-	128
Location Conv1D	31	1	15	32
Location FC	-	-	-	128
Weight FC	-	-	-	1
Decoder				
PreNet FC 1	-	-	-	512
PreNet FC 2	-	-	-	256
Decoder LSTM	-	-	-	1024
Linear Projection FC	-	-	-	80
PostNet				
Conv1D 1	5	1	2	512
Conv1D 2	5	1	2	512
Conv1D 3	5	1	2	512
Conv1D 4	5	1	2	512
Conv1D 5	5	1	2	80

A. Encoder

The multitask CNN (MTCNN) [71] is used to detect face landmarks from raw videos. We crop only the face region (112×112 pixels) and smooth all frame landmarks since

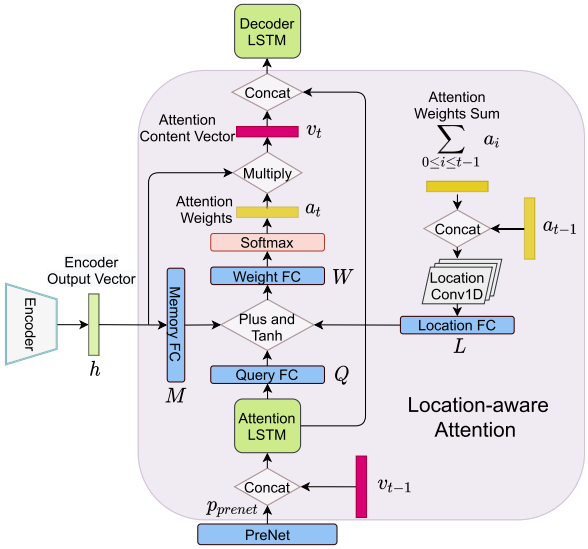

Fig. 3. Computational flow of location-aware attention at time step t.

low-resolution videos or profile faces lead to detection failures sometimes and landmark smoothing can eliminate frame skip in adjacent images. The cropped face sequences are then fed into 3-D CNN blocks and each block is based on a 3-D CNN, batch normalization, ReLU activation, max pooling, and dropout, as shown in Fig. 2. Then, two bidirectional LSTM layers follow which capture the long-distance dependence from the left and right context.

B. Location-Sensitive Attention

We use location-aware attention [72] to bridge the encoder and the decoder. The image sequence input $i = (i_0, \ldots, i_n)$ is first embedded into the latent space representation vector $h = (h_1, \ldots, h_n)$ by the encoder with the same dimension n in time, and then, the intermediate vector h is decoded into the mel spectrogram $o = (o_0, \ldots, o_m)$. At time step t $(0 \le t \le m)$, the attention weight a_t can be obtained by the following equations:

$$a_t = \text{Softmax}(W \cdot \tanh(M \cdot h + Q \cdot x + L \cdot y)) \quad (1)$$

$$x = \text{LSTM}(h \cdot a_{t-1}, p_{\text{prenet}}) \quad (2)$$

$$y = \text{Conv}\left(a_{t-1}, \sum_{0\le i\le t-1} a_i\right) \quad (3)$$

where W, M, Q, and L are the matrices learned by weight FC, memory FC, query FC, and location FC, respectively. In (3), the sum of attention weights of all previous steps is integrated, which enables the current step attention to be aware of the global location and move forward monotonically. Fig. 3 visualizes the computational flow of the attention mechanism. The attention content vector v_t can be obtained by multiplying the encoder output by the normalized attention weights (see the following equation):

$$v_t = a_t \cdot h. \quad (4)$$

C. Decoder

The decoder module consists of one unidirectional LSTM layer and one linear projection layer. The decoder LSTM consumes the attention content vector and the output from attention LSTM to generate one frame at a time. Subsequently, the linear projection layer maps the decoder LSTM outputs to the dimension of the mel-scale filter bank. During training, we use ground-truth mel-spectrogram frames as PreNet inputs, and during inference, the predicted frames from previous time steps are used. Since the decoder only receives past information at every time step, after decoding, five Conv1D layers (postnet) are used to further improve the model performance by smoothing the transition of adjacent frames and using future information, which is not available when decoding.

D. Training Objective

The loss function is the sum of two mean square errors (MSEs), as shown in (5), i.e., the MSE between the decoder output O_{dec} and the target mel spectrogram M_{tar} and the MSE between the postnet output O_{post} and the target mel spectrogram

$$\text{Loss} = \text{MSE}(O_{\text{dec}}, M_{\text{tar}}) + \text{MSE}(O_{\text{post}}, M_{\text{tar}}). \quad (5)$$

E. WaveGlow

We use WaveGlow [26], which combines the approach of the glow-based generative model [73] and the architecture insight of WaveNet [74] to transform the estimated mel spectrogram back to audio. WaveGlow abandons autoregression [74] and speeds up the procedure of waveform synthesis in high quality and resolution. We train WaveGlow from scratch using the same settings as original work [26] but in 16k sampling rate on the LJSpeech dataset [75] to meet the requirement of following up ASR models. To our surprise, the WaveGlow model that is trained with only one female voice can effectively generalize to any unseen voices and stably perform waveform reconstruction.

F. Acoustic Model and Language Model

The Jasper [27] speech recognition system, which is a fully convolutional architecture trained with skip connections and CTC loss, is adopted to directly predict characters from speech signals. We pretrain the Jasper DR 10×5 model[1] on 960 h LibriSpeech and 1000 h AISHELL-2 corpora, which achieves 3.61% word error rate (WER) and 10.05% character error rate (CER) on the development set for English and Chinese, respectively.

Beam search is utilized to decode the output character possibilities from Jasper and a 6-g KenLM [76] language model[2] into grammatically and semantically correct words on sentence level [77].

IV. Experimental Setup

A. Dataset

All datasets used in this article are summarized in Table II and random frames from audio-visual ones are presented

[1]https://nvidia.github.io/OpenSeq2Seq/html/speech-recognition.html
[2]https://github.com/PaddlePaddle/DeepSpeech

TABLE II
OVERVIEW OF ALL CORPORA USED IN THIS ARTICLE. SPK: SPEAKERS. UTT: UTTERANCES. VOCAB: VOCABULARY

Language	Dataset	#Spk.	#Utt.	#Vocab.	#hours	Usage	Modality
Multi-Language	VoxCeleb2 [78]	6112	1.1M	-	2442	LipSound2 pre-training	Audio-Visual
English	GRID [79]	51	33k	51	27.5	LipSound2 fine-tuning	
	TCD-TIMIT [31]	59	5.4k	5.9k	7		
	LJSpeech [75]	1	13.1k	-	24	WaveGlow training	Audio
	LibriSpeech [81]	2484	292.3k	-	960	Acoustic model pre-training	
Chinese	CMLR [80]	11	102k	3.5k	87.7	LipSound2 fine-tuning	Audio-Visual
	AISHELL-2 [82]	1991	-	-	1000	Acoustic model pre-training	Audio

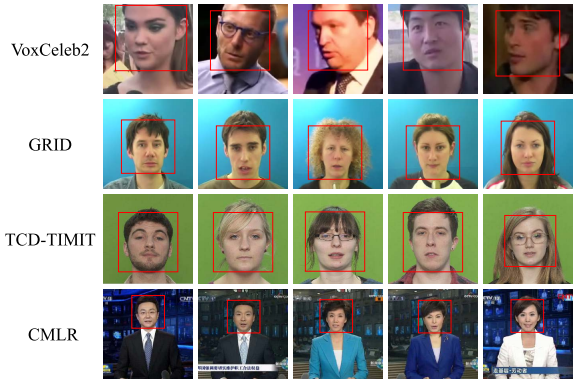

Fig. 4. Random face samples from audio-visual corpora. Only the face region is cropped during training and test. Samples from audio-visual corpora [31], [78], [79], [80].

in Fig. 4. VoxCeleb2 is a large-scale audio-visual corpus, extracted from YouTube videos, containing over one million utterances and more than 6k different speakers from around 145 nationalities and languages. It includes noisy and unconstrained conditions; specifically, the audio stream may be recorded with background noise, such as laughter and room reverberation, and the vision part may contain variable head poses (e.g., frontal faces and profile), variable lighting conditions, and low image quality, while the GRID and TCD-TIMIT datasets are in controlled experimental environments with fixed frontal face angle and clean background in audio and vision. It is worth mentioning that the GRID dataset is designed to contain only a fixed six-word structure and all sentences are generated by a restricted artificial grammar: command + color + preposition + letter + digit + adverb, for example, set blue in Z three now. CMLR is collected from videos by 11 hosts of the Chinese national news program *News Broadcast*, which contains frontal faces and covers a large amount of Chinese vocabulary. We first pretrain LipSound2 on VoxCeleb2 and then fine-tune the model on GRID, TCD-TIMIT, and CMLR for video to mel-spectrogram reconstruction.

LibriSpeech and AISHELL-2 are the current largest open-source speech corpora and widely used speech recognition benchmarks for English and Chinese, respectively. LibriSpeech is derived from audiobooks, containing 460 h of clean speech and 500 h of noisy speech. AISHELL-2 consists of 1000-h different domain speech, for instance, voice command and smart home scenario, and includes various accents from different areas of China. We use LibriSpeech and AISHELL-2 to pretrain the Jasper acoustic model to boost the performance of waveform-to-text transformation. The generated speech on GRID, TCD-TIMIT, and CMLR is used for further fine-tuning to perform lip reading (video-to-text) experiments.

The LJ Speech dataset with only one female voice is especially designed for speech synthesis tasks, which is used for WaveGlow training, in this article, to transform mel spectrogram back to waveforms.

B. Evaluation Metrics

We evaluate the generated speech quality and intelligibility with perceptual evaluation of speech quality (PESQ) [83] and extended short-time objective intelligibility (ESTOI) [84], respectively. The speech-to-text results are measured with WER and CER, the ratio of error terms, i.e., substitutions, deletions, and insertions, to the total number of words/characters in the ground-truth sequences.

C. Training

We only describe the training settings of LipSound2 pre-training, LipSound2 fine-tuning, and Jasper acoustic model fine-tuning. More details about Japser[1] pre-training acoustic model, KenLM[2] language model, and WaveGlow[3] can be found on the open-source websites.

1) Vision Stream: Face landmarks are detected using MTCNN [71] from all video frames and only the face area is cropped and reshaped to size of 112 × 112 as inputs. We also add one "visual period"—an empty frame with all values of 255—at the end of every visual stream to help the decoder stop decoding at the right time. A max decoder step threshold of 1000 is activated to terminate decoding when the decoder fails to capture the "visual period."

[3]https://github.com/NVIDIA/waveglow

TABLE III
SPEAKER-DEPENDENT SPEECH RECONSTRUCTION RESULTS ON GRID AND TCD-TIMIT DATASETS

	GRID		TCD-TIMIT		
Model	ESTOI	PESQ	ESTOI	PESQ	Parameters
Vid2Speech [18]	0.335	1.734	0.298	1.136	0.9M
Lip2AudSpec [19]	0.352	1.673	0.316	1.254	13.0M
Vougioukas et al. [33]	0.361	1.684	0.321	1.218	not available
Ephrat et al. [30]	0.376	1.825	0.310	1.231	9.2M
Lip2Wav [35]	0.535	1.772	0.365	1.350	not available
vid2voc-M-VSR [36]	0.455	1.900	-	-	5.1M
LipSound2	**0.592**	**2.328**	**0.372**	**1.490**	8.5M

2) Audio Stream: We first divide the raw waveforms by the max value to normalize all audios to [0, 1] and then extract the magnitude using the short-time Fourier transform (STFT) with 1024 frequency bins and a 64-ms window size with 16-ms stride. The mel-scale spectrograms are obtained by applying an 80-channel mel filter bank to the magnitude, followed by dynamic range clipping with a minimum value of $1e^{-5}$ and log dynamic range compression.

3) LipSound2 Pre-Training: Horizontal image flipping, gradient clipping with a threshold of 1.0, early stopping, and scheduled sampling [85] are adopted to avoid overfitting. Linear and convolutional layers are initialized with Xavier [86] and tanh functions, respectively. We use the cosine learning rate decay strategy with an initial value of 0.001. Our LipSound2 model has around 100M parameters. The audio and visual sequences are both high-dimensional data, so we conduct all experiments on four NVIDIA Quadro RTX 6000 GPUs with 24-GB memory in parallel to enable a big batch size. The entire pre-training procedure took around 25 days.

4) Fine-Tuning: Pre-trained LipSound2 is fine-tuned on GRID, TCD-TIMIT, and CMLR videos to conduct speech reconstruction experiments. Afterward, the produced speech for English (GRID and TCD-TIMIT) and Chinese (CMLR) is fine-tuned on the pre-trained English (LibriSpeech) and Chinese (AISHELL-2) acoustic models to perform lip reading tasks with a ten times smaller learning rate.

V. EXPERIMENTAL RESULTS

A. Lip-to-Speech Reconstruction

1) Speaker-Dependent Result: We report the generated speech results in two perspectives, i.e., speech quality (PESQ) and speech intelligibility (ESTOI). For a fair comparison, we keep the same settings as previous works. For speaker-dependent tasks, all datasets are randomly split into 90:5:5 for training, validation, and test sets on GRID, respectively (Speaker $S1 - S4$) and TCD-TIMIT (Lipspeaker $1 - 3$). Different from previous works that build one model for each individual speaker, we train only one model on all speakers.

As shown in Table III, our LipSound2 system, which is first pre-trained on the VoxCeleb2 dataset and then fine-tuned on the specific dataset, achieves the highest scores on both PESQ and ESTOI, which reveals the effectiveness of our proposed method. The last column in Table III compares the number of LipSound2 model parameters against those of baseline systems, showing that its best performance is obtained while staying well in the existing range of numbers of parameters.

TABLE IV
SPEAKER-INDEPENDENT SPEECH RECONSTRUCTION RESULTS ON GRID AND TCD-TIMIT DATASETS

	GRID		TCD-TIMIT		
Model	ESTOI	PESQ	ESTOI	PESQ	Parameters
Vougioukas et al. [33]	0.198	1.24	-	-	not available
vid2voc-M-VSR [36]	0.227	1.23	-	-	5.1M
vid2voc-F-VSR [36]	0.210	1.25	-	-	5.2M
LipSound2	**0.363**	**1.72**	**0.30**	**1.31**	8.5M

TABLE V
SPEECH RECONSTRUCTION RESULTS FOR CHINESE ON THE CMLR DATASET

	Speaker-dependent		Speaker-independent	
Model	ESTOI	PESQ	ESTOI	PESQ
LipSound2	**0.36**	**1.43**	**0.28**	**1.21**

2) Speaker-Independent Result: For speaker-independent cases, we follow the same setups for GRID [33] and TCD-TIMIT [31].

LipSound2 achieves the best results on both metrics on the GRID dataset. Moreover, by listening to the reconstructed audios, we find that our model is capable of producing similar voices as ground-truth speakers, instead of generating a weird voice or one of the voices in the training set as occurring in previous works. The model has implicitly learned the mapping between voices and faces. We highly recommend readers to listen to the produced samples on our demo website.[4]

Furthermore, we find substitution errors occurring on segment level (vowels and consonants) because the context information is still not sufficient to disambiguate the phonemes that share the same visible organs, such as lips and tongue, but are different in the invisible ones.

To the best of our knowledge, we are the first to tackle the speaker-independent case on the TCD-TIMIT dataset since TCD-TIMIT consists of limited samples (∼370) for each speaker but with large-scale vocabulary (∼5.9K), which makes the tasks on TCD-TIMIT quite challenging. The speaker-independent results reported in Table IV show considerable performance, for example, the PESQ result is even better than some results reported on speaker-dependent settings (as shown in Table III), which suggests that the large-scale self-supervised pre-training enables the model to successfully generalize to unseen speakers.

3) Speech Reconstruction for Chinese: To explore the effectiveness of our proposed architecture, we further perform speech reconstruction in Chinese. For the speaker-dependent case, we keep the same training and test splits used in CSSMCM [80] for lip reading; for the speaker-independent case, $S1$ (male) and $S6$ (female) are used for testing and the remaining speakers are used for training and validation.

In Table V, only LipSound2 results are reported since we make a first attempt at tackling speech reconstruction in Chinese. After checking the generated audio samples, we find that, besides the confusion on segments, there are some tone errors. One of the reasons is that Chinese is a tonal language

[4]https://leyuanqu.github.io/LipSound2/

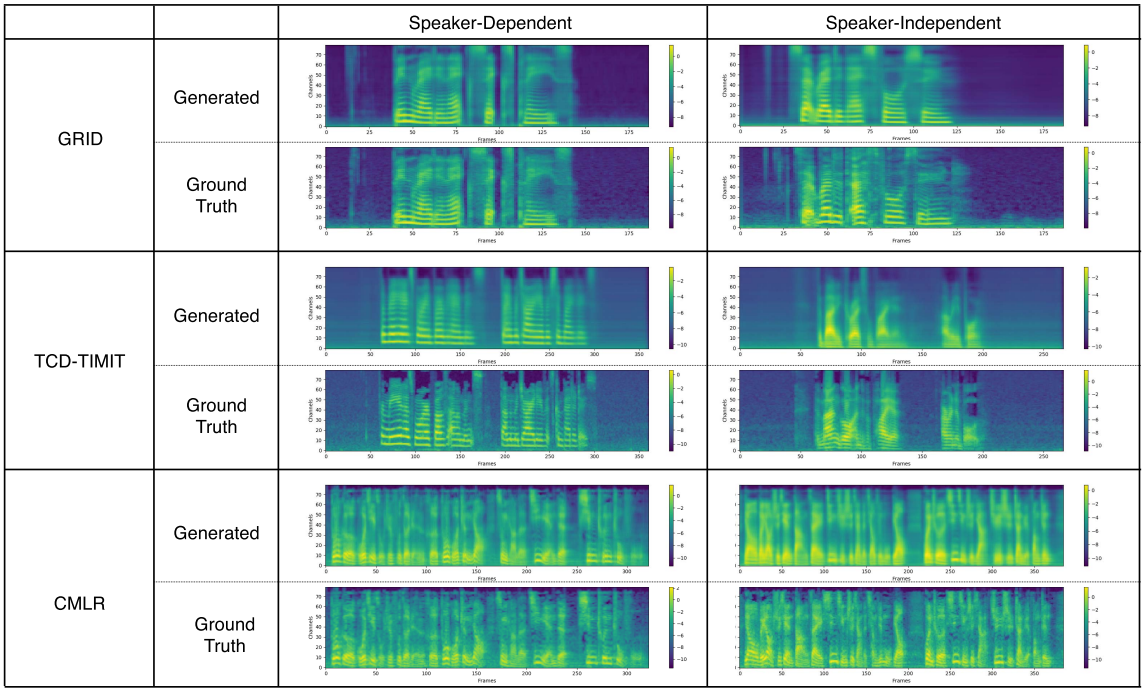

Fig. 5. Comparison between generated mel spectrogram and ground truth in speaker-dependent and speaker-independent settings for English and Chinese [79], [31], [80].

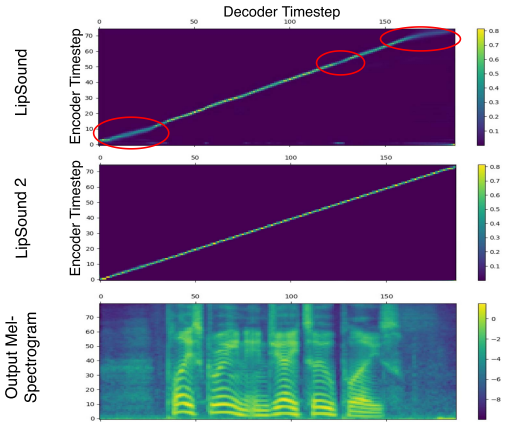

Fig. 6. Attention alignment comparison on the GRID dataset.

in which lexical tones play an important role for semantic discrimination. The fundamental frequency (F0), which is produced by the vibration of vocal cords, is not visible in the input videos (face area), and it is reported that the visual features have a weak correlation to F0 [28]. Another reason is that the VoxCeleb2 dataset mainly consists of nontonal languages, e.g., British English, American English, and German, which makes the pre-training pay little attention to tone production.

4) Attention Alignment: We compare the attention alignments learned by LipSound [20], which is only trained on the GRID dataset and LipSound2 (this article). As shown in Fig. 6, the LipSound attention weights are fuzzy at nonverbal areas and at short pauses between words, which may mislead the decoder into focusing on irrelevant encoder timesteps, whereas the attention weights learned by LipSound2 are intensive and more robust to silence or short pauses.

B. Lip Reading Results

Different from conventional methods which directly transform videos into text, we perform lip reading experiments in two steps, i.e., video-to-wav and wav-to-text.

1) Lip Reading Results for English: We follow the same splits as previous works for training and test on the GRID [44] and TCD-TIMIT [87] datasets. The comparison with related results is listed in Table VI. We report the WER of GRID and TCD-TIMIT audio test sets on pre-trained acoustic models (audio gold standard) and the results fine-tuned on the training audio samples (+Fine-Tuning), which is treated as the upper boundary of lip reading.

Our LipSound2 model achieves the state-of-the-art performance on both GRID and TCD-TIMIT datasets. Fine-tuning the acoustic model pretrained on 960-h LibriSpeech with generated audios can not only significantly boost the model performance but also accelerate training time.

Further improvement can be achieved when an external language model is integrated. The benefit from the language model on the GRID dataset is not as much as on TCD-TIMIT since the sentence structure in GRID is designed by an artificial grammar. The language model can only help to correct misspelled words but cannot contribute grammatically or semantically.

2) Lip Reading Results for Chinese: We also explore lip reading performance in Chinese, as shown in Table VII. Audio gold standard is directly evaluating the CMLR test set on a pre-trained acoustic model trained on a 1000-h AISHELL2 dataset. After fine-tuning with CMLR training audios, we get 3.88% CER and 4.89% CER for speaker-dependent and speaker-independent cases, respectively.

In comparison to other work, our LipSound2 model achieves better results. CER further drops when decoding with an

TABLE VI
LIP READING RESULTS ON THE GRID AND TCD-TIMIT DATASETS ON WER. SPK-DEP: SPEAKER-DEPENDENT. SPK-INDEP: SPEAKER-INDEPENDENT. LM: LANGUAGE MODEL

	GRID		TCD-TIMIT	
Model	Spk-Dep	Spk-Indep	Spk-Dep	Spk-Indep
Audio Gold Standard	22.36	21.88	15.86	15.21
+Fine-tuning	0.15	0.35	5.42	6.73
LipNet [44]	5.6	13.6	-	-
LipNet+LM [44]	4.8	11.4	-	-
PCPG+LM [88]	-	11.2	-	-
TVSR-Net [89]	-	9.1	-	-
WAS [2]	3.0	-	-	-
LCANet[90]	2.9	-	-	-
DualLip [91]	2.7	-	-	-
LipSound [20]	2.5	-	-	-
CD-DNN [87]	-	-	51.26	57.03
MobiLipNetV2 [92]	-	-	-	53.01
LipSound2	1.9	7.3	41.37	46.29
LipSound2 + LM	**1.5**	**6.4**	**39.77**	**43.53**

TABLE VII
LIP READING RESULTS FOR CHINESE ON THE CMLR DATASET. CER: CHARACTER ERROR RATE

Model	Spk-dep	Spk-indep
Audio Gold Standard	19.25	16.2
+Fine-tuning	3.88	4.89
WAS [2]	38.93	-
CSSMCM [80]	32.48	-
LIBS [93]	31.27	-
LipSound2	25.03	36.56
LipSound2 + LM	**22.93**	**33.44**

external language model. Besides, we build a new baseline for CMLR in speaker-independent settings.

VI. DISCUSSION

Although the proposed LipSound2 model pre-trained on a large-scale dataset achieves considerable performance on both speech reconstruction and lip reading tasks, it still generates error speech due to the visual similarity on pronunciation, for example, "pill" is easy to be misrecognized as "bill" in English and "ji zhi" is mistaken as "qi zhi" in Chinese. In addition, our model can generate quite similar voices as the ground truth in speaker-dependent settings, while the model is inclined to predict a voice existing in the training set sometimes in speaker-independent cases. For details and demonstrations, we refer also to the demo video on the project website.[5] How to stop the fine-tuning procedure at the appropriate time and avoid the model overfitting on downstream tasks is an important direction for future research since the MSE loss always declines when using teacher forcing during training, which hardly indicates whether the model is overfitting or not. Besides, a possible solution could be using voice embeddings as additional inputs that can efficiently help models learn speaker identity information, as we found in our previous work [4].

VII. CONCLUSION

In this article, we have proposed LipSound2 that directly predicts speech representations from raw pixels. We investigated the effectiveness of self-supervised pre-training for speech reconstruction on large-scale vocabulary datasets, particularly for speaker-independent settings. Moreover, state-of-the-art results are achieved by fine-tuning the produced audios on a well-pretrained speech recognition model for both English and Chinese lip reading experiments since our two-step method benefits not only from the large-scale crossmodal supervision which enables the model to learn more robust representations and more different content information but also from the advanced speech recognition architecture (acoustic and language models), which is pre-trained on abundant labeled data.

Although we have made great progress on speech reconstruction in controlled environments, there is still a significant gap to the requirements of real-world scenarios. Future work will focus on more realistic configuration, such as the variety of light conditions, moving head poses, and different background environments. Moreover, the current lip reading experiments are separately conducted in two steps in which the error generated in the first step (video-to-wav) will be propagated to the second step (wav-to-text). How to jointly train the two tasks in an end-to-end fashion could be another direction. Besides, we are also interested in integrating our LipSound2 model into active speaker detection, speech enhancement, and speech separation tasks to boost the performance of speech recognition systems in human–robot interaction.

[5]https://leyuanqu.github.io/LipSound2/

ACKNOWLEDGMENT

The authors would like to thank Katja Kösters for improving the language of this article.

REFERENCES

[1] J. Besle, A. Fort, C. Delpuech, and M.-H. Giard, "Bimodal speech: Early suppressive visual effects in human auditory cortex," *Eur. J. Neurosci.*, vol. 20, no. 8, pp. 2225–2234, Oct. 2004.
[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3444–3453.
[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2018, doi: 10.1109/TPAMI.2018.2889052.
[4] L. Qu, C. Weber, and S. Wermter, "Multimodal target speech separation with voice and face references," in *Proc. Interspeech*, Oct. 2020, pp. 1416–1420.
[5] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-visual speech separation using still images," in *Proc. Interspeech*, Oct. 2020, pp. 3481–3485.
[6] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach," in *Proc. Interspeech*, Sep. 2016, pp. 3414–3418.
[7] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5020–5024.
[8] J. Macdonald and H. McGurk, "Visual influences on speech perception processes," *Perception Psychophys.*, vol. 24, no. 3, pp. 253–257, May 1978.
[9] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 337–351, Sep. 1996.
[10] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, 2010.
[11] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10, pp. 2448–2458, Oct. 2010.

[12] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257–267, Feb. 2007.

[13] A. Tsiami, P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Far-field audio-visual scene perception of multi-party human–robot interaction for children and adults," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6568–6572.

[14] R. Tscharn, M. E. Latoschik, D. Löffler, and J. Hurtienne, "'Stop over there': Natural gesture and speech interaction for non-critical spontaneous intervention in autonomous driving," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 91–100.

[15] B. Gick, I. Wilson, and D. Derrick, *Articulatory Phonetics*. Hoboken, NJ, USA: Wiley, 2012.

[16] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*. Dordrecht, The Netherlands: Springer, 1990, pp. 131–149.

[17] S. Goto, K. Onishi, Y. Saito, K. Tachibana, and K. Mori, "Face2Speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image," in *Proc. Interspeech*, Oct. 2020, pp. 1321–1325.

[18] A. Ephrat and S. Peleg, "Vid2Speech: Speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5095–5099.

[19] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2Audspec: Speech reconstruction from silent lip movements video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2516–2520.

[20] L. Qu, C. Weber, and S. Wermter, "LipSound: Neural mel-spectrogram reconstruction for lip reading," in *Proc. Interspeech*, Sep. 2019, pp. 2768–2772.

[21] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, nos. 3–4, pp. 187–207, Apr. 1999.

[22] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[23] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7763–7774.

[24] P. Morgado, Y. Li, and N. Nvasconcelos, "Learning representations from audio-visual spatial alignment," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–12.

[25] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[26] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.

[27] J. Li *et al.*, "Jasper: An end-to-end convolutional neural acoustic model," in *Proc. Interspeech*, 2019, pp. 71–75.

[28] T. L. Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. Interspeech*, Sep. 2015, pp. 1–6.

[29] T. Le Cornu and B. Milner, "Generating intelligible audio speech from visual speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 9, pp. 1751–1761, Sep. 2017.

[30] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 455–462.

[31] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.

[32] Y. Kumar, R. Jain, K. M. Salik, R. R. Shah, Y. Yin, and R. Zimmermann, "Lipper: Synthesizing thy speech using multi-view lipreading," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 2588–2595.

[33] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," in *Proc. Interspeech*, Sep. 2019, pp. 4125–4129.

[34] J. Shen *et al.*, "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4779–4783.

[35] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13796–13805.

[36] D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen, "Vocoder-based speech synthesis from silent videos," in *Proc. Interspeech*, Oct. 2020, pp. 3530–3534.

[37] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.

[38] J. A. Gonzalez *et al.*, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2362–2374, Dec. 2017.

[39] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.

[40] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Proc. 7th Interspeech*, 2002, pp. 1–4.

[41] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. Int. Conf. Image Process.*, 1998, pp. 173–177.

[42] G. Sterpu and N. Harte, "Towards lipreading sentences with active appearance models," 2018, *arXiv:1805.11688*.

[43] D. Parekh, A. Gupta, S. Chhatpar, A. Y. Kumar, and M. Kulkarni, "Lip reading using convolutional auto encoders as feature extractor," 2018, *arXiv:1805.12371*.

[44] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence-level lipreading," in *Proc. GPU Technol. Conf.*, 2017, pp. 1–13. [Online]. Available: https://github.com/Fengdalu/LipNet-PyTorch

[45] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6115–6119.

[46] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Proc. Interspeech*, Aug. 2017, pp. 3652–3656.

[47] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–103.

[48] S. Yang *et al.*, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.

[49] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading," 2019, *arXiv:1905.02540*.

[50] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6319–6323.

[51] K. Thangthai and R. Harvey, "Improving computer lipreading via DNN sequence discriminative training techniques," in *Proc. Interspeech*, Aug. 2017, pp. 1–5.

[52] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *Proc. Interspeech*, 2017, pp. 2415–2419.

[53] B. Shillingford *et al.*, "Large-scale visual speech recognition," in *Proc. Interspeech*, 2018, pp. 4135–4139.

[54] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: A comparison of models and an online application," in *Proc. Interspeech*, Sep. 2018, pp. 3514–3518.

[55] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. ICLR*, 2018, pp. 1–16.

[56] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[57] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 69–84.

[58] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 649–666.

[59] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2794–2802.

[60] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 527–544.

[61] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 391–408.
[62] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[63] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI Blog, 2018. Accessed: Aug. 6, 2020. [Online]. Available: https://openai.com/blog/language-unsupervised
[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[65] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.
[66] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
[67] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 609–617.
[68] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3965–3969.
[69] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 435–451.
[70] B. Chen *et al.*, "Multimodal clustering networks for self-supervised learning from unlabeled videos," 2021, *arXiv:2104.12671*.
[71] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[72] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 577–585.
[73] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10215–10224.
[74] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop (SSW)*, 2016, p. 125.
[75] K. Ito and L. Johnson. (2017). *The LJ Speech Dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/
[76] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proc. 6th Workshop Stat. Mach. Transl.*, 2011, pp. 187–197.
[77] S. Wermter and V. Weber, "SCREEN: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks," *J. Artif. Intell. Res.*, vol. 6, pp. 35–85, Jan. 1997.
[78] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
[79] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustic Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.
[80] Y. Zhao, R. Xu, and M. Song, "A cascade sequence-to-sequence model for Chinese Mandarin lipreading," in *Proc. ACM Multimedia Asia*, 2019, pp. 1–6.
[81] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[82] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming Mandarin ASR research into industrial scale," 2018, *arXiv:1808.10583*.
[83] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, May 2001, pp. 749–752.
[84] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[85] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1171–1179.
[86] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[87] K. Thangthai, H. L. Bear, and R. Harvey, "Comparing phonemes and visemes with DNN-based lipreading," 2018, *arXiv:1805.02924*.
[88] M. Luo, S. Yang, S. Shan, and X. Chen, "Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading," 2020, *arXiv:2003.03983*.
[89] C. Yang, S. Wang, X. Zhang, and Y. Zhu, "Speaker-independent lipreading with limited data," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2181–2185.
[90] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 548–555.
[91] W. Chen, X. Tan, Y. Xia, T. Qin, Y. Wang, and T.-Y. Liu, "DualLip: A system for joint lip reading and generation," 2020, *arXiv:2009.05784*.
[92] A. Koumparoulis and G. Potamianos, "MobiLipNet: Resource-efficient deep learning based lipreading," in *Proc. Interspeech*, Sep. 2019, pp. 2763–2767.
[93] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song, "Hearing lips: Improving lip reading by distilling speech recognizers," in *Proc. AAAI*, 2020, pp. 6917–6924.

Leyuan Qu received the M.Sc. degree in computer science from Beijing Language and Culture University, Beijing, China, in 2017, and the Ph.D. degree from the Department of Informatics, University of Hamburg, Hamburg, Germany, in 2021.

His main research interests include robust speech recognition, audio-visual speech recognition, speech enhancement, speech separation, lip reading, and self-supervised learning.

Cornelius Weber received the Diploma degree in physics from the University of Bielefeld, Bielefeld, Germany, in 1995, and the Ph.D. degree in computer science from the Technische Universität Berlin, Berlin, Germany, in 2000.

He was a Post-Doctoral Fellow of brain and cognitive sciences with the University of Rochester, Rochester, NY, USA. From 2002 to 2005, he was a Research Scientist of hybrid intelligent systems with the University of Sunderland, Sunderland, U.K. He was a Junior Fellow with the Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany, until 2010. He is currently a Laboratory Manager with the Knowledge Technology Group, University of Hamburg, Hamburg, Germany. His current research interests include computational neuroscience with a focus on vision, unsupervised learning, and reinforcement learning.

Stefan Wermter (Member, IEEE) is currently a Full Professor with the University of Hamburg, Hamburg, Germany, where he is also the Director of the Department of Informatics, Knowledge Technology Institute. Currently, he is a co-coordinator of the International Collaborative Research Centre on Crossmodal Learning (TRR-169) and a coordinator of the European Training Network TRAIL on transparent interpretable robots. His main research interests are in the fields of neural networks, hybrid knowledge technology, cognitive robotics, and human–robot interaction.

Prof. Wermter has been an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He is an Associate Editor of *Connection Science* and *International Journal for Hybrid Intelligent Systems*. He is on the Editorial Board of the journals *Cognitive Systems Research*, *Cognitive Computation*, and *Journal of Computational Intelligence*. He is serving as the President for the European Neural Network Society.