

MoDiCoL: A Modular Diagnostic Continual Learning Dataset for Robust Speech Recognition

Theresa Pekarek Rosin¹, Matthias Kerzel¹, Stefan Wermter¹

Knowledge Technology, Department of Informatics, University of Hamburg, Germany

{theresa.pekarek-rosin, matthias.kerzel, stefan.wermter}@uni-hamburg.de

Abstract

Modern Automatic Speech Recognition (ASR) systems have made remarkable progress on standard benchmarks, yet performance gaps have emerged under real-world distribution shifts, caused by recording conditions, accents, speech impairments, and noise. Existing datasets and benchmarks typically isolate these factors, which overlooks their co-occurrence in real-world applications. In this paper, we argue that model robustness can be treated as a dynamic capability that continually develops, and we introduce MoDiCoL, a Modular Diagnostic Continual Learning dataset designed for controlled analysis of linguistic content, speaker characteristics, and acoustic environments. Furthermore, we propose a real-world-inspired continual learning curriculum to simulate incremental updates and study how robustness is acquired, transferred, and forgotten. We evaluate three continual learning strategies and provide detailed insights into robustness under evolving conditions.

Index Terms: speech recognition, continual learning, audio dataset, speech resources

1. Introduction

Modern Automatic Speech Recognition (ASR) models have improved significantly on standard benchmarks over the last few years [1], yet they still struggle with speech that differs from the standard, fixed setups. Challenges include impaired or disfluent speech [2], accents or dialects [3], varying noise or environments [4], and elderly or children [5] speakers. Even after the utilization of large-scale datasets for pretraining, these performance gaps still persist [6].

Existing benchmarks or datasets are typically built for specific factors, e.g., a model's robustness to noise [4, 7], accents [8], or impairment [9, 10], and evaluated separately. This does not reflect real-world conditions, where such distribution shifts accumulate over time. Even datasets with rich metadata, built for evaluation, such as Casual Conversations v2 [11], remain limited by uncontrollable factors, such as recording conditions and speaker selection, making it harder to isolate the factor with the greatest impact. Additionally, since real-world applications often involve accumulating changes, we argue that robustness could be studied as a continual capability that evolves across sequential distribution shifts.

Continual learning (CL) is a sequential learning paradigm that preserves previously acquired knowledge [12, 13]. Therefore, CL enables a simulation of real-world incremental model updates and an examination of how robustness is acquired during training and whether it can be transferred across different factor drifts. While CL has been applied to speech for domain adaptation, such as accent [14] and language adaptation [15], its potential as a diagnostic method to identify where forgetting

occurs in pretrained ASR models has not yet been explored in the same way as it has been, for example, for Vision-Language models [16].

To address these identified knowledge gaps, we introduce **MoDiCoL**¹, a **Modular Diagnostic Continual Learning** dataset, in which we offer a controlled evaluation of different factors that influence ASR performance: linguistic content, speaker characteristics, and acoustic environment. We implement a real-world-inspired CL curriculum for the dataset, using three different CL strategies to analyze the causes and effects of distribution drift across these factors and mitigate the resulting robustness degradation.

2. MoDiCoL: A modular diagnostic continual learning dataset

MoDiCoL is a speech dataset designed to study the robustness of ASR models to different drift factors in a controlled, continual setting. We construct MoDiCoL using a systematic factorial design that enables a rigorous evaluation of linguistic, speaker, and acoustic variation with clearly defined experimental runs. By combining real-world and synthetic speech with a configuration-dependent augmentation pipeline, MoDiCoL allows a targeted evaluation of how different types of factorial drift affect model adaptation and forgetting.

2.1. Factors

As shown in Table 1, MoDiCoL covers three factor categories: Linguistic Content, Speaker Characteristics, and Acoustic Environment. Each of these categories describes a set of factors expressed at different levels.

Linguistic content refers to the language-related characteristics of the dataset. This includes the vocabulary domain, which specifies the subject area (medical, air traffic control, or other domains), and the speech style, which captures how speech is produced (read, spontaneous, or conversational).

Speaker characteristics describe demographic and speech-production factors of the respective speakers in the dataset. These include age group (child, adult, elderly), accent (English, South-East Asian, or other), and health (healthy or impaired). Additionally, we include pauses and disfluencies as separate factors, since they also occur in healthy speech.

Acoustic environment describes the recording conditions under which the speech data was captured. It includes the noise type (clean, babble, or fan noise), the signal-to-noise ratio (SNR) level (clean, 10dB, or 20dB), and the distance between the speaker and the microphone.

¹<https://huggingface.co/datasets/TPekarekRosin/modicol>

Table 1: *The factors covered in MoDiCoL and their levels. Linguistic content describes the factors that account for variation in speaking patterns and vocabulary within the recording context. Speaker characteristic factors cover the speaker-dependent variations, and acoustic environment factors describe the recording environment.*

LINGUISTIC CONTENT	
Vocabulary Domain	Medical, Air Traffic Control, Other
Speech Style	Read, Spontaneous, Conversational
SPEAKER CHARACTERISTICS	
Age	Child, Adult, Elderly
Accent	English, South-East Asian, Other
Health	Healthy, Impaired
Pauses	Yes, No
Disfluencies	Yes, No
ACOUSTIC ENVIRONMENT	
Noise Type	Clean, Babble, Fan
SNR Level	Clean, 10dB, 20dB
Distance	Close, Far

2.2. Structure

We create the dataset with an L27 orthogonal array (OA) following the Taguchi design method [17] and two foldover dimensions F_0 and F_1 to accommodate the six three-dimensional factors (vocabulary domain, speech style, age, accent, noise type, and SNR level) and the four two-dimensional factors (health, pauses, disfluencies, and distance).

The foldover dimensions create four slices, which we use to perform cyclic permutation, yielding 108 (27×4) distinct run configurations in total. We fill each of these runs with 75 samples, a standard heuristic for mixed-level fractional factorial designs [18], resulting in a dataset of **8100** samples overall. Within these runs, some factor-level combinations require special handling:

1. medical (vocabulary domain), elderly/child (age)
2. ATC (vocabulary domain), elderly/child (age)
3. medical/ATC (vocabulary domain), impaired (health)
4. clean (SNR level) and babble/fan (noise type)

We differentiate between combinations unlikely to occur in real-world datasets, but that can still provide insight into model behavior (1–3) and truly infeasible combinations that violate logical constraints (4). The first type is accommodated using synthetic speech generation and voice cloning, while we repair the experimental design by changing the noise type to “clean” for the second type.

2.3. Dataset compilation

The OA of the dataset contains many combinations that lack available real-world data (e.g., impaired speech with non-English accents). We therefore populate as many runs as possible with open-source speech datasets and use text datasets to generate synthetic speech for missing combinations.

Healthy children’s speech is sourced from Non-native Children English Speech (NNCES) [19], Children Speech Recordings [20], and Ultrax Typically Developing Children (UXTD) [21, 22], while impaired children’s speech comes from the Cleft dataset [10]. Healthy adult and elderly speech is taken from Svarah [8], Trustworthy Intent in Speech (TIS) [23], TED-

LIUMv2 [24], and Common Voice [25], with DailyDialog [26] providing conversational data.

For domain-specific speech, ATC data is obtained from the ATCO-2 1-hour test split [27] and the UWB-ATCC dataset [28], which we also use as transcripts for the synthetic speech generation alongside related texts from Wikipedia. Medical speech is collected from United-Syn-Med [29] and the Eka Medical ASR Evaluation Dataset [30], supplemented with medical texts from MedRAG [31], Medical Transcriptions [32], MedDialog [33], NoteChat [34], and simulated medical exams [35].

The synthetic speech is generated using XTTS-v2 [36], a multilingual zero-shot text-to-speech (TTS) model that uses reference audio to match speaker characteristics. We collect such reference files across age, accent, f_0 , and health conditions from datasets including the Perceptual Voice Quality Database [37], the Voice Cloning Toolkit (VCTK) [38], Cleft [10], Common Voice 11.0 [25], Children Speech Recordings [20], NNCES [19], Ultrax Speech Sound Disorders (UXSSD) [21], and TORGO [9].

Overall, MoDiCoL contains 18.79 h of speech, of which 14.08 h are synthetic, with an average sample length of 8.35 s.

2.4. Augmentation pipeline

We resample all files to 16 kHz and save them as wave files. To control for background noise and other aspects of the acoustic environment, as well as disfluencies and pauses, we augment both real and generated speech samples using the following pipeline, as specified in the respective run configuration.

1. **Denoising.** We first suppress background noise for all samples using a DNN-HMM hybrid system [39], initialized from a publicly available checkpoint on Hugging Face².
2. **Disfluency Insertion.** If no natural disfluencies are present, we synthesize filler words using the XTTSv2 model, conditioning on the original audio to preserve speaker identity. Filler words are inserted using the same procedure as pause insertion (see Step 4).
3. **Impairment Simulation.** For speakers with South-East Asian accents and impaired speech, suitable open-source reference recordings were unavailable. In these cases, we either use the healthy reference voices for TTS or start from healthy real speech and simulate impairment. Impairment is introduced via prosodic and spectral modifications, including tremor, jitter, shimmer, and temporal perturbations.
4. **Pause Insertion/Removal.** If the file contains no pauses, they are inserted by detecting existing silent segments (e.g., between words) or, if none are available, by randomly selecting insertion points. If pauses are disabled in the configuration, noticeable silent segments are removed instead.
5. **Distance Simulation.** To simulate increased distance between speaker and microphone, we add reverberation, effectively modeling a larger room and a longer acoustic path.
6. **Noise Injection.** Background noise is added according to the predefined signal-to-noise ratio (SNR) and noise type. For babble and fan noise, we sample from the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [7], which provides dedicated noise recordings for each type. For clean conditions, synthetic random noise is generated. The selected noise is mixed with the speech signal at the specified SNR level.

²<https://huggingface.co/speechbrain/mtl-mimic-voicebank>

3. Methodology

3.1. Continual learning curriculum

Continual learning (CL) can be defined as follows: $\mathcal{T}$ is the set of tasks that is learned by the model in sequence $\mathcal{T} := t_1, t_2, \dots, t_N$, where $\mathcal{D} = \mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_N$ is the corresponding stream of data with unknown distributions over an input-output space $\mathcal{X} \times \mathcal{Y}$ [13, 40]. The model receives a batch of samples at each step, belonging to the current task t_i , from data $\mathcal{D}_t$ with its distinct distribution.

We sample from LibriSpeech [41] to define a control setting t_0 against which all drift types are compared: adult, healthy, read, no pauses, no disfluencies, clean, close, other vocabulary, English accent. We then define a CL curriculum of disjoint run sets that represent different types of data drift:

- t_1 : **Acoustic Environment Drift**. Introduction of environmental variation through babble and fan noises, varied SNR levels, and microphone distances.
- t_2 : **Speaker Characteristics Drift**. Inclusion of new speaker groups, i.e., children, the elderly, and impaired speakers with different accents. Acoustic conditions are kept moderate.
- t_3 : **Linguistic Content Drift**. Exposure to new domains and speaking styles, i.e., spontaneous and conversational speech, and specialized vocabulary (medical, air traffic control).
- t_4 : **Compound Drift**. Combination of multiple drift types: elderly or impaired speakers at a far distance in noisy environments, with conversational or domain-specific speech.

To study robustness transfer across drift types, we evaluate whether robustness acquired on one drift generalizes to others and whether curriculum order influences model performance. We generate two additional permutations of the displayed curriculum for the evaluation while keeping task t_4 fixed.

3.2. Experimental setup

In our experimental setup, we employ three continual learning (CL) methods to analyze different aspects of sequential drift: rehearsal-based stabilization with Experience Replay [42], representation preservation with Representation-level Regularization [43], and gradient subspace isolation with Orthogonal Gradient Descent [44]. As the backbone model, we use a pretrained `whisper-small.en` model [1].

Experience Replay (ER) [42] is a rehearsal-based approach that maintains a memory buffer containing a fraction of data from previous tasks. During fine-tuning on a new task, buffered samples are interleaved with the current training data. We evaluate small (5%) and medium (10%) buffer sizes, with an equal number of samples per past task, to assess whether stability regarding drifts depends on rehearsal mechanisms.

For **Representation-level Regularization (RLR)** [43], we introduce a regularization term that constrains the encoder to preserve previously learned representations. We retain a frozen copy of the encoder after each task and minimize the cosine similarity loss between the mean-pooled encoder outputs of the current and the frozen models. Using RLR, we assess how the curriculum affects the model’s acoustic representation and whether forgetting arises from representational drift.

Orthogonal Gradient Descent (OGD) [44] mitigates forgetting by constraining parameter updates to directions that do not interfere with prior tasks. We apply OGD to the encoder and compute an average gradient over a small subset of the task data, then normalize it to obtain the task-specific direction. During training, gradients are projected onto the orthogonal com-

Table 2: The WER and BERTScore (F1) results of evaluating MoDiCoL on `whisper-small.en` **before** the curriculum training.

SET	A-WER (↓)		F1 (↑)	
	MEAN	MEDIAN	MEAN	MEDIAN
t_0	7.42	0.0	98.28	99.90
t_1	47.62	14.29	95.80	97.03
t_2	87.28	28.57	92.42	94.38
t_3	141.73	42.86	89.54	91.35
t_4	43.37	20.00	94.28	95.63
REAL	86.32	25.00	92.43	94.55
SYNTHETIC	69.97	13.33	95.02	96.90
ALL	75.04	16.67	94.22	96.37

plement of all stored directions to prevent interference through parameter updates. We measure task similarity by computing the cosine similarity between current gradients and stored directions to quantify geometric overlap between gradient subspaces and entanglement between drift factors.

Performance across tasks is measured using Average Word Error Rate (A-WER) and Average Incremental WER (AI-WER). Memory stability is assessed via Forgetting Measure (FM) and Backward Transfer (BWT), while learning plasticity is quantified using Forward Transfer (FWT) and Intransigence Measure (IM), following Wang et al. [13]. Since we use an error rate as the evaluation metric, the interpretation of some metrics changes, which we indicate accordingly.

To approximate real-world conditions for model adaptation, we perform these evaluations in an **online** and **streaming** CL setting, where the batch size is limited to one, and the model sees every sample only once [40]. We split each drift into train (70%), validation (20%), and test (10%) sets and train with a learning rate of $1e-5$. And to quantify the impact of distribution drift, we include a sequential **fine-tuning** (FT) lower baseline without any CL mechanism, which captures task interference, and a **joint training** upper baseline, in which all drift splits are trained simultaneously.

4. Results

Before the CL curriculum, we first evaluate the dataset’s overall quality by computing WER and F1 scores with BERTScore [45] as a semantic similarity metric for `whisper-small.en`. All metrics are reported on the normalized transcripts.

As Table 2 shows, the unadapted model performs best on the control setting t_0 , where it achieves low WER and a high F1 score, but seems to deteriorate substantially under the distribution shifts. For acoustic drift (t_1), the performance drops moderately, but degrades even more severely for speaker (t_2) and linguistic drift (t_3). Here, the WER increases dramatically, indicating a strong sensitivity to changes in speaker characteristics and language content. Interestingly, compound drift (t_4), despite combining all previous drift factors, does not produce the worst performance, suggesting that drift effects do not accumulate in difficulty when applied together.

Overall, there appears to be a large difference between the mean-averaged WER and the median WER of the unadapted model. This indicates that the model generally transcribes the audio with a reasonable error rate (16.67%), but that a few severe outliers are distorting the mean. By examining the transcripts, we see that hallucinations are a major contributor to the

Table 3: The results of the CL curriculum, averaged over the three permutations ($\pm$ standard deviation). Since WER is used as a performance measure, some metric interpretations are reversed, e.g., positive BWT scores indicate forgetting. We mark the optimal direction next to the name. JOINT denotes the joint training baseline, and FT the sequential fine-tuning baseline.

	A-WER ($\downarrow$)	AI-WER ($\downarrow$)	FM (target = 0)	BWT ($\downarrow$)	FWT ($\downarrow$)	IM ($\uparrow$)
ER-5%	25.75 $\pm$ 11.32	23.40 $\pm$ 1.42	-12.89 $\pm$ 14.99	12.89 $\pm$ 14.99	-19.51 $\pm$ 9.04	26.52 $\pm$ 0.10
ER-10%	17.31 $\pm$ 0.48	22.83 $\pm$ 3.84	-1.95 $\pm$ 1.20	1.78 $\pm$ 0.93	-31.31 $\pm$ 0.51	26.64 $\pm$ 0.22
RLR	34.28 $\pm$ 8.80	24.30 $\pm$ 5.31	-22.34 $\pm$ 13.32	22.33 $\pm$ 13.32	-20.95 $\pm$ 8.78	25.09 $\pm$ 1.82
OGD	26.87 $\pm$ 9.84	21.19 $\pm$ 1.18	-12.75 $\pm$ 14.30	12.16 $\pm$ 14.80	-23.81 $\pm$ 10.25	24.87 $\pm$ 1.26
JOINT	27.24 $\pm$ 2.25	-	-	-	-	-
FT	34.14 $\pm$ 8.56	23.73 $\pm$ 3.35	-24.55 $\pm$ 10.99	24.55 $\pm$ 10.99	-27.85 $\pm$ 4.25	26.88 $\pm$ 0.33

high WER, but since the semantic similarity scores (F1) remain high throughout, they do not seem to affect the semantic alignment between the model output and the ground truth. Additionally, synthetic speech generally performs better than real-world speech, likely due to the greater variability in the real-world subset (speakers, impairments, etc.).

The results of the CL curriculum (Table 3) show a clear benefit of utilizing continual learning for the introduction of different drift types. ER with a replay buffer size of 10% is the most balanced approach, with an A-WER of 17.31 ± 0.48 , which outperforms not only the other CL methods but also the joint training baseline. An FM score close to zero and a comparatively low BWT score suggest minimal forgetting during training and, consequently, strong retention of robustness across drift types. The low negative FWT and high IM score also indicate that exposure to previous drifts overall benefits the learning of new tasks. ER with a replay buffer size of 5% still reduces forgetting compared to the FT baseline and outperforms the joint baseline on A-WER, but overall shows greater instability, as evidenced by the large variance, and therefore insufficient replay capacity to stabilize learning across different permutations.

On the other hand, RLR exhibits large forgetting (-22.34 ± 13.32) and BWT scores, while OGD outperforms RLR across most metrics and achieves the best AI-WER score (21.19 ± 1.18) overall. OGD enforces orthogonal gradient updates, which we ensure by computing the cosine similarities between the stored task gradients, which are all close to zero, ranging from 10^{-3} to 10^{-6} . This indicates that task gradients lie in distinct subspaces for OGD and suggests that forgetting arises from destructive updates to these gradients rather than from representational drift. However, because we use mean pooling for RLR, which enforces high-level or global similarity, some performance-critical local structures might be missed.

If we compare the WER progression during the CL curriculum in Figure 1, we can see that t_0 remains stable and even profits from the sequential training across all methods. For the compound drift t_4 , we can observe an interesting phenomenon: While the training of t_4 mostly hurts existing knowledge, especially for t_2 and t_3 , for t_1 , the learning of compound drift factors can be beneficial and even restore performance. Also noteworthy is that, for ER-10%, overall performance recovers.

Overall, we find that robustness across drift types and permutations benefits most from moderately sized memory and replay mechanisms. The high IM scores and large negative FWT scores across all methods, including the FT baseline, indicate that drift types benefit from sequential updates, which is also supported through Figure 1. However, the large variance in FM and BWT scores across all methods except ER-10% indicates that the task order must be carefully chosen.

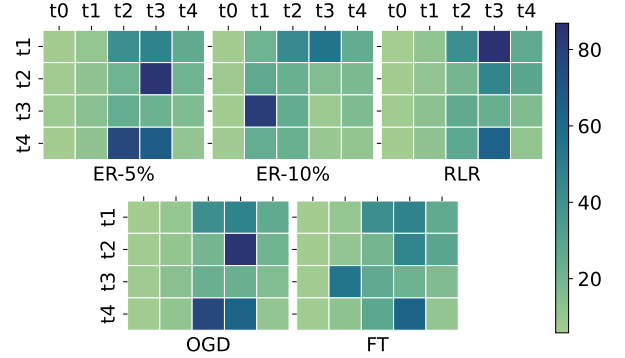


Figure 1: The progression of WER across different methods. Rows correspond to training steps in the curriculum, and columns show the development of per-task performance.

5. Conclusion

We introduced MoDiCoL, a Modular Diagnostic Continual Learning dataset for studying ASR robustness to distributional shifts. Built with a factorial design, MoDiCoL evaluates the impact of linguistic content, speaker characteristics, and acoustic environment drift on performance. We created distinct run configurations with real and synthetic speech, augmented to match defined factor levels, to create a continual learning (CL) curriculum across different drift types. To evaluate adaptation, forgetting, and interference in the model, we applied three CL methods: Experience Replay (ER), Representation-level Regularization (RLR), and Orthogonal Gradient Descent (OGD).

ER with a buffer of 10% of previous task data stabilized training on new tasks and even outperformed the joint training baseline, indicating that replay is beneficial for preserving robustness across drifts. Using RLR and OGD, we examined whether forgetting was due to interfering parameter updates or representational drift. Since OGD outperformed RLR, which exhibited substantial forgetting, we conclude that interference within gradient subspaces is one factor contributing to performance loss. We found that a sequential introduction of drifts improved the model’s learning plasticity, whereas memory stability was highly dependent on task order and the CL method.

Overall, we showed that CL can not only preserve model robustness to distributional drift but also serve as a diagnostic tool for investigating forgetting mechanisms in pretrained models. We make MoDiCoL, its augmentation pipeline, and the CL curriculum available on Hugging Face. In the future, we plan to expand it with additional runs to include more variations of real-world data, enabling further analysis of model behavior.

6. Acknowledgments

The authors gratefully acknowledge funding from Horizon Europe under the MSCA grant agreements No 101072488 (TRAIL), No 101226624 (GREET), and No 101168792 (SWEET).

7. Generative AI Use Disclosure

Generative AI tools have been used to polish the language in the manuscript.

8. References

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, Honolulu, Hawaii, USA, 2023, pp. 28 492–28 518.
- [2] D. Mujtaba, N. Mahapatra, M. Arney, J. Yaruss, H. Gerlach-Houck, C. Herring, and J. Bin, "Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 4795–4809.
- [3] A. S. Dhanjal and W. Singh, "A comprehensive survey on automatic speech recognition using neural networks," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23 367–23 412, Aug. 2023.
- [4] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi, I. Gurvich, S. Peer, X. Xiao, B. M. Elizalde, N. Kanda, X. Wang, S. Shaer, S. Yagev, Y. Asher, S. Sivasankaran, Y. Gong, M. Tang, H. Wang, and E. Krupka, "Notsofar-1 challenge: New datasets, baseline, and tasks for distant meeting transcription," in *Proceedings of Interspeech 2024*. ISCA, Sep. 2024, pp. 5003–5007.
- [5] S. Feng, B. M. Halpern, O. Kudina *et al.*, "Towards inclusive automatic speech recognition," *Computer Speech and Language*, vol. 84, no. C, p. 101567, 2024.
- [6] H. Hwang, E. Jordan, D.-H. Kim-Dufor, C. Lemey, and M. Al-rahabi, "Evaluating ASR in a clinical context : What whisper misses," in *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*. Southern Denmark University, Odense, Denmark: Association for Computational Linguistics, Aug. 2025, pp. 374–378.
- [7] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proceedings of Interspeech 2019*, pp. 1816–1820, 2019.
- [8] T. Javed, S. Joshi, V. Nagarajan, S. Sundaresan, J. Nawale, A. Raman, K. S. Bhogale, P. Kumar, and M. M. Khapra, "Svarah: Evaluating english ASR systems on indian accents," in *Proceedings of Interspeech 2023*. ISCA, 2023, pp. 5087–5091.
- [9] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, Dec. 2012.
- [10] J. Cleland, S. Lloyd, L. Campbell, L. Crampin, J.-P. Palo, E. Sugden, A. Wrench, and N. Zharkova, "The impact of real-time articulatory information on phonetic transcription: Ultrasound-aided transcription in cleft lip and palate speech," *Folia Phoniatrica et Logopaedica*, vol. 72, pp. 120–130, 2019.
- [11] B. Porgali, V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas, "The casual conversations v2 dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 10–17.
- [12] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, May 2019.
- [13] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362–5383, 2024.
- [14] S. Vander Eeck and H. Van hamme, "Rehearsal-Free Online Continual Learning for Automatic Speech Recognition," in *Proceedings of Interspeech 2023*, 2023, pp. 944–948.
- [15] L. Della Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli, "CI-masr: A continual learning benchmark for multilingual asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4931–4944, 2024.
- [16] K. Ahrens, L. Bengtson, J. H. Lee, and S. Wermter, "Visually grounded continual language learning with selective specialization," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Dec. 2023, pp. 7037–7054.
- [17] G. Taguchi, G. Taguchi, and Y. Yokoyama, *Taguchi Methods: Design of Experiments*, ser. Quality engineering series. ASI Press, 1993.
- [18] D. Montgomery, *Design and Analysis of Experiments*, ser. Design and Analysis of Experiments. John Wiley & Sons, 2008.
- [19] K. Radha, M. Bansal, and R. B. Pachori, "Automatic speaker and age identification of children from raw speech using sincnet over erb scale," *Speech Communication*, vol. 159, no. C, Apr. 2024.
- [20] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Children speech recording (english, spontaneous speech + pre-defined sentences)," Dec. 2016.
- [21] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. M. Scobbie, and A. Wrench, "Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions," in *Proceedings of Interspeech 2018*, 2018, pp. 1888–1892.
- [22] M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals, "Ultrasound tongue imaging for diarization and alignment of child speech therapy sessions," in *Proceedings of Interspeech 2019*, 2019, pp. 16–20.
- [23] C. Maltezos-Papastilianou, R. Scherer, and S. Paulmann, "Human voices communicating trustworthy intent: A demographically diverse speech audio dataset," *Scientific Data*, vol. 12, no. 1, p. 921, May 2025.
- [24] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," in *Proceedings of LREC 2014*, Reykjavik, Iceland, 2014, pp. 3935–3939.
- [25] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: ELRA, 2020, pp. 4218–4222.
- [26] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995.
- [27] J. P. Z. Gomez, K. Vesely, I. Szöke, A. Blatt, P. Motlicek, M. Kocour, K. Choukri, I. Nigmatulina, C. Cevenini, A. Tart, J. Cernocký, and D. Klakow, "ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications," *Journal of Data-centric Machine Learning Research*, 2024.
- [28] L. Šmídl, "Air traffic control communication," 2011, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

- [29] United-We-Care, “United-medsyn: Medical speech dataset for asr,” 2024.
- [30] A. Rana, R. Badami, A. Kumar, N. R. Badge, K. Singh, R. Srivathsav, A. Kumar, and S. SN, “Eka medical asr evaluation dataset,” 2025.
- [31] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, “Benchmarking retrieval-augmented generation for medicine,” in *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 6233–6251.
- [32] MTSamples.com, “Medical transcriptions,” 2019.
- [33] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen, and P. Xie, “MedDialog: Large-scale medical dialogue datasets,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9241–9250.
- [34] J. Wang, Z. Yao, Z. Yang, H. Zhou, R. Li, X. Wang, Y. Xu, and H. Yu, “NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes,” in *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 183–15 201.
- [35] F. Fareez, T. Parikh, C. Wavell, S. Shahab, M. Chevalier, S. Good, I. De Blasi, R. Rhouma, C. McMahon, J.-P. Lam, T. Lo, and C. W. Smith, “A dataset of simulated patient-physician medical interviews with a focus on respiratory cases,” *Scientific Data*, vol. 9, no. 1, p. 313, Jun. 2022.
- [36] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, “Xtts: a massively multilingual zero-shot text-to-speech model,” in *Proceedings of Interspeech 2024*, 2024, pp. 4978–4982.
- [37] P. R. Walden, “Perceptual voice qualities database (PVQD),” 2022.
- [38] J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vtk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2019.
- [39] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, “Spectral feature mapping with mimic loss for robust speech recognition,” in *IEEE Conference on Audio, Speech, and Signal Processing (ICASSP)*, 2018.
- [40] G. M. van de Ven, N. Soures, and D. Kudithipudi, “Continual learning and catastrophic forgetting,” in *Learning and Memory: A Comprehensive Reference (Third Edition)*, third edition ed. Oxford: Academic Press, 2025, pp. 153–168.
- [41] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [42] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, “Experience replay for continual learning,” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.
- [43] J. Pomponi, S. Scardapane, V. Lomonaco, and A. Uncini, “Efficient continual learning in neural networks with embedding regularization,” *Neurocomputing*, vol. 397, pp. 139–148, 2020.
- [44] M. Farajtabar, N. Azizan, A. Mott, and A. Li, “Orthogonal gradient descent for continual learning,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 108. PMLR, 26–28 Aug 2020, pp. 3762–3773.
- [45] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.