LOW-BURDEN DATA AUGMENTATION FOR DYSARTHRIC ASR VIA ZERO-SHOT VOICE CLONING

Satwinder Singh[⋆], *Qianli Wang*[⋆], *Zihan Zhong*[⋆], *Clarion Mendes*[†], *Mark Hasegawa-Johnson*[†], *Waleed Abdulla*[⋆], *Seyed Reza Shahamiri*[⋆]

[⋆]University of Auckland, Auckland, New Zealand
[†]University of Illinois Urbana-Champaign, USA

ABSTRACT

Automatic Speech Recognition (ASR) systems struggle with dysarthric speech due to high inter- and intra-speaker variability and severe data scarcity. Collecting real dysarthric speech is labor-intensive and expensive. In this work, we investigate zero-shot voice cloning as a scalable way to synthesize dysarthric speech for ASR training. We utilized Higgs Audio V2 to clone the voices of dysarthric speakers from the TORGO dataset, using a single reference utterance per speaker. We synthesize linguistically diverse prompts drawn from the Speech Accessibility Project (SAP) to preserve speaker-specific pathology while expanding lexical coverage. We fine-tune Whisper-medium on the cloned data and evaluate on real TORGO test utterances. Our results show a 14.7% relative word error rate (WER) reduction (32.96% → 28.12%) compared to the baseline. We also demonstrate that by adding only 1.55 hours of real dysarthric speech, the model achieves 57.59% and 50.28% relative WER reduction compared to the baseline and clone-only model, respectively. Overall, cloned dysarthric speech is a viable augmentation strategy that reduces dependence on costly data collection; its benefits are maximized when paired with even modest amounts of real dysarthric speech.

Index Terms— dysarthria, whisper, automatic speech recognition, zero-shot voice cloning

1. INTRODUCTION

Automatic speech recognition (ASR) has achieved remarkable performance on typical speech, driven by large-scale datasets and transformer-based architectures such as Whisper [1] and wav2vec 2.0 [2]. Yet these gains have not extended equitably to individuals with atypical speech, particularly those with dysarthria, a motor speech disorder caused by neurological conditions such as Parkinson's disease (PD), cerebral palsy (CP), or amyotrophic lateral sclerosis (ALS) [3, 4]. Dysarthric speech is characterized by imprecise articulation, irregular pacing, and unstable phonation, resulting in high *inter-speaker* variability across etiologies and significant *intra-speaker* variability due to fatigue or disease progression [5, 6]. Consequently, even state-of-the-art ASR systems, including those fine-tuned on dysarthric data, exhibit substantial performance degradation, especially for moderate-to-severe cases [4, 7].

A primary bottleneck for developing a robust ASR for dysarthric speech is a data scarcity [8, 9]. As Yue et al. [10] document, collecting dysarthric speech corpora involves protracted recruitment, speaker fatigue requiring frequent breaks, labor-intensive transcription, and burdensome clinical annotation (e.g., severity scoring, intelligibility ratings). These constraints render traditional data collection unsustainable at scale. While benchmark datasets like UASpeech [11] and TORGO [3] offer valuable resources, they remain small (<20 speakers). Even large-scale initiatives like the Speech Accessibility Project (SAP) [12], while expanding speaker coverage, cannot approach the scale of typical speech datasets, severely limiting the development of robust ASR systems.

Earlier studies have explored using synthetic speech to address limited dysarthric corpora. Wagner et al. [13] proposed a personalized fine-tuning setup in which synthetic speech is produced by TTS conditioned on speaker x-vectors, capturing broad speaker traits for adaptation. In parallel, zero-shot voice-cloning systems such as VALL-E [14], F5-TTS [15], and Higgs Audio [16] can reproduce a speaker's vocal identity and prosody from only seconds of reference audio, leveraging discrete speech representations and large-scale audio-text pretraining without speaker-specific fine-tuning.

Most dysarthric synthesis work, however, conditions on fixed embeddings (e.g., x-vectors), which may under-represent dysarthria-salient cues such as irregular timing, coarticulatory reduction, and phonation instability. In contrast, we condition cloning directly on the input waveform using Higgs Audio V2, exposing the generator to the full acoustic signal (temporal dynamics and spectral distortions) and enabling richer preservation of speaker-specific pathology.

We present a systematic study of waveform-conditioned zero-shot cloning for personalized dysarthric ASR. We synthesize 14.94 hours of speech by cloning 8 TORGO speakers from a single reference utterance (average 7.2 seconds) per speaker and prompting with linguistically diverse text from the SAP-240430 dataset [17]. Whisper-medium is fine-tuned on the cloned corpus and evaluated exclusively on real dysarthric test utterances. Cloning fidelity is assessed via `pyannote.audio` speaker-embedding cosine similarity [18] and dynamic time warping (DTW) alignment; downstream performance is measured by WER. Training on cloned data alone reduces WER from 32.96% to 28.12% (14.7% relative improvement). Adding just 1.55 hours of real dysarthric speech further reduces WER to 13.98% (57.6% relative vs. baseline). These results support waveform-conditioned cloning as a scalable, low-burden augmentation strategy that reduces dependence on costly real data.

2. RELATED WORK

2.1. Synthetic Data for Dysarthric ASR

To mitigate data scarcity, prior work has explored synthetic data generation through text-to-speech (TTS) and voice conversion techniques. Recently, TTS models have been investigated [19–21]. These approaches have shown promise in generating dysarthric-like speech. However, they suffer from critical limitations, such as many requiring speaker-specific fine-tuning or multi-utterance enrollment, reintroducing the data collection bottleneck they aim to solve. Also, trained predominantly on typical speech, they often ”normalize” pathological prosody, smoothing out irregular pauses, breathiness, or articulatory patterns, resulting in synthetic data that fails to reflect true dysarthric acoustics.

Similarly, voice conversion based methods [8, 22–24], which transform typical speech into dysarthric-like speech, face even greater constraints. They typically require parallel recordings (identical utterances from source and target speakers), a near-impossible requirement for dysarthric populations due to fatigue and variability, or rely on complex, pathology-specific acoustic mappings that do not scale across speakers or disorders. Moreover, both TTS and VC approaches often generate linguistically constrained outputs (e.g., limited to pre-recorded phrases or narrow vocabularies), reducing lexical diversity and diminishing their effectiveness as training data for real-world ASR applications.

2.2. Zero-Shot Voice Cloning

A key advantage of zero-shot voice cloning is its ability to synthesize out-of-vocabulary (OOV) utterances while preserving the speaker’s pathological prosody. Exposing the ASR model to synthetic speech of novel phrases under dysarthric constraints improves recognition of unseen lexical items during testing. This effectively expands the lexical coverage without additional real data, as shown in [25]. Recent zero-shot cloning models such as VALL-E [14], YourTTS [26], F5-TTS [15], and Higgs Audio [16] represent a significant advance. Trained on huge, diverse speech datasets, they can generate novel utterances in a target speaker’s voice using a few seconds of reference audio, without speaker-specific fine-tuning. These models leverage discrete speech representations and transformer-based decoders to capture fine-grained timbre, rhythm, prosody, and dysarthric characteristics [27]. However, to the best of our knowledge, no prior work has systematically investigated whether cloned dysarthric speech, generated from a single utterance per speaker, can effectively augment ASR training and improve recognition performance on real dysarthric speech. Our work fills this research gap by evaluating Higgs Audio based voice cloning as a practical, low-burden augmentation strategy for dysarthric ASR.

3. PROPOSED FRAMEWORK

3.1. Higgs Audio for Voice Cloning

Our proposed framework is shown in Figure 1. First, we carefully select a single reference text and corresponding audio (average 7.2 seconds) for each speaker in the TORGO dataset for cloning. We choose the phonetically rich sentence ”*The quick brown fox jumps over the lazy dog*” as the reference text. Along with this, we pass out-of-domain text prompts from SAP-240430 [17] to be cloned using the Higgs Audio V2 model [16]. It is a large-scale (5B parameters), open-source state-of-the-art audio foundation model trained on over 10 million hours of diverse audio-text pairs. Designed for zero-shot LLM-driven TTS synthesis, it requires no speaker-specific fine-tuning or post-training adaptation. The architecture employs a dual feed-forward network (Dual-FFN) based audio adapter, which processes acoustic (prosodic/timbral) representations, enabling the modeling of vocal characteristics. Operating at 24 kHz, the model supports system prompt-based control for lexical and stylistic variation through the semantic encoder. It maintains temporal coherence across long-form synthesis, ensuring stable prosody and speaker consistency. These properties make it well-suited for low-resource pathological speech applications, where preserving atypical timing, phonation, and speaker-specific traits is critical for downstream tasks like ASR. To balance naturalness and expressive variability during cloning, we used a sampling configuration with a `temperature=1.0`, `top_k=50`, and `top_p=0.95`.

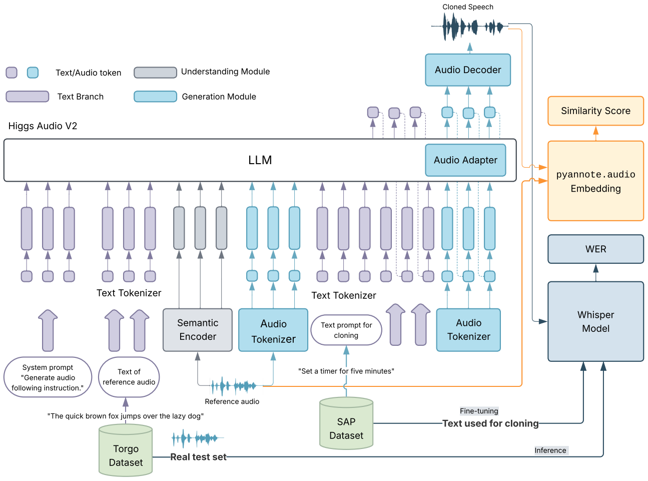

Fig. 1. Overview of the proposed Higgs Audio based voice cloning framework. System prompts are instructions to control tone or style (not used in this work).

3.2. The Whisper Model for ASR Task

Next, we fine-tuned Whisper [1] on the cloned speech utterances and corresponding text. Following our prior work [4, 7], we employed the multilingual Whisper-medium model (769M parameters) as our ASR backbone. The Whisper model adopts an encoder-decoder Transformer architecture tailored for large-scale speech recognition. The encoder processes 80 log-Mel spectrogram features extracted from raw audio, using a series of multi-head self-attention layers to capture both local acoustic patterns and long-range temporal dependencies. The decoder, conditioned on the encoder’s hidden states, autoregressively generates text tokens with cross-attention layers linking the acoustic and linguistic representations. To enhance robustness, Whisper is trained on a multitask objective, including speech recognition, translation, and language identification. This enables the decoder to handle multilingual inputs and adapt to diverse tasks within the same architecture. Unlike conventional ASR pipelines that require separate acoustic, pronunciation, and language models, Whisper integrates these components into a single end-to-end framework. As the baseline, we evaluated the pretrained model in a zero-shot setting, without any fine-tuning, leveraging Whisper’s strong cross-domain generalization capabilities demonstrated in [1]. Fine-tuning was performed with an effective batch size of 32, a `learning_rate=5e-6` and `weight_decay=0.01`. After fine-tuning, we evaluated performance by running inference on the real TORGO test set. We used beam search with a `num_beams=10`, set `no_repeat_ngram_size=3` to discourage repetitive outputs.

In parallel, we estimated speaker similarity score by computing the cosine similarity between embeddings of the cloned utterances and the reference audio, extracted with the `pyannote.audio` speaker-embedding model [18]; further details are given in Section 3.3.2.

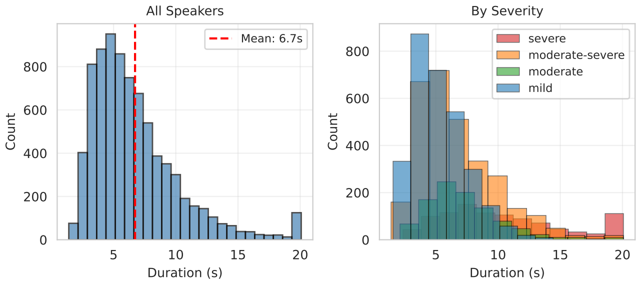

Fig. 2. Duration distribution of cloned utterances.

3.3. Methodology

3.3.1. Datasets

TORGO dataset [3], developed by the University of Toronto, provides clinically annotated speech recordings from individuals with dysarthria due to cerebral palsy (CP) and amyotrophic lateral sclerosis (ALS). The corpus contains approximately 23 hours of speech from 8 dysarthric speakers (5 male, 3 female) and 7 neurotypical controls. Recordings include non-words (e.g., /iy-p-ah/), isolated words (e.g., ”yes”, ”no”), phonetically balanced restricted sentences, and unrestricted spontaneous speech, offering rich phonetic and prosodic coverage. Dysarthric speakers are clinically categorized by severity: severe, moderate-severe, moderate, and mild. For this study, we focus exclusively on restricted speech utterances to ensure linguistic complexity and natural prosody, for robust evaluation of zero-shot voice cloning under realistic communicative conditions.

SAP-240430 dataset [17], produced by the University of Illinois Urbana-Champaign, was established to advance research and development in ASR and related machine learning tasks for individuals with speech disabilities. We employed the dataset version released on April 30, 2024, as the source of text prompts for voice cloning. It was specifically curated to provide broad linguistic coverage, encompassing digital assistant commands, syntactically novel sentences, and spontaneous speech prompts. The text prompts were normalized to expand contractions (e.g., let's → let us), convert digits to words (e.g., 7 → seven), and sentences containing fewer than three words were filtered out to ensure sufficient phonetic and prosodic content for meaningful voice cloning.

Synthesized Cloned Corpus contains 1000 synthesized utterances per speaker present in the TORGO dataset, resulting in a total of 8000 cloned utterances (14.94 hours). The mean utterance duration is 6.72 seconds ($\sigma = 3.49$ s), reflecting natural variation in phrase length and speaker-specific prosody. The distribution of synthesized speech across severity categories is depicted in Figure 2 and is as follows: severe (2.90 hours), moderate-severe (5.58 hours), moderate (2.02 hours), and mild (4.45 hours). Although utterance count is fixed per speaker, total duration varies from 1.39 hours (F03, mild) to 2.89 hours (M04, severe), due to two key factors: (1) inherent variation in SAP's sentence length, and (2) speaker-specific prosodic pacing, including dysarthria-induced pauses and articulation rate.

3.3.2. Evaluation Metrics

Speaker Similarity: To evaluate the fidelity of speaker identity preservation in cloned utterances, we compute the cosine similarity between speaker embeddings extracted using `pyannote.audio` speaker embedding model [18] from the original reference audio and cloned utterances.

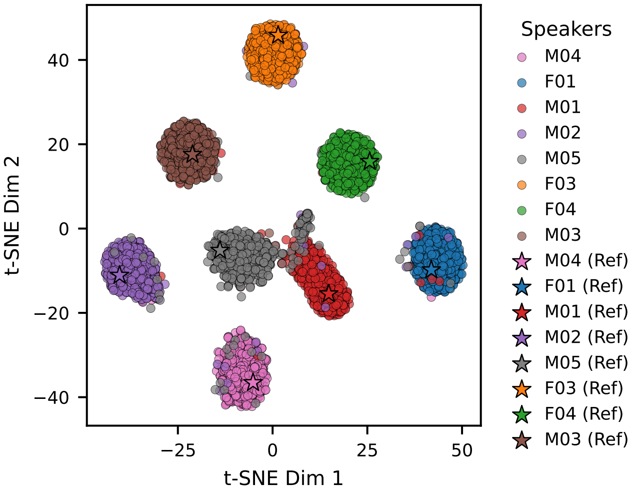

Fig. 3. t-SNE projection of `pyannote.audio` speaker embeddings of TORGO speakers. Stars ($\star$) denote embeddings from original/reference dysarthric speech in the TORGO dataset, while dots ($\bullet$) represent embeddings from cloned speech using our zero-shot voice cloning framework.

Let $\mathbf{r} \in \mathbb{R}^D$ be the embedding of the real utterance, and $\mathbf{c}_j \in \mathbb{R}^D$, $j = 1, \ldots, M$ the embeddings of its cloned utterances. $\varepsilon > 0$ prevent division by zero if a vector is (near) all-zeros ($\varepsilon \approx 10^{-8}$):

$$s_j = \frac{\mathbf{r}^\top \mathbf{c}_j}{\max\big(\|\mathbf{r}\|_2,\, \varepsilon\big)\ \max\big(\|\mathbf{c}_j\|_2,\, \varepsilon\big)} \tag{1}$$

The average similarity across all clones is

$$\bar{s} = \frac{1}{M} \sum_{j=1}^{M} s_j \tag{2}$$

Word Error Rate (WER): To evaluate the effectiveness of cloned speech as training data for dysarthric ASR, we report WER on held-out real dysarthric test utterances from the TORGO dataset. The text prompts used to generate cloned speech and the transcripts of the real TORGO test utterances are lexically disjoint, ensuring no phrase or sentence appears in both training and test sets. This strict separation validates that performance gains reflect true generalization.

4. RESULTS AND DISCUSSION

4.1. Speaker Similarity Analysis

Figure 3 depicts the distribution of speaker embeddings extracted using the `pyannote.audio` toolkit[1]. The clustering of cloned embeddings ($\bullet$) tightly around their corresponding reference speaker's embedding ($\star$) for each speaker (e.g., F03, F04, M03) demonstrates that the Higgs Audio model successfully preserves the unique vocal identity of the source speaker. We observed that some speakers, notably M04 and M05, exhibit greater dispersion in their cloned

[1] `https://huggingface.co/pyannote/embedding`

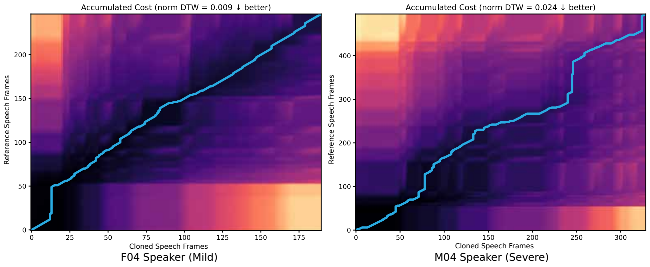

Fig. 4. Comparative dynamic time warping (DTW) alignment for mild (F04) vs. severe (M04) dysarthria.

embeddings, indicating higher variability in the synthesized output for these individuals, correlating with the high severity of their dysarthria.

Figure 4 contrasts the frame-level acoustic alignment between cloned and reference speech for a mild dysarthria speaker (F04, left) and a severe dysarthria speaker (M04, right). The accumulated DTW cost, which measures the total dissimilarity along the optimal alignment path, is significantly lower for F04 (normalized cost = 0.009) than for M04 (normalized cost = 0.024). This quantitative difference visually manifests in the alignment path: F04's path is smoother and closer to the diagonal, indicating a more precise, frame-by-frame match between the clone and reference. In contrast, M04's path exhibits more deviation and jitter, reflecting the model's greater difficulty in replicating the irregular timing, prolonged segments, and unstable phonation characteristic of severe dysarthria. This result provides direct evidence that the acoustic fidelity of zero-shot voice cloning degrades with increasing severity of the speech disorder.

4.2. Downstream ASR Task Performance

Table 1 summarizes the results of dysarthric ASR. Our baseline Whisper-medium model, evaluated without fine-tuning, performs poorly (32.96% WER), confirming that general-purpose ASR systems are ill suited for dysarthric speech. The FT–Clone-only condition, where Whisper-medium is fine-tuned exclusively on 14.94 hours of zero-shot cloned speech and evaluated on real TORGO test utterances, achieves an overall WER of 28.12%, representing a 14.7% relative improvement over baseline (32.96%). This gain confirms that synthetic dysarthric speech, generated from just one reference utterance per speaker, can be a highly effective, low-burden augmentation strategy. Improvements are most pronounced for moderate-to-severe speakers: for example, F01's WER drops from 70.21% to 43.40%, and M04's from 75.85% to 63.54%, indicating that cloned data helps counteract the model's inherent bias toward typical speech patterns.

Interestingly, augmenting the cloned dataset with only 1.55 hours of real dysarthric speech (FT–Clone+Real) from the TORGO train set yields a dramatic further reduction in overall WER to 13.98%, a 57.59% relative improvement over baseline and a 50.28% improvement over FT–Clone-only. This suggests that cloned speech provides scalable, linguistically diverse acoustic coverage, while minimal real data acts as a high-leverage acoustic calibrator, enabling the model to capture better pathological prosody, irregular pauses, and unstable phonation. For instance, severe speaker M04 (22.2 min of real speech) improves from 63.54% (FT–Clone-only) to 32.97% (FT–Clone+Real), and moderate speaker M05 (5.4 min of real speech) improves from 42.19% to 19.05%.

Table 1. WER across training conditions and dysarthria severity levels. All models were evaluated on the real TORGO test set.

Severity	**Speaker**	**WER (%) ↓**		
		Baseline	**FT–Clone-only**	**FT–Clone+Real**
Severe	M04	75.85	63.54	32.97
Mod-severe	F01	70.21	43.40	25.00
	M01	58.10	44.55	13.62
	M02	47.62	35.30	21.25
	Average	58.64	41.08	19.96
Moderate	M05	39.88	42.19	19.05
Mild	F03	14.02	17.74	9.52
	F04	4.02	5.61	2.63
	M03	1.67	3.66	2.29
	Average	6.57	9.00	4.81
Overall WER		**32.96**	**28.12**	**13.98**

Critically, these findings reframe the role of synthetic data: cloned speech is not a substitute for real pathological recordings, but a force multiplier. It reduces the logistical burden of data collection, accelerates model development, and, when combined with even tiny amounts of real speech, enables rapid, cost-effective adaptation of ASR systems to underserved clinical populations. This hybrid paradigm offers a practical, scalable, and scientifically grounded pathway toward equitable and accessible speech technology for individuals with dysarthria.

5. CONCLUSION

This work demonstrates that zero-shot voice cloning can serve as a scalable, low-burden method to augment training data for personalized dysarthric ASR. Fine-tuning Whisper-medium on synthetic speech synthesized from a single utterance per speaker reduces WER by 14.7% relative to the zero-shot baseline. When combined with just 1.55 hours of real dysarthric speech, pooled across speakers, performance improves further, representing a 50.28% relative improvement compared with the clone-only model. These results suggest that synthetic data does not replace real recordings but can meaningfully complement them, particularly in low-resource settings. Our approach directly applies to scenarios where a new speaker provides only a few minutes of enrollment data. That data can be combined with synthetic speech cloned from the same speaker, and optionally with data from other dysarthric speakers, to improve ASR performance without requiring extensive recordings. This is highly relevant for clinical or assistive applications involving new users with dysarthria, where rapid personalization with minimal data is essential.

6. REFERENCES

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning (ICML)*. PMLR, 2023, pp. 28492–28518.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[3] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

[4] Satwinder Singh, Qianli Wang, Zihan Zhong, Clarion Mendes, Mark Hasegawa-Johnson, Waleed Abdulla, and Seyed Reza Shahamiri, "Robust Cross-Etiology and Speaker-Independent Dysarthric Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[5] Heejin Kim, Katie Martin, Mark Hasegawa-Johnson, and Adrienne Perlman, "Frequency of consonant articulation errors in dysarthric speech," *Clinical linguistics & phonetics*, vol. 24, no. 10, pp. 759–770, 2010.

[6] Pam Enderby, "Disorders of communication: Dysarthria," *Handbook of clinical neurology*, vol. 110, pp. 273–281, 2013.

[7] Satwinder Singh, Zihan Zhong, Qianli Wang, Clarion Mendes, Mark Hasegawa-Johnson, Waleed Abdulla, and Seyed Reza Shahamiri, "A Comprehensive Performance Evaluation of Whisper Models in Dysarthric Speech Recognition," in *International Conference on Neural Information Processing*. Springer, 2024, pp. 75–90.

[8] Wei-Zhong Zheng, Ji-Yan Han, Chen-Yu Chen, Yuh-Jer Chang, and Ying-Hui Lai, "Improving the efficiency of dysarthria voice conversion system based on data augmentation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4613–4623, 2023.

[9] Qianli Wang, Zihan Zhong, Satwinder Singh, Clarion Mendes, Mark Hasegawa-Johnson, Waleed Abdulla, and Seyed Reza Shahamiri, "Dysarthric speech conformer: Adaptation for sequence-to-sequence dysarthric speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[10] Zhengjun Yue, Mara Barberis, Tanvina Patel, Judith Dineley, Willemijn Doedens, Lottie Stipdonk, YuanYuan Zhang, Elke de Witte, Erfan Loweimi, Djaina Satoer, et al., "Challenges and practical guidelines for atypical speech data collection, annotation, usage and sharing: A multi-project perspective," in *Proc. Interspeech 2025*, 2025, pp. 3943–3947.

[11] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon R Gunderson, Thomas S Huang, Kenneth L Watkin, and Simone Frame, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[12] Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwilling, Marie Moore Channell, Laura Mattie, Heather Hodges, et al., "Community-supported shared infrastructure in support of speech accessibility," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4162–4175, 2024.

[13] Dominik Wagner, Ilja Baumann, Natalie Engert, Seanie Lee, Elmar Nöth, Korbinian Riedhammer, and Tobias Bocklet, "Personalized Fine-Tuning with Controllable Synthetic Speech from LLM-Generated Transcripts for Dysarthric Speech Recognition," in *Interspeech*, 2025, pp. 3294–3298.

[14] Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei, "Vall-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers," *arXiv preprint arXiv:2406.05370*, 2024.

[15] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024.

[16] Boson AI, "Higgs Audio V2: Redefining Expressiveness in Audio Generation," `https://github.com/boson-ai/higgs-audio`, 2025, GitHub repository. Release blog available at `https://www.boson.ai/blog/higgs-audio-v2`.

[17] Xiuwen Zheng, Bornali Phukon, Jonghwan Na, Ed Cutrell, Kyu J Han, Mark Hasegawa-Johnson, Pan-Pan Jiang, Aadhrik Kuila, Colin Lea, Bob MacDonald, et al., "The Interspeech 2025 Speech Accessibility Project Challenge," in *Proc. Interspeech 2025*, 2025, pp. 3269–3273.

[18] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "Pyannote.Audio: neural building blocks for speaker diarization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020.

[19] Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze, "Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis," in *Proc. Interspeech 2024*, 2024, pp. 2494–2498.

[20] Mohammad Soleymanpour, Michael T Johnson, Rahim Soleymanpour, and Jeffrey Berry, "Accurate synthesis of dysarthric speech for asr data augmentation," *Speech Communication*, vol. 164, pp. 103112, 2024.

[21] Enno Hermann and Mathew Magimai Doss, "Few-shot dysarthric speech recognition with text-to-speech data augmentation," in *Proc. Interspeech 2023*, 2023, pp. 156–160.

[22] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu, "Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition," in *Interspeech*, 2018, pp. 471–475.

[23] Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, and Tomoki Toda, "Towards identity preserving normal to dysarthric voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6672–6676.

[24] Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6009–6013.

[25] John Harvill, Dias Issa, Mark Hasegawa-Johnson, and Changdong Yoo, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6428–6432.

[26] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.

[27] Krishna Gurugubelli, Anil Kumar Vuppala, et al., "Fairness in Dysarthric Speech Synthesis: Understanding Intrinsic Bias in Dysarthric Speech Cloning using F5-TTS," in *Proc. Interspeech 2025*, 2025, pp. 2750–2754.