Formant-Guided Speech Repair for Enhanced Comprehension of Dysarthric Speech

Xin-Yu Chen [1], *Jing-Tong Tzeng* [1], *Carlos Busso* [2], *Chi-Chun Lee* [1]

[1] Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2] Language Technologies Institute, Carnegie Mellon University, USA

xychen@gapp.nthu.edu.tw, roger37890426@gmail.com, busso@cmu.edu, cclee@ee.nthu.edu.tw

Abstract

Dysarthria severely degrades speech intelligibility, limiting effective communication and social participation. While neural reconstruction shows promise, most methods operate as black boxes that overlook known acoustic patterns, e.g., vowel formants that are fundamental to human speech perception. We propose a modular framework featuring a *formant-aligned spectral transformation (FAST)* module. Unlike implicit approaches, *FAST* explicitly regularizes the acoustic space by correcting distorted formants prior to synthesis. Integrated with dysarthria-adapted ASR and speaker-adaptive TTS, our system enhances intelligibility while preserving identity and prosody. Experiments on Mandarin and English datasets demonstrate strong generalization. On the primary corpus, the framework achieves a 72.4% relative reduction in CER. Subjective evaluations further indicate over 90% gains in intelligibility, comprehension, and fluency, with listening effort reduced by 60.9%.

Index Terms: Dysarthric Speech, Speech Repair, Speech Intelligibility, Speech Reconstruction, Formant Correction

1. Introduction

Dysarthria is a neuromotor speech disorder resulting from damage to the central or peripheral nervous system, impairing the strength, coordination, and timing of speech musculature. These deficits substantially degrade speech intelligibility, voice quality, and prosody [1, 2], thereby limiting daily communication and social participation. Although clinical studies emphasize that rehabilitation should restore effective communicative participation [3], most existing assistive technologies focus on *automatic speech recognition* (ASR) [4, 5]. Recent *self-supervised learning* (SSL) methods have improved ASR robustness to pathological speech [6, 7]; however, ASR produces text-only outputs that discard paralinguistic cues essential for natural interaction. *Dysarthric speech reconstruction* (DSR), therefore, adopts a speech-in, speech-out paradigm (closer to real human communication) to enhance intelligibility.

Despite recent progress, DSR methods face a persistent trade-off between intelligibility and naturalness. *Voice conversion* (VC) approaches preserve speaker characteristics but provide limited intelligibility gains for severe dysarthria due to complex non-linear distortions [8]. In contrast, *text-to-speech* (TTS)-based systems achieve high intelligibility but often generate robotic speech that fails to retain the speaker's original prosody [9]. Recent *state-of-the-art* (SOTA) systems based on SSL discrete units [10], latent diffusion models [11], or disentangled style transfer [12] show promising results. However, these approaches often rely on textual transcripts or function as black-box model, offering limited acoustic interpretability and being primarily evaluated on isolated English words.

Clinical phonetic studies consistently identify vowel acoustics as a key determinant of speech intelligibility. Impaired articulatory control in dysarthria leads to a collapsed vowel spaces, reducing the spectral contrast required for reliable phonetic perception [13–15]. Token-based vowel measures have been shown to better predict intelligibility than dynamic trajectory measures [16]. This issue is particularly critical in Mandarin, where vowel quality is closely coupled with tonal realization [17]. Moreover, neurophysiological evidence indicates that vowel sounds are encoded in the auditory cortex through a two-dimensional formant representation [18], highlighting the necessity of explicitly restoring formant structure.

To bridge the gap between clinical insight and neural speech synthesis, we propose *formant-aligned spectral transformation (FAST)*, a linguistically grounded dysarthric speech repair framework that explicitly restores vowel-level acoustic structure prior to synthesis. Unlike purely data-driven approaches, *FAST* performs formant-aligned spectral correction by mapping distorted vowel acoustics toward healthy references, thereby providing interpretable and articulatory-aware guidance for neural reconstruction. We evaluate the proposed framework on multiple dysarthric speech datasets spanning different languages and severity levels [19–22]. Experimental results demonstrate the robustness of our approach, achieving recognition error rate reductions ranging from 29% to 72% across these corpora. These objective gains are further validated through comprehensive human listening evaluations, which confirm substantial improvements in perceptual intelligibility and speech naturalness. In addition, vowel space analyses provide clear acoustic evidence that the proposed method effectively repairs collapsed articulatory representations, confirming that explicit acoustic correction is a critical component for effective dysarthric speech repair.[1]

2. Methodology

The overall architecture of our proposed system is illustrated in Fig. 1. It consists of three main modules: (1) **Dysarthria-Adapted ASR**, which converts dysarthric speech into linguistic representations; (2) ***FAST***, a formant-aligned spectral transformation module that corrects the vowel space to improve intelligibility; and (3) **Speaker-Adaptive TTS**, which synthesizes natural speech while preserving the speaker's identity. This modular design enables effective repair of dysarthric speech while balancing clarity and speaker characteristics. The implementation is publicly available.[2]

2.1. Dysarthria-Adapted ASR

The Dysarthria-Adapted ASR module converts dysarthric speech into linguistic representations by leveraging pre-trained

[1] Demo Page: https://xinyu0308.github.io/FAST-SR-Demo/
[2] Code: https://github.com/xinyu0308/FAST-SR

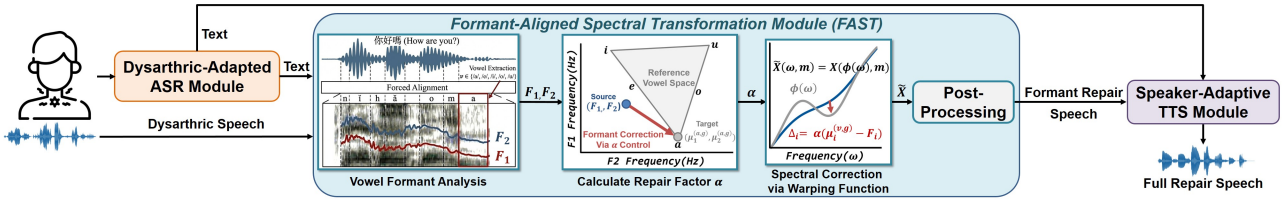

Figure 1: *The overall architecture of our proposed system.*

Wav2vec 2.0 embeddings [23] as input features. A Conformer-based acoustic model processes these embeddings to capture both global and local spectro-temporal dependencies [24]. On top of this, a recurrent decoder autoregressively generates sub-word sequences. To improve training, we adopt a hybrid objective that combines complementary loss functions:

$$\mathcal{L}_{\text{ASR}} = \lambda_{\text{CTC}}\mathcal{L}_{\text{CTC}} + (1 - \lambda_{\text{CTC}})\mathcal{L}_{\text{CE}}. \tag{1}$$

where λ_{CTC} balances the two terms. $\mathcal{L}_{\text{CTC}}$ enforces stable alignment between speech and text, while the attention-based cross-entropy $\mathcal{L}_{\text{CE}}$ enhances predictive accuracy. The resulting transcriptions are used for both forced alignment in the *FAST* module and as linguistic inputs to the TTS module.

2.2. *Formant-Aligned Spectral Transformation (FAST)*

The *FAST* module takes as input the source dysarthric speech signal and the corresponding predicted transcriptions from the ASR system. Phone boundaries are obtained using the Montreal Forced Aligner [25]. Formant reference values $\mu_i^{(v,g)}$ ($i \in \{1, 2\}$) are derived from the AISHELL-1 corpus [26] for Mandarin and the health control part of the TORGO database [22] for English, stratified by vowel class $v \in \{$/a/, /e/, /i/, /o/, /u/$\}$ and speaker gender $g \in \{\text{male}, \text{female}\}$. For instance, $\mu_1^{(/\text{a}/,\text{female})}$ denotes the average F_1 of the vowel /a/ derived from female reference speakers.

For each vowel segment, formants were estimated using Praat's Burg method [27]. Instead of a single midpoint, we sampled multiple points across the central 80% of the segment and took the median of F_1 and F_2 to improve robustness against local estimation errors. The resulting (F_1, F_2) values were then matched to the corresponding vowel class v and gender g, so that deviations could be computed against the corpus-derived references $\mu_i^{(v,g)}$:

$$\text{dev} = \max\left\{\frac{|F_1 - \mu_1^{(v,g)}|}{\mu_1^{(v,g)}}, \frac{|F_2 - \mu_2^{(v,g)}|}{\mu_2^{(v,g)}}\right\}. \tag{2}$$

A repair factor α was applied to scale correction strength:

$$\alpha = \begin{cases} 0, & \text{if } \kappa \cdot \text{dev} < \tau, \\ \min\{1,\ \kappa \cdot \text{dev}\}, & \text{otherwise.} \end{cases} \tag{3}$$

where κ controls sensitivity, τ avoids trivial adjustments.

First, we transform the input speech signals into the *short-time Fourier transform* (STFT) domain to obtain spectral representations $X(\omega, m)$. We then apply formant corrections using a Gaussian-weighted warping function [28]:

$$\phi(\omega) = \omega + \sum_{i=1}^{2} \Delta_i \exp\left(-\frac{(\omega - F_i)^2}{2\sigma_i^2}\right), \tag{4}$$

$$\text{where } \Delta_i = \alpha(\mu_i^{(v,g)} - F_i).$$

Table 1: *Statistics of the dysarthric speech datasets.* ***Eff. Dur.*** *denotes the effective duration after filtering.*

Dataset	Language	Pathology	Content	# Spk	Eff. Dur.
CDSD [19]	Mandarin	CP & Others	Segmented Narrative	44	29.4h
MDSC [20]	Mandarin	CP & Degeneration	Spoken Commands	21	9.1h
MSDM [21]	Mandarin	Subacute Stroke	Isolated Sentences	62	2.2h
TORGO [22]	English	CP & ALS	Isolated Sentences	8	2.5h

where σ_i is set proportional to the original formant frequency to control the bandwidth of local warping. The warped spectrum $\widetilde{X}(\omega, m)$ was obtained by remapping the original spectrum:

$$\widetilde{X}(\omega, m) = X(\phi(\omega), m). \tag{5}$$

Finally, inverse STFT is performed to reconstruct the time-domain waveform. Additional post-processing includes frame-wise spectral cross-fading for smooth transitions, RMS normalization for energy consistency, and high-frequency enhancement for perceptual clarity. The output of the *FAST* module is the source dysarthric speech with corrected vowel formants, which is then fed into the Speaker-Adaptive TTS module along with the ASR-predicted transcriptions to perform complete speech restoration.

2.3. Speaker-Adaptive TTS

The Speaker-Adaptive TTS module, built upon the XTTS v2 framework [29], reconstructs intelligible speech while preserving the original speaker's identity. Instead of standard reference audio, the synthesis is conditioned on the *FAST*-corrected speech, which serves two critical roles: (1) providing prosodic and paralinguistic features via the conditional encoder, and (2) defining the target timbre through the speaker encoder. These acoustic conditions are fused with the text embeddings from the ASR transcriptions to guide the generative decoder.

To effectively capture the characteristics of dysarthric speakers without overfitting, we employ a lightweight speaker adaptation strategy. We freeze the core TTS backbone and fine-tune only the speaker encoder. Given an utterance x_u, the speaker encoder f_{enc} is optimized to generate an embedding close to the target speaker centroid μ_s, ensuring a stable and faithful speaker representation:

$$\mathcal{L}_{\text{spk}} = \|f_{\text{enc}}(x_u) - \mu_s\|^2. \tag{6}$$

This approach allows the system to leverage the robust generative priors of the pre-trained model while adapting specifically to the user's voice using minimal data.

3. Experiments

3.1. Datasets and Preprocessing

We evaluated the proposed framework on pathological speech from four datasets: Part A of the *Chinese dysarthric speech database* (CDSD) [19], the *Mandarin dysarthria speech corpus* (MDSC) [20], the *Mandarin subacute stroke dysarthria multimodal* (MSDM) database [21], and the TORGO database [22]. These corpora encompass diverse neurological disorders including *cerebral palsy* (CP), hepatolenticular degeneration, subacute stroke, and *amyotrophic lateral sclerosis* (ALS).

Table 2: *Comparison of CER/WER, UTMOS, and Speaker Similarity across different dysarthric speech repair methods.* ***Bold*** *and underline indicate the best and second-best results, respectively.*

Method	Input	Input+ *FAST*	ASR+TTS	Ours (w/o SPK)	Ours (w/o *FAST*)	**Ours (Full)**	Oracle TTS	DiffDSR [11]	RnV [30]	Liu et al. [31]
SPK	-	-	✗	✗	✓	✓	GT Text		**Baselines**	
FAST	-	✓	✗	✓	✗	✓				
CER(%) / WER(%) ↓										
CDSD	80.97	69.91	29.95	23.49	28.52	**22.33**	23.80	95.78	89.89	88.19
MDSC	88.92	76.95	46.91	42.94	45.38	**38.44**	39.74	130.51	99.09	82.91
MSDM	48.65	44.98	36.84	37.28	35.60	34.50	**32.38**	98.12	56.13	59.57
TORGO	32.06	32.90	20.67	15.76	23.18	15.42	**13.65**	89.65	52.50	50.62
UTMOS (MOS Score) ↑										
CDSD	1.576	1.546	2.083	2.245	2.076	**2.250**	2.061	1.626	1.604	2.182
MDSC	1.820	1.781	2.173	2.244	2.164	2.257	2.149	1.861	1.792	**2.268**
MSDM	2.219	2.223	2.456	2.486	2.454	**2.595**	2.468	2.024	2.146	2.550
TORGO	2.204	2.044	2.597	2.666	2.571	2.668	2.667	1.993	**2.776**	2.328
Speaker Similarity ↑										
CDSD	0.745	0.736	0.631	0.619	0.632	0.620	0.632	0.704	**0.743**	0.695
MDSC	0.767	**0.764**	0.681	0.664	0.681	0.666	0.683	0.761	0.745	0.751
MSDM	0.768	**0.761**	0.662	0.648	0.664	0.651	0.664	0.682	0.689	0.650
TORGO	0.758	**0.752**	0.684	0.673	0.684	0.675	0.686	0.582	0.670	0.735

For experimental consistency, all audio recordings were resampled to 16 kHz, with corresponding transcripts manually verified. To focus on sentence-level intelligibility, we further applied a selection criterion, excluding utterances shorter than 1.5 seconds or containing fewer than three lexical units (characters for Mandarin, words for English).

Table 1 summarizes the statistics of the processed datasets. As shown, the corpora exhibit significant diversity in linguistic content and speaking styles. MDSC is specifically designed for voice wake-up tasks (e.g., smart-home controls), whereas CDSD and MSDM encompass broader linguistic contexts, covering daily communication and narrative contexts. Regarding TORGO, only sentence-level recordings were utilized to facilitate cross-lingual evaluation. This multi-dimensional diversity spanning languages, pathologies, and linguistic styles provides a rigorous testbed to demonstrate the robustness and generality of our framework.

3.2. Experimental Setup

3.2.1. Model Architecture and Implementation

We developed a fully cascaded dysarthria-adaptive speech processing framework integrating ASR, FAST, and TTS modules. The ASR system, implemented in ESPnet [32], features a Conformer encoder and an RNN decoder utilizing frozen Wav2Vec 2.0 front-ends (WenetSpeech [33] for Mandarin, LibriSpeech [34] for English). For TTS, we employed XTTS v2 [29], fine-tuning only the speaker encoder to adapt to dysarthric characteristics while preserving identity. To ensure rigorous evaluation, all datasets were partitioned into training, validation, and test sets with a ratio of 7:2:1 in a speaker-independent manner, with the exception of MDSC, which followed its original predefined split($\approx$ 6:1:3). ASR training minimized a hybrid CTC/attention objective with $\lambda_{\mathrm{CTC}} = 0.3$ (Eq. 1), determined empirically via grid search.

3.2.2. Objective Evaluation Metrics

The pipeline is evaluated using *character error rate* (CER), *word error rate* (WER), *UTokyo-SaruLab MOS prediction system* (UTMOS), and Speaker Cosine Similarity. CER and WER measures the transcription accuracy, where lower scores indicate better performance:

$$\mathrm{CER/WER} = \frac{S + D + I}{N}, \tag{7}$$

where S, D, and I denote the number of substitutions, deletions, and insertions, respectively, and N represents the total number of characters (for CER) or words (for WER) in the reference transcript.

For speech quality, we utilize UTMOS [35] to predict listener-perceived naturalness. Speaker similarity is measured by computing the cosine similarity between the embeddings of the synthesized speech and the original input speech using Resemblyzer [36] to ensure identity preservation.

3.2.3. Acoustic Metrics for Vowel Correction

To assess the effectiveness of our vowel correction, we employ two complementary acoustic metrics. The *formant centralization ratio* (FCR) [37] is a ratio-based index designed to quantify the centralization of vowel formants towards the neutral position, independent of anatomical variability:

$$FCR = \frac{F2_u + F2_a + F1_i + F1_u}{F2_i + F1_a}, \tag{8}$$

where $F1$ and $F2$ denote the first and second formant frequencies of the corner vowels $/a/$, $/i/$, and $/u/$, respectively. A lower FCR indicates reduced centralization and more distinct vowel targets. In contrast, the *vowel space area* (VSA) measures the dynamic range of articulatory movements by calculating the absolute acoustic area formed by these vowels:

$$\mathrm{VSA} = \tfrac{1}{2} \times |F1_i(F2_a - F2_u) + F1_a(F2_u - F2_i) + F1_u(F2_i - F2_a)|. \tag{9}$$

A larger VSA signifies a wider range of motion and an expanded articulatory working space.

3.2.4. Subjective Evaluation

To assess the perceptual efficacy of the proposed *FAST* module, subjective evaluations were conducted by 18 native Mandarin listeners on 120 samples. The test set comprised four gender-balanced speakers, one from each CDSD severity level. For each speaker, 10 randomly selected utterances were processed across three conditions: *Original*, *Ours*, and *Ours w/o FAST*. Listeners rated the samples on a 5-point MOS scale for Intelligibility (clarity of articulation), Comprehension (content understanding), Fluency, and Listening Effort (cognitive load required). Samples were presented in a randomized order to minimize bias. Results are reported with 95% confidence intervals, and paired significance tests compared system performance.

3.2.5. Baselines

- **DiffDSR** [11]: A latent diffusion-based system utilizing an SSL content encoder for phoneme restoration and an in-context speaker encoder to preserve identity.
- **RnV** [30]: A dysarthric-to-healthy conversion framework that extends the Rhythm and Voice model with syllable-based rhythm modeling to correct temporal distortions.
- **Liu et al.** [31]: A two-stage voice conversion approach combining KNN-VC with same-gender retrieval for initial repair, followed by *so-vits-svc* for timbre restoration.

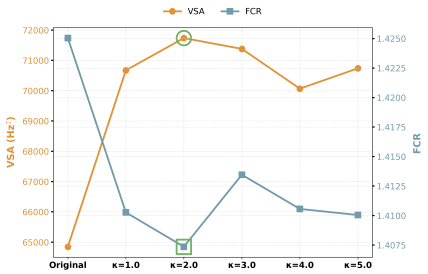

Figure 2: *The FAST module effectively repairs the compressed vowel space, achieving optimal performance at* $\kappa = 2.0$

3.3. Results and Analysis

3.3.1. Speech Repair Performances

Table 2 compares our method with TTS-based and recent speech repair baselines. Oracle TTS uses ground-truth transcripts as a topline, while ASR+TTS synthesizes speech from ASR outputs. We also include DiffDSR [11], RnV [30], and Liu et al. [31], together with ablation variants removing *speaker adaptation* (SPK) or *FAST*. The approach referred to as Ours (Full) integrates all proposed components

Ours (Full) consistently achieves the lowest CER/WER among all repair methods across datasets. Compared to the ASR+TTS baseline, our method yields substantial intelligibility gains, with relative error rate reductions of 25.4% on CDSD and 18.0% on MDSC. In contrast, external baselines exceed 50% error rates, indicating that conventional voice conversion and diffusion techniques struggle to preserve linguistic content when dealing with severe articulatory distortions.

Remarkably, our method outperforms the topline Oracle TTS on both CDSD and MDSC datasets. This counter-intuitive result can be attributed to the *FAST* module. While Oracle TTS relies on ground-truth text, its synthesis quality is still bounded by the acoustic characteristics of the dysarthric reference audio. Ours (Full) explicitly regularizes the vowel space via *FAST* before synthesis. This result highlights that correcting acoustic features is as crucial as correct textual conditioning for severe dysarthria.

Ablation studies confirm the necessity of each component. The Input+*FAST* column shows high error rates, indicating that spectral transformation alone is insufficient and must be coupled with TTS reconstruction. Removing *FAST* or speaker adaptation degrades CER/WER by approximately 4–7% absolute, confirming their complementary roles. Regarding quality, our method achieves the best or second-best UTMOS. This superior performance indicates that the reconstructed speech preserves high perceptual naturalness and acoustic fidelity. Regarding speaker similarity, we observe a strategic trade-off: a slight reduction is exchanged for massive gains in intelligibility. This result reflects the model's ability to filter out dysarthric characteristics while retaining the speaker's essential vocal timbre, resulting in highly articulate and fluent speech that is easier for listeners to understand.

3.3.2. Vowel Space Analysis and Validation

To validate the effectiveness of the *FAST* module, we analyzed both the quantitative metrics and the acoustic vowel space. First, regarding parameter selection, Fig. 2 identifies $\kappa = 2.0$ as the optimal operating point. At this setting, the system achieves the most substantial improvement, increasing the VSA by approximately 10.6% while minimizing the FCR.

To visually corroborate this restoration, Fig. 3 illustrates the acoustic vowel space for female and male speakers. Consistent with the metric gains, the vowel space progressively ex-

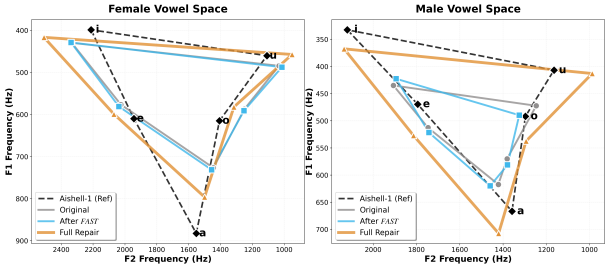

Figure 3: *Effect of FAST Module on Vowel Space Expansion in Mandarin Dysarthric Speech*

Table 3: *Subjective MOS results (Mean ± SD).*

Method	Intell. (↑)	Comp. (↑)	Fluency (↑)	Effort (↓)
Original	2.31 ± 1.35	2.20 ± 1.42	2.27 ± 1.07	3.81 ± 1.46
w/o *FAST*	4.38 ± 1.00	4.25 ± 1.14	4.11 ± 0.98	1.66 ± 1.11
Proposed	$\mathbf{4.60 \pm 0.76}$	$\mathbf{4.40 \pm 1.00}$	$\mathbf{4.32 \pm 0.88}$	$\mathbf{1.49 \pm 0.92}$

pands after applying *FAST*, moving away from the collapsed dysarthric state and approaching the normative Aishell reference. The distribution further enlarges after the backend TTS processing ("Full Repair"), demonstrating that our method effectively restores articulatory distinctiveness.

3.3.3. Subjective MOS Analysis

To evaluate the perceptual effectiveness of the proposed system, we conducted a subjective MOS evaluation on the CDSD dataset. Table 3 compares the original dysarthric speech, the repair system without *FAST*, and the full proposed method. Compared with the original recordings, the proposed system achieves substantial improvements across all metrics, nearly doubling the Intelligibility and Comprehension scores while reducing Listening Effort by more than 60%. These results indicate a pronounced enhancement in speech clarity and a significant reduction in the effort required for understanding.

Furthermore, the inclusion of the *FAST* module provides consistent perceptual gains beyond the backend model alone. As shown in Table 3, *FAST* yields relative improvements of approximately 5.0% in Intelligibility and 5.1% in Fluency, and further reduces Listening Effort by 10.2%. These results suggest that vowel formant correction effectively alleviates listeners' cognitive load and contributes to more fluent and intelligible speech.

4. Conclusion and Future Work

In this work, we presented a modular dysarthric speech repair framework that integrates a dysarthria-adapted ASR, the proposed *FAST* vowel space repair module, and a speaker-adaptive TTS backend. Extensive evaluations across diverse languages and severity levels show that our system consistently outperforms recent repair approaches, achieving the lowest CER and WER. Notably, our method even surpasses the Oracle TTS topline on several datasets, demonstrating that correct textual conditioning alone is insufficient when acoustic realizations remain distorted. By explicitly regularizing the vowel formant structure prior to synthesis, the *FAST* module restores articulatory distinctiveness and plays a crucial role in improving intelligibility. Vowel space analysis and subjective MOS evaluations further provide interpretable and perceptual evidence that acoustic correction effectively reduces listeners' cognitive load while substantially enhancing speech clarity, albeit with a minor trade-off in speaker similarity.

Future work will focus on extending articulatory repair beyond vowels and exploring adaptive control to better handle varying dysarthria severities, while improving speaker identity preservation for practical assistive communication scenarios.

5. Generative AI Use Disclosure

Generative AI tools were used solely for language editing and stylistic refinement of this manuscript. All technical content, experimental design, analysis, and conclusions were developed by the authors. The authors take full responsibility for the content of this paper.

6. References

[1] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2012.

[2] R. D. Kent, "Research on speech motor control and its disorders: A review and prospective," *Journal of Communication disorders*, vol. 33, no. 5, pp. 391–428, 2000.

[3] A. D. Palmer, J. T. Newsom, and K. S. Rook, "How does difficulty communicating affect the social relationships of older adults? an exploration using data from a national survey," *Journal of communication disorders*, vol. 62, pp. 131–146, 2016.

[4] N. Gohider and O. A. Basir, "Recent advancements in automatic disordered speech recognition: A survey paper," *Natural Language Processing Journal*, vol. 9, p. 100110, 2024.

[5] K. Rosero, A. N. Salman, S. Chandra, B. Sisman, C. Van't Slot, A. A. Kane, R. R. Hallac, and C. Busso, "Advancing pediatric asr: The role of voice generation in disordered speech," in *Proc. Interspeech 2025*, 2025, pp. 2890–2894.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[7] A. Hernandez, P. A. Pérez-Toro, E. Noeth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Interspeech 2022*, 2022, pp. 51–55.

[8] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu, and H. Meng, "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7744–7748.

[9] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech 2019*, 2019, pp. 4115–4119.

[10] X. Chen, D. Yang, D. Wang, X. Wu, Z. Wu, and H. Meng, "CoLM-DSR: Leveraging neural codec language modeling for multi-modal dysarthric speech reconstruction," in *Proc. Interspeech 2024*, 2024, pp. 4129–4133.

[11] X. Chen, D. Yang, W. Wu, M. Wu, J. Xu, X. Wu, Z. Wu, and H. Meng, "DiffDSR: Dysarthric speech reconstruction using latent diffusion model," *arXiv preprint arXiv:2506.00350*, 2025.

[12] K. Rosero, E. Yeo, D. Mortensen, C. Van't Slot, R. Hallac, and C. Busso, "Finding my voice: Generative reconstruction of disordered speech for automated clinical evaluation," *ArXiv e-prints (arXiv:2509.19231)*, pp. 1–5, September 2025.

[13] F. Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *Proceedings of the second workshop on speech and language processing for assistive technologies*, 2011, pp. 11–21.

[14] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 187–194, 2011.

[15] K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: mapping to perception," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 1, pp. 68–81, 2014.

[16] A. Thompson, M. E. Hirsch, K. L. Lansford, and Y. Kim, "Vowel acoustics as predictors of speech intelligibility in dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 66, no. 8S, pp. 3100–3114, 2023.

[17] H.-M. Liu, F.-M. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3879–3889, 2005.

[18] Y. Oganian, I. Bhaya-Grossman, K. Johnson, and E. F. Chang, "Vowel and formant representation in the human auditory speech cortex," *Neuron*, vol. 111, no. 13, pp. 2105–2118, 2023.

[19] Y. Wan, M. Sun, X. Kang, J. Li, P. Guo, M. Gao, and S.-J. Wang, "CDSD: Chinese dysarthria speech database," in *Proc. Interspeech 2024*, 2024, pp. 4109–4113.

[20] M. Gao, H. Chen, J. Du, X. Xu, H. Guo, H. Bu, J. Yang, M. Li, and C.-H. Lee, "Enhancing voice wake-up for dysarthria: Mandarin dysarthria speech corpus release and customized system design," in *Proc. Interspeech 2024*, 2024, pp. 2465–2469.

[21] J. Liu, X. Liu, Y. Yang, R. Ruzi, X. Zuo, C. Xu, C. Li, X. Li, R. Su, A.-m. Hu, Y.-M. Zhang, S. Zhao, X. Du, L. Wang, and N. Yan, "The open-access Mandarin subacute stroke dysarthria multimodal (MSDM) database for intelligent assessment," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 131–135.

[22] 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, 2012.

[23] P. Guo and S. Liu, "chinese_speech_pretrain," https://github.com/TencentGameMate/chinese_speech_pretrain, gitHub repository.

[24] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent developments on ESPnet toolkit boosted by conformer," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5874–5878.

[25] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[27] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[28] E. Snelson, Z. Ghahramani, and C. Rasmussen, "Warped Gaussian processes," *Advances in neural information processing systems*, vol. 16, 2003.

[29] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, "XTTS: a massively multilingual zero-shot text-to-speech model," in *Proc. Interspeech 2024*, 2024, pp. 4978–4982.

[30] Karl El Hajal and Enno Hermann and Sevada Hovsepyan and Mathew Magimai Doss, "Unsupervised rhythm and voice conversion to improve ASR on dysarthric speech," in *Interspeech 2025*, 2025, pp. 2760–2764.

[31] D. Liu, Y. Lin, H. Bu, and M. Li, "Two-stage and self-supervised voice conversion for zero-shot dysarthric speech reconstruction," in *2024 International Conference on Asian Language Processing (IALP)*. IEEE, 2024, pp. 423–427.

[32] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "ESPnet: End-to-end speech processing toolkit," *INTERSPEECH 2018, Hyderabad, India*, 2018.

[33] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "WenetSpeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6182–6186.

[34] 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)*. IEEE, 2015, pp. 5206–5210.

[35] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," *Interspeech 2022*, 2022.

[36] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[37] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant Centralization Ratio: A proposal for a new acoustic measure of dysarthric speech." *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 1, pp. 114–125, 2010.