DARS: Dysarthria-Aware Rhythm-Style Synthesis for ASR Enhancement

Minghui Wu*, Xueling Liu†, Jiahuan Fan†, Haitao Tang†, Yanyong Zhang*, Yue Zhang‡
* University of Science and Technology of China, China
E-mail: wmhsky@mail.ustc.edu.cn, yanyongz@ustc.edu.cn
† iFlytek Co., Ltd., China
‡ Huawei Technology, China

***Abstract*—Dysarthric speech exhibits abnormal prosody and significant speaker variability, presenting persistent challenges for automatic speech recognition (ASR). While text-to-speech (TTS)-based data augmentation has shown potential, existing methods often fail to accurately model the pathological rhythm and acoustic style of dysarthric speech. To address this, we propose DARS, a dysarthria-aware rhythm-style synthesis framework based on the Matcha-TTS architecture. DARS incorporates a multi-stage rhythm predictor optimized by contrastive preferences between normal and dysarthric speech, along with a dysarthric-style conditional flow matching mechanism, jointly enhancing temporal rhythm reconstruction and pathological acoustic style simulation. Experiments on the TORGO dataset demonstrate that DARS achieves a Mean Cepstral Distortion (MCD) of 4.29, closely approximating real dysarthric speech. Adapting a Whisper-based ASR system with synthetic dysarthric speech from DARS achieves a 54.22% relative reduction in word error rate (WER) compared to state-of-the-art methods, demonstrating the framework's effectiveness in enhancing recognition performance.**

I. Introduction

Dysarthria is a motor speech disorder caused by neurological conditions and is typically characterized by slurred articulation, reduced speech rate, and abnormal prosody. These manifestations significantly impair verbal communication and limit the social participation of dysarthric patients [1], [2]. Automatic speech recognition (ASR), as an assistive tool, holds promise for assisting individuals with dysarthria [3], [4]. However, current ASR systems still exhibit limited recognition accuracy for dysarthric speech. The main challenges arise from two key factors: (1) substantial speaker variability stemming from heterogeneous dysarthric patterns [5], [6], and (2) severe data scarcity due to the high cost and complexity of data collection. These issues hinder the development of robust ASR systems tailored to dysarthric speech [7], [8].

To address data scarcity, recent studies have explored text-to-speech (TTS)-based data augmentation to reduce the notable prosody mismatch between synthesized and real dysarthric speech [9]–[12], particularly in terms of rhythm and style distribution. Wagner *et al.* [10] integrate large language models with controllable TTS and x-vector-based speaker adaptation, achieving performance gains for severe dysarthria. However, the method relies heavily on prompt engineering and offers limited controllability over speech style. Leung et al. [11] use a diffusion-based approach to synthesize pathological speech with high-speech-quality and semantic consistency, but lack precision in modeling dysarthric rhythm patterns. Although Soleymanpour *et al.* [9] introduce severity-conditioned style controller and pause insertion mechanism, their approach remains limited in its ability to capture subtle speech characteristics in mild dysarthric cases.

To address these limitations in rhythm modeling precision and pathological style controllability, we propose the **D**ysarthria-**A**ware **R**hythm-**S**tyle (DARS) synthesis framework, based on the Matcha-TTS architecture. DARS comprises two key mechanisms:

1) A **multi-stage rhythm predictor** that follows a pause-then-duration strategy, guided by contrastive preference optimization (CPO) between normal and dysarthric speech to better capture fragmented rhythmic patterns;
2) A **dysarthria-aware conditional flow matching** mechanism that incorporates pathological style vectors to constrain the synthesis process, thereby enhancing modeling of dysarthric acoustic variations.

These two mechanisms work synergistically to significantly improve the prosody consistency of synthesized dysarthric speech, generating highly representative training data for dysarthric ASR systems.

II. Related Works

A. MATCHA-TTS

Given the computationally intensive sampling process in diffusion probabilistic model (DPM)-based TTS systems [13], [14], we adopt Matcha-TTS [15] as our backbone. This non-autoregressive encoder-decoder (ED) framework enables high-fidelity speech synthesis with significantly reduced inference steps, while achieving automatic alignment between input text and dysarthric speech without manual annotation.

The encoder employs a transformer-based architecture [16] to extract high-level semantic representations from text. The monotonic alignment search (MAS) [17] provides forced alignment information as targets for the duration predictor. During inference, the duration predictor's outputs are used to upsample the encoder outputs, yielding the conditional mean values μ. The decoder adopts U-Net-based flow-prediction network [18], consisting of downsampling, middle and upsampling modules. Each module is composed of transformer layers, and

979-8-3315-7206-8/25/$31.00

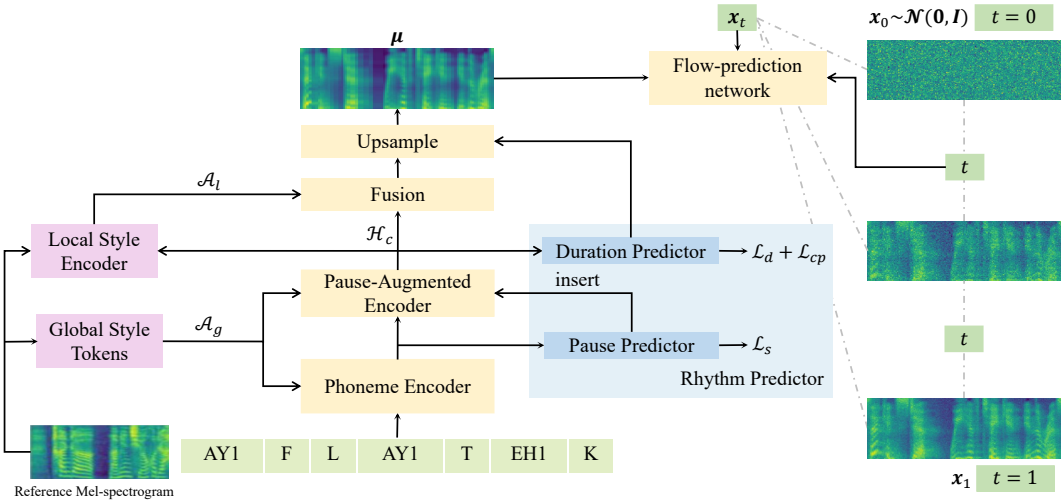

Fig. 1. DARS framework based on Matcha-TTS model

the Snake-beta activation from BigVGAN [19] is used to accelerate convergence.

During training, Matcha-TTS model eliminates the redundant multi-step denoising process in DPMs by employing an Optimal-Transport Conditional Flow Matching (OT-CFM) strategy [14], [15], [20]. This method simplifies the mapping from random noise to target speech by modeling the conditional flow-matching vector field. Acoustic features are generated by solving the corresponding ordinary differential equation (ODE) using a first-order Euler forward method [21], [22], reducing more than 50% inference steps.

B. Whisper-Based ASR

Whisper is one of the most widely adopted end-to-end speech recognition frameworks [23], featuring a transformer-based ED architecture. The encoder receives mel-spectrograms as input and produces high-level semantic representations, while the decoder generates the corresponding text sequence in an autoregressive manner. The model is trained on approximately 680,000 hours of multilingual, weakly supervised speech data, enabling strong generalization across various speech-related tasks. Whisper is released in multiple variants with different parameter scales. Among them, Whisper-large [24], with approximately 1.543 billion parameters, achieves leading performance on a range of ASR benchmarks and exhibits strong robustness in low-resource languages, noisy environments, and multi-speaker scenarios.

However, since Whisper is pretrained almost entirely on normal speech, its performance degrades significantly when applied to dysarthric speech. To address this limitation, we employ the proposed DARS-TTS system to generate large-scale dysarthria-prosody synthetic data for Whisper fine-tuning. We explore two adaptation strategies: full-parameter fine-tuning and parameter-efficient Low-Rank Adaptation (LoRA) [25], to facilitate effective adaptation to dysarthric speech.

III. PROPOSED METHOD

Building upon Matcha-TTS, we introduce rhythm and style adaptation mechanisms specifically designed for dysarthric speech. The following sections elaborate on each of these mechanisms in detail.

A. Multi-Stage Rhythm Predictor Optimized by Dysarthria-Guided Contrastive Preference

The duration predictor in Matcha-TTS is used to estimate durations from the input text. However, since the text is typically preprocessed based on normal speech, its pause patterns often differ significantly from those in dysarthric speech. To address this, we introduce a dysarthric speech rhythm predictor (blue box in Fig. 1), which integrates multi-stage prediction and CPO to precisely control the distinctive rhythmic patterns.

1) Multi-Stage Rhythm Prediction: The multi-stage rhythm predictor models abnormal rhythmic patterns in pathological speech through a sequential processing pipeline. First, the raw phoneme sequence is encoded by the phoneme encoder without pause information. A pause predictor then classifies potential pause types between phonemes and inserts corresponding pause embeddings into the sequence. The augmented sequence is re-encoded by the pause-augmented encoder to obtain contextual representations enriched with pause cues. Finally, these representations are fed into a duration predictor to estimate phoneme-level durations. This cascaded design integrates pauses explicitly into duration modeling, enabling more accurate reconstruction of dysarthric speech prosody.

Let the high-level phonemic representation sequence from the phoneme encoder be $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N]$. The pause predictor f_θ^s processes $\mathbf{E}$ to output pause class distributions $\hat{\mathbf{S}} = [\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \cdots, \hat{\mathbf{s}}_N]$. Ground-truth labels $\mathbf{S} = [s_1, s_2, \cdots, s_N]$ are derived from alignments obtained using the Montreal Forced Aligner (MFA) [26]. Specifically, pauses are grouped into K discrete categories according to their durations: class 0 indicates no pause, and classes 1 through $K-1$ represent

increasingly longer pauses. We use the cross-entropy loss to guide pause prediction:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{I}(s_i = k)\log \hat{s}_{i,k} \tag{1}$$

where $\hat{s}_{i,k}$ denotes the predicted probability that the i-th position corresponds to pause class k.

Following pause prediction, pause embeddings (corresponding to $\hat{\mathbf{S}}$) are inserted into $\mathbf{E}$, creating an augmented sequence $\mathbf{E}_{\text{aug}}$. The pause-augmented encoder processes $\mathbf{E}_{\text{aug}}$ to obtain contextual representations enriched with pause information. These representations are then fed into the duration predictor f_θ^t to estimate durations. We train the duration predictor using mean squared error (MSE) on log-scale durations:

$$\mathcal{L}_d = \frac{1}{N'}\sum_{i=1}^{N'}(\log d_i - \log \hat{d}_i)^2 \tag{2}$$

where $\hat{d}_i$ is the predicted duration of the i-th element, d_i is the reference duration obtained via the MAS alignment method [13], [15], and $N' \geq N$ denotes the sequence length after pause insertion.

2) Dysarthria-Guided Contrastive Preference Optimization: Considering the complexity and variability of dysarthric speech patterns, we introduce a dysarthria-guided contrastive preference learning mechanism built upon the multi-stage rhythm prediction module. It encourages the synthesized rhythmic style to better match the distribution of real dysarthric speech, while deviating from that of normal speech. We formulate the contrastive preference loss function $\mathcal{L}_{cp}$ as follows:

$$\mathcal{L}_{cp} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \max\Big(0,\, \mathrm{D}(\hat{d}_i, d_i) - \mathrm{D}(\hat{d}_i, d_i^n) + m\Big) \tag{3}$$

$$w_i = \begin{cases} \alpha \cdot \hat{s}_{i,s_i} & \text{if } s_i > 0 \\ \beta \cdot \hat{s}_{i,0} & \text{if } s_i = 0 \end{cases} \tag{4}$$

where $\mathrm{D}(\hat{d}_i, d_i)$ represents the distance between the predicted duration $\hat{d}_i$ and the ground-truth dysarthric duration d_i, while $\mathrm{D}(\hat{d}_i, d_i^n)$ is the distance between $\hat{d}_i$ and the duration of normal speech d_i^n. In our experiments, this distance is measured by the absolute difference. The margin m is a predefined threshold, typically set to 0.75. The contrastive loss $\mathcal{L}_{cp}$ enforces that the predicted duration should be closer to the dysarthric ground truth than to the normal-speech counterpart by at least m. The dynamic weight w_i is defined based on the model's predicted class probabilities. For positions labeled as pauses ($s_i > 0$), the weight is proportional to $\hat{s}_{i,s_i}$, the predicted probability of the true pause class. For non-pause positions ($s_i = 0$), the weight depends on $\hat{s}_{i,0}$, the probability assigned to the non-pause class. Parameters α and β control the relative weights assigned to pause and non-pause positions in the loss function. We empirically set $\alpha \geq \beta$ to emphasize accurate pause modeling, which is critical for capturing the atypical rhythmic patterns characteristic of dysarthric speech.

B. Dysarthria-aware Acoustic Conditional Flow Matching

Matcha-TTS predicts the flow matching vector field solely based on the conditional mean values μ, derived from the phoneme sequence $\mathbf{E}$ and speaker identity $\mathbf{v}$. However, in the context of dysarthric speech synthesis, such a design may overlook rich pathological prosodic patterns embedded in the acoustic features. To address this limitation, we propose a dysarthria-aware acoustic conditional flow matching mechanism, which introduces additional global and local acoustic style vectors, $\mathcal{A}_g$ and $\mathcal{A}_l$, which are fused with content-related conditions to jointly modulate the flow matching process.

Specifically, we extract global style tokens (GSTs) [27] from the reference mel-spectrogram as the global prosodic representation $\mathcal{A}_g$. Concurrently, we model frame-level style variations through a local style encoder adapted from [28], yielding $\mathcal{A}_l$, which employs a vector quantization bottleneck to extract style representations from reference mel-spectrograms and aligns them with the pause-augmented encoder's output hidden states $\mathcal{H}_c$ via attention. The aligned local style features $\mathcal{A}_l$ are then fused with $\mathcal{H}_c$ to condition the generation of the conditional mean values μ. The complete vector prediction network is defined as follows:

$$\mu = \text{Encoder}(\mathbf{E}, \mathcal{A}_g, \mathcal{A}_l) \tag{5}$$

$$\mathbf{v}_t(\mathbf{x} \mid \mu; \theta) = \text{Decoder}(\mathbf{x}, \mu, t) \tag{6}$$

where Decoder$(\cdot)$ is a U-Net-based vector field prediction network, θ denotes the model parameters, $\mathbf{x}$ is the latent variable, and $t \in [0, 1]$ represents the flow matching step.

Ultimately, our model adopts the conditional flow matching objective $\mathcal{L}_{\text{CFM}}$ from Matcha-TTS [15], which aims to learn a vector field that maps samples from a simple prior distribution to dysarthric acoustic features. Beyond content-related information, our approach further incorporates dysarthric acoustic style representations into the vector field prediction process, thereby enhancing the model's ability to reconstruct typical acoustic abnormalities found in pathological speech.

IV. EXPERIMENTS

A. Experimental Setup

The TORGO database is employed as the dysarthric speech dataset [29]. It contains recordings from 8 speakers with dysarthria and 7 healthy control speakers. The 8 dysarthric speakers have been diagnosed with either Amyotrophic Lateral Sclerosis (ALS) or Cerebral Palsy (CP). As shown in Table I, these speakers are categorized into four severity levels based on their speech impairments: Severe, Mod.-Sev., Moderate, and Mild. In the speaker descriptions, "F" and "M" indicate female and male genders, respectively, while the accompanying numbers denote participant IDs within the dataset.

TABLE I
DYSARTHRIA SEVERITY FOR THE TORGO DATABASE.

	Severe				Mod.-Sev.	Moderate	Mild	
Participant	F01	M01	M02	M04	M05	F03	F04	M03
Number of Utterances	228	739	772	659	610	1097	675	806

Given that the TORGO dataset is resource-limited, we investigate training strategies for the Matcha-TTS model to minimize data usage as much as possible:

- All-Speaker (ASp): A single TTS model is trained using the combined training data from all dysarthric speakers.
- Single-Speaker (SSp): Eight individual models are trained separately for each dysarthric speaker (F01∼M03), using only their corresponding speech data.
- Dysarthria-Severity-Group Speaker (DSpG): Four TTS models are trained independently on data grouped by severity levels of dysarthria.

For the Matcha-TTS model architecture and training strategies, we adopt the MAT-10 configuration as described in [15]. However, we expand the parameter size of the text encoder, which now comprises 8 layers with 4 attention heads and 1024 filter channels per layer. In our proposed DARS framework, the phoneme encoder and the pause-augmented encoder follow architectures similar to those of the original Matcha-TTS encoder. Specifically, the phoneme encoder is a smaller variant consisting of only 2 layers with 4 attention heads and 512 filter channels per layer, whereas the pause-augmented encoder is larger, matching the previously described text encoder in scale. The pause and duration predictors share identical network structures. Further details regarding the GST encoder can be found in [27], while the implementation of the frame-level local style encoder follows [28]. Prior to training the DARS model, we employ an MFA model [26] trained on the full TORGO dataset to extract pause labels. In CPO, the duration predictor for normal speech is pre-trained on the LibriSpeech English dataset [30] to obtain the reference duration corresponding to the input text.

For dysarthric ASR, we fine-tune the Whisper-Large model on the synthesized speech data and evaluate its performance on the original data. Since the TORGO dataset does not provide predefined splits for training, validation, and evaluation, we employ a multi-sampling strategy, selecting 10%, 20%, and 30% of the data as evaluation sets. These subsets are chosen to ensure minimal performance variation across the three splits when evaluated with the publicly available Whisper-Large model [23]. The remaining data is then divided into training and validation sets with a 9:1 ratio, which are also used for training the Matcha-TTS model. Subsequently, the trained Matcha-TTS is used to synthesize the training data, which is employed for Whisper-Large adaptation.

We use Mean Cepstral Distortion (MCD) [11], [31] to evaluate the similarity between the synthesized Mel-spectrogram and the ground-truth Mel-spectrogram. In speech synthesis evaluations, MCD has been shown to correlate with subjective evaluation results. For ASR performance evaluation, we use the Word Error Rate (WER) [23] as the primary metric. Additionally, an Overall WER is calculated by averaging the WER scores across individual speakers.

B. Experimental Results

Evaluation Set Stability Across Sampling Ratios. Table II presents the performance of the open-source Whisper-Large model on TORGO test sets sampled at different proportions. These evaluation sets (E1–E3) were sampled and recognized multiple times, demonstrating robust stability across sampling ratios. Based on this consistency, we select the 20% evaluation set for all subsequent experiments. The remaining 80% of the data serves as training and validation data for both speech synthesis and recognition tasks.

TABLE II
WER COMPARISON ON TORGO EVALUATION SETS BY SENTENCE COUNT.

ID	Whisper-Large	WER (%)				
		Severe	Mod.-Sev.	Moderate	Mild	Overall
E1	10%	125.67	182.53	32.78	13.91	83.21
E2	20%	**123.36**	**184.14**	**34.69**	**12.71**	**83.19**
E3	30%	121.91	186.25	35.47	11.98	83.24

TABLE III
MCD FOR TTS MODELS TRAINED ON TORGO DATA.

ID	Models	α	β	ASp	DSpG	SSp
E4	Grad-TTS [11]	–	–	6.61	6.71	6.81
E5	MATCHA-TTS (baseline)	–	–	6.25	6.43	6.64
E6	DARS (E5 w/ rhythm)	–	–	6.09	6.32	6.57
E7	DARS w/ CPO	0.5	0.5	5.83	6.07	6.21
E8		0.7	0.3	5.72	5.93	6.08
E9	DARS w/ CPO + style	0.7	0.3	**4.29**	**4.46**	**4.61**

Impact of Modeling Mechanisms on Synthesis Quality. Table III presents the MCD results on the validation set for synthesized speech generated by TTS models under three training strategies: ASp, DSpG, and SSp. E4 corresponds to the results from Grad-TTS. E5 serves as the baseline system using the original Matcha-TTS. E6 incorporates multi-stage rhythm prediction modules into E5. E7 and E8 further integrate the CPO strategy based on E6. E9 introduces acoustic style vectors into the CPO-enhanced models.

We observe that the ASp training strategy achieves the best performance, which can be attributed to its ability to utilize a larger amount of training data. Incorporating pause and duration modeling, together with the CPO strategy, significantly reduces the acoustic discrepancy between synthesized and real dysarthric speech. Furthermore, the integration of acoustic style vectors further improves synthesis quality by providing speaker-specific acoustic guidance. In the following experiments, we analyze the synthesized data based on the three training strategies derived from E9.

Effectiveness of Synthesized Speech in Enhancing ASR. Table IV compares the recognition performance of ASR systems trained on dysarthric speech synthesized by TTS models using different training strategies versus systems trained on real speech. We primarily adopt two adaptation strategies for E9-synthesized data: full-parameter fine-tuning and LoRA-based fine-tuning. It can be observed that full-parameter fine-tuning outperforms LoRA. This performance gap likely stems from the Whisper model's original pre-training on non-dysarthric speech. As a result, LoRA's localized parameter adaptation

TABLE IV
WER PERFORMANCE COMPARISON ACROSS DATA TYPES AND TRAINING STRATEGIES USING THE WHISPER-LARGE MODEL.

ID	Data Type	training strategy		WER (%)				
		FT	LORA	Severe	Mod.-Sev.	Moderate	Mild	Overall
E2	-			123.36	184.14	34.69	12.71	83.19
E10	RealData	✓		18.63	11.88	2.93	2.43	8.85
E11			✓	22.35	13.66	3.24	2.54	10.09
E12	ASp	✓		**18.64**	**11.88**	**2.94**	**2.44**	**8.87**
E13			✓	22.37	13.67	3.24	2.56	10.12
E14	DSpG	✓		21.06	13.07	3.15	2.50	9.96
E15			✓	24.23	15.45	3.21	2.54	11.32
E16	SSp	✓		22.18	13.55	3.21	2.54	10.31
E17			✓	26.10	16.64	3.30	2.61	12.19

TABLE V
WER COMPARISON WITH SOTA SYSTEMS.

ID	DataSet	WER (%)				
		Severe	Mod.-Sev.	Moderate	Mild	Overall
E18 [9]	TORGO	55.88	49.6	36.8	12.6	39.2
E19 [11]	TORGO	23.3	13.98	3.27	2.57	16.93
E12	TORGO	18.64	11.88	2.94	2.44	8.87
E20	TORGO+ LibriSpeech Text	**13.98**	**9.79**	**2.62**	**2.31**	**7.75**

shows limited modeling capacity when applied to dysarthric speech, which exhibits significant distributional differences from typical speech patterns.

It can be observed that the ASp training strategy (E12) generates dysarthric speech that closely matches real speech in acoustic characteristics and demonstrates strong generalization capability. ASR models trained on E12 achieve recognition performance comparable to those trained on real speech (E10) across all dysarthria severity levels. Moreover, compared to the baseline model (E2), E12 achieves an 89.33% relative reduction in Overall WER. Notably, greater performance gains are observed for more severe impairment categories, highlighting the enhanced effectiveness and adaptability of DARS in addressing significant speech impairments.

Comparison with SOTA Systems and DARS Model. Table V presents a comparison of recognition performance across different severity levels on the TORGO dataset between DARS and state-of-the-art (SOTA) systems. In E18 [9], the TTS model is FastSpeech2 [32] and the ASR model is a DNN-HMM [33] system. In E19 [11], the TTS model is Grad-TTS [13] and the ASR model is Whisper.

To evaluate DARS's generalization ability, we perform dysarthric speech synthesis using text from the LibriSpeech dataset, with reference speech randomly selected from TORGO. After adapting the ASR model using the augmented dysarthric speech (E20), we achieve an Overall WER of 7.75%, representing a 54.22% relative WER reduction over E19.

V. CONCLUSIONS

This study demonstrates that fine-tuning the Whisper model with DARS-synthesized speech achieves significant WER reductions across all dysarthria severity levels, with a relative reduction of 93.54% in Mod.-Sev. cases. Compared to advanced systems combining Grad-TTS and Whisper, our approach delivers a 54.22% relative WER reduction, demonstrating superior prosodic adaptation and generalization capabilities. Cross-corpus synthesis experiments using LibriSpeech text further confirm the robustness of our method in out-of-domain scenarios. These results indicate that joint modeling of pathological rhythm and acoustic styles not only enhances the realism and controllability of synthetic speech, but also significantly improves the effectiveness and generalizability of data augmentation for dysarthric speech recognition.

REFERENCES

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

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

[3] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition.," in *Interspeech*, 2018, pp. 471–475.

[4] Z. Qian and K. Xiao, "A survey of automatic speech recognition for dysarthric speech," *Electronics*, vol. 12, no. 20, p. 4278, 2023.

[5] M. B. Mustafa, F. Rosdi, S. S. Salim, and M. U. Mughal, "Exploring the influence of general and specific factors on the recognition accuracy of an asr system for dysarthric speaker," *Expert Systems with Applications*, vol. 42, no. 8, pp. 3924–3932, 2015.

[6] H. P. Rowe, S. E. Gutz, M. F. Maffei, K. Tomanek, and J. R. Green, "Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective," *Frontiers in computer science*, vol. 4, p. 770 210, 2022.

[7] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[8] C. Bhat and H. Strik, "Speech technology for automatic recognition and assessment of dysarthric speech: An overview," *Journal of Speech, Language, and Hearing Research*, vol. 68, no. 2, pp. 547–577, 2025.

[9] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing dysarthric speech using multi-speaker tts for dysarthric speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7382–7386.

[10] D. Wagner, I. Baumann, N. Engert, *et al.*, "Personalized fine-tuning with controllable synthetic speech from llm-generated transcripts for dysarthric speech recognition," *arXiv preprint arXiv:2505.12991*, 2025.

[11] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, "Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis," *arXiv preprint arXiv:2406.08568*, 2024.

[12] K. El Hajal, E. Hermann, A. Kulkarni, and M. M. Doss, "Unsupervised rhythm and voice conversion of dysarthric to healthy speech for asr," in *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, IEEE, 2025, pp. 1–5.

[13] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International conference on machine learning*, PMLR, 2021, pp. 8599–8608.

[14] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2595–2605.

[15] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-tts: A fast tts architecture with conditional flow matching," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 341–11 345.

[16] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[19] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[20] Z. Ye, W. Xue, X. Tan, J. Chen, Q. Liu, and Y. Guo, "Comospeech: One-step speech and singing voice synthesis via consistency model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1831–1839.

[21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[22] M. S. Albergo and E. Vanden-Eijnden, "Building normalizing flows with stochastic interpolants," *arXiv preprint arXiv:2209.15571*, 2022.

[23] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 12, pp. 2649–2658, 2012.

[24] H. Ma, Z. Peng, M. Shao, J. Li, and J. Liu, "Extending whisper with prompt tuning to target-speaker asr," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 516–12 520.

[25] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource asr," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 29, 2024.

[26] 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.

[27] Y. Wang, D. Stanton, Y. Zhang, *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 5180–5189.

[28] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 10 970–10 983.

[29] 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, pp. 523–541, 2012.

[30] 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.

[31] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion.," in *SLTU*, 2008, pp. 63–68.

[32] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[33] L. Li, Y. Zhao, D. Jiang, *et al.*, "Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition," in *2013 Humaine association conference on affective computing and intelligent interaction*, IEEE, 2013, pp. 312–317.