



Facilitating Personalized TTS for Dysarthric Speakers Using Knowledge Anchoring and Curriculum Learning

Yejin Jeon¹, Solee Im¹, Youngjae Kim¹, Gary Geunbae Lee^{1,2}

¹GSAI, POSTECH, South Korea

²CSE, POSTECH, South Korea

jeonyj0612@postech.ac.kr, solee0022@postech.ac.kr, yj122198@postech.ac.kr,
gblee@postech.ac.kr

Abstract

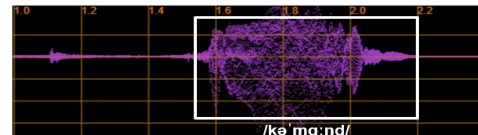
Dysarthric speakers experience substantial communication challenges due to impaired motor control of the speech apparatus, which leads to reduced speech intelligibility. This creates significant obstacles in dataset curation since actual recording of long, articulate sentences for the objective of training personalized TTS models becomes infeasible. Thus, the limited availability of audio data, in addition to the articulation errors that are present within the audio, complicates personalized speech synthesis for target dysarthric speaker adaptation. To address this, we frame the issue as a domain transfer task and introduce a knowledge anchoring framework that leverages a teacher-student model, enhanced by curriculum learning through audio augmentation. Experimental results show that the proposed zero-shot multi-speaker TTS model effectively generates synthetic speech with markedly reduced articulation errors and high speaker fidelity, while maintaining prosodic naturalness.

Index Terms: personalized speech synthesis, speech disorders, domain transfer

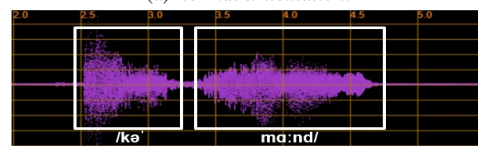
1. Introduction

Dysarthric speech emerges from a range of etiological factors, including cerebrovascular incidents such as strokes, as well as neuromuscular disorders linked to multiple sclerosis, cerebral palsy, and Parkinson’s disease [1, 2, 3]. These conditions compromise the neuromuscular control of speech production mechanisms, resulting in slurred, unintelligible, and phonetically distorted speech, which poses significant barriers to verbal communication. Given that speech is a fundamental medium for self-expression and social interaction, individuals with dysarthria frequently experience frustration, social isolation, and a diminished quality of life [4]. In response to these communication challenges, research has taken two primary directions: improving comprehension from the listener’s perspective, and enhancing speech production from the speaker’s perspective. Most existing research has prioritized the former by conducting research focusing on the enhancement of automatic speech recognition (ASR) systems so as to facilitate more effective communication by enabling caregivers, medical professionals, and downstream assistive technologies to interpret dysarthric speech with greater accuracy [5, 6, 7, 8].

While interpretation of affected speech is important, it is also crucial to empower dysarthric individuals by enhancing their ability to produce more intelligible speech and enhance their quality of life. Given that articulation deficits are a defining characteristic of dysarthria, early speaker-centric approaches sought to improve intelligibility by substituting impaired speech segments with unaffected ones [9, 10, 11]. However, these approaches resulted in the loss of the speaker’s



(a) Normal articulation.



(b) Dysarthric articulation.

Figure 1: Visualizations of the articulation differences between a highly intelligible female speaker and a low-intelligibility dysarthric speaker. The dysarthric speaker exhibits elongated speech production (2.0s vs. 0.6s) and segmented articulation.

unique vocal identity as the average model voice was utilized. To mitigate this issue, voice banking emerged as an alternative, allowing individuals to pre-record speech samples before the onset of dysarthria for later reconstruction of their natural voice [12, 13]. Although effective for progressive degenerative conditions, this method is infeasible for individuals with sudden-onset dysarthria (e.g., stroke) as they lack pre-recorded samples.

Multi-speaker text-to-speech (TTS) synthesis offers an efficient alternative by generating speech from textual input while simultaneously cloning the speaker’s voice from reference audio. However, training such models from scratch or even employing few-shot learning using dysarthric speech presents significant challenges. Dysarthric speech often contains articulation errors, which if used as training data, result in models that produce unstable and unintelligible synthetic speech. Additionally, TTS models typically require large-scale, high-quality training datasets, which makes it impractical to rely solely on dysarthric recordings. To circumvent these issues, prior works have explored hybrid approaches that sequentially utilize a single-speaker TTS and voice conversion model [14, 15].

In order to address the aforementioned limitations that are inherent in few-shot and from-scratch approaches, we focus on zero-shot multi-speaker TTS [16, 17] since just a single reference recording of the target speaker is required to conduct voice-preserving speech synthesis [16, 17]. Yet, extending this approach to the domain of dysarthric speech is not straightforward because there exists a fundamental mismatch between the audio samples used for model training and inference. This domain discrepancy is twofold; not only do variations in articulation exist, which range from highly intelligible to less intelligi-

ble speech, but dysarthric speech samples are often limited to single-word utterances [18] due to the speakers’ physical difficulties in speech production (Figure 1). This setting is in stark contrast to the much longer and well-articulated sentences typical of standard TTS datasets [19, 20]. Consequently, this challenge can be reframed as a zero-shot domain transfer problem, wherein the model must immediately and effectively extract speaker-specific vocal characteristics while remaining robust to articulation distortions in the input reference audio, which differ significantly from those encountered in the training data.

To resolve the dual domain transfer challenge seen in zero-shot multi-speaker TTS for dysarthric speakers, we take inspiration from pedagogical and physical therapy rehabilitation paradigms, where an expert plays a fundamental role in *guiding* a learner in the acquisition of a skill. To the best of our knowledge, we propose the first teacher-student framework for this task, in which a teacher model generates an anchored representation that stabilizes learning and guides a student model throughout a structured learning process. Furthermore, a key aspect of this structured training is its adherence to a curriculum learning strategy, where the student model is gradually introduced to progressively shorter inputs. As a result, the student model learns to generalize effectively to the short and highly variable nature of dysarthric speech during inference. In essence, our methodology enables the model to disentangle speaker identity from speech articulation distortions, thereby facilitating the generation of highly intelligible speech that retains the distinct voice of the target dysarthric speaker. The efficacy of our approach is substantiated through both objective and subjective evaluations, and demonstrates its potential to enhance personalized communication for dysarthric individuals.

2. Methodology

Our multi-speaker TTS model is composed of three main components: a module for generating a speaker representation from reference audio, a backbone text-to-mel-spectrogram acoustic model, and a vocoder that converts the mel-spectrogram into audio. The backbone acoustic model adheres to the FastSpeech2 architecture [21], which consists of a text encoder, a variance adaptor, and a decoder. To align with the objectives of multi-speaker TTS, we integrate the speaker representation that has been produced by the speaker encoder into the backbone TTS text encoder and decoder as in [22]. Moreover, HiFi-GAN is employed as the pretrained vocoder [23] for mel-to-audio generation. Given that the ultimate goal is to achieve zero-shot extraction and utilization of speaker representations from dysarthric reference audio, we focus on the speaker encoder, which is structured around two key components: knowledge anchoring by a teacher for the student model, and curriculum learning specifically designed for the student model (Figure 2).

2.1. Knowledge Anchoring

Unlike conventional multi-speaker TTS systems, where the reference audio input during inference remains within the same domain as the training data, the task at hand requires effective extraction of timbre-specific features while remaining resilient to articulation distortions, as well as the scarcity of speaker information inherent in short audio segments at inference time. However, attempting to simultaneously address this dual-task objective using a single speaker encoder as in typical multi-speaker TTS systems [24, 25, 26], is difficult. Therefore, to simplify this task, we separate it by introducing two collaborative

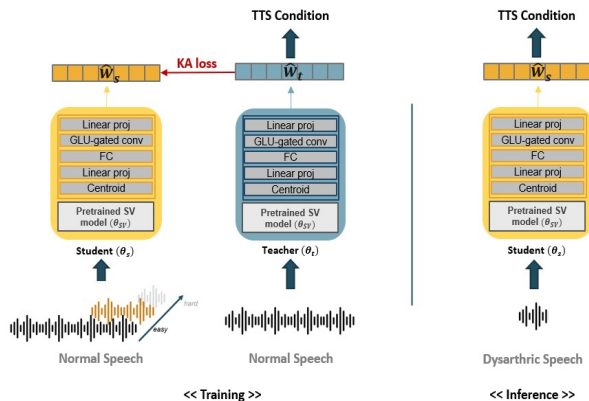


Figure 2: *Training and inference processes. During training, the teacher model conditions the backbone TTS model while serving as an anchor for the student model. The student model is trained using curriculum learning.*

teacher-student modules. A teacher speaker encoder initially captures the comprehensive information from the reference audio; this not only serves as an *anchor* for the student module to learn from, but also as a preliminary *filter* that removes irrelevant acoustic properties. Both teacher θ_t and student θ_s models share an identical architecture, where the mel-spectrogram of the input audio $mel_{i,j}$ is passed through a speaker verification¹ θ_{SV} that is pretrained to compute the average embedding of each speaker. Formally, this is denoted as

$$e_{i,j} = \theta_{SV}(mel_{i,j}) \quad (1)$$

$$c_i = \frac{1}{M} \sum_{j=1}^M e_{i,j} \quad (2)$$

where c_i is the centroid for speaker i , and $e_{i,j}$ denotes the embedding of the j -th utterance from speaker i . The resulting embedding $c \in \mathbb{R}^{256}$ is then used as the input to pass through a linear projection layer, a fully connected (FC) block with Mish activation [28], and a GLU-gated convolution stack with residual connections [29]. Another FC layer and temporal averaging results in two N -dimensional speaker style vectors $\hat{w}_t = \{y_1, y_2, \dots, y_N\}$ and $\hat{w}_s = \{x_1, x_2, \dots, x_N\}$ from each of the teacher and student modules, respectively. The following knowledge anchoring relationship is subsequently formed:

$$\mathcal{L}_{MAE}(\hat{w}_s, \hat{w}_t) = \frac{1}{N} \sum_{n=1}^N \|x_n - y_n\|_1 \quad (3)$$

2.2. Curriculum Learning (CL)

Simply training the teacher and student models as described in the previous subsection presents two key challenges: (1) there is no direct connection between the training and inference domains, and (2) a trivial solution can emerge since both models share the same architecture and receive identical input data. To mitigate this, the student model is progressively trained with shorter segments of audio compared to the teacher model, enforcing a non-trivial learning process. Formally, let S denote the total number of training steps and C the number of progressive

¹<https://github.com/resemble-ai/Resemblyzer>

Algorithm 1: Multi-view Augmentation for CL

Input : Audio \mathcal{A} , Total Training Steps S , Number of Cropping Stages C , Current Step s

Output: Student Representation \hat{w}_s

Step 1: Compute Cropping Parameters

Define steps of each cropping stage : $S_c = S/C$;

Compute cropping ratios for each stage:

$$r = \left[1 - \frac{1}{C+1}, 1 - \frac{2}{C+1}, \dots, 1 - \frac{C}{C+1}\right];$$

Step 2: Progressive Audio Cropping

stage $\leftarrow \lfloor \frac{s}{S_c} \rfloor$;

$\mathcal{A}' \leftarrow \mathcal{A}[: r[\text{stage}] \times \text{len}(\mathcal{A})]$;

return $\hat{w}_s \leftarrow \theta_s(\mathcal{A}')$;

cropping stages applied to the input audio. The number of training steps allocated to each cropped audio stage is then given by $\frac{S}{C}$. An overview of this multi-view audio augmentation process is provided in Algorithm 1.

Moreover, when training the entire TTS framework, the style representation \hat{w}_t that is generated by the teacher model is employed as the condition for the backbone TTS encoder and decoder, while at inference time, the style representation generated by the student model \hat{w}_s is utilized. Conditioning is implemented as in [22], where text input feature of the backbone text encoder $P_{text} = \{p_1, p_2, \dots, p_K\}$ of K -dimensions is fused with style representation w using its gain g and bias b .

$$\mu = \frac{1}{H} \sum_{k=1}^H p_k, \sigma = \sqrt{\sum_{k=1}^H (p_k - \mu)^2} \quad (4)$$

$$N(p) = \frac{p - \mu}{\sigma} \quad (5)$$

$$Fusion(P_{text}, \hat{w}) = g(\hat{w}) \odot N(p) + b(\hat{w}) \quad (6)$$

The total loss is defined as a combination of the mel reconstruction loss between the synthesized audio and the original full reference audio used for the teacher model, and the loss between the style vectors that were computed by each of the teacher and student speaker encoders.

$$\mathcal{L}_{Total} = \mathcal{L}_{MAE}(\hat{mel}, mel) + \mathcal{L}_{MAE}(\hat{w}_s, \hat{w}_t) \quad (7)$$

3. Experimental Settings

To validate our proposed methodology, we compare its performance against three zero-shot baseline models. The first employs discriminators trained with style prototypes and episodic training [22]. While originally designed for normal speech, its method was also intended to generalize to unseen speakers, even those of short audio samples. On the other hand, the second and third baselines specifically target dysarthric speech: the former extracts a speaker representation using a pretrained speaker verification model [30], while the latter adopts a hybrid approach akin to [14]. In this framework, a single-speaker Fast-Pitch [31] model generates speech corresponding to the target input text, and then re-synthesized to match the target dysarthric speaker’s voice using the FreeVC [32] voice conversion model.

To ensure fair comparisons, all experiments were conducted in identical conditions. Training was performed on a single

Table 1: Comparisons with Adaptive [22], Conditional [30], and Hybrid [31, 32] baseline models. 95% confidence intervals are reported for MOS.

Model	Obj. Metrics		Subj. Metrics	
	PER (↓)	Spk Sim (↑)	MOS-Nat	MOS-Spk
Adaptive	64.455	0.570	2.908 ± 0.23	2.753 ± 0.24
Conditional	47.696	0.647	2.839 ± 0.20	2.906 ± 0.21
Hybrid	31.017	0.534	3.371 ± 0.15	3.731 ± 0.22
Proposed	14.254	0.619	3.601 ± 0.18	3.909 ± 0.21

Table 2: Subjective evaluations across different dysarthric speaker intelligibility (Int.) groups.

		Very Low Int.	Low Int.	Middle Int.	High Int.
MOS-Spk	Adaptive	2.722	3.053	3.000	3.000
	Conditional	2.789	2.947	2.895	2.763
	Hybrid	3.293	3.395	3.461	3.395
	Proposed	3.338	3.763	3.618	3.882
MOS-Nat	Adaptive	2.789	2.711	2.789	2.697
	Conditional	2.932	2.947	2.829	2.895
	Hybrid	3.699	3.737	3.763	3.750
	Proposed	3.692	4.066	3.869	4.171

GPU for the proposed 33M-parameter model using the Librispeech [19] dataset with multi-view audio augmentation as described in Section 2.2. Specifically, audio cropping was conducted three times over 500,000 training steps (i.e., approximately every 160,000 steps). Zero-shot synthesis was conducted on the UASpeech dysarthria dataset [18] and evaluated by employing both subjective and objective metrics. MOS ratings were gathered from 19 Amazon Mechanical Turk workers who assessed naturalness in terms of prosodic intonation, and vocal similarity. Moreover, objective evaluations were conducted with phoneme error rate (PER), and speaker similarity, which was computed using cosine similarity between synthesized and ground truth speech representations as in prior research [33].

4. Results and Discussion

4.1. Main Results

The results presented in Table 1 underscore the efficacy of the proposed methodology in mitigating phonemic articulation errors while maintaining high speaker fidelity. PER demonstrates that the proposed methodology achieves a substantial reduction of over 50 points compared to the Adaptive [22] baseline, and over 15 points compared to highest performing baseline model. Notably, this improvement of articulation accuracy does not come at the cost of speaker similarity, as the proposed model maintains a high speaker similarity score of 0.619, which is comparable to or surpasses existing approaches. Additionally, subjective evaluations reveal that the proposed method achieves the highest perceived naturalness in terms of intonation and pronunciation (MOS-Nat: 3.601) and speaker similarity (MOS-Spk: 3.909), which reinforces its effectiveness in preserving both articulation clarity and speaker identity.

Beyond overall performance, Table 2 provides a breakdown of subjective evaluations across dysarthric speaker groups categorized by different intelligibility levels (Very Low, Low, Middle, and High). The proposed model consistently outperforms other approaches in speaker similarity across all intelligibility groups. In terms of naturalness, while the proposed model shows marginally lower scores than that of the Hybrid model

Table 3: Ablation studies on Knowledge Anchoring and Curriculum Learning. ‘Stu’ denotes the student model, while ‘w/out CL’ indicates direct training on the shortest audio inputs. CL is applied to the student model when present; otherwise, it is applied to the teacher model.

Model	Teacher		Student	
	PER (↓)	Spk Sim (↑)	PER (↓)	Spk Sim (↑)
w/out Stu. w/out CL	26.428	0.618	-	-
w/out Stu. w/ CL	22.846	0.623	-	-
w/ Stu w/out CL	21.935	0.647	15.579	0.613
w/ Stu. w/ CL	20.559	0.646	14.254	0.619

Table 4: Student-Teacher comparisons across different dysarthric speaker groups.

Model	Teacher		Student	
	PER (↓)	Spk Sim (↑)	PER (↓)	Spk Sim (↑)
Very Low Int.	20.528	0.586	15.157	0.556
Low Int.	19.463	0.616	15.586	0.584
Middle Int.	23.314	0.643	15.243	0.610
High Int.	18.668	0.708	12.266	0.690
Average	20.559	0.646	14.254	0.619

for the Very Low category, the difference is only 0.007. Moreover, it can be seen that the proposed model surpasses the other baseline models for all remaining intelligibility levels.

4.2. Knowledge Anchoring and CL

The importance of the knowledge anchoring framework can be assessed by comparing models trained with and without the student model while maintaining curriculum learning in its most restrictive form (w/out CL in Table 3). Note that this setting does not completely eliminate curriculum learning. Rather, it represents a scenario in which the speaker encoder is trained directly using the shortest audio from the multi-view augmentation process, and bypasses the gradual transition from longer to shorter references. Under this condition, sole utilization of the teacher model as the TTS condition yields a high phoneme error rate (PER) of 26.428 (w/out Stu. w/out CL). However, introducing the student model (w/ Stu. w/out CL) leads to a substantial reduction in PER to 15.579, which marks an improvement of 10.849 points. Comparable speaker similarity scores are also retained between the teacher and student models. A plausible explanation for these improvements is that the teacher model truly functions as a filter, which enables the student model to focus more effectively on speaker-specific timbre attributes. Thus, this reduces content leakage, since target timbre features do not inadvertently blend with miscellaneous acoustic factors, which would otherwise interfere with the target text’s pronunciation when conditioning the backbone TTS model. These findings highlight the necessity of distinct teacher and student learning processes for enhancing phonemic articulation.

While the teacher-student framework alone offers substantial benefits, its effectiveness is further amplified by curriculum learning. By structuring training such that the student speaker encoder initially learns from longer, more informative references before adapting to shorter, more constrained inputs, this gradual adaptation results in a PER that decreases from 26.428 (w/out Stu. w/out CL) to 22.846 (w/out Stu. w/ CL) for the teacher-only model, and from 15.579 (w/ Stu. w/out CL) to 14.254 (w/ Stu. w/ CL) for the knowledge-anchored model. Speaker similarity also increases from 0.613 to 0.619 for the

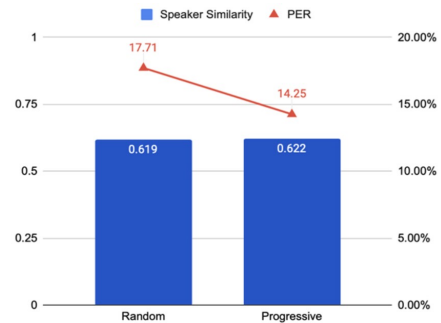


Figure 3: Variations of multi-view augmentation. Incrementally shortening audio significantly reduces PER while preserving speaker similarity comparable to the Random setting.

student speaker encoder. To further analyze model performance across dysarthric speaker groups, Table 4 presents results stratified by intelligibility level. A clear trend emerges; both teacher and student models achieve lower PER and higher speaker similarity as intelligibility increases. When using the teacher model to condition the backbone TTS, PER decreases from 20.528 (Very Low Intelligibility) to 18.668 (High Intelligibility), with speaker similarity improving from 0.586 to 0.708. Similarly, the student model’s PER drops from 15.157 to 12.266, with speaker similarity rising from 0.556 to 0.690. Notably, across all intelligibility levels, the student model consistently yields the lowest articulation errors. These results align well with the fundamental goal of speech synthesis, which is to foremost produce speech with accurate pronunciation. These trends are also intuitive, as higher intelligibility speakers exhibit fewer articulation distortions that obscure speaker-specific characteristics, which in turn leads to easier timbre extraction by the speaker encoder.

4.3. Multi-view Audio Augmentation for CL

In the proposed approach, the student model is trained using progressively shorter audio inputs to facilitate structured adaptation. To evaluate the efficacy of this augmentation strategy, we conduct an additional experiment using a randomized setting, where input audio utilized for the student speaker encoder is randomly cropped at each training step using one of the predefined cropping ratios. As shown in Figure 3, the proposed progressive strategy leads to a substantial reduction in PER while preserving speaker similarity. Although the randomized setting exhibits a marginal increase in speaker similarity of 0.003, this is likely attributable to the greater proportion of longer audio inputs encountered during training compared to the structured progressive setting that was originally utilized.

5. Conclusion

In this paper, we present a zero-shot multi-speaker TTS approach to enhance personalized communication for dysarthric speakers. Unlike conventional single-encoder models, we introduce a knowledge anchoring framework with dual speaker encoders. Additionally, multiple views of the input reference audio is constructed and employed incrementally for curriculum learning. This approach effectively addresses the dual challenge of mitigating articulation artifacts while extracting robust speaker representations from minimal speech input. Comprehensive evaluations confirm that the synthesized speech remains highly intelligible, natural, and speaker-consistent.

6. Acknowledgements

This research was supported by Smart HealthCare for Police Officers Program(www.kipot.or.kr) through the Korea Institutes of Police Technology(KIPoT) funded by the Korean National Police Agency(KNPA, Korea)(No. RS-2022-PT000186; 47.5%), by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2024-00437866; 47.5%), and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH); 5%).

7. References

- [1] D. Mulhari, G. Meoni, M. Marini, and L. Fanucci, "Machine learning assistive application for users with speech disorders," in *Applied Soft Computing*, 2021.
- [2] F. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," in *Journal of Speech and Hearing Research*, 1969.
- [3] Y. Yunusova, G. Weismer, J. R. Westbury, and M. J. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," in *Journal of Speech and Hearing Research*, 2021.
- [4] J. Mertl, E. Žáčková, and B. Řepová, "Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis," in *Disabil Rehabil Assist Technol.*, 2018.
- [5] S. Leivaditi, T. Matsushima, M. Coler, S. Nayak, and V. Verkhdanova, "Fine-Tuning Strategies for Dutch Dysarthric Speech Recognition: Evaluating the Impact of Healthy, Disease-Specific, and Speaker-Specific Data," in *Interspeech*, 2024.
- [6] I.-T. Hsieh and C.-H. Wu, "Dysarthric Speech Recognition Using Curriculum Learning and Articulatory Feature Embedding," in *Interspeech*, 2024.
- [7] S. Wang, S. Zhao, J. Zhou, A. Kong, and Y. Qin, "Enhancing Dysarthric Speech Recognition for Unseen Speakers via Prototype-Based Adaptation," in *Interspeech*, 2024.
- [8] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, "Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric-Speech Synthesis," in *Interspeech*, 2024.
- [9] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based Speech Synthesis to Reconstruct the Voice of Individuals with Degenerative Speech Disorders," in *Interspeech*, 2024.
- [10] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, "Building personalised synthetic voices for individuals with severe speech impairment," in *Computer Speech & Language*, 2013.
- [11] Y. Wang, X. Wu, D. Wang, L. Meng, and H. Meng, "UNIT-DSR: Dysarthric Speech Reconstruction System Using Speech Unit Normalization," in *ICASSP*, 2024.
- [12] H. T. Bunnell, J. Lilley, C. Pennington, B. Moyers, and J. Polikoff, "The ModelTalker System," in *The Blizzard Challenge*, 2010.
- [13] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," in *Acoustical Science and Technology*, 2012.
- [14] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, and Y. Shiga, "High-Intelligibility Speech Synthesis for Dysarthric Speakers with LPCNet-Based TTS and CycleVAE-Based VC," in *ICASSP*, 2021.
- [15] R. Nanzaka and T. Takiguchi, "Hybrid Text-to-Speech for Articulation Disorders with a Small Amount of Non-Parallel Data," in *APSIPS Annual Summit and Conference*, 2018.
- [16] Y. Jeon, Y. Kim, and G. G. Lee, "Enhancing Zero-Shot Multi-Speaker TTS with Negated Speaker Representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [17] B. J. Choi, M. Jeong, J. Y. Lee, and N. S. Kim, "SNAC: Speaker-Normalized Affine Coupling Layer in Flow-Based Architecture for Zero-Shot Multi-Speaker Text-to-Speech," in *IEEE Signal Processing Letters*, 2023.
- [18] H. Kim, M. Hasegawa-Johnson, drienne Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric Speech Database for Universal Access Research," in *Interspeech*, 2008.
- [19] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wum, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech*, 2019.
- [20] K. Ito and L. Johnson, "The LJ Speech Dataset," in <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [21] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *NeurIPS*, 2019.
- [22] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech : Multi-speaker adaptive text-to-speech generation," in *International Conference on Learning Representations (ICLR)*, 2021.
- [23] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *NeurIPS*, 2020.
- [24] N. Kumar and A. N. andBrejesh Lall, "Zero-Shot Normalization Driven Multi-Speaker Text to Speech Synthesis," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.
- [25] E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. G'olge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *Proceedings of Machine Learning Research*, 2022.
- [26] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, and N. Chen, "Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings," in *ICASSP*, 2020.
- [27] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *ICASSP*, 2018.
- [28] D. Misra, "Mish: A Self Regularized Non-Monotonic Activation Function," in *BMVC*, 2020.
- [29] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language Modeling with Gated Convolutional Networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML, 2017*.
- [30] K. Azizah, "Zero-Shot Voice Cloning Text-to-Speech for Dysphonia Disorder Speakers," *IEEE Access*, vol. 12, pp. 63 528–63 547, 2024.
- [31] A. Łańcucki, "FastPitch: Parallel Text-to-speech with Pitch Prediction," in *ICASSP*, 2021.
- [32] J. Lii, W. Tu, and L. Xiao, "Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion," in *ICASSP*, 2023.
- [33] Y. Zhou1, C. Song, X. Li, L. Zhang, Z. Wu, Y. Bian, D. Su, and H. Meng, "Content-Dependent Fine-Grained Speaker Embedding for Zero-Shot Speaker Adaptation in Text-to-Speech Synthesis," in *Interspeech*, 2022.