



Data Augmentation using Speech Synthesis for Speaker-Independent Dysarthria Severity Classification

Minseop Kim¹, Minsu Han¹, Seokyoung Hong¹, Myoung-wan Koo¹

¹Department of Artificial Intelligence, Sogang University, South Korea

jik2196@naver.com, minsu083@sogang.ac.kr, hsy960925@sogang.ac.kr, mwkoo@sogang.ac.kr

Abstract

Accurate dysarthria severity classification is essential for assessing motor speech disorders, and automation can improve efficiency and accessibility in clinical settings. While deep learning has significantly advanced this field, recent studies have increasingly leveraged large foundation ASR models. However, most studies focus on speaker-dependent (SD) classification, leaving speaker-independent (SI) classification as a major challenge due to limited datasets. SI classification is crucial in real-world scenarios where patient-specific information is unavailable. To address this, we applied two types of speech synthesis models for the first time in this task. We explore various strategies for integrating zero-shot text-to-speech (ZS-TTS) and voice conversion (VC) models to enhance SI classification and propose the most effective utilization settings. Our approach significantly improves the SI severity classification performance, paving the way for further research in this area.

Index Terms: Dysarthric Speech, Severity Classification, Speech Synthesis, zero-shot TTS

1. Introduction

Dysarthria is a motor speech disorder caused by neurological impairment, affecting articulation, phonation, resonance, prosody, and respiration. Recent advancements in automatic dysarthria severity classification have leveraged machine learning models with a cascaded structure of feature extraction followed by classification. Conventional approaches extract spectrum-based features, such as MFCCs, i-vectors, and Mel-spectrograms, and acoustic-based features related to prosody, articulation, and phonation. These extracted features serve as critical inputs for improving severity classification accuracy.

Recently, the emergence of large foundation models for Automatic Speech Recognition (ASR) has facilitated the use of self-supervised representations, such as Wav2vec2 [1] and HuBERT [2], for dysarthria severity classification. These self-supervised learning models have demonstrated significant improvements across various speech-related tasks, including severity classification [3]. Furthermore, the weakly supervised learning model Whisper[4] has achieved state-of-the-art (SOTA) performance in ASR, and features extracted from this model have shown superior effectiveness in dysarthria severity classification [5]. Additionally, custom-designed features, such as speech recognition-driven features [6] and language-unique features [7], have also been explored for severity classification.

In terms of classifiers, previous studies have implemented both machine learning (ML)-based and deep neural network (DNN)-based approaches [8, 9]. ML-based classifiers have primarily utilized methods such as support vector machines (SVM) and random forests (RF). In contrast, DNN classifiers

have mainly employed architectures including feed-forward networks, Convolutional Neural Networks (CNN), ResNet, etc. While these methods have demonstrated significant effectiveness in severity classification, most studies have focused on speaker-dependent (SD) classification. Due to the limited availability of large datasets, speaker-independent (SI) severity classification—where the speakers in the training and validation sets differ from those in the test set—remains a significant challenge. Speaker-independent severity classification is crucial, as patient-specific information is often unavailable in real-world scenarios.

Given the inherent limitation of dysarthria datasets, numerous studies have explored data augmentation techniques. Traditional approaches have employed methods such as temporal and speed modification in the spectral domain [10] and vocal tract length perturbation (VTLP) [11]. While these methods have improved performance, they struggle to capture the distinct characteristics of dysarthric speech. [12] combined these traditional approaches to enhance severity classification but did not demonstrate effectiveness in SI classification. Subsequently, generative models have been applied to dysarthric speech processing. GAN-based methods transformed normal speech into dysarthric speech [13] or synthesized dysarthric-like corpora via voice conversion [14]. Diffusion-based approaches have also been explored: TTS diffusion models were used to enhance dysarthria ASR [15], while diffusion-based voice conversion (VC) models generated atypical speech [16]. Additionally, transformer-based TTS models [17] and end-to-end TTS models [18] have been studied for dysarthria ASR.

However, most studies on dysarthria data augmentation focus on ASR, with little research addressing SI severity classification. To overcome the challenges of limited datasets, we applied two types of speech synthesis models—ZS-TTS and VC models—to SI severity classification. As these models do not inherently preserve severity-specific features, we conducted extensive experiments to determine the most effective utilization strategies.

Our overall procedure and framework are described in Figure 1. Below are the key contributions of this paper:

- Propose a novel data augmentation framework designed to enhance SI dysarthria severity classification on imbalanced dataset.
- Explore various mixture ratios of augmented data and original data to optimize their integration for improving SI severity classification.
- Compare two different types of speech synthesis models to analyze their respective strengths and weaknesses, aiming to optimize their effective utilization for enhancing SI dysarthria severity classification.

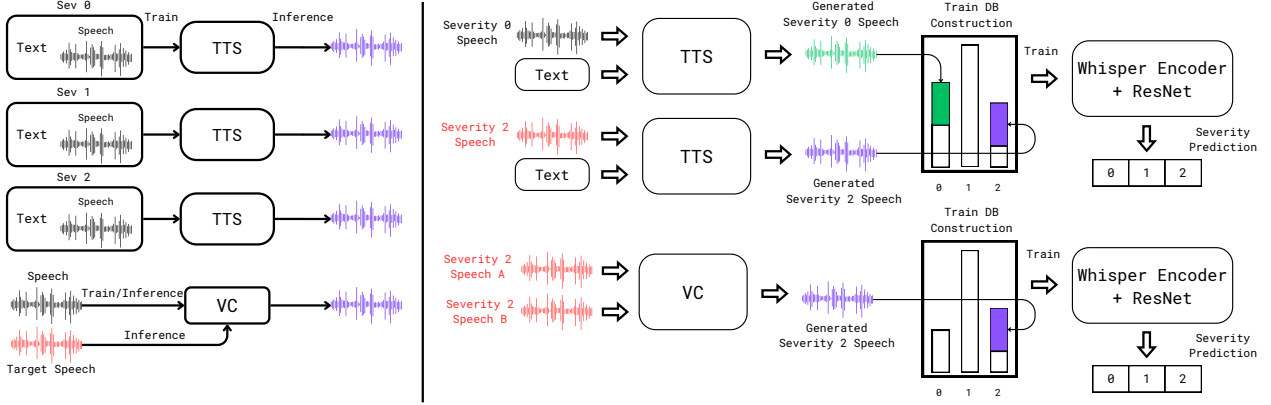


Figure 1: *Left* : Train and Inference procedure of Dysarthria Speech Synthesis models for data augmentation. The TTS model was fine-tuned exclusively on severity levels. *Right* : Our overall framework to enhance Dysarthria Severity Classification Model for an unbalanced dataset. For TTS, severity 0 and 2 data performed best, while for VC, only severity 2 data did.

2. Selecting Models for Classification and Speech data synthesis

2.1. Classification Model

For our classification model, we referred to [3, 5] as baselines. Both studies conducted extensive experiments on the TORGO dataset for severity classification. Specifically, [3] focused on SI severity classification and demonstrated that HuBERT is the most effective feature extraction module for this task. In contrast, [5] showed that Whisper achieved the best performance in SD classification.

To determine the optimal feature extraction module for SI classification, we conducted experiments on both SD and SI classification on TORGO dataset and SI classification on Korean dataset (described in Section 3). For the classification module, we adopted ResNet following [5], assuming that ResNet generally outperforms the simple CNN model used in [3].

For SI classification on the TORGO dataset, we designed a simplified version of the experiment conducted in [3]. Specifically, instead of performing 18 iterations of leave-each-speaker-out testing per severity level, we conducted only 4 iterations. Additionally, we fixed speakers M03, M05, and M04 from each severity level as part of the training set. To compare with similar parameter size, we used base model for Wav2vec2 and HuBERT. As a result, we confirmed Whisper as the best model for the feature extraction module, which showed best performance in every settings. Our following experiments on classification were conducted based on Whisper + ResNet structure. The composition of ResNet was simplified following [5]. The results of the comparison are shown in Table 1.

Table 1: *Selection of classification models for SI severity classification on Korean Dataset. Metric: Balanced Accuracy*

Classification Model	TORGO		Korean
	SD	SI	SI
Wav2vec2 (base) + ResNet	98.75	59.76	62.90
HuBERT (base) + ResNet	98.74	60.74	66.39
Whisper (Small) + ResNet	99.73	62.40	68.50

2.2. Speech Synthesis Models

We selected xTTS v2.0[19] for our TTS model due to its support for Korean and zero-shot TTS capability, which enables ef-

ficient fine-tuning even with a limited dataset. Following the official guidelines from the xTTS GitHub[20] repository, we fine-tuned only the GPT-2 encoder to effectively model dysarthric speech characteristics. To ensure the generated speech accurately reflects the distinct articulatory patterns associated with different severities of dysarthria, we fine-tuned the model separately for each severity level in the Korean dysarthric dataset.

For the voice conversion model, we selected Hierspeechpp[21] which has shown state-of-the-art results on English and Korean. Hierspeechpp has been trained on a mixture of English and Korean data, for 2,700 hours. Based on official github, which offered only inference code, we reconstructed train code for fine-tune. Compared to xTTS, we trained all of our Korean dysarthric dataset on a single model and conducted generation for each severity.

3. Experiment

In this study, we conducted experiments on a Korean dysarthric speech dataset with two speech synthesis models to improve classification performance through speech augmentation. For this task, we employed the Whisper-ResNet model for dysarthria severity classification, as it has demonstrated state-of-the-art (SOTA) performance in SD and SI severity classification tasks at previous section.

Table 2: *Detailed Composition of the Korean Dysarthric Dataset. Total Dataset is about 11 hours with 371 speakers.*

Severity	w/o Dysarthria	Mild-to-Moderate	Severe	Total
DDK (sec)	2492	9403	1589	13483
Sentence (sec)	2347	20408	3146	25901
# of Male	30	182	18	230
# of Female	29	102	10	141

3.1. Dataset

The Korean dysarthric speech dataset is based on the dysarthria data collected on [22], and we collected extra data for Diadochokinetic(DDK) tasks. Our new dataset consists of recordings from dysarthric speakers performing two tasks: DDK rate assessment and sentence reading. Each speaker produced four DDK utterances corresponding to "pa," "ta," "ka," and "pa-ta-ka" repetitions, along with six standardized sentences commonly used for speech assessment in Korean dysarthric patients.

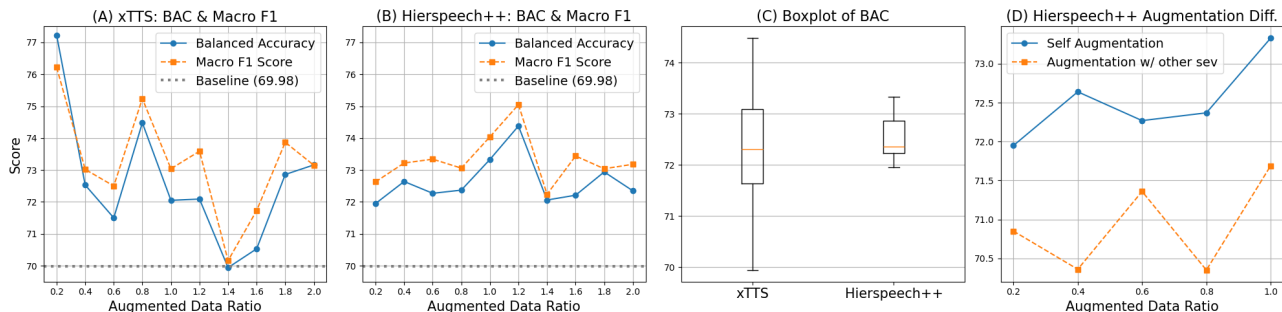


Figure 2: Analysis of Experimental Results by Augmentation Method. Figures (A) and (B) illustrate the Balanced Accuracy (BAC) and Macro F1 Score for different ratios of augmented data to real data. (A) shows the results using xTTS for data augmentation, while (B) presents those with Hierspeech++. (C) presents a boxplot comparing the performance of different augmentation methods. (D) compares two approaches to data augmentation using Hierspeech++: Self Augmentation, where both the source and target are fixed as severity 2, and Augmentation with other severity, where the source is fixed as severity 2 while severity 0 and 1 are used as targets.

These six sentences were carefully designed to encompass all Korean consonants and vowels based on their frequency of occurrence. In total, each patient contributed ten speech samples.

The severity levels in this dataset were evaluated and assigned by professional speech-language pathologists based on the NIH Stroke Scale, ensuring a clinically reliable classification of dysarthric severity. However, a notable characteristic of this dataset is class imbalance, where the number of samples in severity level 0 (sev. 0), level 1 (sev. 1), and level 2 (sev. 2) varies significantly. Throughout this paper, we consistently refer to non-dysarthric speech as sev. 0, mild-to-moderate dysarthria as sev. 1, and severe dysarthria as sev. 2 for clarity and consistency. This imbalance poses challenges in model training and evaluation, which we address through augmentation strategies. Details of our dataset is described in Table 2.

In addition to the Korean dataset, the TORGO dataset, a widely used resource in dysarthria research, also includes DDK and sentence reading tasks. However, due to limitations in speaker diversity and dataset composition, this study focuses exclusively on the Korean dataset for our experiments.

3.2. Experimental Setup

3.2.1. xTTS setup

The corpus is split into training and validation sets at a 0.85:0.15 ratio for each severity level (Table 2). Training uses the AdamW optimizer (learning rate: $5e-6$, weight decay: $1e-2$) with a MultiStepLR schedule, decaying by $1e-2$ at [90,000, 2,700,000, 5,400,000] steps. The batch size is 8 (batch group: 48), with evaluation also using a batch size of 8. Gradient accumulation is set to 1, and weight decay applies only to weights in multi-GPU training. The model was trained for 30 epochs on four NVIDIA A100 GPUs (80GB each).

3.2.2. Hierspeechpp setup

Following the original paper, we fine-tuned Hierspeechpp using AdamW optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay $\lambda = 0.01$. We applied learning rate schedule by the decay of $0.999^{1/8}$ with an initial learning rate of $1e-4$ for Korean dysarthric dataset, and the batch size was 20 with 1,500k steps on two A100 GPUs. Following original paper, we sliced the audio with 61,440 frames for training, and adopted windowed generator training for generator with 9600 frames.

3.2.3. Classification Model Setup

We follow most of the experimental settings from [5] for our classification model. While the prior work employs the SGD

optimizer, we adopt the AdamW optimizer for more stable convergence with default parameter settings. Additionally, to mitigate the impact of class imbalance, we apply the class weight option, which assigns higher weights to underrepresented classes to ensure they contribute more significantly during training. We train only ResNet with a batch size of 8 for 10 epochs on a single A100 GPU.

4. Data Augmentation using Speech Synthesis models

4.1. Training Strategies for Dysarthric Speech Synthesis

For VC model, we simply trained all of our Korean dysarthric speech, as we believe just controlling source and target speech at inference could successfully generate the speech of each severity. Conversely, training TTS models on both sev. 1 and sev. 2 speech introduced a data imbalance issue, which resulted in suboptimal generation of sev. 2 speech. Specifically, when trained on a mixture of sev. 1 and 2 dysarthric speech, the TTS model predominantly synthesized sev. 1 dysarthric speaker speech, struggling to generate sev. 2 dysarthric speaker speech effectively.

4.2. Data Generation using VC and TTS

In our VC-based data augmentation approach, we conducted experiments by fixing sev. 2 speech as the source while varying the target between sev. 0 and sev. 1. Similarly, we explored the opposite configuration, where sev. 2 was set as the target, and sev. 0 or sev. 1 was used as the source. However, in both cases, the generated speech failed to effectively preserve key dysarthric articulation characteristics, such as pauses and pronunciation errors, leading to suboptimal augmentation results. To address this issue, we adopted a self-augmentation strategy, where both the source and target speech were fixed as sev. 2 samples. This approach ensured that the generated speech retained dysarthric features more effectively, ultimately yielding superior augmentation performance.

For TTS-based data augmentation, we initially conducted experiments using sev. 2 dysarthric speech as the reference audio during inference. However, this approach resulted in poor output quality due to the inherent noisiness of sev. 2 speech. To address this issue, we experimented with using sev. 1 dysarthric speech as the reference audio while still generating sev. 2 dysarthric speech. This experiment demonstrated that using an excessively noisy reference in zero-shot TTS adversely affects

the model’s performance.

5. Classification Results and Analysis

5.1. Overall Performance and Model Comparison

Figure 2 illustrates the impact of our data augmentation on SI classification with respect to the baseline, which was trained on a dataset with severe data imbalance (sev. 0 and 2) using direct replication. As shown in Figure 2 (A) and (B), our proposed framework consistently outperformed the baseline.

To analyze the impact of the amount of augmented data compared to original data, we analyzed the classification performance across different original-to-augmented data ratios with both synthesis models. Figure 2 (A) shows that TTS-augmented classification demonstrated improvements over the baseline across different original-to-augmented data ratios. Although the overall trend showed high variance, we could observe gradually decreasing performance as the original-to-augmentation ratio grows up. This high variance is particularly evident in the 2 (C) box plot, where the xTTS model, under a 0.2 ratio condition, exhibited substantial variability across 10 fragments, as indicated by a high standard deviation. This suggests that while the TTS model can generate high-quality augmented data, it also produces inconsistent results.

As shown in Figure 2 (B), in all cases, the VC model outperformed the baseline. Regarding the variance of the quality of augmented data, the classification result gradually increases until the original-to-augmented data ratio reached 1:1, and gradually decreased after that. Notably, as observed in the Figure 2 (C) box plot, when the total augmented data was split in an isolated manner at a 0.2 ratio, the mean and standard deviation analysis indicated that the augmented data maintained consistent and high-quality characteristics.

Figure 3 illustrates detailed classification results in confusion matrix. As shown in upper left, our baseline struggled classifying sev. 2. Most of sev. 2 were classified as sev. 1. On the upper right side, with simple duplication on sev. 2 data, the results got slightly better, but didn’t have enough impact. On the other hand, lower side of the figure shows the best results of our approach, demonstrating the effectiveness of proposed augmentation method. Despite the slight decrease on sev. 1 classification, sev. 2 classification gets remarkably better. Consistent with our assumption, the data augmentation models couldn’t completely generate the speech of corresponding severity. However, the increase in low-severity data—despite inherently not being an exact match—still improved classification performance. To be specific, xTTS-augmented classification showed slightly better performance compared to approach using Hierspeechpp.

5.2. Inference Architecture and Quality Variability

The observed inconsistencies in TTS-based augmentation stem from differences in inference architecture and input constraints. Unlike the VC model, which directly utilizes source and target speech from sev. 2 dysarthric speakers, TTS models generate speech using a combination of reference audio and input text. Given the impracticality of obtaining human-annotated transcriptions from dysarthric patients, we relied on original text for fine-tuning and inference, which contributed to variability in data quality. Additionally, when sev. 1 dysarthric speech was used as reference audio in xTTS, the model exhibited high variance in outputs due to the wide spectral distribution of sev. 1, unlike more conventional datasets such as TORGO.

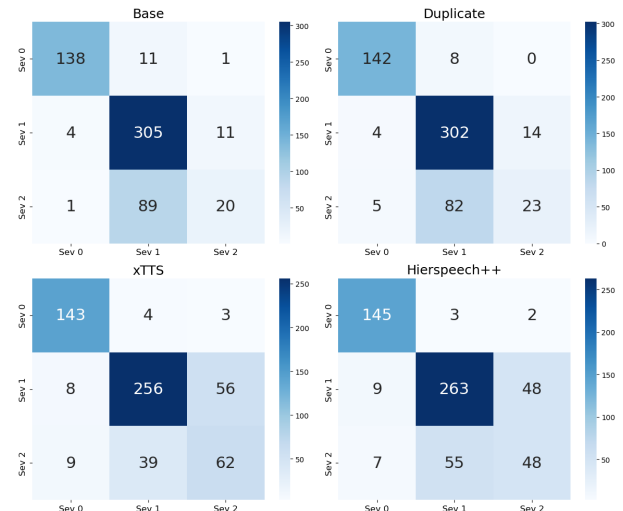


Figure 3: Confusion Matrix of severity classification result. Upper left: Baseline. Upper right: Simple duplication of sev. 2. Lower left and right: data augmented with xTTS and Hierspeechpp respectively.

While the VC model consistently maintains higher quality, its limitations become evident in Figure 2 (D). This figure compares the experimental results when sev. 0 or sev. 1 was set as the target speech with sev. 2 as the source speech, against the scenario where both source and target were sev. 2. As previously discussed in Section 4.2, the first approach exhibited challenges associated with preserving key dysarthric articulation characteristics. The key takeaway from these results is that while the VC model demonstrates consistent performance, its effectiveness is significantly influenced by the source and target speech characteristics. This dependency highlights the advantage of self-augmentation, where both the source and target are sev. 2, as it ensures that the generated speech retains essential dysarthric features more reliably.

In summary, both VC and TTS models effectively improve data augmentation for dysarthric speech synthesis. However, the VC model produces more consistent and reliable data, whereas the TTS model, though capable of high-quality generation, suffers from variability due to its dependency on reference audio and transcription constraints.

6. Conclusion

Our experiments utilizing speech synthesis have generally led to performance improvements in the SI classification. Additionally, we observed meaningful changes in augmentation ratios across different speech synthesis models. However, this study has certain limitations. Specifically, we were unable to conduct an expert evaluation of the quality of augmented data generated by the xTTS and Hierspeech models. While our results demonstrate the effectiveness of augmentation in enhancing classification performance, further validation is needed to ensure that the synthesized data is not only beneficial for model training but also meets human annotation standards.

Furthermore, our experiments revealed an imbalance in the quality of augmented data across different Synthesis models. Identifying and selecting only high-quality augmented samples remains an open challenge. Future research should explore effective methods for filtering high-quality synthetic data to maximize both model performance and annotation reliability.

7. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00621, RS-2022- II220621, Development of artificial intelligence technology that provides dialog-based multi-modal explainability)

8. References

- [1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” 2020. [Online]. Available: <https://arxiv.org/abs/2006.11477>
- [2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.07447>
- [3] F. Javanmardi, S. R. Kadiri, and P. Alku, “Pre-trained models for detection and severity level classification of dysarthria from speech,” *Speech Communication*, vol. 158, p. 103047, 2024.
- [4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [5] S. Rathod, M. Charola, A. Vora, Y. Jogi, and H. A. Patil, “Whisper features for dysarthric severity-level classification,” *Small*, vol. 12, no. 768, p. 12, 2023.
- [6] Y. Choi, J. Lee, and M.-W. Koo, “Speech recognition-based feature extraction for enhanced automatic severity classification in dysarthric speech,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 953–960.
- [7] E. J. Yeo, K. Choi, S. Kim, and M. Chung, “Cross-lingual dysarthria severity classification for english, korean, and tamil,” in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 566–574.
- [8] A. Al-Ali, S. Al-Maadeed, M. Saleh, R. C. Naidu, Z. C. Alex, P. Ramachandran, R. Khoodeeram, and R. K. M., “Classification of dysarthria based on the levels of severity. a systematic review,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.07264>
- [9] M. Suresh and J. Thomas, “Review on dysarthric speech severity level classification frameworks,” in *2023 International Conference on Control, Communication and Computing (ICCC)*, 2023, pp. 1–6.
- [10] B. Vachhani, C. Bhat, and S. K. Kopparapu, “Data augmentation using healthy speech for dysarthric speech recognition,” in *Inter-speech 2018*, 2018, pp. 471–475.
- [11] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, “Investigation of data augmentation techniques for disordered speech recognition,” *arXiv preprint arXiv:2201.05562*, 2022.
- [12] B. Karumuru, P. Sapkota, and H. Kathania, “In-domain data augmentation to enhance severity level classification of dysarthria from speech,” in *2024 International Conference on Signal Processing and Communications (SPCOM)*, 2024, pp. 1–5.
- [13] Z. Jin, M. Geng, X. Xie, J. Yu, S. Liu, X. Liu, and H. Meng, “Adversarial data augmentation for disordered speech recognition,” 2021. [Online]. Available: <https://arxiv.org/abs/2108.00899>
- [14] W.-Z. Zheng, J.-Y. Han, C.-Y. Chen, Y.-J. Chang, and Y.-H. Lai, “Improving the efficiency of dysarthria voice conversion system based on data augmentation,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4613–4623, 2023.
- [15] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, “Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.08568>
- [16] H. Wang, T. Thebaud, J. Villalba, M. Sydnor, B. Lammers, N. Dehak, and L. Moro-Velazquez, “Duta-vc: A duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.10588>
- [17] E. Hermann and M. Magimai-Doss, “Few-shot dysarthric speech recognition with text-to-speech data augmentation,” in *Inter-speech 2023*, 2023, pp. 156–160.
- [18] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, “Synthesizing dysarthric speech using multi-talker tts for dysarthric speech recognition,” 2022. [Online]. Available: <https://arxiv.org/abs/2201.11571>
- [19] E. Casanova, K. Davis, E. Gölge, G. Gökner, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, “Xtts: a massively multilingual zero-shot text-to-speech model,” *arXiv preprint arXiv:2406.04904*, 2024.
- [20] E. Gölge and T. C. T. Team, “Coqui tts,” 2021, a deep learning toolkit for Text-to-Speech, battle-tested in research and production. [Online]. Available: <https://github.com/coqui-ai/TTS>
- [21] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, “Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis,” *arXiv preprint arXiv:2311.12454*, 2023.
- [22] J. Lee, Y. Choi, T.-J. Song, and M.-W. Koo, “Inappropriate pause detection in dysarthric speech using large-scale speech recognition,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 486–12 490.