Synthetic Dysarthric Speech: A Supplement, Not a Substitute for Authentic Data in Dysarthric Speech Recognition

Jingting Li[1,2], *Keyi Feng*[1,3], *Xinran Zhao*[1,3], *Yan Wang*[1,2], *Su-Jing Wang**[1,2]

[1]State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, China

[2]Department of Psychology, University of the Chinese Academy of Sciences, China

[3]School of Computer, Jiangsu University of Science and Technology, China

wangsujing@psych.ac.cn

Abstract

Dysarthric speech recognition (DSR) is an emerging field that can enhance social interactions and mental health for individuals with dysarthria. However, the lack of sufficient Chinese dysarthric speech data and challenges like ambiguity and individual differences hinder performance improvements. Text-to-speech (TTS) technology is well-established in normal speech recognition and can also supplement dysarthric speech data. This study explores the impact of TTS-based Chinese dysarthric speech generation on DSR performance. Speaker-dependent experiments show that synthetic dysarthric speech alone does not effectively improve DSR performance. Through statistical analysis of acoustic features, we reveal the disparities between synthetic and authentic speech in dysarthria and highlight the limitations of synthetic data for DSR. These findings provide insights for future improvements in speech generation methods.

Index Terms: Dysarthric speech recognition, Synthetic dysarthric speech, Acoustic feature difference

1. Introduction

Dysarthria, a motor speech disorder, affects many individuals and includes a range of conditions caused by various factors such as cerebral palsy, Parkinson's disease, amyotrophic lateral sclerosis, and stroke [1]. Those with dysarthria often experience articulatory impairments like slower articulation, imprecise pronunciation, inappropriate pauses, and decreased clarity [2]. Accurately recognizing speech from individuals with dysarthria and enabling effective communication with computers could significantly improve their daily lives. A reliable speech recognition system for individuals with dysarthria would not only facilitate tasks such as using smart home devices and voice assistants, but also enhance quality of life, reduce dependency on caregivers, and promote meaningful interaction with the outside world.

While significant progress has been made in Automatic Speech Recognition (ASR) systems [3], recognizing speech from individuals with dysarthria remains a major challenge. Specifically, Dysarthric Speech Recognition (DSR) presents significant challenges due to two key factors: the scarcity of sufficient training data, and the unique and highly variable speech characteristics of each individual. Inter-individual variability in dysarthria severity complicates the development of universal DSR models.

Meantime, speaker-dependent DSR systems can enhance recognition accuracy for individual speakers, making them highly applicable in real-world scenarios. However, in the Chinese context, building such systems also faces the challenge of limited sample size. For example, although the release of the CDSD database [4] (currently the largest Chinese dysarthric speech database) has helped address the shortage of Chinese dysarthric speech data, the fact that each individual has only up to 10 hours of data remains insufficient for training accurate models that can handle the diverse and complex characteristics of dysarthric speech.

To overcome this limitation of scarce data, text-to-speech (TTS) technology has been increasingly used as a data augmentation method. This approach helps reduce the burden on the dysarthric speech community by generating large-scale synthetic data, ultimately improving DSR performance. However, there is no existing research exploring whether, in the Chinese context, synthetic speech can replace authentic speech to directly enhance recognition performance.

In this study, we investigate the impact of speech synthesized through TTS technology, using Mandarin phonetics, on the performance of DSR. The experiment result shows that TTS-synthesized speech data could serve as a supplement but not a substitute of authentic data for DSR. Furthermore, we compare the differences between synthetic and authentic speech through feature visualization and statistical analysis. By exploring the effects of different characteristics of synthetic samples, we demonstrate that TTS-synthesized speech data does not fully capture the diverse range of speech characteristics present within the dysarthria population. The contribution of this work can be summarized as follows:

- This study focuses on TTS-based speech generation data augmentation in Chinese DSR, a field that has received limited attention previously.
- We demonstrate that synthetic dysarthric speech cannot replace authentic speech in improving DSR performance.
- The study highlights the differences between synthetic and authentic dysarthric speech across various acoustic features, providing insights for future DSR data augmentation using speech generation.

2. Related Works

2.1. Dysarthric Speech Databases

The development of speech recognition models heavily relies on large-scale speech datasets. However, the advancement of DSR technology is currently constrained by the issue of small sample sizes. Chinese dysarthric speech data is relatively limited, both when compared to normal speech data and when compared to dysarthric speech data in English. For instance, the commonly used Chinese speech recognition dataset WenetSpeech [5] includes multilingual Chinese speech data across multiple domains, with a scale of over 10,000 hours, which is much larger than the existing dysarthric speech databases. Internationally, the English dysarthric database Whitaker was

established as early as 1993 [6]. Over the span of 28 years, from 1993 to 2021, six dysarthric speech databases have been established abroad, among which the largest is the Euphonia database from 2021, with 1,300 hours of speech data [7]. The most widely used dysarthric speech databases currently include UA-Speech [8] and Torgo [9].

Given the significant differences in pronunciation, language morphology, and sentence structure between Chinese and English, and considering that Chinese has more homophones and polyphonic words than English, existing English dysarthric speech databases do not meet the specific requirements of Chinese DSR tasks. The establishment of Chinese dysarthric speech databases started later. In 2015, Wong et al. established the Cantonese dysarthric speech corpus CUDYS [10], in 2024, Liu et al established the Mandarin Subacute Stroke Dysarthria Multimodal Database MSDM [11], Gao et al. released the Mandarin Dysarthria Speech Corpus (MDSC) [12], with all three databases containing less than 10 hours of data. The largest public database is the CDSD database[1] published by Wang et al., which includes 144 hours of dysarthric speech data [4]. However, overall, the total scale of current dysarthric speech data is relatively small.

The main difficulty in collecting dysarthric speech data is that individuals with dysarthria face challenges when speaking. For the same length of text, individuals with dysarthria require more time and effort, making it difficult to collect speech data from them. Additionally, because there are significant individual differences in dysarthric speech, current dysarthric speech data cannot be automatically annotated, making manual annotation of this speech data even more challenging. This further complicates the construction of dysarthric speech databases. Therefore, how to provide sufficient training samples for DSR is an urgent problem that needs to be addressed.

2.2. Data Augmentation in DSR

Over the past several years, numerous data augmentation methods have been proposed to overcome the scarcity of dysarthric speech data for DSR. In 2018, Vachhani et al. augmented dysarthric ASR training by applying temporal and speed modifications to healthy speech [13]; also in 2018, Jiao et al. simulated dysarthric speech through adversarial training to generate training data for clinical speech applications [14]. In 2021, Soleymanpour et al. proposed a prosodic transformation and time-feature masking method for subword end-to-end ASR augmentation [15]. In 2022, Celin Mariyaet al. introduced virtual microphone array synthesis with multi-resolution feature extraction to augment continuous dysarthric speech [16], and Bhat et al. presented a two-stage augmentation scheme combining deep autoencoder modifications with SpecAugment [17]. In 2023, Jin et al. advanced personalized adversarial data augmentation using speaker-dependent GANs on dysarthric and elderly speech [18], while Soleymanpour et al. developed a modified multi-talker TTS system with dysarthria severity controls for the TORGO corpus [19], and Baali et al. improved Arabic DSR by combining signal-based speed/tempo modifications with a Parallel Wave Generative adversarial model [20]. In 2024, Leung et al. demonstrated diffusion-based text-to-dysarthric-speech synthesis for augmenting training data for models like Whisper [21], and Wang et al. compared various adversarial augmentation strategies to enhance fine-tuning of pre-trained ASR systems [22], while Naeini et al. showed

[1] Application website: http://melab.psych.ac.cn/CDSD.html

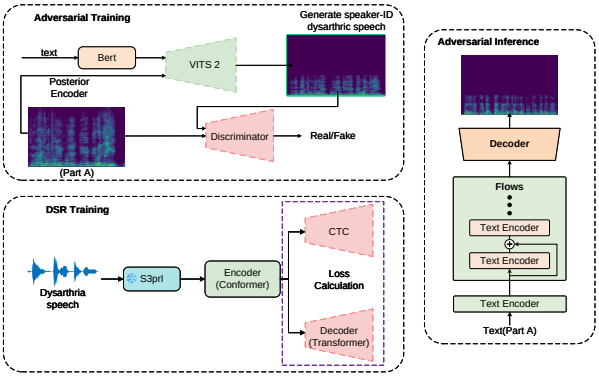


Figure 1: *Overview of Speech generation and DSR in out study*

that emulated and synthetic augmentation methods can significantly improve dysarthric speech segmentation across neurological disorders [23].

In our research, we focus on TTS-synthesized speech effect on DSR using a Chinese dysarthric speech database, distinct from the languages and datasets typically used in previous studies. Furthermore, previous studies have not explored whether synthetic speech can replace authentic dysarthric speech and analyze its acoustic features. Our findings reveal that synthetic speech can serve as a supplementary method to improve performance. However, relying solely on synthetic data does not lead to good results, as it cannot replace authentic speech data and there are differences in multiple acoustic features.

3. Method

In the technical implementation, as illustrated in Fig. 1, we selected two widely used methods for TTS and DSR, i.e., BERT-VITS2 [24] and ESPnet [25]. The BERT-VITS2 was selected for its proven capability in modeling prosodic variations (8.2k stars on Github), while the Conformer-Transformer architecture in ESPnet provides SOTA DSR performance [4, 20, 22].

3.1. TTS

In this study, we trained a BERT-VITS2 speech synthesis model using the CDSD Database to more accurately capture the speech characteristics of dysarthric individuals. The trained TTS model was then used to generate a large amount of synthetic impaired speech data for data augmentation. The BERT-VITS2 model consists of two main components: an autoregressive text encoder and a non-autoregressive acoustic model.

Precisely, the BERT modelserves as the autoregressive text encoder. BERT is based on a bidirectional Transformer architecture and uses self-attention mechanisms to capture dependencies between words in a given context. The VITS2 architecture [26], is responsible for generating the speech waveform. It combines Variational Autoencoders, adversarial training, and Monotonic Alignment Search. The encoder is built with six layers of dilated convolutions, where each layer has a different dilation rate (1, 2, 4, 8, 16, and 1), and outputs the mean and variance of the latent variables. The decoder adopts an improved WaveNet architecture, consisting of 40 residual blocks, each containing two dilated convolutions, and uses dynamic convolutions and skip connections to generate high-quality waveforms. The alignment module uses a bidirectional LSTM to compute the alignment between text and speech, ensuring that the timing

and content of the synthetic speech correspond to the input text. The discriminator processes waveforms at multiple periods and scales to ensure realistic and high-quality speech. A random duration predictor introduces variability into speech timing by injecting random noise into the model.

During the integration of BERT and VITS2, BERT features are concatenated with the output of the encoder and passed through a 1×1 convolution for fusion. In the alignment module, the concatenated BERT and Mel-spectrogram features are input into a bidirectional LSTM to compute the alignment matrix. Throughout training, BERT is used as a static feature extractor to enhance the quality of speech synthesis.

3.2. DSR

Our DSR pipeline leverages the ESPnet framework, utilizing Conformer as the encoder and Transformer as the decoder. Specifically, The frontend is configured with S3PRL [27], a pre-trained feature extraction model that processes raw audio data and generates feature representations suitable for subsequent model training. In particular, the Chinese-HuBERT-large.pt model, a HuBERT-based feature extractor, is used. HuBERT [28] is a self-supervised learning model for speech representations that produces high-quality feature embeddings for speech recognition tasks. For the pre-encoder, a linear configuration is used. In terms of data augmentation, SpecAugment [29] is applied to the spectrograms. This technique performs specific time-frequency transformations to increase the diversity of the training data, which in turn enhances the model's robustness.

4. Experiment

This study investigates the impact of TTS-based speech generation on Chinese DSR performance, including how different scales of data generation affect performance and comparing the acoustic features of TTS-synthesized and authentic speech.

4.1. Configuration

The CDSD database contains 133 hours of recordings from 44 speakers. The dataset is divided into two parts: Part A includes 44 hours of audio data from all 44 speakers, with each speaker contributing one hour of speech. Part B includes 80 hours of audio data, where eight speakers from Part A each contribute 10 hours of speech. Our study used Part B for speaker-independent DSR. Due to an error in the annotation information for one participant, we only used data from seven participants. The speech data from the seven speakers is divided into training, validation, and test sets in an 8:1:1 ratio. The training set is further randomly split into two equal subsets, Set A and Set B.

We trained the BERT-VITS2 model using Set A data from seven speakers, creating seven speaker-specific TTS models. For each speaker, two synthetic speech data sets were generated from Set A transcriptions: $\hat{A}^{-}$ from the authentic text, and $\hat{A}^{+}$ with a new text set 7 to 9 times the volume of $\hat{A}^{-}$. The resulting training sets are listed in Table 1.

For the BERT-VITS2 model, we used the chinese-roberta-wwm-ext-large pre-trained language model [30]. The model was trained for 1000 epochs with a batch size of 12, an initial learning rate of 0.0002, and the Adam optimizer (betas 0.8 and 0.99). The learning rate decayed at a rate of 0.99995. TTS model convergence occurred between 5,000 and 6,000 steps, trained on a 24GB 4090 GPU. For DSR model training, the initial learning rate was 0.0005, with a warmup for the first 30,000 steps, then adjusted per the schedule. The batch size was set with a maximum bin limit of 12,000,000 per batch, and training was done on three 24GB 3090 GPUs.

4.2. DSR Comparison

As listed in Table. 2, we tested the DSR performance based on seven group of training sets. The trends of the results are visually presented in the bar chart in the Supp. Materials[2]. It can be seen that, when the synthetic speech sample size does not significantly exceed that of Set A+B, on the basis of Set B, neither using the data synthesized from Set A alone nor using both the synthetic and authentic data from Set A achieves better DSR performance than using Set A+B. This emphasizes the differences between synthetic and authentic data, especially in the speech data of the dysarthric group. Synthetic speech not only fails to fully represent the diversity of authentic speech but also suffers from unpredictable quality issues.

For the cases of Set B+A+$\hat{A}^{+}$ and B+$\hat{A}^{+}$, where the synthetic data quantity is significantly larger, reaching 7-9 times the amount of authentic data, the model performance only slightly exceeds that of Set B+A. However, this improvement is far from significant. Compared to training with authentic data, the performance boost from synthetic data remains marginal. Firstly, synthetic data cannot fully capture the diversity and complexity of authentic speech. Dysarthric speech often exhibits more variation and irregularity in pronunciation, and the quality of synthetic speech may fail to accurately simulate these characteristics, negatively impacting model performance during training. Additionally, the diversity of synthetic data is limited, meaning it cannot cover all the potential variations present in authentic speech. This limitation causes the model to produce higher error rates when faced with authentic data.

As illustrated in Fig. 2(a), further visualization of the features from both data sets shows significant differences in the feature space between synthetic and authentic speech. These differences highlight the inconsistencies in speech features between synthetic and authentic data. Furthermore, from the spectrogram (Fig. 2(b)), there are noticeable differences between the synthetic and authentic speech. In terms of frequency distribution and energy intensity, the authentic speech exhibits a rich low-frequency component with a complex energy distribution, while the synthetic speech, although similar, shows different energy patterns in certain frequency bands. Regarding dynamic features, the frequency components of the authentic speech change diversely over time, whereas the synthetic speech exhibits different patterns.

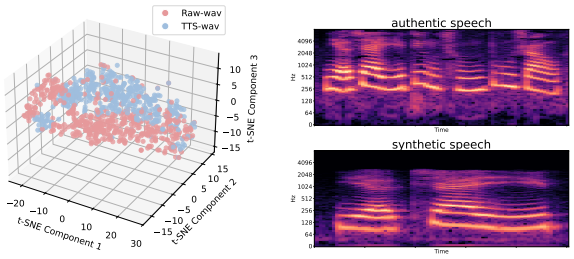
4.3. Similarity Comparison

To understand the differences, we analyze the acoustic features [31] of synthetic and authentic speech. We randomly select 100 pairs of speech samples with the same text from each speaker. The comparison metrics include: fundamental Fre-

[2]Supp. Materials link: https://doi.org/10.5281/zenodo.15511438

Table 1: *Training Set Combinations. Scale is based on sample size of Set A.*

Comb	1	2	3	4	5	6	7
Set	B	B+A	B+ $\hat{A}^{-}$	B+ $\frac{1}{2}\hat{A}^{+}$	B +$\hat{A}^{+}$	B+A+$\hat{A}^{-}$	B+A+$\hat{A}^{+}$
Scale	1	2	2	5-6	8-10	3	9-11



(a) t-SNE result (b) Spectrum difference

Figure 2: *Speech feature difference (from Speaker 01)*

quency correlation coefficient (F0), fundamental frequency jitter difference (F0JD), pitch shimmer difference (PSD), loudness difference (LD), root mean square error correlation coefficient (RM-C), cosine similarity of MFCC, perceptual evaluation of speech quality (PESQ) score, and average speaker similarity (ASS). The mean and t-test values of these metrics are calculated to assess whether the differences are statistically significant, as presented in Table 3. P-value analysis can be found in Supp. Materials. Besides, metrics outside the [0,1] range were normalized for better interpretability (see Supp. Materials).

Dysarthric speech shows notable pitch instability with pronounced variation patterns. Our results reveal significant differences in pitch-related features between synthetic and authentic speech, especially F0 and F0JD. These differences align with dysarthric speech's irregular pitch caused by speech production difficulties and poor muscle coordination. While synthetic speech mimics some pitch features, large discrepancies remain, highlighting that complex pitch changes in dysarthric speech are not fully captured. The numerical differences of PSD among different speakers and in the overall data are also quite notable, indicating that there are substantial differences in fine-grained stability between synthetic and authentic speech.

For dysarthric speech, loudness fluctuations occur due to unstable airflows or poor motor control. Although the differences between synthetic and authentic speech seem relatively small on the LD mean values, there are significant differences in statistical metrics. Furthermore, the weak RM-C presents the discrepancies in signal characteristics. These differences reflect inherent irregularities in loudness and signal errors in dysarthric speech that generation methods fail to fully address.

The MFCC cosine similarity is above 0.998, showing a high similarity in spectral envelope features, suggesting that the timbre of synthetic and authentic speech is alike. However, the significant statistical differences suggest that subtle frequency

Table 2: *DSR performance (CER↓) based on different training sets*

Speaker ID	01	04	06	08	09	10	12
B	25.6	48.8	72.5	49.1	49.4	36.3	44.0
B+A	16.9	42.9	52.3	32.2	35.4	25.9	35.1
B+$\hat{A}^-$	17.9	44.6	55.9	36.9	37.2	29.1	34.9
B+$\frac{1}{2}\hat{A}^+$	18.4	44.4	51.7	36.5	38.5	24.3	35.8
B+$\hat{A}^+$	15.5	43.0	48.9	30.7	35	21.0	37.9
B+A+$\hat{A}^-$	17.5	45	53.9	31.4	34.9	26.7	35.6
B+A+$\hat{A}^+$	**14.1**	**39.7**	**48.5**	**25.4**	**28.8**	**19.0**	**33.1**

Table 3: *Statistic analysis on difference between TTS-synthesized and authentic speech data. The direction of the arrow indicates that the similarity between the data is higher.*

Metric	Speaker ID	1	4	6	8	9	10	12	Overall
Mean	F0 ↑	0.7194	0.6657	0.6061	0.6696	0.6589	0.6581	0.7659	0.6777
	F0JD ↓	0.9971	0.9843	1.0000	1.0000	1.0000	1.0000	0.9976	0.9970
	PSD ↓	0.7995	0.5644	0.2668	0.6224	0.4582	0.4044	0.3109	0.4895
	LD ↓	0.2157	0.1837	0.0757	0.0987	0.0457	0.0823	0.0464	0.1069
	RM - C ↑	0.6642	0.5847	0.6829	0.6685	0.6198	0.6431	0.7095	0.6532
	MF - C ↑	0.9999	0.9998	0.9998	0.9982	0.9999	0.9998	0.9997	0.9996
	PESQ ↑	0.3418	0.3466	0.3388	0.3380	0.3412	0.3533	0.3639	0.3462
	ASS ↑	0.9364	0.9486	0.9421	0.9475	0.9386	0.9569	0.9469	0.9453
T - test	F0	-14.75	-19.93	-17.02	-17.83	-18.68	-22.54	-14.60	-45.32
	F0JD	339.75	88.02	inf	inf	inf	inf	416.99	589.48
	PSD	46.25	36.37	15.67	35.60	17.64	25.44	13.47	50.16
	LD	43.77	30.48	15.10	31.83	14.09	26.02	9.13	37.00
	RM - C	-19.33	-28.17	-17.75	-22.88	-24.95	-22.95	-17.79	-56.03
	MF - C	-5.62	-4.00	-4.39	-1.16	-4.36	-3.44	-3.15	-1.76
	PESQ	-216.02	-139.96	-253.15	-203.77	-259.23	-147.88	-156.79	-469.47
	ASS	-61.57	-60.08	-48.22	-43.13	-57.67	-52.30	-53.33	-119.03

or temporal variations still prevent full reproduction of the authentic speech.

The low PESQ score suggests poor speech quality similarity. This reflects that while synthetic speech captures some spectral features, it cannot fully replicate the clarity and fluency of natural speech. The complexity and irregularity of dysarthric speech make the reproduction difficult.

The ASS is 0.9453, indicating a high degree of similarity in speaker characteristics. Despite the differences between dysarthric and normal speech, synthetic speech retains some speaker-specific features, showing its potential for personalized speech reconstruction. However, due to individual differences in dysarthric speech, significant discrepancies remain, particularly across samples.

Dysarthric speech exhibits unique characteristics across various dimensions, including pitch instability, loudness fluctuations, and poor speech quality. These features present challenges when attempting to generate speech that mimics dysarthric speech. While synthetic speech can resemble authentic dysarthric speech in certain aspects (such as spectral features and speaker similarity), significant differences remain across multiple dimensions, particularly in the finer aspects of speech quality and fluency. Therefore, while synthetic speech can serve as a supplement for data augmentation, it cannot fully replace authentic speech, especially in terms of detailed speech quality and smoothness.

5. Conclusion

DSR is an important tool for improving the quality of life and promoting social interaction within the dysarthric community. However, challenges such as the difficulty in collecting speech data have limited the development of DSR technologies. Speech generation-based data augmentation is an effective method to enhance DSR performance. This paper explores the impact of synthetic speech on DSR performance and presents a statistical analysis of the feature differences between synthetic and authentic speech. The results demonstrate that synthetic speech based on dysarthric speech can serve as a supplementary data source but cannot replace authentic speech, as there are significant differences in dimensions such as pitch and loudness. These findings offer insights for future research, particularly in improving the quality of synthetic speech to better mimic the specific speech characteristics of the dysarthric population. Additionally, exploring how to utilize synthetic speech data more effectively to enhance DSR performance remains an important direction for future investigation.

6. Acknowledgment

This research was partially funded by 1) the National Natural Science Foundation of China (62476269, 62276252); 2) the Youth Innovation Promotion Association CAS.

7. References

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

[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] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep speech 2 : End-to-end speech recognition in english and mandarin." in *International Conference on Machine Learning*, 2016, pp. 173–182.

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

[5] 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.

[6] J. Deller Jr, M. Liu, L. Ferrier, and P. Robichaud, "The whitaker database of dysarthric (cerebral palsy) speech," *The Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.

[7] R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson *et al.*, "Disordered speech data collection: Lessons learned at 1 million utterances from project Euphonia." in *Interspeech*, vol. 2021, 2021, pp. 4833–4837.

[8] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. R. Gunderson, T. S. Huang, K. L. Watkin, and S. Frame, "Dysarthric speech database for universal access research." in *Interspeech*, vol. 2008, 2008, pp. 1741–1744.

[9] 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.

[10] K. H. Wong, Y. T. Yeung, E. H. Chan, P. C. Wong, G.-A. Levow, and H. Meng, "Development of a cantonese dysarthric speech corpus," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] J. Liu, X. Liu, Y. Yang, R. Ruzi, X. Zuo, C. Xu, C. Li, X. Li, R. Su, H. An-ming, 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)*, 2024, pp. 131–135.

[12] 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 *Interspeech 2024*, 2024, pp. 2465–2469.

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

[14] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6009–6013.

[15] M. Soleymanpour, M. T. Johnson, and J. Berry, "Dysarthric speech augmentation using prosodic transformation and masking for subword end-to-end asr," in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2021, pp. 42–46.

[16] T. Mariya Celin, P. Vijayalakshmi, and T. Nagarajan, "Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 601–622, 2023.

[17] C. Bhat, A. Panda, and H. Strik, "Improved asr performance for dysarthric speech using two-stage dataaugmentation." in *INTERSPEECH*, 2022, pp. 46–50.

[18] Z. Jin, M. Geng, J. Deng, T. Wang, S. Hu, G. Li, and X. Liu, "," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[19] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Accurate synthesis of dysarthric speech for asr data augmentation," *Speech Communication*, vol. 164, p. 103112, 2024.

[20] M. Baali, I. Almakky, S. Shehata, and F. Karray, "Arabic dysarthric speech recognition using adversarial and signal-based augmentation," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, 2023, pp. 1558–1562.

[21] 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 *Proceedings of Interspeech 2024*. Sheffield, 2024.

[22] H. Wang, Z. Jin, M. Geng, S. Hu, G. Li, T. Wang, H. Xu, and X. Liu, "Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 311–12 315.

[23] S. A. Naeini, L. Simmatis, D. Jafari, Y. Yunusova, and B. Taati, "Improving dysarthric speech segmentation with emulated and synthetic augmentation," *IEEE Journal of Translational Engineering in Health and Medicine*, 2024.

[24] https://github.com/fishaudio/Bert-VITS2.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *Interspeech 2018*, 2018.

[26] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," *arXiv preprint arXiv:2307.16430*, 2023.

[27] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang *et al.*, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[28] W. Chen, X. Chang, Y. Peng, Z. Ni, S. Maiti, and S. Watanabe, "Reducing barriers to self-supervised learning: Hubert pre-training with academic compute," in *Interspeech 2023*, 2023, pp. 4404–4408.

[29] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, 2019, pp. 2613–2617.

[30] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9599397

[31] "Vocal acoustic analysis – jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013, cENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.