# Zero- and One-Shot Data Augmentation for Sentence-Level Dysarthric Speech Recognition in Constrained Scenarios

Shiyao Wang[0009−0006−5786−7109], Shiwan Zhao[0000−0001−5068−025X], Jiaming Zhou[0009−0002−4819−4572], and Yong Qin[0009−0000−2748−3020]*✉

Nankai University, China
wangshiyao@mail.nankai.edu.cn
zhaosw@gmail.com
zhoujiaming@mail.nankai.edu.cn
qinyong@nankai.edu.cn

**Abstract.** Dysarthric speech recognition (DSR) research has witnessed remarkable progress in recent years, evolving from the basic understanding of individual words to the intricate comprehension of sentence-level expressions, all driven by the pressing communication needs of individuals with dysarthria. Nevertheless, the scarcity of available data remains a substantial hurdle, posing a significant challenge to the development of effective sentence-level DSR systems. In response to this issue, dysarthric data augmentation (DDA) has emerged as a highly promising approach. Generative models are frequently employed to generate training data for automatic speech recognition tasks. However, their effectiveness hinges on the ability of the synthesized data to accurately represent the target domain. The wide-ranging variability in pronunciation among dysarthric speakers makes it extremely difficult for models trained on data from existing speakers to produce useful augmented data, especially in zero-shot or one-shot learning settings. To address this limitation, we put forward a novel text-coverage strategy specifically designed for text-matching data synthesis. This innovative strategy allows for efficient zero/one-shot DDA, leading to substantial enhancements in the performance of DSR when dealing with unseen dysarthric speakers. Such improvements are of great significance in practical applications, including dysarthria rehabilitation programs and day-to-day common-sentence communication scenarios.

**Keywords:** Data augmentation · Dysarthric speech recognition · Zero-/one-shot · Sentence-level · Unseen speakers

## 1 Introduction

Dysarthria, a speech disorder characterized by impaired vocal control, is commonly associated with neurological conditions such as cerebral palsy and Parkin-

* Corresponding author.

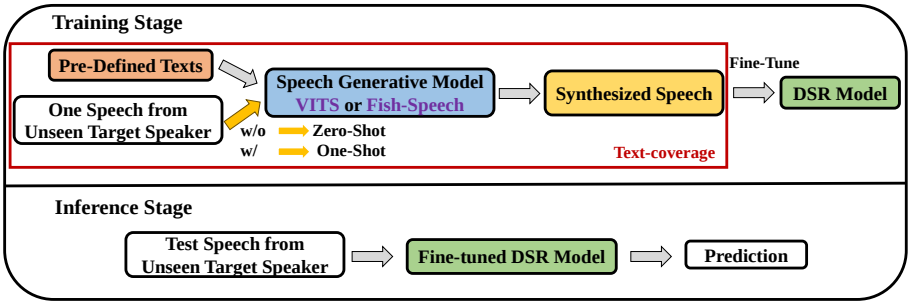


**Fig. 1.** The overview of our Zero-/One-shot DDA pipeline. Training stage: DSR model fine-tuned with **text-coverage** data. Inference stage: Fine-tuned DSR model used for prediction.

son's disease. It leads to speech that exhibits pronunciation errors and irregular prosody, significantly hindering automatic speech recognition (ASR) systems trained on typical speech [1]. Research in dysarthric speech recognition (DSR) primarily focuses on word recognition [2, 3]. Current methodologies include speaker-independent models [4, 5], which often struggle to generalize to unseen dysarthric speakers due to pronunciation variability, and speaker-dependent models [6], which necessitate continuous data collection and costly fine-tuning to maintain optimal performance. Wang et al. [7] proposed a prototype-based, fine-tuning-free classification method for word-level DSR that utilizes speaker-specific pronunciation patterns for rapid and effective adaptation. However, DSR now faces the challenge of achieving sentence-level understanding. Sentence-level DSR is particularly sensitive to pronunciation patterns, including breathing sounds, abnormal pauses (which can lead to insertion errors), and mispronunciations (resulting in replacement errors). Additionally, sentence-level DSR encounters challenges related to larger vocabularies and coarticulation effects, requiring more targeted data for effective adaptation.

Acquiring data from dysarthric speakers presents challenges due to participant limitations and recording difficulties. The unique characteristics of dysarthric speech further complicate the annotation process, requiring specialized expertise. Consequently, few open-source dysarthric datasets exist, primarily at the word or command level [2, 8]. While some sentence-level dysarthric datasets are available [9, 10], their quantity remains limited. To address this data scarcity, dysarthric data augmentation (DDA) is essential. Current sentence-level DDA methods employ either Zero-Shot (simulating general dysarthric characteristics without using target speaker speech data) or All-Test-Data [11, 12] (utilizing all available test data from the target speaker to train speech generative models and then using these models to synthesize training data) settings. The All-Test-Data setting exhibits the best performance but represents an idealized scenario. Our work focuses on the more practical Zero-Shot and One-Shot settings with speech generative models.

Data augmentation in ASR, which typically involves synthesizing speech from texts in the training set, encounters challenges when applied to DSR. The limited size of dysarthric datasets often results in content mismatches between the training and target domains. Additionally, pronunciation variations between unseen target speakers and those in the training set exacerbate this mismatch. Consequently, synthetic data generated from training set texts using speech generative models trained on either typical speech or existing dysarthric speakers may not accurately capture the characteristics of the target domain. Rosenberg et al. [13] established a performance benchmark by synthesizing speech from test set texts, suggesting that while this data augmentation setting may be impractical for general ASR, it holds value for specific domain applications. We consider sentence-level DSR, a newly developed and challenging task, to be one such application. Chen et al. [14] emphasized the necessity for the texts of synthesized speech to closely align with the target domain to enhance ASR performance. Yang et al. [15] highlighted that text content is more crucial than speaking style in speaker adaptation, positing that performance improvements arise from synthesized speech standardizing the ASR model's language understanding. Based on these insights, we propose a text-coverage strategy (see Fig. 1), evaluating its effectiveness in Zero-Shot and One-Shot DDA settings. This strategy utilizes speech generative models to synthesize speech data directly from pre-defined texts. Although limiting the target domain content may not be ideal for general ASR, it offers advantages for applications such as dysarthria rehabilitation or scenarios where dysarthric speakers use common sentences for communication, particularly given the prior focus of DSR on word recognition. Considering the limited dysarthric data and the prosody variability of dysarthric speech, we selected VITS [16] for Zero-Shot and Fish-Speech [17] for One-Shot as our DDA models (see Section 3 for selection analysis). To our knowledge, this is the first application of these models in the context of DDA.

We evaluated our text-coverage strategy on two sentence-level dysarthric speech datasets: the English Torgo dataset [9] and the Chinese CDSD dataset [10]. On the Torgo dataset, the Zero-Shot text-coverage strategy achieved a 3.29% absolute reduction in average WER, while the One-Shot text-coverage strategy resulted in a 5.98% absolute WER reduction. For the CDSD dataset, we implemented a stepwise filtering mechanism, ceasing further DDA for speakers with high intelligibility (CER $<$25%; [10] test set CER: 26.46%). Finally, 20 out of 24 speakers to achieve a CER below 25%. By utilizing **a maximum of one sample** per speaker, the average final CER was 19.525% (calculated by averaging the boldfaced values in Table 2).

Our text-coverage strategy leverages prior knowledge of the target domain's content and incorporates dysarthric acoustic knowledge from the speech generative model. Related work in text-only domain adaptation [18, 19] typically employs language models (LMs) trained on target domain text data to improve ASR performance via shallow fusion of ASR and LM outputs during inference. This approach enhances the fluency of predicted text and adapts to the target domain's linguistic characteristics. We compared this approach with ours on the

CDSD dataset and found that while it can improve performance, our method yields more significant gains.

The main **contributions** of this work are as follows:

- We systematically analyzed existing DDA methods, highlighting their advantages and limitations.
- We present a novel and efficient text-coverage strategy that enables the first sentence-level zero-shot and one-shot DDA based on speech generative models.
- We evaluated the effectiveness of our proposed DDA strategy and selected models using the Torgo and CDSD dysarthric datasets.

## 2 Previous methods

**Pure signal processing approaches** for dysarthric data augmentation (DDA) encompass implementations such as: (1) modifying speaking rate using PSOLA and pitch using linear transformation [20], (2) manipulating the log-Mel spectrogram through time warping, frequency and time masking [21], (3) perturbing vocal tract length and adding reverberation [22], and (4) applying various spectral-temporal transformations [23]. While these methods simulate general characteristics of dysarthric speech, providing interpretability, their limited prosodic enhancement restricts the generation of authentic, diverse speech, hindering optimal performance.

**GANs** Given the significant pronunciation variability among dysarthric speakers, each speaker's pronunciation pattern can be considered a unique distribution. This characteristic has led to the use of various GANs for DDA, including DCGAN [24], CycleGAN [25], VAE-GAN [26], and Spectral basis GAN [27]. While GAN-based DDA excels at fitting data distributions, it requires substantial target speaker data; insufficient data can cause model collapse and limit output diversity.

**Reinforcement learning** Jin et al. [28] proposed a dynamic DDA method that utilizes reinforcement learning to optimize SpecAugment parameters during DSR training. While this approach demonstrates good performance through a speech chain, it is dependent on existing data and necessitates sufficient training data, along with further exploration of diverse settings.

**Reusing speech generative models** Recent DDA research has adapted Voice Conversion (VC) and Text-to-Speech (TTS) models, originally designed for typical speech, by training or fine-tuning them using dysarthric data. Models that have been adopted include: (1) attention-based VC [29], (2) sparse-structured spectral envelope transformation-based VC [30], (3) autoencoder-based VC [31], (4) diffusion-based VC [32] and TTS [12]. Beyond directly reusing models, Soleymanpour et al. [11] modified a multi-speaker TTS system by incorporating a dysarthria severity level coefficient and a pause insertion model, enabling the synthesis of dysarthric speech across varying severity levels. While these methods show promise, Hu et al. [33] demonstrated that some TTS models struggle to generate natural-sounding dysarthric speech, requiring larger datasets or longer

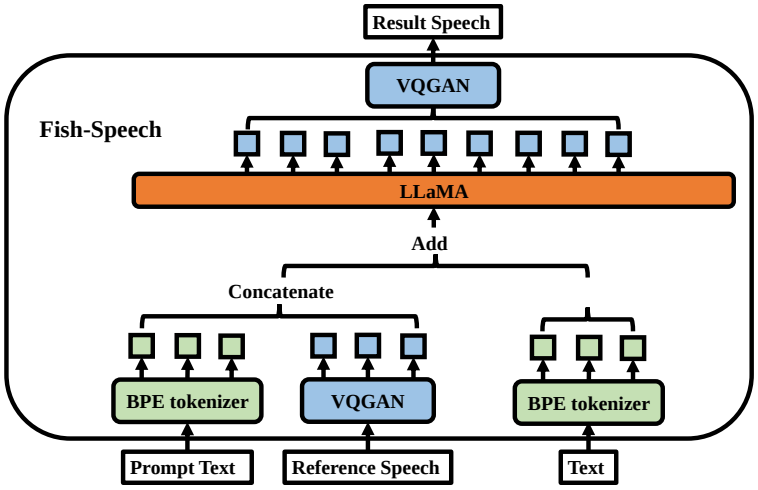


**Fig. 2.** Fish-Speech Inference Framework.'Prompt Text' corresponds to 'Reference Speech' and 'Text' is the target texts for synthesis. BPE: byte pair encoding.

training times. Furthermore, with the exception of [11, 12], these works primarily focus on enhancing word recognition, and this approach has not fully kept pace with the progress of typical speech TTS models, nor has it explored Voice Cloning technology.

## 3 Strategy design and model selection

This section details the rationale for our dysarthric data augmentation (DDA) strategy design and DDA models selection.

**Text-coverage strategy** To enhance the performance of Dysarthria Speech Recognition (DSR) models for unseen dysarthric speakers, improving the generalization ability of the DSR model or adapting the model is crucial. However, constructing a speaker-independent and text-independent DSR model requires large and diverse training data. To our knowledge, the number of speakers and texts covered by the widely-used Torgo and CDSD datasets remains limited. Constructing a speaker-dependent DSR model requires a large amount of comprehensive speech data from the target speakers to capture their unique pronunciation patterns and prosodic variations, but collecting and annotating dysarthric data are very difficult. Building on the findings discussed in [13–15] in Section 1, we propose the text-coverage strategy (see Fig. 1). This strategy involves synthesizing speech data based on the pre-defined texts content of the target scenario, allowing the DSR model to adapt proactively and enhance its performance.

**VITS** We chose the end-to-end TTS model VITS because it eliminates the need for separate vocoder adaptation, a process requiring extensive data. This

**Table 1.** WER (%) on Torgo. 'M', 'L', and 'VL' denote moderate, low, and very low severity, respectively. 'FT' indicates fine-tuning. 'AVG' is the speaker-level average WER. See Section 4.2 for definitions of **V**, **F1**, and **F2**.

| **Settings** | | **Severity Level** | | | **AVG** |
|---|---|---|---|---|---|
| | | **M** | **L** | **VL** | |
| **Without DDA** | [36] | 54.46 | 33.89 | 10.35 | 40.86 |
| | [37] | 48.40 | 31.07 | 8.66 | 36.30 |
| | [11] | 56.34 | 46.60 | 13.85 | 44.50 |
| | [12] | 85.35 | 28.54 | **3.44** | 57.77 |
| | LOSO | **39.56** | **24.60** | 6.30 | **29.38** |
| **Zero-Shot** | **V** FT LOSO | 34.38 | 24.40 | 6.20 | 26.09 ($\Delta$ **3.29%**) |
| **One-Shot** | **F1** FT LOSO | 33.36 | 21.50 | 5.85 | 25.00 |
| | **F2** FT LOSO | **31.06** | **20.90** | **5.50** | **23.40** ($\Delta$ **5.98%**) |
| **All-Test-Data** | [11] | 50.14 | 36.80 | 12.60 | 39.09 |
| | [12] | 26.37 | 26.14 | 3.82 | 20.70 |

is particularly advantageous given the limited availability of dysarthric speech data, as the vocoder significantly impacts generated speech quality and accuracy. Furthermore, VITS's architecture, incorporating GANs, can effectively learn the pronunciation distribution of dysarthric speakers through adversarial training. Moreover, as a conditional variational autoencoder, VITS introduces variability through latent variables and a random duration predictor, generating synthesized speech with diverse prosody for the same input text. This matches the prosodic variations in dysarthric speech and improves DSR model generalization.

**Fish-Speech** We selected Fish-Speech, a free and open-source voice cloning model. Its inference framework is illustrated in Figure 2, which we created based on the code from [17][1]. Our decision was influenced by several factors. First, Fish-Speech's UTF-8 text encoding and byte pair encoding tokenizer bypasses the phoneme system, eliminating the need for language-specific preprocessing and training, an advantage for our experiments involving two languages. Second, its architecture, a recently evolving neural codec-based architecture [34], can efficiently generate speech with diverse and realistic prosody. Finally, preliminary testing on the Torgo dataset indicated that its cloning performance for dysarthric speech, evaluated through sampling and manual assessment, is acceptable and could be further improved by fine-tuning LLaMA [35] of Fish-Speech.

## 4 Experiments

### 4.1 Dataset

**Torgo dataset** The Torgo dataset comprises English speech from 8 dysarthric speakers and 7 control speakers. After filtering out samples with transcription

[1] The code used is from August 2024.

errors and unrestricted content, a total of 16,582 data segments remained, with an approximate duration of 13.67 hours. Using the data from the 8 dysarthric speakers, we fine-tuned the pre-trained model using the leave-one-speaker-out (LOSO) method to construct the LOSO model. For the DDA models, all data except that of the target dysarthric speaker were used during training or fine-tuning, including data from the control speakers.

**CDSD dataset** We utilize the initial release of the CDSD dataset, comprising 29,466 Chinese recordings from 24 dysarthric speakers. Using the data from these 24 speakers, we also fine-tune the pre-trained model with the LOSO method to construct a LOSO model for DSR. For the DDA models, we incorporate 12,885 recordings (14.38 hours) from 31 dysarthric speakers with intelligibility annotations from the AISHELL-6B[2] dataset during training or fine-tuning

### 4.2 Experimental settings

**DDA models** We implemented the VITS model using the open-source framework[3], which has 39.68M parameters. We followed the original VITS codebase for English text processing, while performing word segmentation and Pinyin parsing for Chinese text. Hyperparameters were adapted from the vctk_base.json: for Torgo, batch size 16, n_speakers 14, 64k training steps; for AISHELL-6B, batch size 64, n_speakers 31, 36k training steps. We used the pre-trained Fish-Speech model[4] (532.07M parameters, 150k hours of pre-training data: 50k hours each of English, Chinese, and Japanese, and 1.5k hours of mixed data for supervised fine-tuning). Pre-trained Fish-Speech produced good timbre similarity but poor prosody. Since VQGAN [38] in Fish-Speech affects timbre and LLaMA affects prosody, we fine-tuned the LLaMA component using LoRA [39] with a learning rate of 1e-10 for 2000 steps.

**DSR models** We trained models using the ESPnet framework[5]. Training was set for a maximum of 200 epochs, with early stopping based on validation loss. For Torgo, an ASR model (70.47M parameters) pre-trained on the 960-hour LibriSpeech dataset[6] was fine-tuned using the LOSO approach. We synthesized speech for the test set texts using VITS, pre-trained Fish-Speech, and fine-tuned Fish-Speech, creating settings **V**, **F1**, and **F2**. Synthesized data was used to further fine-tune the LOSO models. For CDSD, an ASR model (116.91M parameters) pre-trained on the 10k-hour WenetSpeech dataset [40] was fine-tuned with LOSO, followed by additional fine-tuning using the DDA data. To mitigate potential negative impacts of DDA on high-intelligibility speakers (CER $<25\%$), we implemented a stepwise filtering mechanism. This involved progressing through LOSO models with Zero-Shot (**V**), One-Shot (**F2**), and All-Test-Data (**F3**) settings. Note that **F3** is identical to **F2** except that it includes all test data from

[2] https://www.aishelltech.com/AISHELL_6B
[3] https://github.com/jaywalnut310/VITS
[4] huggingface.co/fishaudio/fish-speech-1.2-sft
[5] https://github.com/espnet/espnet
[6] https://www.openslr.org/12/

the target speaker during the Fish-Speech fine-tuning stage. When a model's CER on a target speaker's speech is less than 25%, the LOSO model for that speaker is excluded from further DDA testing.

**Text-only domain adaptation** uses shallow fusion of ASR and language model (LM) predictions during inference:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \log p(y|x) + \lambda \log p_{LM}(y), \tag{1}$$

where $x$ represents the speech features, $y$ is the label sequence, $p(y|x)$ is the ASR model probability, $p_{LM}(y)$ is the LM probability, and $\lambda$ weights the LM's contribution. Due to the limited amount of sentence data in Torgo (25.5%) and space constraints, we only evaluated text-only domain adaptation on CDSD to assess the impact of text adaptation. We trained ESPnet LMs (54.06M parameters) for 100 epochs using the training and test set texts (representing source and target domains) in the LOSO setup. We explored the effects of $\lambda$ values of 0.3, 0.6, and 0.8.

### 4.3 Results and discussions

The experimental results on the Torgo dataset are presented in Table 1. Espana Bonnet et al. [36] compared the performance of various DNN architectures on Torgo, but none of the models achieved optimal results for all speakers. For our comparison, we selected the best performance for each speaker from [36]. Joy et al. [37] improved DSR performance by adjusting DNN architecture parameters. Soleymanpour et al. [11] employed a DNN-HMM architecture for DSR, while Leung et al. [12] utilized the large speech model Whisper [41] for DSR. We include the results from [11, 12] both before and after DDA. DDA methods used in [11, 12] are described in Section 2. Our LOSO models demonstrate strong performance without DDA. The Zero-Shot setting, which employs the DDA method that combines the VITS model with the text-coverage strategy, achieves a significant 3.29% absolute decrease in average WER. Further improvements are observed in the One-Shot setting, where Fish-Speech utilizes a single speech sample from the target dysarthric speaker to generate data that is more aligned with the target domain, leading to enhanced performance across all severity levels. Compared to the results of [11, 12] enhanced under the All-Test-Data setting, our One-Shot results significantly outperform [11] (p-value $<0.05$) and achieve comparable performance to [12]. Notably, we achieve these results using only one sample from the unseen target speaker, indicating that our method of leveraging Fish-Speech to learn the speaking style while aligning with the target text domain is an effective DDA approach.

For the CDSD dataset, lacking prior work using the LOSO setup for comparison, we first validated our model settings by building a DSR model. This was achieved by fine-tuning the pre-trained model using the data split described in [10]. Our model achieved a CER of 20.4% on the test split, a 6.06% absolute reduction compared to the results reported in [10]. As shown in Table 2, our LOSO models reduced the CER for 13 speakers to below 21%. The effectiveness

**Table 2.** CER (%) on CDSD. 'LM*' denotes enhancing the LOSO models via text-only domain adaptation.

| **Speaker ID** | **04** | **05** | **07** | **09** | **12** | **13** | **14** | **15** | **19** | **20** | **23** | **25** | **26** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pre-train ASR | 38.8 | 34.9 | 33.3 | 63.2 | 21.5 | 7.0 | 22.1 | 57.3 | 19.0 | 46.3 | 47.5 | 54.5 | 58.2 |
| LOSO models | **3.0** | **7.1** | **7.7** | **18.8** | **2.8** | **1.5** | **3.0** | **15.3** | **14.5** | **11.8** | **9.6** | **18.2** | **20.9** |
| **Speaker ID** | **01** | **02** | **10** | **16** | **03** | **08** | **17** | **06** | **11** | **18** | **21** | | |
| Pre-train ASR | 30.3 | 73.8 | 76.6 | 54.1 | 63.8 | 67.1 | 48.1 | 96.0 | 73.8 | 96.0 | 59.8 | | |
| LOSO models | 26.7 | 40.3 | 43.0 | 36.5 | 40.1 | 49.3 | 37.5 | 87.1 | 56.8 | 94.6 | 43.5 | | |
| **Zero-Shot** | **8.7** | **13.3** | **17.6** | **14.6** | 29.4 | 28.1 | 29.1 | 71.3 | 40.9 | 92.7 | 36.0 | | |
| **One-Shot** | — | — | — | — | **17.8** | **22.2** | **15.5** | **63.5** | **38.2** | **90.9** | **32.1** | | |
| **All-Test-Data** | — | — | — | — | — | — | — | 58.0 | 36.4 | 90.7 | 31.1 | | |
| $LM_{source}$ $\lambda = 0.3$ | 20.4 | 32.5 | 32.1 | 38.6 | 36.5 | 36.4 | 40.7 | 86.9 | 48.5 | 94.4 | 37.5 | | |
| $LM_{source}$ $\lambda = 0.6$ | 19.1 | 28.3 | 25.5 | 42.8 | 36.5 | 27.9 | 45.8 | 83.9 | 45.9 | 94.2 | 35.0 | | |
| $LM_{source}$ $\lambda = 0.8$ | 19.8 | 28.7 | 24.1 | 47.5 | 37.6 | 26.0 | 49.7 | 83.0 | 46.8 | 94.2 | 35.7 | | |
| $LM_{target}$ $\lambda = 0.3$ | 24.8 | 37.3 | 40.6 | 34.7 | 39.1 | 45.0 | 36.4 | 87.5 | 52.7 | 93.4 | 41.3 | | |
| $LM_{target}$ $\lambda = 0.6$ | 23.9 | 35.2 | 38.4 | 33.9 | 38.8 | 42.6 | 37.0 | 84.5 | 51.6 | 92.2 | 40.8 | | |
| $LM_{target}$ $\lambda = 0.8$ | 24.1 | 35.4 | 38.7 | 34.1 | 39.4 | 42.5 | 38.5 | 83.9 | 52.8 | 92.1 | 41.4 | | |

of our DDA methods is evident from Table 2: (1) The Zero-Shot setting further reduced the CER of 4 additional speakers to below 18% ($p < 0.05$), demonstrating the benefit of combining a multi-speaker dysarthric TTS model with our text-coverage strategy. (2) The One-Shot setting, which combines voice cloning and the text-coverage strategy, further improved DSR performance for the remaining 7 speakers ($p < 0.05$). Overall, we reduced the CER to below 25% for 20 speakers. However, speakers 06 and 18 (both children) and speakers 11 and 21 (both elderly men) presented greater challenges due to higher severity levels of dysarthria. This highlights the need for exploring new solutions specifically tailored to these more severely affected speakers.

We applied text-only domain adaptation to enhance LOSO models on the CDSD dataset (results in Table 2). By tuning the LM weight, $\lambda$, we consistently achieved performance improvements over the baseline LOSO models. Although the ESPnet framework defaults to a $\lambda$ of 0.3, increasing $\lambda$ generally improved performance for most speakers, highlighting the difficulty LOSO models face with unseen dysarthric speech. Furthermore, for most speakers, LMs trained on the training set texts (source domain) outperformed those trained on test set texts (target domain), likely due to the limited size of the test sets in the LOSO setup. However, our proposed method, which leverages only test set texts and acoustic information from the DDA model, yielded significantly greater improvements. We attribute this to our method's focus on directly improving the LOSO model itself.

## 5 Conclusions

We introduce a novel text-coverage strategy to improve sentence-level dysarthric speech recognition for unseen speakers in constrained scenarios. This strategy, combined with VITS and Fish-Speech, enables the first sentence-level zero-/one-shot dysarthric data augmentation using speech generative models. We have validated our strategy on the English dysarthric dataset Torgo and the Chinese dysarthric dataset CDSD.

**Acknowledgements** This work has been supported in part by NSF China (Grant No.62271270).

## References

1. Luigi De Russis and Fulvio Corno, "On the impact of dysarthric speech on contemporary asr cloud platforms," *Journal of Reliable Intelligent Environments*, vol. 5, pp. 163–172, 2019.
2. Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.
3. Zhaopeng Qian, Kejing Xiao, and Chongchong Yu, "A survey of technologies for automatic dysarthric speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, pp. 48, 2023.
4. Chitralekha Bhat, Ashish Panda, and Helmer Strik, "Improved asr performance for dysarthric speech using two-stage data augmentation," *Proc. Interspeech 2022*, pp. 46–50, 2022.
5. Eungbeom Kim, Yunkee Chae, Jaeheon Sim, and Kyogu Lee, "Debiased Automatic Speech Recognition for Dysarthric Speech via Sample Reweighting with Sample Affinity Test," in *INTERSPEECH*, 2023, pp. 1508–1512.
6. Katrin Tomanek, Katie Seaver, Pan-Pan Jiang, Richard Cave, Lauren Harrell, and Jordan R Green, "An analysis of degenerating speech due to progressive dysarthria on asr performance," in *ICASSP*. IEEE, 2023.
7. Shiyao Wang, Shiwan Zhao, Jiaming Zhou, Aobo Kong, and Yong Qin, "Enhancing dysarthric speech recognition for unseen speakers via prototype-based adaptation," in *Interspeech 2024*, 2024, pp. 1305–1309.
8. Rosanna Turrisi, Arianna Braccia, Marco Emanuele, Simone Giulietti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga, and Leonardo Badino, "Easycall corpus: a dysarthric speech dataset," *arXiv preprint arXiv:2104.02542*, 2021.
9. Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, 2012.
10. Mengyi Sun, Ming Gao, Xinchen Kang, Shiru Wang, Jun Du, Dengfeng Yao, and Su-Jing Wang, "Cdsd: Chinese dysarthria speech database," *arXiv preprint arXiv:2310.15930*, 2023.
11. Mohammad Soleymanpour, Michael T Johnson, Rahim Soleymanpour, and Jeffrey Berry, "Synthesizing dysarthric speech using multi-speaker tts for dysarthric speech recognition," in *ICASSP*. IEEE, 2022.

12. Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze, "Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis," in *Interspeech*, 2024.
13. Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu, "Speech recognition with augmented synthesized speech," in *ASRU*. IEEE, 2019, pp. 996–1002.
14. Zhehuai Chen, Andrew Rosenberg, Yu Zhang, Gary Wang, Bhuvana Ramabhadran, and Pedro J Moreno, "Improving speech recognition using gan-based speech synthesis and contrastive unspoken text selection.," in *Interspeech*, 2020, pp. 556–560.
15. Karren Yang, Ting-Yao Hu, Jen-Hao Rick Chang, Hema Swetha Koppula, and Oncel Tuzel, "Text is all you need: Personalizing asr models using controllable speech synthesis," in *ICASSP*. IEEE, 2023, pp. 1–5.
16. Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*. PMLR, 2021, pp. 5530–5540.
17. Tianyu Li Shijia Liao, "Fish speech v1," `https://github.com/fishaudio/fish-speech`, 2024.
18. Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.
19. Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *ICASSP*. IEEE, 2018, pp. 1–5828.
20. Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *ICASSP*. IEEE, 2018, pp. 6009–6013.
21. Lidan Wu, Daoming Zong, Shiliang Sun, and Jing Zhao, "A sequential contrastive learning framework for robust dysarthric speech recognition," in *ICASSP*. IEEE, 2021, pp. 7303–7307.
22. Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng, "Investigation of data augmentation techniques for disordered speech recognition," in *Interspeech 2020*, 2020, pp. 696–700.
23. Shansong Liu, Mengzhe Geng, Shoukang Hu, Xurong Xie, Mingyu Cui, Jianwei Yu, Xunying Liu, and Helen Meng, "Recent progress in the cuhk dysarthric speech recognition system," *TASLP*, vol. 29, pp. 2267–2281, 2021.
24. Zengrui Jin, Mengzhe Geng, Jiajun Deng, Tianzi Wang, Shujie Hu, Guinan Li, and Xunying Liu, "Personalized adversarial data augmentation for dysarthric and elderly speech recognition," *TASLP*, 2023.
25. Protima Nomo Sudro, Rohan Kumar Das, Rohit Sinha, and SR Mahadeva Prasanna, "Significance of data augmentation for improving cleft lip and palate speech recognition," in *APSIPA ASC*. IEEE, 2021, pp. 484–490.
26. Zengrui Jin, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shujie Hu, Jiajun Deng, Guinan Li, and Xunying Liu, "Adversarial data augmentation using vae-gan for disordered speech recognition," in *ICASSP*. IEEE, 2023, pp. 1–5.
27. Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu, "Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation," in *ICASSP*. IEEE, 2024, pp. 12311–12315.

28. Zengrui Jin, Xurong Xie, Tianzi Wang, Mengzhe Geng, Jiajun Deng, Guinan Li, Shujie Hu, and Xunying Liu, "Towards automatic data augmentation for disordered speech recognition," in *ICASSP*. IEEE, 2024, pp. 10626–10630.
29. John Harvill, Dias Issa, Mark Hasegawa-Johnson, and Changdong Yoo, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *ICASSP*. IEEE, 2021, pp. 6428–6432.
30. Yunxin Zhao, Minguang Song, Yanghao Yue, and Mili Kuruvilla-Dugdale, "Personalizing tts voices for progressive dysarthria," in *BHI*. IEEE, 2021.
31. Marc Illa, Bence Mark Halpern, Rob van Son, Laureano Moro-Velazquez, and Odette Scharenborg, "Pathological voice adaptation with autoencoder-based voice conversion," in *SSW 11*, 2021, pp. 19–24.
32. Helin Wang, Thomas Thebaud, Jesús Villalba, Myra Sydnor, Becky Lammers, Najim Dehak, and Laureano Moro-Velazquez, "Duta-vc: A duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model," in *INTERSPEECH 2023*, 2023, pp. 1548–1552.
33. Andrew Hu, Dhruv Phadnis, and Seyed Reza Shahamiri, "Generating synthetic dysarthric speech to overcome dysarthria acoustic data scarcity," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 6751–6768, 2023.
34. Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee, "Towards audio language modeling-an overview," *arXiv preprint arXiv:2402.13236*, 2024.
35. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
36. Cristina Espana-Bonet and José AR Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *IberSPEECH*. Springer, 2016, pp. 97–107.
37. Neethu Mariam Joy and Srinivasan Umesh, "Improving acoustic models in torgo dysarthric speech database," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, 2018.
38. Patrick Esser, Robin Rombach, and Bjorn Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12873–12883.
39. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
40. Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al., "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP*. IEEE, 2022, pp. 6182–6186.
41. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. PMLR, 2023, pp. 28492–28518.