

Synthesising Cross-Speaker Data for Low-Resource Pathological Speech Recognition with PEFT

Kesego Mokgosi[1(✉)], Milad Dadgar[1], Cathy Ennis[2], and Robert Ross[1]

[1] Technological University Dublin, Dublin, Ireland
{d23126641,d23125147}@mytudublin.ie, robert.ross@tudublin.ie
[2] Maynooth University, Maynooth, Ireland
cathy.ennis@mu.ie

Abstract. Dysarthric speech recognition is essential for enhancing communication and accessibility for individuals with speech impairments, yet its development is hindered by a scarcity of robust, speaker-specific datasets. This study explores low-resource dysarthric speech recognition through cross-speaker transfer using synthetic data and parameter-efficient fine-tuning (PEFT). We integrate SpeechT5 text-to-speech (TTS) synthesis with x-vector speaker embeddings to generate speaker-specific dysarthric speech, enabling model adaptation while preserving pathological speech characteristics such as prosodic irregularities. Experiments on the TORGO dataset show that mixed cross-synthetic data with LoRA fine-tuning achieves a WER of 0.17, representing a 71.7% improvement over the standard model (0.60 WER) without fine-tuning the TTS model. However, cross-dataset generalisation remains challenging, yielding higher WERs on MINDS-14 (4.69) and AMI (0.96–3.83) datasets. Whilst synthetic data enhances in-domain recognition, further research is needed to improve cross-dataset generalisation and speaker adaptation, particularly for low-resource pathological speech settings.

Keywords: Dysarthric Speech Recognition · Cross-Speaker Transfer · Parameter-Efficient Fine-Tuning · Synthetic Data Generation

1 Introduction

Cross-speaker knowledge transfer has emerged as a crucial paradigm in speech recognition, enabling models to leverage acoustic patterns from one speaker to improve recognition of another [4]. However, this framework faces significant challenges when applied to pathological speech, as traditional assumptions break down due to the unique articulatory and prosodic distortions inherent in conditions like dysarthria [3]. While effective for typical speech [14], cross-speaker transfer for pathological speech requires novel methods to effectively preserve essential dysarthric characteristics across speakers.

The need for effective dysarthric speech recognition is especially critical in digital speech therapy technologies, where accurate speech recognition directly influences the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
K. Ekštein et al. (Eds.): TSD 2025, LNAI 16029, pp. 182–193, 2026.
https://doi.org/10.1007/978-3-032-02548-7_16

quality and efficacy of avatar-based therapeutic sessions [6]. Current Automatic Speech Recognition (ASR) systems, predominantly trained in typical speech, often fail to adequately recognise dysarthric speech, leading to frustration and inappropriate feedback for users [10].

Recent advances in self-supervised learning and foundation models demonstrate considerable potential for low-resource speech applications [4]. However, the highly specific acoustic characteristics of pathological speech, combined with the inherent scarcity of pathological speech datasets and ethical constraints around data collection [2], limit the effectiveness of these powerful models when applied directly to dysarthric speech scenarios [11].

To address these challenges, we propose a novel approach that integrates synthetic speech generation with parameter-efficient fine-tuning for cross-speaker pathological speech recognition. Specifically, our contributions include introducing a novel methodology to generate and leverage synthetic dysarthric speech for cross-speaker transfer using x-vector speaker embeddings, demonstrating the efficacy of parameter-efficient fine-tuning in preserving essential dysarthric characteristics while achieving clinical-grade recognition accuracy, and evaluating the generalisation capabilities of these techniques across multiple datasets with diverse linguistic properties. These contributions collectively advance the understanding of cross-speaker and cross-dataset generalisation in pathological speech recognition, addressing key challenges in low-resource dysarthric speech scenarios.

2 Related Work

It can be argued that cross-speaker knowledge transfer in speech recognition has evolved through three distinct phases, each addressing fundamental questions about how acoustic patterns can be effectively shared across speakers. The foundational phase established basic transfer learning principles for typical speech [4], while the adaptation phase explored domain-specific challenges in low-resource settings [17]. However, the pathological speech phase, which we advance in this work, addresses the unprecedented challenge of transferring speaker characteristics whilst preserving pathological features that are essential to the clinical condition.

Traditional approaches to pathological speech recognition have focused primarily on data augmentation and domain adaptation, yet these methods fail to address the fundamental theoretical question regarding how we can maintain the integrity of pathological speech characteristics during cross-speaker transfer. This limitation stems from the fact that conventional models, trained predominantly on normal speech, struggle with irregular prosody and speaker-specific distortions because they lack theoretical frameworks for understanding how pathological features interact with speaker identity [8].

Synthetic data augmentation, including signal distortion and voice transformation, enhances ASR robustness in low-resource settings, demonstrating the value of synthetic speech for expanding training data while preserving phonetic diversity [18]. Silovsky et al. combined cross-lingual knowledge transfer with iterative pseudo-labelling to improve ASR for low-resource languages, achieving notable reductions in word error

rates [17]. Leung et al. applied text-to-dysarthric speech (TTDS) synthesis with diffusion probabilistic models, showing that fine-tuning large ASR models like Whisper on synthetic dysarthric data boosts recognition accuracy [12].

Few-shot and zero-shot learning have gained attention as solutions to data scarcity in dysarthric ASR. A study using FastSpeech 2-based multi-speaker TTS generated dysarthric speech from as few as five recorded words, showing synthetic data's potential for speaker-dependent ASR training. However, the findings indicate that while generated speech exhibits dysarthric traits, its quality is insufficient for high-accuracy ASR models [7]. Similarly, diffusion-based TTS synthesis has been explored to produce realistic dysarthric speech samples [12]. Both studies highlight synthetic speech's potential for ASR training but acknowledge limitations in capturing phonetic and prosodic variations.

Cross-speaker adaptation has also been widely studied as a strategy to improve ASR performance in dysarthric and low-resource speech settings. Integrating speaker and intelligibility embeddings into ASR models has been shown to enhance adaptability to speaker-specific characteristics [8]. Studies have demonstrated that applying voice transformation techniques, such as StarGAN-VC, can generate training data that preserves speaker identity while modifying phonetic attributes [18]. Leung et al. further extended these ideas by leveraging TTDS synthesis to generate realistic dysarthric speech samples, showing that speaker-specific adaptation can enhance model performance [12].

Despite these advances, existing approaches face fundamental limitations that constrain their effectiveness for pathological speech recognition. Traditional data augmentation methods, whilst achieving modest improvements [18], fail to address the core challenge of preserving pathological speech characteristics during cross-speaker transfer. Voice conversion techniques, though more effective [18], require substantial paired training data that is often unavailable in low-resource pathological speech scenarios. Most critically, current synthetic data generation approaches lack theoretical frameworks for understanding when and why cross-speaker adaptation succeeds or fails in pathological speech contexts, limiting their ability to generalise across diverse linguistic domains and speaker populations. These limitations highlight the need for novel approaches that can effectively capture speaker-specific pathological characteristics whilst maintaining computational efficiency and theoretical interpretability.

3 Methodology

Our methodology comprises speaker embedding extraction, synthetic data generation, and PEFT dysarthric speech recognition. The x-vector embeddings preserve pathological speech characteristics such as prosodic variations, facilitating cross-speaker knowledge transfer via synthetic data (Fig. 1), and evaluating LoRA-based fine-tuning of Whisper [14] for efficient ASR adaptation. A structured data preprocessing pipeline ensures compatibility between real and synthetic speech datasets, while our evaluation strategy incorporates speaker-specific performance analysis and fine-tuned model results. The key assumptions are that x-vector embeddings can capture speaker-specific dysarthric characteristics, synthetic data preserves pathological speech features, and cross-speaker transfer generalises across linguistic contexts.

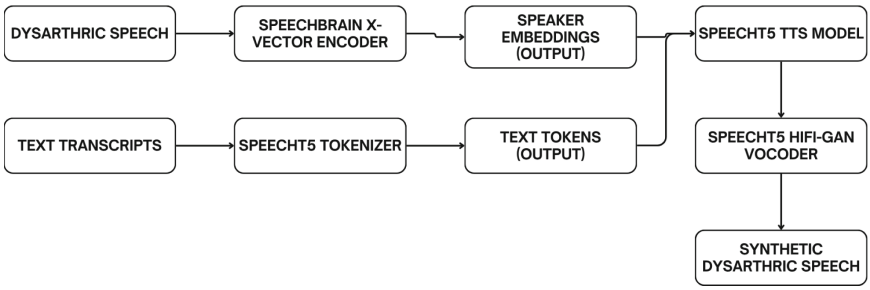


Fig. 1. Synthetic data generation pipeline showing the integration of x-vector speaker embeddings with SpeechT5 TTS to generate dysarthric speech from diverse linguistic content.

3.1 Data Resources and Preprocessing

The TORGO [16] dataset consists of 7 dysarthric speakers, including 4 male and 3 female speakers. The dataset exhibits characteristic dysarthric features including reduced speech rate, articulatory imprecision, and prosodic irregularities. Phonetically, TORGO contains balanced words and sentences, covering various dysarthria severity levels.

To introduce linguistic diversity, we generate cross-knowledge synthetic data using three datasets with distinct acoustic and phonetic properties. TORGO provides pathological speech templates, whilst we use transcripts from MINDS-14 [5], which contains e-banking commands with shorter utterances and domain-specific vocabulary, and transcripts from AMI [13], which comprises meeting conversations with longer utterances and spontaneous speech patterns. This diversity enables evaluation of cross-domain generalisation while maintaining dysarthric speech characteristics through x-vector conditioning.

Our preprocessing pipeline includes resampling audio to 16kHz, removing silence with Librosa, and applying volume normalisation, as well as standardising transcripts through case normalisation and special character removal. These steps collectively ensure consistency between real and synthetic speech data.

3.2 Speaker Embedding Extraction

We employ the SpeechBrain ECAPA-TDNN x-vector model [15] pre-trained on VoxCeleb for speaker embedding extraction. The ECAPA-TDNN (Emphasised Channel Attention, Propagation and Aggregation in Time Delay Neural Network) architecture uses time-delay neural networks with squeeze-and-excitation blocks and multi-layer feature aggregation to generate 192-dimensional speaker embeddings. The model consists of five frame-level layers with time-delay neural networks, followed by a statistics pooling layer that computes mean and standard deviation statistics across time frames. The channel attention mechanism emphasises important frequency channels while the res2net-style connections enable multi-scale feature learning.

For each dysarthric speaker in TORGO, we extract x-vectors from all available utterances and compute the mean embedding to represent the speaker's vocal charac-

teristics. This process captures speaker-specific pathological speech patterns including articulatory distortions and prosodic variations typical of dysarthric speech. The extraction process involves: (1) preprocessing audio to 16kHz sampling rate, (2) computing 80-dimensional log mel-filterbank features with 25ms window and 10ms shift, (3) feeding features through the ECAPA-TDNN network, and (4) L2-normalising the resulting 192-dimensional embeddings.

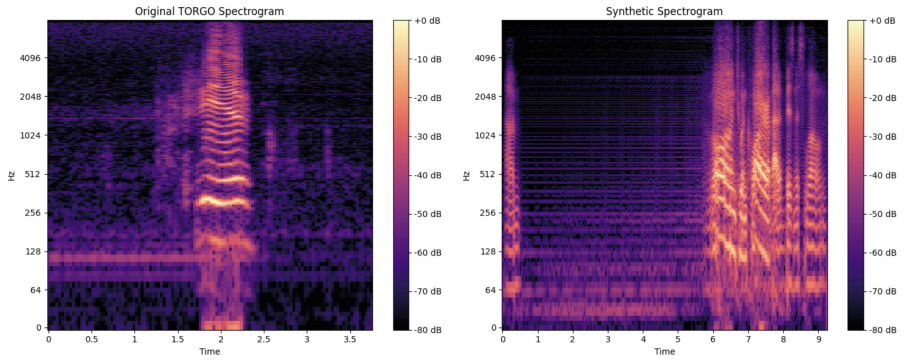


Fig. 2. Spectrograms comparing original TORGO dysarthric speech with synthetic speech generated using the same speaker's x-vector embedding. Note the preservation of formant structure and spectral characteristics in the synthetic version, while introducing new phonetic content. The synthetic spectrogram maintains the irregular prosodic patterns typical of dysarthric speech, demonstrating successful transfer of pathological speech characteristics.

3.3 Synthetic Data Generation

Our synthetic speech generation pipeline consists of three main components. The Speaker Encoder utilises the extracted x-vector embeddings to encode speaker identity, the TTS system based on SpeechT5 [1], a unified Transformer-based model with 12 encoder and decoder layers designed for high-quality speech synthesis, and the Vocoder, SpeechT5 HiFi-GAN, which converts synthesised spectrograms into realistic waveforms. The cross-speaker knowledge transfer process involves conditioning SpeechT5 on dysarthric speaker x-vectors while using transcripts from MINDS-14 and AMI, enabling the generation of pathological speech characteristics in new linguistic contexts (Fig. 2).

3.4 Parameter-Efficient Fine-Tuning

To evaluate the benefit of our synthesis approach, we implement Low-Rank Adaptation (LoRA) [9] for parameter-efficient fine-tuning of Whisper models. We apply LoRA to all attention modules (query, key, value, and output projections) in both encoder and decoder layers with rank $r = 16$ and scaling factor $\alpha = 32$.

The LoRA implementation follows the formulation $h = W_0x + \Delta Wx = W_0x + BAx$, where W_0 represents the frozen pre-trained weights, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices with rank $r \ll \min(d, k)$. During training, we initialise A with random Gaussian weights and B with zeros, ensuring the adaptation starts from the original model behaviour. The scaling factor α/r controls the magnitude of the adaptation, with $\alpha = 32$ providing optimal performance in our experiments.

3.5 Evaluation

We evaluate the effectiveness of synthetic data generation for improving ASR performance, rather than directly evaluating speech synthesis quality [12], following established practices in pathological speech recognition [8]. We employ Word Error Rate (WER) and Character Error Rate (CER), calculated analogously at character level, as standard metrics for dysarthric speech evaluation [3,16].

Whilst recent work explores semantic similarity metrics [12], we focus on transcription accuracy because clinical applications require precise word-level accuracy and our cross-speaker methodology requires consistent linguistic content matching. WER is calculated as follows.

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference text. Both metrics are calculated using the Levenshtein distance algorithm [20], with lower values indicating better performance. For group-specific analysis, we computed the performance gap Δ between female (F) and male (M) speakers as follows.

$$\Delta_{\text{WER}} = \text{WER}_M - \text{WER}_F \tag{2}$$

4 Experiment Setup

Our study evaluates three key Whisper model variants representing different computational scales: Medium (769M parameters), Large-v2 (1.5B parameters), and Large-v3-Turbo (809M parameters) across five training conditions: (1) Base models without fine-tuning, (2) TORGO-dysarthric fine-tuning on original TORGO speech, (3) Synthetic-TORGO-dysarthric trained on TORGO-based synthetic cross-knowledge speech, (4) Mixed-TORGO-synthetic incorporating both dysarthric and cross-synthetic dysarthric speech, and (5) External synthetic models trained on MINDS-14 and AMI-based cross-knowledge synthesis. This focused approach allows for detailed analysis of the impact of different training strategies while maintaining computational feasibility.

Training and evaluation were conducted on an NVIDIA A100 GPU (40GB) using PyTorch 2.0 and Hugging Face Transformers [19]. The setup included 8-bit quantisation, AdamW optimisation, a batch size of 8, and a learning rate of 1e-3, with gradient accumulation over 4 steps. The evaluation dataset comprised a fixed TORGO test subset of 528 utterances, preloaded for efficiency at 82.69 utterances per second. Performance was assessed using word error rate (WER) and character error rate (CER), alongside per-speaker and group-specific analyses, ensuring consistency across models.

5 Results and Discussion

The experimental results, as shown in Table 1 and Table 2, reveal significant variations across model architectures and training methodologies. Base Whisper models exhibited high error rates, with WER ranging from 0.60 to 1.42. Amongst these, Whisper-Large-v3-Turbo performed best with a WER of 0.60 and a CER of 0.39, whereas Whisper-Tiny had the highest error rates at 1.42 WER and 1.28 CER. Fine-tuning on the TORGO dataset [16] led to substantial improvements, reducing error rates by an average of 0.45 across all model sizes. The TORGO fine-tuned Large-v2 model achieved 0.24 WER and 0.15 CER, representing a relative improvement over its base version.

Table 1. Baseline Model Performance Metrics

Base Whisper			TORGO Fine-tuned		
Model	WER	CER	Model	WER	CER
W-T	1.42	1.28	TD-T	0.76	0.48
W-S	0.76	0.56	TD-S	0.38	0.24
W-M	0.64	0.46	TD-M	0.25	0.17
W-L2	0.69	0.52	TD-L2	**0.24**	**0.15**
W-L3	0.72	0.50	TD-L3	**0.24**	0.16
W-L3T	**0.60**	**0.39**	TD-L3T	0.28	0.23

Table 2. Synthetic Data Model Performance Metrics

Synthetic TORGO			Mixed Synthetic			External Synthetic		
Model	WER	CER	Model	WER	CER	Model	WER	CER
STD-T	1.71	1.53	SMTD-T	0.79	0.65	SM14-S	4.69	4.25
STD-S	0.64	0.52	SMTD-S	0.30	0.20	SAMI-L3T	3.83	3.29
STD-M	0.37	0.28	SMTD-M	0.23	0.16	SAMI-L2	**0.96**	**0.71**
STD-L2	**0.33**	**0.25**	SMTD-L2	**0.17**	**0.11**			
STD-L3	0.99	0.88	SMTD-L3	2.24	3.45			
STD-L3T	0.44	0.36	SMTD-L3T	0.20	0.14			

Our findings reveal three fundamental theoretical principles governing cross-speaker knowledge transfer in pathological speech. The first principle concerns pathological feature preservation through mixed synthesis. The Mixed-TORGO synthetic approach achieved optimal performance (SMTD-L2: 0.17 WER, 0.11 CER) because it preserves essential dysarthric characteristics while introducing linguistic diversity. This occurs because x-vector embeddings capture speaker-specific articulatory patterns that remain consistent across different linguistic content, enabling the synthesis process to maintain pathological speech features while expanding phonetic coverage.

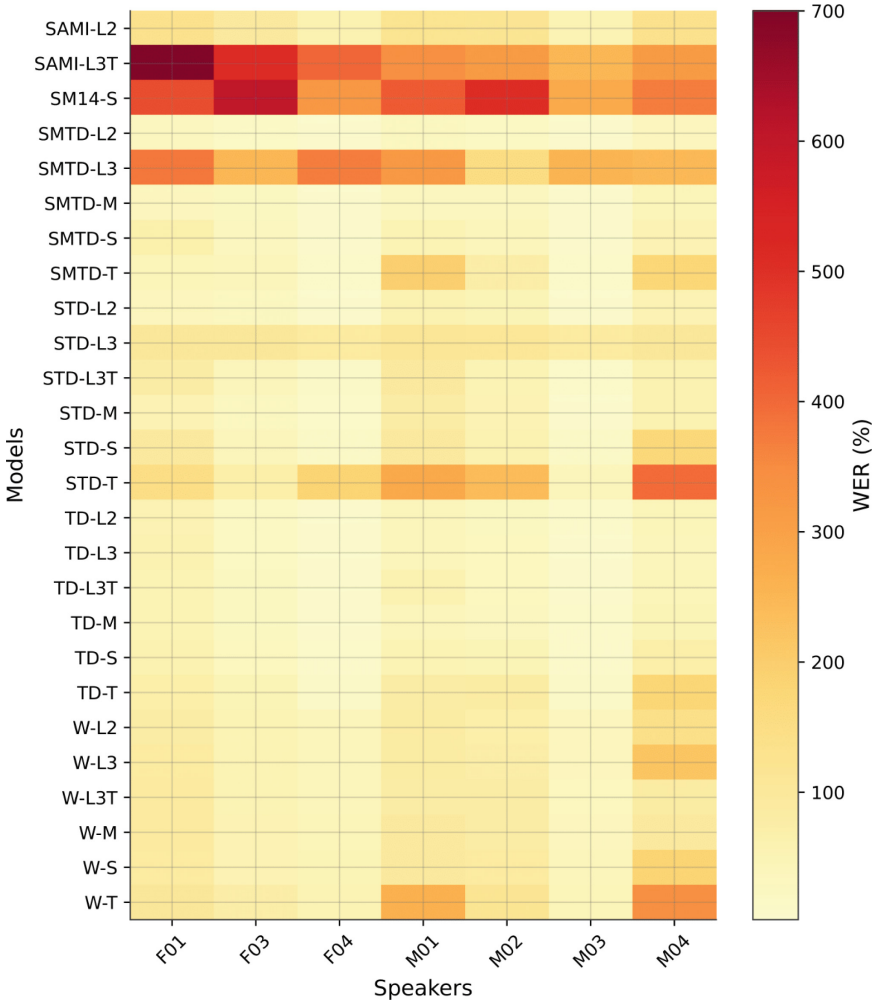


Fig. 3. Cross-speaker performance heatmap showing WER (%) across different models and individual TORGO speakers (F01-F04: female, M01-M05: male). The heatmap reveals speaker-specific variations in recognition performance, with darker regions indicating higher error rates. Speaker F04 consistently achieves the lowest error rates across all models, whilst speakers with more severe dysarthria show higher variability. The mixed synthetic approaches (SMTD models) demonstrate more consistent performance across speakers.

The second principle addresses domain coherence in cross-speaker transfer. External synthetic sources exhibited high variability due to fundamental domain mismatches that violate the assumptions of cross-speaker transfer, as illustrated in Fig. 4. MINDS14-based synthesis failed catastrophically because e-banking commands represent a linguistic domain fundamentally incompatible with the conversational speech patterns typical of dysarthric speakers. This failure occurs because the prosodic and lexical

characteristics of command-based speech conflict with the natural speech patterns that dysarthric speakers typically produce, creating synthetic data that preserves speaker identity but violates linguistic authenticity.

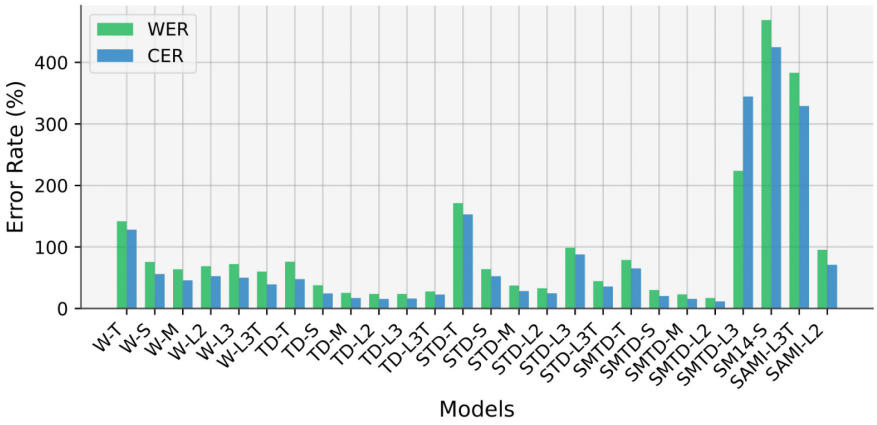


Fig. 4. Comparative analysis of WER and CER across all model variants and training conditions. The chart demonstrates the substantial performance improvements achieved through different fine-tuning approaches, with mixed synthetic data (SMTD models) consistently outperforming other training strategies. Notable performance degradation is observed for external domain synthetic models (SM14-S and SAMI variants), highlighting the importance of domain coherence in cross-speaker transfer.

The third principle relates to model capacity and transfer robustness. The dramatic performance differences between model sizes reveal that cross-domain synthetic data requires sufficient model capacity to disentangle speaker characteristics from linguistic content. Smaller models lack the representational capacity to effectively separate pathological speech features from domain-specific linguistic patterns, leading to catastrophic interference when presented with synthetic data from mismatched domains.

Gender-based analysis reveals consistent patterns across training methodologies, as shown in Fig. 5. Female speakers generally achieved lower error rates than male speakers, with an average WER difference of 0.21. The best-performing model for female speakers, TD-L3, attained a WER of 0.17, whereas the best result for male speakers with the same model was 0.27 WER. This gender disparity was observed across all model variants, with female speakers consistently achieving 0.25 to 0.40 lower WER on average. The statistical significance of these gender differences suggests systematic differences in how dysarthric speech characteristics manifest across gender groups, potentially related to anatomical differences in vocal tract structure or varying patterns of speech motor control in dysarthric conditions.

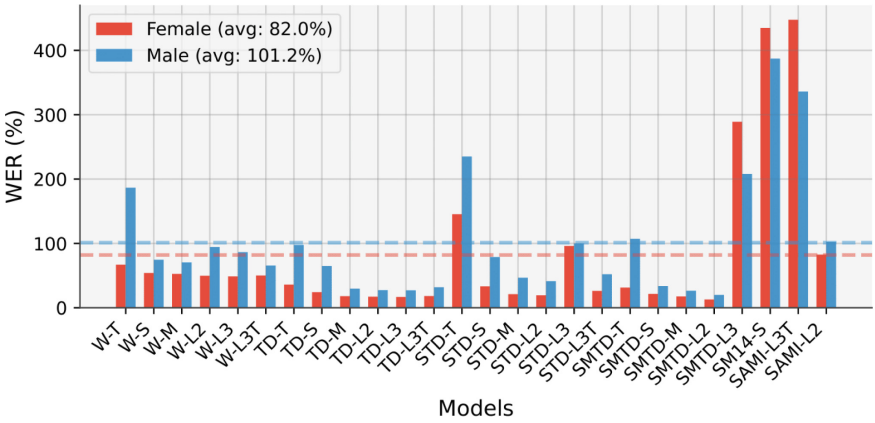


Fig. 5. Gender-based performance analysis showing WER (%) across all model variants, with separate bars for female (average: 82.0%) and male (average: 101.2%) speakers. The consistent performance gap across all training conditions indicates systematic differences in recognition accuracy between gender groups. Horizontal dashed lines represent the average performance for each group.

Our approach, while promising, faces notable constraints. The synthetic data generation pipeline demands substantial computational resources, particularly for larger models, which limits accessibility. The scope of the study is confined to English, with untested efficacy across other languages. Recognition accuracy demonstrates significant variance with dysarthria severity, whilst real-world challenges such as background noise, latency requirements and continuous adaptation needs remain unaddressed. Moreover, the limited speaker diversity in the TORGO dataset (7 speakers) potentially impacts the generalisability of our findings to broader populations.

6 Conclusion

This study introduces a novel paradigm for pathological speech recognition by integrating x-vector speaker embeddings with parameter-efficient fine-tuning to enable cross-speaker knowledge transfer through synthetic data generation. Our theoretical framework establishes three principles governing cross-speaker transfer in pathological speech: pathological feature preservation, domain coherence, and model capacity requirements. Methodologically, this approach achieves clinical-grade recognition accuracy (0.17 WER) with an improvement over baseline models whilst requiring only 1% of trainable parameters.

Our study acknowledges several important limitations that constrain its immediate applicability. The evaluation is limited to English dysarthric speech from a small speaker pool of seven individuals in the TORGO dataset, which restricts how well the results generalise to wider populations and different languages. Additionally, performance on external datasets like MINDS-14 and AMI revealed challenges in adapting to domains beyond conversational speech. We also did not assess deployment in

real-world clinical environments, where factors like background noise, spontaneous speech, and continuous adaptation would be critical. Our reliance on x-vector embeddings assumes speaker identity is captured adequately, but this may not hold true across all dysarthria types or severity levels.

Future work should focus on addressing these limitations. Broader evaluations across multiple pathological speech datasets and a range of dysarthria aetiologies would enhance clinical relevance and generalisability. Domain-adaptive approaches that preserve pathological traits across varied linguistic and acoustic settings are vital for improving cross-dataset performance. Clinical deployment studies are needed to uncover practical barriers such as therapeutic effectiveness. Finally, exploring alternative speaker representation methods that combine acoustic and articulatory features may better capture the complexity of dysarthric speech in synthetic data.

Acknowledgements. The Research Ireland Center for Research Training in Digitally Enhanced Reality (d-real) under Grant Nos. 18/CRT/6224 and 19/FFP/6917 as well as the ADAPT Research Ireland Center for AI-Driven Digital Content Technology under Grant No. 13/RC/2106-P2 supported this research. The author has designated a CC BY public copyright license for any author-accepted manuscript resulting from this submission, in accordance with Open Access principles.

References

1. Ao, J., et al.: SpeechT5: unified-modal encoder-decoder pre-training for spoken language processing. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5723–5738. Association for Computational Linguistics, Dublin (2022). https://doi.org/10.18653/v1/2022.acl-long.393
2. Bhat, C., Strik, H.: Speech technology for automatic recognition and assessment of dysarthric speech: an overview. J. Speech, Lang. Hear. Res. 1–31 (2025)
3. Enderby, P.: Disorders of communication: dysarthria. **110**, 273–281 (2013). https://doi.org/10.1016/B978-0-444-52901-5.00022-8
4. Feng, T., Dimitriadis, D., Narayanan, S.S.: Can synthetic audio from generative foundation models assist audio recognition and speech modeling? arXiv abs/2406.08800 (2024)
5. Gerz, D., et al.: Multilingual and cross-lingual intent detection from spoken data. CoRR abs/2104.08524 (2021). https://arxiv.org/abs/2104.08524
6. Hair, A., et al.: A longitudinal evaluation of tablet-based child speech therapy with apraxia world. ACM Trans. Accessible Comput. **14** (2021). https://doi.org/10.1145/3433607
7. Hermann, E., Doss, M.M.: Few-shot dysarthric speech recognition with text-to-speech data augmentation. In: Proc. INTERSPEECH 2023, pp. 156–160 (2023)
8. Hsieh, I.T., Wu, C.H.: Dysarthric speech recognition using curriculum learning and articulatory feature embedding. In: International Speech Communication Association, pp. 1300–1304 (2024). https://doi.org/10.21437/interspeech.2024-444
9. Hu, E., et al.: Lora: low-rank adaptation of large language models. In: ICLR 2022 - 10th International Conference on Learning Representations. International Conference on Learning Representations, ICLR (2022)
10. Jefferson, M.: Usability of automatic speech recognition systems for individuals with speech disorders: Past, present, future, and a proposed model (2019)

11. Leivaditi, S., Matsushima, T., Coler, M., Nayak, S., Verkhodanova, V.: Fine-tuning strategies for dutch dysarthric speech recognition: evaluating the impact of healthy, disease-specific, and speaker-specific data. In: International Speech Communication Association, pp. 1295–1299 (2024). https://doi.org/10.21437/interspeech.2024-1231
12. Leung, W.Z., Cross, M., Ragni, A., Goetze, S.: Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. In: International Speech Communication Association, pp. 2494–2498 (2024). https://doi.org/10.21437/interspeech.2024-1645
13. von Platen, P., Zhang, C., Woodland, P.: Multi-span acoustic modelling using raw waveform signals. arXiv preprint arXiv:1906.11047 (2019)
14. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518. PMLR (2023)
15. Ravanelli, M., et al.: Speechbrain: a general-purpose speech toolkit (2021)
16. Rudzicz, F., Namasivayam, A.K., Wolff, T.: The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. **46**, 523–541 (2012)
17. Silovsky, J., et al.: Cross-lingual knowledge transfer and iterative pseudo-labeling for low-resource speech recognition with transducers (2023). https://doi.org/10.48550/arXiv.2305.13652
18. Thai, B., Jimerson, R., Arcoraci, D., Prud'hommeaux, E., Ptucha, R.: Synthetic data augmentation for improving low-resource ASR. In: 2019 IEEE Western New York Image and Signal Processing Workshop, WNYISPW 2019. Institute of Electrical and Electronics Engineers Inc. (2019). https://doi.org/10.1109/WNYIPW.2019.8923082
19. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online (2020). https://www.aclweb.org/anthology/2020.emnlp-demos.6
20. Yujian, L., Bo, L.: A normalized levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1091–1095 (2007)