Improved Dysarthric Speech to Text Conversion via TTS Personalization

Péter Mihajlik*‡, Éva Székely†, Piroska Barta*, Máté Soma Kádár‡§, Gergely Dobsinszki§*, László Tóth¶
*Department of Telecommunications and Artificial Intelligence, Budapest University of Technology, Hungary
†Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
‡Hungarian Research Centre for Linguistics, HUN-REN, Hungary
§SpeechTex Ltd., Hungary
¶Institute of Informatics, University of Szeged, Hungary

***Abstract*—We present a case study on developing a customized speech-to-text system for a Hungarian speaker with severe dysarthria. State-of-the-art automatic speech recognition (ASR) models struggle with zero-shot transcription of dysarthric speech, yielding high error rates. To improve performance with limited real dysarthric data, we fine-tune an ASR model using synthetic speech generated via a personalized text-to-speech (TTS) system. We introduce a method for generating synthetic dysarthric speech with controlled severity by leveraging premorbidity recordings of the given speaker and speaker embedding interpolation, enabling ASR fine-tuning on a continuum of impairments. Fine-tuning on both real and synthetic dysarthric speech reduces the character error rate (CER) from 36–51% (zero-shot) to 7.3%. Our monolingual FastConformer_Hu ASR model significantly outperforms Whisper-turbo when fine-tuned on the same data, and the inclusion of synthetic speech contributes to an 18% relative CER reduction. These results highlight the potential of personalized ASR systems for improving accessibility for individuals with severe speech impairments.**

***Index Terms*—automatic speech recognition, dysarthric speech, text to speech synthesis, few-shot learning**

I. Introduction

Certain medical conditions can significantly impact an individual's speech production over the long term. For instance, stroke-related cerebral injuries may lead to dysarthric speech, which can be challenging to understand. Medically, dysarthria is an umbrella term for any neuromotor disorder that results in abnormal speech control. It might influence the speed, strength, accuracy, range, tone, duration etc. of the speech signal, depending on its actual etiology [1]. Although dysarthric speech often exhibits a significantly altered articulation, leading to low or even near-zero general intelligibility, the speech errors tend to be consistent. As a result, close acquaintances can often learn to understand the altered speech patterns. Therefore, adapting Automatic Speech Recognition (ASR) systems for the transcription of dysarthric speech is a promising approach, as evidenced by numerous previous studies [2]. The challenges of dysarthric ASR, however, are not restricted to the disordered speech signal: data collection can be particularly difficult not only because speech production can be slow and tiresome to the speakers affected but also because the human transcription (and its verification) needs special skills and the active involvement of people suffering from dysarthria. As a result, only a limited amount of dysarthric speech data is available for research and much of it consists of isolated words or expressions [3].

Data scarcity poses a significant challenge, particularly when developing a general speech-to-text system for dysarthric speech, especially in languages with relatively few speakers. Moreover, dysarthria manifests in multiple forms [2], making it practically infeasible to design a single ASR system that accommodates all variations. For assistive technology, personalizing ASR for a single user is a practical approach, as the system is designed solely for their speech patterns [4].

In this study, we present our results in developing an assistive Hungarian-language speech-to-text system for a person with dysarthria, adapting the ASR system using only a few minutes of dysarthric speech. To mitigate overfitting in large end-to-end neural acoustic models fine-tuned on limited data, we first develop a personalized text-to-speech (TTS) system to generate dysarthric speech data for augmentation purposes. Results show that even in this few-shot scenario, the high initial error rate of a foundational ASR model can be significantly reduced. Further improvements are achieved by incorporating TTS-generated speech, synthesized using multiple methods to reflect different levels of dysarthria and intelligibility. Consequently, we achieved a Character Error Rate (CER) as low as 7.3% on an independent evaluation set from the same speaker. This suggests that our end-to-end ASR model is already suitable for real-life dysarthric speech transcription and could pave the way for assistive ASR technology for others with similar speech disorders.

II. Related work

Earlier studies focused on increasing intelligibility [5] – a direction still active today [6]. As regards ASR, the first attempts aimed at recognizing isolated words only [7], [8]. With the advent of the deep neural technologies the the demand for larger amounts of training data has further increased, leading to the ubiquitous use of data augmentation and the utilization synthetic data [9], [10]. Hence, many of the current papers on dysarthric ASR focus on how to convert normal speech into 'dysarthric' [11], or even to synthesize dysarthric speech to use it for the purpose of data augmentation in ASR.

This work was supported by the Hungarian NRDI Fund through projects NKFIH K143075/K135038, NKFIH-828-2/2021 and by KIFÜ Kommondor.

In the context of dysarthric speech, the benefit of TTS augmentation was shown in [12], in a few-shot learning scenario. Using FastSpeech 2-based multi-speaker TTS [13] with a learned dysarthria embedding, it was shown that synthetic dysarthric speech improves isolated word recognition of an unseen dysarthric speaker. Notably, while synthetic speech alone was not sufficient to train a model from scratch, combining a small amount of real dysarthric audio with abundant synthetic speech outperformed using only the limited real data. This finding underlines that TTS-generated data can quickly adapt ASR to new dysarthric speakers or commands, especially when paired with even a few authentic samples. [14] applied a multi-speaker TTS approach that injects dysarthria-specific controls into a neural TTS system. Their method adds a dysarthria severity level embedding and a pause-insertion mechanism to an end-to-end TTS, enabling the synthesis of dysarthric speech at various severity levels. Augmenting a dysarthria-tailored ASR system with such synthetic data significantly reduced error rates: a DNN-HMM ASR model trained with the synthetic dysarthric speech achieved a 12.2% WER reduction over the baseline, and the explicit severity/pause modeling provided an additional 6.5% WER decrease. [15] trained a diffusion-based Grad-TTS [16] model from scratch on dysarthric data to create synthetic samples without any parallel typical recordings. Using these samples to fine-tune Whisper led to significant improvements: for the TORGO dysarthric dataset [17], adding generated data reduced WER from 56.1% (using only real data) to 20.1%.

While most augmentation studies aim to improve generic ASR models across a variety of dysarthric speakers [18], an alternative research focus is tailoring ASR to individual users. Personalized dysarthric ASR systems – trained on a single user's speech patterns – have been shown to outperform generalized models on the given user's speech [4] for short phrases, or using small natural dysarthric corpora [19]. Our approach aligns with this personalized paradigm, specifically for assistive technology. By leveraging a pre-dysarthria speech corpus from the user (to capture their voice identity) and adapting a foundation TTS model to produce dysarthric speech, we effectively reproduce the user's impaired voice for data augmentation. This strategy, inspired by the successes of multi-speaker and fine-tuned TTS in prior work, allows us to bootstrap high-accuracy general speech-to-text conversion with minimal real dysarthric recordings.

III. Data

A. Dysarthric Speech Corpora

Altogether **less than an hour** of in-domain, Hungarian language transcribed dysarthric speech was available for the research. For the training set, phonetically rich sentences were read by the target (stroke survivor) speaker. For the validation set, a short story was read. As for the evaluation, sentences of everyday life were read from the same speaker. Statistics of the dysarthric speech data sets are shown on Table I.

B. Fluent Speech Corpora

In our particular case, the target speaker suffering from dysarthria recorded **13 hours** of spontaneous speech in **lecture materials** (as part of his profession as a university teacher) prior to the stroke event. Although such amount of healthy speech might not be typical for people developing dysarthria in general, due to the wide spread use of social media, similar scenario might become more common in the future. This collection of lecture speech is, however, unlabeled.

Additionally, the same **195 phonetically rich sentences** read in the Train-dys set were recorded previously (before stroke) as well by the target speaker. This is the only supervised, i.e., exactly transcribed, fluent speech data set from the given subject.

C. Synthetic Dysarthric Speech Corpora

As the amount of in-domain (dysarthric) speech was minuscule – even for speaker-specific ASR – and because of the domain mismatch regarding the fluent (unlabeled) speech from the same speaker, we have decided to create more dysarthric data by speech synthesis utilizing the previous data sources and a foundational TTS model. The TTS system XTTS-v2 [20] was selected for the generation of dysarthric Hungarian speech for the target speaker, because it is an open-sourced multilingual deep neural network based model that includes support for the Hungarian language[1].

1) Adaptation to the target speaker's original voice: The corpus of 13 hours of lectures was segmented into utterances delineated by breath events following [21]. To select the utterances from the lecture recordings which are most suitable for TTS training (regarding fluency, intelligibility) we employed an ASR ensemble technique using automatic transcriptions created with both BEAST2 [22] and Whisper [23]. A subset of the lecture corpus was created by comparing the two ASR transcriptions and selecting the utterances where the two transcriptions had an edit-distance of max 4 characters. This yielded 1242 utterances, corresponding to **2 hours of lecture speech** associated with automatic transcriptions (pseudo-labels).

2) Fine-tuning on dysarthric speech: To approximate the increased variability of dysarthric speech resulting from speakers articulating words with varying degrees of success, we leverage premorbidity recordings of the same speaker to generate synthetic dysarthric speech at different severity levels. Since these recordings provide a clean baseline of the

[1]https://huggingface.co/coqui/XTTS-v2

TABLE I
SINGLE-SPEAKER DYSARTHRIC SPEECH DATA SETS

	Train-dys	Val-dys	Test-dys
Duration [min]	21	6	17
# segments	195	40	107
# words	1406	363	1382
# chars	9523	2331	8326

speaker's healthy voice, we fine-tune the speaker-adapted TTS model on the 195 recorded sentences of dysarthric speech (Train-dys) to gradually shift the synthesis away from fluent articulation toward impaired speech. By controlling the extent of fine-tuning and interpolating between fluent and dysarthric embeddings, we aim to create a continuum of severity that better reflects the fluctuating nature of dysarthric speech. To enable synthesis with different severity levels of dysarthria, we saved 3 checkpoints:

- *FT-Dys-Undertrained* (training stopped after 2000 iterations, while validation loss was still decreasing)
- *FT-Dys-Best* (optimally trained, validation loss minimised after 3200 iterations)
- *FT-Dys-Overtrained* (trained for 3000 iterations after validation loss minimised; overfitted to dysarthric speech)

Additionally, the speaker-adapted XTTS model was fine-tuned on the 195 recorded sentences of both dysarthric speech and on the 195 recorded sentences of typical speech using 2 separate speaker embeddings until the validation loss was minimized; we call this model FT-Dys-Co-trained. The goal of this approach was to be able to scale the level of dysarthria by weighting the two embeddings.

3) Synthesizing dysarthric Speech: We carefully selected the text material for synthesis: 5000 sentences were filtered from the spoken language subset of the Hungarian Gigaword corpus [24] with lengths of 9–11 words not containing foreign words or abbreviations.

We synthesized these sentences with *FT-Dys-Undertrained, FT-Dys-Best, FT-Dys-Overtrained*, using temperature = 0.9, repetition-penalty = 6 parameters. To ensure maximum variability, the 195 dysarthric sentences in the training data were looped through as reference audio so each of them was used 12-13 times. [20] reports that both speaker and style characteristics of the reference audio are copied during synthesis.

FT-Dys-Co-trained was used to synthesize 4 additional datasets. We modified the interpolation function of XTTS – which originally performed an equal-weight blend of speaker embeddings –, to support explicit weighted combinations. The new function accepts arbitrary weight vectors – when the weights sum to one, they yield a convex combination that precisely controls the relative contributions of each speaker embedding. Moreover, by permitting negative weight values, the function facilitates *extrapolation* in the speaker embedding space, thereby enabling modulation of speaker characteristics beyond the range defined by the reference embeddings. Using the same looping method for reference audios from each corpus [typical, dysarthric], we synthesized the 5000 sentences with settings A:[0.2, 0.8], B:[0.0, 1.0], C:[-0.5, 1.5], D:[-1.5, 2.5], (temperature=0.9, repetition-penalty=6). The total duration of the resulted synthesized data sets varied between 10 and 12 hours.

IV. Zero-Shot ASR Experiments

A. Comparison of ASR Engines on Dysarthric Speech

The aim of these experiments was to assess the difficulty of automatic transcription of dysarthric speech in Hungarian from the target speaker by comparing various ASR engines. As Table II shows, state-of-the-art multilingual and monolingual deep neural approaches fail to deliver a transcription with any practical usability. The last one 'FastConformer_Hu', is our in-house ASR model [25] trained on 5k hours of Hungarian broadcast and conversational speech following the 'FastConformer-BPE-CTC' recipe of the NVIDIA NeMo toolkit [26] and using the FastConformer architecture [27]. As it consistently achieved the lowest error rates, we focused on the FastConformer_Hu ASR model (115 Million parameters), but control experiments were conducted with the Whisper-large-v3-turbo model (Whisper-turbo, in short, with 809 Million parameters), as well.

TABLE II
Zero-shot (baseline) ASR results [%] on single speaker Hungarian dysarthric speech data sets

ASR engine	Train-dys		Val-dys		Eval-dys	
	WER	**CER**	**WER**	**CER**	**WER**	**CER**
Gemini 2.0 Flash[2]	101	66	74	45	76	51
Whisper-turbo [23]	107	69	62	34	68	44
BEAST2 [22]	76	44	68	38	68	43
FastConformer_Hu	72	36	58	22	65	36

B. Intelligibility-related ASR tests on Synthesized Data

Using a well-trained general ASR model for re-transcribing synthesized speech and evaluating WER/CER is a common practice to estimate the intelligibility of various TTS approaches. Based on the previous results, we have decided to apply the FastConformer_Hu model for this purpose. In Table III it can be clearly seen that the under- and overtrained XTTS models gave lower and higher WER/CER, respectively, as expected. By comparing the Co-trained A-D approaches again, a monotonic increase (decreasing intelligibility) can again be observed. To verify the appropriateness of the FastConformer_Hu ASR model, typical (premorbidity) speech data sets from the target speaker were additionally transcribed and evaluated, confirming that fluent read and lecture speech is the easiest to transcribe (and understand). Based on the WER/CER, we can conclude that the various synthetic corpora resemble different severity levels of dysarthric speech. However, as expected with fine-tuning approaches, the WER/CER of the natural dysarthric speech is still higher (Train-dys, see Table II).

V. ASR Model Fine-tuning on Dysarthric Speech

In this set of experiments the FastConformer_Hu and Whisper-turbo ASR models used in the previous (zero-shot) experiments were fine-tuned on various dysarthric speech data sets. The Fastconformer_Hu was fine-tuned in the NeMo environment [26], keeping the original recipe by default. The learning-rate was optimized for each setup, and the following hyperparameters were kept constant: batch size=16, number of

[2]https://deepmind.google/technologies/gemini/flash/

[3]Pseudo-labels were used as references.

TABLE III
ZERO-SHOT FASTCONFORMER_HU ASR RESULTS [%] ON SYNTHESIZED DYSARTHRIC SPEECH AND ON HEALTHY CONTROL SPEECH

Data set	WER	CER
FT-Dys-Undertrained	44	16
FT-Dys-Best	46	17
FT-Dys-Overtrained	60	26
FT-Dys-Co-trained-A	35	12
FT-Dys-Co-trained-B	45	17
FT-Dys-Co-trained-C	54	23
FT-Dys-Co-trained-D	57	26
Fluent phon. rich sentences	4.1	0.3
2 hours of lectures[3]	9.8	2.4

TABLE IV
FINE-TUNED **FASTCONFORMER_HU** ASR RESULTS [%] MEASURED ON DYSARTHRIC SPEECH

	Val-dys		Eval-dys	
Fine-tuning data sets	**WER**	**CER**	**WER**	**CER**
Train-dys	28.4	9.2	21.9	8.9
Train-dys + FT-Dys-Undertrained	24.0	8.1	22.5	9.2
Train-dys + FT-Dys-Best	22.8	7.7	19.5	7.7
Train-dys + FT-Dys-Overtrained	**22.6**	**7.4**	19.0	7.8
Train-dys + FT-Dys-Co-trained-A	24.0	7.8	21.0	8.8
Train-dys + FT-Dys-Co-trained-B	24.2	8.3	22.7	9.4
Train-dys + FT-Dys-Co-trained-C	24.3	8.9	22.7	9.5
Train-dys + FT-Dys-Co-trained-D	25.9	8.9	23.3	10.0
Train-dys + FT-Dys-Best + Over	26.7	8.6	**18.3**	**7.3**
Train-dys + All (7) FT-Dys	25.1	9.0	19.7	8.3
Train-dys + 2 hours of lectures	27.6	9.1	26.0	10.7
Fluent phon. rich sentences	54.8	22.3	59.3	31.8

iterations=16k, AdamW optimizer with betas [0.8, 0.9], weight decay=0.02, cosine annealing, minimum learning rate=1e-7. In case of Whisper-turbo we used the HuggingFace Transformer library and recipe[4]. Similarly to the other ASR model, learning rate was optimized for each setup and we fixed the maximum number of steps to 5000 for fine-tuning. All experiments were run on one RTX 5000 Ada GPU.

We found that leaving the original dysarthric train set (Train-dys) out always deteriorated the results, so, by default, it was added to the fine-tune sets. Beyond investigating the effect of fine-tuning on each ∼10 hours long synthetic dysarthric speech data set (along with Train-dys), we also applied combinations of synthetic sets (Best + Overtrained vs. all of them). In the experiments, by default, the ratio of original/synthesized speech (21 min/10 hours) was kept constant. Finally we conducted control experiments with mixed fine-tune sets (21 min. from Train-dys + 2 hours of fluent lectures) and by fine-tuning on fluent speech only. For results of FastConformer_Hu and Whisper-turbo, see Table IV and Table V, respectively.

VI. RESULTS AND DISCUSSION

As the results show, ASR models fine-tuned on dysarthric speech gave dramatically lower error rates on both the in-domain validation and evaluation sets containing speech with severe dysarthria than the baseline systems (Table II). Moreover, many of the synthesized dysarthric corpora further reduced the error rates on the validation, evaluation sets, or both. Considering our most important metric, evaluation CER, however, the best results were achieved when combining synthetic dysarthric data sets in addition to real dysarthric speech. We focus on CER rather than WER, as it better reflects the manual effort required to correct ASR output and provides a more stable statistical measure. The improvements from combining synthetic dysarthric data with real dysarthric fine tuning are evident for both the FastConformer_Hu and Whisper-turbo models. As Figure 1 shows, the best results are outside of the confidence intervals (calculated using the tool of Ferrer&Riera [28]) of the 'train-dys' only fine-tuning setup, confirming the effectivity of our approach.

Comparing the foundational ASR models FastConformer_Hu and Whisper-turbo, the conclusions are straightforward: while Whisper offers high accuracy and accessibility for English, our well-trained monolingual model could significantly outperform it in the given Hungarian language task. Finally, the control experiment results (the last two rows of Table IV) show that adding typical (premorbidity) data from the target speaker, instead of synthetic dysarthric data, did not generally improve the results. This confirms that the observed improvements were not due to general speaker-specific characteristics (such as vocal tract length or timbre) but rather to the effectiveness of few-shot learning in modeling dysarthric articulation. The high error rates from fine-tuning only on typical (and supervised) speech indicate that adapting to typical speech has little impact on recognizing dysarthric speech.

VII. CONCLUSION

This study demonstrates that a personalized ASR system designed for assistive technology can enable accurate speech-to-text conversion for individuals with severe dysarthria, even with minimal dysarthric speech data in a relatively low-resource language. By fine-tuning a monolingual FastConformer-based ASR model and augmenting it with synthetic dysarthric speech, we reduced the character error rate (CER) from 36–51% (zero-shot) to 7.3% and the word

TABLE V
FINE-TUNED **WHISPER-TURBO** ASR RESULTS [%] MEASURED ON DYSARTHRIC SPEECH

	Val-dys		Eval-dys	
Fine-tuning data sets	**WER**	**CER**	**WER**	**CER**
Train-dys	41.1	14.9	35.0	14.7
Train-dys + FT-Dys-Undertrained	36.6	13.3	35.2	15.2
Train-dys + FT-Dys-Best	35.8	12.1	32.2	13.1
Train-dys + FT-Dys-Overtrained	35.6	11.8	31.8	13.1
Train-dys + FT-Dys-Co-trained-A	35.5	11.2	33.4	14.1
Train-dys + FT-Dys-Co-trained-B	**33.9**	11.4	34.7	14.2
Train-dys + FT-Dys-Co-trained-C	36.9	11.8	34.0	14.0
Train-dys + FT-Dys-Co-trained-D	35.5	**10.9**	35.5	14.6
Train-dys + FT-Dys-Best + Over	**33.9**	11.4	32.0	12.3
Train-dys + All (7) FT-Dys	34.7	**10.9**	**30.4**	**12.2**

[4]https://huggingface.co/blog/fine-tune-whisper

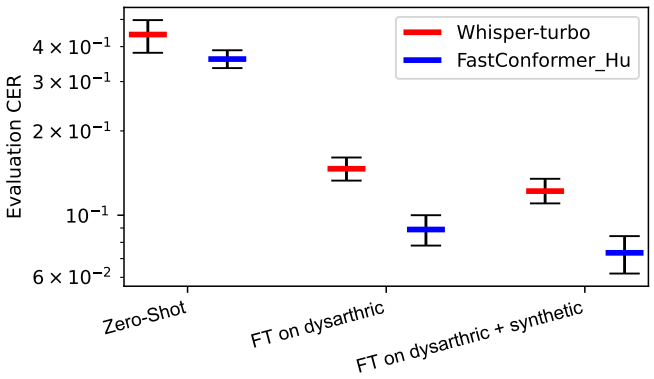

Fig. 1. Best CER results on dysarthric speech (Eval-dys) for different ASR approaches: zero-shot, fine-tuning (FT) only on real dysarthric speech, and FT on real + synthesized dysarthric data (with confidence intervals).

error rate (WER) from 76–65% to 18.3%. The monolingual model outperformed Whisper-turbo, highlighting the importance of language-specific adaptation. We used a speaker-adapted TTS system trained on premorbidity recordings to generate dysarthric speech with controlled severity. Synthetic data improved ASR performance, contributing to an 18% relative CER reduction. Control experiments confirmed that adaptation to fluent speech alone did not enhance dysarthric ASR, reinforcing the need for targeted modeling of impaired articulation. Given the promising results of zero-shot TTS which can utilize as little as a few seconds of unaffected recordings from a speaker [20], our results are likely reproducible with much less typical speech data than the what was used in this study, which will be subject of future work.

References

[1] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier, 2013.

[2] C. Bhat and H. Strik, "Speech technology for automatic recognition and assessment of dysarthric speech: An overview," *Journal of Speech, Language, and Hearing Research*, vol. 68, no. 2, pp. 547–577, 2025.

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

[4] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner *et al.*, "Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases." in *Proc. Interspeech*, 2021, pp. 4778–4782.

[5] A. B. Kain, J.-P. Hosom, X. Niu, J. P. Van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech communication*, vol. 49, no. 9, pp. 743–759, 2007.

[6] D. Liu, Y. Lin, H. Bu, and M. Li, "Two-stage and self-supervised voice conversion for zero-shot dysarthric speech reconstruction," in *Proc. IALP*, 2024, pp. 423–427.

[7] M. Fried-Oken, "Voice recognition device as a computer interface for motor and speech impaired people." *Archives of physical medicine and rehabilitation*, vol. 66, no. 10, pp. 678–681, 1985.

[8] P. D. Green, J. Carmichael, A. Hatzis, P. Enderby, M. S. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers." in *Proc. Interspeech*, 2003, pp. 1189–1192.

[9] A. Xiao, W. Zheng, G. Keren, D. Le, F. Zhang, C. Fuegen, O. Kalinli, Y. Saraf, and A. Mohamed, "Scaling ASR improves zero and few shot learning," in *Proc. Interspeech*, 2022, pp. 5135–5139.

[10] G. Yang, F. Yu, Z. Ma, Z. Du, Z. Gao, S. Zhang, and X. Chen, "Enhancing low-resource ASR through versatile TTS: Bridging the data gap," *arXiv preprint arXiv:2410.16726*, 2024.

[11] W.-C. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion," in *Proc. ICASSP*, 2022, pp. 6672–6676.

[12] E. Hermann and M. M. Doss, "Few-shot dysarthric speech recognition with text-to-speech data augmentation," in *Proc. Interspeech*, 2023, pp. 156–160.

[13] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc ICLR*, 2021.

[14] 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.

[15] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, "Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis," in *Proc. Interspeech*, 2024, pp. 2494–2498.

[16] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proc. ICML*, 2021, pp. 8599–8608.

[17] 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.

[18] H. Wang, V. Ravichandran, M. Rao, B. Lammers, M. Sydnor, N. Maragakis *et al.*, "Improving fairness for spoken language understanding in atypical speech with text-to-speech," in *Proc. NeurIPS Workshop on Synthetic Data Generation with Generative AI*, 2023.

[19] J. Tobin and K. Tomanek, "Personalized automatic speech recognition trained on small disordered speech datasets," in *Proc. ICASSP*, 2022, pp. 6637–6641.

[20] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart *et al.*, "XTTS: a massively multilingual zero-shot text-to-speech model," in *Proc. Interspeech*, 2024, pp. 4978–4982.

[21] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector," *Proc. ICASSP*, pp. 6925–6929, 2019.

[22] M. S. Kádár, G. Dobsinszki, K. Mády, and P. Mihajlik, "Feeding the BEAST – the enhancement of the BEA speech transcriber and its integration with neural language model /in Hungarian/," in *XIX.Conference on Hungarian Computational Linguistics*, 2023, pp. 135–143.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[24] C. Oravecz, T. Váradi, and B. Sass, "The Hungarian Gigaword corpus," in *Proceedings of LREC'14*. Reykjavik, Iceland: ELRA, May 2014, pp. 1719–1723. [Online]. Available: https://aclanthology.org/L14-1536/

[25] G. Dobsinszki, M. S. Kádár, T. Fegyó, K. Mády, and P. Mihajlik, "Training a Hungarian speech recognizer with tens of thousands of hours of speech /in Hungarian/," in *XXI. Conference on Hungarian Computational Linguistics*, 2025, pp. 87–96.

[26] E. Harper, S. Majumdar, O. Kuchaiev, J. Li, Y. Zhang, E. Bakhturina *et al.*, "NeMo: a toolkit for conversational ai and large language models." [Online]. Available: https://nvidia.github.io/NeMo/

[27] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *In arXiv: 2305.05084, eess.AS*, 2023.

[28] L. Ferrer and P. Riera, "Confidence intervals for evaluation in machine learning." [Online]. Available: https://github.com/luferrer/ConfidenceIntervals