forum acusticum 2026
8-12 September · Graz, Austria

TEXT-TO-SPEECH DATA AUGMENTATION FOR DYSARTHRIC CHILD SPEECH RECONSTRUCTION

Moritz Pfeiler[1], Benedikt Mayrhofer[1], Barbara Schuppler[1], Martin Hagmüller[1]
[1]*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

The goal of Dysarthric Speech Reconstruction is to improve intelligibility by converting dysarthric into healthy speech. In the future, this technology could become a valuable communication aid for people with this speech impairment. The development of systems performing complex speech tasks typically requires large amounts of data, which is especially sparse for pathological child speech. This paper explores the potential of Text-to-Speech data augmentation to improve performance in resource constraint scenarios. In this augmentation technique, synthetic dysarthric child speech is generated from text. To evaluate the effectiveness, a Voice Conversion model is trained with just 2.37 hours of real speech in two configurations: (1) only real speech (2) pre-trained with synthetic speech and fine-tuned on real speech. While the overall intelligibility remains low, data augmentation leads to noticeable improvements, which encourages further research. Future work should focus on optimizing both Text-to-Speech and Voice Conversion models for this task.

Keywords: *dysarthric speech reconstruction, voice conversion, text-to-speech, data augmentation*

1. Introduction

Dysarthria is a speech disorder in which the control of muscles for speech production is impaired. This condition can be caused by genetic birth defects, neurological diseases (e.g. Huntington's disease, Parkinson's disease) or brain tumors and injuries. Typical consequences are a lower speaking rate and changes in phonation, articulation and prosody, which leads to reduced intelligibility[1]. The rapidly evolving technologies Automatic Speech Recognition (ASR), Voice Conversion (VC) and Text-to-Speech (TTS) have the potential to enable people with dysarthria to communicate more effectively[1].

VC systems transform speech to sound as if spoken by another speaker while preserving the linguistic content. They typically consist of three stages: analysis, which extracts an intermediate speech representation; mapping, which modifies speaker characteristics to match a target speaker; and reconstruction, which converts the modified representation back into a speech waveform [2]. The rise of deep learning brought significant improvements to VC. Neural networks, such as Variational Auto Encoders (VAEs) or encoder blocks from ASR systems, allow for a more sophisticated separation of speaker identity/characteristics and linguistic content. These embeddings can then be manipulated or replaced more effectively allowing also zero-shot VC. To synthesize audio from intermediate representations, Generative Adversarial Networks (GANs) are commonly used [3].

Deep learning VC models have also been adapted and evaluated for dysarthric speech intelligibility enhancement. Treating the problem as a style transfer task for 2D images, GAN-based models, trained to transform spectrograms, have been a major focus [4]. More recently, diffusion-based approaches were proposed, promising superior VC performance [5].

In low-resource scenarios, transfer learning from other speech related tasks such as ASR or TTS has proven effective [6]. Also pre-training and fine-tuning techniques are an option. Expressive VC models can be pre-trained on large amounts of speech data and then fine-tuned to work well on a specific speaker where little data is available. While there are large speech datasets and pre-trained models for healthy speech, the substantial differences between healthy adult speech and dysarthric child speech lowers the expected effectiveness of these approaches [7].

More promising appears TTS-based data augmentation, where a large parallel pre-training dataset is generated synthetically. This approach was studied in

Corresponding author: *moritz.pfeiler@student.tugraz.at*

[1]https://www.asha.org/public/speech/disorders/dysarthria/

a VC context both for healthy speech [8] and electrolaryngeal speech [7], and has also proven effective in dysarthric ASR. The works presented in [9] and [10] all report an improvement of word and phoneme error rates by incorporating synthetic dysarthric speech in the training routine.
This paper explores the potential of sequence-to-sequence VC trained with parallel data as a means for Dysarthric Speech Reconstruction. Working with a limited amount of training data, we investigate whether TTS data augmentation can increase reconstruction quality and robustness in a low-resource setting. We identify suitable TTS models for this task and discuss our efforts towards creating an identity-preserving target voice.

2. Methods

2.1. Data

2.1.1. Source Samples: Dysarthric Speech

The dataset (2.37h of netto speech in 878 utterances) used in this project consists of annotated dysarthric speech recordings of a child diagnosed with GLUT1 deficiency syndrome. The majority of the data is read speech, but there is also spontaneous speech and read digits available. For a more detailed analysis see [11]. The distribution of sample lengths is shown in Figure 1. The spike at low lengths is caused by single word utterances collected during speech therapy sessions.

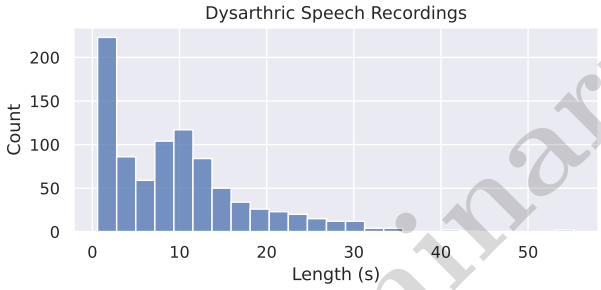

Figure 1: Histogram of the lengths of samples in the dysarthric speech dataset (N=878).

2.1.2. Target Samples: Healthy Speech

Choosing an appropriate voice for a child is a complex task of its own, even if technical challenges are not considered. Using TTS to generate target samples strikes a sensible balance between speech quality and the effort required to gather large amounts of data. Parallel data, which is necessary for many VC models, can be easily generated this way.
While creating a voice that perfectly fits the child is beyond the scope of this research, we still defined some general directions for the synthetic voice.
(1) The speaking rate should be rather slow. (2) The voice should have an Austrian accent. (3) The pitch and tone should match the original dysarthric speech.
The GRASS corpus [12] includes voices satisfying (1) and (2). This corpus includes both read and spontaneous speech recordings in studio quality with text annotations. Once a suitable baseline voice is created with appropriate speaking rate and accent, (3) can be easily achieved with acceptable quality using FreeVC [13] and the original dysarthric speech as the target.
Read speech is most suitable for TTS training, as the speaking style is typically consistent. Unfortunately, there is not enough data per speaker to train a robust TTS model. To overcome this issue, other speech data must be included in the training. Fusing the data of all suitable GRASS speakers with a larger German dataset and training a multi-speaker TTS model, yields better results.
The model of choice for this task is YourTTS [14]. Starting from a publicly available checkpoint trained on the CML dataset [15], the model is first trained on the HUI-Audio-Corpus-German CLEAN [16] and thorsten-de [17] datasets for 240k steps. In a second stage, the recordings from the selected GRASS speakers are added and the model is fine-tuned on the combined dataset for another 240k steps. Further training results in overfitting. While the synthesized speech gets livelier and closer to the targets in terms of expressiveness, the rhythm/timing becomes unnatural and vowel combinations (diphthongs) are often mispronounced.
In the end, the YourTTS replication of GRASS speaker F021 was chosen as the base for the target voice. The pitch and tone are changed to better match the child with FreeVC.

2.1.3. Text-to-Speech Data Augmentation

Training a robust VC model with only 142 minutes of data is difficult.
For this work we use VITS [18] to augment the dysarthric speech dataset. The model is first pre-trained with the thorsten-de dataset for 100k steps. The model is then fine-tuned for 20k steps with the available dysarthric speech samples and picks up the speaker characteristics of the child. Beyond 20k steps, the model overfits to the noisy data.
With TTS for both source and target set up, the 20k shortest sentences from the HUI corpus are used to generate the pre-training dataset. This amounts to 19 hours of healthy speech and 73 hours of dysarthric speech.

2.2. Voice Conversion Model Selection

The model choice was based on the results of preliminary experiments with various VC methods. We tested the open source models FreeVC, OpenVoice [19] and kNN-VC [20], which are trained on healthy speech and don't alter temporal features. They are not able to address patterns in dysarthric speech that limit intelligibility, such as insertions or slow articulation [11]. In an attempt to introduce knowledge about dysarthric speech, we paired a whisper ASR model [21], that was fine-tuned in [11] on the target dysarthric dataset, with XVC [22], which achieved good results in a recent study on pathological VC [23]. Despite using fine-tuned content features, the conversion results did not improve upon previous exper-

iments. Based on these results, our assumption is that we need a model that is capable of modifying both the temporal and spectral structure. The sequence-to-sequence model **AAS-VC** meets these requirements and is said to perform well even with small datasets [24]. The source code is available on GitHub as part of the seq2seq toolbox[2]. AAS-VC is set up with Conformers as encoder and decoder, which are responsible for spectral changes, and the flow-based probabilistic duration predictor, which controls time stretching during inference to match the speaking rate and rhythm of the target. While the paper recommends phonetic posteriorgrams (PPGs) as the input feature, we use mel spectrograms, because finding a performant PPG encoder for german dysarthric child speech, was beyond the scope. During training with parallel data, an alignment between source and target features is calculated using a monotonic alignment search algorithm. From this alignment, durations for each segment of the input sequence are derived, which are used to train the duration predictor. Because only integer durations are possible, the source sequence must be shorter than the target sequence [24].

2.3. Text-to-Speech Models

TTS is used to generate synthetic source and target speech for a parallel dataset. We use models from the Coqui TTS toolbox[3], where pre-trained checkpoints and training scripts are available.

VITS is an end-to-end single-speaker TTS model that is trained with parallel data i.e. audio-text pairs. It uses VAE to connect the vocoder with a Prior Encoder conditioned on the text input. This allows the model to learn a suitable latent representation. Similar to AAS-VC, the model features a stochastic duration predictor that expands the text encoder outputs to audio length. Adversarial training and variational inference are used for better synthesis quality and more natural and diverse rhythm respectively [18]. **YourTTS** is an extension of VITS, adding speaker and language embeddings, which allow for multi-speaker training and zero-shot inference [14].

2.4. Experiments

To evaluate the effects of pre-training on synthetic data, the AAS-VC model is trained with 2 different setups: (1) Real data only and (2) pre-training on synthetic data + fine-tuning on real data.

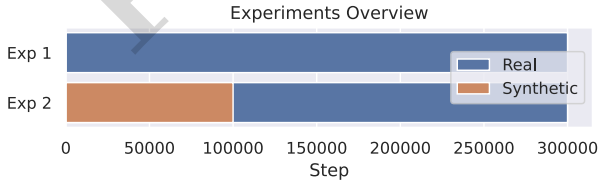

Figure 2: Experimental setup

Training Data and Preprocessing: The alignment module of AAS-VC imposes the constraint that the source input must be shorter than the target input for all training samples. Because the dysarthric speech samples are significantly longer than the healthy targets, the ratio must be corrected. This can be done inside the model with the *post encoder reduction factor*. That way the encoder still works on the original length dysarthric input, which is very resource intensive. For increased computational efficiency, we speed up the dysarthric recordings beforehand by a factor of 2 and reduce the *post encoder reduction factor* accordingly. The findings in [25] suggest that this substantial temporal compression won't harm the reconstruction performance. The authors show that speeding up dysarthric speech to healthy speaking rates improves results in a phoneme recognition task. The train/validation/test split is 95%/1%/4% and the effective batch size is 16.

Hyperparameters: The hyperparameters for all training runs are taken from the *arctic* example (mel-mel-mel) provided in the AAS-VC repository. Only the post encoder reduction factor is adapted to fit the duration ratio of source and target samples. All models are trained for a total of 300k steps.

Training Dynamics: Figure 3 shows the evolutions of train losses for the different experiments. The blue curve shows the experiment 1. Training AAS-VC for 300k steps using only real data. The orange curve shows the pre-training with the synthetic dataset. Convergence is significantly slower compared to the much smaller real dataset. Green represents the loss of the fine-tuning phase of experiment 2. Starting from the pre-trained checkpoint, increases the initial convergence speed, compared to experiment 1. After about 50k steps the differences in convergence speed vanish.

While the convergence process is not fully completed at the ends of the experiments, we decided against further training, because of diminishing returns and high computational effort. 100k training steps take about 44 hours on an NVIDIA Tesla V100 16GB GPU.

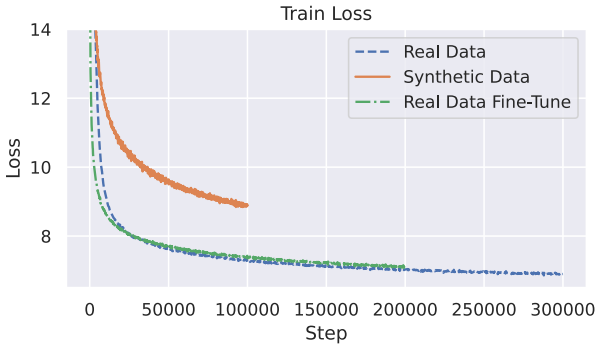

Figure 3: Losses for all training phases. Fine-tuning starts at 100k steps but the loss curve was shifted to the start for comparability.

[2]https://github.com/unilight/seq2seq-vc
[3]https://github.com/idiap/coqui-ai-TTS

3. Results and Discussion

At this stage, a systematic intelligibility evaluation is not yet meaningful. All models struggle to produce intelligible outputs.
Instead, we compare the experiments through qualitative evaluations at different training stages as well as ASR word and character error rates (WER, CER) across 16 unseen utterances.

3.1. Experiment 1 (real data only)

Training only on real data results in choppy speech which is mostly unintelligible. Even with knowledge of the linguistic content only few speech sections can be identified as words. The prosody, specifically the rhythm, is unnatural. More training iterations reduce the noisiness of the generated speech. Other aspects stay the same and intelligibility does not improve. This is a clear sign of overfitting. Due to the small training dataset, the model fails to generalize to unseen data.

3.2. Experiment 2 (synthetic + real data)

Immediately after pre-training, when the model has only seen synthetic data, the results are breathy and smeared. The quality is low but comparable to results from experiment 1. This is an indication that the characteristics of synthetic speech are similar to real dysarthric speech.
Fine-tuning on the real dataset brings audible improvements to both the naturalness of the prosody and the intelligibility. Converted speech samples are less choppy than in experiment 1. However, the rhythm is still rather synthetic. The robustness of the duration predictor definitely improved with more training data, but there is still room for improvement. While overall intelligibility is still poor, individual words are frequently pronounced in an intelligible way.
Beyond 100k fine-tuning steps, signs of overfitting appear. The 300k step model yields more choppy results and the speaking rate is faster, which lowers intelligibility slightly.

3.3. ASR Performance

We evaluate the performance across four german ASR models available on Hugging Face[4], which were also used in [11]. WERs and CERs are computed for the unprocessed dysarthric speech samples as well as the outputs of experiment 1 (300k steps) and experiment 2 (200k steps).
Comparing WERs does not yield much insight in regards to intelligibility, as they are all close to or above 100%. The CERs in Table 1 support the subjective evaluation. Compared to experiment 1, experiment 2 achieves 23% lower CERs averaged across ASR models. For the wav2vec2 models [26], [27] the experiment 1 performs worse than the original, which can be attributed to overfitting. The high CERs of the whisper models on the original dysarthric speech are due to hallucinations.

[4]https://huggingface.co/

Table 1: Average CER in % (N=16)

Model	Original	Exp 1	Exp 2
bofenghuang/whisper-small-cv11-german	219	103	76
bofenghuang/whisper-medium-cv11-german	382	102	76
sharrnah/wav2vec2-bert-CV16-de	62	83	60
aware-ai/wav2vec2-xls-r-300m-german	66	81	65

3.4. Duration Predictor

To visualize the duration predictors performance, we show spectrograms of source, target and model outputs of a test sample in Figure 4. The annotation for this recording is "*Heute hatten wir Turnunterricht in der Schule*". The colored sections correspond to the sentence, while gray-scale indicates background noise or the models interpretation of this section. White vertical lines mark word boundaries.
Both models reduce the length of the input by about a factor of 4. Comparing the durations of the linguistic content (colored sections), we see that the model from experiment 1 is able to predict the duration of the target best. Also the word boundaries match better. Experiment 2 's output is slightly longer as the first four words are spoken slower. The lower speaking rate also contributes to the improved intelligibility.
Model 1 handles appended noise in the input better. While both models transform the background noise into speech-like sounds, model 1 leaves a short pause after the last word *Schule*. Model 2 completely removes this pause and *Schule* merges with the noise, making it unintelligible.

4. Conclusion

The objective of this work has been the evaluation of TTS data augmentation as way to improve robustness of dysarthric child speech reconstruction in a single-speaker low-resource scenario. First, a TTS model was trained on a small set of dysarthric child speech recordings. With this model, a large synthetic pre-training dataset was generated. To analyze the effect of this data augmentation technique, two experiments were conducted. In the experiment 1, a VC model was trained on just the small real dysarthric dataset. For experiment 2, the VC model was first pre-trained using the synthetic dataset and then fine-tuned with the real dataset.
While the achieved intelligibility was unsatisfactory across all experiments, the pre-training and fine-tuning approach brought a noticeable improvement in both intelligibility and naturalness for unseen data. This also reflects in a 23% reduction in CER compared to the experiment with no augmentation. The results show that TTS data augmentation can

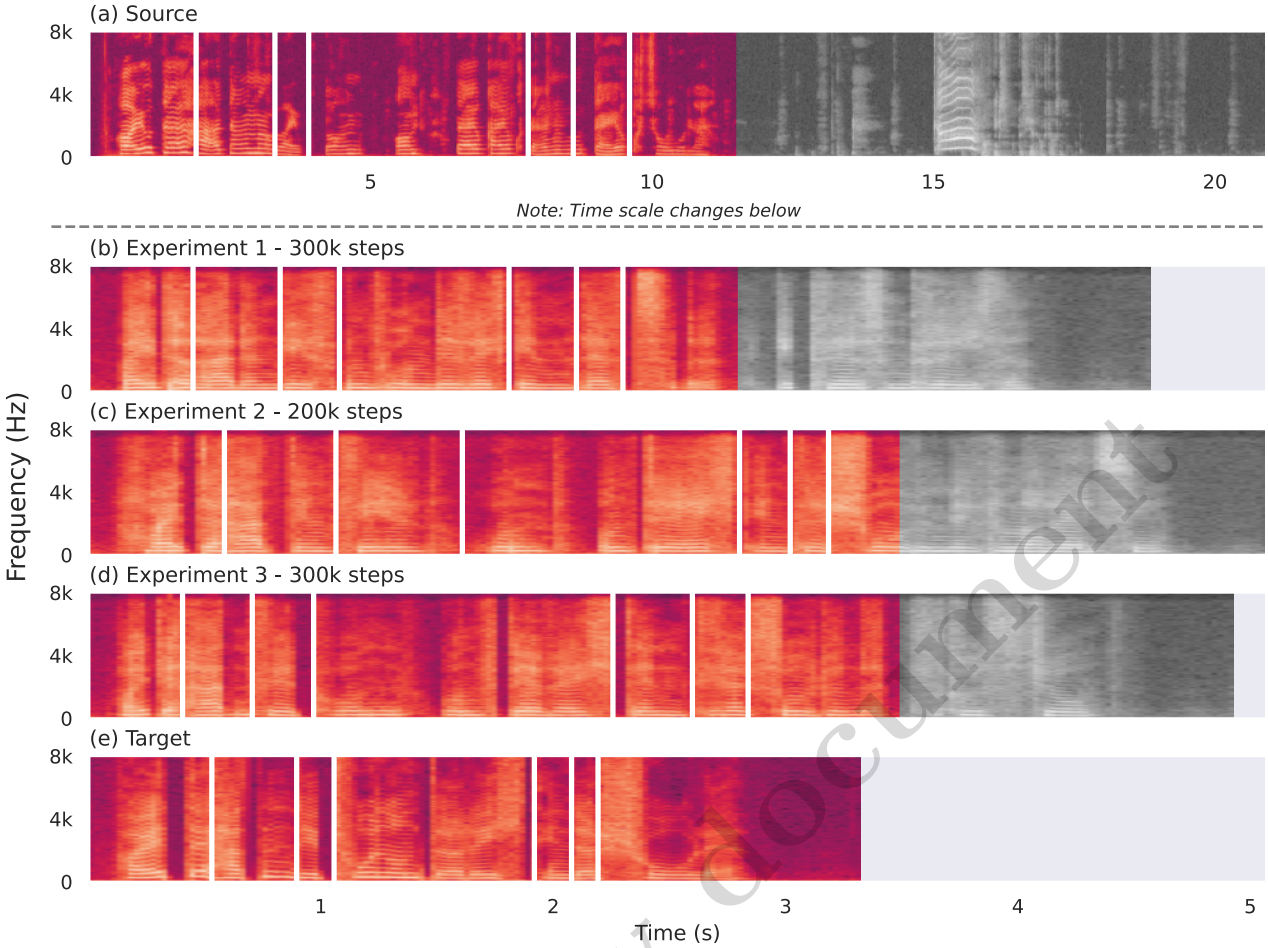

Figure 4: Spectrograms of Source (a), Target (d) and best Models (b), (c). Gray-scale indicates noise. White vertical lines mark word boundaries.

be an effective tool for dysarthric child speech and encourages further research in this direction.

From this research arise ideas for improvements. Especially with small datasets the quality of recordings and annotations are important. As part of a recently concluded project, parts of the dataset were phonetically segmented and annotated by a linguist [28]. This phonetic information can be used to train a TTS model that more closely matches the child's voice. Training with cleaner data, is also likely to improve VC performance.

Further research should also focus on improving the model and explore different architectures. Using more specialized speech representations as the model's input, such as features from whisper or other speech foundation models, could be the next step. For parallel training, VAE-based end-to-end approaches have the potential of improving synthesis quality. Diffusion-based models could pose an alternative to parallel training.

5. Acknowledgments

This research was funded in part by the Austrian Science Fund (FWF) [10.55776/PAT5948223]. This research was approved by the TU Graz ethics board.

References

[1] A. Hamza, D. Addou, and I. Hadjadji, "Enhancing Dysarthric Speech Intelligibility: A Review of Techniques," in *Proc. 3rd Int. Conf. on Electronics, Energy and Measurement (IC2EM)*, 2025. doi: 10.1109/IC2EM63689.2025.11101176.

[2] B. Sisman, J. Yamagishi, S. King, and H. Li, "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, Aug. 2020, doi: 10.1109/TASLP.2020.3038524.

[3] T. Walczyna and Z. Piotrowski, "Overview of Voice Conversion Methods Based on Deep Learning," *Applied Sciences*, vol. 13, no. 5, 2023, doi: 10.3390/app13053100.

[4] H. Mehrez, M. Chaiani, and S. A. Selouani, "Using StarGANv2 Voice Conversion to Enhance the Quality of Dysarthric Speech," in *Proc. Int. Conf. on Artificial Intelligence in Information and Communication (ICAIIC)*, 2024. doi: 10.1109/ICAIIC60209.2024.10463241.

[5] X. Chen *et al.*, "DiffDSR: Dysarthric Speech Reconstruction Using Latent Diffusion Model,"

in *Proc. Interspeech*, 2025. doi: 10.21437/Interspeech.2025-1770.

[6] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Pretraining Techniques for Sequence-to-Sequence Voice Conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021, doi: 10.1109/TASLP.2021.3049336.

[7] D. Ma, L. P. Violeta, K. Kobayashi, and T. Toda, "Pretraining and Fine-Tuning Techniques for Electrolaryngeal Speech Enhancement Based on Sequence-to-Sequence Voice Conversion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, 2025, doi: 10.1109/TASLPRO.2025.3577374.

[8] D. Ma, W.-C. Huang, and T. Toda, "Investigation of Text-to-Speech-based Synthetic Parallel Data for Sequence-to-Sequence Non-Parallel Voice Conversion," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, 2021, pp. 870–877.

[9] 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. doi: 10.21437/Interspeech.2024-1645.

[10] E. Hermann and M. Magimai-Doss, "Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation," in *Proc. Interspeech*, 2023. doi: 10.21437/Interspeech.2023-2481.

[11] L. Eckert and B. Schuppler, "Automatic Speech Recognition for a Dysarthric Child Speaking Austrian German," in *Proc. Forum Acusticum Euronoise*, 2025. doi: 10.61782/fa.2025.0168.

[12] B. Schuppler, M. Hagmüller, J. A. Morales Cordovilla, and H. Pessentheiner, "GRASS: The Graz Corpus of Read and Spontaneous Speech," in *Proc. 9th Conf. on Language Resources and Evaluation (LREC)*, 2014, pp. 1465–1470.

[13] J. Li, W. Tu, and L. Xiao, "Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion," in *Proc. IEEE ICASSP*, 2023. doi: 10.1109/ICASSP49357.2023.10095191.

[14] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *Proc. 39th Int. Conf. on Machine Learning*, 17 2022, pp. 2709–2720.

[15] F. S. Oliveira, E. Casanova, A. C. Junior, A. S. Soares, and A. R. Galvão Filho, "CML-TTS: A Multilingual Dataset for Speech Synthesis in Low-Resource Languages," in *Text, Speech, and Dialogue*, 2023, pp. 188–199.

[16] P. Puchtler, J. Wirth, and R. Peinl, "HUI-Audio-Corpus-German: A High Quality TTS Dataset," in *KI 2021: Advances in Artificial Intelligence: Proc. 44th German Conference on AI*, 2021. doi: 10.1007/978-3-030-87626-5_15.

[17] T. Müller and D. Kreutz, *Thorsten - Open German Voice (Neutral) Dataset.* (Feb. 2021). Zenodo. doi: 10.5281/zenodo.5525342.

[18] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proc. of the 38th Int. Conf. on Machine Learning*, 18 2021, pp. 5530–5540.

[19] Z. Qin, W. Zhao, X. Yu, and X. Sun, "OpenVoice: Versatile Instant Voice Cloning." [Online]. Available: https://arxiv.org/abs/2312.01479

[20] M. Baas, B. van Niekerk, and H. Kamper, "Voice Conversion With Just Nearest Neighbors," in *Proc. Interspeech*, 2023. doi: 10.21437/Interspeech.2023-419.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." 2022. doi: 10.48550/arXiv.2309.07598.

[22] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, "Using Joint Training Speaker Encoder With Consistency Loss to Achieve Cross-Lingual Voice Conversion and Expressive Voice Conversion," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8. doi: 10.1109/ASRU57964.2023.10389651.

[23] B. Mayrhofer, M. Hagmüller, F. Pernkopf, and P. Aichinger, "Evaluating State of the Art Voice Conversion Models for Dysphonic and Electro-Larynx Speech," in *Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze University Press, 2025, pp. 33–36.

[24] W.-C. Huang, K. Kobayashi, and T. Toda, "AAS-VC: On the Generalization Ability of Automatic Alignment Search based Non-autoregressive Sequence-to-sequence Voice Conversion." 2023. doi: 10.48550/arXiv.2309.07598.

[25] L. Prananta, B. Halpern, S. Feng, and O. Scharenborg, "The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition," in *Proc. Interspeech*, 2022, pp. 36–40. doi: 10.21437/Interspeech.2022-190.

[26] Y.-A. Chung *et al.*, "w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 244–250. doi: 10.1109/ASRU51503.2021.9688253.

[27] A. Babu *et al.*, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech*, 2022. doi: 10.21437/Interspeech.2022-143.

[28] M. Galović, "Phonetic Analysis of Dysarthric Child Speech," in *Book of Abstracts: 3rd Graz-Vienna Speech Workshop*, Medical University of Vienna, 2025, pp. 15–16.