University of Zurich and ETH Zurich
Institute of Neuroinformatics
Department of Computational Linguistics

Master's Thesis/Semester Project Opportunity
Personalized Voice Cloning and Synthetic Data Generation for Impaired Speech

Supervisors

Prof. Dr. Benjamin Grewe	Primary PI	
Roman Boehringer	Supervisor	roman@ini.ethz.ch
Niclas Pokel	Supervisor	npokel@ethz.ch
Pehuen Moure	Supervisor	pehuen@ini.uzh.ch
Dr. Yingqiang Gao	Supervisor	yingqiang.gao@uzh.ch

Abstract

Automatic Speech Recognition (ASR) for individuals with impaired speech is severely hampered by data scarcity. This project addresses this problem by developing a **personalized Text-to-Dysarthric-Speech (TTDS) model** to serve as an advanced data augmentation method. Unlike assistive technologies that aim to correct speech impairments, the primary goal here is to faithfully clone a speaker's unique impaired speech patterns. Using state-of-the-art generative audio models (e.g., VITS), the system will learn to generate synthetic yet realistically impaired speech data from very few recordings. A key innovation will be to leverage phoneme uncertainty analyses from prior work [5] to guide the synthesis process , enabling the targeted generation of more realistic phonetic deviations. This project is designed for a highly motivated, independent individual ready to take ownership of a challenging research topic.

Research Goals and Problem Statement

Standard Text-to-Speech (TTS) models are trained on vast amounts of normative speech and fail to reproduce the complex, inconsistent, and highly variable acoustic characteristics of impaired speech. Simply training these models on impaired data often leads to poor-quality results. The central research question is therefore: **How can we develop a high-fidelity generative model that learns to synthesize speech with a speaker's specific impaired acoustic characteristics, based on a very limited amount of data?**

The primary goals of this thesis are:

1. To develop a generative model that learns both the vocal identity (timbre, pitch) and the specific articulatory patterns of a speaker's impairment.

2. To create a system that generates large quantities of synthetic yet realistic impaired speech from text. This data will serve as training material to improve the robustness of separate ASR systems.

3. To investigate how Phoneme Difficulty Scores (PhDScore) from our prior research [5] can be used as a conditioning signal to control the type and severity of synthesized impairments, enabling more targeted data augmentation.

Proposed Work, Expected Outcomes, and Results

- Conduct a comprehensive literature review of state-of-the-art TTS, voice conversion, and voice cloning models, with a focus on few-shot learning and applications to atypical speech.
- Select and implement a suitable baseline architecture (e.g., VITS, YourTTS) in PyTorch.
- Adapt the model and training procedure to effectively learn from the impaired speech data available in the UA-Speech and BF-Sprache datasets. This may involve novel conditioning mechanisms or loss functions.
- Integration: Incorporate the Phoneme Difficulty Score (PhDScore) from our previous work [5] as an additional conditioning signal. The hypothesis is that this will allow the model to generate more realistic errors on phonemes that are particularly difficult for the speaker.
- Design and conduct a rigorous evaluation protocol: The primary success metric will be the improvement of a downstream ASR model.
    1. Train a baseline ASR model (e.g., VI-LoRA) on the original, small datasets.
    2. Augment the training data with the newly synthesized impaired speech.
    3. Retrain the ASR model on the augmented dataset and measure the reduction in Word/Character Error Rate (WER/CER).
- **Expected Outcome:** A functional TTDS system and empirical proof of its effectiveness in improving ASR performance. The results will be summarized in a high-quality manuscript suitable for submission to a top-tier conference (e.g., Interspeech, ICASSP).

Discussion of Prior Work

This project is motivated by the need to overcome data scarcity in the field of impaired speech recognition. Our lab's previous work [4, 5, 6] focused on *recognizing* impaired speech, providing a robust ASR framework (VI-LoRA), valuable datasets (BF-Sprache), and analysis methods (PhDScore) that will be instrumental for the evaluation and methodological foundation of this new generative project.

The direct methodological basis for this project is recent work in TTS-based data augmentation for dysarthric ASR [3, 7]. These studies have demonstrated the feasibility of synthesizing impaired speech. Our project will extend this approach by integrating uncertainty-based conditioning signals to enable more targeted and potentially more effective data augmentation. The underlying

technology will be inspired by advances in zero-shot TTS and voice cloning, as seen in models like VITS [2] and YourTTS [1].

Recommended Background Reading:

- [3, 7] - For context on synthesizing dysarthric speech.
- [2] - For understanding the VITS architecture.
- [1] - For zero-shot multi-speaker TTS.
- [5] - To understand the Phoneme Difficulty Score (PhDScore) to be used as a conditioning signal.

Qualifications

This project requires a high degree of independence and a passion for tackling challenging, real-world problems. The ideal candidate will be driven by the opportunity to create a novel assistive technology and publish their work.

- Excellent programming skills in Python and deep familiarity with PyTorch.
- Strong background in deep learning, particularly generative models (VAEs, GANs, Diffusion Models).
- Experience with or strong interest in audio signal processing and speech technology.
- Highly self-motivated, proactive, and capable of taking ownership of a research project.
- A clear ambition to produce a scientific publication for a top-tier conference.

References

[1] Edresson Casanova, Christopher Shulby, Çağlar Gülçehre, Erick M. de T. e Souza, Julio C. M. de Oliveira, and Anderson de L. H. Santos. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, 2022.

[2] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[3] 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 *Proc. Interspeech 2024*, pages 2494–2498, 2024.

[4] Niclas Pokel, Pehuén Moure, Roman Boehringer, and Yingqiang Gao. Adapting foundation speech recognition models to impaired speech: A semantic re-chaining approach for personalization of german speech. In *12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)*, pages 82–86, 2025. doi: 10.21437/DiSS.2025-17.

[5] Niclas Pokel, Pehuén Moure, Roman Boehringer, and Yingqiang Gao. Data-efficient asr per-

sonalization for non-normative speech using an uncertainty-based phoneme difficulty score for guided sampling, 2025. URL https://arxiv.org/abs/2509.20396.

[6] Niclas Pokel, Pehuén Moure, Roman Boehringer, Shih-Chii Liu, and Yingqiang Gao. Variational low-rank adaptation for personalized impaired speech recognition, 2025. URL https://arxiv.org/abs/2509.20397.

[7] Mohammad Soleymanpour, Michael T Johnson, Rahim Soleymanpour, and Jeffrey Berry. Accurate Synthesis of Dysarthric Speech for ASR Data Augmentation. *Speech Communication*, 164:103112, 2024.