University of Stuttgart
Germany

Speech Enhancement in the Context of Neural Vocoders

Thesis	FA/MA
Supervisor	Tobias Raichle
Examiner	Prof. Dr.-Ing. Bin Yang

Motivation

Speech enhancement studies improving the quality of spoken language and finds applications as a front-end in automatic speech recognition, telecommunication or hearing aids. Generally, speech enhancement covers multiple types of corruptions (noise, reverberations, echoes, compression artifacts etc.) but most previous works focus on denoising.

A common bottleneck of time-frequency domain speech enhancement systems is the spectrogram's phase component. Recent publications found that the phase information is vital for natural sounding speech enhancement [1]. However, it is difficult to accurately estimate, due to its lack of structure. On the other hand, text-to-speech systems typically rely on a neural vocoder [2, 3] to convert the estimated spectrogram's magnitude into a waveform signal. While there is a publication using vocoders for speech enhancement [4], their approach did not include end-to-end fine-tuning of the complete speech enhancement system. In this thesis, we want to further explore the use of vocoders as an output stage of speech enhancement systems.

Objectives

- Implement a framework for using vocoders as the output stage for speech enhancement systems
- Fine-tune the complete system
- Evaluate different existing architectures
- Evaluate the model on a range of benchmarks and compare with existing models
- Identify problems in the approach and find solutions

Prerequisites

- Took the Deep Learning exam with good results
- Took the Detection and Pattern Recognition exam with good results
- Good programming skills in Python
- Experience in ML-frameworks (Preferably PyTorch)
- *Optional*: Experience in sequence modelling
- *Optional*: Participated in the ISS Deep Learning Lab

If this topic has sparked your interest, write me an email and we can discuss the proposal in more detail. Please include your current transcript and CV.

References

[1] Ruizhe Cao, Sherif Abdulatif, and Bin Yang. "CMGAN: Conformer-based metric GAN for speech enhancement". In: *arXiv preprint arXiv:2203.15149* (2022).

[2] Zhifeng Kong et al. "Diffwave: A versatile diffusion model for audio synthesis". In: *arXiv preprint arXiv:2009.09761* (2020).

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis". In: *Advances in neural information processing systems* 33 (2020), pp. 17022–17033.

[4] Haohe Liu et al. "VoiceFixer: Toward general speech restoration with neural vocoder". In: *arXiv preprint arXiv:2109.13731* (2021).

Tobias Raichle
tobias.raichle@iss.uni-stuttgart.de

10. September 2024