University of Stuttgart
Germany

Speech Enhancement in the Context of Foundation Models

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.

Foundation models (e.g. [1]) are very large models that are trained on a vast amount of data, spanning a wide range of domains. Generally, they are trained without a direct task in mind.
Whereas foundation models have deeply permeated the computer vision domain, they have not had the same impact in speech processing beyond speech recognition. While [2] explored using speech foundation models [3, 4] for speech enhancement, they found that the models did not perform well with noisy inputs and did not explore how to use foundation models for speech enhancement in depth. In this thesis, we want to explore how speech foundation models can be leveraged for speech enhancement and how fine-tuning can improve results.

Objectives

- Implement a framework for using foundation models for speech enhancement
- Fine-tune the foundation models using multiple datasets
- Evaluate the model on a range of benchmarks and compare with existing models
- Identify problems in the approach and find solutions
- Explore new approaches for using foundation models in speech enhancement

Prerequisites

- Took the Deep Learning 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] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[2] Shu-wen Yang et al. "A Large-Scale Evaluation of Speech Foundation Models". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[3] Alexei Baevski et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460.

[4] Wei-Ning Hsu et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units". In: *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021), pp. 3451–3460.

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

10. September 2024