Master Thesis

Speech Enhancement with Deep Learning Techniques

Speech enhancement (SE) is the task of maximizing the perceptual quality of speech signals, in particular by removing background noise and data compression artefacts, restoring frequency content or de-reverberation. Most recorded speech signals contain some form of the above mentioned artefacts that reduce intelligibility and listening preferences. The task of speech enhancement is particularly interesting for applications like audio/video calls, voice notes, hearing aids, automatic speech recognition systems or automatic processing of audio/video content, for example podcasts, tutorials or lectures. SE is also featured in the annually reoccurring DNS-Challenge at ICASSP, where leading research teams are competing and presenting their results [1]. While this field of research was previously dominated by digital signal processing methods, more recent and state-of-the-art approaches include deep learning techniques and architectures like Autoencoder UNets [2], GANs [3, 4] or Diffusion Probabilistic Models [5] both in the frequency- and time-domain. Major drawbacks of SE-methods are the introduction of so-called musical artifacts and discrepancies in phase-alignment of the estimated speech signal. To prevent those drawbacks, a variety of post-processing steps like filtering, thresholding or networks have been proposed [6, 7]. SE-techniques are evaluated on a broad set of quantitative metrics like SDR, SAR, SIR, LSD, PESQ, STOI, DNSMOS etc. as well as qualitative listening tests. For real-time usage on mobile or desktop devices, SE-models further need to work with low-latency constraints (20-30ms) and a small processing and memory footprint [8].

This thesis aims to reproduce and expand the above mentioned deep learning techniques to improve state-of-the-art SE-methods. A dataset based on the VCTK [9] and AudioSet [10] corpus with additional audio artifacts like background noise, codec compression, reverberation and down-sampling is already prepared and available. Proposed SE-methods will be benchmarked and evaluated both on qualitative and quantitative metrics (see above) under real-time processing constraints. The thesis offers the possibility to be published as a scientific publication and can be written in English.

Literature

[1] Reddy, C.K., Dubey, H., Gopal, V., Cutler, R., Braun, S., Gamper, H., Aichner, R. and Srinivasan, S., 2021, June. Icassp 2021 deep noise suppression challenge. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6623-6627). IEEE.
[2] Defossez, A., Synnaeve, G. and Adi, Y., 2020. Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*.
[3] Liu, H., Kong, Q., Tian, Q., Zhao, Y., Wang, D., Huang, C. and Wang, Y., 2021. VoiceFixer: Toward General Speech Restoration With Neural Vocoder. *arXiv preprint arXiv:2109.13731*.
[4] Greshler, G., Shaham, T.R. and Michaeli, T., 2021. Catch-A-Waveform: Learning to Generate Audio from a Single Short Example. *arXiv preprint arXiv:2106.06426*.
[5] Kong, Z., Ping, W., Huang, J., Zhao, K. and Catanzaro, B., 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.
[6] Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N. and Mitsufuji, Y., 2017, March. Improving music source separation based on deep neural networks through data augmentation and network blending. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 261-265). IEEE.
[7] Wiegratz, R., 2021. Adding Context Information to Deep Neural Network based Audio Source Separation.
[8] Braun, S., Gamper, H., Reddy, C.K. and Tashev, I., 2021, June. Towards efficient models for real-time deep noise suppression. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 656-660). IEEE.
[9] Yamagishi, J., Veaux, C. and MacDonald, K., 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).
[10] Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M. and Ritter, M., 2017, March. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 776-780). IEEE.

Requirements

Knowledge of:

- Machine Learning and Deep Learning theory and practice
- Audio DSP theory and practice
- Python programming
- Machine Learning frameworks like Tensorflow and/or Pytorch

Supervision

Fabian Seipel, f.seipel@campus.tu-berlin.de
Corvin Jaedicke, jaedicke@campus.tu-berlin.de
Prof. Dr. Stefan Weinzierl, stefan.weinzierl@tu-berlin.de