DL Project Proposal: Speech Denoising

Hector Martel, Samuel Pegg, Toghrul Abbasli

April 2021

Topic: An Application of Deep Learning to a Real World Problem in audio.

1 Problem Statement and Motivation

Audio denoising, and Speech denoising in particular, have been classical signal processing problems. There are many applications in which it is beneficial to remove unwanted noises from audio or recover a clean signal from a corrupted one. For example, it has applications in voice calls, video conferencing or audio postproduction. With the advances of Deep Learning the performance has drastically improved, allowing for more intelligible results and real time processing.

2 Literature Review

To get an overview of the topic, the existing solutions and the current challenges, we will review some research literature on speech enhancement [1, 2, 3], audio restoration [4, 5] and source separation [6, 7, 8]. Recently, a method from image denoising has been successfully applied to the audio domain [9]. There is a significant overlap between these 3 tasks, thus, models proposed to solve one can benefit the others as well. Detailed references are included at the end of this document.

3 Datasets

There are some widely-used datasets of speech recordings such as LibriSpeech[10] and WSJ[11] can be used in conjunction with datasets of ambient noises/sounds like UrbanSound8K[12] to create the training and evaluation data. Alternatively, we can use existing dedicated noise suppression datasets such as DNS (Deep Noise Suppression) that have been recently proposed [13]. We are also planning to apply audio data augmentation methods to our datasets.

4 Methods and Existing Solutions

As for the algorithms/methods we are interested in reproducing and comparing time domain and frequency domain-based approaches and studying their advantages and disadvantages. The focus will be on end-to-end speaker-independent models. Some methods include UNet[14], Wave-UNet[6] or ConvTasNet[15] just to name a few. Another direction to consider is to incorporate other architectures like the transformer, which have proven to be successful in other tasks.

5 Evaluation

The evaluation of the task can be objective and subjective. For the purpose of this project we are going to focus on objective evaluation metrics to assess the audio signal quality. This makes it easier to compare our results with other methods. We are interested in observing the final scores provided

by the different metrics (tables) together with the convergence of each method (plots of loss w.r.t. epochs) in the evaluation section.

Any Audio	Speech Audio Only
SNR (Signal to Noise Ratio)	SSNR
SDR (Signal to Distortion Ratio)	PESQ (WB, NB)
SI-SDR (Scale-Invariant SDR)	STOI

Table 1: Objective metrics commonly used in speech enhancement and source separation tasks.

To inspect the results and present them in the report, we can plot the spectrograms of the noisy mixture and the estimations of clean and noise audios. The results can also be inspected directly by listening. If we have time, we plan to include a website to listen to the results as supplementary material.

References

[1] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," 2021.

[2] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," 2021.

[3] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," 2020.

[4] V. S. Narayanaswamy, J. J. Thiagarajan, and A. Spanias, "On the design of deep priors for unsupervised audio restoration," 2021.

[5] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, p. 2362–2372, Dec 2019.

[6] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," 2018.

[7] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," 2020.

[8] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," 2020.

[9] M. M. Kashyap, A. Tambwekar, K. Manohara, and S. Natarajan, "Speech denoising without clean training data: a noise2noise approach," 2021.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[11] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, (USA), p. 357–362, Association for Computational Linguistics, 1992.

[12] A. Diment, A. Mesaros, T. Heittola, and T. Virtanen, "Tut rare sound events, development dataset urban sounds," Mar. 2017. The license terms are specified in the LICENSE.txt file.

[13] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," 2021.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[15] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, p. 1256–1266, Aug 2019.