Project Proposal: Evaluating Classical Image Denoisers for Time-Frequency Domain Speech Enhancement

Jack Irish
EE367
Winter 2025

1. Background/Related Work

Speech Enhancement (SE) is a class of problems that relate to improving the quality of audio signals containing human speech. Denoising is one of the most studied speech enhancement tasks, with effective algorithms required for the success of technologies like VOIP and speech recognition. Interestingly, most modern SE algorithms do not directly operate on the time domain waveform, but instead use the Short-Time Fourier Transform (STFT) to transform the signal into a time-frequency spectrogram. The magnitude of the STFT is treated as a 2D image, where methods like a neural network trained on clean-noisy image pairs are used to produce a denoised spectrogram. At this point, the clean audio signal may be recovered by adding back the noisy signal's phase and performing the inverse STFT. As with other image processing tasks, CNNs are a popular architecture because of their smaller size and ability to take advantage of spatial (spectro-temporal in the case of STFT magnitudes) structure.

The Redundant Convolutional Encoder-Decoder (R-CED) architecture (Park and Lee, 2016) is a novel design where the encoder stage maps the input image to a higher-dimensional space and the decoder compresses the higher dimensional features back down into lower dimensions, opposite to how traditional encoder-decoder networks are designed. The authors found that this architecture performed well for speech enhancement, outperforming other leading neural network designs.

A different perspective on time-frequency domain denoising involves learning a mask to be applied to the noisy STFT instead of learning the mapping directly from noisy STFT to clean STFT. The CNN-GRU model (Hasannezhad et al. 2020) employs a CNN for feature extraction cascaded with an LSTM stage to predict the real and imaginary parts of a mask to be applied to the noisy STFT.

In terms of general image denoising, classical filtering methods saw great popularity before the advent of deep neural networks and still see some use today due to their reliability and lower computational complexity. Wavelet thresholding is a well-studied denoising technique where coefficients of the wavelet transform of the noisy image are modified to produce a cleaner image after inverting the transform. One of the more popular variants of this method involves adaptively changing the thresholding parameters for each wavelet subband based on estimated statistics. (Chang et al., 2000).

Non-local denoising methods like BM3D (Dabov, Kostadin et al. 2007) rose to the forefront of traditional denoising methods due to their superior performance. BM3D collects similar-looking image patches into 3D groups that are filtered then separated back into their original positions.

2. Project Description

My goal for this project is to determine if classical denoising algorithms for natural images (wavelet thresholding, BM3D, etc) are still effective when applied to time-frequency domain images produced from noisy speech signals. I would like to test these traditional methods both as standalone denoisers and as a preprocessing stage to potentially improve the effectiveness of modern networks like Park and Lee's R-CED. I plan on evaluating the classical denoisers as well as the hybrid architectures (R-CED trained on spectrograms pre-processed with classical denoisers) against the standalone R-CED architecture as outlined by Park and Lee.

3. Project Timeline

My first task will be to find a suitable dataset of clean and noisy speech (MS-SNSD looks promising) as well as choose appropriate metrics to compare the various methods' denoising performance. Most SE literature uses both objective (SNR, etc.) and subjective (i.e. large-scale surveys) metrics to evaluate audio quality, but I will likely only use objective statistics due to time constraints.

The majority of my time will be spent implementing and training the R-CED model from Park and Lee's 2016 paper. I intentionally chose this particular slightly outdated model because it is not too large, so it should be reasonable to train on a short time scale. Some time will be spent implementing Chang's wavelet thresholding method as it requires a slight modification of commonly available wavelet thresholding tools. BM3D is already implemented in both MATLAB and Python, so will be relatively easy to apply to the dataset.

Ideally, I should be left with about a week to analyze the performance of the denoising methods and prepare my report.

References

Chang, S. Grace, Bin Yu, and Martin Vetterli. "Adaptive wavelet thresholding for image denoising and compression." *IEEE transactions on image processing* 9.9 (2000): 1532-1546

Dabov, Kostadin, et al. "Image denoising by sparse 3-D transform-domain collaborative filtering." *IEEE Transactions on image processing* 16.8 (2007): 2080-2095.

Hasannezhad, Mojtaba, et al. "An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement." *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020.

Park, Se Rim, and Jinwon Lee. "A fully convolutional neural network for speech enhancement." *arXiv preprint arXiv:1609.07132* (2016).

Yuliani, Asri Rizki, et al. "Speech enhancement using deep learning methods: A review." *Jurnal Elektronika dan Telekomunikasi* 21.1 (2021): 19-26.