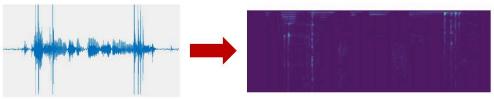
Evaluating Classical Image Denoising for Speech Enhancement

Jack Irish
Stanford University

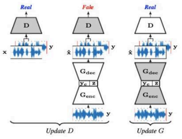
Motivation

- Speech enhancement (SE) is an important task required for the success of technologies like VOIP and speech recognition
- Many SE algorithms act on signals in the time-frequency domain (spectrogram) to take advantage of underlying structure
- With the problem transformed into the denoising of a 2D image, techniques like CNN denoisers can be used
- If classical, non-data-driven denoising methods can be applied to this domain of images, they could offer time and memory savings over deep neural networks



Related Work

- R-CED architecture is an early example of a CNN used for spectral SE [1]
- Alternate approach predicts "noise mask" from spectrum instead of mapping directly from noisy to clean spectrum [2]
- Time-domain methods like Sergan [3] address lack of phase information in magnitude spectrum



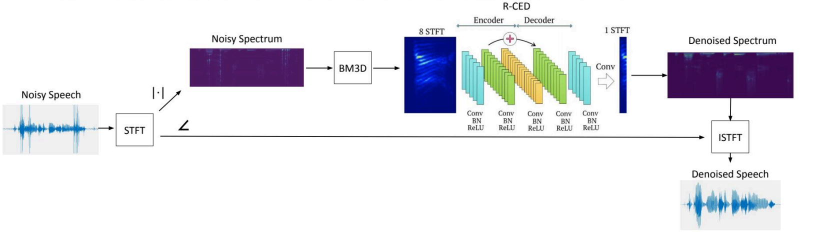
Sergan, Baby and Verhuist [3]

References

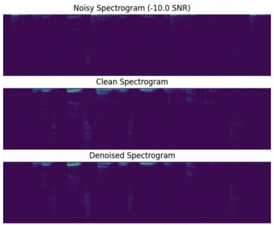
[1] Park, Se Rim, and Jinwon Lee. "A fully convolutional neural network for speech enhancement." arXiv preprint arXiv:1609.07132 (2016).
[2] Z. Xu, S. Elshamy, and T. Fingscheidt, "Using separate losses for speech and noise in mask-based speech enhancement," in 2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2020
[3] D. Baby and S. Verhulst, "Sergan: Speech Enhancement Using Relativistic Generative Adversarial Networks with Gradient Penalty," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK (2019)
[4] Dabov, Kostadin, et al. "Image denoising by sparse 3-D transform-domain collaborative filtering." IEEE Transactions on image processing 16.8 (2007)
[5] Chandan K. A. Reddy et al. "A scalable noisy speech dataset and online subjective test framework," Proc. Interspeech (2019)

Method

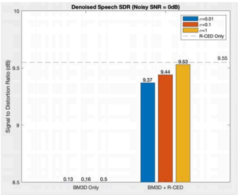
- Train R-CED [1] on speech spectrum preprocessed with classical denoising algorithm (BM3D [4]) to produce denoised spectrum
- MS-SNSD dataset [5] used for training on various noise types at SNR of 0db and -10dB
- Compare performance against applying BM3D only or R-CED only to noisy spectra
- Experimental std estimates for BM3D chosen as ~(0.1x, 1x, 10x) average spectrum pixel value

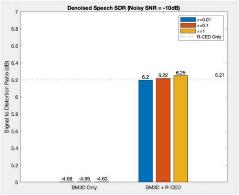


Experimental Results



-10.0dB SNR noise level, BM3D std=1.0





- BM3D preprocessing enhances CNN performance at higher noise levels, potentially enabling the use of smaller networks
- However, BM3D incurs additional computational cost for minimal performance gains
- As an idea for future work, perhaps a simpler preprocessing method like local filtering could have similar benefits without the complexity of BM3D