Evaluating Classical Image Denoising for Speech Enhancement

Jack Irish

Abstract—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. Many 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, which is typically denoised via supervised deep learning, then used to reconstruct the denoised time domain signal. This project aims to determine if lighter-weight classical denoising techniques designed for natural images can be extended to this class of spectrum magnitude images. A proposed method using the BM3D algorithm as a preprocessing stage for a CNN denoiser shows performance improvements in high-noise environments, potentially allowing for the use of smaller networks without sacrificing denoising effectiveness.

1 Introduction

The recovery of speech signals from noisy environments is an important problem with applications in technologies such as mobile phones, speech recognition and hearing aids. This is a difficult problem due to the rapid, time-varying nature of speech signals and the variety of noise sources in real environments.

Speech enhancement algorithms range from common filtering techniques to deep neural networks. A popular class of methods first use the Short Time Fourier Transform (STFT) to produce a two dimensional time-frequency magnitude spectrum to denoise instead of the raw speech signal (Figure 1). This approach does not make full use of the noisy signal's phase information, but under the assumption that the human ear is not sensitive to minor phase variations, the denoising task is greatly simplified. With the problem transformed into the denoising of a single-channel 2D image, image processing techniques like CNN denoisers can be used to either map directly to a clean spectrum or learn a mask to apply to the noisy image. Even with such an unusual class of images, convolutional neural networks are effective denoisers because they take advantage of an image's spatial structure (local time and frequency structure in the case of STFT magnitudes) to learn sophisticated mappings with a small number of parameters. The lesser memory and computational cost of a CNN is desirable for embedded applications like mobile devices that are often used in crowded, noisy environments.

Considering the value of a computational and memory efficient algorithm, applying a simpler image denoising method to speech spectra could enable efficient, real-time processing on even more resource constrained devices. The primary difficulty in this approach is that many of these lightweight image denoising algorithms were designed for use on natural images and are only optimal under well modeled additive Gaussian noise. The proposed method addresses this by employing a classical denoising method as a preprocessing stage for a CNN denoiser. Experiments demonstrate that this hybrid technique outperforms the non-neural network method and the CNN when used on their own.

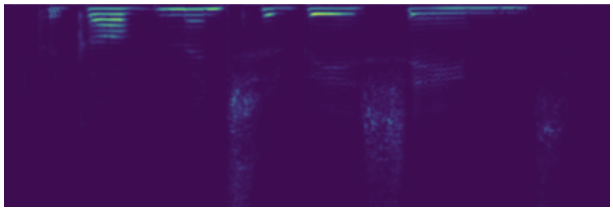

Fig. 1. STFT Magnitude Spectrum of Speech Signal

2 Related Work

Early speech enhancement methods like Wiener filtering or signal subspace SE [1] applied traditional signal processing methods to the time domain noisy audio signal. These methods often relied on a simple additive noise model and were less robust to real noise sources.

Deep neural networks have become very popular for speech enhancement due to their ability to generalize to more complicated noise sources. It is also easy to generate large amounts of synthetic training data by combining clean speech samples with a variety of recorded noise samples, both of which are readily available due to the growth of speech recognition and speech enhancement research. The majority of speech enhancement networks either operate on the raw noisy audio signal or on the aforementioned STFT magnitude spectrum, which sacrifices phase information for simpler data that leads to more stable training.

SERGAN [2] is an example of a time-domain speech enhancement method that uses a generative adversarial network to learn a mapping directly between noisy speech directly and a clean signal (Figure 2). The authors introduce a relativistic generative adversarial model that addresses the instability issues that come with time domain speech enhancement networks and generative adversarial networks in general. SERGAN was shown to outperform other time

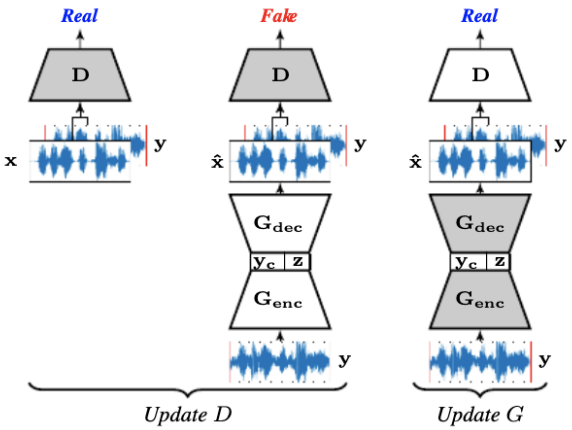

Fig. 2. SERGAN Training Methodology

domain and spectral speech enhancement networks, at the cost of a large and complex model.

Among time-frequency domain methods, models typically either learn a noise mask or a direct mapping to a denoised spectrum. In the first case, algorithms are inspired by classical filtering methods that estimate a frequency domain weighting rule that can be applied to the noisy spectrum to suppress noisy components. Using a deep network like a CNN to learn these masks overcomes the limitations that classical filtering algorithms encounter when the noise distribution is not particularly well behaved. For example, Xu, Elshamy and Fingscheidt [3] propose a novel loss function that balances noise reduction with signal gain and use it to train a CNN to predict masks from noisy STFT magnitude spectra.

The latter class of time-frequency domain methods involves training a model to directly map from noisy magnitude spectra to clean magnitude spectra. Unsurprisingly, CNN architectures developed for image denoising are also able to effectively denoise these spectrograms. The Redundant Convolutional Encoder-Decoder (R-CED) network, proposed by Park and Lee [4], is an encoder-decoder network where the encoder maps the input image into a higher dimensional feature space instead of a lower dimensional one. The decoder then projects back down into the input space. The authors note that the use of a higher dimensional intermediate feature space allows for more effective inference, working like the kernel trick. Taking full advantage of the memory efficiency of convolutional networks, the R-CED model required far fewer parameters to achieve the same performance as competing architectures.

In general, time-frequency speech enhancement models are memory and compute efficient, as well as more stable during training, but fall slightly behind methods that make better use of the noisy signal's phase information in terms of raw performance.

3 Method

The proposed method aims to improve a time-frequency domain speech denoising model by preprocessing noisy spectra with a classical image denoising algorithm. The hybrid method was evaluated against the denoising network on its own at different noise levels to determine if and when such a hybrid approach offers performance improvements. For each noise level and preprocessing method, an identical denoising model was retrained from scratch to provide a fair comparison. Additionally, the classical algorithm by itself will be applied to the same test data to see if it is feasible to forego the data-driven approach entirely.

3.1 Denoising Network

Park and Lee's fully convolutional R-CED network [4] will be used as the time-frequency denoising model in the following experiments. Specifically, the 33K parameter R-CED variant with 16 convolutional layers and skip connections every other layer was chosen because it outperformed all other versions of the R-CED model except for the 100K parameter, 20 layer design. The largest R-CED architecture was not used due to constraints on training time. As shown in Figure 3, the R-CED network takes a noisy STFT magnitude frame along with the seven preceding frames as input and produces a single clean STFT magnitude frame as output. The entire noisy spectrum is passed through the R-CED model, segment by segment, to produce the denoised spectrum. Input and output spectrum segments are normalized to have zero mean and unit variance. During inference, training set statistics are used to normalize input spectra and denormalize output spectra.

The input spectra are produced from the noisy speech signals by applying a 256-point Short Time Fourier Transform with a 64-point window shift. Following Park and Lee's method, the Hamming window function will be used. Only the magnitude of the complex STFT spectrum is denoised.

The denoised speech signal is reconstructed by combining the denoised magnitude with the corresponding noisy STFT phase components, applying the inverse transform. Because the human ear is largely insensitive to phase, the use of noisy phase does not have a major impact on model performance. However, very large phase error ($> 45°$) can lead to noticeable degradation in quality, so Park and Lee's 'phase aware scaling' (Equation 1) is applied to target clean magnitudes during training to ensure that the model learns to attenuate STFT components with extreme differences between clean and noisy phase.

$$\mathbf{s}_{phase\ aware} = \mathbf{s}_{clean} |\cos(\theta_{clean} - \theta_{noisy})| \quad (1)$$

Once again following the authors' outlined method, the R-CED model was trained using Adam with a batch size of 64. The learning rate was started at $lr = 0.0015$ and decreased to $\frac{lr}{2}$, $\frac{lr}{3}$ and $\frac{lr}{4}$ after each epoch where the validation loss fails to improve. After the next epoch with no validation improvement, training is stopped.

3.2 Preprocessing Stage

BM3D [5] was chosen as the classical preprocessing stage. BM3D is a non-local image denoising algorithm that is among the state of the art of non-data driven methods in terms of noise attenuation and visual quality. This particular choice was made because BM3D should be able to exploit

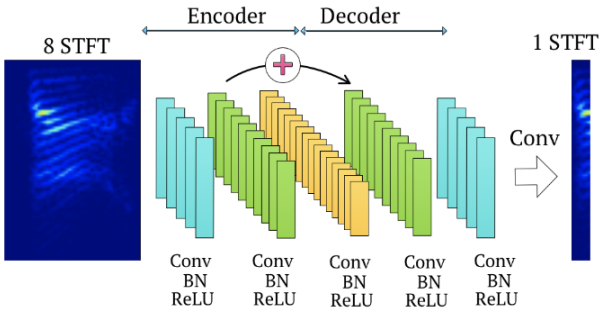

Fig. 3. Redundant Convolutional Encoder-Decoder Network

the patterns and structure of magnitude spectra better than simpler local denoising methods.

The BM3D algorithm requires an estimate of the noise standard deviation. Since the time-frequency domain audio noise cannot be approximated well by a Gaussian, three estimates were tested to see which would perform best: $\sigma = 0.01$, $\sigma = 0.1$ and $\sigma = 1.0$, roughly equal to $\frac{\mu}{10}$, μ and 10μ where μ is the average pixel value of a noisy magnitude spectrum.

BM3D was used both as the sole spectrum denoising method and as a preprocessing stage to apply to samples before training the R-CED model (Figure 4).

3.3 Training Data

Ideally these experiments would have been conducted using the same dataset as Park and Lee's original R-CED work. However, the TIMIT speech corpus that was used for clean speech samples is not freely available to the public, so an alternative source of data had to be found.

Instead, the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [6] was used to generate pairs of noisy and clean speech signals. The dataset consists of a collection of clean speech of about a sentence in length and a variety of noise samples commonly found in crowded environments like air conditioner hum or the babble of background conversation. Samples are automatically combined to form sets of clean and noisy speech at any specified SNR. The following experiments were conducted using approximately 6 hours of speech at two signal to noise ratios of 0 and -10 dB. 20% of the generated data was left aside as a test set.

3.4 Evaluation Metrics

Three objective metrics were used for comparison: Signal to Distortion Ratio (SDR) (Equation 2), Short Time Objective Intelligibility (STOI) [7], and Perceptual Evaluation of Speech Quality (PESQ) [8]. All three of these metrics were used to compare the reconstructed time domain denoised audio signal with its target clean counterpart. SDR is measured in decibels and larger values indicate better denoising performance. STOI is measured from -1 to 1 with higher values indicating more intelligible speech. PESQ is measured from 1 to 5 with higher values indicating higher quality speech.

$$\text{SDR} = 10\log_{10}\left(\frac{||\mathbf{y}||^2}{||f(\mathbf{x}) - \mathbf{y}||^2}\right) \quad (2)$$

4 Experimental Results

The BM3D denoiser, the R-CED denoiser and the hybrid denoiser with BM3D preprocessing were evaluated on the MS-SNSD test set for noisy speech with an SNR of 0 dB and -10 dB. The BM3D and hybrid denoisers were tested across a range of three standard deviation estimates. The two methods involving the R-CED model were trained from scratch in each case to produce a separate model for each noise level and, in the case of the hybrid architecture, a separate model for each of the three BM3D parameter estimates.

TABLE 1
BM3D Test Set Performance

	0dB SNR Speech			-10dB SNR Speech		
BM3D σ	SDR (dB)	STOI	PESQ	SDR (dB)	STOI	PESQ
0.01	0.13	0.73	1.20	-4.69	0.66	1.14
0.1	0.16	0.73	1.21	-4.69	0.66	1.14
1.0	0.5	0.73	1.22	-4.63	0.66	1.14

TABLE 2
R-CED Test Set Performance

0dB SNR Speech			-10dB SNR Speech		
SDR (dB)	STOI	PESQ	SDR (dB)	STOI	PESQ
9.55	0.76	1.39	6.21	0.69	1.24

TABLE 3
Hybrid Method Test Set Performance

	0dB SNR Speech			-10dB SNR Speech		
BM3D σ	SDR (dB)	STOI	PESQ	SDR (dB)	STOI	PESQ
0.01	9.37	0.75	1.37	6.20	0.69	1.24
0.1	9.44	0.75	1.37	6.22	0.68	1.24
1.0	9.53	0.75	1.39	6.25	0.68	1.24

4.1 BM3D Denoiser

Table 1 shows the performance of BM3D as the sole spectrum denoising method. The performance is uniformly poor across both noise levels. The example in Figure 5 demonstrates how BM3D performs some minor smoothing of portions of the spectrum most impacted by approximately white noise, but fails to attenuate the less random undesirable parts of the spectrum.

4.2 R-CED Denoiser

Table 2 shows the performance of the R-CED model trained at the two different noise levels. The CNN spectrum denoiser performs considerably better in all metrics when compared to BM3D. As seen in Figure 6, the network is able to identify and eliminate structured portions of the spectrum that are not associated with speech.

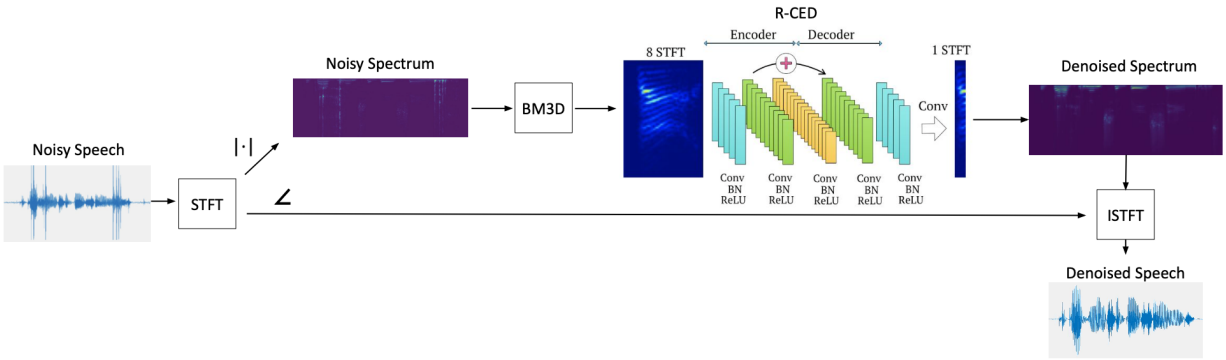

Fig. 4. Hybrid R-CED Speech Denoiser with BM3D Preprocessing

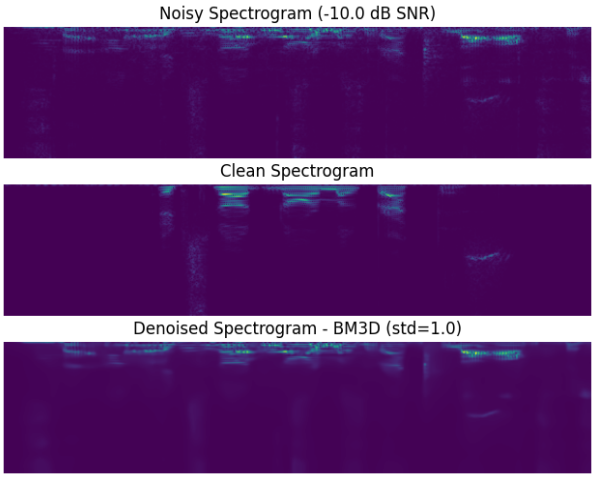

Fig. 5. BM3D Denoising Example

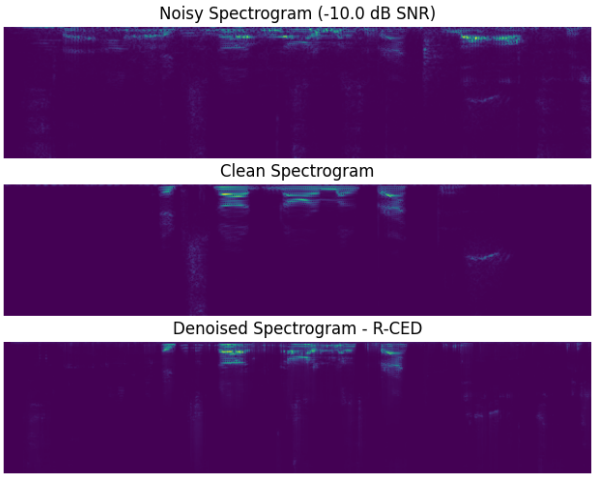

Fig. 6. R-CED Denoising Example

4.3 Hybrid Denoiser

The hybrid denoiser's performance is outlined in Table 3. This method performs very similarly to the lone R-CED network at a 0 dB signal to noise ratio and slightly outperforms the R-CED network at the higher -10 dB SNR noise level. The example in Figure 7 shows how the R-CED network with BM3D preprocessing produces a similar, but slightly 'smoother' looking denoised spectrum than the R-CED on its own. For both the hybrid method and the pure BM3D denoiser, the best performance was observed when the noise standard deviation estimate σ was chosen to be 1.

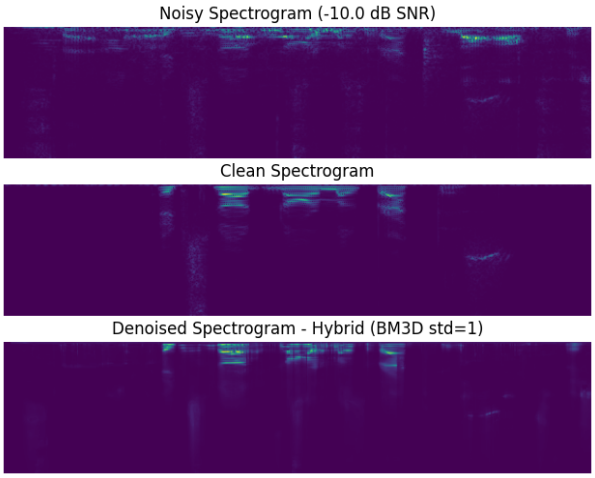

Fig. 7. Hybrid Denoising Example

4.4 Summary

The experimental results for all three methods have been compiled in Table 4. Only the results for the best BM3D standard deviation estimate ($\sigma = 1$) are included for the BM3D and hybrid denoisers. The best performance for each metric and noise level is highlighted in bold print.

Figures 8 and 9 compare the resulting signal to distortion ratios at an SNR of 0 and -10 dB respectively and illustrate how both R-CED based methods perform similarly and vastly outperform BM3D on its own.

5 Discussion

As one might expect given the non-natural domain of spectral magnitude images, the fully convolutional R-CED denoiser handily outperforms the classical BM3D denoising algorithm. BM3D is able to smooth out portions of noisy spectra that can be well approximated by additive Gaussian noise, but the algorithm is incapable of removing noise with

TABLE 4
Experimental Performance Overview

	0dB SNR Speech			-10dB SNR Speech		
Method	SDR (dB)	STOI	PESQ	SDR (dB)	STOI	PESQ
BM3D	0.5	0.73	1.22	-4.63	0.66	1.14
R-CED	**9.55**	**0.76**	**1.39**	6.21	**0.69**	**1.24**
Hybrid	9.53	0.75	**1.39**	**6.25**	0.68	**1.24**

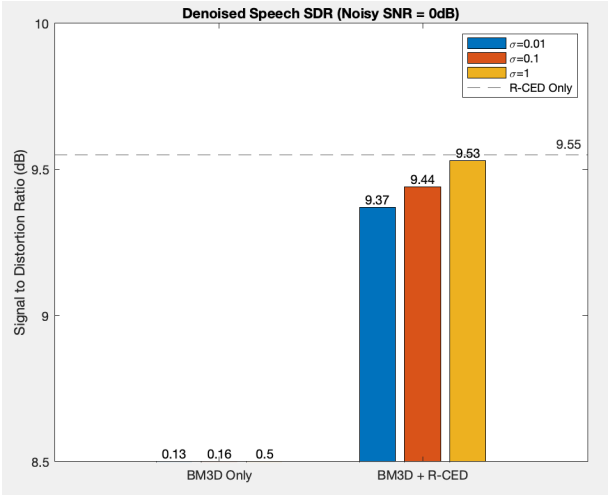

Fig. 8. Signal to Distortion Ratio at 0dB SNR

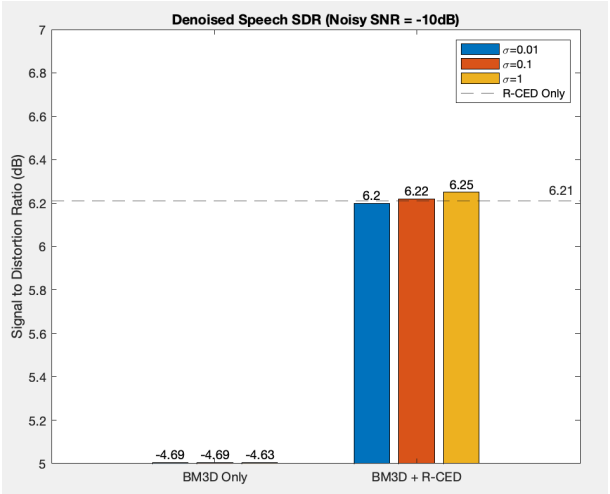

Fig. 9. Signal to Distortion Ratio at -10dB SNR

any sort of structure. However, BM3D's attenuation of white noise leads to a slight performance improvement when used as a preprocessing stage before the R-CED network. This improvement is only seen at the higher -10 dB SNR noise level, likely due to the more prominent presence of approximately white noise, especially in the sparse regions of the noisy spectrum.

It is important to note that objective speech quality metrics such as the three used in this experiment are not perfect indicators of denoising performance. For instance, the minor SDR improvement obtained from BM3D preprocessing is almost imperceptible to the ear. Some speech enhancement studies also employ subjective metrics to better quantify the quality of reconstructed audio. These metrics typically involve large scale anonymous surveys where listeners score speech samples on some numerical scale. Due to time constraints, the three experimental techniques were evaluated using only objective metrics, potentially preventing deeper insights.

Despite this, a small increase in an objective metric like signal to distortion ratio could lead to noticeable improvements in applications like speech processing even if the difference in the speech signals was not noticeable to the human ear. If the proposed hybrid speech enhancement method were employed as part of a speech recognition algorithm, the difference in objective signal quality could heighten a neural network's ability to classify the speech correctly.

6 Conclusion

In summary, classical image denoisers like BM3D show promise as an effective preprocessing stage for time-frequency domain speech enhancement networks. At high noise levels, the hybrid denoising method with the fully convolutional R-CED network trained on BM3D output outperformed an identical R-CED network trained on raw noisy magnitude spectra. In theory, this could allow a smaller model with fewer parameters to achieve the same level of performance a larger network without classical preprocessing.

However, these improvements are minor - just a 0.4 dB increase of signal to distortion ratio and nearly no noticeable improvement to perceptual speech metrics like STOI and PESQ. In a realistic embedded system, the additional computational cost of a sophisticated non-local algorithm like BM3D is likely too great to consider use of a hybrid method over a purely data driven technique.

This does not necessarily mean that other computationally cheaper classical denoising algorithms cannot provide the same performance benefits to a spectrum denoising network. In future work, alternate preprocessing methods like Wiener filtering could be evaluated, and may demonstrate the same ability to reduce any approximately white noise without the relatively extreme computational strain of BM3D. Additionally, further testing with a larger dataset and subjective speech quality metrics collected via survey might lead to more actionable results.

References

[1] W. P. Hermus, K. and H. Van hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Adv. Signal Process*, 2007.

[2] 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)*, pp. 106–110, 2019.

[3] Z. Xu, S. Elshamy, and T. Fingscheidt, "Using separate losses for speech and noise in mask-based speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7519–7523.

[4] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," 2016. [Online]. Available: https://arxiv.org/abs/1609.07132

[5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[6] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *CoRR*, vol. abs/1909.08050, 2019.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[8] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.