Speech Enhancement for Signals with Additive White Noise

Gregory R. Hessler
Electrical and Computer Engg.
Georgia Tech
Atlanta, GA, United States
gregory.hessler@gatech.edu

Kshama Kodthalu Shivashankara
Electrical and Computer Engg.
Georgia Tech
Atlanta, GA, United States
kshamaks@gatech.edu

Yue Teng
Electrical and Computer Engg.
Georgia Tech
Atlanta, GA, United States
yteng38@gatech.edu

***Abstract*—This project addresses the problem of suppressing additive white noise in speech signals. In this project, we implemented two single channel short-time Fourier transform (STFT) based speech enhancement methods, namely Wiener filtering and spectral subtraction. The experiment results are evaluated based on signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ). The Wiener filtering algorithm leads to PESQ improvments of up to 0.84, whereas The spectral subtraction algorithm leads to PESQ improvements of up to 0.35.**

***Index Terms*—speech enhancement, additive white noise, Wiener filter, spectral subtraction**

I. Introduction

With any real-world signal, noise is an inevitable component of the acquired signal; speech is no different. To minimize the noise is a goal in just about in signal processing problem. In order to do so, one must have an understanding of both the type of signal they are trying to measure and of the type of noise they are likely to encounter.

Common noise types tested with speech signals include:

1) White Noise
2) Babble Noise (i.e. background speech)
3) Factory floor noise
4) Cockpit noise

For this project, we will focus on the case of additive white noise. While this type of noise may not necessarily be the most widely applicable, the mathematical properties of additive white noise yield several algorithmic techniques for speech enhancement. Here we refer to speech enhancement as the altering of the signal to achieve a signal which has less noisy, higher clarity, and/or less fatigue while listening to.

One object metric to judge the quality of a speech signal is via the Perceptual Evaluation of Speech Quality (PESQ) method [1]. PESQ was developed to model the subjective test on voice quality perceived by humans. The metric can be analyzed via a full reference algorithm which compares the enhanced speech signal to the ground truth speech signal. The metric ranges from 1 (bad) to 5 (excellent).

In the report that follows, we consider two techniques: the Wiener Filter and the Spectral Subtraction methods. We compare their performance for varying levels of noise and draw conclusions on their utility.

II. Methodology

We consider the noisy speech signal $y[n]$ composed of the clean speech signal $s[n]$ and additive white noise $d[n]$ forming

$$y[n] = s[n] + d[n]. \tag{1}$$

Since speech signals are non-stationary, we process the signal on a frame-by-frame basis. Hence in the spectral domain we can write

$$Y(\omega, k) = S(\omega, k) + D(\omega, k) \tag{2}$$

where ω is the normalized radian frequency and k is an index for the frame number. With uncorrelated background noise, it follows that the power spectrum has no cross-terms meaning

$$|Y(\omega, k)|^2 = |S(\omega, k)|^2 + |D(\omega, k)|^2. \tag{3}$$

Our goal is to estimate the power spectrum of the speech signal $|\hat{S}(\omega, k)|^2$.

A. Wiener Filter

The Wiener filter has frequency response

$$H(\omega, k) = \frac{|S(\omega, k)|}{|Y(\omega, k)|} = \frac{|Y(\omega, k)| - |D(\omega, k)|}{|Y(\omega, k)|}.$$

While we do not know the true noise power $|D(\omega, k)|$ we assume the expectation $\mathbb{E}\left(|D(\omega, k)|\right)$ is equal to the power in the unvoiced regions. This yields

$$H(\omega, k) = \frac{|S(\omega, k)|}{|Y(\omega, k)|} = \frac{|Y(\omega, k)| - \mathbb{E}\left(|D(\omega, k)|\right)}{|Y(\omega, k)|}. \tag{4}$$

Thus the enhanced signal can be estimated by

$$\hat{S}(\omega, k) = H(\omega, k) Y(\omega, k)$$

Taking the inverse STFT and combining the segments results in the estimate $\hat{s}[n]$.

B. Spectral Subtraction

Our goal is once again to estimate the power spectrum of the speech signal $|\hat{S}(\omega,k)|^2$. However, while we have the power spectrum of the noisy speech signal $|Y(\omega,k)|^2$, we do not know the power spectrum of the noise and hence will have another estimate $|\hat{D}(\omega,k)|^2$. This spectral subtraction method thus can be expressed as

$$|\hat{S}(\omega,k)|^2 = |Y(\omega,k)|^2 - |\hat{D}(\omega,k)|^2. \tag{5}$$

where the core of the algorithm is to compute the estimate of the noise spectrum $|\hat{D}(\omega,k)|^2$. Although the noise spectrum cannot be directly computed, however, in practice we can estimate based on the spectrum of the unvoiced regions.

A common spectral subtraction technique [2] introduces α, an over-subtraction factor,

$$|\hat{S}(\omega,k)|^2 = |Y(\omega,k)|^2 - \alpha|\hat{D}(\omega,k)|^2, \tag{6}$$

where α is defined as

$$\alpha = \begin{cases} 5 & SSNR < 5 \\ 4 - \frac{3}{20}(SSNR) & -5 \leq SSNR \leq 20 \\ 1 & SSNR > 20 \end{cases}$$

While not a necessity in white-noise systems, a more robust algorithm known as multi-band spectral subtraction [3] can more accurately apply the appropriate over-subtraction factor should there be differing power of noise in different spectra. Hence we can consider a given speech signal to be broken up into spectrum bands according to

$$|\hat{S}_i(\omega,k)|^2 = |Y_i(\omega,k)|^2 - \alpha_i\delta_i|\hat{D}_i(\omega,k)|^2 \tag{7}$$
$$b_i \leq \omega \leq e_i,$$

where b_i and e_i are the beginning and end of a frequency band, α_i is the over-subtraction factor in the ith band and δ_i is an additional parameter we can choose to adjust the noise removal in each band. Each α_i is computed as

$$\alpha_i = \begin{cases} 5 & SSNR_i < 5 \\ 4 - \frac{3}{20}(SSNR_i) & -5 \leq SSNR_i \leq 20 \\ 1 & SSNR_i > 20 \end{cases}$$

where segmental signal-to-noise ratio $SSNR_i$ is computed as

$$SSNR_i = 10\log_{10}\left(\frac{\sum_{\omega=b_i}^{e_i}|Y_i(\omega,k)|^2}{\sum_{\omega=b_i}^{e_i}|\hat{D}_i(\omega,k)|^2}\right)$$

When computing $|\hat{S}_i(\omega,k)|^2$ from (7), we note that it's possible in this formulation for the difference to be negative, hence we have to set a bottom threshold

$$|\hat{S}_i(\omega,k)|^2 = \begin{cases} |\hat{S}_i(\omega,k)|^2 & |\hat{S}_i(\omega,k)|^2 > 0 \\ \beta|Y_i(\omega,k)|^2 & else \end{cases} \tag{8}$$

where $\beta = 0.002$ is the spectral floor parameter. From this estimate of $|\hat{S}_i(\omega,k)|^2$, we can combine this magnitude with the phase of $Y(\omega,k)$, take the inverse STFT, and combining the individual windows yielding the enhanced signal $\hat{s}[n]$. This entire process can be visualized in a block diagram in figure 1.

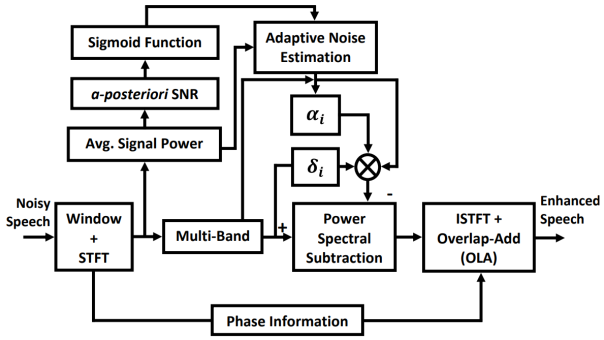

Fig. 1: Block diagram summarizing the multi-band spectral subtraction method [4]

C. Evaluation

The three methods are tested on five female utterances and five male utterances from the TIMIT database with 5 dB, 10 dB, and 15 dB additive white Gaussian noise. The waveforms and spectrograms of the clean signal, the noisy signal, and the enhanced signal are plotted to visualize the results. In addition, a quantitative metric for speech quality i.e. SNR and PESQ are used to evaluate the results.

III. Results

A. Wiener Filter

Using the Wiener Filter method increases the SNR of the 5 dB test data set by 4.39 dB, the SNR of the 10 dB test data set by 3.33 dB, and the SNR of the 15 dB test data set by 1.50 dB, on average. The figures below shows the waveforms and spectrograms of a clean signal, the noisy versions (with 5 dB, 10 dB, and 15 dB SNR), and the reconstructed signals. The results comparing the PESQ metric for varying noise levels is shown below. A higher PESQ value indicates higher perceptual quality.

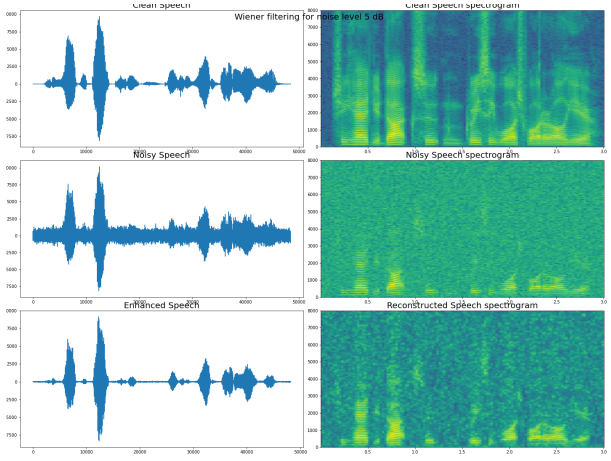

Fig. 2: Wiener Filter with 5 dB Noise

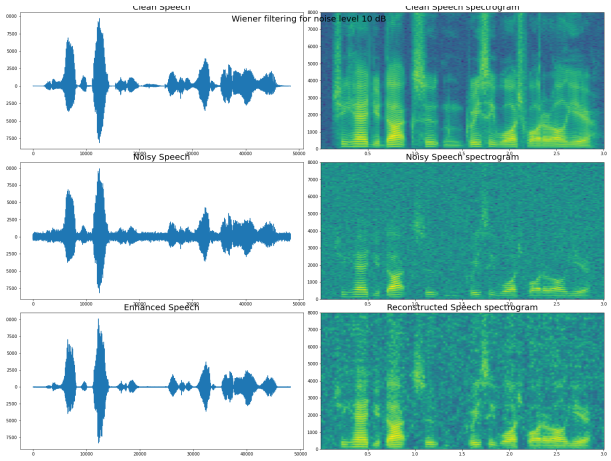

Fig. 3: Wiener Filter with 10 dB Noise

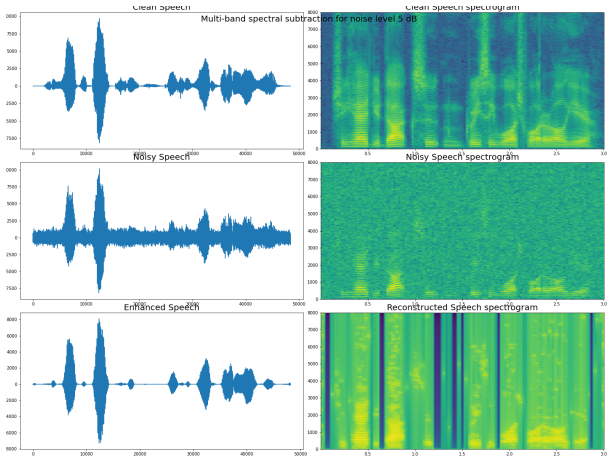

Fig. 5: Multiband spectral subtraction with 5 dB Noise

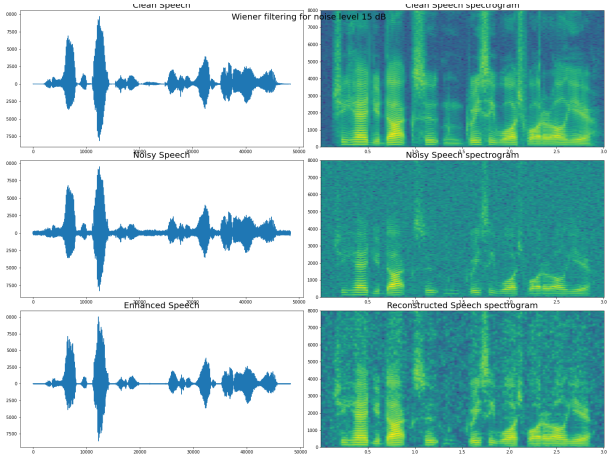

Fig. 4: Wiener Filter with 15 dB Noise

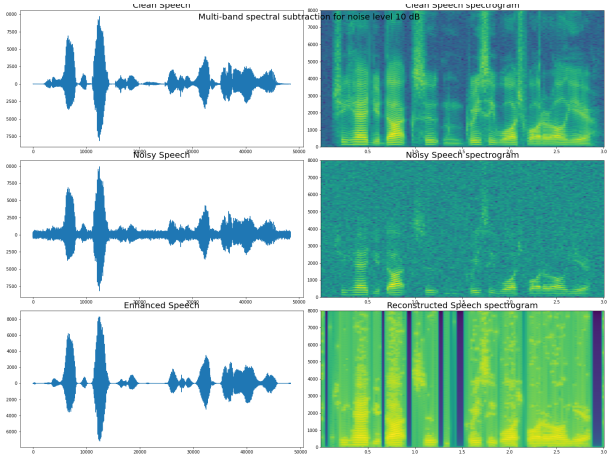

Fig. 6: Multiband spectral subtraction with 10 dB Noise

TABLE I: PESQ evaluation for 5 dB noise

Avg PESQ - Noisy	Avg PESQ - Reconstructed	Avg PESQ improvement
1.4268	1.9721	0.5453

TABLE II: PESQ evaluation for 10 dB noise

Avg PESQ - Noisy	Avg PESQ - Reconstructed	Avg PESQ improvement
1.7127	2.4973	0.78466

TABLE III: PESQ evaluation for 15 dB noise

Avg PESQ - Noisy	Avg PESQ - Reconstructed	Avg PESQ improvement
2.1184	2.9535	0.8351

B. Spectral Subtraction

The spectral subtraction method increases the SNR of the 5 dB data set by 3.50 dB and the SNR of the 10 dB by 2.24 dB on average. Spectral subtraction does not enhance the 15 dB speech signals in terms of the SNR. However, spectral subtraction does make the high frequency harmonic lines more visible in the spectrograms.The figures below shows the waveforms and spectrograms of a clean signal, the noisy versions (with 5 dB, 10 dB, and 15 dB SNR), and the reconstructed signals. The results comparing the PESQ metric for varying noise levels is shown below. A higher PESQ value indicates higher perceptual quality.

TABLE IV: PESQ evaluation for 5 dB noise

Avg PESQ - Noisy	Avg PESQ - Reconstructed	Avg PESQ improvement
1.4268	1.4426	0.0158

TABLE V: PESQ evaluation for 10 dB noise

Avg PESQ - Noisy	Avg PESQ - Reconstructed	Avg PESQ improvement
1.7127	2.0171	0.30442

TABLE VI: PESQ evaluation for 15 dB noise

Avg PESQ - Noisy	Avg PESQ - Reconstructed	Avg PESQ improvement
2.1184	2.4659	0.3475

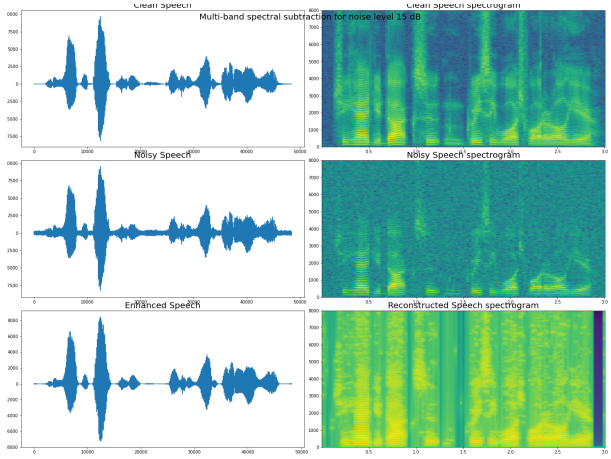

Fig. 7: Multiband spectral subtraction with 15 dB Noise

IV. Conclusion

We have thus compared two different methods for speech enhancement, namely Wiener Filtering and Multiband spectral subtraction. The noise being used is mainly additive white noise. Both the above mentioned methods approximate the noise in the initial frames to be the noise in the silent segments. Hence for higher noise levels, we see a drop in performance, as it becomes more difficult to distinguish between silent and voiced frames. Correspondingly, we see a drop in PESQ for higher noise levels.

We see that Wiener filtering performs better than Spectral subtraction for all noise levels. In spectral subtraction, as an estimate of noise spectrum is computed from segments of speech absence, and is subtracted from noisy speech spectrum, the method is not efficient for speech corrupted with non-stationary noise such as car noise, babble noise, helicopter noise. Plain spectral subtraction is also not efficient for higher SNRs as it degrades the actual signal too. Wiener filter is a better approach in such cases.

V. Contribution

- Gregory Hessler
- Kshama Kodthalu Shivashankara
- Yue Teng

References

[1] 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, pp. 749–752 vol.2, 2001.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211, 1979.

[3] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, vol. 4, 05 2002.

[4] N. Upadhyay and A. Karmakar, "An improved multi-band spectral subtraction algorithm for enhancing speech in various noise environments," *Procedia Engineering*, vol. 64, pp. 312–321, 2013. International Conference on Design and Manufacturing (IConDM2013).