Speech Denoising and Audio Enhancement

Gore Kao
Electrical and Computer Engineering
gorek@andrew.cmu.edu

Yuhao Liu
Electrical and Computer Engineering
yuhaoliu@andrew.cmu.edu

Yifan Peng
Electrical and Computer Engineering
yifanpen@andrew.cmu.edu

Yunyang Zeng
Electrical and Computer Engineering
yunyangz@andrew.cmu.edu

Abstract

Noise in speech signals is often present due to distractors such as radio, television or white noise etc, therefore speech denoising tasks have become ubiquitous to improve the integrity of the desired signal. In this work, we explore two methods of speech denoising: Non-Negative Matrix Factorization (NMF) and Wiener Filtering. In addition to standard NMF procedures, we attempt to add regularization to improve results. We also included the concept of A Prior SNR in addition to the regular Wiener Filtering method.

1 Introduction

The objective of speech denoising and audio enhancement is to remove any interference from a degraded speech signal in order to improve the clarity and quality of information contained in the signal. In traditional speech recognition tasks, this is especially important as a preprocessing step to obtain clean data for the following downstream tasks. In the past decades, significant improvements have been made on the performance of Automatic Speech Recognition (ASR). However, in order to achieve high performance, the input speech signals are required to be recorded in noiseless environments, which is usually impractical. In this project, we explore some effective algorithms to reduce the ambient noise in real-life speech recordings.

2 Related Work

Previous work from Ding et al. presents a Discrete Cosine Time (DCT) based system for speech enhancement [1]. The work states that the DCT has been shown to be a good approximation to the Karhunen–Loeve Transform (KLT) while also providing better energy compactness. While KLT has shown is a popular algorithm and has seen success in speech enhancement, there is a high computational complexity with the implementation. The DCT offers an energy compact solution and is shown to provide a solution that improves the signal-to-noise ratio (SNR). The work by Soon et al. further highlights the advantages of DCT as an effective and energy efficient solution [2]. Furthermore, previous works show that Non-negative Matrix Factorization (NMF) works well for separating sounds when the building blocks for different sources are sufficiently distinct [3]. Also, a conventional denoising technique, Wiener Filtering is proved to perform well on stationary environments [4].

Additionally, a spectral filtering method by Maher has also shown success in audio enhancement. The method extends upon previous systems coupling a time domain level detection with a frequency domain filtration step by distinguishing between the coherent audio and incoherent noise of the desired signal components [5]. The work utilizes the short time Fourier transform (SFTF) to identify

parts of the signal that behave consistently over a window frame and parts that fluctuate, indicating unwanted noise.

3 Dataset

For current experiments, we used the clean speech from TIMIT [6], which contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. To synthesize noisy signals, we used the background noise from Google Speech Commands [7] and some noise recordings from [8]. The mixing script is adapted from an open-source tool [1].

4 Results

4.1 Evaluation metric:

In signal processing, the signal-to-noise ratio (SNR) is widely used to measure the quality of a degraded signal. It is defined as the ratio of the power of a signal to the power of noise: SNR $= \frac{P_{\text{signal}}}{P_{\text{noise}}}$. We use a python package *pysepm*[2] to calculate the segmental SNR.

4.2 Non-Negative Matrix Factorization (NMF)

For a given signal, V, we attempt to reconstruct the signal by learning a set of bases (building blocks) and weights, W and H such that $V \approx WH$. We utilize a standard Non-negative Matrix Factorization (NMF) technique to learn these building blocks given clean training signals and assume these building blocks will be able to represent the clean portions of a signal well in a given corrupted signal (mixed with noise). For our mixed signal, we mix the original clean signal with white noise to produce a new signal with a specific SNR. The general objective function we attempt to minimize is shown in Figure 1.

$$
\begin{aligned}
&0.5 * ||X - WH||^2_{loss} \\
&+alpha\_W * l1_{ratio} * n\_features * ||vec(W)||_1 \\
&+alpha\_H * l1_{ratio} * n\_samples * ||vec(H)||_1 \\
&+0.5 * alpha\_W * (1 - l1_{ratio}) * n\_features * ||W||^2_{Fro} \\
&+0.5 * alpha\_H * (1 - l1_{ratio}) * n\_samples * ||H||^2_{Fro}
\end{aligned}
$$

Figure 1: Objective function to be minimized for NMF (using KL divergence loss) [9].

During training, we utilize the availability of a clean speech spectrogram V_{speech}, and a noisy spectrogram (without clean speech), V_{noise}. Previous work in audio separation has shown that using the Kullback-Leibler (KL) divergence as the objective function has worked well [10], therefore we attempt to minimize the KL Divergence during our training phase. In particular, the KL Divergence between the original signal, V, and the reconstructed signal, WH, can be written as $D(V||WH)$. The building blocks for both the clean speech signal and noisy signal are learned separately to obtain W_{speech} and W_{noise} respectively.

To reconstruct the denoised spectrogram, we first fix the learned building blocks from the training phase and concatenate the bases to form $W_{mixed} = [W_{speech} W_{noisy}]$. Following that, we can learn the weights for the mixed signal H_{mixed} corresponding to the fixed building blocks. To obtain the denoised signal, we take the rows in H_{mixed} corresponding to the columns in W_{speech} to get a reconstruction $V_{speech} = W_{speech} H_{mixed;1:n_b}$, where n_b is the number of columns in W_{speech} [3]. The entire flow diagram in which we follow our baseline implementation can be seen in Figure 2 [11].

In addition to standard NMF techniques, which can be obtained by setting alpha and l1 ratio hyper parameters in the objective function above to 0 and minimizing with the KL divergence, we also

[1] https://github.com/Sato-Kunihiko/audio-SNR

[2] https://github.com/schmiph2/pysepm

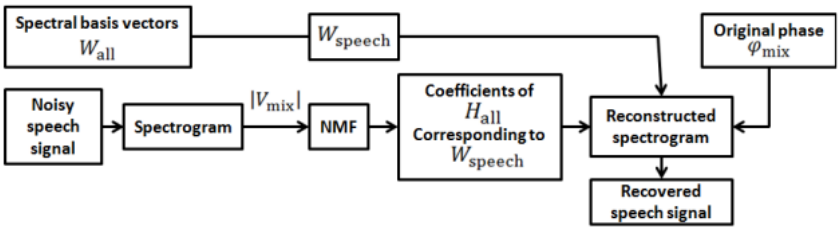

Figure 2: NMF training procedure block diagram Ludena et al.[11]

explore a few different values for alpha and l1 ratio to see how reconstruction is affected by the regularization terms. Our results show that a mix between the l1 and l2 ratio can help reconstruction error, as well a small value of alpha. There are minimal improvement as alpha decreases too much so we do not experiment past that point. The results of the mean reconstruction error are shown in the table detailed in Figure 3.

		l1 ratio		
		0	0.5	1
alpha	0	11.348	9.638	10.146
	0.1	10.984	8.937	9.865
	0.01	9.351	7.838	8.691
	0.001	9.501	7.935	8.788

Figure 3: Mean reconstruction error with NMF techniques including regularization.

To visualize the reconstructed signals, we also show the spectrogram of signals comparing the clean speech, noisy mixed speech, and reconstructed speech shown in Figure 4 (this can be compared with the spectrogram reconstruction of Wiener Filtering in the following subsection).

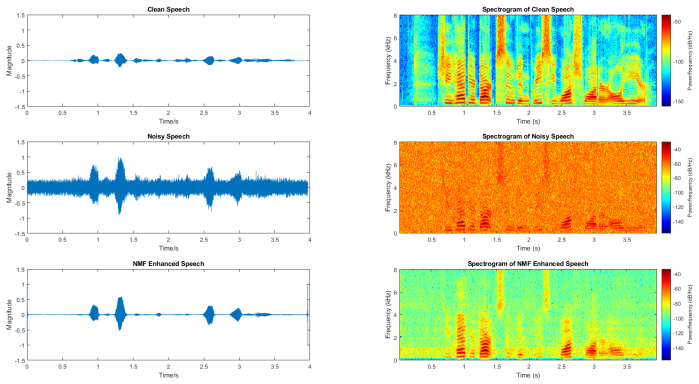

Figure 4: Spectrogram of signals using NMF with white, stationary, and additive noise

4.3 Wiener Filtering

A classical speech denoising method, the Wiener Filtering method is also studied. The Wiener Filter is a classical statistical approach of estimating an unknown signal using a correlated known signal as input to a filter to produce the estimate as output. The Wiener Filter is optimal in terms of the mean squared error. In Speech enhancement applications the Wiener Filter takes a noisy speech as input, estimates the clean speech based on minimum mean squared error optimization rule.

The Wiener Filtering Algorithm we studied is propsed by P. Scalart and J.V. Filho.[12]. Let $s(t)$ and $b(t)$ denote the speech and the additive noise process. The observed signal $x(t)$ is given by

$x(t) = s(t) + b(t)$. Denote the amplitude of kth spectral component of the signal $s(t)$, the noise $b(t)$ and the observation $x(t)$ as A_k, B_k, R_k, where $S_k = A_k e^{ja_k}$, $X_k = R_k e^{jv_k}$ in the analysis interval [0,T] and quasi-stationarity is guaranteed over the period T. The amplitude estimate $\hat{A}_k$ is obtained by multiplying X_k with a non-linear gain function defined as $G(f_k) = \frac{\hat{A}_k}{X_k}$. Based on the minimum mean squared error estimation, the optimal gain function $G_w(f_k) = \frac{SNR^k_{prior}}{1+SNR^k_{prior}}$. The estimate of SNR^k_{prior}, $S\hat{N}R^k_{prior} = \alpha\frac{|\hat{A}_{k-1}(w)|^2}{|\hat{B}_{k-1}(w)|^2} + (1-\alpha)\max[S\hat{N}R^k_{post} - 1, 0]$ and $S\hat{N}R^k_{post} = \frac{|R_k(w)|^2}{|\hat{B}_k(w)|^2}$. In these equations, the $|\hat{A}_{k-1}(w)|^2$ is the power spectral density estimation of the $k-1$th analysis frame of the clean signal. $|R_k(w)|^2$ is the power spectral density of the kth analysis frame of the noisy signal. $|\hat{B}_k(w)|^2$ is the power spectral density estimation of the kth analysis frame of the noise signal. α is a smoothing factor in prior update.

During testing ,the noisy speech signal is synthesized by adding a clean speech with a stationary white gaussian noise at 0dB SNR. The Wiener Filter has been tested on 760 synthesized noisy speech files with duration around 3 seconds. The average SNR improvement is 7.39dB. In theory, the Wiener Filtering method performs well when the additive signals (in this case both the clean speech and the noise) are stationary random processes. We will show in the following discussion that the performance of the Wiener Filtering method degrades when the signals are non-stationary. Figure 5 shows the performance of using Wiener Filter to enhance a speech corrupted by white gaussian noise.

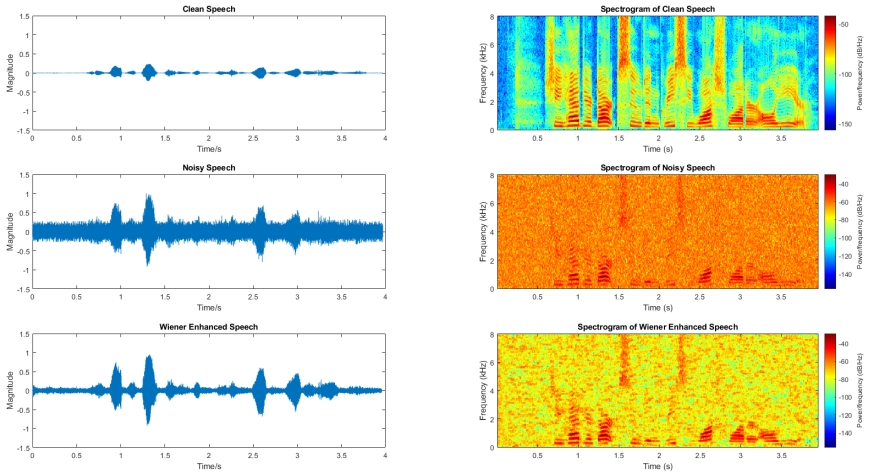

Figure 5: Spectrogram of signals using Wiener Filtering with white, stationary, and additive noise.

We also tested the Wiener Filtering algorithm using non-stationary noises. Specifically, the noise signal is recorded in a meeting environment [8], so it contains both random noise and human speech from other speakers. The Wiener Filtering algorithm is able to reduce random noise but cannot extract the speech from the target speaker, as shown in Figure 6. The SNR before enhancement is -2.8982, and the SNR after enhancement is -1.1707. The SNR is improved slightly, because the random part of the noise signal has been reduced. However, the non-stationary part cannot be suppressed.

5 Discussion and Analysis

In this project, we applied a few machine learning and signal processing algorithms to speech denoising. Generally, to recover a clean speech signal from a corrupted signal, we need to utilize the characteristics of both clean and noise signals.

In NMF-based methods, we first learn a set of basis from the training corpus and then reconstruct the clean signal using these basis. This method works well if the desired signal has very different basis compared with noises. However, if there are some common building blocks, it will be challenging to

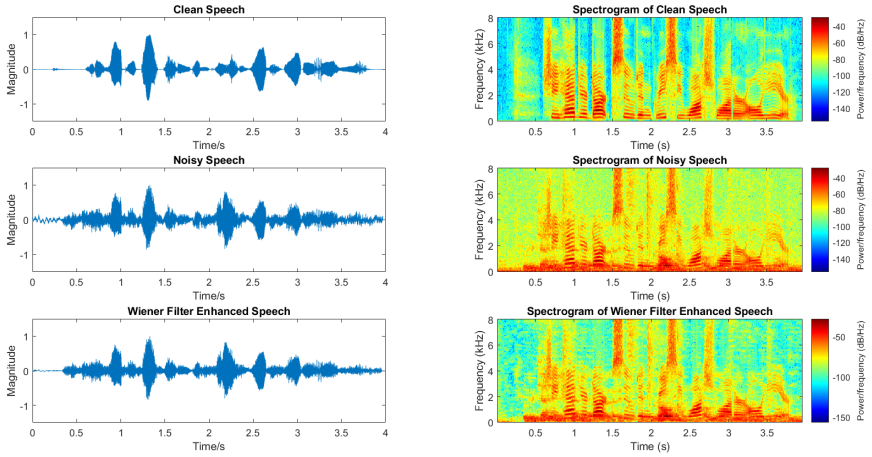

Figure 6: Spectrogram of signals with Wiener Filtering under conditions of non stationary noise. The waveform of the enhanced signal is closer to the noisy speech but is very different from the original clean speech.

distinguish between these two signals. In our experiments, we mainly used random noise, which is irrelevant to the desired clean speech, so this method achieved good performance.

In Wiener filtering, we apply an adaptive filter to the observed corrupted signal and minimize the mean squared error to estimate the original clean speech. In this algorithm, the additive noise is assumed to be stationary with known statistical or spectral characteristics. This approach generates good results when white noise is used. But when we add non-stationary noise recorded in a real meeting environment, the performance degrades as expected.

Because the NMF method has a training stage and can benefit from environment specific noise models, we do not consider a direct comparison between the performance of NMF method and Wiener Filtering method to be useful. However, it is noticeable that for white gaussian noise, from figure 4 and figure 5 the NMF method has higher SNR performance than the Wiener Filtering method. It is also worth to notice that by listening to the two files in our *demo*[3], one can hear that the NMF suppresses sibilants such as 's' and 'sh' to a higher extent than the Wiener Filter does. We conclude that for white gaussian noise, the NMF has better performance in terms of source to noise ratio but worse source to distortion ratio than the Wiener Filtering method.

To improve the performance of speech denoising in a practical setting, we may consider the following directions in the future. First, we need to learn better representations for clean and noise signals so that we can separate them more accurately. Second, we have only considered single-channel signals. However, in a real environment, a microphone array is often used, which generates multi-channel speech signals. By using multiple channels, the desired signal can be recovered or separated more effectively and efficiently.

[3]https://github.com/toxidol/MLSP-FA2021/tree/main/demo

References

[1] Huijun Ding, Yann Soon, and Chai Kiat Yeo. A dct-based speech enhancement system with pitch synchronous analysis. *IEEE Transactions on audio, speech, and language processing*, 19 (8):2614–2623, 2011.

[2] Yann Soon, Soo Ngee Koh, and Chai Kiat Yeo. Noisy speech enhancement using discrete cosine transform. *Speech communication*, 24(3):249–257, 1998.

[3] Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4029–4032. IEEE, 2008.

[4] "speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. *Tech. Rep. ETSI ES*, 202 050 V1.1.5, 2007-01.

[5] Robert C Maher. Audio enancement using nonlinear time-frequency filtering. In *Audio Engineering Society Conference: 26th International Conference: Audio Forensics in the Digital Age*. Audio Engineering Society, 2005.

[6] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 11 1992.

[7] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[8] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, June 2013. URL `https://doi.org/10.5281/zenodo.1227121`. Supported by Inria under the Associate Team Program VERSAMUS.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[11] Jimmy Ludeña-Choez and Ascensión Gallardo-Antolín. Speech denoising using non-negative matrix factorization with kullback-leibler divergence and sparseness constraints. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 207–216. Springer, 2012.

[12] P. Scalart and J.V. Filho. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632 vol. 2, 1996. doi: 10.1109/ICASSP.1996.543199.