Project Report

Team 32

Srujana Vanka (2020102005)

Sankeerthana Venugopal (2020102008)

Suppression of Acoustic Noise in Speech Using Spectral Subtraction

Abstract

A noise suppression algorithm is developed to reduce the spectral effects of acoustically added noise in speech. Spectral Subtraction is a computationally efficient approach to digital speech analysis. The noise is removed by subtracting the average stationary noise spectrum obtained during non speech activity, from the noise corrupted speech spectrum. Further, various methods are employed to bring the speech estimate closer and closer to the clean speech signal.

Introduction

We use speech communications on a daily basis. A speaker, a listener, and several communication equipment are all engaged in speech communication. Random noises frequently obstruct communication between speaker and listener, which degrades the speech performance, lowers quality and intelligibility. In the literature, a range of noise suppressing strategies capable of reducing background noise have been studied. These include- using noise-cancelling microphones, internal modification of the voice processor algorithms to explicitly compensate for signal contamination, or pre-processor noise reduction.

The goal of this project is to create a noise suppression technique and implement a computationally efficient algorithm. Spectral subtraction is a processor-independent, computationally efficient way to arrive at effective digital speech analysis. The assumption is that the noise is a stationary or a slowly varying process, and that the noise spectrum does not change significantly in-between the update periods. Spectral subtraction needs only noisy speech as input. For this, an estimator is obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The signal collected during non-speech activity provides the spectral information needed to define the noise spectrum.

The Additive Noise Model

Let x(t) be a noise corrupted signal. We model x(t) as:

$x(t) = s(t) + n(t)$, where $s(t)$ is the windowed clean speech signal and $n(t)$ is the windowed unwanted background noise.

This can be discretized as

$$x[k] = s[k] + n[k]$$

Taking Fourier Transform on both sides,

$X(e^{jw}) = S(e^{jw}) + N(e^{jw})$

where $X(e^{jw}) = \sum_{k=0}^{L-1} x[k]e^{-jwk}$

Spectral Subtraction

$X(e^{jw}) = |X(e^{jw})|e^{j\theta_x}$ and $N(e^{jw}) = |N(e^{jw})|e^{j\theta_n}$

$S(e^{jw}) = X(e^{jw}) - N(e^{jw})$

$N(e^{jw})$ is generally unknown

Here, we model noise to be stationary. Hence, it can be replaced with $\mu(e^{jw})$, the Fourier Transform of the average value of n[k] taken during non-speech activities over different time frames

$$\hat{S}(e^{jw}) = [|X(e^{jw})| - \mu(e^{jw})]e^{j\theta_x}$$

Spectral Error

The spectral error can be calculated as

$$\epsilon(e^{jw}) = \hat{S}(e^{jw}) - S(e^{jw}) = N(e^{jw}) - \mu(e^{jw})e^{j\theta_x}$$

A number of simple modifications are available to reduce the auditory effects of this spectral error.

This error can be reduced by the following four methods:

1) Magnitude Averaging

From above, we know that the spectral error is the difference between the spectral subtraction estimator and the speech signal. This spectral error can be reduced by taking the local average of the spectral magnitudes, $i.e$, by replacing $|X(e^{jw})|$ with $\overline{|X(e^{jw})|}$ where,

$$\overline{|\mathrm{X}(\mathrm{e}^{\mathrm{jw}})|} = \frac{1}{M}\sum_{i=0}^{M-1}\overline{|\mathrm{X_i}(\mathrm{e}^{\mathrm{jw}})|}$$

This is basically the window sliding technique. Therefore, $X_i(e^{jw}) = i^{th}$ time windowed transform of $x(k)$.

Hence, on magnitude averaging, $\hat{S}(e^{jw}) = [\overline{|\mathrm{X}(\mathrm{e}^{\mathrm{jw}})|} - \mu(e^{jw})]e^{j\theta_x(e^{jw})}$

This in turn makes the spectral error almost equal to $\overline{|\mathrm{N}|} - \mu$, where $\overline{|\mathrm{N}|}$ is the local average of the noise spectrum magnitude. On taking a longer average, the spectral error reduces as $\overline{|\mathrm{N}|}$ almost coincides with μ.

2) Half-wave rectification

For each frequency w, if $|X(e^{jw})| < \mu(e^{jw}),$ then $\hat{S}(e^{jw}) = 0, i.e,$ if the spectral magnitude is less than the average noise spectrum magnitude, the output is equal to zero. To arrive at this, we use half wave rectification method on the spectral subtraction filter $H(e^{jw}), i.e,$

$$H_R(e^{jw}) = \frac{H(e^{jw}) + |H(e^{jw})|}{2} \quad and \quad \hat{S}(e^{jw}) = H_R(e^{jw})X(e^{jw})$$

As a result, half-wave rectification causes the magnitude spectrum at each frequency w to be biased down by the noise bias established at that frequency.

The drawback of half rectification can be seen in situations where the noisy signal at a frequency w is lower than $\mu(e^{jw})$. This in turn will give incorrect speech information at that frequency.

3) Residual Noise Reduction

After half-wave rectification, the spectrum lying above $\mu(e^{jw})$ remains. If there is no speech activity , the noise residual, $N_R = N - \mu e^{j\theta_n}$ takes values between zero and a maximum.

This noise residual will be random from frame to frame. Hence, whenever

$|\hat{S}(e^{jw})| < N_{R_{MAX}}$, (implies an absence of speech activity), the noise residual is suppressed by taking the minimum magnitude value from the three adjacent analysis frames;

$$\hat{S}_j(e^{jw}) = min[\hat{S}_{j-1}(e^{jw}), \hat{S}_j(e^{jw}), \hat{S}_{j+1}(e^{jw})] \qquad if, |\hat{S}_j(e^{jw})| < N_{R_{MAX}}$$

4) Additional Signal Attenuation During Non Speech Activity

The ratio of $\hat{S}(e^{jw})$ to $\mu(e^{jw})$ reflects the absence or presence of speech activity in a given analysis frame. If there is no speech activity, $\hat{S}(e^{jw})$ contains only the noise residual.

This residual noise can be set to zero. However, having some signal present during non speech activity was judged to give the higher quality result. So, wherever the signal to average noise ratio was below a certain threshold value, the output was attenuated by a factor c.

$$T = 20\log_{10}[\frac{1}{2\pi}\int_{-\pi}^{+\pi}|\frac{\hat{S}(e^{jw})}{\mu(e^{jw})}|dw]$$

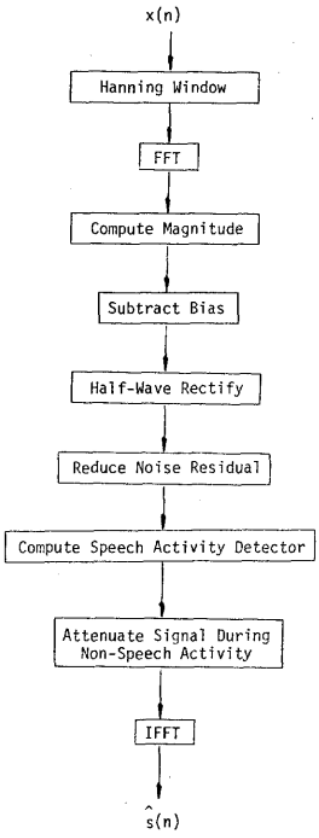
Thus, the output spectral estimate including output attenuation during non-speech activity is given by:

$$\hat{S}(e^{jw}) = \hat{S}(e^{jw}) \qquad \text{if} \quad T > -12dB$$

$$\hat{S}(e^{jw}) = cX(e^{jw}) \qquad \text{if} \quad T < -12dB$$

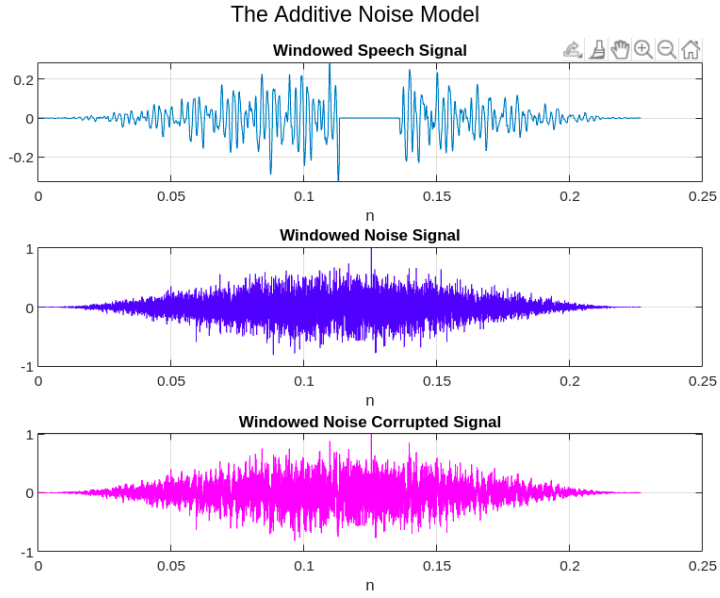
Implementation and results

- The system block diagram for the implementation is given below:

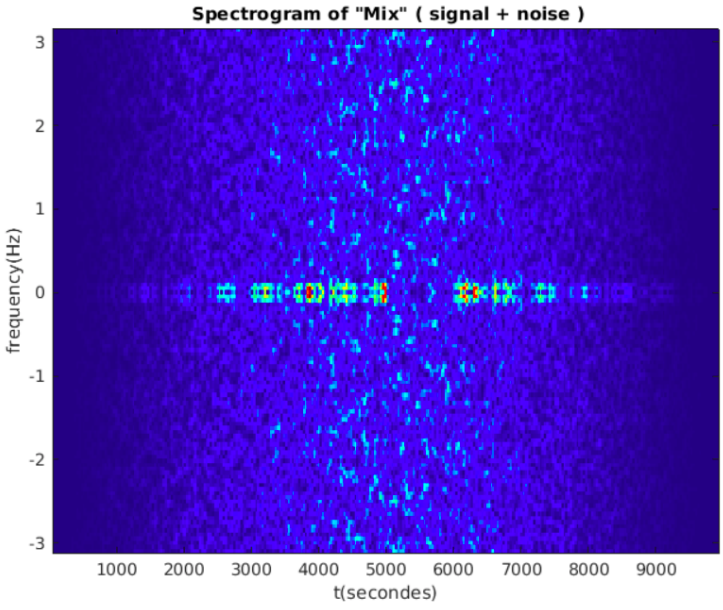


- The additive noise model is created by taking the sum of the windowed (hanning window) speech signal and the windowed noise signal, where speech signal is taken to be an audio file, noise signal is gaussian noise that is generated using randn() function on MATLAB. It is assumed that the background noise is acoustically or digitally added to the speech, and the background noise environment remains locally stationary to the degree that the expected value of it's spectral magnitude just prior to speech activity equals its expected value during speech activity.

The plot of the additive noise model in time domain is given below:



The same additive noise model is given in the form of a spectrogram:

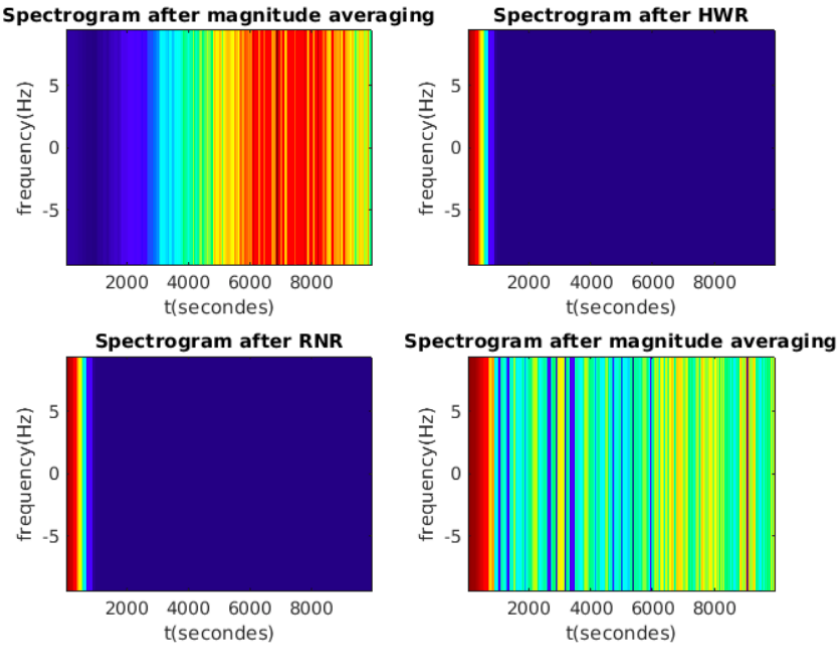


Visual representation of frequencies of a given signal with time is called Spectrogram. In a spectrogram representation plot — one axis represents the time, the second axis represents frequencies and the colors represent magnitude (amplitude) of the observed frequency at a particular time.

- Once the additive noise model is generated, the short time fourier tranform of the noisy signal is taken and the spectral subtraction estimator is determined:

$$\hat{S}(e^{jw}) = [|X(e^{jw})| - \mu(e^{jw})]e^{j\theta_x}$$

- After obtaining the spectral estimate, we can compute the spectral error: $\epsilon(e^{jw}) = \hat{S}(e^{jw}) - S(e^{jw}) = N(e^{jw}) - \mu(e^{jw})e^{j\theta_x}$ and then error reduction methods are implemented.
- The following spectrogram shows the spectral estimate after each error reduction method:

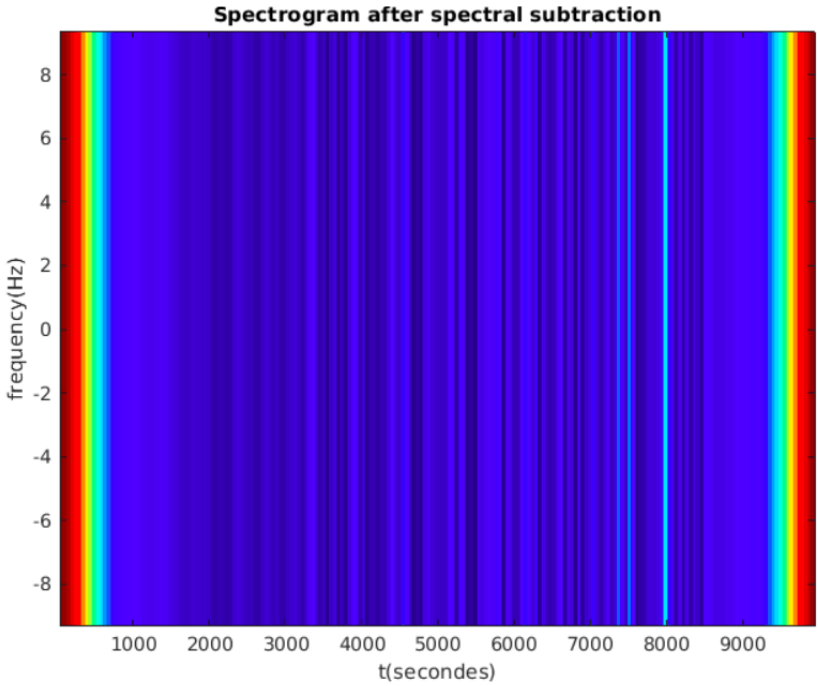


- **Some of the challenges in the error reduction methods:**

1. Speech is non-stationary, so in case of magnitude averaging, averaging over a longer period of time will decrease intelligibility.
2. In half wave rectification, areas where the noisy speech at a given frequency is less the than estimated noise, the speech information gets incorrectly removed at that frequency.

3. In residual noise reduction, more storage is required to store the minimum magnitude values of the adjacent frames.

- To reconstruct the noise reduced signal back in the time domain, we take the inverse fourier transform. The spectrogram obtained is shown below:



Conclusion

In this project, noise removal from noisy speeches has been studied and analyzed by evaluating graphs and simulation results. The study includes methods for removing noise from noisy speeches using spectral Subtraction. Spectral estimates for the background noise were obtained from the input signal during non-speech activity. The algorithm can be implemented using a single microphone source and requires about the same computation as a high-speech convolution. Results indicate overall significant improvements in quality and intelligibility.

Acknowledgement

We would like to think our professor Mr. Santosh for giving us the opportunity to do this project. The project was a profound learning experience for us which also helped us in doing a lot of research. We would also like to thank our TAs for helping and guiding us whenever necessary. Lastly, we would like to thank my teammate for their constant support and contributions that led to the completion of this project.

References

[1] Cole, C., Karam, M., & Aglan, H. (2008), “ Spectral Subtraction of Noise in Speech Processing Applications” IEEE

[2] S. F. Boll, “Suppression of noise in speech using the SABER method,” in Proc. ZEEE Znt. Conf. on Acoust., Speech, Signal Processing, Tulsa, OK, Apr. 1978, pp. 606-609.

[3] To check the reference/bibliography please open sci-hub, search you paper there by writing in the format as mentioned above.