IOSR Journal of Engineering
Mar. 2012, Vol. 2(3) pp: 376-381

Speech Enhancement using Spectral Subtraction

N. Siddiah[1], T.Srikanth[2] and M. Venkatesh Varma[3]

[1]Dept. of E.C.E., Mekapati Raja Mohana Reddy Institute of Technology & Science, Udyagiri, A.P., India.
[2]Dept. of E.C.E., Kallam Haranadha Reddy Institute of Technology, Guntur, A.P., India.
[3]Dept. of E.C.E., Chalapathi Institute of Technology, Guntur, A.P., India.

Abstract

This paper address several problems associated with Automatic Speech Recognition Systems (ASR) and study a speech enhancement technique that could possibly reduce the inefficiencies that ASR systems encounter. Spectral Subtraction (SS) is a method used to reduce the amount of noise acoustically added in the speech signals. Our goal is to implement the SS algorithm to provide speech enhancement while researching Automatic Speech Recognition, to discover whether SS can enhance the efficiency of ASR systems. Spectral Subtraction is an algorithm designed to reduce the degrading effects of noise acoustically added in speech signals.

This paper focuses on the removal of white noise in speech signals, and attempts to explain how SS can improve ASR systems. The importance of the SS method over other methods is also explored. As our day-to-day lives become more complicated, ASR provides a hands free way to complete a variety of duties by simply speaking. Used effectively ASR systems can optimize most tasks and allow a user to complete them at a rate that is substantially faster. In addition these systems can enhance the way the hearing impaired communicate, improve security, and can provide authentication for many applications. For these and many more reasons, there is an obvious requirement for adequate ASR systems and their integration into our everyday life.

Keywords: Speech Enhancement, Speech Recognition, Spectral Subtraction, Windowing techniques, Noise reduction.

I. INTRODUCTION

Many systems rely on automatic speech recognition (ASR) to carry out their required tasks. Using speech as its input to perform certain tasks, it is important to ensure that background noise will not degrade the performance of systems or ultimately completely inhibited. Spectral Subtraction (SS) is an algorithm designed to reduce the degrading effects of noise acoustically added in speech signals. With applications from speech and language development in young children to aiding individuals with hearing impairments ASR is becoming increasingly popular and the demand for efficient systems is more evident. While humans are the best examples of ASR, the term as we know it usually means the process in which a computer recognizes and/or identifies spoken words. Not with standing any task that involves interfacing with a computer can potentially use ASR, the following applications are the most common right now: Dictation, Command and Control, Mobile, Personal Accessories and medical or disability.

The spectral subtraction algorithm is historically one of the first algorithms proposed for noise reduction [1, 2], and is perhaps one of the most popular algorithms. It is based on a simple principle. Assuming additive noise, one can obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum of the noisy speech spectrum. The noise spectrum can be estimated and updated during periods when the signal is absent. The enhanced signal is obtained by calculating the inverse discrete Fourier transform spectrum of the signal estimated by the phase of the signal with noise. The algorithm is computationally simple, since it only involves a single step forward and inverse Fourier transform.

The simple subtraction processing comes with a price. The subtraction process needs to be done carefully to avoid any speech distortion. If too much is subtracted, then some speech information might be removed, while if too little is subtracted then much of the interfering noise remains. Many methods have been proposed to alleviate, and in some cases, eliminate some of the speech distortion introduced by the spectral subtraction process [3]. Some suggested over-subtracting estimates of the noise spectrum and spectral flooring (rather than setting to zero) negative values [4]. Others suggested dividing the spectrum into a few contiguous frequency bands and applying different non-linear rules in each band [5, 6]. Yet, others suggested using a psychoacoustical model to adjust the over-subtraction parameters so as to render the residual noise inaudible [7].

The derivation of the equations spectral subtraction is based on the assumption that the cross terms involving the phase difference between signals clean and noise are zero. The cross terms is assumed to be zero because the speech signal is uncorrelated with noise interference. Several attempts have been made to take into account or other wise compensate the cross-terms [8, 9, 10], spectral subtraction. The study in [10] evaluated the effect of neglecting the cross terms on the performance of speech recognition.

This paper focuses on present day problems associated with speech recognition, especially the removal of white noise in speech signals, and attempts to explain how SS can improve ASR systems. White noise is a type of noise that is produced by combining sounds of all different frequencies together. Because it contains all frequencies white noise can drown or mask other sounds, which may contain significant information, needed for input into an ASR system. If a reasonable estimate of white noise contained in a given speech signal can be obtained and removed from a signal, then we should see an improvement in the quality of the speech and efficiency for most ASR systems.

Other methods used to reduce the amount of noise in speech signals include: Noise cancelling microphones, although essential for extremely high noise environments such as the helicopter cockpit, they offer little or no noise reduction above 1 kHz.

Another and one of the most efficient techniques to improve robustness of speech recognition systems on additive noise consists in training the acoustic models with data corrupted by noise at different signal-to-noise ratios (SNR). However as it is stated this method requires training by individuals in different environments, which may or may not be available in all situation.

II. SPEECH ENHANCEMENT TECHNIUES

Speech enhancement (SE),i.e,ways that a speech signal, subject to certain degradations (e.g ,additive noise,interfering talkers, bandlimiting), can be processed to increase its intelligibility (the likelihood of being correctly understood) and/or its quality.

There are three classes of SE methods, each with its own advantages and limitations:

1. Harmonic Filtering
2. Parametric Resynthesis
3. Spectral Subtraction

A. Harmonic Filtering

This method works only for voiced speech, requires an Fo estimate, and suppresses spectral energy between desired harmonics.

The harmonic SE method attempts to identify the Fo (and hence harmonics) either of the desired speech or of interfering sources. If the desired sound is the strongest component in the signal, its frequencies can be identified and other frequencies may then be suppressed; otherwise a strong interfering sound's frequencies can be identified and suppressed, with the remaining frequencies presumably retaining some of the desired speech source. Such simple weiner filtering **(**suppressing wide band noise between harmonics) improves SNR but has little effect on intelligibility.

B. Parametric Resynthesis:

This method adopts a specific speech production model (e.g., from low-rate coding), and reconstructs a clean speech signal based on the model, using parameter estimates from the noisy speech.

The parametric resynthesis SE method improves speech signals by parametric estimation and speech resynthesis. Speech synthesizers generate noise-free speech from parametric representations of either a vocal tract model or previously analysed speech.

Most synthesizers employ separate representations for vocal tract shape and excitation information, coding the former with about 10 spectral parameters and coding the later with estimates of intensity and periodicity (e.g. Fo). Such synthesis suffers from the same mechanical quality as found in low-rate speech coding and from degraded parameter estimate (due to noise).

C. Spectral Subtraction (SS):

Spectral subtraction (SS) is an algorithm which is used to reduce the amount of noise acoustically added in the speech signal. In this method we subtract the noise power spectrum from noisy signal power spectrum.

In the case of negative signal-to-noise ratio(SNR)(i.e., more energy in the interference than in the desired speech),this method works well for both general noise and interfering speakers, although musical tone or noise artifacts often occur at frame boundaries in such reconstructed speech. SS generally reduces noise power (improving quality), but often reduces intelligibility (especially in low SNR situations), due to suppression of weak portions of speech (e.g., high frequency formats and unvoiced speech).

Segmenting the Data

The data from the signal are segmented and windowed, such that if the sequence is separated into half-overlapped data buffers, then the sum of these windowed sequences adds back up to the original sequence. 10ms windows of data were used in this analysis.

Windowing is the multiplication of a speech signal S(n) by a window W(n), which yields a set of speech samples X(n) weighted by a shape of window.

Where S(n) = speech signal
W(n) = windowing function
X(n) = Noisy speech signal

W(n) may have an infinite duration but most practical windows have finite length to simplify computation. Many applications prefer some speech averaging, to yield an output parameter contour (vs. time) that represents some slowly varying physiological aspects of vocal tract movements.

Types of Windows:

1. Hamming window

2. Hanning window

3. Kaiser window

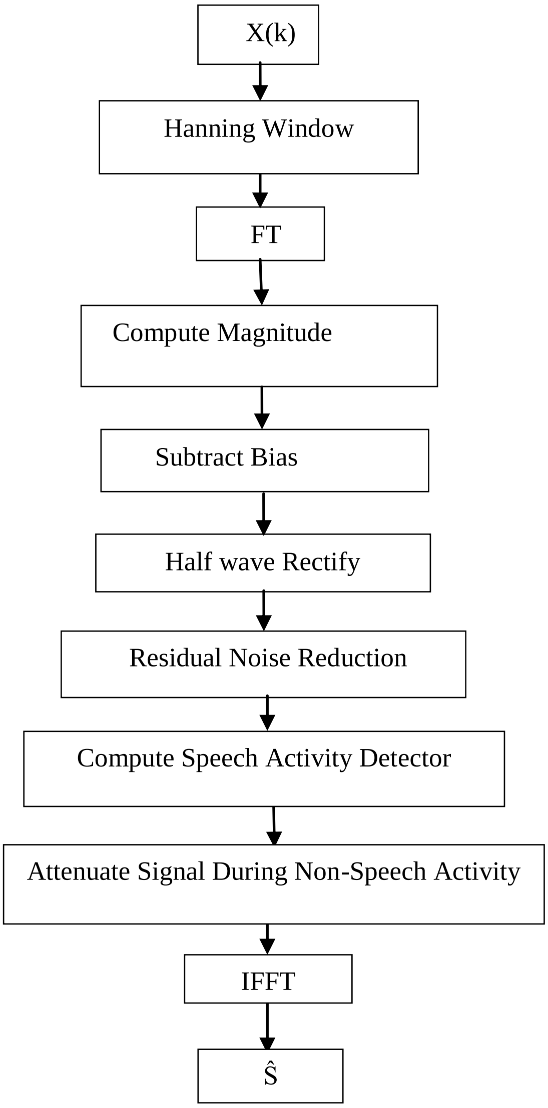

Fig. 1: Flow chart for spectral subtraction

Hamming Window

For Hamming window the attenuation coefficient (α) is 0.54.At low frequencies the stop band attenuation is high, so the ripples presented in stop band is more when compared to hanning window. The hamming window results in both pass band and stop band of the filter.

$W_{hm}(n) = 0.54+0.46\cos(2\Pi n/N-1)$,
$-(N-1)/2<=n<=(N-1)/2$
$= 0$,
otherwise

For Hanning window the attenuation coefficient(α) is 0.5.At high frequencies, the stop band attenuation is high, and at low frequencies, the stop band attenuation is low, so the ripples presented in stop band also easy to eliminate when compared to Hamming and Kaiser.

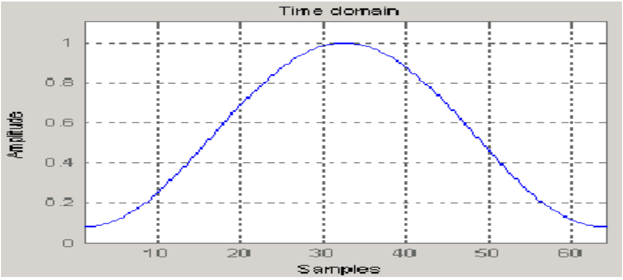

Fig. 2: Shape of Hamming Window

Hanning Window

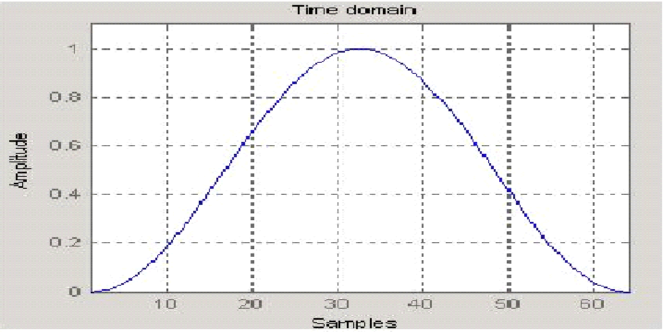

Fig. 3: Shape of Hanning Window

$W_{hn}(n)=0.5+0.5\cos(2\Pi n/N-1)$,
$-(N-1)/2<=n<=(N-1)/2$
$= 0$, Otherwise.

Kaiser Window

For Kaiser window the attenuation coefficient 0, 5.4414,8.885.

- At α=0,the output =1,kaiser becomes rectangular window.
- At α=5.4414,kaiser becomes hamming window.
- At α=8.885, kaiser becomes Blackmann window.

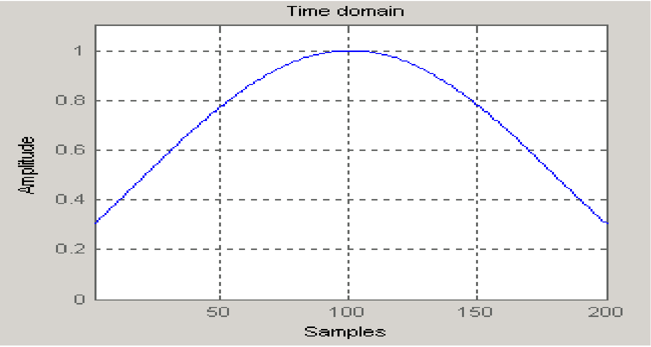

Fig. 4: Shape of Kaiser Window

Fourier Transform

Let a windowed speech signal and noise signal be represented by *s(k)* and *n(k)* respectively. The sum of the two is then denoted by *x(k)*,

$$x(k) = s(k) + n(k) \qquad (1)$$

Taking the Fourier Transform of both sides gives

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega}) \qquad (2)$$

Where

$$\mathrm{x(k)} \longleftrightarrow X(e^{j\omega})$$

$$X(e^{j\omega}) = \sum_{K=0}^{L-1} e^{-j\omega k} \qquad (3)$$

Compute Noise Spectrum Magnitude

To obtain the estimate of the noise spectrum the magnitude $N(e^{j\omega})$ of $N(e^{j\omega})$ is replaced by its average value $\mu(e^{j\omega})$ taken during the regions estimated as "noise only". For this analysis the first 50ms were used as the "noise-only".

The phase $\theta N(e^{j\omega})$ of $N(e^{j\omega})$ is replaced by the phase $\theta x(e^{j\omega})$ Of $X(e^{j\omega})$, due to the fact that the two signals are assumed to have the delay.

Through manipulation and substitution of equation (2) we obtain the spectral subtraction estimator $\hat{S}(e^{j\omega})$:

$$\hat{S}(e^{j\omega}) = [\ X(e^{j\omega})\ -\mu(e^{j\omega})]e^{j\theta x\,(e^{j\omega})} \qquad (4)$$

The error that results from this estimator is given by

$$\varepsilon(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = N(e^{j\omega} - \mu(e^{j\omega})e^{j\theta}) \qquad (5)$$

In efforts to reduce this error local averaging is used because $\varepsilon(e^{j\omega})$ is simply the difference between $N(e^{j\omega})$ and its mean μ. Therefore $X(e^{j\omega})$ is replaced with $|\overline{X(e^{j\omega})}|$

Where

$$|\overline{X(e^{j\omega})}| = \sum_{i=0}^{M-1} X_i(e^{j\omega})$$

$X_i(e^{j\omega})$ = ith time-windowed transform of *x(k).*

By substitution in equation (4) we have

$$\hat{S}_A(e^{j\omega}) = [X(e^{j\omega}) - \mu(e^{j\omega})]e^{j\theta x\,(e^{j\omega})} \qquad (6)$$

The spectral error is now approximately

$$\epsilon(e^{j\omega}) = \hat{S}_A(e^{j\omega}) - \hat{S}(e^{j\omega}) = |\overline{N(e^{j\omega})}| - \mu(e^{j\omega}) \qquad (7)$$

Where

$$|\overline{N(e^{j\omega})}| = \frac{1}{M}\sum_{i=0}^{M-1} N_i(e^{j\omega})$$

Thus, the sample mean of $N\ (e^{j\omega})$ will converge to $\mu(e^{j\omega})$ as a longer average is taken.

It has also been noted that averaging over more than three half-overlapped frames, will weaken intelligibility. The reason for this is because the noise magnitude estimate has been assumed to stay constant throughout and by underestimating we take less risk of removing any important speech information.

Half-Wave Rectification

For frequencies where $|\overline{X(e^{j\omega})}|$ is less than $\mu(e^{j\omega})$, the estimator $\hat{S}(e^{j\omega})$ will become negative, therefore the output at these frequencies is set to zero. This is half-wave rectification.

The advantage of half-wave rectification is that the noise floor is reduced by $\mu(e^{j\omega})$. When the speech plus the noise is less than $\mu(e^{j\omega})$ this leads to an incorrect removal of speech information and a possible decrease in intelligibility.

Residual Noise Reduction

While half-wave rectification zeros out the speech plus noise that is less than $\mu(e^{j\omega})$,speech plus noise above $\mu(e^{j\omega})$ still remain. When there is no speech present in a given signal the difference between N and $\mu e^{j\theta n}$ is called noise residual and will demonstrate itself as disorderly spaced narrow bands of magnitude spikes. Once the signal is transformed back into the time domain, these disorderly spaced narrow bands of magnitude spikes will sound like the sum of tone generators with random frequencies.

This is a phenomenon known as the .musical noise effect. Because the magnitude spikes fluctuate from frame to frame, we are able to reduce the audible effects of the noise residual by replacing the current values from each frame with the minimum values chosen from the adjacent frames.

The motivation behind this replacement scheme is threefold: first, if the amplitude of $\hat{S}(e^{j\omega})$ lies below the maximum noise residual, and it varies radically from analysis frame to frame, then there is a high probability that the spectrum at that frequency is due to noise; therefore, suppress it by taking the minimum value; second if $\hat{S}(e^{j\omega})$ lies below the maximum but has a nearly constant value, there is a high probability that the spectrum at that frequency is due to low energy speech; therefore, taking the minimum will retain the information; and third, if $\hat{S}(e^{j\omega})$ is greater than the maximum, there is speech present a that that frequency; therefore, removing the bias is sufficient.

Residual Noise Reduction is implemented as:

$$|\hat{S}_i(e^{j\omega})| = |\hat{S}_j(e^{j\omega})|\ for|\hat{S}_i(e^{j\omega})| \geq \max(|N_R(e^{j\omega})|)$$

Attenuate Signal during Non-Speech Activity

The amount of energy in $\hat{S}(e^{j\omega})$ compared to $\mu(e^{j\omega})$ supplies an indication of the presence of speech activity contained inside a given analysis frame. Empirically, it was determined that the average (before versus after) power ratio was down at least 12dB .

This offered an estimate for detecting the absence of speech given by:

$$T = 10\ log_{10}\left[\frac{1}{2\mu}\int_{-\pi}^{\pi}\left|\frac{\hat{S}(e^{j\omega})}{\mu(e^{j\omega})}\right| d\omega\right]$$

If T was less than -12dB for a particular frame, it was classified as having no speech and attenuated by a factor *c*, where

$20\ log_{10}C = -30dB$. -30dB, was found to be a reasonable, but not optimum amount of attenuation.

The output of the spectral estimate including signal attenuation is given by:

$$\hat{S}(e^{j\omega}) = \begin{cases} \hat{S}(e^{j\omega}) \quad T \geq -12db \\ CX(e^{j\omega})T \leq \ -12db \end{cases}$$

III. RESULTS

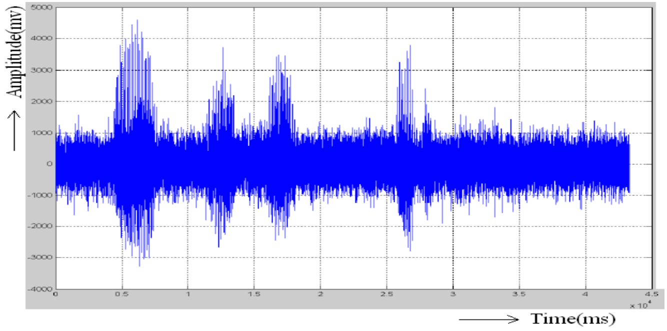

Fig. 5: Time waveform of speech utterance .

This is the input signal which is ready for enhancing its frequency range is typically 300Hzs to 4000 Hzs which is in between audio range. It has 5 speech contents with 50ms time duration. For this signal time is taken in (ms) across X-axis and amplitude is taken in (mv) across Y-axis.

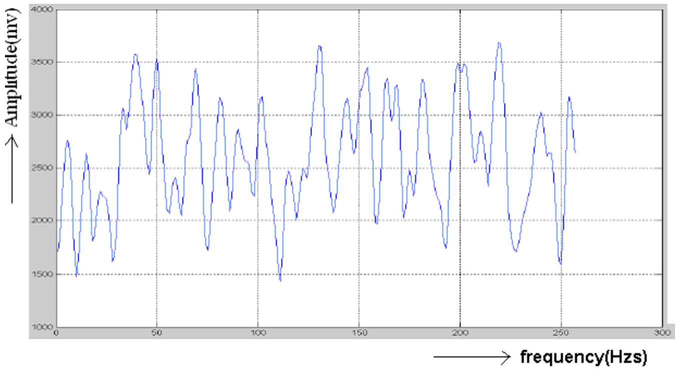

Fig. 6: Average noise magnitude of speech utterance

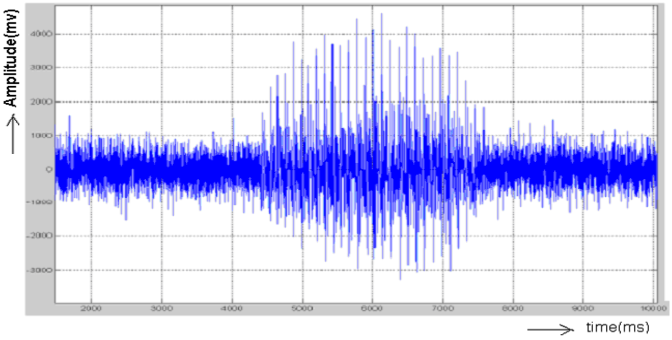

Fig. 7: Time waveform of speech utterance for single noisy speech content.

This is the single noisy speech content which is taken from the input noisy signal for enhancing. Before applying the total input signal to the spectral subtraction ,it is the example application of enhancing process. The total duration of this speech content is 10ms.

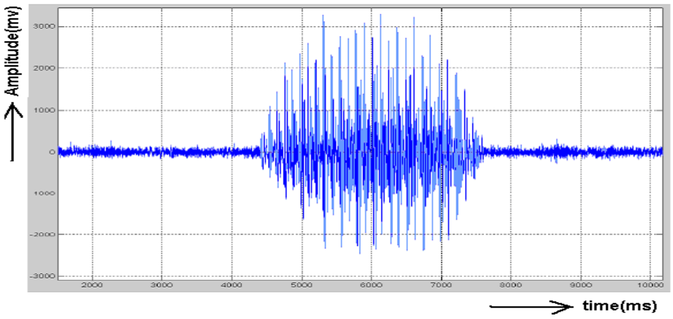

Fig. 8: Time waveform of Enhanced speech content.

This is the enhanced single speech content &is the output of Spectral Subtraction when we apply the input for SS as the single speech content. For this signal time is taken in ms across X-axis and amplitude is taken in mv across Y-axis & the duration of this signal is 10ms.

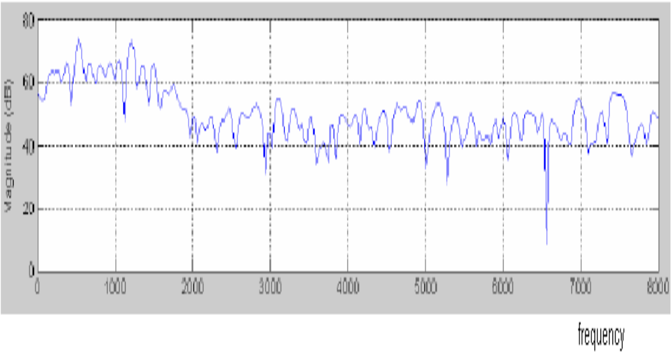

Fig. 9: Spectrum of Noisy speech signal $\left|\overline{X(e^{j\omega})}\right|$

Spectrum is the waveform of the signal magnitude with respect to frequency. The above signal is the spectrum of input noisy signal, it was obtained after Fourier Transform. For this signal frequency is taken in Hzs across X-axis and magnitude is taken in dBs across Y-axis.

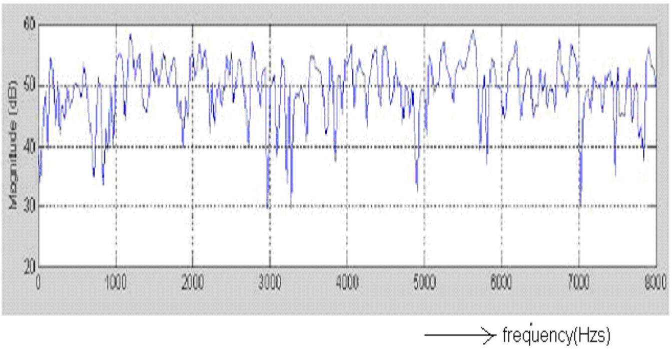

Fig. 10: Avg. noise magnitude μ($e^{j\omega}$)

This is the spectrum of spectral estimator which is estimated from noisy speech signal. . For this signal frequency is taken in Hzs across X-axis and magnitude is taken in dBs across Y-axis.

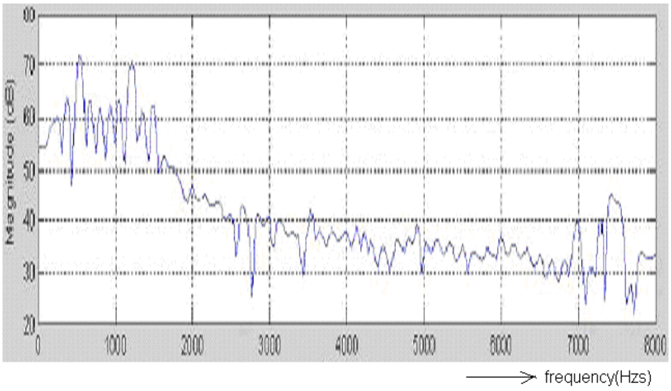

Fig. 11: Spectrum of Enhanced speech $\hat{S}(e^{j\omega})$

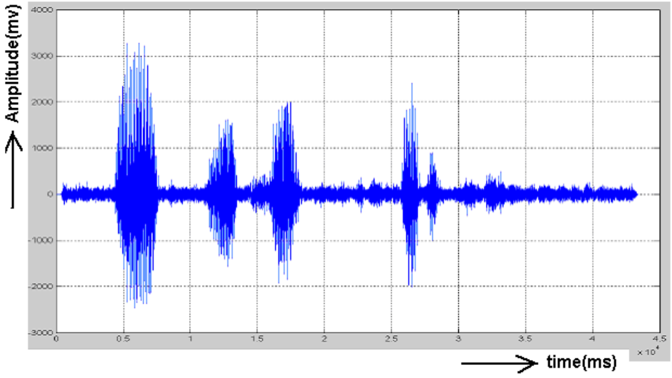

Fig. 12: Time wave form of speech utterance after SS.

The above signal is the speech utterance after bais removal, half wave rectification, frame averaging ,residual noise reduction , non-speech activity and signal reconstruction. For this signal time is taken in ms across X-axis and amplitude is taken in mv across Y-axis.

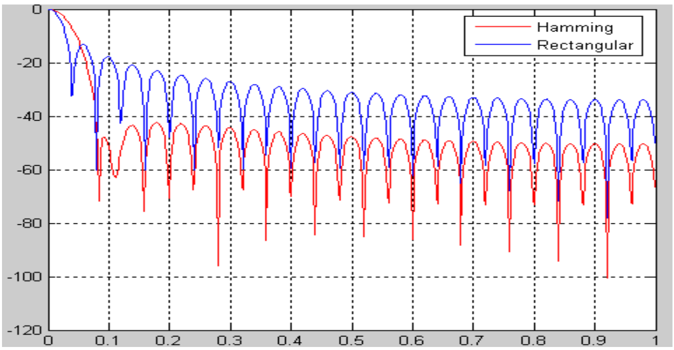

Fig. 13: waveforms for hamming and rectangular window techniques

The enhanced speech signals were played back and demonstrated considerable improvement from the original signals. Some tribulations encountered during this implementation were discovered during the speech activity detector step, the algorithm only detected the first five and last two frames as having no speech, all other frames were found to contain speech information. This is an extremely low number of frames to be classified as no speech and was quite unexpected. In addition, due to randomly spaced narrow bands of noise residual, the final results exhibited the phenomenon known as the musical noise effect.

IV. COCLUSION

SS, a noise removal algorithm has been successfully implemented and tested. Sufficient estimates of noise spectra were determined from initially noisy speech signals and effectively removed throughout the signal to produce an enhanced speech signal. Overall the results display a considerable improvement in the quality of speech signals, which should increase the performance in ASR recognition systems.

REFERENCES

[1]. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (2), 113–120.

[2]. Study and the development of the INTEL technique for improving speech intelligibility. Technical Report NSC-FR/4023, Nicolet Scientific Corporation.

[3]. Speech Enhancement: Theory and Practice. CRC Press LLC, Boca Raton, FL.

[4]. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, pp. 208–211.

[5]. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

[6]. Experiments with a Non-linear Spectral Subtractor (NSS) Hidden Markov Models and the projections for robust recognition in cars. Speech Commun 11 (2–3), 215–228.

[7]. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. 7 (3), 126–137.

[8]. Improving performance of spectral subtraction in speech recognition using a model for additive noise. IEEE Trans. Speech Audio Process. 6 (6), 579–582.

[9]. Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task. In: Proc. Internat. Conf. Spoken Language Processing, pp. 477–480.

[10]. An assessment on the fundamental limitations of spectral subtraction. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, Signal Processing, Vol. I. pp. 145–148.