International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 5, May 2016

Enhancing Speech Signal Using Multiband Spectral Subtraction Method

[1] Shruthi O R, [2] Jennifer C Saldanha
[1] Student, M.Tech (DECS), [2] Assistant Professor,
St. Joseph Engineering College Mangaluru, India
[1] shruthior@gmail.com, [2] jennifers@sjec.ac.in

***Abstract*—Aim of speech enhancement is to improve the intelligibility and quality of the speech. By recording the speech signal in the noisy environment, clean speech signals are degraded. Speech enhancement reduces the noise without distorting the original (clean) signal. In this concept, Multi-Band Spectral Subtraction (MBSS) method is used to enhance the noisy speech signal. This approach, takes into account the fact that colored noise affects the speech spectrum differently at various frequencies. Then the properties of this method are derived in terms of input and output SNR.**

***Index Terms*— Intelligibility, MBSS, Signal to Noise Ratio (SNR).**

I. INTRODUCTION

Speech enhancement aims to improve speech quality by using various algorithms. It may sound simple, but it is meant by the word quality. It can be at least clarity and intelligibility, pleasantness, or compatibility with some other method in speech processing. Intelligibility and pleasantness are difficult to measure by any mathematical algorithm. The central methods for enhancing speech are the removal of background noise, echo suppression and the process of artificially bringing certain frequencies into the speech signal. First of all, every speech measurement performed in a natural environment contains some amount of echo.

In the current telephone networks speech is band limited between 300–3400 Hz. Artificial bandwidth expansion can be utilized to restore the frequencies that disappear on the route. These methods are also useful in speech compression. When the background noise is suppressed, it is crucial not to harm or garble the speech signal. The delay must be kept very small to avoid producing more noise instead of cancelling the existing noise. The operation of all the speech enhancement methods in the following sections is based on the spectra calculated from adjacent frames of speech. In practice, the frames are a little bit overlapping and the frame size is a couple of dozens of milliseconds. The windowed speech frame is padded with zeros to make its length equal to the nearest power of two. In modern hands free speech communication environments often occurs the situation that the speech signal is superposed by background noise. This is particular the case if the speaker is not located as close as possible to the microphone. The speech signal intensity decreases with growing distance to the microphone. It is even possible that background noise sources are captured at a higher level than the speech signal. The noise distorts the speech and words are hardly intelligible. In order to improve the intelligibility and reduce the listeners (FES) stress by increasing the signal to noise ratio a noise reduction procedure also called speech enhancement algorithm is applied.

II. DESCRIPTION

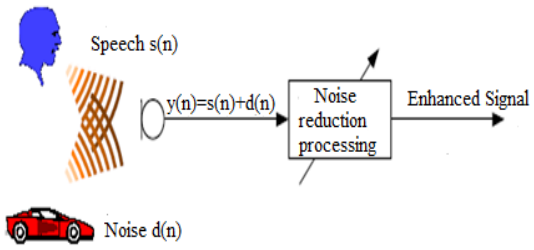

Figure 1: Overview of Speech enhancement technique.

Spectral subtraction is a method for restoration of the power or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. It is the most common of the subtractive type algorithms, which form a family of methods based on subtraction of the noise estimate from the original speech. These systems form a category of algorithms that operate in the frequency domain. The noise spectrum is estimated, and updated, from the periods when the signal is absent and only the noise is present. The assumption being that noise is stationary or a slowly varying process, and that the noise spectrum does not change significantly between the

updating periods. For restoration of the time-domain signal, an estimate of the instantaneous magnitude spectrum is combined with the phase of the noisy signal, and then transformed via a inverse discrete Fourier transform to the time domain. The phase of the noisy signal is not modified, as not only is it hard to get an estimation of the phase as compared to the magnitude spectrum, it is also believed that from perceptual point of view the phase does not carry any useful information for noise suppression. Thus, assume that y(n) the discrete noise corrupted input signal is composed of the clean speech signal s(n) and d(n) the uncorrelated additive noise signal, then it the noisy signal can be represented as:

$$y(n)=s(n)+d(n) \quad (1)$$

Where y(n), s(n) and d(n) are the corrupted speech signal, clean speech signal and the noise respectively. The power Where y(n), s(n) and d(n) are the corrupted speech signal, clean speech signal and the noise respectively. The power spectrum of the corrupted speech can be approximately estimated as:

$$|Y(k)|^2=|S(k)|^2+|D(k)|^2 \quad (2)$$

Where S(k) and D(k) are the magnitude spectra of the clean speech and the noise respectively. Since the noise spectrum cannot be directly obtained, an estimate $\hat{D}(k)$ is calculated during periods of silence. The estimate of the clean speech spectrum is obtained as:

$$|\hat{S}(k)|^2 = |Y(k)|^2 - \alpha|\hat{D}(k)|^2 \quad (3)$$

Where α is an over-subtraction factor, this is a function of the segmental SNR. This implementation assumes that the noise affects the speech spectrum uniformly and the over-subtraction factor α subtracts an over-estimate of the noise over the whole spectrum. That is not the case, however, with real-world noise (e.g., car noise, cafeteria noise, etc.).

To take into account, fact that colored noise affects the speech spectrum differently at various frequencies; a multi-band approach is used in spectral subtraction. The speech spectrum is divided into N non-overlapping bands, and spectral subtraction is performed independently in each band. So, the estimate of the clean speech spectrum in the i^{th} band is obtained by:

$$|\hat{S}_i(k)|^2 = |Y_i(k)|^2 - \alpha_i \delta_i |\hat{D}_i(k)|^2 \quad , \quad b_i \le k \le e_i \quad (4)$$

Where b_i and e_i are the beginning and ending frequency bins of the i^{th} frequency band, α_i is the over-subtraction factor of the i^{th} band and δ_i is a tweaking factor that can be individually set for each frequency band to customize the noise removal properties. The band specific over subtraction factor α_i is a function of the segmental SNR_i of the i^{th} frequency band which is calculated as:

$$SNR_i(dB) = 10 \log_{10}\left(\frac{\sum_{k=bi}^{ei}|Y_i(k)|^2}{\sum_{k=bi}^{ei}|\hat{D}_i(k)|^2}\right) \quad (5)$$

Using the SNR_i value, α_i can be determined as:

$$\alpha_i=\begin{cases} 5, & SNR_i < -5 \\ 4-\frac{3}{20}(SNR_i), & -5 \le SNR_i \le 20 \\ 1, & SNR_i > 20 \end{cases} \quad (6)$$

While the use of the over-subtraction factor α_i provides a degree of control over the noise subtraction level in each band, the use of multiple frequency bands and the use of the δ_i weights provide an additional degree of control within each band. The negative values in the enhanced spectrum were floored to the noisy spectrum as:

$$|\hat{S}_i(k)|^2=\begin{cases} |\hat{S}_i(k)|^2 & |\hat{S}_i(k)|^2 > 0 \\ \beta|\hat{Y}_i(k)|^2 & \text{else} \end{cases} \quad (7)$$

Where the spectral floor parameter is set to $\beta = 0.002$.

III. BLOCK DIAGRAM

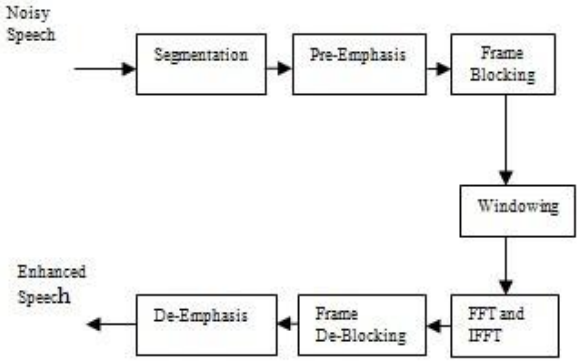

Figure 2: Block Diagram.

Segmentation

There is a need for segmentation while processing a speech signal in the frequency domain. Any speech signal is a non stationary random process; hence it is not Fourier transformable. For a Non-Stationary Random Process the auto-correlation between its 2 sample is given by $R_{xx}(\tau, t_{ij})$. It is not only function of time elapse (τ) between the sampling instances but also a function of t_{ij}. Hence it is non deterministic function, since the autocorrelation is random in nature. Fourier transform $R_{xx}(\tau, t_{ij})$, which is spectral

density of x(t) given by $S_{xx}(f, t_{ij})$ is also random in nature and hence highly non deterministic. Fourier transform of $R_{xx}(\tau, t_{ij})$ is given by:

$$\int_{-\infty}^{\infty} R_{xx}(\tau, t_{ij}) e^{-j2\pi f\tau} d\tau \quad (8)$$

The spectral density $S_{xx}(f)$ is given by $|X(f)|^2$, where X(f) is a spectral content of x(t). Interestingly a speech signal behaves as a wide sense stationary random process within a duration of 35msec maximum. There is need for segmenting a speech signal of duration, say 10msec. Before, the spectral estimation is done by FFT or otherwise in order to perform frequency domain based speech signal processing. Let us consider segmentation duration as 10 msec with f_s as 8 KHz, there will be 80 samples in each segment of a speech signal.

B. Pre-Emphasis

The high frequency components in any section of the spectrum will have less energy compared to the low frequency components. Certain applications like speech recognition and others need the features of high frequency components. Hence to extract the features at high frequency region, need to energize the high frequency components, by pre-emphasizing. The process of pre-emphasizing also reduces the noise in the high frequency region especially when the noise is stationary. The process of pre-emphasizing is given by

$$y(n) = x(n) + \alpha\, x(n-1) \quad (9)$$

Where x (n) is the present input sample to the pre-emphasize, y(n) is the pre-emphasized sample, α is the pre-emphasizing factor.

C. Frame Blocking

The input signal is windowed using a smooth trapezoid window in which the first D samples of the input frame buffer d(m) are over lapped form the last D samples of the previous frame. This overlap is described as;

$$D(m,n) = d(m-1, L+n);\ 0 \leq n < D \quad (10)$$

Where m is the current frame, n is the sample index to the buffer d(m),L-80 is the frame length, and D=24 is the overlap in samples. This results in the input buffer containing L+D-104 samples in which last D samples are the pre emphasized over lap form the previous frame, and the following L samples are pre emphasized input from the current frame.

D. Windowing

The speech signal is segmented and sent to the FFT block; effectively a rectangular window is applied on the speech signal. Segmenting a speech with a rectangular window results in convolution between spectral content of the speech signal and sinc function.

$$s(n) * r(n) \quad \text{(Time Domain)} \quad (11)$$

$$S(W) \times R(W) \quad \text{(Frequency Domain)} \quad (12)$$

This convolution is equivalent to introducing certain distortion in the spectral content of a signal called spectral leakage. The spectral leakage distortion is due to the high amplitude side lobes of the sinc function. The spectral leakage distortion can be contained by the use of a non-rectangular window like hanning, hamming, barlet etc. The course of the spectral leakage distortion is mainly due to the abrupt edges of the rectangular window. Therefore any window which is symmetrical in nature and having smooth variation at the edges is preferred in the place of rectangular window. Then FFT and IFFT will takes place.

E. Frame De-Blocking

In the frame De-Blocking, the first 48 samples out of 128 samples of IFFT outputs are algebraically added with the last 48 samples of the previous frame. The next 32 samples of the current frame are joined (concatenated) without any processing. The last 48 samples of the current frame are given for the next frame, so out of 128 samples of the every frame output of the IFFT's 80 samples is obtained with the above procedure.

F. De-Emphasis

This is the reverse process of Pre-Emphasis.

IV. EXPERIMENTAL RESULTS

A. MATLAB Result

The speech signal with fan noise, i.e "hello hello +fan noise" of 3sec duration is taken as the input signal.

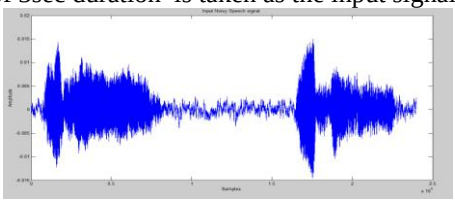

Figure 3: Input Noisy speech signal.

The result of multiband spectral subtraction method is given in figure 2.

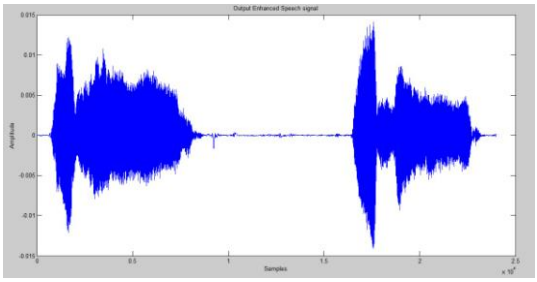

Figure 4: Enhanced speech signal.

Another example with fan noise of 6sec duration is given in figure 5.

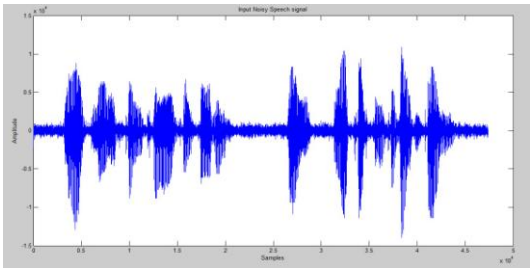

Figure 5: Input Noisy speech signal of 6sec duration.

Enhanced signal of figure 5 is given below in figure 6.

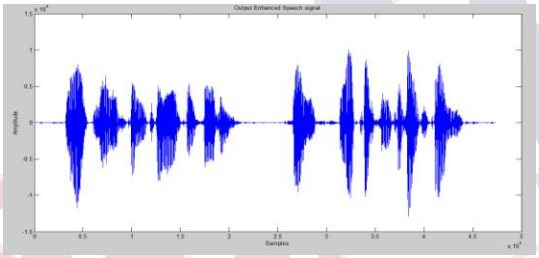

Figure 6: Enhanced speech signal of 6sec duration.

B. Spectrogram Analysis

The spectrogram of the figure 3 is given in figure 7.

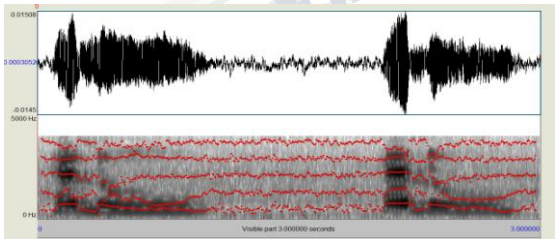

Figure 7: Spectrogram of noisy speech signal of 3sec duration.

In figure 7 red lines represents the formants of the speech Signal. Blue line represents the pitch of the utterance "hello hello".

Spectrogram of figure 4 is given below in figure 8.

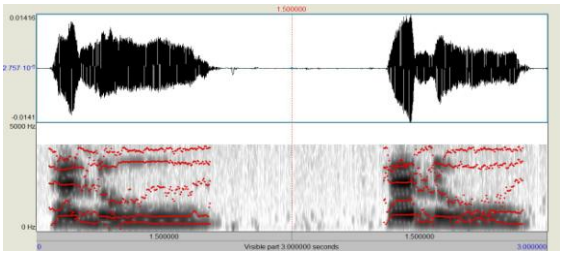

Figure 8: Spectrogram of enhanced speech signal of 3sec duration.

Table 1: SNR Result

Method	Input SNR(dB)	Output SNR(dB)	Improvement in SNR (dB)
MBSS	103.44	105.17	1.72

Table 2: Spectrogram Analysis

Method	Formants(Hz)		Pitch(Hz)		Intensity(dB)	
			Min	Max	Min.	Max.
MBSS	F1	649	99.56	488.4	3.42	34.86
	F2	1441.40				
	F3	2316.46				
	F4	3207.49				

V. CONCLUSION

In this paper, concept of Multi-Band Spectral Subtraction method is used to enhance the noisy speech signal. Here 3sec and 6sec long duration speech signal sampled at 8 KHz is used. Experimental result shows that, SNR of the enhanced speech signal is improved by 1.72 dB with respect to input SNR.

REFERENCES

[1] B Shahnaz, Fattah S A, "Speech Enhancement Based on a Modified Spectral Subtraction Method," IEEE 57th International Midwest Symposium on Circuits and Systems, Aug.3-6, 2014, pp. 1085–1088

[2] K. Wu and P. Chen, "Efficient speech enhancement using spectral subtraction for car hands-free application", International Conference on Consumer Electronics, vol.2,pp.220-221,2001.

[3] Sunil D. Kamath, Philipos C. Loizou, "A Multiband Spectral Subtraction Method for Enhancing Speech

Corrupted by Coloured Noise", University of Texas at Dallas.

[4] Lalchhandami, Rajat Gupta " Different Approaches of Spectral Subtraction Method for Speech Enhancement",International Journal of Mathematical Sciences, Technology and Humanities 95 (2013) 1056 – 1062.

[5] Suraj P. Patil , Uday P Mithapelli," Speech Enhancement in Terms of Intelligibility Using Modified Multiband Spectral Subtraction", International Journal of Emerging Engineering Research and Technology Volume 2, Issue 2, May 2014, PP 96-100.