A Study of Multiband Spectral Subtraction for Speech Enhancement in Magnitude and Power Spectral Domains

[1]Naik D C, [2]Sreenivasa Murthy A

[1]Research Scholar, [2]Professor
[12]Department of Electronics and Communication Engineering
[12]University Visvesvaraya College of Engineering, Bengaluru, India

Abstract: Over the last 50 years or so researches/engineers have proposed a number of speech enhancement algorithms. The goal of speech enhancement algorithm is to obtain the clean speech to a level that is, appreciated by human listeners be impaired or normal and that improves the performance of speech-activated machines. Spectral subtraction is probably the first approach proposed in this direction. It is so attractive in terms of both comprehension and implementation that even after 40 years of its existence, researches looking at modifications to the basic approach to get a better signal in the sense of improved intelligibility and quality. In most spectral subtraction based speech enhancement approach, the estimated noise spectrum is assumed to affect the entire short time speech spectrum uniformly. However for most of the real world noise scenario, these assumptions turn out to be invalid to alleviate this problem, the spectrum of each frame is divided into a number of sub-bands and weighted noise estimates are subtracted from each of these sub-bands. The noise estimates are obtained from the sub-band specific SNRs. The multi-band spectral subtraction alleviates this problem by subtracting the weighted noise estimates from the non-overlapping bands of the noisy speech spectrum. In this research activity, we have used multi-band spectral subtraction in magnitude spectral domain and tested for babble noise, random noise, car noise and helicopter noise using both subjective and objective measures. The result is compared with those obtained by power spectral domain. It is observed that the method results in reduced background noise and improved speech quality for babble noise and random noise.

Index Terms - **spectral subtraction, multi-band spectral subtraction, subjective and objective performance measures, spectrograms.**

I. Introduction

One of the very easy to understand and simple to implement a method to reduce the additive white background noise is spectral subtraction [1-2]. As the white Gaussian noise has a categorical spectrum, the noise impacts speech spectrum uniformly. However, the spectrum of naturally occurring noise is not categorical. For example, in the babble noise low-frequency region comprises more energy than the high- frequency region. It is difficult to estimate the necessary amount of noise spectrum to be removed from different frequency bands while retaining excepted properties of speech spectrum [6][10]. In the method, proposed by Kamath et al, [6] noise is estimated from the periods of speech absence and in the method proposed by Upadhyay [10] noise is estimated by decision direct approach. In these two methods the noisy speech spectrum is split up into non-overlapping bands and the different amount of noise to be subtracted from each band is calculated based on sub-band SNR.

The method proposed in [1] implements the above method in the power spectral domain in a different manner, considerably reduces the background noise from the noisy speech. However, these methods bring an annoying distortion in the enhanced speech signal called musical noise. This is because of inaccurate noise estimate and also due to flooring negative values to zero or to some empirical threshold value.

In this paper, we discuss a multiband spectral subtraction method implemented in magnitude and power domain to understand and compare their performances. We show that the magnitude domain multiband spectral subtraction method reduces more background noise while maintaining considerable speech quality and intelligibility for random noise and babble noise.

Algorithms were implemented by framing the sentences using 20-ms duration hamming windowed 50% overlap between adjacent frames. Section II presents the multiband spectral subtraction method in power and magnitude domain [13], section III presents the experimental results, followed by conclusion and references.

II. Multi-band spectral subtraction

Assuming the additive model, we can express noisy speech signal in terms of a clean speech signal and noise signal as follows:

$$y(n) = s(n) + n(n) \qquad (1)$$

Where s(n), n(n) and y(n) are the clean speech signal, noise signal, and noisy speech signal respectively. Taking the DFT on both sides

$$Y(k) = S(k) + N(k) \qquad (2)$$

Where Y (k), S(k) & N(k) can be written in polar form as $Y(k) = |Y(k)| exp\left(j\theta_y(k)\right)$,$S(k) = |S(k)| exp\left(j\theta_y(k)\right)$, $N(k) = |N(k)| exp\left(j\theta_n(k)\right)$ respectively.

Since the short time phase spectrum is not important for speech enhancement. We write equation 2 as

$$|Y(k)| = |S(k)| + |N(k)| \qquad (3)$$

The power spectrum of the noisy speech signal can be obtained as

$$Y(k).Y(k)^* = \left(S(k) + N(k)\right).\left(S(k)^* + N(k)^*\right)$$

$$|Y(k)|^2 = |S(k)|^2 + |N(k)|^2 + 2Re\{S(k)N(k)\} \qquad (4)$$

We will take the expectation operator on both sides, since the terms $|N(k)|^2$ and $S(k).N(k)$ cannot be estimated directly and are approximated as, $E\{|N(k)|^2\}$ and $E\{S(k).N(k)\}$, where $E\{.\}$ the expectation operator with zero mean and uncorrelated, the term is $E\{S(k).N(k)\}$ becomes zero. Therefore the estimated speech signal written as

$$\left|\hat{S}(k)\right|^2 = |Y(k)|^2 - \left|\widehat{N}(k)\right|^2 \qquad (5)$$

$\left|\widehat{N}(k)\right|^2$ is approximated as average values from the non-speech activity frames of a noisy speech signal.

In general, the spectral subtraction algorithm can be directly obtained by altering the power spectrum from a variable p

$$\left|\hat{S}(k)\right|^p = |Y(k)|^p - \left|\widehat{N}(k)\right|^p \qquad (6)$$

Where $p = 2$ correspond to the power spectrum domain and $p = 1$ correspond to the magnitude spectrum domain.

By Berouti et.al. [1] Implementation scenario, at some spectral components $\widehat{N}(k)$ may be smaller than $N(k)$. Hence $\widehat{N}(k)$ is multiplied with α (over-subtraction factor), therefore the estimation of the clean speech spectrum is as follows:

$$\left|\hat{S}(k)\right|^2 = |Y(k)|^2 - \alpha\left|\widehat{N}(k)\right|^2 \qquad (7)$$

Suppose, if right hand side becomes negative that spectral component is replaced by $\beta\,|Y(k)|^2$, where β is called spectral floor factor. Thus α removes most of the speech component, but β reduces the amount of musical noise perceived. Experimentally it was found that at 0dB SNR 3≤α≤6 and 0.05≤β≤0.1 gave good results.

Where α can be termed as subtraction factor, which is a function of the segmental SNR which is greater than 1. This scenario assumes that the noise impact on the speech spectrum uniformly over the whole spectrum and the subtraction factor subtracts an overestimate of the noise from the whole spectrum. In order to reduce the speech distortion, we decided to set the value of α varying from frame to frame within the same sentence.

To take into consideration that the real world noise scenario the speech spectrum varies differently at different frequencies depending on the noise condition, Kamath.et.al proposed a multi-band approach of spectral subtraction, where the noisy speech spectrum is divided into N (preferably 3) non-overlapping bands and for each band spectral subtraction [5] is performed independently. Hence, the estimation of the clean speech spectrum in the j^{th} band is obtained as

$$|(\hat{S}_j(k)|^p = \left|Y_j(k)\right|^p - \propto_j \delta_j\left|\widehat{N}_j(k)\right|^p, sf_j \le k \le ef_j \qquad (8)$$

Where sf_j and ef_j are the starting and ending frequency indices of the j_{th} frequency band. For each frequency band, the subtraction factor $\propto_j$ and tweaking factor δ_j has to be set individually (by equation 10 and 11) to customize the noise removal properties.

The $\propto_j$ is a band specific subtraction factor, which is a function of segmental SNR_j of the j_{th} frequency band which can be calculated as:

$$SNR_j(dB) = 10log_{10}\left(\frac{\sum_{k=sf_i}^{ef_j}\left|Y_j(k)\right|^2}{\sum_{k=sf_i}^{ef_j}\left|\widehat{N}_j(k)\right|^2}\right) \qquad (9)$$

Using the SNR_j value calculated in the above equation $\propto_j$ can be fixed as:

$$\propto_j = \begin{cases} 5 & , SNR_j < -5 \\ 4 - \frac{3}{20}(SNR_j) & , -5 \le SNR_j \le 20 \\ 1 & , SNR_j > 20 \end{cases} \qquad (10)$$

To remove the musical noise [12], the value of α is as approximated by empirically determined for best noise reduction. We then decided that the value of α should not be varied with sentences with different SNR but also across frames of the same sentence. The reason for changing α within a sentence is that the segmental SNR varies from frame to frame as the noise level is invariant for random noise. By considering the real-time experiment we could conclude that the value of α should vary within a sentence.

The values for δ_j were empirically determined and set to:

$$\delta_j = \begin{cases} 1 & , f_j \le 1kHz \\ 2.5 & , 1kHz < f_j \le \frac{Fs}{2} - 2kHz \\ 1.5 & , \frac{Fs}{2} - 2kHz \end{cases} \qquad (11)$$

Where f_j is the upper frequency of the jth band, and Fs is the sampling frequency. To get control over the noise subtraction level in each band $\propto_j$ is used and δ_j is used to remove the additional noise characteristic from the frequency bands.

The negative measure in the enhanced speech spectrum in equation (8) was considered as follows:

$$|\hat{S}_j(k)|^p = \begin{cases} \left|\hat{S}_j(k)\right|^p, & \left|\hat{S}_j(k)\right|^p > 0 \\ \beta\left|Y_j(k)\right|^p, & else \end{cases} \qquad (12)$$

The need for using smaller δ_j values is to reduce speech distortion while estimating the enhanced speech spectrum. The algorithm was implemented in both magnitude and power domain.

$$\hat{S}(k) = |\hat{S}_j(k)|\, e^{j\theta_x k} \qquad (13)$$

The enhanced speech spectrum within each band is combined and for power domain, we have to take the square root of the estimated enhanced speech spectrum [14]. The enhanced signal is obtained by acquiring the inverse Fourier transform of the enhanced speech spectrum by multiplying with the phase of the noisy signal and then, the overlap-add method is applied to obtain enhanced speech signal.

III. RESULTS

In this section, we are going to discuss objective [3, 7] and subjective performance measures [9] for the comparison of algorithms proposed for magnitude and power spectral domain, from the results we conclude that the magnitude domain gives good results when compared to power domain. We have used the corpus "In the fall of 1991 he took a coaching job at a high school" of 16000 sampling frequency.

These methods are not suitable for car and helicopter noise types as they give bad results in terms of objective, subjective measures and poor enhanced speech signal in terms of quality and intelligibility.

Table I: Power Domain

Noise	I/P SNR	segSNR	fwsegSNR	dWSS
Random Noise	10	8.36	5.73	200.74
	5	7.40	5.21	205.32
	0	3.44	4.80	204.87
	-5	2.43	4.60	203.70
Babble Noise	10	11.35	7.42	110.12
	5	6.73	6.31	147.11
	0	1.79	5.02	195.27
	-5	-2.12	1.75	234.11
Car Noise	10	4.23	5.11	98.46
	5	2.29	3.22	102.23
	0	0.06	1.27	104.46
	-5	-2.12	0.99	107.33
Helicopter Noise	10	2.24	3.22	120.26
	5	2.00	2.12	122.72
	0	1.23	1.99	127.86
	-5	-3.76	0.66	140.24

Table II: Magnitude Domain

Noise	I/P SNR	segSNR	fwsegSNR	dWSS
Random Noise	10	15.44	9.56	148.91
	5	12.64	7.72	180.21
	0	10.49	6.20	195.50
	-5	6.95	4.59	211.59
Babble Noise	10	12.01	8.82	163.07
	5	8.72	7.32	211.57
	0	5.79	5.60	246.70
	-5	3.10	3.73	287.78
Car Noise	10	5.77	5.06	104.23
	5	4.23	3.16	107.76
	0	2.21	1.12	110.26
	-5	1.99	0.19	120.86
Helicopter Noise	10	2.98	4.23	103.26
	5	2.16	2.49	120.44
	0	0.98	1.22	129.17
	-5	-3.02	0.10	133.23

3.1 Objective Measures

The objective measures are computed by the mathematical equations, some of the measures are average Segmental $SNR(segSNR)$, frequency-weighted segmental SNR $(fwsegSNR)$ and weighted spectral slope $(dwss)$. The measured values are shown in table I and II. There are many other objective measures are there but why, we are considering only those measures because we can easily compare the results with other domains and these measures can give the exact comparison of results and with other implemented algorithms.

3.1.1 Average Segmental SNR

One of the widely used objective measures is the average segmental SNR. As the value of segSNR is higher the enhanced speech signal has more signal power compared to noise power, the average segmental SNR is given by,

$$segSNR = \frac{1}{M} \sum_{m=0}^{M-1} 10\, log_{10} \frac{\sum_{l=lm}^{lm+l-1} S(l)^2}{\sum_{l=lm}^{lm+l-1} [\hat{S}(l)-S(l)]^2} \qquad (14)$$

S and $\hat{S}$ are the clean and enhanced speech signal, M denotes the number of frames, l denotes frame length. For different noise types and for different input SNR values, the average segmental SNR was computed by positioning the clean and enhanced speech signals. For stationary as well as non-stationary noise types the magnitude domain achieves the best in average segmental SNRs measure.

3.1.2 Frequency-Weighted SNR Measures

The frequency-weighted segmental SNR $(fwsegSNR)$ computed using the following equation:

$$fwsegSNR = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{k} W(j,m) log_{10} \frac{|S(j,m)|^2}{\left(|S(j,m)|-|\hat{S}(j,m)|\right)^2}}{\sum_{j=1}^{k} W(j,m)} \qquad (15)$$

where M is the total number of frames in the signal, k is the number of bands, $W(j,m)$ is the weight placed on the j^{th} frequency band, $|S(j,m)|$ and $|\hat{S}(j,m)|$ are the weighted clean signal spectrum and weighted enhanced signal spectrum in the j^{th} frequency band at the m^{th} frame. We considered the magnitude spectrum of the clean signal raised to a power γ is considered to find the weighted function. i.e.

$$W(j,m) = |S(j,m)|^{\gamma} \qquad (16)$$

For maximum correlation γ is varied, in this measure γ is varies from 0.1 to 2 and obtained maximum correlation with γ = 0:2
The spectra $|S(j,m)|$ of a clean speech signal are obtained by dividing the signal bandwidth into either 25 bands or 13 bands with respect to the ears critical bands. In our performance measure, we use 25 critical bands. The weighted spectra were obtained by multiplying the fast spectra with overlapping Gaussian-shaped windows and adding up the weighted spectra within each band.

3.1.3 Weighted Spectral Slope

The $dwss$ distance measure [9] computes the weighted difference between the spectral slopes of a clean speech signal and enhanced speech signal in each frequency band. The spectral slope is found to be a difference between adjacent spectral magnitudes in decibels. The WSS measure is as calculated by the following equation

$$d_{WSS} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{k} W(j,m)\left(\left(S_c(j,m)-S_e(j,m)\right)\right)^2}{\sum_{j=1}^{k} W(j,m)} \qquad (17)$$

Where $W(j,m)$ the weights are computed as per [8], $S_c(j;m)$ is the spectral slope of a clean speech signal and $S_e(j;m)$ is the spectral slope of an enhanced speech signal. In our implementation, the number of bands was set to $k = 25$. As the average segmental SNR and frequency weighted segmental SNR values are high and lower the values of $dwss$ means enhanced speech signal is good.

3.2 Subjective Measure

This is one of the performance measure test, where we will tell the listeners to listen to the noisy and enhanced speech signals to rate the quality and intelligibility of the signals from 1 to 5. A score of 1 represents poor and 5 represents the best. Here we conducted the test with 10 listeners comprising 5 male and 5 female, among them 3 are known about the speech signals rest of them not.

Table III: Power Domain

Noise	I/P SNR	SIG	BAK	OVL
Random Noise	10	3.8	4	4
	5	3.7	3.6	3.6
	0	3	3	2.9
	-5	2.5	2.2	2.3
Babble Noise	10	3.6	3.6	3.7
	5	2.7	2.7	3
	0	2.6	1.7	1.8
	-5	1	1.2	1.1

Table IV: Magnitude Domain

Noise	I/P SNR	SIG	BAK	OVL
Random Noise	10	4.4	4.6	4.6
	5	3.6	3.8	3.7
	0	2.7	3.2	3
	-5	1.9	2.7	2.3
Babble Noise	10	4.4	3.9	4.2
	5	3.5	3.4	3.5
	0	2.6	2.9	2.8
	-5	1.6	1.5	1.5

While us conducting the experiment the listeners have to differentiate among SIG (signal distortion), BAK (background intrusiveness) and OVL (overall quality). OVL refers to what the overall quality of the speech signal, similarly, SIG and BAK refer only to signal and background noise respectively. The subjective measures for power and magnitude domain are as shown in table III and IV.

3.3 Spectrograms

Spectrogram can be described as the 3D spectral information represented on a 2D plane with the x-axis as time and y-axis as frequency and third dimension denoting intensity or gray value. Darkness represents the presence of energy and the signal strength in those tract systems for the given sound unit, these resonances are also called as formant frequencies which comprises the high energy portions in the frequency spectrum of a speech signal. As we can see from Figure 1 and 2, the background noise is almost removed and the format frequencies are clearly visible in the magnitude domain as compared to the spectrogram of the power domain, we can see that in the lower regions, the darkness is more means there is a more energy of speech signal in those regions.

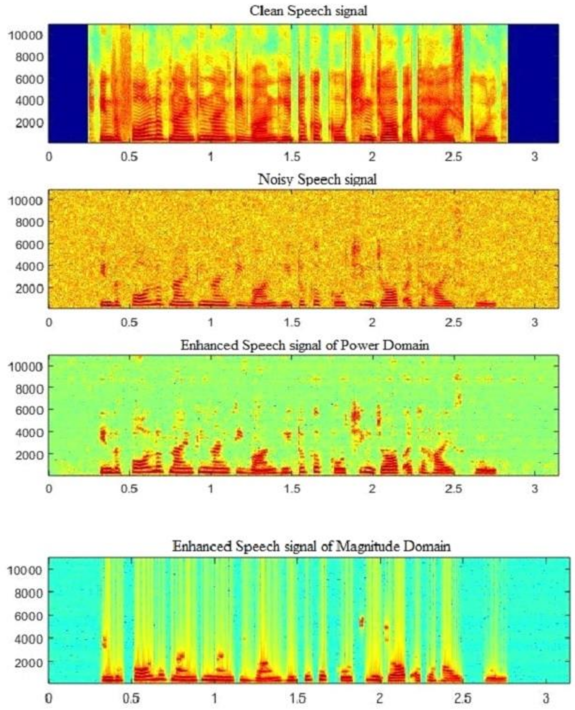

Figure 1: Spectrogram of the clean, noisy and enhanced speech signal of random noise of 5dB in power domain and magnitude domain.

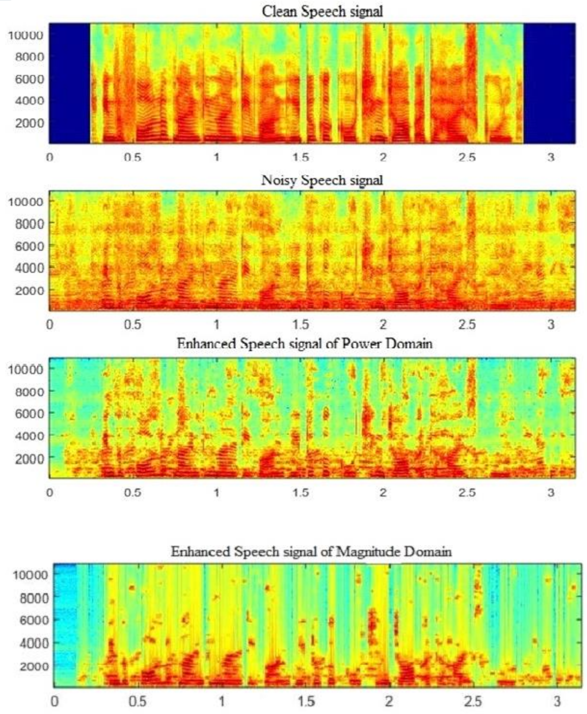

Figure 2: Spectrogram of the clean, noisy and enhanced speech signal of babble noise of 5dB in power and magnitude domain.

IV. Conclusion

The multiband spectral subtraction algorithm is implemented in the power and magnitude domain. We report the results in terms of objective measures, subjective measures & spectrograms. The values obtained for the subjective measures like SIG, BAK & OVL are higher for magnitude domain when compared to that of the power domain. Also, objective measures like segSNR, fwsegSNR & dWSS show a similar trend. The spectrograms indicate that the magnitude domain processing removes noise to a larger extent as compared to the power domain. Hence we can conclude that multiband spectral subtraction in magnitude domain performs better than the power domain.

References

[1] M.Berouti, R. Schwartz and J.Makhoul, ”Enhancement of Speech Corrupted by Acoustic NOise” Proc.IEEE Int Conf Acoust., Speech, Signal Process., pp.208-211, Apr 1979.
[2] S.Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction” IEEE Trans. Acoust, Speech, Signal Process, Vol.27, pp-113-120, Apr 1979.
[3] J.Hansen and B.Pellom,”An Effective Quality Evaluation Protocol for Speech Enhancements Algorithm” Inter.conf.on Spoken Language Processing, Vol 7, pp.2819-2822, Sydney, Australia, Dec 1998.

[4] C.He and G.Zweing,"Adaptive two-band Spectral Subtraction with Multi-window Spectral Estimation" ICASSP,vol 2,pp.793-796,1999

[5] K.Wu and P.Chan,"Efficient Speech Enhancement Using Spectral Subtraction for car Hands-free Application" International Conference on Consumer Electronics, Vol 2, pp.220-221, 2001.

[6] Sunil D Kamath and P. Loizou,"A Multi-band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise" in Proceedings Int.Conf. Acoustic, Speech, Signal Processing, Orlando, USA, May 2002

[7] Y.Hu and P.Loizou,"Subjective Comparison of Speech Enhancement Algorithms" in Proc.IEEE Int.Conf. Acoust, Speech, Signal; process, Vol 1, pp.153-156, 2006.

[8] P.Loizou Speech Enhancement, Theory and Practice.Boca Raton, FL: CRC, 2007.

[9] H.Yi and P.C.Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement" Audio, Speech and Language Processing, IEEE Transactions, vol.16,pp.229-238, Jan 2008.

[10] Navneet Upadhyay and Abhijit Karmakar, "The Spectral Subtractive type Algorithms for Enhancement of Noisy Speech: A review", Int. Journal of Research and Reviews - Signal Acquisition and Processing, Vol. 1, no. 3,pp-43-49, Sept 2011.

[11] Navneet Upadhyay and Abhijit Karmakar,"An Improved Multi-Band Spectral Subtraction Algorithm for Enhancing Speech in Various Noise Environments" International Conference on Design and Manufacturing, IConDM 2013.

[12] Siddala Vihari, Dr.A.Sreenivasa Murthy, Priyanka Soni, and Naik.D.C,"Comparision of Speech Enhancement Algorithms" in Proc.ICISP, pp.666-676, Aug 2016.

[13] Julien Basco and Eric Plourde,"Speech Enhancement Using both Spectral and Spectral Modulation Domains" 2017 IEEE 30th Canadian conference on Electrical and Computer Engineering, CCECE-2017.

[14] Naik.D.C, Dr.A.Sreenivasa Murthy and Ramesh Nuttaki," Modified Magnitude Spectral Subtraction Methods for Speech Enhancement" 2017 international conference on Electrical, Electronics, Communication, Computer, and Optimization techniques(ICEECCOT) ,pp.274-279,2017