International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Noise Removal in Speech Signal using Modified Berouti Spectral Subtraction for Emotion Recognition

May Mon Lynn[1], Chaw Su[2]

University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar

Abstract: ***In this paper, Berouti Spectral subtraction is used for reducing noise from noisy speech signals. It calculates the spectrum of the noisy speech using the combination of Fast Fourier Transform (FFT) and spectral flux and then the noise spectrum is subtracted from the noisy speech spectrum. The performance of this paper was measured by calculating the Signal to Noise Ratio (SNR). This paper proposes a new parameter for speech enhancement. The idea is to apply the spectral subtraction with spectral flux for estimating the noise more precisely. The new parameter is used for spectral subtraction in unvoiced speech frames and the existing power factor in spectral subtraction method is improved.***

Keywords: Speech Enhancement, Spectral Subtraction, Spectral Flux, Emotion Recognition

1. Introduction

Among the various trends of speech recognition, the problematic issues of recognizing emotion recognition has become popular in research area. In many speech communication systems, the quality of speech is degraded due to the background noise. The enhancement process aims to improve the speeches overall quality; to increase the speech intelligibility in order to reduce the listener fatigue, ambiguity etc depending on specific application. The enhancement system may be designed only to achieve one of these aims or several. There are many enhancement methods based on wavelets domain. Among them, selecting the best threshold for reduction of noise in the wavelet domain is the challenging subject [7,10,11,12]. Donoho [7] proposed a novel approach for noise reduction using the wavelet thresholding. In references [2,3,4] have studied spectral subtraction is also popular and simple method for speech enhancement. The main problem of the basic spectral subtraction is that it makes musical noise after subtraction. Reduction of the musical noise is also one of research fields [3,4]. In Berouti Spectral Subtraction [4], the estimated noise spectrum is subtracted from noisy speech spectrum is made by introducing over-subtraction and spectral-floor factors. The spectral floor controls the amount of remaining residual noise and over-subtraction controls the amount of perceived musical noise.

First of all, this paper is an extension of work originally presented in [14]. The speech emotion is recognized using created dataset and standard dataset are reviewed. This paper also described the significant enhanced feature with berouti spectral subtraction which is more informative for the recognition. It can recognize six types of emotion namely; angry, disgust, fear, happy, surprise and sad. The modelling techniques of this system is speaker independent and text independent. It can be seen that, the recognition rate is more accurate by using this significant feature.
This paper modify the generalized spectral subtraction method proposed by M. Berouti, R. Schwartz, and J. Makhoul [4] for reducing the noise in speech signal. The new parameter is also introduced for enhancement of input signal.

2. Power Spectral Subtraction

The assumption of original spectral subtraction is subtracting the estimated noise signal from the original signal to get the clean speech signal. This assumption is modeled by following equation:

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

Where s(n) is clean speech signal, d(n) is noise and G is the term for SNR control. We assume that the noise signal is uncorrelated:

$$r_d(\eta) = D_0\delta(\eta) \quad (2)$$

Where r_d is autocorrelation function of noise signal and D_0 is a constant [1]. Because d(n) is uncorrelated process, we can show:

$$\Gamma_y(\omega) = \Gamma_s(\omega) + \Gamma_d(\omega) \quad (3)$$

Where Γis the power spectral density (PSD). So, if we can estimate $\Gamma_d(\omega)$, we will be able to estimate the $\Gamma_s(\omega)$:

$$\hat{\Gamma}_s(\omega) = \hat{\Gamma}_y(\omega) - \hat{\Gamma}_d(\omega) \quad (4)$$

Note that noise is estimated from silence frames. Because of PSD is related to Discrete-Time Fourier Transform (DTFT) [1] as:

$$\Gamma y(\omega) = \frac{Y(\omega)Y^*(\omega)}{N^2} = \frac{|Y(\omega)|^2}{N^2} \quad (5)$$

It can conclude from (4) & (5):

$$\left|\hat{S}(\omega)\right|^2 = \left|Y(\omega)\right|^2 - \left|\hat{D}(\omega)\right|^2 \quad (6)$$

or estimating the speech signal frame, the other factor $\hat{\varphi}s(\omega)$ is estimated the phase spectrum of speech frame. The

practical applications shown by Boll [3] that the noisy phase spectrum is sufficient for the estimation of clean speech phase spectrum:

$$\hat{\varphi}_s(\omega) = \hat{\varphi}_y(\omega) \quad (7)$$

Therefore from equations (6) & (7), we can obtain the estimated speech frame:

$$\hat{S}(\omega) = \left[|Y(\omega)|^2 - |\hat{D}(\omega)|^2 \right]^{1/2} .e^{j\varphi_y(\omega)} \quad (8)$$

And a generalization for (8) we have:

$$\hat{S}(n) = IDTFT\left\{ \left[|Y(\omega)|^2 - |\hat{D}(\omega)|^{\gamma} \right]^{1/\gamma} .e^{j\varphi_y(\omega)} \right\} \quad (9)$$

The value of power exponent "γ" can be optimized by tuning from 1 to 2.

The main problem of this method is getting negative value after subtractions process. This problem is being during the estimation of the noise spectrum. Therefore this problem can be removed by two methods [1]. One is half-wave rectification:

$$|\hat{S}(\omega)|^2 = \begin{cases} |\hat{S}(\omega)|^2 & \text{if } |\hat{S}(\omega)|^2 > 0 \\ 0 & \text{else} \end{cases} \quad (10)$$

and next one is full-wave rectification:

$$|\hat{S}(\omega)|^2 = abs(|\hat{S}(\omega)|^2) \quad (11)$$

The above two methods can arise a new noise namely "musical" noise. This is becoming the main problem of spectral subtraction methods.

3. Generalized Spectral Subtraction

This section also describes the idea of applying the spectral subtraction method with spectral features and details coefficients. The first step is to apply spectral flux to the noisy signal frame, so the approximations and details coefficients are acquired. Also spectral flux is applied to the noise estimated from noisy (music) frames to acquire the estimated approximations and details coefficients of noise. In the next step, the GSS algorithm proposed by M. Berouti, R. Schwartz, and J. Makhoul [4] has been improved and the improved algorithm is applied to both of approximations and details of noisy signal in parallel. The improved algorithm is described as:

$$|\hat{S}(\omega)|^{\gamma} = \begin{cases} |\hat{Y}(\omega)|^{\gamma} - \theta.\alpha|\hat{D}(\omega)|^{\gamma} \text{ if } |\hat{Y}(\omega)|^{\gamma} > (\alpha+\beta)|\hat{D}(\omega)|^{\gamma} \\ \beta|\hat{D}(\omega)|^{\gamma} \text{ else} \end{cases}$$

The modified Berouti Spectral Subtraction block diagram is shown in the following figure.

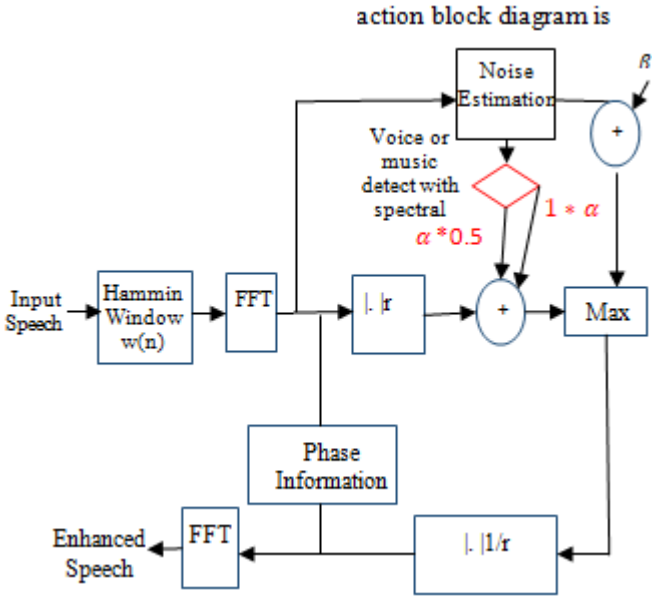

Figure 1: Block diagram of Modified Berouti Spectral Subtraction

Where $|\hat{S}(\omega)|^{\gamma}$ is the spectrum of enhanced signal, $|\hat{Y}(\omega)|^{\gamma}$ is the spectrum of noisy signal and $|\hat{D}(\omega)|^{\gamma}$ is the spectrum of estimated noise signal. The modified over-subtraction factor α is determined as:

$$\alpha = \begin{cases} \alpha_0 + 1 & SNR_i \le -5db \\ \alpha_0 - \frac{1}{5}SNR & -5db \le SNR_i \le 20db \\ \alpha_0 - 3 & SNR_i \ge 20db \end{cases} \quad (13)$$

M. Berouti, R. Schwartz, and J. Makhoul [4] recommended for α0 should be between 3 & 6. The experimental results of this paper have shown that α_0 =4 is appropriate. The segmental SNR of the i^{th} noisy signal frame (SNR_i) is calculated as:

$$SNR_i = 10\log_{10} \frac{\sum_{k=b}^{e} |Y(k)|^2}{\sum_{k=b}^{e} |\hat{D}(k)|^2} \quad (14)$$

Where b and e are the beginning and ending frequency bins of the i th noisy signal frame. We proposed a new factor θ which is determined as:

$$\theta = \begin{cases} 0.5 & \text{if the frame is music} \\ 1 & \text{else} \end{cases} \quad (15)$$

Seok [11] proposed that the algorithm for determining the voiced/unvoiced decision. The unvoiced regions where the speech is active (not silence frames) are considered as "unvoiced speech frames". Therefore a voice activity detector (VAD) is needed for it. This paper proposes the following algorithm in order to determine that the noisy frame is unvoiced speech frame:

$$\text{VADSNR} = 10\log_{10}\left(\frac{\text{noisy frame energy}}{\text{estimated noise energy}}\right) \quad (16)$$

if (the input frame is unvoiced) and (VADSNR>1)

$\theta = 0.5$ (the input is unvoice speech frame)

else

$\theta = 1$

end

The spectral-floor factor has been determined as β=0.01 through experiments. The power factor γ is determined by optimization, so we varied it from 1 to 2 for corrupted speech by white Guassian noise with global signal-to-noise ratios (SNR's) of -10db to 10db (by 5db steps). The SNR's of enhanced speech is depicted in Fig.1. Therefore, the best value was chosen γ=1.5 .

At the end of this algorithm which is applied to both approximations and details signals in parallel, the spectrum of enhanced approximations and details signals is calculated as:

$$\hat{S}(\omega) = |\hat{S}(\omega)|e^{j\hat{\varphi}(\omega)} \quad (17)$$

where φ (phase spectrum of enhanced approximations and details signals) is determined from equation (7). The final step is to apply inverse DWT to reconstruct the enhanced speech:

$$\hat{S}(n) = IDFT(\hat{S}(\omega)) \quad (18)$$

4. Experimental Results

The proposed speech enhancement algorithm has been tested on the spoken English sentence which has been chosen from SSAVEES database. The sampling frequency of all the sentences are 44.1 kHz and spoken by four male speaker. First, for optimization of power factor "γ", we varied γ from 1 to 2 for corrupted speech by white Gaussian noise with global SNR of -10db to 10db (by 5db steps). The output enhanced speech was checked by observing SNR improvement, then the best value γ =1.5 for power factor was chosen. Figures 2(a) and 2(b) show the temporal results of original and enhanced speech by the proposed method.

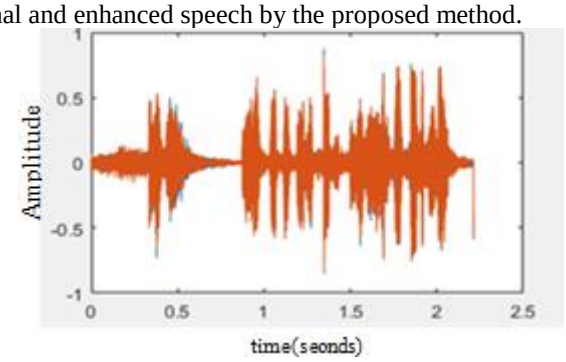

Figure 2(a): Original Signal

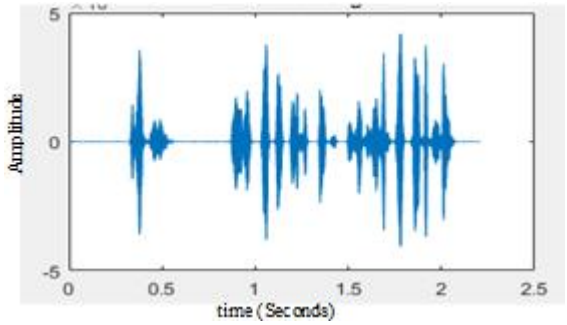

Figure 2(b): Enhanced Signal

The proposed modified Berouti spectral subtraction has been implemented to enhance the noisy speech with several SNR's on four speakers. The following table shows the enhancement performances.

Table 1: SNR comparison results for four person with different input SNR

Noisy (SNR db)	*Enhanced (SNR db)*			
	1st person	*2nd person*	*3rd person*	*4th person*
-10	4.4	3.4	3.2	3.7
-5	6.1	5.9	5.4	6.8
0	8.3	8.1	8.7	9.3
5	11.5	11.8	10.5	12.1
10	15	16.5	15.9	16.3
15	20.3	19.3	19.5	19.8
20	24.5	23.3	22.5	23.2

According to the previous analysis, the analysis results of two feature sets (Original Berouti with MFCC combine Feature and Modified Berouti with MFCC combine Feature) are shown in the following table 2 and figure 3.

Table 2: Comparison Results of Recognition Rate on Two Feature Dataset

Type of Emotion	*Recognition Rate (%) (Original Berouti with MFCC combine Feature)*	*Recognition Rate (%) (Modified Berouti with MFCC combine Feature)*
angry	75	87.5
disgust	55	80
fear	40	60
happy	40	60
sad	60	80
disgust	50	75

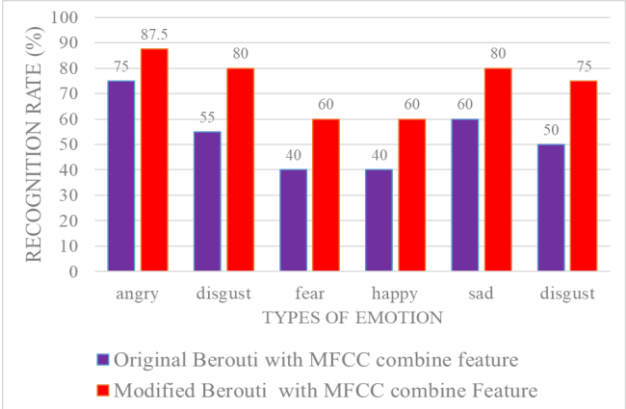

Figure 3: Classification accuracy rate on testing samples

The following table describes the comparison results of enhanced features based on created dataset with different classifiers.

Table 3: Recognition Rate with different Classifiers on the Created Dataset

Classification Method	*Overall Accuracy (%)*		
	KNN	*SRC*	*Random Forest*
Recognition Rate (%) (Coventional Berouti with MFCC combine Feature)	51.6	52	53.3
Recognition Rate (%) (Modified Berouti with MFCC combine Feature)	72	71.5	73.8

The analysis results of recognition rate for the different classifiers plot on the following figure. It also presents the comparison results of conventional Berouti and Modified Berouti method.

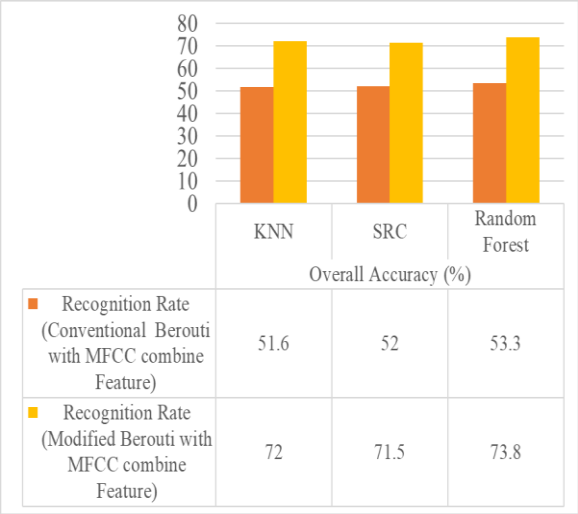

Figure 4: Recognition Rate with different Classifier on the Created Dataset

5. Conclusions

In the current works, the modified Berouti spectral subtraction algorithm is applied to spectral flux approximations and details signals in parallel. The proposed method has shown that it can enhance the noisy speech and remove the musical noise. In order to improve enhancement of unvoiced speech frames, this system proposed a new factor that showed to work better relative to old algorithm. Experimental results have shown considerable improvement in the signal-to-noise ratios.

This paper apply for the area of automatic emotion recognition from speech signals. In this system, the input signal is enhanced with the help of modified Berouti spectral subtraction and the best feature extraction are carried out with combination of MFCCs and spectral features. The analysis result is carried out with random forest classifier. The enhanced feature recognition rate is about 68.3% in previous works.

According to the analysis results, it can be seen that the extracted features after modifying the enhancement process are more efficient for the proposed system. Currently the recognition rate is raised to 73.8%.

References

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete-time processing of speech signals, 2nd edition, IEEE Press, 2000.
[2] S. Kamath, and P. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, Proceedings of ICASSP-2002, Orlando, FL, May 2002.
[3] S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. on Acoust. Speech & Signal Processing, Vol. ASSP-27, April 1979, pp. 113-120.
[4] M. Berouti, R. Schwartz, and J. Makhoul, Enhancement of speech corrupted by acoustic noise, Proc. IEEE ICASSP , Washington DC, April 1979, pp. 208-211.
[5] P. S. Whitehead, D. V. Anderson, M. A. Clements, Adaptive acoustic noise suppression for speech enhancement, Proceedings of 2003 International Conference on Multimedia and Expo (ICME '03), Vol. 1, 6-9 July 2003, pp. 565 -568.
[6] H. Sameti, H. Sheikhzadeh, Li Deng, R. L. Brennan, HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise, IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 5, September 1998.
[7] D. L. Donoho, De-noising by soft-thresholding, IEEE Transactions on Information Theory, Vol. 41, No. 3, May 1995, pp. 613-627.
[8] K. Y. Lee, B. G. Lee, S. Ann,Adaptive filtering for speech enhancement in colored noise, IEEE Trans. on Signal Processing Letters, Vol. 4, October 1997, pp. 277-279.
[9] M. Klein and P. Kabal, Signal subspace speech enhancement with perceptual post-filtering, Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Orlando, FL, May 2002, pp. I-537-I-540.
[10] Ing Yann Soon, Soo Ngee Koh, Chai Kiat Yeo, Wavelet for Speech Denoising, TENCON 97, Brisbane, Australia, 1997, pp. 479-482.
[11] J. Seok , K. Bae, Speech enhancement with reduction of noise components in the wavelet domain, Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal 2011 IEEE 7th International Colloquium on Signal Processing and its Applications Processing (ICASSP '97), Volume 2, April 21-24, 1997, pp. 1323-1326.
[12] S. Chang, Y. Kwon, S. Yang, I. Kim, Speech enhancement for non-stationary noise environment by adaptive wavelet packet, Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002), Vol. 1, 2002, pp. 561-564.
[13] Y.Ghanbari, M.R. Karami 'A new approach for speech enhancement based on adaptive thresholding of wavelet packets' Speech communication 48 (2006) 927-940

[14] May Mon Lynn, Chaw Su, Kyi Kyi Maw," Efficient Feature Extraction for Emotion Recognition System", Proceeding of 4th I2CT IEEE Conferences, SDMIT Ujire, Mangalore, India, 27th -28th , October, 2018.