Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 6, No. 3, August 2015), Pages: 87-95
www.jacr.iausari.ac.ir

Proposing a New Speech Enhancement Method based on Spectral Subtraction and Binary Masking With a Bank of 128 Gamma-tone Filters

Mozhgan Monjizade[1]✉,Saeed Ayat[2]

1) M. Sc. Student, Department of Computer Engineering and Information Technology, Payame Noor University, Iran

2) Associate Professor, Department of Computer Engineering and Information Technology, Payame Noor University, Iran

Monjizade2020@gmail.com; dr.ayat@pnu.ac.ir

Received: 2014/6/21; Accepted: 2014/8/8

Abstract

In order to the enhancement of the quality of speech corrupted by additive noise, *a speech enhancement method has been put forward based on the combination of spectral subtraction and binary masking. Spectral subtraction is a powerful method for removing noise from speech and binary masking provides essential elements to be used in monaural speech segregation. In the proposed combined method, first, spectral subtraction is used for reduction of noise in noisy speech and then binary masking is used for monaural speech segregation from musical noise introduced by spectral subtraction. The binary masking method, isolates the basic principle of target voice from other signals by using time frequency decomposition, energy masking and unites grouping. This masking is like what the human ear does in noisy environment. In our implementation for binary masking, an auditory filter (Gamma-tone) is divided into different frequency sub-bands. From these sub-band channels, channels 1,2,4,8,16,32,64,128 have been used from this bank of 128 Gamma-tone filter for implementing the binary masking. Evaluations show that the proposed combined method can improve the signal to noise ratio from 5 to 19 db for experimented signals and have better performance than binary masking or spectral subtraction in most situations, especially when noise and speech have not similar power spectrum.*

Keywords: *speech enhancement, spectral subtraction, binary masking, Gamma-tone filter bank, musical noise.*

1. Introduction

Voice signal is an important medium for information dissemination and human's most effective way of communication as well. However, voice signal in the communication process is often influenced by the surrounding environment and the transmission medium. The noise that leads in voice quality decline deteriorates voice processing system performance dramatically. For example, speaker recognition systems work well at ideal conditions or in areas with weak noise, but their performance drops rapidly when exposed to noisy environment [1].

Several methods have been developed to remove the noise, thereby improving the quality and/or intelligibility of speech signal. Some of these methods are spectral subtraction, Wiener filtering, Kalman filtering, and some others [2, 6]. For these techniques to work properly, an accurate estimate of the additive noise must be used [5]. One of the best methods is the spectral subtraction. By now, different versions of spectral subtraction have been proposed to increase its performance such as [3-5, 19]. Despite its high noise removal, it can cause an annoying noise called musical noise. Hence it can reduce overall signal quality. Musical noise is produced because the noise spectral cannot be calculated exactly, so has been used their estimations [6].

Another method is binary masking that uses noisy speech decomposition in time-frequency (T-F) units of signal, energy masking, and units grouping, like what the human ear does to describe the brain's information by nerves in noisy environments [7]. Sub-bands in the human ear are broader for greater frequencies. Therefore, in periphery processing of the signal, mixed signals can use through an auditory filter with distribution of different frequency sub-bands; a rectangle window is used to calculate its energy as a T-F unit [8]. Gamma-Tone filter has been used to create time frequency units upon estimated noise in spectral subtraction method. Speech spectrum is used when the speech value is bigger than the noise value and in other points zero or an alternate number is used. In Section 2, spectral subtraction and binary mask methods are mentioned. Section 3 describes the proposed method and Section 4 reports the conclusion.

2. Related Works

The first spectral subtraction method was described by Steven Boll [8] and the other methods are expressed in [9-10] to improve the spectral subtraction method on noise reduction. Boll subtracted an estimate of the noise spectrum from the noisy speech spectrum. The noise spectrum is obtained from the signal spectrum measured during non-speech activity. After the spectral estimator is implemented, the spectral error is computed and the other methods are introduced to reduce it.

In other hands, in the design and implementation of computation auditory systems (used to simulate the human auditory system) the method of peripheral analysis can be used. To implement this method, a computational model for peripheral auditory scene analysis should be used first and this involves time frequency units.

The Importance of separating these areas is because of using frequency analysis and specific frequencies related to properties of the human ear. These properties are used to simulate the outer and middle ears; simulation implements the frequency range through the cochlear and passing through hair cells of ear.

2.1. Spectral subtraction method

At first stage, the speech is transformed to digital form by windowing signal and overlapping. After this stage, the magnitude spectra of the windowed data are calculated and the noise spectrum bias calculated during non-speech activity is subtracted off. Then resulting negative amplitudes are replaced with another value.

The windowed noise signal $n\,(k)$ has been added to a windowed speech signal s (k), and their sum denoted by x (k). Then

$$\mathbf{x(k)} = \mathbf{s(k)} + \boldsymbol{n(k)} \quad (1)$$

Taking the Fourier transform gives

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

The magnitude $|N(e^{j\omega})|$ of $N(e^{j\omega})$ is replaced by its average value $\mu(e^{j\omega})$ taken during non-speech activity, and the phase $N(e^{j\omega})\ of\theta_N(e^{j\omega})$ is replaced by the phase $\theta_N(e^{j\omega})$ of $X(e^{j\omega})$ [3] and substitutions result in the spectral subtraction estimator $\hat{S}(e^{j\omega})$ in (3) formula:

$$\hat{S}(e^{j\omega}) = [X(e^{j\omega}) - \mu(e^{j\omega})]e^{j\theta_x(e^{j\omega})} \quad (3)$$

The error $\epsilon(e^{j\omega})$ created in calculating the estimation spectrum is given by:

$$\boldsymbol{\epsilon(e^{j\omega})} = \mathbf{\hat{S}(e^{j\omega})} - \mathbf{S(e^{j\omega})} \quad (4)$$

Some simple modifications are available to improve the auditory effects of spectral error such as magnitude averaging, residual noise reduction, half-wave rectification and additional signal attenuation during non-speech activity.

In residual noise reduction method, noise residual will have a magnitude between zero and a maximum value measured during non-speech activity. In fact, the noise can be suppressed by replacing its current value with its minimum value chosen from the adjacent analysis frames. Taking the minimum value is used only when the magnitude of $\hat{S}(e^{j\omega})$ is less than the maximum noise residual calculated during non-speech activity. The motivation behind this replacement scheme is three conditions: first, if the amplitude of $\hat{S}(e^{j\omega})$ lies below the maximum noise residual, and it varies radically in frame to frame analysis, then there is a high probability that the spectrum at that frequency is due to noise; thereby suppressing it by taking the minimum; 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 speech information; and third, if $\hat{S}(e^{j\omega})$ is greater than the maximum of power spectrum, there is speech in that frequency.The amount of noise reduction using this replacement scheme was judged equivalent to that obtained by averaging over three frames. The disadvantage to the scheme is that more storage is required to save the maximum noise residuals and the magnitude values for three adjacent frames. The residual noise reduction scheme is implemented as:

$$\begin{cases} |\mathbf{\hat{S}_i(e^{j\omega})}| = |\mathbf{\hat{S}_i(e^{j\omega})}| & \mathbf{for} |\mathbf{\hat{S}_i(e^{j\omega})}| \geq \mathbf{max}|\mathbf{N_R(e^{j\omega})}| \\ |\mathbf{\hat{S}_i(e^{j\omega})}| = \mathbf{min}\{|\mathbf{\hat{S}_i(e^{j\omega})}| \ \ \mathbf{j = i-1, i, i+1}\} & \mathbf{for} |\mathbf{\hat{S}_i(e^{j\omega})}| < \boldsymbol{max}|\mathbf{N_R(e^{j\omega})}| \end{cases} \quad (3)$$

2.2. Binary masking with Gamma-tone filter bank

Time-Frequency decomposition

To generate the time-frequency units, the domain impulse response function has been

used with the inverse correlation function peak in the words. To produce these units, a Gamma-tone filter bank has been used that calculates frequency response based on the equations in a point of time.

Gamma-tone Filter

The gamma-tone filter is widely used in models of the auditory system and is physiologically motivated to mimic the structure of peripheral auditory processing stage. The gamma-tone function is defined in time domain by its impulse response:

$$\mathbf{g(t) = \alpha t^{n-1} \exp(-2\pi bt) \cos(2\pi ft + \emptyset)} \quad (4)$$

Where,n is the order,b is the bandwidth of the filter,α is the amplitude,f is the filter center frequency, and $\emptyset$ is the phase of gamma-tone filter.

Patterson et al. (1992) show that the impulse response of the gamma-tone functions of order 4 provides an excellent fit to the human auditory filter.

The Gamma-tone filter bank takes its name from observation in the form of a carrier wave (tone) $\cos(2\pi f_0 t + \varphi)$ amplitude modulated with an envelope that is proportional to $t^{n-1}\exp(-2\pi bt)$which is the same functional form as the Gamma distribution in statistics.

$$\mathbf{ERB = 24.7(4.37.10^{-3}f + 1)} \quad (5)$$

Where f is signal center frequency rate. When the order of the filter is 4, the bandwidth of the gamma-tone filter is 1.019 ERB and the filter center frequencies are distributed across frequency in proportion to their bandwidth, known as the Equivalent Rectangular Bandwidth (ERB) scale. The ERB scale is approximately logarithmic, on which the filter center frequencies are equally spaced.

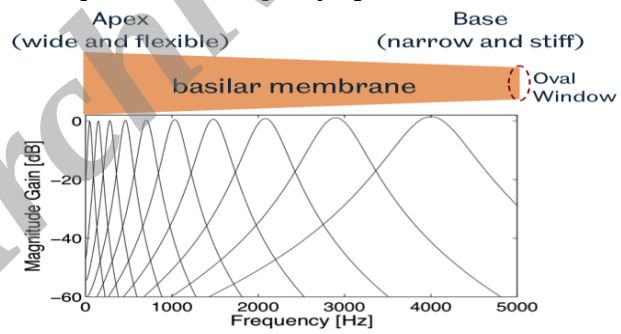

Figure1.Frequency responses of a gamma-tone filter bank with ten filters whose center frequencies are equally spaced between 50 Hz and 4 kHz on the ERB-rate scale [14].

Binary mask method

Ideal binary mask is one of the essential components in speech segregation from silence or noise in monaural speech signals that has the optimum performance in segregation of time frequency units. Performance of ideal binary mask in energy distribution examination of speech signal is based on the human auditory and automatic speech recognition systems and evaluation is carried out based on signal to noise ratio computation.

To analyze the binary mask method, Computational Auditory Scene Analysis (CASA) is used; it is a new approach to solve speech enhancement problem [8, 12]. CASA approach includes two main stages: segmentation and grouping [15].

In the binary mask method, SNR amount of the signal's different components is compared with each other such that if SNR is greater than a specific amount (e.g.θ), the binary mask in that point of time frequency gets 1 (or a given value) otherwise it gets 0 [12].

$$\mathbf{LC} = \mathbf{log}^{\frac{|\mathbf{s(t,f)}|^2}{|\mathbf{n(t,f)}|^2}} \quad (6)$$

Where $S_{(t,f)}$ is the time frequency unit in clean signal and $N_{(t,f)}$ is the T-F unit in estimated noise signal. Target formula for time frequency unit calculation in [16] is given by:

(7)

$$\mathbf{Em(t,f)} = \begin{cases} \mathbf{1} & \mathbf{if\ s(t,f) > \boldsymbol{n}(t,f)\ and\ s(t,f) > \boldsymbol{\theta}} \\ \mathbf{0} & \mathbf{other} \end{cases}$$

Θ is the threshold of the binary masks. In IBM, when Θ is equal to 1(or non-zero value), it indicates $|s(t,f)|^2 \geq |n(t,f)|^2$ and otherwise 0.

In binary masking method, the signal is divided to m segments that k is the number of samples in one segment. $S_{\emptyset}$ is the target signal energy; S_d is estimation of target signal segregation [15,18].

$$\mathbf{SNR_{SEG}} = \frac{\mathbf{10}}{\mathbf{M}} \sum_{\mathbf{m=0}}^{\mathbf{M-1}} \mathbf{log} \frac{\sum_{\mathbf{k=Km}}^{\mathbf{Km+K-1}} \mathbf{S_{\emptyset}^2(k)}}{\sum_{\mathbf{k=Km}}^{\mathbf{Km+K-1}} \left[\mathbf{S_d(k) - S_{\emptyset}(k)}\right]^2} \quad (8)$$

If SNR in (10) can be calculated in time frequency units, the formula changes as follows:

$$\mathbf{SNR_{SEG}} = \mathbf{10log} \frac{\sum_{\mathbf{k}} \mathbf{S_{(t,f)}^2(k)}}{\sum_{\mathbf{k}} \left[\mathbf{X_{(t,f)}(k) - S_{(t.f)}(k)}\right]^2} \quad (9)$$

When computing the binary mask, an increase in quantity of threshold might cause a direct problem leading to lose the target speech.

Another way of ideal binary mask estimation is to use statistical methods. Considering that most of the energy is contained in voiced segments, firstly, only information of local time frequency units is used; secondly, the correlation information among time frequency units is omitted; and thirdly, the prior probabilities of the reliable/unreliable classes estimated from training data are almost equal in high frequency channels; the prior probabilities improve classifier very limited[17].

3. Proposed Method

The following assumptions have been used to solve this problem in this paper: 1) the non-correlated noise has been combined with clean speech, 2) the background noise environment remains locally stationary in small time scope (about 20 ms) and the silent scope in time range of 300 rpm is used for stationery noise computation. If the noise

changes to a new state, a limited time (about 300 rpm) is sufficient to estimate the noise spectrum at the signal onset (before speech). Finally, noise reduction is possible by removing noise spectrum from the noisy speech signal.

At the first stage, the noisy speech passes through the spectral subtraction method. In this article, Hanning window with a length of 256 and 50% overlapping has been used and a new method of spectral subtraction with 8 windows has been adopted to estimate the noise power spectrum.

At the second stage, decomposition of an input signal in the frequency domain has been used with 128 Gamma-tone filters [11],which are equal to rectangular bandwidth rate scale from 50 to 8000 Hz (see [12] for details). In each filter channel, the output is divided into 20-ms time frames with 10-ms overlapping between consecutive frames. Unlike previous approaches from Channels 1,2,4,8,16,32,64,128 have been used from a bank of 128 Gamma-tone filter for implementing the binary mask and then the output has been calculated.These channels are selected via more SNR calculation as compared with the other channels :

$$\mathbf{p(f)} = \frac{\sum_t \mathbf{E(f,t)}}{\sum_f \sum_t \mathbf{E(f,t)}} \quad (10)$$

At the final stage, noise and target speech obtained from the spectral subtraction pass through the binary mask. The evaluations indicate that the proposed method mainly improves the clarity and intelligibility of speech signal as compared with the binary mask when it is applied to all channels.

4. Experimental Result

A speech from TIMIT is selected as a clean speech that is a “good service needs to be rewarded by big tips”. Three models of the noise from NOISEX database (white, colored, F16) are selected to be added to the speech. The noisy speech crosses the spectral subtraction method but output speech has a musical noise. Signal Noise Ratio (SNR) is calculated by the following formula:

$$\mathbf{SNR_{imp} = SNR_{out} - SNR_{in}} \quad (11)$$

In formula (13), SNR_{in} and SNR_{out} indicate the SNR of the initial noisy signal (y) and SNR_{imp} of the enhanced signal $(\hat{s})$ respectively:

$$\mathbf{SNR_{in} = 10log_{10}} \frac{\sum \mathbf{s^2(n)}}{\sum (\mathbf{y(n) - s(n)})^2} \quad (12)$$

$$\mathbf{SNR_{out} = 10log_{10}} \frac{\sum \mathbf{s^2(n)}}{\sum (\hat{\mathbf{s}}\mathbf{(n) - s(n)})^2} \quad (13)$$

Table 1 indicates the SNR_{imp} in spectral subtraction method with three models of the noise from NOISEX database. The IS parameter is initial silence length duration in second and its default value is .25 sec.

Table 1.SNR_{imp} by Spectral subtraction method with different values of IS

SNR(Signal Noise Ratio)	White Noise	Colored Noise	F16 Noise
IS = 0.1	7.3264	20.3188	5.0784
IS = 0.2	7.3274	20.3469	5.0766

IS = 0.25	7.3272	20.3511	5.0764
IS = 0.3	7.3273	20.3964	5.0772
IS = 0.5	7.3276	20.3832	5.0588

To implement and decompose the combined signal into the time frequency domain, a bank of 128 gamma-tone filter is used. Considering Formula 9the useful Time-Frequency unit is set to "1", whose amount is directly decided by the energy threshold in Table 2.

Table 2.Useful "T-F" units amount in different threshold

Threshold(Θ)	"1" Time-Frequency Units	"0" Time-Frequency Units	Proportion (%)
0	5557504	0	100
0.00001	5557504	0	100
0.001	5546744	10760	99
0.01	5211783	345721	94
0.2	3220479	2337025	58

SNR_{imp} by ideal binary masking method presented in table 3:

Table 3.$SNR_{imp}$by Ideal binary Masking

noise Type	Colored	White	F16
SNR	3.7019	3.3821	5.6217

The Channels 1, 2, 4, 8, 16, 32, 64 and 128 are used from a bank of 128 gamma-tone filters for implementing the binary mask and then calculate the output and in this method, the target speech has an enhancement shown at the end of Page.

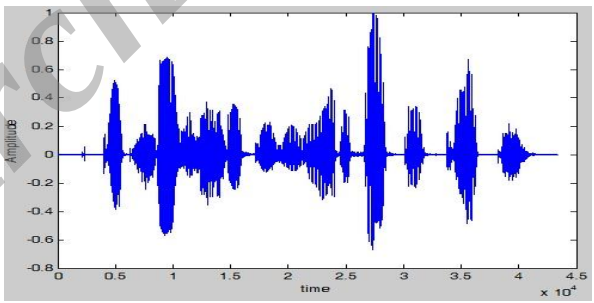

Figure 2.The clean signal

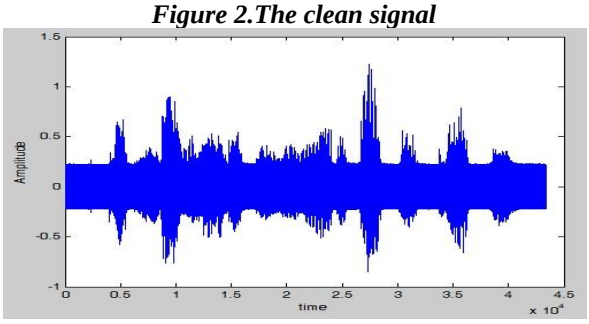

Figure 3.The noisy signal with colored noise.

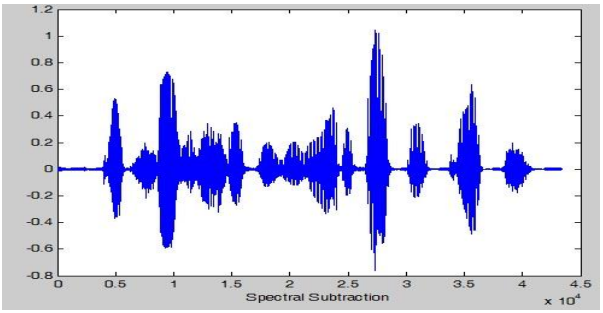

Figure 4.The enhancement signal with spectral subtraction method.

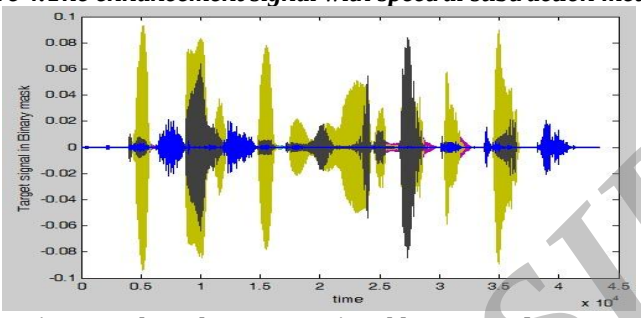

Figure5. The enhancement signal by proposed system.

Table 4 presents the SNR_{in}, SNR_{out} and SNR_{imp} values for different noise types:

Table 4.SNR calculation for Ideal binary Masking

noise Type	Colored	White	F16
Colored	-5.4225	14.1392	19.5621
White	-4.8211	3.4676	8.2887
F16	-7.0899	-2.3058	5.7841

Table 4 shows that the performance of the proposed system is the best for colored noise and it is weak for F16 noise.

4. Conclusion

The proposed speech enhancement method combined spectral subtraction and binary masking. Spectral subtraction algorithm is used for reduction of noise in noisy speech and then binary masking method is used for monaural speech separation from musical noise introduced by spectral subtraction. The experiments showed that this method has its best performance for colored noise when the similarity of noise power to the target signal is least.

References

[1] H. Liu, X. Yu, W. Wan and R. Swaminathan, "An Improved Spectral Subtraction Method," Audio, Language and Image Processing, pp. 790-793, china, 2012.

[2] J. Deller, J. Proakis, H. Hansen, "Discrete Time Processing of Speech Signals," Macmillan, New York, 1993.

[3] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans Acoustics, Speech and Signal Processing, vol. ASSP-27, No. 2, pp.113-120, 1979.

[4] H. Hu, F. Kuo, H. Wang, "Supplementary Schemes to Spectral Subtraction for Speech Enhancement," Speech Communication, 2002.

[5] H. Gustafsson, S. Nordholm, "Speech Subtraction using Reduces Delay Convolution and Avaptive Averaging," IEEE Trans Speech and Audio Processing, vol. 9, No. 8, pp. 799-807, 2001.

[6] S. Ayat, M. Manzuri, R. Dianat, "An Improved Spectral Subtraction Speech Enhancement System by Using an Adaptive Spectral Estimator," IEEE Canadian Conference on Electrical and Computer Engineering, pp. 261-264, 2005.

[7] G. Brown, M. Cooke, "Computational Auditory Scene Analysis," Computational Speech and Language, vol. 8, No. 4, pp. 297-336, 1994.
[8] Y. Jiang, H. Zhou, "An Algorithm Combined with Spectral Subtraction and Binary Masking for Monaural Speech Segregation," Signal Processing, Communication and Computing, pp. 1-4, 2011.
[9] Y. Cai, C. Hou, "Sub-band Spectral Subtraction Speech Enhancement Based on the DFT modulated Filter Banks," Signal Processing, pp. 571-574, 2012.
[10] L. Cao, T. Zhang, H. Cao, "Multi-band Spectral Subtraction Method Combined with Auditory Masking Properties for Speech Enhancement," International Conference on image and signal processing, 2012.
[11] R. Patterson, J. Holdsworth, I. Nimmo-smith, P. Rice, "An Efficient Auditory Filter bank based on the Gamma-tone Function," MRC Applied Psychology Unite, 1988.
[12] D. Wang, G. Brown, E. Hoboken, "Computational Auditory Scene Analysis: principles, Algorithms, and Applications," Wiley and IEEE Press, 2006.
[13] H. Guoning, D. Wang, "A Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation," Audio, Speech and Language Processing, IEEE Press, pp. 2067-2079, 2010.
[14] http://www.dcs.shef.ac.uk/~ning/, October-2-2012.
[15] Y. Jiang, H. Zhou, Z. Feng, "Performance Analysis of Ideal Binary Masks in Speech Enhancement," Image and Signal Processing, IEEE Press, pp. 2422-2425, 2011.
[16] L. Yipeng, "On the Optimality of Ideal Binary Time-Frequency Masks," Speech Communication, pp. 230-239, 2009.
[17] S. Liang, W. Liu, "Binary Masking Estimation for Voiced Speech Segregation Using Bayesian Method," IEEE Trans, Pattern Recognition, pp. 345-349, 2011.
[18] T. De Souza, G. Regrigues, H. Yehia, "Binary Spectral Masking for Speech Recognition System," Telecommunication and Signal Processing, pp. 432-436, 2012.
[19] W. Santosh, P. Prem, T. Nitya, "Speech Enhancement Using Spectral Subtraction and Cascades-Median Based Noise Estimation for Hearing Impaired Listeners," IEEE Conference, India, 2013.