TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.6, June 2014, pp. 4475 ~ 4484
DOI: 10.11591/telkomnika.v12i6.5485

A Dual-Microphone Speech Enhancement Algorithm for Close-Talk System

Yi Jiang*[1], Zhenming Feng[1],Yuanyuan Zu[2], Xi Lu[2]
[1]Department of Electronic Engineering, Tsinghua University
Beijing 100084, P.R. China, +86-62781702
[2]Quartermaster Equipment Research Institute, General Logistics Department
Beijing 100082, P.R. China, +86-62276675
*Corresponding author, e-mail: jiangyi09@mails.tsinghua.edu.cn

Abstract

While human listening is robust in complex auditory scenes, current speech enhancement algorithms do not perform well in noisy environments, even close-talk system is used. This paper addresses the robustness in dual microphone embedded close talk system by employing a computational auditory scene analysis (CASA) framework. The energy difference between the two microphones is used as the primary separation cue to estimate the ideal binary mask (IBM). We also use voice activity detection to find the noise periods, and update the separation critical value. Generalization interference locations and reverberant conditions are used to examine performance of the proposed system. Evaluation and comparison show that the proposed system outperforms other two systems on the test conditions.

Keywords: *energy difference, close-talk system, speech enhancement, binary mask*

1. Introduction

Speech enhancement in noisy and reverberant environments is a very challenging problem.Even a close-talk microphone is used to collect the target speech. The performance gap between human listener and speech enhancement algorithms remains large [1]. In recent decade, Computational auditory scene analysis (CASA) provides a new approach to solve this speech enhancement problem directly depending on the human listening processing [2]. It uses two dimension ideal binary masks (IBM) to segregate the speech domination time-frequency (T-F) units and can improve the intelligibility of the noisy speech for both normal-hearing and hearing-impaired listeners dramatically [3, 4].

Based on CASA, one microphone speech enhancement algorithms use monaural feature, fundamental frequencies (F0), onset/offset, GFCC [5], and so on, to segregate the target speech from the noise. These systems are hard to work in noisy conditions, for the noise distorts the monaural feature. Monaural speech separation is particularly difficult as one has access only to a single-channel noisy signal.

With two ears, human listening is robust under both noisy and reverberant conditions. Binaural cues contribute to auditory scene analysis [6]. So far, dual-microphone system often employs binaural feature, such as interaural time differences (ITD) and interaural intensity differences (IID). These systems have yielded significant improvement in speech separation [7, 8]. Kernel density, GMM with SVM [9, 10], multilayer perceptron (MLP) [11] and deep neural network (DNNs) [12] are used to model this features and classifier the target speech. But the model training and classification is too complex and time consumption to use in real time embedded system. Other two or multi microphone array systems are also used in portable system [13, 14], but can'tdeal with the unsteady noise.

In this paper, using the computational auditory scene analysis (CASA) as a framework, we propose a speech separation approach for close-talk system. We just use the dual-microphone energy difference (DMED) as the cues to separate the close speech and the far noise [15]. We also use the DMED to detect the noise period, and update the separation critical value. The system does not need train, and runs on real time. It is simple to integrate in the embedded system.

Received December 24, 2013; Revised February 19, 2014; Accepted March 5, 2014

The rest of this paper is organized as follows. In section 2, we present an overview of the speech separation algorithm for dual-microphone close-talk system.section 3 describes how to extract DEMD feature and estimate the IBM. The systematic evaluation and comparison is present in section 4. Finally, we conclude the paper in section 5.

2. System Overview

The proposed dual-microphone embedded close talk system is shown in Figure1. Two microphones are used to collect the close mouth speech and the far noise simultaneous. The same two gammatone filterbanks are used to decompose the two microphones input into T-F domain representations. A T-F unit corresponds to a certain channel in a filterbank at a certain time frame. It is a similar manner as the human ears do.

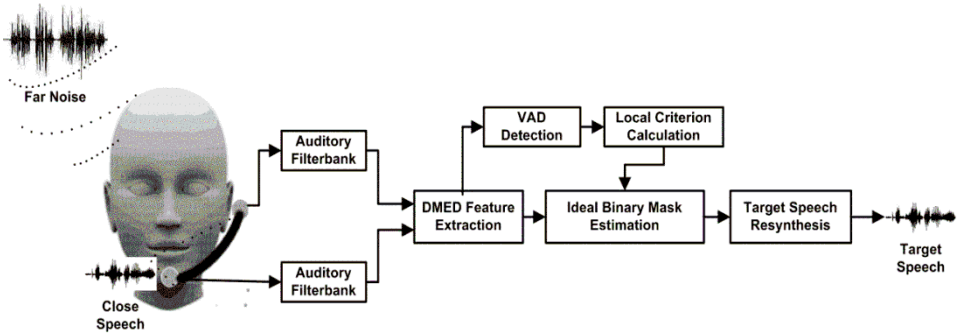


Figure 1. Schematic Diagram of the Dual-microphone Close-talk System

We extract the dual microphone energy difference (DMED) feature in each T-F unit pair, which is respect to the locations of the sounds. The DMED is used as the cue to estimate the IBM for the T-F units, where 1 indicates the target signal dominates the corresponding time-frequency (T-F) unit and 0 otherwise.

The DMED feature is also used in voice activity detection (VAD) to find the noise only period. In this period, we update the local SNR criterion (LC) value for the IBM estimation algorithm. As two microphone energy difference features vary with frequency channels [16, 17], the LC is calculated for each channel separately.

In resynthesis processing, the T-F units with the target label (unity) comprise the segregated target stream.

3. Feature Extraction and Speech Segregation

As aclose-talk system, one microphone indexed as 1 is placed in front of the mouth only several centimeters away. It mainly collects the target speech. Another microphone indexed as 2 is placed near the left ear, which used to collect the target speech and interference equally.

3.1. Gammatone Filterbanks

The auditory filterbank is used to decompose input mixture signal into small requency band respectively. In this paper, a simplified implementation of a cochlear model gammatone filterbanks proposed by Roy Patterson is provided for the embedded system. Each filter bank is designed as a set of parallel equivalent rectangular bandwidth (ERB) band pass filter, described in time domain as Equation (1).

$$g(f_c,t) = t^{n-1}e^{-2\pi b(f_c)t}\cos(2\pi f_c t + \phi) \quad (1)$$

The channel number of the gammatone filterbanks c is set to 32 for embedded system. The filter center frequencies f_c are organized from high frequencies at the base of the cochlea to

low frequencies at the apex. In order to simulate human auditory behavior, the central frequencies of the filter bank in this paper f_c is from 50Hz to 8000Hz. And the band width of each cochlear filter $b(f_c)$ is described by 1.019 times the ERB on each central frequency. The order, n is set to 4 for embedded system. ϕ is the phase. We set ϕ to zero. We also introduce a simplified gammatone filter form to do this work [18]:

$$G(s) = \frac{6[(s+B_c)^4 - 6(s+B_c)^2\omega_c^2 + \omega_c^4]}{[(s+B_c)^2 + \omega_c^2]^4} \quad (2)$$

Where $B_c = 2\pi b(f_c)$ and $\omega_c = 2\pi f_c$. Each filterbank setup with four second-order filters.

3.2. DMED Feature Extraction

With gammatone filterbank, the signals received by microphones are divided into various frequency channels. Then a time frame window is used to segment the signal to small units called T-F units.

$$X(c,m) = g(f_c, t_m) = \left[x_{f_c,mt1,}, x_{f_c,mt2}, ...x_{f_c,mtk} \right] \quad (3)$$

The filter frequency channel index is c, center frequency is f_c. The frame index is m, The t is the time section of frame m, signal contains k data points. In this paper, the T-F unit is 20-ms time frames with 10-ms overlapping between consecutive frames. k equal to 320.

The energy of the one T-F unit is calculate by:

$$\left\| X(c,m) \right\|^2 = \sum_t X_{mt,f_c}^2 \quad (4)$$

The energy difference between the two microphones in each T-F units is calculated by the energy ratio.

$$DMED(c,m) = \frac{\left\| X_1(c,m) \right\|^2}{\left\| X_2(c,m) \right\|^2} \quad (5)$$

The DMED value indicate the distance difference between the sound sources and the two microphones. We also use $DMED_S(c,m)$ and $DMED_N(c,m)$ to indicate DMED value of the target sound sourceand noise source respectively.

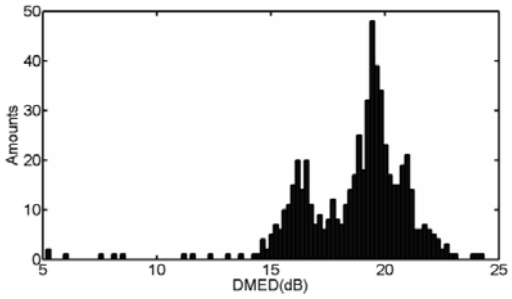
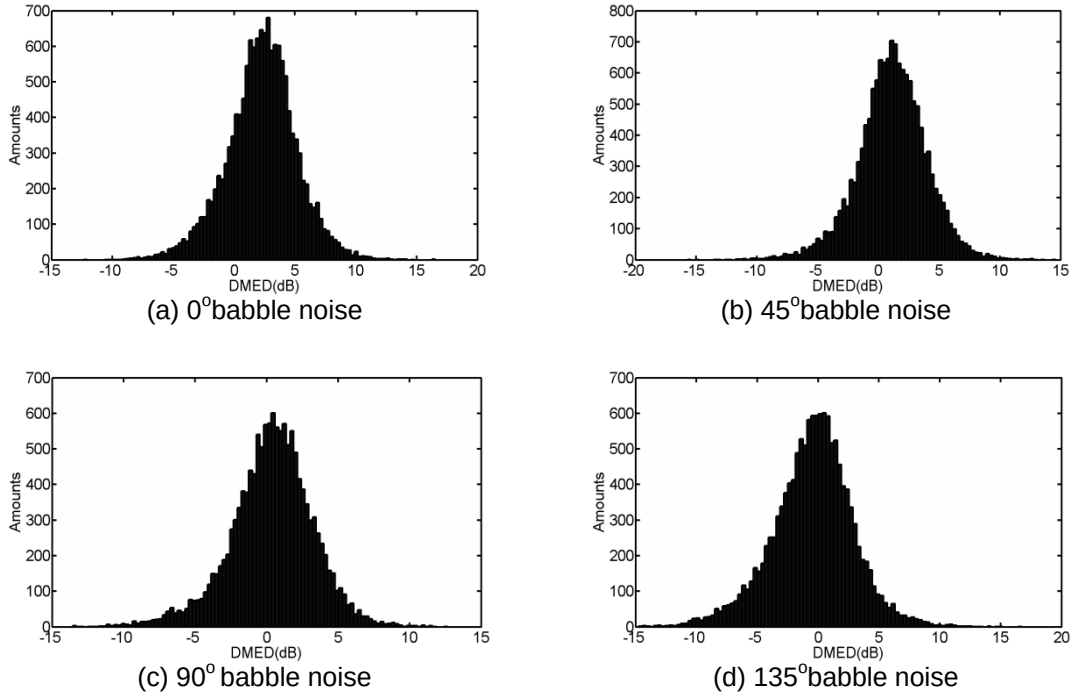


Figure 2. The histogram of $DMED_S$

As shown in Figure 2. In the 16th channel $DMED_s$ of the near sound source are distributed around 20dB. The smallest value is above 10dB, and the largest is more than 24dB. The two big peaks indicate the a microphone location changing in the talking period.

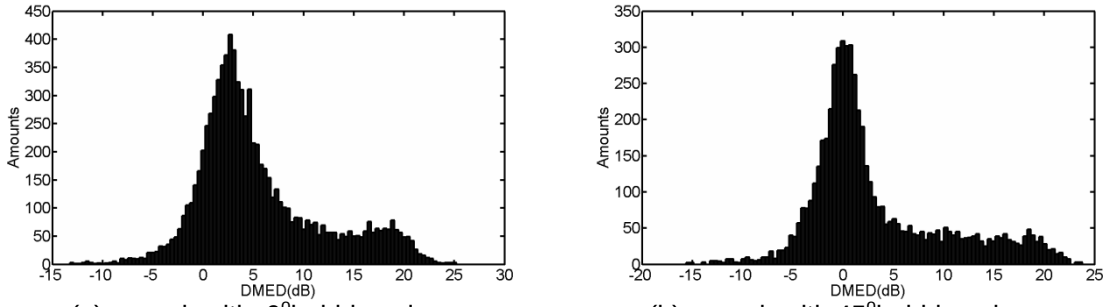


(a) 0°babble noise (b) 45°babble noise

(c) 90° babble noise (d) 135°babble noise

Figure 3. The Histogram of $DMED_N$ at Various Azimuths

Figure 3 shows the 16th channel $DMED_N$ of a babble noise. The interference locat at azimuth 0°, 45°, 90° and 135°. As shown in Figure 3, the $DMED_N$ values are close to 0dB. Compare to the distance between the far interference to the two microphones. The distance difference between interference and two microphpones is verysmall. At different locations, the $DMED_N$ change small.



(a) speech with 0°babble noise (b) speech with 45°babble noise

Figure 4. DMED of Mixtures

The DMED of two mixture signals are shown in Figure 4. The mixture speech is actual recording sentence, with a babble noise present at azimuth 0°(a) and 45°(b). There are obviously two peaks in (a) and (b). The peaks on the left (close to 0dB) represent the DMED offar noise, while the peaks on the right (close to 20dB) represent the DMED of target speech. Due to the effect of the human head's shape and the microphone location, the DMED is robust with various noise locations.

As a close-talk systems, the difference between $DMED_N$ and $DMED_S$ is significant. The DMED is used as thecue to separate the target speech.

3.3. Ideal Binary Mask Estimation

In this close-talk systems, the IBM is estimated by the DMED cues, and used to separate the target signal by fellows:

$$BM(c,m)=\begin{cases}1 & if\ DMED(c,m)>LC(c,m)\\ 0 & others\end{cases} \tag{6}$$

Where BM is the estimated binary mask value of the T-F units in frequency channel c and time frame m. The 1 indicates T-F units that are dominated by the target speech, and 0 indicates the T-F units that are dominated by noise. LC is the local separation critical. Using the DMED as the speech separation cues, the LC is calculated as:

$$LC(c,m)=\frac{2}{\frac{1}{DMED_S(c,m)}+\frac{1}{DMED_N(c,m)}} \tag{7}$$

In usually close-talk implementation (Figure 2) and given the conclusion of HRTF, $DMED_S$ is always over 100. $DMED_N$ from the far noise is around 1, much smaller than $DMED_S$. The difference between them is significant. Considered the Equation (7), the $LC(c,m)$ is decided by the smaller valuebetween $DMED_S$ and $DMED_N$. Obviously, the $DMED_N$ is the decisive factor. We calculate the LC as:

$$LC(c,m)=\frac{2}{\frac{1}{DMED_N(c,m)}+0.01} \tag{8}$$

In this paper we update the LC in noise only period. A voice activity detection (VAD) is used to distinguish the noise only sectionin the mixture.

$$VAD(m)=\begin{cases}1 & N_m<\alpha\ and\ DMTD(m,\tau)<\theta\\ 0 & others\end{cases} \tag{9}$$

Where N_m is the number of speech including channels in frame m. It count the channel number, which $DMED(c,m)>10$. And α set to 2, indicates the target speech exist only in very limited channels. The $DMTD(m,\tau)$ is the time difference between two microphonesignals. For the $DMTD$ of thetarget speech is larger than the noise signal in most conditions. The θ sets to 6 based on oure experience.

To calculate the $DMTD(m,\tau)$, we use the two microphone signals x_1 and x_2 as whole. The normalized correlogram between two microphones in each frame $Corr(m,\tau)$ is calculated with delay τ by the following cross correlation function.

$$Corr(m,\tau)=\frac{\sum_{n=1}^{k}(x_{m,1}(n)-\bar{x}_{m,1})(x_{m,2}(n-\tau)-\bar{x}_{m,2})}{\sqrt{\sum_{n=1}^{k}(x_{m,1}(n)-\bar{x}_{m,1})^2}\sqrt{\sum_{n=1}^{k}(x_{m,2}(n-\tau)-\bar{x}_{m,2})^2}} \tag{10}$$

$$DMTD(m,\tau)=\max_{\tau}Corr(m,\tau) \tag{11}$$

Where k is the frame size in sampling point, and equal to 320 in this study. The range of τ is from -1ms to 1ms.

4. Evaluation and Comparison

4.1. Test Corpous

An actual recording data of a dual-microphone system is used to test the algorithm's performance in office environment. we record the target speech and noise separately.

In the recording test corpus, the target speech includes 600 short Chinese utterances involving 200 Chinese names, 200 stock names and 200 place names, which were collected in quiet office rooms by two male speakers and one female speaker. The noise sounds include babble, white, m109 and machinegun come from NOISE 92 database. The interference presents at a distance of 1.5m from the listener, azimuth at 0°, 45°, 90°, 135° and 180°. All at 0° elevation, unless otherwise specified.

We use recording dual-microphone clean speech and various locations noise to generate the mixture signal with defined SNR. At this condition, the clean speech is fixed on original magnitude. We adjust the energy of recording noise to get the defined SNR.

Another simulated test corpus is also employed, which is created by various clean speech signals with four different noises and2 room reverberant configurations.

We use a set of binaural impulse responses (BIRs) to generate the transfer function from interference to two microphones. To the close-talk system, it is hard to get the transfer function from mouth to two microphones. We use the same sentences with different amplitude to simulate the target speech signals of the two microphones. The speech materials of are chosen from TIMIT corpus randomly. Four noises come from NOISE 92 database too.

We use the ROOMSIM package to generate a library of BIRs. The reflection paths of a particular sound source are obtained using the image reverberation model for a small rectangular office room. Reflection coefficients of the wall surfaces are set to be equally. The room size is 6m×4m×3m. The two microphones locate at 2.5m×2.5m×2m. The distance between the two microphones is 8cm. Seventeen sound sources locate at a same distance of 1.5 m from the two microphones. The azimuth isbetween -180° and 180°, and the elevation is between 0° and 90° by step of 45°.

Noises are drawn randomly from the data base and are convolved with a select BIRs to generate the mixture with the speech utterances. The interference number is randomly 1 to 5. The interferences are locate at 17 positions randomly. All SNR of the mixture signals are -5dB.

Finally, we generate a set of 1000 simulated mixtures to evaluate the performance of the dual microphone speech enhancement algorithm.

4.2. Comparison Systems

In the experiments below, reuslts of the proposed method are compared with two existing methods from the literature [14, 19]. The system proposed in [14], denoted PLD, is a coherence-based algorithm. The energy level difference and coherence function is used to get the target sound in noisy environment. The distance between the two microphones is small, which make it hard to be used in close-talk system. The algorithm estimates the power spectral density of the noise and reduce it, which makes it hard to eliminate the non-steady noise. In this paper,we use the first 100ms signal of the mixture to estimate the noise. The second comparison system used is the joint localization and segregation approach presented in [19], dubbed MESSL, and is representative of the spatial clustering approach to localization. The system requires specification of the number of sources and iteratively fits GMM models of inter-aural phase difference (IPD) and ILD to the observed data using an EM procedure. Across frequency integration is handled by tying GMM models in individual frequency bands to a principal ITD. We give the MESSL the number of the sound source in these testing.This algorithm is also hard to use on line. We use an implementation of the DLP and MESSL provided by the algorithm authors. We also show the results of the IBM as a baseline.

4.3. Evaluation Results

1) SNR performance with recording data

Table 1 shows the speech segregation results with one babble noise invarious SNR conditions. The interference fixed at azimuth 0°. The The proposed algorithm gets the best performance in all conditons, and close to the results of IBM. Almost all systems get positive results on this test conditions, especially in low SNR conditions. Because the locations of the target speech and noise are not fixed in the recording period, the IPD and ILD are changing

from time to time. The MESSL get the worst results. The proposed algorithm suit to deal with daily noise.

Table 1. SNR (dB) Performace with Babble Noise

SNR(dB)	-5	0	5	10
IBM	6.76	9.74	12.91	16.37
Proposed	2.28	6.32	10.49	14.43
DLP	1.97	5.45	9.32	13.43
MESSL	1.31	4.47	5.11	8.87

The SNR performance with various noise type is also evaluated. The SNR of the mixture signal is 0dB. The interference locate atazimuth 180°, elevation 0°. As shown in Table 2, in mostconditions, the proposed algorithm gets better result than compare systems. With machinegun noise, the MESSL get a little higher score than the proposed system. DLP lose ability on machinegun noise.

Table 2. SNR (dB) Performance with Various Noise Types

Noise type	BABBLE	M109	MACHINEGUN	WHITE
IBM	9.99	9.56	11.98	10.29
Proposed	6.46	5.47	8.25	7.00
DLP	5.67	4.66	0.00	6.58
MESSL	5.37	5.30	8.51	6.17

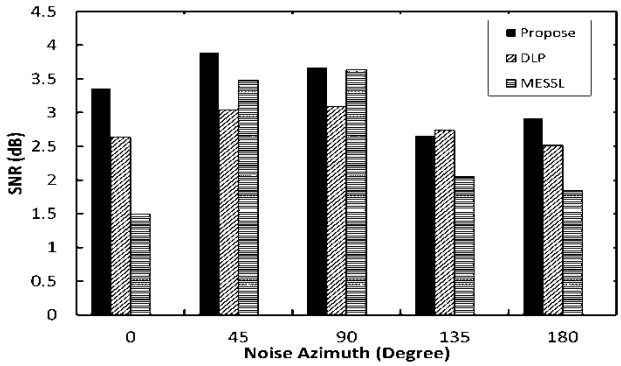


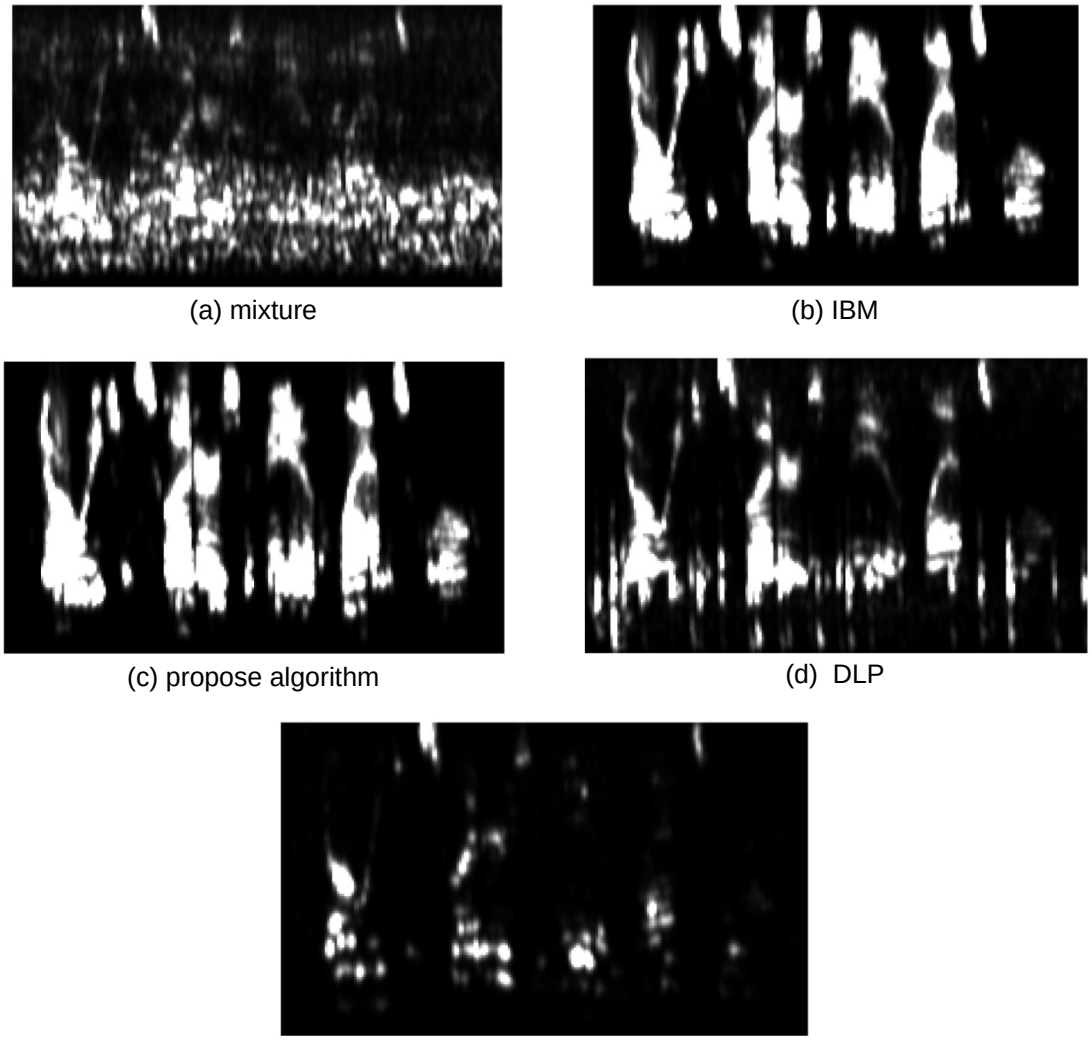
Figure 5. The SNR Performances with Babble Noise at Various Azimuths

White noise locatesat various azimuths is used to evaluate the SNR performance of systems. The SNR of the mixture signal is -5dB. As shown in Figure 5. The proposed algorithm gets the highest SNR improvements in most conditons. It is also more robustness than the compare two algorithms.The proposed DMED algorithm can improve the SNR on various azimuths an elevations. It also has good performance on various frequencies.

2) Performance evaluation with simulated data

Figure 6 illustrates the results of the three speech segregation algorithms. This is a five interferences and -5dB SNR test conditions. The five interferences randomly locate at 17 positions. (a) shows the spectrograms of the reverberant mixture. The noise signal disorder the target speech serious. It is hard to discriminate the target speech from noise signal. The (b) and (c) is the spectrograms of the signal that resynthesized by the ideal binary mask and the binary mask estimated by the proposed algorithm. Compare to the mixture signal, this two methods get the target speech and decrease the noise significant. The perform of the proposed system is very close to the result of IBM. The estimation errors make the result of DMED little worse than the result of IBM. The (d) is the spectrograms output of the PLD based algorithm. It reduces most of noise, and damage the target speech at the same time. There are also some noise

retain obviously. The result of MESSL algorithm is shown in (e). It removes most noise, and rebuilds the target speech. It also damages the target signal significant, especially on some certain signal frequency.



(a) mixture (b) IBM

(c) propose algorithm (d) DLP

(e) MESSL

Figure 6. The Spectrograms of Speech Segregation Results

We evaluate the system performance with various number of babble noises. All input mixture SNRs are -5dB. We calculate the SNR improvements. Results are given in Table 3. The ideal binary mask (IBM) is also used as the baseline.

Table 3. SNR Improvements with 1 to 5 Interferences

Noise sound number	1	2	3	4	5
IBM	8.64	7.46	7.16	7.13	6.92
Proposed	7.84	6.63	6.38	6.37	6.13
DLP	2.72	1.95	1.99	2.13	2.20
MESSL	7.64	5.88	4.34	3.69	3.05

As shown in Table 3, all algorithms have a positive result. The worst result is 1.95 from DLP. The proposed algorithm gets the best performance in all conditionsand close to the baseline IBM results. It has a gradually decreasing with the sound number increasing. The multiple sound sources are the main reason for the worse performance of MESSL algorithm. MESSL gets very high score in just one babble noise, but drop quickly with the interferences number increasing. We also find the system perform better with simulated date than recording data, forthe recording data provides more complex auditory scenes than the simulated data.

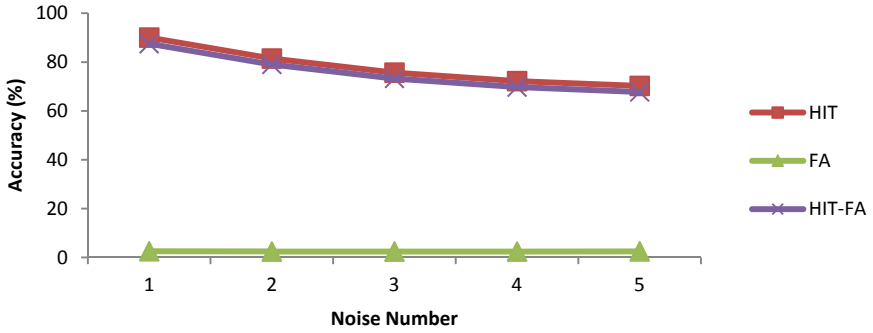


Figure 7. The Performance of Speech Intelligibility with Various Interferences

To measure classification-based separation performance, we use HIT-FA as our main evaluation criterion, which has been shown to be well correlated to human intelligibility [3]. The HIT rate is the percent of correctly classified target-dominant T-F units in the IBM. The FA (false-alarm) rate is the percent of wrongly classified interference-dominant T-F units. As shown in Figure 7. With the noise number increasing the intelligibility decrease slowly. The HIT-FA rate is almost 70% with five babble noises, and better than most of the one microphone algorithms [5].

5. Summary Concluding Remarks

The performance of the proposed algorithm with various interferences in reverberation and noisy environments are evaluated by SNR and speech intelligibility. The results indicate the proposed system has better performance than other comparison algorithms. The proposed speech separation approach is suit for the close-talk system, not only high performance but also simple complex and real time. The monaural feature, such as pitch, GFCC and MFCCis potentially benefit speech detection and segregation, which can be used to improve the performance of the proposed algorithm. This is a topic that will be addressed in future work.

References

[1] WG Yan, GY Xiang, ZX Qun. A signal subspace speech enhancement method for various noises, *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(2): 726-735.
[2] DL Wang, GJ Brown. *Eds.* Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. New Jersey: JOHN WILEY & SONS. 2006.
[3] N Li, PC Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *J Acoust Soc Am.* 2008; 123(3): 1673-1682.
[4] Y Jiang, W Liang, H Zhou, ZM Feng. Performance of binary time-frequency masks in low signal to noise ratio environments. *J. Tsinghua Univ.* 2012; 52(5): 636-641.
[5] YX Wang, K Han, DL Wang. Exploring monaural features for classification-based speech segregation. *IEEE Trans. On Audio Speech Lang Process.* 2013; 21(2): 270-279.
[6] AS Bregman. Auditory Scene Analysis. MIT Press, Cambridge, MA. 1990.
[7] J Woodruff, DL Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Trans. on Audio Speech Lang Process*. 2012; 20(5): 1503-1512.
[8] S Keronen, H Kallasjoki, U Remes, GJ Brown, JF Gemmeke, KJ. Palomaki. Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment. *Computer Speech & Language*. 2013; 27(3): 798-819.

[9] B Jing, W Jie, ZX Ying. A Parameters Optimization Method of nu-support Vector Machine and Its Application in Speech Recognition. *Journal of Computers*. 2013; 8(1): 113-120.
[10] Y Bo, L Haifeng, F Chunying. Speech Emotion Recognition based on Optimized Support Vector Machine. *Journal of Software*. 2012; 7(12): 2726-2733.
[11] ZZ Jin, DL Wang. A supervised learning approach to monaural segregation of reverberant speech. *IEEE Trans. On Audio Speech Lang Process.* 2009; 17(4): 625-638.
[12] G Hinton, S Osindero, Y The. A fast learning algorithm for deep belief nets. *Neural Comput.* 2006; 18(7): 1527–1554.
[13] H Zhou, Y Jiang, M Jiang, Q Chen. Energy difference based speech segregation for close-talk system. *Applied Mechanics and Materials*. 2012; 223-231: 1738-1741.
[14] NYousefian, PC Loizou. A dual-microphone speech enhancement algorithm based on the coherence function. *IEEE Trans. on Audio Speech Lang Process*. 2012; 20(2): 599-609.
[15] Y Jiang, M Jiang, Y Zu, H Zhou, Feng. Using energy difference for speech separation of dual-microphone close-talk system.*Sensors and Transducers*. 2013; 21(5): 122-127.
[16] J Blauert. Spatial Hearing The Psychophysics of Human Sound Localization, Cambridge, MA: MIT Press. 1997.
[17] N Roman, DL Wang, GJ Brown. Speech segregation based on sound localization. *J Acoust Soc Am.* 2003; 114(4): 2236-2252.
[18] Y Jiang, YY Zu, X Chen, H Zhou. Performance evaluation of a gammatone filterbank for the embedded system. *Applied Mechanics and Materials*. 2013; 336-338: 1459-1462.
[19] MI Mandel, RJ Weiss, DPW Ellis. Model-based expectation-maximization source separation and localization. *IEEE Trans. on Audio Speech Lang. Process.* 2010; 18(2): 382-394.