Sensors & Transducers, Vol. 21, Special Issue, May 2013, pp. 122-127

www.sensorsportal.com

Sensors & Transducers

http://www.sensorsportal.com

Using Energy Difference for Speech Separation of Dual-microphone Close-talk System

[1] Yi Jiang, [2] Ming Jiang, [3] Yuanyuan Zu, [3] Hong Zhou, [1] Zhenming Feng

[1] Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P. R. China

[2] Deparment of optoelectronic science and engineering, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, P. R. China

[3] Quartermaster Equipment Research Institute, Beijing, 100010, P. R. China

E-mail: yijiang2013@yeah.net

Received: 23 March 2013 /Accepted: 14 May 2013 /Published: 30 May 2013

Abstract: Using the computational auditory scene analysis (CASA) as a framework, a novel speech separation approach based on dual-microphone energy difference (DMED) is proposed for close-talk system. The energy levels of the two microphones are calculated in time-frequency (T-F) units. The DMEDs are calculated as the energy level ratio between the two microphones, and used as a cue to estimate the signal to noise ratio (SNR) and ideal binary mask (IBM) for mix-acoustic of the close-to-mouth microphone. The binary masked units are grouped to generate the target speech. Test with speeches and different noises show that the algorithm is more than 95 % accurate. As the T-F units' length increase, the accuracy increase as well. Using automatic speech recognition (ASR) analysis, we show that the proposed algorithm improves speech quality in actual close talk system. *Copyright © 2013 IFSA.*

Keywords: Speech separation, Computational auditory scene analysis (CASA), Ideal binary mask (IBM), Close-talk system, Dual-microphone energy difference (DMED).

1. Introduction

Given the popularity of portable devices, people can communicate anywhere and anytime. Background noise is one of the primary factors in decreasing the performance of portable communication systems and robust automatic speech recognition (ASR) systems. Close-talk equipment, such as mobile phones or headsets, often uses a nearby microphone to improve the quality of speech collection. Even if the microphone is close enough to the mouth, obtaining clean speech is also difficult in complex auditory scenes, especially in noisy environments such as railway stations, airports and the subway.

In recent years, great progress has been made in the study of the computational auditory scene analysis (CASA) algorithm for speech separation [1], ASR [2], and robust speaker identification [3] from mixture acoustic signals. Using CASA as the framework, the acoustic input is divided into auditory segments as time frequency (T-F) units by gammatone filters. Each T-F unit likely comes from one single source [4]. Wang proposes the ideal binary masks (IBM) as the critical computational goal for a CASA based system. Many studies have confirmed the good performance of IBM in different noise conditions and low SNR conditions [5]. The key point of CASA methods is to find proper cues to assign each T-F unit to different sources. The main cues in the monaural speech segregation system include pitch [6] and onset/offset [7], which are too complex or sensitive to be used in real live application systems. An inter-aural time

differences (ITD) and inter-aural intensity differences (IID) cues of dual-microphone system is used as a locator to estimate the IBM [8]. The dual-microphone system based on CASA attempts to explain the mechanism of the human ears other than speech enhancement.

Another distinguished class of dual-microphone speech enhancement techniques is the coherence-based algorithm. In a dual-microphone hearing aids system, the energy level difference and coherence function are used to get the front target sound in noisy environment [9, 10]. The aids system estimates the power spectral density (PSD) of the noise, which makes it hard to reduce the non-stationary noise. The distance between the two microphones in hearing aid system is also small, which make it hard to be used in close-talk system. A dual–microphone mobile phone system uses spectral subtraction to get the target speech [11]. The noise difference between the two microphones reduces the mobile phone's performance.

In the close-talk system, one microphone is near the mouth. The present study positions another microphone far from the mouth. Both the theoretical calculations and the experiments indicate that the energy difference between the two microphones increases substantially for a lateral sound source as distance decreases. Then, the DMED difference between the close talk and far noise can be used to separate the target speech from the noise.

2. Dual-microphone Speech Separation

The structure of the dual-microphone system is shown in Fig. 1. Two microphones in different positions are used to independently collect the target speech and noise. Using the energy level difference as separation cue, the complex audio scene can be viewed as two sound sources: a close target speech and a far environment noise. The aim of the system is to separate the target speech signal from the mixture signal of the close microphone A.

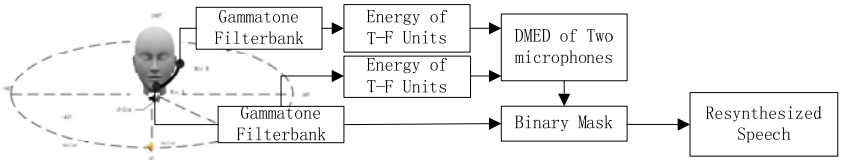

Fig. 1. Schematic diagram of the dual-microphone system.

With the framework of computational auditory scene analysis (CASA), the proposed closed-talk speech segregate processing consists of two parts: the same auditory filter bank is used to decompose the input mixture signal. Then energy is calculated in each frame as T-F units respectively. Then the energy difference between microphone A and B is used as cue to generate the binary mask. Subsequently, the binary masks are affected on the decomposed signal of microphone A to group the target speech.

3. Binary Mask Estimation

Background noise acoustically mixed with clean speech is additive in this paper. This assumption is described by the following equation:

$$X_A = S_A + N_A, \tag{1}$$

$$X_B = S_B + N_B, \tag{2}$$

where x_A and x_B refers to the mixture signal obtained by the dual-microphone A and B, respectively, which compose of target speech and environment noise. In this paper, the position of microphone A is close to the target speech. s_A and s_B refers directly to the target speech signal reaching microphone A and B, respectively. N_A and N_B is the noise signal received by microphones. The distance between A and B is less than 10 cm. the time delay of the sound between the two microphones is less than 0.3 ms, and is omitted in the energy calculation.

The energy of the mixture signal can be calculated as

$$\|X_A\|^2 = \|S_A\|^2 + \|N_A\|^2 + 2\|S_A\|\|N_A\|\cos\theta_A, \tag{3}$$

$$\|X_B\|^2 = \|S_B\|^2 + \|N_B\|^2 + 2\|S_B\|\|N_B\|\cos\theta_B, \tag{4}$$

where θ_A and θ_B indicate the angle between the vector of target speech and noise in microphones, respectively. Based on CASA, the signals received by microphones are divided into a time sequence of T-F units by gammatone filterbank and subsequent time windowing. In each T-F unit k points or k-dimensional vectors are present in time sequence. The signal of microphone A can be described as

$$X_A(t, f) = [x_1 x_2 ... x_k], \tag{5}$$

where t and f index are the time and frequency dimension. The energy of one T-F unit can be calculated as

$$\|X_A(t,f)\|^2 = \|S_A(t,f)\|^2 + \|N_A(t,f)\|^2 + 2\|S_A(t,f)\|\|N_A(t,f)\|\cos\theta_A \quad (6)$$

$$\|X_B(t,f)\|^2 = \|S_B(t,f)\|^2 + \|N_B(t,f)\|^2 + 2\|S_B(t,f)\|\|N_B(t,f)\|\cos\theta_B, \quad (7)$$

In practice, $\cos\theta_A$ and $\cos\theta_B$ is usually small, $2\|S_A(t,f)\|\|N_A(t,f)\|\cos\theta_A$ and $2\|S_B(t,f)\|\|N_B(t,f)\|\cos\theta_B$ can be ignored, especially with the increase of dimension k. Then the energy in the system is equal to

$$\|X_A(t,f)\|^2 = \|S_A(t,f)\|^2 + \|N_A(t,f)\|^2, \quad (8)$$

$$\|X_B(t,f)\|^2 = \|S_B(t,f)\|^2 + \|N_B(t,f)\|^2, \quad (9)$$

The value of DMED calculates as

$$DMED_{AB}(t,f) = \frac{\|X_A(t,f)\|^2}{\|X_B(t,f)\|^2} = \frac{\frac{\|S_A(t,f)\|^2}{\|N_A(t,f)\|^2}+1}{\frac{\|S_B(t,f)\|^2}{\|N_A(t,f)\|^2}+\frac{\|N_B(t,f)\|^2}{\|N_A(t,f)\|^2}}, \quad (10)$$

The DMED value of the target speech signal and noise can be described separately as

$$DMED_S(t,f) = \frac{\|S_A(t,f)\|^2}{\|S_B(t,f)\|^2}, \quad (11)$$

$$DMED_N(t,f) = \frac{\|N_A(t,f)\|^2}{\|N_B(t,f)\|^2}, \quad (12)$$

The $DMED_S(t,f)$ indicate the DMED value of the close sound in frame t and frequency f, and the $DMED_N(t,f)$ indicate the DMED value of the far noise in framea. In close-talk system, they can be fixed to certain value as $DMED_S$ and $DMED_N$. Then the dual-microphone energy difference is.

$$DMED_{AB}(t,f) = \frac{\|X_A(t,f)\|^2}{\|X_B(t,f)\|^2} = \frac{\frac{\|S_A(t,f)\|^2}{\|N_A(t,f)\|^2}+1}{\frac{\|S_A(t,f)\|^2}{\|N_A(t,f)\|^2}\frac{1}{DMED_S}+\frac{1}{DMED_N}}, \quad (13)$$

where $\frac{\|S_A(t,f)\|^2}{\|N_A(t,f)\|^2}$ indicates the SNR in each microphone A T-F units. Thus $DMED_{AB}(t,f)$ relates to the SNR.

In CASA, the single microphone IBM is generated based on the signal energy and noise energy in the mixed signal. The output of CASA segregation is in the form of a binary T-F mask that indicates whether a particular T-F unit is dominated by speech or background noise.

$$M(t,f) = \begin{cases} 1 & if\ \|S_A(t,f)\|^2 \geq \|N_A(t,f)\|^2 \\ 0 & others \end{cases}, \quad (14)$$

where $M(t,f)$ is the binary mask value to the T-F unit. The variable "1" indicates T-F unit that belongs to the target speech. The variable "0" indicates that the T-F unit is dominated by noise and belongs to the noise.

In this paper, we use the cues of DMED to estimate the IBM of the nearby microphone A, and $\|S_A(t,f)\|^2 = \|N_A(t,f)\|^2$ is also the separation threshold of the T-F units of microphone A. The separation threshold would be

$$DMEDT_{AB} = \frac{2}{\frac{1}{DMED_S}+\frac{1}{DMED_N}}, \quad (15)$$

This indicates that in the dual-microphone system, the harmonic mean of the DMED can be used to generate the binary mask.

The difference of the two microphones can also be described as

$$DMED_{AB}(t,f) = DMED_S + \frac{1-\frac{DMED_S}{DMED_N}}{\frac{\|S_A(t,f)\|^2}{\|N_A(t,f)\|^2}\frac{1}{DMED_S}+\frac{1}{DMED_N}} \quad (16)$$

Combined with the result of HRTF and microphone location of the close-talk system, $\frac{DMED_S}{DMED_N} > 1$. The value of $DMED_{AB}(t,f)$ increases with the increasing of $\frac{\|S_A(t,f)\|^2}{\|N_A(t,f)\|^2}$ in each T-F unit.

The binary mask for close microphone is estimated

$$DM(t,f) = \begin{cases} 1 & if\ DMED_{AB}(t,f) \geq DMEDT_{AB} \\ \gamma & others \end{cases}, \quad (17)$$

γ sets to zero to estimate the IBM. In common application, we can adapt the value from zero to one to retain part of the noise mainly units.

4. Performance and Comparison

The DMED based separation algorithm transfers the IBM of one microphone system to the dual-microphone system.

A testing corpus is employed, which created with one clean speech and different noises. The speech materials are chosen from TIMIT corpus, and noise materials come from noise 92. The mask accurate between IBM and DMED is compared in different SNR conditions. We also use actual recordings to evaluate it performance with standard ASR system.

4.1. Testing Corpus Setup

1) A simulated testing corpus.

A simulated testing corpus is created as follows to conduct an SNR evaluation:

$$X_A = S_A + N_A, \tag{18}$$

$$X_B = \frac{S_A}{a} + N_A, \tag{19}$$

where A and B is the index of two microphones. $a > 1$ indicates weakening of the target speech energy between microphone A and microphone B, which is 10 in this paper. The noise is always far away from microphone A and B, and so the energy level is almost the same to microphones A and B. The time delay or the time difference of the two microphones is therefore not considered.

The mixture signal of microphone A with different SNRs is generated to test the performance of the DMED-based algorithm.

$$SNR_A(t) = 10\log_{10}{\sum_t S_A^2(t) / \sum_t N_A^2(t)}, \tag{20}$$

The S_A is certain and fixed at "sx198" and is chosen from TIMIT test sets. Then the power of NA is adjusted to generate the mixture signal in different SNRs. The NA and weaken speech signal are used to generate the mixture signal of microphone B as equation (19).

2) Actual recordings of a dual-microphone system.

The actual close-talk recording system with two-microphone is set up as shown in Fig. 1. Microphone A is about 2 centimeters away from the mouth. Microphone B is posed near the left ear on one head-set. The distance between microphone A and B is almost 10 cm. A noise source is placed about 1.5 m away from the test person.

4.2. Binary Masks Estimation

IBM is one goal of CASA system. Thus, the proposed algorithm is evaluated by SNR estimation and IBM comparison.

1) SNR estimation.

The main principle of IBM is to calculate the SNR of each T-F units. We use the DMED to estimate the SNR in each T-F units.

The actual SNR of the mixture is 0dB with babble noise. The true SNR is calculated by the target signal and far noise signal directly. The predicted SNR is calculated by the equation (13), and the $DMED_S$ is 100, $DMED_N = 1$. As shown in Fig. 3, the DMED based algorithm provides a good estimate (prediction) of the true SNR value.

2) IBM estimation.

The similarity between IBM and the binary masks is calculated as classification accuracy:

$$\text{Accuracy} = \frac{\sum_t \sum_f (DM(t,f) == M(t,f)}{\sum_t \sum_f 1} \times 100\%. \tag{21}$$

where $M(t,f)$ refers to the binary masks generated by equation (14). $DM(t,f)$ refers to the binary masks generated by the algorithm proposed as equation (17), where γ is equal to zero. $\sum_t \sum_f 1$ is the number of total units. The variable t and f, indicates the time frame and frequency channel of the T-F units. A higher accurate would result in better separate performance.

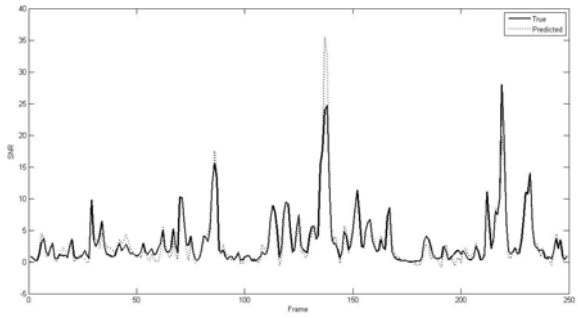

Fig. 3. Comparison between the true SNR values and its' predicted values in T-F units. The channel center frequency of the T-F units is 1000 Hz.

The similarity of the binary mask between the two algorithms is shown. Four types of noise signal are used to generate the mixture signals, which SNR levels various from -30 dB to 30 dB at -5 dB intervals. The accuracies are more than 95 % in all conditions.

The differences between the IBM and DMED-based binary masks are less than 5 % in all conditions. The cue of DMED is robust in different SNRs, especially in higher or lower SNR conditions. Fig. 4 shows that DMED performs better with machine gun noise than with babble, si762, and m109 noise. Stronger correlation between target source and noise, a larger effect of the additional factor $2\|S_A(t,f)\|\|N_A(t,f)\|\cos\theta_A$ and $2\|S_B(t,f)\|\|N_B(t,f)\|\cos\theta_B$, and greater difficulties in separating the mixture signals.

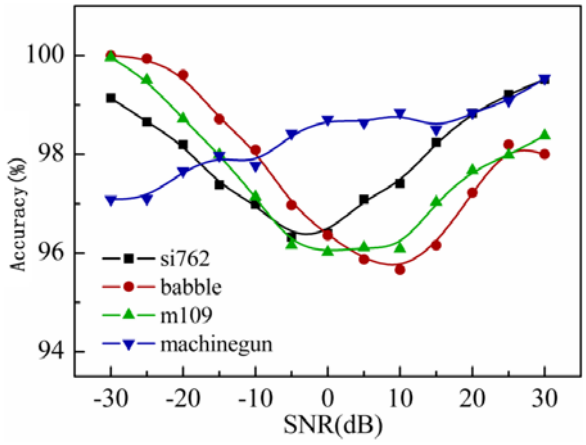

Fig. 4. Accuracy of the DMED-based binary mask.

3) System Performance with various lengths of T-F units

The performance of the proposed method with different lengths of T-F units is given in Fig. 5. Four types of noise and speech "sx198" were used to generate the mixture signal at the SNR level of -5 dB.

By increasing the frame length from 2 ms to 256 ms, Accuracy is increased as well. The best performance is obtained at 256 ms above 97 %. Given the T-F units' increase in length, the correlation between signal and noise are decreased. The smaller the value of $2\|S_A(t,f)\|\|N_A(t,f)\|\cos\theta_A$, $2\|S_B(t,f)\|\|N_B(t,f)\|\cos\theta_B$.

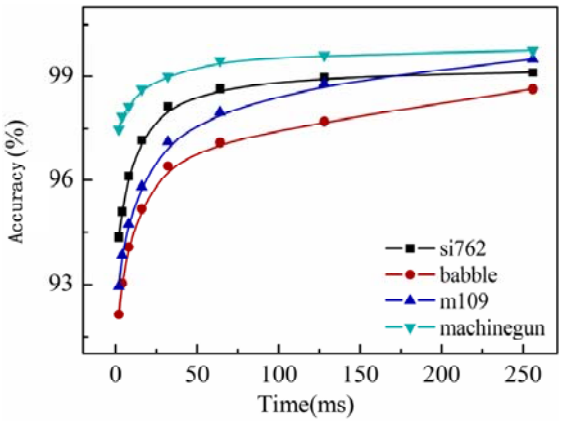

Fig. 5. DMED Performance with various T-F units' lengths.

4.3. ASR Performance with Actual Recordings of a Dual-microphone System

The training dataset is from the standard Mandarin speech database collected under the state-sponsored 863 research program, which involves 127 hours of reading speech data. The test data consist of recordings of two male speakers and one female speaker, which collected in office rooms with babble noise 1.5 m away from the speaker. Each speaker speaks 600 short Chinese utterances involving 200 Chinese names, 200 stock names and 200 Chinese place-names. The acoustic model of the ASR baseline system is based on the structure of GMM-HMM and cross-word mono-phones modeled in 3 states left-to-right HMMs. Each state density is 10 component Gaussian mixture models with diagonal covariance. The baseline acoustic model is trained by the standard HTK3.4 toolkit.

The two microphones system was used to collect the signal as section 4.1. We got 3734 test sentences.

Table 1 shows results of ASR accuracy over 3734 sentences. For this evaluation, the SNR of the mixture signals are from -5 dB to 20 dB with babble, m109 and single speech noise. The sentence accuracy and word accuracy is improved almost 10 % as average by the proposed algorithm. The wiener and spectral subtract algorithm has the lower accuracy, and they would damage the target speech when remove the noise. The dual-microphone PLD algorithm improves the ASR accuracy with the coherence between two microphones.

Table 1. ASR accuracy (%) of the actual recordings.

Algorithm	Sentence Accuracy (%)	Word Accuracy (%)
Original Mixture	64.70	70.78
Spectral subtract [12]	63.63	71.32
Wiener [13]	52.89	61.09
PLD [10]	69.04	76.20
Proposed	73.67	80.28

In Fig. 6. The SNR is estimated from the mixture signal of microphone A. The data of wiener and spectral subtract is got from the close microphone A. power level difference based Dual-microphone algorithm is named as PLD.

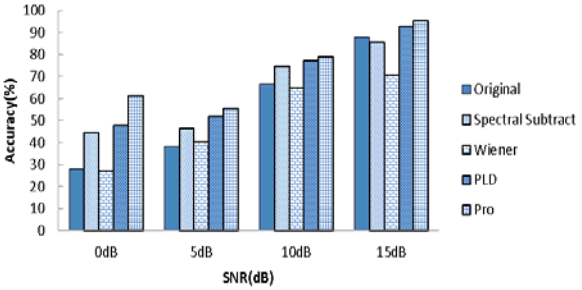

Fig. 6. Recognition accuracy with babble noise.

We observe the proposed algorithm outperforms the single channel wiener and spectral subtract algorithm and the dual-microphone PLD, especially in low SNR conditions. The proposed algorithm can improve the intellective of target speech in noisy environments.

6. Conclusions

An extended algorithm to separate the target speech from far noise is proposed. Compared with the IBM for single microphone, the DMEDs can be used to obtain the optimal binary masks for two microphone systems. Systematic evaluation shows that the proposed algorithm based on DMED performs similarly well to the IBM. In all conditions, the accuracies are more than 95 %. Better performance can be obtained by increasing frame length, which would be a problem in the real-time application. ASR test shown that the proposed algorithm performance better than the other system in babble noisy environments. Obtaining DMED of the target sound

and noise is the key point. Fortunately, in the close-talk system, the great difference of DMED between the close target speech and far noise sound source make it simplify. More work should be done to get more accurate DMED value to improve the performance of this algorithm.

References

[1]. Chao Ling Hsu, De Liang Wang, J. S. R. Jang, Ke Hu, A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No. 5, 2012, pp. 1482-1491.

[2]. Narayanan, A., Xiaojia Zhao, De Liang Wang, Fosler-Lussier, Robust speech recognition using multiple prior models for speech reconstruction, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2011)'*, Prague, Czech Republic, 22-27 May 2011, pp. 4800-4803.

[3]. Xiaojia Zhao, Yang Shao, De Liang Wang, CASA-Based Robust Speaker Identification, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No. 5, 2012, pp. 1608-1616.

[4]. G. J. Brown, and Martin Cooke, Computational auditory scene analysis, *Computer Speech And Language*, Vol. 8, No. 4, 1994, pp. 297-336.

[5]. Yi Jiang, Hong Zhong, Zhenming Feng, Performance analysis of ideal binary masks in speech enhancement, in *Proceedings of the 4th International Congress on Image and Signal Processing (CISP' 2011)'*, Shanghai, China, 15-17 October 2011, pp. 2422-2425.

[6]. Guoning Hu, Deliang Wang, A Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 8, 2010, pp. 2067-2079.

[7]. Guoning Hu, Deliang Wang, Auditory Segmentation Based on Onset and Offset Analysis, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 2, 2007, pp. 396-405.

[8]. N. Roman, D. L. Wang, and G. J. Brown, Speech segregation based on sound localization, *Journal Of The Acoustical Society of America*, Vol. 114, No. 41, 2003, pp. 2236-2252.

[9]. N. Yousefian, A. Akbari, and M. Rahmani, Using power level difference for near field dual-microphone speech enhancement, *Applied Acoustics*, Vol. 70, No. 11, 2009, pp. 1412-1421.

[10]. N. Yousefian, and P. C. Loizou, A Dual-Microphone Speech Enhancement Algorithm Based on the Coherence Function, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 2, 2012, pp. 599-609.

[11]. F. Kallel, M. Frikha, M. Ghorbel, A. Ben Hamida, and C. Berger-Vachon, Dual-channel spectral subtraction algorithms based speech enhancement dedicated to a bilateral cochlear implant, *Applied Acoustics*, Vol. 73, No. 1, 2012, pp. 12-20.

[12]. D. S. Brungart, W. M. Rabi Nowitz, Auditory localization of nearby sources: Head-related transfer functions, *Journal of The Acoustical Society of America*, Vol. 106, No. 3, 1999, pp. 1465-1479.

[13]. D. O. Kim, B. Bishop, S. Kuwada, Acoustic Cues for Sound Source Distance and Azimuth in Rabbits, a Racquetball and a Rigid Spherical Model, *Jaro-Journal of the Association for Research in Otolaryngology*, Vol. 11, No. 4, 2010, pp. 541-557.