2017 5th International Conference on Mechanics and Mechatronics (ICMM 2017)
ISBN: 978-1-60595-541-4

Using an Adjustment Training and a Smoothing Mask for Speech Segregation

Yi JIANG[1,a], Run-Sheng LIU[2] and Yuan-Yuan ZU[1]

[1]*Quartermaster Equipment Research Institute, CPLA, 100082 Beijing, China*
[2]*Department of Electronic Engineering, Tsinghua University, 100084, China*

Abstract. This paper focuses on the improvement of speech intelligibility and nature auditory perception. A dual microphone computational auditory scene analysis (CASA) based speech segregation system is proposed. A deep neural network (DNN) is equipped to estimate the parameter mask, which is used to train a smoothing mask to segregate the target speech from the mixture. A mask smoothing method is proposed to reduce the musical noise, which is caused by estimation errors. The performance of the proposed method is systematic evaluated with the simulated and recording data. The tests show that the proposed method improves the signal to noise ratio (SNR), suppress the musical noise, and has good performance on untrained locations and reverberant test conditions too.

1 Introduction

Talking and speech is the most convenient and efficient way of communication. But the speech is always damaged by noise around. The hearing is also very sensitive of the noise. The long periods of noise occur time also decrease the ability of human hearing. In most applications, the aim of speech enhancement is to improve some perceptual aspect of the degraded speech, and try to maintain the energy of the target speech. More and more, the speech quality such as intelligibility improvement is highly desirable especially when the listener is exposed to high levels of noise. The naturalness of speech is also discussed too.

Speech enhancement algorithms which reduce or suppress the background noise are introduced as the traditional method, and get big success in noisy signal processing. With the primary goal of improving speech quality, spectral subtractive algorithms, statistical-model-based algorithms, subspace algorithms and binary mask algorithms have been proposed in the literature for speech enhancement [1]. There are two main targets for speech segregation, which include reducing the noise distribution and containing the target information. One application is for automatic speech recognition system, the other one is for human listening or hearing aids. To human listener, the speech intelligibility and the listening comfortable is the same important. With deep neural networks, the research simulates the speech enhancement processing as human hearing system do. Also human auditory system is more complex than our understanding, and powerful to deal with noise environments. To a human listening system, it is more desirable to enhance the intelligibility and a comfortable level rather than only improving the quality of speech.

With the ideal of computational auditory scene analysis (CASA), the signal processing simulates human listening. The speech enhancement system focus on the domination part in the mixture as human hearing system does [2]. The main sound energy can mask the other sound as the auditory scientist found [3]. A binary mask is used to label a sound segment as 1 the target signal, 0 else the interference. A noise level depend parameter mask is also used to estimate the probability of the target speech in the mixture. The CASA based methods get success on speech enhancements, not only signal to noise ratio improvement, but also speech intelligibility for normal hearing and hearing-impaired listeners. The deep neural network (DNN) is always used as a classifier to estimate the binary mask 0 and 1, which get a great success not only on research but also applications [4].

With speech segregation system widely used in our daily life, more and more application pay more attention on the speech's naturalness for well listener feeling than traditional speech enhancement methods. Our prior work has proved the good performance of the DNN based speech segregation system [5]. There are two main problems in the future real applications. The first one is the unmatched training and using conditions. The second one is the remained musical noise, which will damage the feeling of sound and make sound heard uncomfortable. For the DNN and CASA based masking methods, the remained noise is always boring and hardly calculated by the signal noise ratio (SNR). In some application cases, this musical noise can be more disturbing to the listener than the original distortions caused by the interfering noise. Our prior research also finds the binary and

[a]Corresponding author: jiangyi09@tsinghua.org.cn

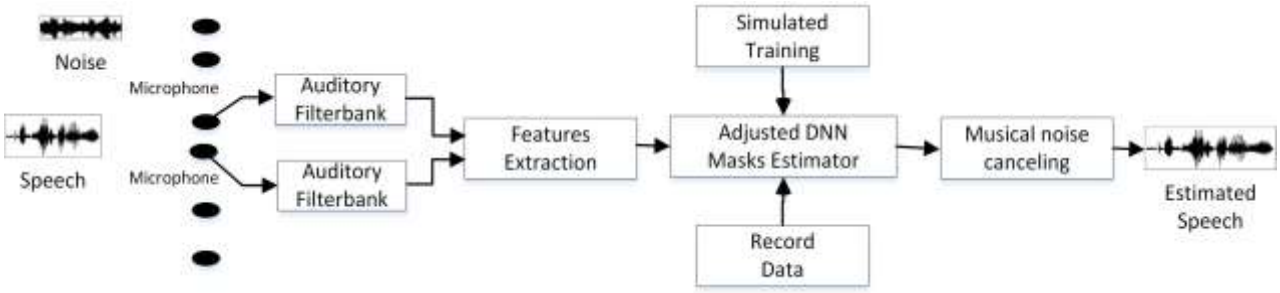

Figure 1. Schematic diagram of proposed algorithm.

parameter mask leads to some uncomfortable remained noise. In this study, we try to reduce the so called musical noise distortion at the same time. In this paper, the proposed system uses the estimated masks to detect the voice absence, and reduce the music noise with human hearing masking effect.

The rest of this paper is organized as follows. In Section 2, we give an overview of the proposed online speech segregation system. The experiments hardware, software, training and test data set is also introduced. In Section 3, we describe the adjustment training method. The smoothing mask generate method is introduced too. The evaluation is given in Section 4. Finally, we conclude this work in Section 5.

2 System Overview

In Fig. 1, the proposed dual microphone system uses two microphones to collect the noise and speech at the same time. The distance between two microphones is flexible, and can be changed with applications requirement. The size of the system is also flexible for different use. Two auditory filter banks are used to decompose the input signal to 64 frequency channels from 50Hz to 8000Hz. We use the gamma tone as the auditory filter bank to simulate the human auditory system; more detail can be found in [5]. Then a 20ms time frame with 10ms overlap is employed in each frequency channel to segregate the signal to time-frequency (T-F) units. All fellow signals processing such as feature extraction and energy masking is based on T-F units.

The dual microphone feature (indicates as the DF), as the interaural time difference (ITD), interaural intensity difference (IID), and monaural feature gammatone frequency cepstrum coefficient (GFCC), is extract from each T-F pairs, which is used as the cues to separate the target speech from the interference.

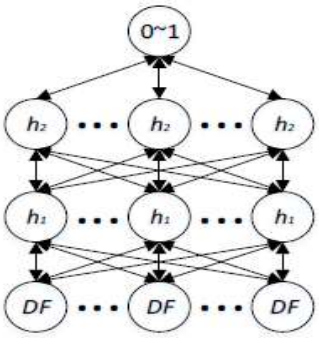

Figure 2. The DNNs with two hidden layers.

These dual microphone features are used as the input of the DNN masks estimator, which is first trained with simulated data, then adjustment training with recording data specifically. We used a DNN estimator with two hidden layers as shown in Fig. 2. Each hidden layer has 256 nodes named as h1 and h2. The output layer has only one node named pm. It is the estimated parameter mask, and is a real number. The parameter mask indicates energy proportion of the target speech in the mixture.

Finally, a mask smoothing method is proposed for musical noise cancelling, and in order to improve the speech comfort level for human listener. The principle ideal is deleting the isolated noise and contains the noise with little influence to the target speech.

2.1 Experiments Hardware

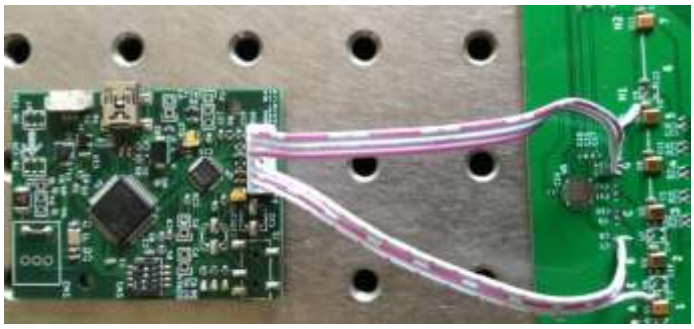

Figure 3. Hardware of the dual microphone system.

A six microphones matrix is shown in Fig. 3, is designed as the input part. The same type silicon microphones are used with high quality and consistency. The distance between these microphone are from 1 cm to 6 cm with 1cm interval. Two of the six microphones is choice each time to build the dual microphone system.

Two microphones can be selected each time to setup various dual microphone systems for different applications.

Two high speed signal processing chips sample the signal, and transfer the data to computer with mini USB port. The sample rate sets to 16 kHz, and can up to 96 kHz. We separate the microphone board from the main processing board for the easy and flexible configuration and reduce the signal disturbance.

2.2 Training and Test Configurations

The software platform is based on MATLAB software. The data preparing part is employed to generate the training and test data by simulated method or recording directly. The training program provides the monaural and binaural feature extraction codes, DNN training codes. The test program has the recording codes, DNN estimator codes, and finally output the enhanced target speech. We use MATLAB to analyse the smoothing filter performance. We also use the SNR calculate codes to evaluate the performance of the system.

We train the DNN in two scenarios on 0 dB conditions. In the simulated training scenario, we generate binaural mixtures that simulate pickup of multiple speech sources in a quiet office by the ROOMSIM package [6]. The TIMIT [7] corpus and NOISEX92 [8] is used as the target speech and interfere noise at the same time. We place the target speech on 0 degree. Then the interference sources are systematically varied between 10 degree and 350 degree, except 180 degree as the symmetry location of 0 degree, spaced by 10 degree to simulate different noise.

In the recording scenario, the target speech is fixed on 0 degree, and the interference location is changing from 10 degree to 350 degree except 180 degree with 10 degree step. Then the dual microphone system records the speech and noise separately. The mixture signal is generate by the recording noise and speech with prior setting SNR. This is an addictive noise with a small reverberant of the office.

The test dataset comes from another recording data, with the SNR various from -10 dB to 10 dB. It contains 50 sentences.

3 Adjustment Training and Smoothing Masks

3.1 The Adjustment Training Method

We train the DNN with 150 scenes random choose from the two scenarios training datasets. The sentence number choose from the simulate scenario is 150, 140, 120, 100, 50 and 0. The remains scenes are chosen from the recording scenario.

In the test processing, the data comes from the online recording. There are 50 test sentences. The output of the proposed system is used to calculate the SNR improvements, and evaluate the performance of the proposed system. The test results are list on Table 1.

On all conditions, the proposed dual microphone speech enhancement system gets positive results, as describe on our prior work [9]. With the adjustment training, the more recording data including, the better performance on test recording dataset. The gap is about 2.5 dB on 0 dB conditions. So the more convenient method to transfer the laboratory system to real application is training the system with extra real data. The results also show the strong generalization ability of the DNN method.

Table 1. The performance of the adjust method.

Training Data	Input SNR				
	-10	***-5***	***0***	***5***	***10***
150+0	-1.56	0.61	3.20	6.50	10.82
140+10	-0.81	0.84	3.40	6.76	10.49
120+30	-0.78	1.03	3.79	7.23	10.90
100+50	-0.77	1.19	4.09	7.54	11.01
50+100	-0.50	1.37	4.34	7.77	11.02
0+150	0.80	3.07	5.70	8.59	11.94

3.2 The Smoothing Mask

Based on human auditory masking effect, the much noise can be retained follow by the target speech. On the other sides, the noise should be suppressed before the target speech. With this principle, we propose the smoothing method to improve the target speech comfortability and reduce the musical noise.

We introduce a target speech active detection method by the estimated masks. Then the masks are smoothed to reduce the musical noise distortion. The smoothed masks provide the information of the speech energy in the noisy speech, and are used to recover the target speech.

The musical noise is smoothed by two steps. We first make sure the voice or speech is absence. Then the suppression or keeping is decided. The speech absence is calculated on each T-F units by Equation 1.

$$V(f,t)=\begin{cases}1 & emask(f,t)>=0.5\\ 0 & emask(f,t)<0.5\end{cases} \tag{1}$$

The v indicates the voice absence or not in frequency channel f and time frame the mask is the mask value, which is the output of the DNN estimator. The equation chooses these T-F units, which cantinas target speech energy more than the noise. We count the *V (f,t)* number on time frame t of all 64 frequency f, then calculate the probability of voice absence $V_P(t)$ with Equation 2.

$$V_P(t)=\sum_f V(t,f)/64 \tag{2}$$

Then the speech absence in time frame t is calculated by $V_A(t)$ as follows Equation 3.

$$V_A(t)=\begin{cases}1 & V_P(t)>=0.6\\ 0 & V_P(t)<0.6\end{cases} \tag{3}$$

In this method, we estimate the voice absence by the target speech energy in all frequency channels. It is an energy based voice active detection (VAD) method, which can be applied very easily.

A novel low-pass filter is introduced. The basic ideal of the smoothing filter is The amplitude-frequency curve of the filter is shown in Fig. 4.

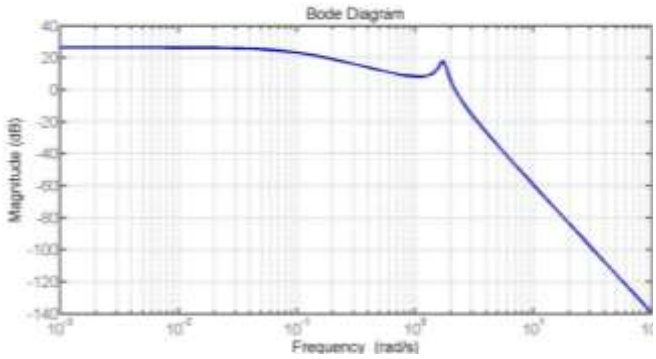

Figure 4. The Bode diagram of the mask smoothing filter.

The basic ideal of the smoothing filter is reducing the musical noise and retaining the noise fellow the traget speech to improve the intelligibility and comfortable level of the speech.

4 Evaluations with Recording Data

In this test, we use a speaker to broadcast news around the target speaker as the interference. It simulates common application scenes.

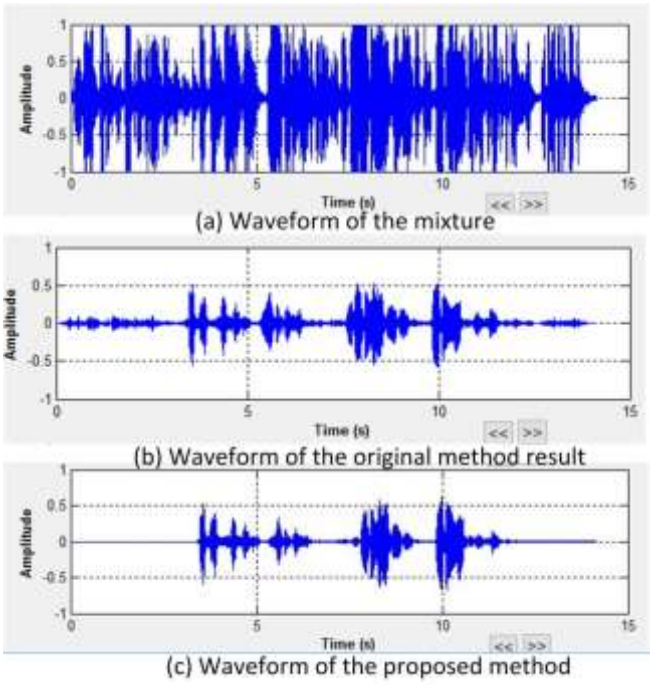

Figure 5. The signal waveform results.

As show in Fig. 5, the mixture is full of noise. It is very difficult to separate the target speech to the noise. As the interference is similar to the target speech, it is hard to be deal with traditional method, such as the spectral subtraction or wiener filter method. With DNN, The original method extracts the target speech from the mixture, as shown in (b). The proposed system has less musical noise in (c) than in (b). The musical noise in (b) is very small in magnitude, but makes the listener uncomfortable. As shown in Fig. 6 (b), the proposed speech enhancement remains most of the target speech energy. We can recognize the F0 and the onset offset of the speech. As shown in the waveform and spectrum, the proposed method suppresses the noise directly and effectively. The result is more nature listening than the results without adjust DNN and smoothing mask.

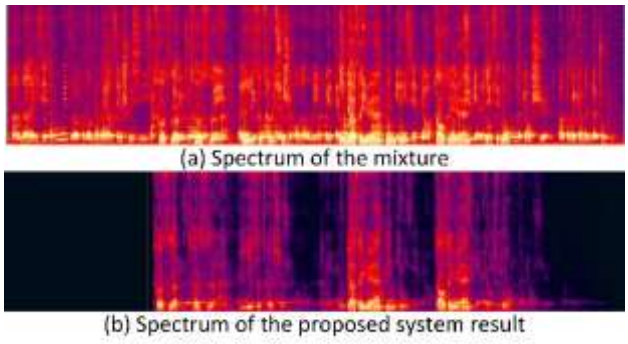

Figure 6. The spectrum of the mixture and the proposed method results.

Conclusions

In this paper, we introduce a new method to apply the laboratory speech enhancement system to real operating system. The big problem in the realization not only the environments difference between the trained conditions and real application, but also the hardware difference. The big gap between the laboratory simulate data and the real world application data will limited the performance of the speech enhancement system, especially to machine learning based system.

With adjustment training and smoothing mask, the proposed method success adopt from the simulated system to real application. The smoothing mask reduces the musical noise significantly. The proposed method balances the speech intelligibility and the musical noise. It improves the naturalness of the segregated speech at the same time, and makes the target speech hearing comfortable.

We will take more effort on reducing the method complexity, train more data, and improve the performance of the dual microphone speech enhancement system in the future research.

Acknowledgment

The work was supported in part by Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).

References

[1] Loizou P C. Speech Enhancement: Theory and Practice. CRC press (NY, 2013).

[2] D.L. Wang, G.J. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications (Wiley/IEEE Press, Hoboken, NJ, 2006).

[3] Bregman A S. Auditory Scene Analysis: The Perceptual Organization of Sound. (MIT press, 1994).

[4] K. Han, and D.L. Wang, IEEE Trans. Audio Speech Lang. Process., **21**, 166-175 (2013).
[5] Y. Jiang, et al., IEEE Trans. Audio, Speech, Lang. Process., **22**, 2112-2121 (2014).
[6] D. R. Campbell, The ROOMSIM User Guide (v3.3). Available: http://media.paisley.ac.uk/campbell/Roomsim/
[7] Garofolo J S, Lamel L F, Fisher W M, et al. DARPA TIMIT acoustic phonetic continuous speech corpus. (1993).
[8] Varga A, Steeneken H J M. Speech Commun, **12**, 247-251 (1993).
[9] Y. Jiang, W. Liang, Y. Zu, H. Zhou, Z. Feng, and Q. Chen, J. Tsinghua Univ., **52**, 636-641 (2012).