MATEC Web of Conferences **31**, 17004 (2015)
DOI: 10.1051/matecconf/20153117004

Parameter masks for close talk speech segregation using deep neural networks

Yi Jiang [1, a], Runsheng Liu [2]

[1] *The Quartermaster Equipment Research Institute, CPLA, Beijing, China*
[2] *Electronic Engineering, Tsinghua University, Beijing, China*

Abstract. A deep neural networks (DNN) based close talk speech segregation algorithm is introduced. One nearby microphone is used to collect the target speech as close talk indicated, and another microphone is used to get the noise in environments. The time and energy difference between the two microphones signal is used as the segregation cue. A DNN estimator on each frequency channel is used to calculate the parameter masks. The parameter masks represent the target speech energy in each time frequency (T-F) units. Experiment results show the good performance of the proposed system. The signal to noise ratio (SNR) improvement is 8.1 dB on 0 dB noisy environment.

1 Introduction

The noise is a big problem in speech communications and sound collections. It distorts the clean speech and reduces its intelligence, and make the speech hard to understanding. How to get the clean target speech in noisy is always a challenge task. We usually use close talk microphone to collect the speech with small distance between the microphone and mouth. Such as mobile phone, headset microphone and so on, they can get better result than faraway microphone. But the close talk system also has bad performance on noisy environments.

While human listening has excellent speech tracing performance on noisy conditions. It separates the target speech and tracks it. With the development of auditory scene analysis (ASA), scientists conclude and simulate the ability of hearing system with the computational auditory scene analysis (CASA) algorithm [1]. In this frame work, the auditory signal is first decomposed to various frequency channels and generated as the time frequency (T-F) units. The auditory features are extracted from T-F unit to discriminate the target speech and environment noise, such as onset/offset, pitch, MFCC and GFCC. The relationship between the features and the sound sources is hard to describe. The supervised and unsupervised methods of the machine learning algorithms are used to train these algorithms [2]. With the rapid development of deep neural networks (DNN) [3], it provides another method for such speech enhancement system, and gets good results in binary classification systems. In this paper, the DNN [4] is used as a signal to noise ratio (SNR) estimator to calculate the energy in each T-F unit with a parameter mask. The calculated SNR value is used to generate the parameter mask for target speech resynthesize.

In the following section, we present an overview of the parameter mask speech segregation algorithm for close talk system. Section 3 describes the auditory features extraction, labels automatic generation and DNN estimator training. We present several experiments in Section 4 and conclude the paper in Section 5.

2 System overview

The proposed close talk parameter masks algorithm consists of four stages. As shown in Figure 1. Two microphones are used to collect the target speech and noise simultaneous. The mouth is closer to one microphone than another microphone. The noise is far away from the two microphones.

In the first stage, the same two auditory filter banks are used to decompose the inputs into T-F units respectively. A gammatone filter bank is used as the auditory filter bank to simulate frequency decomposes processing of the human hearing system. The gammatone filter bank has simple time domain form [5]. In this paper we use a 64 channels gammatone filter banks, which central frequencies are equally distributed on the Bark scale from 50Hz to 8000Hz. A time frame is used to separate the output signals to T-F unit. A T-F unit corresponds to a certain frequency and time frame signal. A rectangle window is used, which has 20ms length and 10ms overlap. In this way, we represent the input signal with a matrix T-F units.

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

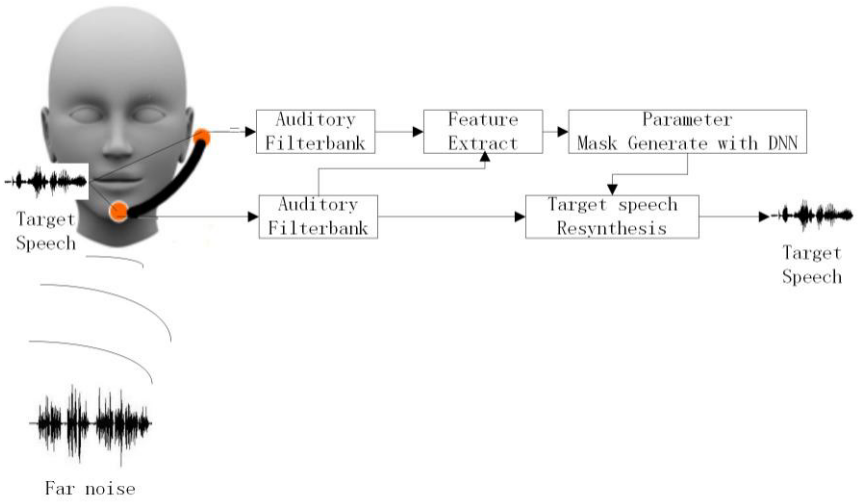

Figure 1. Schematic diagram of the parameter mask speech enhancement

In the second stage, the two-microphone features are extracted from the two microphone T-F unit pairs, which are be used to estimate the parameter mask. The features include the time and energy difference between two microphones. These features are indicate the locations difference between the target speech and noise, which are used as the inputs of the DNN.

In the third stage, a DNN estimator is used on each frequency channel to calculate the parameter mask in each T-F unit with the binaural features. The SNR of the T-F unit means the ratio of the target speech energy in the mixture. The estimated SNR is used to generate a parameter mask, and separate the energy of the target speech from the mixture.

In the final stage, the parameter masks are used with the close mixture signals to resynthesize the target speech.

3 Auditory feature and DNN estimator

The auditory feature is the cue to separate the target speech and noise. The relation between the auditory feature and the target energy is complex. We use the DNN to learn and describe it.

A 64 channels gammatone cochlear model is used to decompose the input signal to various frequency channels signal as the human auditory filter bank do. It is first proposed by Roy Patterson to simulate the hearing function in time domain. Fellow the ideal of interaural intensity difference (IID) and interaural time difference (ITD) in general two microphones systems, the energy and time difference between the two microphones in close talk system is introduced as dual microphone energy difference (DMED), dual microphone cross-correlation (DMCC). The DMED is a 2 dimensions feature, and the DMCC is a 32 dimensions feature.

Above all, the 34 dimensions auditory feature (DMED+DMCC) is extracted from each T-F unit pair, and used as the input of the DNN estimator to estimate the SNR.

3.1. Parameter masks

The speech and noise is additive and uncorrelated. The mixture is described as Eq.1.

$$x=s+n \tag{1}$$

the x indicates the mixture signal. s indicates the target speech signal. n indicates the noise signal.

In order to estimate the speech energy in T-F unit, the ideal parameter masks are calculate from the SNR directly to get energy ratio of the target speech in T-F unit as Eq.2.

$$\mathrm{pm}(c,m)=\sqrt{\frac{S(c,m)}{S(c,m)+N(c,m)}} \tag{2}$$

The pm indicates parameter masks, the c, m indicates the frequency channel and time frame respectively. While S is the speech energy of the target speech. N indicate the noise energy.

In training stage, the ideal parameter masks are calculated by the known target speech and noise energy in T-F unit with simulated training data.

3.2. DNN estimator

A DNN architecture with two hidden layers is used to estimate the parameter masks by the auditory feature, which is illustrated in lecture [3]. The DNN consists one real input layer, two hidden binary layers and a real output layer with one neuron.

In this paper, the 34 dimensions real two microphone auditory features of one T-F units are used as the DNNs real inputs. Each hidden layer has 200 binary hidden neurons. The output is a real neurons, which indicates the SNR in the T-F unit. We follow the approaches in [3] to initial the parameter of DNN, where the restricted Boltzmann machines (RBM) are used with unsupervised learning method together.

The labelled data are provided by the ideal parameter masks. The training goal is to maximize the accuracy rates, which indicates the difference between the labels and the estimated outputs. In test processing, the DNN estimator outputs the label for each T-F units as an estimated ideal parameter masks.

4 Experiments and results

An actual recording data collected with a dual-microphone system in an office environment is used to train the algorithm and evaluate its performance. The recording corpus record the target speeches, include 600 short Chinese utterances. There are 200 Chinese names, 200 stock names and 200 Chinese place names. All materials are collected with two male speakers and one female speaker in quiet office rooms. The babble, white, m109 and machinegun, from NOISE 92 databases is used as the noise. All noise sources are presented at a distance of 1.5 m from the target speaker, located at azimuth 0°, 45°, 90°, 135° and 180° unless otherwise specified.

The recording speech and noise data are split to two half parts. One half part for training, and other half part for test. In training stage, the noise signal is randomly pick up and mixture with speech to generate the 0dB SNR training data. The segment SNR of T-F unit is calculated to generate the ideal parameter masks, which is be used as the label for training and test. In the test stage, the speech and noise is used to generate the test corpus with various SNR from -15dB to 15dB.

Table 1. SNR improvement of the proposed system with various SNR inputs (dB).

Input SNR (dB)	**-15**	**-10**	**-5**	**0**	**5**	**10**	**15**
IPM	2.0	4.3	6.9	9.9	13.3	16.9	20.5
Proposed method	0.0	1.3	4.3	8.1	12.2	16.1	19.8

The performance of the proposed system is show in Table 1. The proposed method gets the positive results on all conditions. Even on -15dB condition, the algorithm gets 15.0dB SNR improvements. Compare to the ideal parameter masks, the proposed method is 2dB lower than the ideal parameter masks (IPM).

5 Conclusion

With DNN estimator, the SNR of each T-F unit is calculated, which is used to generate the parameter masks for the close talk system. The parameter masks indicate the energy of target speech in each T-F unit. Results with recording data show the parameter masks have good performance on low SNR conditions, even on -10dB condition. More test will do in the future work.

References

1. D.L. Wang, and G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley and IEEE Press, NJ, 22-120 (2006).
2. X. Yong, An experimental study on speech enhancement based on deep neural networks, IEEE Signal Processing Letters **21**, 1, 65-68 (2014).
3. Y. Jiang, D.L. Wang, Rusheng Liu, etc, Binaural classification for reverberant speech segregation using deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**, 12, 2112-2121 (2014).
4. G.E. Hinton, and R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science **313**, 5786, 504-507 (2006).
5. Y. Jiang, R.S. Liu, and Z.M. Feng, Dual-microphone speech enhancement algorithm based on the auditory features for a close-talk system, Journal of Tsinghua University **54**, 9, 1179-1183 (2014).