A Speech Enhancement Algorithm Using Computational Auditory Scene Analysis with Spectral Subtraction

Cong Guo, Like Hui, Wei-Qiang Zhang, Jia Liu
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
Email: wqzhang@tsinghua.edu.cn

Abstract—Computational auditory scene analysis (CASA) system is well used in speech enhancement area in recent years. We propose a new system that combines CASA and spectral subtraction to get better enhanced speech. The CASA part consists of the latest method deep neural networks (DNNs). The original way to reconstruct the denoise signal is to use the estimated masks with direct overlap-add method ignoring the information of noise within the frames. In our system, we estimate self-adapted thresholds for each channel by Gaussian Mixture Model from the estimated ratio masks (ERMs) to separate noise and speech of each channel. In this way, we make full use of the information within frames. The results show increase in both objective and subjective evaluation.

Index Terms—speech enhancement, computational auditory scene analysis (CASA), deep neural network (DNN).

I. INTRODUCTION

Speech technology is becoming a hot study topic in recent years. With the developing of computer science and Internet, the methods are more efficient and the data is enough. Especially after Hinton [1] improved the training algorithm of neural network, almost the whole pattern recognition studies benefit from this.

As for speech enhancement, it is one of the pre-processing module of most of speech techniques. And monaural speech enhancement is the basic research topic of it. With the improving of hardware and software ability, more questions are required nowadays. We focus on improving the speech quality and intelligibility at low SNR, multi-noise situations.

To achieve this goal, ITU [2] proposed an idea with double stages Wiener filter which became a standard in speech communication. But the system is not robust because of the inaccurately of noise spectrum especially at low SNRs. Wang et al. [3] provided an computational auditory scene analysis (CASA) algorithm which simulates the hearing system of humans based on pattern recognition theory. It treats input audio as different streams from different sound sources and separates them with masks on each time-frequency (T-F) unit. And the key point of CASA system is to estimate the masks. With the improving of DNNs, recent studies sent lots of acoustic conditions into training sets to get the masks more accurately. Based on CASA, Narayanan et al. [4] changes the input labels from ideal binary mask (IBM) to ideal ratio mask (IRM) and gets better results. Wang et al. [5] proposed a new audio features multi-resolution cochleagram (MRCG) which contains short-time, long-time and multi-channel information. Yong et al. [6] did all the work in original FFT frequency domain and used DNNs to estimate the frequency domain characteristics directly with the mean-square error (MSE) function.

Our base line system is made by the CASA with DNNs. It is because the data-driven method improves the accuracy rate of estimate masks from the training data sets. Additionally, we replace the traditional DNNs output layer softmax function by sigmoid function in order to solve the estimation problem within one DNN model. This modification reduces the training time and training data spaces.

But the CASA algorithm has a natural weakness. The traditional reconstruct way is to multiply the noisy signal with the masks and add the different channel units in the same frame together. If the mask is larger than 0.5 which belongs to speech unit, the noise residual is kept after multiplying. And if the mask is smaller than 0.5, the speech information in this unit is restrained in this unit. So even the estimate mask is 100 percent correct which is the upper bound of this system, the evaluation results are still not that satisfied.

To overcome this problem, we consider the traditional signal processing ideas which could recover each T-F unit more precisely. One of the idea is using spectral subtraction to get the denoise spectrum for each T-F unit. This requires the estimated noise spectrum for each T-F unit which are the key point at this algorithm and influence the result a lot. So we take advantages of the connection between the IRMs and SNRs and use Gaussian Mixture Model to produce the self-adapted thresholds of ERMs for each channel to separate speech units and noise units. And with the short-time stability of the noise, the noise spectrum for each units could be provided more precisely.

The following of the paper consists four parts. In Section 2, we introduce the DNN-based CASA. Next, we aim at the reconstruction part of the CASA and give the proposed system which combines the CASA and spectral subtraction. Section

This work was supported by the National Natural Science Foundation of China under Grant No. 61370034, No. 61273268 and No. 61403224, and in part supported by China Scholarship Council.

The corresponding author is W.-Q. Zhang.
4 presents the experiment parameters and the results. Section 5 concludes the paper.

II. DNN-BASED CASA SYSTEM

The DNN-based CASA tends to estimate the mask for each speech unit and synthesis the audio. I will list the process of our DNN-based CASA system. First audio will be separate by channel and frame. Every audio will be passed through a 64-channel gammatone filterbank to sub-band signals.

Second we have to calculate the masks for each unit which could be IBM or IRM. In paper [4], the result indicates that using ERM to reconstruct the noisy wav is better than EBM in most of the evaluation. And the IRM mask is defined as [7]

$$IRM(m, c) = \frac{S^2(m, c)}{N^2(m, c) + S^2(m, c)}$$ (1)

where $S^2(m, c)$ and $N^2(m, c)$ are indicated the speech power and noise power of the unit at frame $m$ and channel $c$.

Using IRM we could keep more speech information. So in this paper, we use the IRM as the mask. In order to estimate the IRM more correctly, we build a deep neural network.

In test stage, we load the trained network model to estimate the IRM and synthesize the denoised signal using overlap add method.

A. DNN Architecture Description

The original DNN for ASR mostly creates by four kinds of layers. The input layer connects the input features to the hidden layer. The hidden layer is a long vector for mathematical transformation. Usually the neural networks are trained for pattern recognition problems. So the output layer is used to be Softmax function which let the probability of each node be reasonable. These three are full connected layer. The last is the nonlinear layer which is used to be sigmoid, ReLU or tanh functions.

In speech enhancement situation, the initial design was to match the sub-band numbers with the network model numbers. So each channel has its own DNN model. The benefit is easy to see that for each net it’s only a simple zero-one classification and the accuracy is pretty high. But on the other hand, this design have several weakness.

- The input unit has already passed a filter. The traditional MFCC or FBank feature does not fit very well.
- When using DNNs to solve speech enhancement problem, we could see the results benefit from the two dimension pooling layer design. But the connection between channels is unused.
- Similar with above, the numbers of net model increase with the sub-band filter numbers and due to the black box effect of DNNs, they can not combine. The training time and space is proportional to the model numbers.

So we modify the net architecture by using one network and remove original output layer. (a) The input unit is changed from a T-F input to a frame input. And the model’s label is changed from a 2 dimension IRM to a 64 dimension IRM vector. (b) Each IRM within the IRM vector is no longer a probability of this category. So Softmax layer is replaced by Sigmoid layer. Using Sigmoid but not ReLU is because the output value ranges from 0 to 1 which is the same with Sigmoid. (c) To increase the information in the features we extract, we use MRCG [5] feature so the short-time information, long-time info and interval channel info are all been processed into the neural network.

B. Overlap Add Method

After estimate the ratio mask, both [3] and [4] use a simple way to get the denoised speech unit.

$$y_{de} = \sqrt{ERM} \ast y$$ (2)

where $y$ is the noisy unit signal in time domain and $y_{de}$ is the denoised unit signal.

The ERMs are multiplied with the time domain signal of each unit directly. And the final output for each frame is adding the denoised unit of the same frame together.

And the $SNR$ for each frame could be

$$SNR_{de} = \frac{\sum ERM(m, c) \ast S^2(m, c)}{\sum ERM(m, c) \ast (S^2(m, c) + N^2(m, c))}$$ (3)

The spectrum of pure, noisy audio are demonstrated in Fig.1(a) and Fig.1(b). The denoised wav using (2) and ERM is shown in Fig.1(c) and the IRM ones is Fig.1(d). The results indicate that noise at the low frequency is moved clearly. But the middle frequency which is infected a lot is not recovered as well as the pure one. The noise residual is still maintained in the denoised audio.

III. CASA SYSTEM WITH SPECTRAL SUBTRACTION

As shown in Fig.1(d), the reconstructed audio with IRM has resumed the envelop of the pure one. But it still has a weakness in noise residual and speech wav spectrum structure. The reason why the upper bound of using IRM still has some problem is because the rough reconstruct algorithm. When using (2), the amplitude of each unit is increase and decrease at the same time. That is to say, when IRM is less than 0.5, the speech info in this unit is restrained and when IRM is larger than 0.5, the noise residual is still retained. This two effects are both not welcomed.

To reconstruct the denoised wav more accurately, we consider to introduce signal processing algorithm to replace the simple multiplication way. And one of the effective way is to reproduce the pure wav spectrum. So we use the spectral subtraction to handle each unit. And for each unit, we save the phase of the noisy unit and the amplitude of the denoised spectrum is the noisy spectrum minus the noise spectrum. But the amplitude of the noise spectrum of each unit is still unknown. Our framework is shown in Fig.2. And we have the following inference.

Firstly the noise within a small time period is basically stable. And when a person speaks in this time period, the SNR will fall when he or she stops talking. So the SNR at
least has two states, with one of it represents speech state and the other represents noise state.

Secondly the definition of SNR is

$$SNR = \frac{S^2(m, c)}{N^2(m, c)}$$

(4)

contact this with (1), there is

$$SNR(m, c) = \frac{IRM(m, c)}{1 - IRM(m, c)}$$

(5)

So after we get the ERMs from the neural network for each channel, we could also get the estimated SNRs for each channel.

Thirdly we use Gaussian Mixture Model to count the model for each state of SNR. And the model with the largest mean value denote the speech state. As a matter of experience, the threshold for speech and noise is

$$threshold = \arg\max \left( \frac{SNR_i}{\bar{SNR}_i} - \sqrt{\text{var}_i} \right)$$

(6)

where $\bar{SNR}_i$ and $\text{var}_i$ is the mean and var of the $i_{th}$ GMM model.

For classification, the units which their estimated SNRs are above the threshold are speech units and the rest of them are noise units. Fig.3 plots the self-adapted IRM threshold for a channel.

At last, for each unit, if it is noise unit, the noise spectrum of it is itself and the denoise unit of it is zero. If it is speech unit, the noise spectrum of it is the nearest noise unit spectrum.

The Fig.1(e) and Fig.1(f) indicate that the algorithm reduces the noise residual much. The recovered audio especially the IRM_GMM is very closed to the pure one.

IV. EXPERIMENT AND RESULT

A. Experiment setup

In our experiment, all the signals are resampled at 16 kHz rate and experiments are conducted on TIMIT database. We select 2000 utterances and other 400 utterances from the training folder of TIMIT as training set and development set. And we use the standard 192 utterances in TIMIT to produce the test set. All the data are added with 0dB and -5dB of babble, white, factory noise in NOISEX-92 [8] database. So the final number of each set are 12000, 2400, 1152.

And for adding noise, each noise is separated into two parts. The training set and dev set is mixed with the front halves of the noise at a random time point. And the test set is producing with the same way except the back halves of the interference.

All signals are framed by 20-ms windows and 10-ms frame shifts. The mask for each unit is generated in Gammatone
domain which is more similar with the response of person’s hearing system. And the number of filters designed for sub-band is 64 and the center frequencies span from 50 Hz to 8000 Hz equally. So in this way, the labels are produced according to the definition of IRM.

Due to sub-band of each unit, the original MFCC feature does not work well in NN net training. We prefer MRCG [5] features which contained short-time, long-time, spectral correlation information.

There will be four systems for comparison. ERM represents the base line DNNs CASA. IRM is the upper bound of base line which use the correct masks for reconstruction. ERM\_GMM is the proposed DNNs CASA with spectral subtraction. And IRM\_GMM is the upper bound of the modified system.

For the evaluation, the SNR, segmental Signal-to-Noise Ratio (segSNR in dB) and the log spectral distance (LSD in dB) are the original signal evaluation which reflect the relation between signals and noises. From the human’s hearing side, we choose Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) which are the objective speech quality and intelligibility score. Each input testing set is denoised with the above four systems and we will present their scores in the same sheet.

The results are listed from Table I, II and III. Each Table compares the output of different enhancement systems dealing with the same mixture audio set. All the results are the increment scores comparing with the noisy audio.

Considering the results of ERM and ERM\_GMM, the base line performs better in SNR, PESQ and STOI. ERM\_GMM acts well in segSNR and LSD. It’s because the definition of LSD has the same idea with spectral subtraction and segSNR is counted by frame which our subtraction is by frame and by channel.

And for the comparation of enhancement effect, we contrast the IRM\_GMM with the IRM. It’s observed that in 0dB, the PESQ and STOI of IRM are both higher than that of IRM\_GMM. But in -5dB, all the best evaluations are from IRM\_GMM system. Our proposed system is designed to estimate the noise spectrum precisely and the results indicate that it truly made it.

For the comparison of ERM and IRM, in 0db, the PESQ and STOI is about half of the maximum increase. The SNR, segSNR, LSD is a bit less than the upper bound. But the ERM\_GMM is not robust. The system is more sensitive. It indicates that our DNN needs to improve more to get better results.

Last, for different kinds of noises, the best system IRM\_GMM is higher in white and factory than babble. Because babble noise spectrum overlap much with the human voice spectrum. We are planning to import more kinds of noises to our NN net in the future.

### Table I

<table>
<thead>
<tr>
<th>Babble Type</th>
<th>SNR</th>
<th>segSNR</th>
<th>LSD</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM_GMM</td>
<td>-1.414</td>
<td>1.923</td>
<td>1.445</td>
<td>-0.632</td>
<td>0.019</td>
</tr>
<tr>
<td>ERM</td>
<td>1.776</td>
<td>0.374</td>
<td>1.619</td>
<td>0.192</td>
<td>0.038</td>
</tr>
<tr>
<td>IRM_GMM</td>
<td>1.751</td>
<td>2.219</td>
<td>2.952</td>
<td>0.224</td>
<td>0.085</td>
</tr>
<tr>
<td>IRM</td>
<td>1.989</td>
<td>0.614</td>
<td>2.058</td>
<td>0.388</td>
<td>0.119</td>
</tr>
<tr>
<td>ERM</td>
<td>2.004</td>
<td>0.828</td>
<td>1.844</td>
<td>0.080</td>
<td>0.053</td>
</tr>
<tr>
<td>IRM_GMM</td>
<td>3.726</td>
<td>3.678</td>
<td>4.776</td>
<td>0.331</td>
<td>0.166</td>
</tr>
<tr>
<td>IRM</td>
<td>2.641</td>
<td>1.536</td>
<td>2.586</td>
<td>0.348</td>
<td>0.144</td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>White Type</th>
<th>SNR</th>
<th>segSNR</th>
<th>LSD</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM_GMM</td>
<td>-0.862</td>
<td>1.701</td>
<td>8.133</td>
<td>-0.250</td>
<td>0.033</td>
</tr>
<tr>
<td>ERM</td>
<td>2.318</td>
<td>1.099</td>
<td>3.399</td>
<td>0.267</td>
<td>0.066</td>
</tr>
<tr>
<td>IRM_GMM</td>
<td>0.703</td>
<td>2.365</td>
<td>9.587</td>
<td>0.396</td>
<td>0.033</td>
</tr>
<tr>
<td>IRM</td>
<td>1.761</td>
<td>0.798</td>
<td>4.157</td>
<td>0.512</td>
<td>0.112</td>
</tr>
<tr>
<td>ERM</td>
<td>1.892</td>
<td>2.264</td>
<td>6.444</td>
<td>-0.422</td>
<td>0.056</td>
</tr>
<tr>
<td>IRM_GMM</td>
<td>3.915</td>
<td>3.658</td>
<td>8.224</td>
<td>0.468</td>
<td>0.179</td>
</tr>
<tr>
<td>IRM</td>
<td>2.963</td>
<td>0.728</td>
<td>4.842</td>
<td>0.449</td>
<td>0.152</td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Factory Type</th>
<th>SNR</th>
<th>segSNR</th>
<th>LSD</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM_GMM</td>
<td>-0.827</td>
<td>1.763</td>
<td>4.191</td>
<td>-0.432</td>
<td>0.042</td>
</tr>
<tr>
<td>ERM</td>
<td>2.731</td>
<td>1.552</td>
<td>2.656</td>
<td>0.192</td>
<td>0.052</td>
</tr>
<tr>
<td>IRM_GMM</td>
<td>1.248</td>
<td>2.303</td>
<td>5.532</td>
<td>0.4088</td>
<td>0.1</td>
</tr>
<tr>
<td>IRM</td>
<td>2.538</td>
<td>1.084</td>
<td>3.099</td>
<td>0.451</td>
<td>0.126</td>
</tr>
<tr>
<td>ERM_GMM</td>
<td>1.892</td>
<td>2.264</td>
<td>6.443</td>
<td>-0.429</td>
<td>0.057</td>
</tr>
<tr>
<td>ERM</td>
<td>2.956</td>
<td>1.116</td>
<td>3.125</td>
<td>0.123</td>
<td>0.034</td>
</tr>
<tr>
<td>IRM_GMM</td>
<td>4.023</td>
<td>3.247</td>
<td>8.417</td>
<td>0.468</td>
<td>0.168</td>
</tr>
<tr>
<td>IRM</td>
<td>3.084</td>
<td>0.926</td>
<td>3.829</td>
<td>0.403</td>
<td>0.154</td>
</tr>
</tbody>
</table>

### V. Conclusions

We propose a deep neural network CASA system with spectral subtraction to extract speech from mixture noisy audio. The base line which is deep neural networks CASA is stable when deals with different kinds of noisy data in low
SNRs. The DNNs softmax layer is replaced by sigmoid layer. Hence the training time and spaces is smaller 64 times than the original DNNs design.

But the CASA reconstruction part does not consider the connection between noise in time and even the ideal ratio mask could not restore the speech well. Our motivation is to make use of the time domain info by GMM and spectrum subtraction. The self-adaption threshold proposed by GMM divides the noise unit and speech unit. And from the results and spectrum, the spectral subtraction performs well in precise noise estimation and reduces the noise residual and keeps more speech information in the final outputs. The modification raises the upper bounds of the improvement in almost every evaluations.

But the new algorithm is using the spectral subtraction without any change which involves the familiar "music noise" problem. The DNN structure is also not very complex and could be improved more. The future works needs to integrate the resent studies of spectral subtraction and DNNs to overcome these.

REFERENCES