Speech Enhancement Based on Multi-Stream Model

Yan Xiong*, Qiang Chen*
Department of Computer Science，
Guangdong University of Education
Guangdong, China
e-mail: xiongy@gdei.edu.cn, cq_c@gdei.edu.cn

Fang Xu, Jun Zhang
School of Electronic and Information Engineering,
South China University of Technology,
Guangdong, China
e-mail: 897445793@qq.com, eejzhang@scut.edu.com

***Abstract*—In most of the current speech enhancement systems, speech signals collected by microphone are used as the only input data stream to recover the clean speeches, which will be greatly affected by the acoustic noise levels. Based on the fact that the noises or mismatches do not affect different data streams in similar ways, this paper proposes a new speech enhancement framework which can make use of multi-stream information even when some data streams are not directly related to the speech waveform by employing a multi-stream model based speech filter. A new speech enhancement method is also proposed based on the acoustic and throat microphone recordings. Experimental results show that the proposed method outperforms several conventional single stream speech enhancement methods in different noisy environments.**

Keywords-speech enhancement; multi-stream; model based; throat microphone

I. INTRODUCTION

In practical applications, speech signals are often corrupted by acoustic noises, which will degrade the quality and intelligibility of speeches. Therefore in many situations, speech enhancement systems are required to improve the speech quality, intelligibility or the performance of speech coding and speech recognition systems [1].

Current speech enhancement techniques can be broadly divided into two categories, i.e., the single-channel and multiple-channel approaches [2]. The single channel approaches process the speech signals received from single microphone in certain domains like the time, frequency or wavelet domain, which includes many classical speech enhancement techniques such as spectral subtraction [3], Winner filter [4], MMSE [5], and etc. On the other hand, the multi-channel approaches [6-7] require more than one microphone to exploit spatial information and separate the signal of interest from other interferences. The spatial filtering (or beamforming) is commonly used to form a beam towards the target signal. Comparing with the single-channel approaches, the multi-channel techniques can usually achieve better performance but with higher computational complexity and larger sizes. Therefore the single microphone speech enhancement techniques are still of wide interest in many applications [1].

However, in most of the current speech enhancement systems, clean speeches are recovered only from the signals collected by microphones. They are greatly affected by the acoustic noises therefore suffered from performance degradation in many situations. The multi-stream (MS) approaches [8-10] have been wildly used in the automatic speech recognition (ASR) for many years and proven to be effective ways to improve the recognition accuracy and robustness of the ASR systems. Based on the fact that the noises or mismatches do not affect the different data streams in similar ways, the MS recognizers can usually outperform the single stream ones in various and unpredictable noisy environments by choosing and fusing the complementary data streams properly. The major difficulty of applying multi-stream approach to speech enhancement is that the speech waveform cannot be directly recovered from many kinds of data streams, for example, the visual information of lips.

In this paper, we propose a new multi-stream speech enhancement framework which can make use of multi-stream information even some data streams are not directly related to the speech waveform. The proposed approach is based on the framework of the single-channel model based speech enhancement technique, with the exception that a multi-stream model is employed to classify each noisy speech frame. Based on the classification results, the noisy speech is enhanced by a class dependent filter as the conventional model based enhancement methods do. In this way, the multi-stream information does not required to be used to recover the speech waveform directly, but to improve the robustness of the frame classier, which can outperform the conventional model based enhancement methods with single data stream extracted from the noisy acoustic speech signals.

The rest parts of this paper are organized as follows. In Section 2, the multi-stream speech enhancement framework is introduced. In section 3, a multi-stream enhancement method is proposed based on the acoustic and throat microphone recordings. Section 4 presents the experimental results and Section 5 conclusions the paper.

II. FRAMEWORK

In the conventional single-channel model based speech enhancement systems, the input noisy speech is first segmented into frames, then a model based classier is used to calculate the score of each frame and an optimal filters is constructed for each class according to the current noise model. Finally the noisy speech frames are filtered by a class

* To whom correspondence should be addressed.

This work was supported in part by the National Natural Science Foundation of China (No. 61372083, 61370186), Science and Technology Planning Project of Guangdong Province (No.2014B010116001, 2014A010103040), Science and Technology Planning Project of Guangzhou (No.201604010049, No.201510010203), Appropriative Researching Fund for Professors and Doctors, Guangdong University of Education (No.2015ARF25).

978-1-5090-4400-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDH.2016.54

dependent filter fused by the optimal filters of each class. Fig. 1 (a) shows the block diagram of the conventional single stream model based speech enhancement system. Comparing with non-model based techniques like spectral subtraction, Winner filter and MMSE, the model based methods model the noisy speech in a more precise way, therefore can provide better performance.

The accuracy of the frame classification is important to any model based enhancement method because it determines the performance of the noisy speech filter. However, in the conventional model based approaches, classification errors will occur more frequently when the environment acoustic noise level rises because the classifications only depend on the single stream of acoustic speech. The multi-stream approach is an effective way to deal with acoustic noise, which is wildly used in speech recognition. By introducing other data streams that are less affected by or immune to acoustic noises, such as the visual information, throat or bone microphone, and etc., the classier will be more robust under low acoustic SNRs. The block diagram of the proposed multi-stream model based speech enhancement system is showed in Fig.1 (b), in which the single stream classier in Fig.1 (a) is replaced by a multi-stream classier.

Let $s^i(n)$ denotes the ith input data stream, where $i = 1 \sim N$ and N is the number of the input data streams. The data stream of acoustic speech is necessary in and assumes to be $s^1(n)$. The stream data are segmented into frames and input to a multi-stream classier with L classes. After the feature extraction and classification, the classier outputs the scores of each class as

$$q(k,l) = f\left(\mathbf{o}^1(k), \mathbf{o}^2(k), ..., \mathbf{o}^N(k)\right), \quad (1)$$

where $\mathbf{o}^i(k)$ is the observed feature vector for the kth frame of the ith data stream, $q(k,l)$ is the output score of the lth class for the kth frame, and $f(\bullet)$ is the scoring function of the multi-stream classier. For each class and the current acoustic noise, an optimal filter can be obtained by

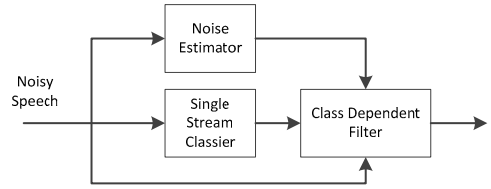
$$\mathbf{h}(k,l) = \arg\max I\left(l, \mathbf{h}(k,l), \mathbf{n}(k)\right), \quad (2)$$

where $\mathbf{h}(k,l)$ is the impulse response of the optimal filter for the lth class under the current noise $\mathbf{n}(k)$, $I(\bullet)$ denotes the object function for optimizing the filter coefficients, for example, the Minimum Mean Square Error (MMSE). Then the input acoustic speech is enhanced by

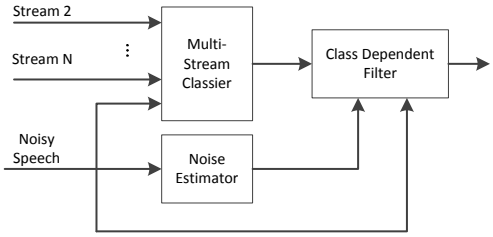
$$\hat{\mathbf{s}}(k) = g\left(\mathbf{s}^1(k), \mathbf{h}(k,l)\right), \quad (3)$$

where $\hat{\mathbf{s}}(k)$ and $\mathbf{s}^1(k)$ denote the enhanced and noisy acoustic speech of the kth frame, $g(\bullet)$ denotes the fusion function of the L optimal filters for each class.

According to the above discussion, there are several key problems for establishing a multi-stream speech enhancement system, i.e., the selection of complementary data streams and the designation of $f(\bullet)$, $I(\bullet)$ and $g(\bullet)$.



(a) Single stream



(b) Multi-stream

Figure 1. Diagram of single and multi-stream model based speech enhancement systems

III. Speech Enhancement with Acoustic and Throat Microphone

Unlike the acoustic microphone (AM), the throat microphone (TM) records speech signal in the form of vibrations through skin-attached piezo-electric sensors. Its recordings are immune to acoustic noises, which show good robustness in low acoustic SNR environments, but represent lower bandwidth speech signal content and perform worse in high acoustic SNR environments compared to the open-air acoustic recordings. Therefore it provides good complementary information for the acoustic microphone recordings when the acoustic SNR changes and is an attractive candidate for the multi-stream speech enhancement system.

The structure of the proposed method is showed in Fig.2. Speech signals from the acoustic and throat microphones are first divided into synchronous frames. Then speech features are extracted from each frame to form two observation streams. Variable kinds of features used in the speech recognition can be employed here.

The classifier is a multi-stream GMM or HMM, in which each Gaussian mixture represents one class of speech frames. The conventional output probability of the multi-stream GMM or the state of multi-stream HMM is modified as

$$p\left(\mathbf{o}^1(k), \mathbf{o}^2(k)\right) = \sum_{m=0}^{M-1} c_m \prod_{i=1}^{2} \left(N\left(\mathbf{o}^i(k), \boldsymbol{\mu}_m^i, \boldsymbol{\sigma}_m^i\right)\right)^{w_m^i}, \quad (4)$$

where $\mathbf{o}^1(k)$ and $\mathbf{o}^2(k)$ denote the observations of acoustic and throat microphone recordings, M is the number

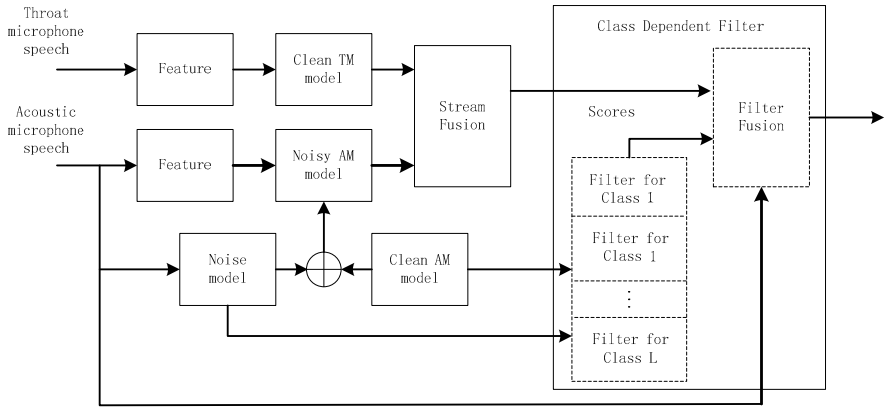


Figure 2. Diagram of the proposed speech enhancement method using acoustic and throat microphone recordings

of the mixture, $N\left(\mathbf{o}^i(k), \boldsymbol{\mu}_m^i, \boldsymbol{\sigma}_m^i\right)$ is the Gaussian function with the mean and variance of $\boldsymbol{\mu}_m^i$ and $\boldsymbol{\sigma}_m^i$, c_m and w_m^i are the weights of Gaussian mixtures and data streams respectively. Then the score of current frame for each class can be given by

$$q(k,l) = c_l \prod_{i=1}^{2} \left(N\left(\mathbf{o}^i(k), \boldsymbol{\mu}_l^i, \boldsymbol{\sigma}_l^i\right)\right)^{w_l^i}. \quad (5)$$

Model compensation techniques like VTS or PMC can be applied to the stream model of acoustic speech to further improve the robustness of the classier. The noise can be modeled using the same method as the speech.

Assume that the current acoustic frame belongs to class *l*. An optimal Winner filter for class *l* can be given by the following formula under the MMSE criterion when the noise model is available:

$$H(l,j) = \frac{\mu_S(l,j)}{\mu_S(l,j) + \mu_N(j)}, \quad (6)$$

where $\mathbf{H}(l)$ is the optimal filter gain of the *l*th class for the noisy speech spectra, $\boldsymbol{\mu}_S(l) = \boldsymbol{\mu}_l^1$ is the means of the acoustic speech spectrum of the *l*th class and $\boldsymbol{\mu}_N$ is the means of the noise spectrum.

The final class dependent filter can be obtained by fusion all the optimal filters together according to the score of each class as:

$$\hat{\mathbf{H}}(k) = \sum_{l=1}^{L} q(k,1)\mathbf{H}(l) / \sum_{l=1}^{L} q(k,1), \quad (7)$$

or

$$\hat{\mathbf{H}}(k) = \mathbf{H}(l), \quad l = \arg\max q(k,l). \quad (8)$$

IV. EXPERIMENTS

In the experiment, about 32 minutes speech pairs from the acoustic and throat microphones with the contents of Chinese news are recorded synchronously in a quiet laboratory environment. All speeches are spoken by 4 adult men and 4 adult women. The speeches are sampled with the sample rate of 16 kHz and quantized with 16 bits. 80 percent of the speech data are used as training set and the rest are testing set.

The white noise, F16 noise, and factory noise from the NOISEX92 database are artificially added to the clean acoustic speeches with SNR levels from -10dB~10dB. The throat microphone recordings are not affected by the acoustic noises and assumed to be clean in the experiment.

The acoustic and throat microphone speeches are segmented into frames with the frame length of 512 samples and the frame shift of 256 samples. For each speech frame, 13 MFCC coefficients are extracted from the log energy outputs of 24 Mel filter. A two-stream GMM with 4 mixtures is trained on the training set with the HTK tools 3.0 [11]. The stream weight of the acoustic and throat microphone recordings are set to 0.1 and 0.9 respectively. The noise models are the means of the noise spectral which is estimated by averaging the spectral of pure noise frames. (8) is used to calculate the final speech filter in the experiment. The PESQ subjective criterion is used to evaluate the performance of the proposed enhancement systems.

Table 1 shows the PESQ scores of spectral subtraction, Winner filter, MMSE, single stream model based enhancement method and the proposed multi-stream model based enhancement method, where "noisy", "SS", "Winner", "MMSE", "SMB" and "MMB" refer to the noisy speech, spectral subtraction, Winner filter, MMSE, single stream model based enhancement method and the proposed multi-stream model based enhancement method respectively. The results show that the proposed method

TABLE I. PESQ SCORES OF THE PROPOSED AND REFERENCE METHODS

	-10dB	*-5dB*	*0dB*	*5dB*	*10dB*
Noisy	1.11	1.38	1.57	1.83	2.19
SS	1.17	1.37	1.64	1.97	2.35
Winner	1.16	1.38	1.66	2.00	2.38
MMSE	1.16	1.38	1.69	2.04	2.44
SMB	1.73	1.95	2.14	2.33	2.53
MMB	1.75	2.03	2.29	2.48	2.66

TABLE II. PESQ SCORES OF DIFFERENT STREAM WEIGHTS

(w_a, w_t)	*-10dB*	*-5dB*	*0dB*	*5dB*	*10dB*
(0.0, 1.0)	1.73	2.01	2.26	2.46	2.64
(0.1, 0.9)	1.75	2.03	2.29	2.48	2.66
(0.3, 0.7)	1.70	2.00	2.27	2.48	2.67
(0.5, 0.5)	1.58	1.94	2.22	2.46	2.66
(0.7, 0.3)	1.31	1.62	2.01	2.32	2.61
(0.1, 0.9)	0.95	1.42	1.84	2.16	2.48
(1.0, 0.0)	1.73	1.95	2.14	2.33	2.53

achieves the best scores in different type of noises and SNR levels. It means that fusing an acoustic noise immune data stream into the classier can improve the performance of model based enhancement system effectively.

Table 2 shows the PESQ scores of the proposed multi-stream method with different stream weights. w_a and w_t denote the stream weights of data streams from the acoustic and throat microphones respectively and $w_a + w_t = 1$. When one of the stream weights is equals to zero, it means that the system turns into a single stream one. The results in Table 2 show that the multi-stream system performs better than the single stream ones. It also can be seen that the performances are stable when $w_t \geq 0.5$ while become worse when $w_t < 0.5$, especially when the SNRs are low. It is because that the throat microphone recordings can provide stable classification information when the acoustic SNR changes.

V. CONCLUSION

In this paper, we propose a new model based multi-stream speech enhancement framework, in which the input noisy speech frames are classified by a multi-stream classifier and then filter by a class dependent filter. The proposed multi-stream enhancement framework can make use of the information provided by data streams from which the speech waveform might not be recovered directly. Based on the new multi-stream enhancement framework, a speech enhancement method using acoustic and throat microphone recordings is also proposed. Experimental results show that the proposed method outperforms several conventional single stream speech enhancement methods. Further improvement will be achieved by using more sophisticated stream weight estimation techniques.

REFERENCES

[1] H. Veisi, H. Sameti, "Speech enhancement using hidden Markov models in Mel-frequency domain," Speech Communication, vol. 55, pp. 205-220, 2013.

[2] S. Y. Low, D. S. Pham, S. Venkatesh, "Compressive speech enhancement," Speech Communication, vol. 55, pp. 757-768, 2013.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustics, Speech and Signal Processing, vol. 27, pp.113-120, 1979.

[4] J. Lim, A. Oppenheim, "All-pole modeling of degraded speech," IEEE Trans. Acoustics Speech Signal Process, vol. 26, pp. 197–210, 1978.

[5] R. Martin, "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors," IEEE Trans. Speech and Audio Processing, vol. 13, pp. 845-856, 2005.

[6] D. P. Jarrett, M. Taseska, E. A. P. Habets, P. A. Naylor, "Noise Reduction in the Spherical Harmonic Domain Using a Tradeoff Beamformer and Narrowband DOA Estimates," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 24, pp. 967-978, 2014.

[7] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, J. Benesty, "A Framework for Speech Enhancement With Ad Hoc Microphone Arrays," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 24, pp. 1038-1051, 2016.

[8] V. Estellers, M. Gurban, and J. Thiran, "On Dynamic Stream Weighting for Audio-Visual Speech Recognition", IEEE Trans. Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1145 – 1157, 2012.

[9] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition", IEEE Signal Processing Letters, vol. 10, no.3, pp. 72-74, 2003.

[10] S. K. Nemala, K. Patil, and M. Elhilali, "A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition", IEEE Trans. Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 416-426, 2013

[11] S. Young, and etc. (2006, December). The HTK Book (for HTK version 3.4) [Online]. Available: http://htk.eng.cam.ac.uk/