A SPEECH ENHANCEMENT ALGORITHM BY ITERATING SINGLE- AND MULTI-MICROPHONE PROCESSING AND ITS APPLICATION TO ROBUST ASR

Xueliang Zhang[1], *Zhong-Qiu Wang*[2] *and DeLiang Wang*[2,3]

[1]Department of Computer Science, Inner Mongolia University, China
[2]Department of Computer Science and Engineering, The Ohio State University, USA
[3]Center for Cognitive and Brain Sciences, The Ohio State University, USA
cszxl@imu.edu.cn, {wangzhon, dwang}@cse.ohio-state.edu

ABSTRACT

We propose a speech enhancement algorithm based on single- and multi-microphone processing techniques. The core of the algorithm estimates a time-frequency mask which represents the target speech and use masking-based beamforming to enhance corrupted speech. Specifically, in single-microphone processing, the received signals of a microphone array are treated as individual signals and we estimate a mask for the signal of each microphone using a deep neural network (DNN). With these masks, in multi-microphone processing, we calculate a spatial covariance matrix of noise and steering vector for beamforming. In addition, we propose a masking-based post-filter to further suppress the noise in the output of beamforming. Then, the enhanced speech is sent back to DNN for mask re-estimation. When these steps are iterated for a few times, we obtain the final enhanced speech. The proposed algorithm is evaluated as a frontend for automatic speech recognition (ASR) and achieves a 5.05% average word error rate (WER) on the real environment test set of CHiME-3, outperforming the current best algorithm by 13.34%.

Index Terms—post-filtering, beamforming, deep neural networks, speech enhancement, spectral masking

1. INTRODUCTION

In real-world environments, ASR systems are severely interfered by background noise. To deal with this problem, speech enhancement or separation [3] is often used as a frontend to suppress noise and reduce the acoustic mismatch between training and testing. However, it is challenging for speech enhancement to suppress noise without introducing speech distortion. Another way is to collect noisy speech as much as possible and directly train an ASR system on large-scale training data. Apparently, it has high costs for collecting and labeling data. A compromise is to combine these two approaches, and such methods show the state-of-the-art environmental robustness for ASR.

Speech enhancement can be divided into single- and multi-microphone processing in general. Traditional single-microphone speech enhancement mainly utilizes the statistical information of speech or noise. Recently, deep learning was introduced to speech enhancement/separation [14], and it typically learns a mapping from noisy features to a time-frequency (T-F) mask or target speech [15]. The biggest issue for supervised speech enhancement is generalization to new noises, since test and training conditions can be quite different. On the other hand, direction information is effective for separating the sound sources coming from different directions, and it can be captured in a multi-microphone situation. Beamforming, or spatial filtering, is the dominant approach for multi-microphone speech enhancement.

A beamformer is often parameterized by a steering vector for the target direction. A common approach to obtaining a steering vector is by using the direction of arrival (DOA) and microphone array geometry with the assumption of plane wave propagation. Recently, masking-based beamforming [9][17] has been studied. The advantage of this method is that it can work without the knowledge of microphone array geometry and shows robustness to real noisy environments. The main idea is to use the principal eigenvector of the target speech covariance matrix as the steering vector. To compute the covariance matrix, a T-F mask is estimated by spatial vector clustering.

We propose a masking-based speech enhancement algorithm by combining single- and multi-microphone processing. In single-microphone processing, a DNN is trained to map the spectral features to a T-F mask. In multi-microphone processing, the estimated mask is used to calculate the noise covariance matrix and steering vector, with which minimized variance distortionless response (MVDR) beamformer [5] is employed for speech enhancement. In this case, the performance of beamforming depends on the accuracy of the mask estimate. On the other hand, the beamformer exhibits stable noise suppression and can help to get a better mask estimate. Based on this observation, our proposed method iterates masking and beamforming. The application of the proposed method to robust ASR shows very good results on the CHiME-3 corpus and it outperforms the current best algorithm [17] significantly.

This research was supported in part by a China National Nature Science Foundation grant (No. 61365006), an AFRL contract (FA8750-15-1-0279), an NSF grant (IIS-1409431), and the Ohio Supercomputer Center. The work was performed while the first author was a visiting scholar at the Ohio State University.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

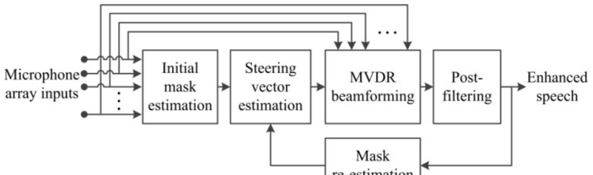

Fig. 1. Schematic diagram of the proposed speech enhancement algorithm.

In the following section, we describe the relation to previous works. Section 3 presents our speech enhancement algorithm. Experimental results are given in Section 4. We conclude the paper in Section 5.

2. RELATION TO PRIOR WORKS

The proposed method can be viewed as a robust beamformer using a T-F mask to estimate a steering vector and spatial noise covariance matrix. In [9], a mask is obtained by an unsupervised method, i.e. spatial vector clustering. In each frequency channel, a complex GMM is used to describe the distribution of the spatial vectors of a microphone array. The GMM has two Gaussian components that describe the noise and noisy speech respectively. With the mask derived from the posterior probabilities of each component, noise and noisy speech covariance matrices are computed. The steering vector corresponds to the principal eigenvector of the target speech covariance matrix obtained by subtracting the noise covariance matrix from the noisy speech covariance matrix. Finally, the enhanced speech is obtained by using MVDR beamforming. In [8], a mask is obtained by a supervised method, where speech-dominated and noise-dominated masks are estimated by using bi-directional Long Short-Term Memory (BLSTM). With these two masks, they enhance the noisy speech by using a generalized eigenvalue (GEV) beamformer [16] with an optional distortion reduction filter.

Different from the above methods, the proposed method employs DNN to perform supervised T-F masking, and the target speech covariance matrix is obtained in an adaptive way. Our iterative procedure plays an important role in improving the mask estimation. In addition, a masking-based post-filtering is proposed to further boost the performance.

3. SYSTEM DESCRIPTION

The proposed system is shown in Fig. 1. In the stage of initial mask estimation, we treat the microphone array as several independent microphones and generate a mask for each microphone input using one DNN model. Then an initial mask is constructed by taking the maximum value across the multiple masks of each T-F unit. With the initial mask, the steering vector and noise covariance matrix are calculated. We obtain the enhanced speech by applying MVDR beamforming and post-filtering to the microphone array signals. To refine the estimated mask, we feed the enhanced speech into the DNN again. This process is iterated for several times to obtain the final enhanced speech.

3.1. Single-microphone Processing

For the supervised speech enhancement, there are three key factors, i.e. training targets, features and learning machines. Here, we employ the IRM as the training target, which is defined as [11]

$$IRM(t,f) = \sqrt{\frac{s^2(t,f)}{s^2(t,f)+n^2(t,f)}} \tag{1}$$

where $s^2(t,f)$ and $n^2(t,f)$ denote the speech and noise energy in a particular T-F unit, respectively. Eq. (1) is closely related to the Wiener filter which is the optimal estimator in the power spectrum domain [11].

Discriminative features are very important for learning machines. In this study, we use a set of complementary features consisting of amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP) and Mel-frequency cepstral coefficients (MFCC). The feature set used here is similar to the one in [15], while we exclude the cochleagram and delta features. Since useful information is carried across time frames, a symmetric 9-frame context window is used to splice adjacent frames into a single feature vector. We train a feed-forward DNN to map the input feature vector to the IRM of the central frame. At the test stage, the estimated IRM is used for beamforming and post-filtering.

3.2. Multi-microphone Processing

3.2.1. MVDR beamformer

The MVDR beamformer is to minimize the noise energy while imposing linear constraints to maintain the energy from the target direction. In the short-time Fourier transform (STFT) domain, the received signal can be expressed as:

$$\mathbf{y}(t,f) = \mathbf{c}(f)s(t,f) + \mathbf{n}(t,f) \tag{2}$$

where $\mathbf{y}(t,f)$ and $\mathbf{n}(t,f)$ are STFT vectors of the received signals and noises of a microphone array at time frame t and frequency channel f, respectively. $s(t,f)$ represents the STFT of the speech source, $\mathbf{c}(f)s(t,f)$ stands for the direct path part of the received speech signal and $\mathbf{c}(f)$ is the steering vector of the microphone array.

For frequency channel f, the MVDR beamformer aims to find a weight vector $\mathbf{w}(f)$ that can minimize the average output power of the beamformer while maintaining the energy along the look direction. This optimization problem can be formulated as Eq. (3). To express it more concisely, we omit the notation of frequency f in the rest of this section.

$$\mathbf{w}_{\text{opt}} = \text{argmin}_{\mathbf{w}}\{\mathbf{w}^H \Phi_{\text{n}} \mathbf{w}\}, \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{c} = 1 \tag{3}$$

where H denotes the conjugate transpose and Φ_{n} is the spatial covariance matrix of the noise. The well known solution of this optimization problem is:

$$\mathbf{w}_{\mathrm{opt}} = \frac{\Phi_{\mathrm{n}}^{-1}\mathbf{c}}{\mathbf{c}^{H}\Phi_{\mathrm{n}}^{-1}\mathbf{c}} \tag{4}$$

The enhanced speech signal $\tilde{s}(t)$ is produced by applying the linear filter $\mathbf{w}_{\mathrm{opt}}$ to the microphone signal vector using Eq. (5). Therefore, the key step of MVDR beamforming is to accurately estimate $\mathbf{c}$ and Φ_{n}.

$$\tilde{s}(t) = \mathbf{w}_{\mathrm{opt}}^{H}\mathbf{y}(t) \tag{5}$$

3.2.2. Steering vector estimation

Steering vector is traditionally obtained by using DOA estimation and microphone array geometry [2]. When SNR level is low, it is pretty hard to estimate the DOA accurately. In contrast, the mask-based approach [9] calculates the spatial covariance matrix of target speech signal and uses its principal eigenvector as the estimated steering vector. We follow this idea but with refinements.

First, we calculate the noisy speech and noise covariance matrix using Eq. (6) and (7), respectively.

$$\Phi_{\mathrm{y}}(t) = \mathbf{y}(t)\mathbf{y}(t)^{H} \tag{6}$$

$$\Phi_{\mathrm{n}}(t) = \frac{1}{\sum_{l=t-L}^{t+L}(1-m(l))}\sum_{l=t-L}^{t+L}\big(1-m(l)\big)\mathbf{y}(l)\mathbf{y}(l)^{H} \tag{7}$$

where m is the estimated IRM from the DNN, and L (set to 10 in this study) is the half window size. The covariance matrix of the target speech is obtained by subtracting the noise covariance matrix from the noisy speech covariance matrix, see Eq. (8).

$$\Phi_{\mathrm{x}} = \Phi_{\mathrm{y}} - \Phi_{\mathrm{n}} = \frac{1}{T}\sum_{t=1}^{T}\Big(\Phi_{\mathrm{y}}(t) - \Phi_{\mathrm{n}}(t)\Big) \tag{8}$$

The estimated steering vector can be obtained by first performing eigenvector decomposition on Φ_{x} and then extracting the eigenvector corresponding to the maximum eigenvalue. In [9], it can be viewed as $\Phi_{\mathrm{n}}(t)$ is calculated on the entire signal or on a fixed-length segment. That may not be a good strategy for non-stationary noises. Instead, we adopt an adaptive way, using the neighboring $(2 \times L + 1)$ frames to calculate $\Phi_{\mathrm{n}}(t)$, as shown in Eq. (7). In our experiments, this adaptive approach leads to much better performance, especially for non-stationary noises.

3.3. Post-filtering

Due to the resolution issue of the MVDR beamformer and the inaccuracy of the steering vector estimate, the noise reduction is not sufficient enough. In contrast, IRM can accomplish higher noise reduction than the MVDR beamformer, therefore its integration in the beamformer output would probably lead to SNR gain. However, directly applying the IRM to the beamformer output is very sensitive to the IRM estimation error. Instead, we propose a post-filtering method. For each frequency channel, we first calculate the global SNR, $cSNR(f)$, using estimated mask by Eq. (9), and then use it to compute a threshold, $\lambda(f)$. Finally, we obtain the gain of the post-filter, $\mathrm{g}(t,f)$, by Eq. (11).

$$cSNR(f) = 10\log_{10}\frac{\sum_{t=1}^{T} m(t,f)\tilde{s}(t,f)^{2}}{\sum_{t=1}^{T}\big(1-m(t,f)\big)\tilde{s}(t,f)^{2}} \tag{9}$$

$$\lambda(f) = \frac{1}{1+\exp\left((cSNR(f)-\alpha)/\beta\right)} \tag{10}$$

$$\mathrm{g}(t,f) = m(t,f)^{\lambda(f)} \tag{11}$$

where $m(t,f)$ and $\tilde{s}(t,f)$ are the estimated IRM and spectrum of the beamformer output. Eq. (10) is a sigmoidal function, so the threshold $\lambda(f)$ is between [0,1]. We use parameters α and β to adjust the shape of the sigmoidal function. Through cross-validation, their values are set to -5 and 2, respectively.

From Eq. (10), we can see that $\lambda(f)$ would be close to 0 when $cSNR(f)$ is high, making $\mathrm{g}(t,f)$ being close to 1 no matter what $m(t,f)$ is. Otherwise, $\mathrm{g}(t,f)$ is close to $m(t,f)$ when $cSNR(f)$ is low. The final enhanced spectrum is the product of $\mathrm{g}(t,f)$ and $\tilde{s}(t,f)$.

4. EXPERIMENTAL RESULTS

We evaluate the proposed speech enhancement algorithm on ASR tasks using the CHiME-3 dataset [2]. The proposed algorithm is used as a frontend for ASR systems.

4.1. Dataset

The CHiME-3 challenge uses a read speech corpus based on the speaker-independent medium-vocabulary (5k) subset of the Wall Street Journal (WSJ0) corpus [6]. There are two types of data in CHiME-3. The first one is "Real data" - speech data recorded in real noisy environments (on a bus, cafe, pedestrian area, and street junction) uttered by actual talkers. The second one is "Simulated data" - noisy utterances that have been generated by artificially mixing clean speech data with noisy backgrounds. The ultimate goal is to recognize the real data. All the submitted systems in the challenge are ranked according to their performance on the real subset of the test data [2].

The training set is composed of 8738 (1600 real + 7138 simulated) noisy utterances. The development set and test set include 3280 (1640 real + 1640 simulated) and 2640 (1320 real + 1320 simulated) noisy utterances in the four different environments, respectively. All of these data are simulated or recorded to form six-channel signals for each utterance. Details of the data sets and regulations can be found in [2].

4.2. DNN Training

As mentioned in subsection 3.1, we build a DNN for mask estimation. The DNN has three hidden layers each with 1024 rectified linear units (ReLUs) [12]. Sigmoidal units are used in the output layer since the IRM is bounded between 0 and 1. The DNN weights are randomly initialized and the adaptive gradient algorithm (AdaGrad) [4] is utilized for optimization. The momentum is set to 0.5 for the first 5 epochs and 0.9 for the remaining 25 epochs (30 epochs in total).

We use the simulated training data to train the DNN, as the IRM can only be derived on simulated data. The multi-microphone input signals are treated as individual single-channel signals. The total number of the noisy utterances for training is therefore 7138×6. We use the simulated development set for early stopping and hyper-parameter tuning. Its total number is 1640×6.

4.3. Speech Recognition

To facilitate the comparison, we use two DNN-based ASR systems, denoted as ASR-1 and ASR-2, to evaluate the ASR performance. ASR-1 is the baseline system provided in the official CHiME-3 package. It is trained on the noisy utterances at the fifth channel by using the standard training procedure in the Kaldi toolkit [13], i.e. pre-training, cross entropy training, state-level Minimum Bayes Risk (sMBR) training and language model re-scoring. The details of the baseline system can be found in the official training recipes[1]. ASR-2 is built in the same way, but using the noisy utterances of all the six channels. Apparently, much more data is used for acoustic modeling, and the resulting acoustic model is found to be more robust.

As frontend processing, the official weighted delay-and-sum beamformer is used for comparison. We test its performance using ASR-1 and ASR-2 backend systems. The weighted delay-and-sum beamformer is implemented using the BeamformIt toolkit [1] where the DOA estimate is obtained from GCC-PHAT [10] and a two-step Viterbi post-processing technique is used to avoid instabilities.

We also compare the proposed approach with two other masking-based beamforming algorithms. The first one, proposed by Higuchi *et al.* [9][17], utilizes a clustering method for mask estimation. Their ASR system uses a more advanced deep convolutional neural network (CNN) for acoustic modeling. A recurrent neural network (RNN) language model is employed for lattice re-scoring. Their acoustic models are trained on the noisy utterances from all the six microphones. After unsupervised speaker adaptation, their system achieves the best results in the CHiME-3 challenge. The second one, proposed by Heymann *et al.* [7][8], utilizes BLSTM for mask estimation. When training BLSTM, they augment the simulated data to get better mask estimation. Their ASR system is the one provided in the official CHiME-3 package, which is almost the same as ASR-1.

The proposed algorithm without iterative procedure is also included for comparison. The results are all listed in Table 1. The proposed speech enhancement algorithm achieves 5.05% WER on the real environment test set using ASR-2 as the backend. It outperforms the previous best system [17] by 13.34% relatively (5.05% vs. 5.83%). Although two ASR systems have different acoustic models, both of them are trained using the same data, i.e. the noisy utterances from all the six microphones, and the language models are both RNNLMs.

When ASR-1 is used as the backend, the performance of the proposed algorithm achieves 6.17% WER on the real environment test set which is still better than the others, except for the system of Higuchi *et al.*. Without the iterative procedure, the proposed algorithm exhibits noticeable performance degradation, i.e. 5.42% WER, on the real test data. It indicates that the iterative procedure is effective for noise reduction.

Table 1. WERs (%) for different systems on CHiME-3

Speech enhancement (frontend)	ASR system (backend)	Development		Evaluation	
		simu	*real*	*simu*	*real*
BeamformIt	*ASR-1*	6.77	5.75	10.91	11.47
	ASR-2	6.08	5.07	9.47	9.88
Heymann et al.[2]	*ASR-1*	5.01	4.53	5.60	7.45
Higuchi et al.	*CNN-based*	3.63	3.45	4.46	5.83
Proposed	*ASR-1*	4.59	4.02	5.01	6.17
	ASR-2	3.53	3.49	3.98	**5.05**
Proposed without iterative procedure	*ASR-2*	3.37	3.46	3.88	5.42

5. CONCLUSION

We have proposed a speech enhancement algorithm that combines single- and multi-microphone processing and evaluated its performance in a robust ASR task. Experimental results show that, as a frontend, the proposed algorithm greatly improves ASR performance. Our ASR results significantly outperform the current best system on the CHiME-3 dataset .

It should be mentioned that our ASR system is relatively simple, and includes no speaker adaptation. By including advanced techniques in ASR, we believe that the ASR performance can be further improved.

6. ACKNOWLEDGEMENT

The authors would like to thank Yuxuan Wang for providing his DNN code.

[1] Available at https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4/s5_6ch

[2] The results are reported at https://github.com/fgnt/nn-gev

7. REFERENCES

[1] Anguera, X., Wooters, C., and Hernando, J., “Acoustic beamforming for speaker diarization of meetings,” *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 15, pp: 2011-2023, 2007.

[2] Barker, J., *et al.*, “The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, ” in *Proc. IEEE ASRU*, pp: 504-511, 2015.

[3] Benesty, J., Makino, S., Chen, J., Eds. *Speech Enhancement*. Springer Science & Business Media, 2005.

[4] Duchi, J., Hazan, E., and Singer, Y., “Adaptive subgradient methods for online learning and stochastic optimization,” *The Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, 2011.

[5] Frost III, O.L., “An algorithm for linearly constrained adaptive array processing, ” in *Proceedings of the IEEE*, vol.60, pp:926-935, 1972.

[6] Garofalo, J., *et al.*, “CSR-I (WSJ0) complete,” Linguistic Data Consortium, Philadelphia, 2007.

[7] Heymann, J., Drude, L., Chinaev, A., and Haeb-Umbach, R., “LSTM supported GEV beamformer front-end for the 3rd CHiME challenge, ” in *Proc. IEEE ASUR* 2015.

[8] Heymann, J., Drude, L., and Haeb-Umbach, R., “Neural network based spectral mask estimation for acoustic beamforming,” in *Proc. ICASSP*, pp: 196-200, 2016.

[9] Higuchi, T., *et al.*, “Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise, ” in *Proc. ICASSP*, pp: 5210-5214, 2016.

[10] Knapp, C., and Clifford Carter, G., “The generalized correlation method for estimation of time delay,” *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 24, pp.320–327, 1976.

[11] Loizou, P.C., *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC, 2007.

[12] Nair, V., and Hinton, G., “Rectified linear units improve restricted boltzmann machines,” in *Proc. ICML 2010* pp: 807-814.

[13] Povey, D., Ghoshal, A., and Boulianne, G., “The Kaldi speech recognition toolkit,” in *Proc. IEEE ASUR* 2011.

[14] Wang, Y., and Wang, D.L., “Towards scaling up classification-based speech separation,” *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, pp. 1381-1390, 2013.

[15] Wang, Y., Narayanan, A., and Wang, D.L., “On training targets for supervised speech separation,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, pp:1849-1858, 2014.

[16] Warsitz, E., and Haeb-Umbach, R., “Blind acoustic beamforming based on generalized eigenvalue decomposition,” *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 15, pp: 1529–1539, 2007.

[17] Yoshioka, T., *et al.*, “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” in *Proc. IEEE ASRU*, pp: 436 - 443, 2015.