INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Speech Enhancement Based on Beamforming and Post-Filtering by Combining Phase Information

Rui Cheng, Changchun Bao

Speech and Audio Signal Processing Laboratory, Faculty of Information Technology,
Beijing University of Technology, Beijing, 100124, China
chengrui@emails.bjut.edu.cn, baochch@bjut.edu.cn

Abstract

Speech enhancement is an indispensable technology in the field of speech interaction. With the development of microphone array signal processing technology and deep learning, the beamforming combined with neural network has provided a more diverse solution for this field. In this paper, a multi-channel speech enhancement method is proposed, which combines beamforming and post-filtering based on neural network. The spatial features and phase information of target speech are incorporated into the beamforming by neural network, and a neural network based single-channel post-filtering with the phase correction is further combined to improve the performance. The experiments at different signal-to-noise ratio (SNR) levels confirmed that the proposed method results in an obvious improvement on speech quality and intelligibility compared to the reference methods.

Index Terms: beamforming, post-filtering, spatial features, phase sensitive, phase correction

1. Introduction

With the development of speech communication technology, the use of microphone array to extract the desired speech from noisy and reverberant environments has become an important research task. One important reason for this is that the microphone array could utilize information about speech source location except for spectral information [1].

Beamforming is one of the most effective methods to implement speech enhancement [2][3][4]. Among them, the minimum variance distortionless response (MVDR) beamforming [5] is a kind of adaptive beamforming method. It can make the interesting signal undistorted when the interesting signal passes through the filter on the desired direction. Meanwhile, it can minimize the variance of residual noise or the filtered speech.

In recent years, traditional beamforming methods have been combined with deep neural networks (DNNs) to improve the performance of speech enhancement. For example, the masking estimation methods [6][7] were proposed by combining the MVDR beamforming and the DNNs. The long short-term memory networks (LSTMs) were used in these methods to estimate the mask for masking noisy speech of each channel so that the spatial covariance matrices (SCMs) of clean speech and noise are calculated for constructing the MVDR beamformer. On the other hand, an iterative-based post-filtering method was proposed in [8], in which the MVDR was constructed using the DNNs to estimate the mask. Thus, a mask-based post-filter was obtained to further suppress the noise of the beamformed speech. Through the iterative operations, i.e., the output speech is sent back to the neural network for re-estimation of the mask, the finally enhanced speech is obtained. Although great progress has been made in the beamforming with the help of the DNNs, there is still much room for improvement and research significance.

In this paper, a multi-channel speech enhancement method based on the spatial and phase information is proposed, in which the DNNs-based beamforming and post- filtering are included. In the beamforming, the efficient spatial features and phase-sensitive-based mask are considered to implement speech enhancement based on the DNNs and the MVDR beamforming. In post-filtering, a single-channel speech enhancement method based on the DNNs with the phase-sensitive-based mask and phase correction technology is performed, which can further improve the performance of the beamformed speech. The experiments showed that the proposed method could increase the speech quality and intelligibility effectively by making better use of the spatial and phase information.

The structure of the paper is organized as follows: The proposed neural network-based beamforming method is described in Section 2. The proposed single-channel post-filtering method is given in Section 3. Experiments and discussion are shown in Section 4. Finally, Conclusions and future work are summarized in Section 5.

2. Neural Network-Based Beamforming

The block diagram of the neural network-based beamforming is illustrated in Figure 1. In this method, only one target speech is considered for reverberation condition in the cases of additive microphone self-noise and isotropic spherically diffuse noise which is generated by acoustic noise field generator [9].

2.1. Input Features of Neural Network

Considering a microphone array with L microphones that captures the reverberant speech with noise. In the short-time Fourier transform (STFT) domain, the received signal at the l^{th} microphone can be represented as $Y_l(n,k)=|Y_l(n,k)|e^{j\angle Y_l(n,k)}$, where n and k are the indexes of time frame and frequency bin, respectively. $|\cdot|$ and $\angle\cdot$ indicate the magnitude and phase operation, respectively. The magnitude of noisy speech plays an indispensable role in the DNNs-based methods due to its time-varying characteristics. So, in this work, the magnitude is used for primary feature.

Benefits from microphone array, the inter-channel phase differences (IPDs) representing the spatial feature of the target speech can be used for the second feature. IPDs can help

http://dx.doi.org/10.21437/Interspeech.2020-990

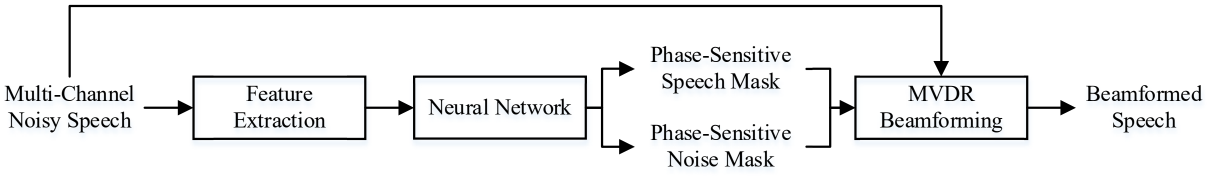

Figure 1: *The block diagram of the neural network-based beamforming.*

neural network to make full use of spatial information and better serve the subsequent beamforming process. When the 1st microphone is chosen as the reference microphone, the IPDs at the i^{th} microphone can be calculated as

$$IPD_i(n,k) = \angle Y_i(n,k) - \angle Y_1(n,k), \quad i = 2,\cdots,L \quad (1)$$

That is, the IPDs is obtained by subtracting the phase of the reference microphone from the phases of other microphones. This operation can eliminate inherent wrap characteristics of the short-term phase and can be used as input features of neural network to capture spatial information of target speech and acoustic characteristics of the room.

As a result, the magnitudes and the IPDs of each microphone are combined together to form input features of neural network, so input feature vector can be expressed as

$$\begin{aligned}&\mathcal{F}(n,k)=\\&\left[|Y_1(n,k)|\cdots|Y_L(n,k)|\, IPD_2(n,k)\cdots IPD_L(n,k)\right]\end{aligned} \quad (2)$$

where $Y_1(n,k)$ ··· $Y_L(n,k)$ are the magnitude of the received signals from the 1st to the L^{th} microphone, and the $IPD_2(n,k)$ ··· $IPD_L(n,k)$ are the IPDs with respect to reference microphone from the 2nd to the L^{th} microphone, respectively.

2.2. Structure of Neural Network

The neural network in beamforming consists of three BiLSTM [10] layers and one output layer with two linear layers. Each layer contains 512 neurons. Activation functions ReLU [11] and tanh are used for hidden layers and output layer, respectively. The Adam optimizer is chosen to update the parameters of the neural network.

2.3. Training Targets of Neural Network

For the supervised speech enhancement methods, in general, mask-based training target can achieve better performance than magnitude-based training target [12]. In order to utilize phase information better, the phase-sensitive mask (PSM) [13] considering phase difference between the expected speech and noisy speech is used in this work. So, the PSMs of speech and noise are given by

$$\begin{aligned}&sPSM(n,k) =\\&\frac{|S_1(n,k)|}{|S_1(n,k)|+|N_1(n,k)|}\cos\left(\angle Y_1(n,k)-\angle S_1(n,k)\right)\end{aligned} \quad (3)$$

$$\begin{aligned}&nPSM(n,k) =\\&\frac{|N_1(n,k)|}{|S_1(n,k)|+|N_1(n,k)|}\cos\left(\angle Y_1(n,k)-\angle N_1(n,k)\right)\end{aligned} \quad (4)$$

where $S_1(n,k)$, $N_1(n,k)$ represent the STFT of speech and noise of the reference microphone, respectively.

In order to obtain more accurate estimation of the PSMs, the PSMs of speech and noise are supervised simultaneously by mean squared error criterion, it is given by

$$\begin{aligned}\varepsilon=&\left\|sPSM^*(n,k)-sPSM(n,k)\right\|_2^2\\&+\left\|nPSM^*(n,k)-nPSM(n,k)\right\|_2^2\end{aligned} \quad (5)$$

where $\|\cdot\|_2$ denotes L2 norm, $sPSM^*(n,k)$ and $nPSM^*(n,k)$ indicate the estimations of $sPSM(n,k)$ and $nPSM(n,k)$, respectively.

2.4. PSM-Based Beamforming

The SCMs of speech and noise $\phi_{ss}(k)$ and $\phi_{nn}(k)$ can be estimated from the $sPSM^*(n,k)$ and $nPSM^*(n,k)$ as follows

$$\phi_{ss}(k) = \frac{\sum_{n=1}^{N} sPSM^*(n,k)\mathbf{y}(n,k)\mathbf{y}^H(n,k)}{\sum_{n=1}^{N} sPSM^*(n,k)} \quad (6)$$

$$\phi_{nn}(k) = \frac{\sum_{n=1}^{N} nPSM^*(n,k)\mathbf{y}(n,k)\mathbf{y}^H(n,k)}{\sum_{n=1}^{N} nPSM^*(n,k)} \quad (7)$$

where N is the total number of the frames. $\phi_{ss}(k)$ is constrained to rank-1 as described in [14]. Additionally, $\mathbf{y}(n,k)$ is expressed a matrix of the STFT magnitudes of the received signals of L microphones

$$\mathbf{y}(n,k) = \left[Y_1(n,k),\ldots,Y_L(n,k)\right]^T \quad (8)$$

where $[\cdot]^{\mathrm{T}}$ is the transpose of a matrix.

With the SCMs of speech and noise, the MVDR beamformer can be calculated as

$$\mathbf{w}(k) = \frac{\phi_{nn}^{-1}(k)\phi_{ss}(k)\mathbf{u}}{\mathrm{Tr}\left[\phi_{nn}^{-1}(f)\phi_{ss}(f)\right]} \quad (9)$$

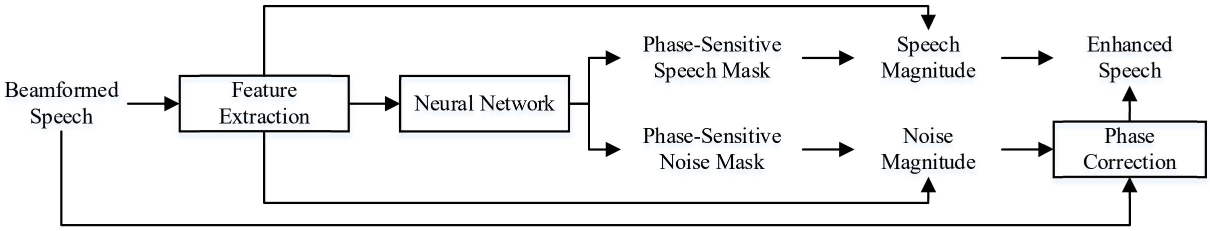

Figure 2: *The block diagram of the single-channel post-filtering.*

where **u** is a column vector whose first element is 1 and other elements are 0 [15]. Tr[·] is the trace of a matrix. So, the beamformed speech $S_{beam}(n,k)$ can be given by

$$S_{beam}(n,k) = \mathbf{w}^H(k)\mathbf{y}(n,k) \tag{10}$$

where $(\cdot)^{\mathrm{H}}$ is the Hermitian transpose.

3. Single-Channel Post-Filtering

In order to further suppress the noise components in the direction of the target speech. a single-channel speech enhancement method based on the DNNs and phase correction is proposed as the post-filtering. The block diagram of this method is shown in Figure 2.

In this method, the beamformed speech $S_{beam}(n,k)$ is fed into the neural network after feature extraction, the structure of the network is the same as the network used in neural network-based beamforming method. The enhanced magnitude $|S_{post}(n,k)|$ and $|N_{post}(n,k)|$ of speech and noise are obtained through multiplying the estimated PSMs of speech and noise by the beamformed speech $S_{beam}(n,k)$. The PSMs are defined similarly to Eq. (3) and Eq. (4), respectively.

Then, the phase correction [16] can be performed to obtain a more accurate phase than directly using phase of the beamformed speech, because there is an error between the phase of the beamformed speech and clean speech. The phase correction function $\Gamma(n,k)$ is defined by [16]

$$\Gamma(n,k) = ce^{-\frac{|S_{beam}(n,k)|^2}{|N_{post}(n,k)|^2}}\Omega(k)\,|\,N_{post}(n,k)\,| \tag{11}$$

where c is a constant. $\Omega(k)$ is a time-invariant anti-symmetry function given by

$$\Omega(k) = \begin{cases} 1 & 0 < k/K < 0.5 \\ -1 & 0.5 < k/K < 1 \\ 0 & else \end{cases} \tag{12}$$

where K is the length of the window used for the STFT. The corrected phase $\theta(n,k)$ can be obtained by the $\angle\cdot$ operator on the summation of the beamformed speech $S_{beam}(n,k)$ and phase correction function $\Gamma(n,k)$

$$\theta(n,k) = \angle\left(S_{beam}(n,k) + \Gamma(n,k)\right) \tag{13}$$

Then, the enhanced speech $S_{post}(n,k)$ can be obtained by combing the corrected phase $\theta(n,k)$ and the magnitude $|S_{post}(n,k)|$ of the speech estimated in post-filtering

$$S_{post}(n,k) = \left|S_{post}(n,k)\right|e^{j\theta(n,k)} \tag{14}$$

After the beamforming with spatial and phase information, this post-filtering can effectively enhance the speech.

4. Experiments and Results

4.1. Datasets

In the experiments, a uniform linear microphone with L=8 microphone is used, its microphone spacing is 4 cm. In order to analyze the effect of spatial position of the target speech and the array on the speech enhancement, 7 different relative positions for the target speech and the array are set in training, whereas 3 arbitrary relative positions are used in testing. For each relative position, the angular range of the array is discretized with a 5° resolution to get 37 different angular positions of relative target speech. The details about data configuration information is shown in Table 1, which is inspired by [17].

For the speech in training, 370 randomly chosen clean speech utterances from the TIMIT corpus [18] are used to match 37 different angular positions of the array, each angular position is related to 10 different speech utterances, and each utterance is about 3 s. 129-dimension magnitude spectrum of speech is used, where the window length used for the STFT is 256 samples and the windows are overlapped by 128 samples. The sampling rate of speech signal is 16 kHz. The white noise used as the microphone self-noise is added to the training data to get a 10 dB input SNR level, and babble noise used as the isotropic spherically diffuse noise is added to generate -6 dB, 0 dB and 6 dB input SNR levels. Clean speech, noise and room impulse responses (RIRs) [19] are combined to simulate the signals received by the microphones in different acoustic conditions. The training set is about 6 hours. For the network in the beamforming, we use entire training set, while for the network in the post-filtering, we only use noisy speech of the first microphone. In testing, another 185 utterances from the TIMIT test set are randomly chosen as the clean speech of testing set. For each relative position of target speech and array in test set, 5 utterances correspond to one angular

Table 1: *Configuration for the generation of the training dataset and test dataset.*

	Training dataset	Test dataset
Speech	TIMIT	
Room size [length, width and height]	[6, 6, 2.7]	[4, 7, 3]
Array positions	7 different positions	3 arbitrary positions
RT_{60}	0.3s	0.38s
SNR	Diffuse babble: -6 to 6 dB, Spatially white: 10 dB	

Table 2: *The average results of quality and intelligibility test.*

Methods			MVDR	IRM-CNN	NN-BF	NN-BF-PF
Δ PESQ	SNR (dB)	-6	0.1027	0.1985	0.2915	**0.3566**
		0	0.1386	0.3126	0.5055	**0.5750**
		6	0.1559	0.3958	0.5517	**0.6631**
Δ STOI (%)	SNR (dB)	-6	0.1111	4.3306	5.7714	**6.0818**
		0	0.1080	3.4616	3.8038	**4.6773**
		6	0.0944	0.5312	2.0740	**3.4631**
Δ SSNR	SNR (dB)	-6	2.5218	3.0299	3.1886	**6.9447**
		0	2.0398	2.5707	2.7303	**5.3444**
		6	1.7745	2.2202	2.1726	**4.1287**

position for 37 different angular positions. The test set is constructed by 3 different SNR levels for the babble noise from -6 to 6 dB at a step of 6 dB, while the white noise is 10 dB SNR level for all the cases. The duration of each speech segment in the test set is about 10 minutes.

4.2. Experimental Results

To evaluate speech quality and intelligibility, the enhanced speech is evaluated by perceptual evaluation of speech quality (PESQ) [20] in wide band (WB) mode, short-time objective intelligibility (STOI) [21] and segment SNR (SSNR) [22]. The proposed neural network-based beamforming method is referred as the NN-BF, and the method combining post-filtering is called the NN-BF-PF. The average results of speech quality and intelligibility are shown in Table 2, where Δ represents the improvement to the received speech at reference microphone.

From the comparison in Table 2, we can see that due to the existence of reverberation and the limitation of the number of microphones, the MVDR method does not achieve better enhancement performance. Compared with the MVDR method, the IRM-CNN method is trained under the coexistence condition of reverberation and noise, and the phase of the noisy speech received by each microphone is also considered in the training process [17], its performance seems to be improved than the MVDR method.

Furthermore, the superiority of the proposed methods can be seen from Table 2. The proposed NN-BF method, whether the IPDs in the input features or the PSMs in the training target, they all take phase information into the construction of beamformer, which better reflects the spatial features of the target speech. They make NN-BF better in the same environment than MVDR and IRM-CNN. The proposed NN-BF-PF method is based on the NN-BF method and fused with post-filtering based on the PSMs and phase correction, which further improves the performance by using phase information.

5. Conclusions

In this paper, a new method for multi-channel speech enhancement was proposed, which is based on the spatial and phase information. The proposed method was divided into MVDR beamforming based on deep neural network and spatial information, and single-channel post-filtering based on deep neural network and phase correction. In beamforming process, a more comprehensive spatial and phase information, such as IPDs and PSMs, of the speech source were considered for building the MVDR beamformer. In the post-filtering, phase correction technology and DNN-based PSMs estimation were combined to further improve enhanced speech performance. In comparison with the conventional speech enhancement methods, the proposed method can significantly improve the speech quality and intelligibility by making full use of spatial features and phase information. In the future work, we will further study how to use DNNs combined with spatial features and phase information to construct a more accurate beamformer, so as to achieve more robust speech enhancement methods.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.61831019, No.61471014).

7. References

[1] S. Gannot, E. Vincent, S. Markovich-Golan and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692-730, April 2017.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," in *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830-1847, July 2004.

[4] A. Spriet, M. Moonen, and J. Wouters, "The impact of speech detection errors on the noise reduction performance of multichannel Wiener filtering and generalized sidelobe cancellation," in *Signal Processing*, vol. 85, no. 6, pp. 1073-1088, June 2005.

[5] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," in *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408-1418, Aug. 1969.

[6] T. Higuchi, N. Ito, T. Yoshioka and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 5210-5214.

[7] J. Heymann, L. Drude and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 196-200.

[8] X. Zhang, Z. Wang and D. Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 276-280.

[9] E. A. P. Habets and S. Gannot, "Generating Sensor Signals in Isotropic Noise Fields," in *Journal Acoust. Soc. of America*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[10] J. Lee, K. Kim, T. Shabestary and H. Kang, "Deep bi-directional long short-term memory based speech enhancement for wind noise reduction," *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, 2017, pp. 41-45.

[11] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Intl. Conf. on Machine Learning (ICML)*, Haifa, Israel, 2010, pp. 807–814.

[12] Y. Wang, A. Narayanan and D. Wang, "On Training Targets for Supervised Speech Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

[13] H. Erdogan, J. R. Hershey, S. Watanabe and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, 2015, pp. 708-712.

[14] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments," in *Computer Speech & Language*, vol. 49, pp.37-51, 2018.

[15] M. Souden, J. Benesty and S. Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260-276, Feb. 2010.

[16] R. Cheng, C. Bao and Y. Xiang, "Speech Enhancement with Phase Correction based on Modified DNN Architecture," *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, HI, USA, 2018, pp. 1222-1227.

[17] S. Chakrabarty, D. Wang and E. A. P. Habets, "Time-Frequency Masking Based Online Speech Enhancement with Multi-Channel Data Using Convolutional Neural Networks," *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, 2018, pp. 476-480.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[19] E. A. P. Habets. (2016) Room Impulse Response (RIR) generator. [Online]. Available: https://github.com/ehabets/RIRGenerator.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Salt Lake City, UT, USA, 2001, pp. 749-752 vol.2.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125-2136, Sept. 2011.

[22] John H L, Bryan H, Pellom L,“ An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms,” *International Conference on Speech & Language Processing*, 1988, pp. 2819-2822.