A DEEP LEARNING BASED BINAURAL SPEECH ENHANCEMENT APPROACH WITH SPATIAL CUES PRESERVATION

Xingwei Sun[⋆†] *Risheng Xia*[⋆] *Junfeng Li*[⋆†] *Yonghong Yan*[⋆†‡]

[⋆]Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Sciences
[†]University of Chinese Academy of Sciences
[‡]Xinjiang Key Laboratory of Minority Speech and Language Information Processing,
Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

ABSTRACT

The studies of binaural hearing indicated considerable benefits of the spatial information of sound sources in speech understanding in noise. In this paper, we propose a binaural speech enhancement approach based on deep neural network. In this approach, the signals at the left and right channels are regarded as the real and imaginary parts of a monaural complex signal, a complex ideal ratio mask is accordingly introduced and then further estimated using the complex deep neural network, followed by applying to the monaural complex signal. Experimental results showed that the suggested binaural speech enhancement approach is able to effectively suppress multiple interfering signals and preserve the binaural cues of target signal.

Index Terms— Binaural speech enhancement, complex deep neural network, complex ideal ratio mask, binaural cue preservation

1. INTRODUCTION

As speech is usually contaminated by background noise and interferences in real environments, speech enhancement techniques have been extensively studied in the past several decades. According to the number of microphones involved, speech enhancement techniques include single-channel approaches which exploit the different characteristics of speech and noise in the time-frequency domain to suppress noise, and multi-channel approaches which additionally utilize the spatial information of target speech and interference [1]. These traditional speech enhancement approaches usually generate single-channel output. However, studies of binaural hearing showed that speech understanding in noise can greatly benefit from the differences in binaural cues of target and interfering signals, and it is necessary for localizing target source by preserving its binaural cues [2].

To enhance target speech and preserve its binaural cues, many studies have been conducted by extending the traditional speech enhancement approaches from monaural to binaural output scenarios, especially in the two-input two-output cases [3, 4, 5, 6, 7]. The constrains of binaural cues preservation at the system outputs have been applied to the generalized sidelobe canceller (GSC) beamformer and the multi-channel Wiener filter (MWF) [3, 5]. Another kind of approaches is to extend single-channel noise reduction technique to the binaural case by introducing some constrains to the noise-reduction filters. The spectral subtraction is extended to binaural noise reduction with a constraint on noise suppression in each frequency band to preserve the spatial cues of the target speech [4]. The time-variant Wiener filter is extended to binaural speech enhancement [6] based on the equalization-cancellation (EC) theory [8]. In [7], the two-channel inputs are first combined into a complex single-channel signal, and the widely linear estimation theory is then applied to derive the optimal noise-reduction filters with the constraint of binaural cues preservation. Similarly, the cue-preserving linear minimum mean-square error (MMSE) filter for based on the general concept of the common spectral gain function is proposed in [9].

In recent years, deep neural network (DNN) based speech enhancement methods have demonstrated the improved performance in single- and multi-channel setups [10]. Inspired by the strongness of the DNN-based speech enhancement methods, in this paper, we propose a novel two-input two-output speech enhancement approach to reduce the interfering signals and preserve binaural cues of the target signal. Specifically, the binaural speech signals are first combined into a complex signal with its real part corresponds to the left channel and the imaginary part corresponds to the right channel. To enhance the target speech in the complex domain, a complex mask is further presented and then estimated using the complex deep neural network (cDNN). The estimated mask is finally applied to the complex input signal to enhance the target signal and preserve its binaural cues. Experimental results showed that this cDNN-based binaural speech enhancement provides the SNR improvement of more than 10dB and preserve the binaural cues well.

2. BINAURAL SIGNAL MODEL

In noisy environments, speech signal is often corrupted by background noise and interfering signals. Consequently, the observed signals, $x_L(k)$ and $x_R(k)$, at the discrete-time index k of left and right ears, can be written as

$$x_L(k) = h_L(k) \otimes s(k) + n_L(k) = s_L(k) + n_L(k), \quad (1)$$

$$x_R(k) = h_R(k) \otimes s(k) + n_R(k) = s_R(k) + n_R(k), \quad (2)$$

where $h_L(k)$ and $h_R(k)$ denote the head-related impulse responses (HRIRs) from the speech source $s(k)$ to two ears, $\otimes$ denotes the linear convolution, and $n_L(k)$ and $n_R(k)$ are the noise signals which might be a combination of multiple interference signals and background noise.

In this paper, we aim to attenuate the noise signals $n_L(k)$ and $n_R(k)$ as much as possible and preserve $s_L(k)$ and $s_R(k)$ with their spatial information. Therefore, it is possible to localize the target sound source after binaural enhancement processing using the preserved binaural cues. In this study, the direction of the target signal is assumed to be known *a priori*. However, no restriction is imposed on the number and location of the interference noise sources.

3. BINAURAL SPEECH ENHANCEMENT USING THE COMPLEX IDEAL RATIO MASK

The block diagram of the proposed binaural speech enhancement system is plotted in Fig. 1. In this proposed system, the binaural signals are first combined into a complex monaural signal, and the combined cIRMs is then estimated using the cDNN model and further applied to the input noisy signals to enhance the target speech.

3.1. Acoustic features

To inherently preserve the binaural cues of target speech, the proposed binaural speech enhancement approach transforms the problem of binaural speech enhancement to the single-channel speech enhancement in the complex domain by combining the binaural signals into a complex monaural signal, defined as

$$x_C(k) = x_L(k) + jx_R(k) = s_C(k) + n_C(k) \quad (3)$$

where $s_C(k) = s_L(k) + js_R(k)$ is the complex desired signal, and $n_C(k) = n_L(k) + jn_R(k)$ is the complex additive noise.

The complex spectra of noisy speech signal at the left and right channels are used as input acoustic features. In the complex monaural signal model, it can be given by

$$X_C(t, f) = X_L(t, f) + jX_R(t, f) \quad (4)$$

where $X_L(t, f)$ and $X_R(t, f)$ are the complex spectra at left and right channels with the real and imaginary components combine together, t and f denote the frame index and the frequency bin index, respectively.

3.2. Complex ideal ratio mask as training target

In the proposed binaural speech enhancement approach, the complex ideal ratio mask (cIRM) [11] is suggested as the training target which is calculated by the complex spectra of noisy and target speech signals.

Let S and X denote the complex spectra of the target and noisy speech signals, M is the cIRM. The cIRM could be derived as (t and f are omitted for simplicity)

$$M = \frac{S}{X} = \frac{S_r + jS_i}{X_r + jX_i} = \frac{X_r S_r + X_i S_i}{{X_r}^2 + {X_i}^2} + j\frac{X_r S_i - X_i S_r}{{X_r}^2 + {X_i}^2} \quad (5)$$

In our binaural speech enhancement approach, the cIRMs at left and right channel are calculated respectively, and further combined into the complex monaural cIRMs which is used as the complex training target given by

$$M_C = M_L + jM_R \quad (6)$$

where M_L and M_R are the cIRMs at left and right channels with the real and imaginary components combine together.

3.3. Estimating cIRMs with complex deep neural network

Since both the acoustic features and the training target are complex, it is therefore suggested to use complex deep neural network for training. The concept of cDNN was originally proposed in [12] and further developed to perform monaural speech enhancement [13]. Similar to the traditional feedforward DNN, the forward pass to compute the $(l+1)$th layer unit value from the lth layer of cDNN is expressed as

$$\begin{aligned} H^{l+1} &= F(W^l H^l + b^l) \\ &= F((W_r^l + jW_i^l)(H_r^l + jH_i^l) + (b_r^l + jb_i^l)) \end{aligned} \quad (7)$$

where W and b denote the wight matrix and bias which are both complex-valued parameters, F is the activation function. The cDNN can be trained using back propagation algorithm as the traditional feedforward DNN.

The enhanced binaural signal is finally obtained by applying the cIRMs estimated by cDNNs to the noisy input signals, followed by the inverse short-time Fourier transform (STFT) process. It is noted that the binaural cues of the target source are well preserved since the two-input signals are processed simultaneously.

4. EXPERIMENTS AND RESULTS

4.1. Sound data

For both training and testing datasets, we use the TIMIT corpus [14] which have been divided into the training and testing sets with each including 4680 and 1620 sentences. To obtain noise-independent models, three noises (m109,

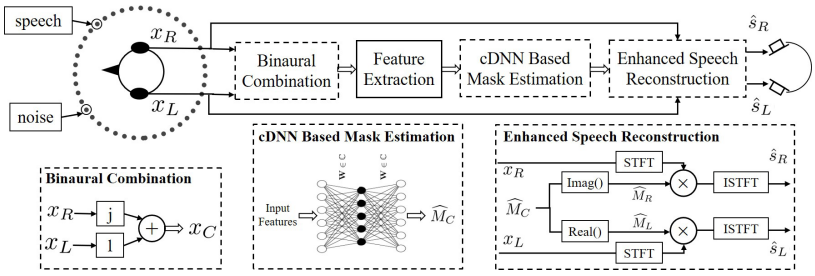

Fig. 1. Block diagram of the proposed binaural speech enhancement system.

leopard and factory1) from the NOISEX92 dataset [15] are used for testing and the other 12 noises are used for training. These signals are then convolved with the HRIRs measured at the MIT Media Laboratory (`http://sound.media.mit.edu/KEMAR.html`) to generate the binaural target and interference signals. We generate seven speech sets with the target source situated at different azimuths $\theta_s = \{-90°, -60°, -30°, 0°, 30°, 60°, 90°\}$, where $0°$ indicates the source in front of the listener and the negative azimuths correspond to the left-hand side while the positive ones correspond to the right-hand side. The binaural noisy signals are generated by adding the scaled binaural interference signals to the binaural target signals with the the signal to noise ratio (SNR) set to 0dB in the ear closest to the noise source. Every sentence in the speech set is contaminated by one of the interfering signals from seven different azimuths $\varphi_n = \{-90°, -60°, -30°, 0°, 30°, 60°, 90°\}$, and every speech set is used to train a cDNN model. By randomly select the noise type and azimuths, we generated 55440 and 5040 sentences in every training and testing set which is about 50 and 4 hours.

4.2. Implementation of the binaural speech enhancement approaches

In the implementation of the proposed binaural speech enhancement approaches, the sampling frequency of the binaural target and interference signals are both 16 kHz. STFTs are computed by first dividing a signal into 64 ms time frames with 32 ms frame shift and then compute the 1024-point fast Fourier transform within each time frame. The cDNN includes 4 hidden layers, each with 1024 complex-valued units. Weights are randomly initialized and bias initialized to zero. The rectified linear unit activation function [16] is used for the hidden layers and the linear activation function is used for the output layer. The Adam optimizer [17] is utilized for back propagation with the learning rate set to 0.0005 and multiply 0.7 every epoch when the loss of validation set didn't decrease. The mean squared error serves as the objective function. We also employ the dropout technique [18] to avoid overfitting with the dropout rate set to 0.5. The batch size is 500 and the total number of training epochs is 20. To incorporate temporal context, we use an input window that spans 7 frames (3 before and 3 after) of the input features to predict one frame of the masks.

As a comparative speech enhancement approach to the cDNN based method, we also trained the traditional feedforward DNN models for two channels separately. In other word, a DNN model is trained with the real parts of the acoustic feature in Eq.(4) and training target in Eq.(6), so as the imaginary parts. This approach is referred as *DNN-mona* in the experiment. The DNN models are trained in the same way with the cDNN models. The binaural speech is reconstructed after the two channels enhanced separately. We also compared our approach with the traditional signal processing method, two-input two-output spectral subtraction approach, proposed in [19] which is referred as *Twoch-ss*. Our proposed cDNN based method is referred as *cDNN-bina*.

4.3. Experimental evaluations for speech enhancement

To examine the efficacy of the proposed algorithm for enhancement, we performed evaluations in single-noise (S_0N_φ) with the interference direction φ varied from $-90°$ to $90°$ in increments of $30°$, and multi-noise (2a: $S_0N_{-60,60}$, 2b: $S_0N_{-30,60}$, 2c:$S_0N_{-60,30}$, 3a:$S_0N_{-60,0,60}$, 3b:$S_0N_{-30,60,90}$, 4a:$S_0N_{-60,-30,30,60}$) conditions. In our experiments, the improvement in SNR ($\triangle SNR$) is used [20]. The $\triangle SNR$ is defined as the difference value of SNRs between the output enhanced signal and the input noisy signal. A higher $\triangle SNR$ means a higher improvement in speech quality by speech enhancement processing.

Fig.2 portrays the $\triangle SNRs$ averaged across all utterances of the testing set, as processed using the three different approaches. The results in Fig.2(a) and (b) show that the cDNN based binaural speech enhancement method produce greatest $\triangle SNRs$, and that these $\triangle SNRs$ vary with the incoming direction of the interference signal in the single-noise condition. Specifically, the $\triangle SNRs$ are much higher when the interference signal is close to the ear with which the enhanced signal is under evaluation. This is the case in which the input signals are more noisy with low SNRs. Fig.2(c) and (d) indicate

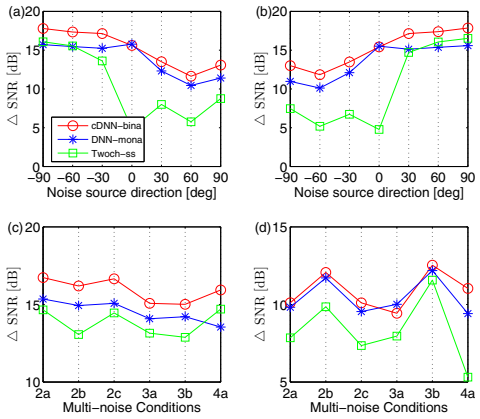

Fig. 2. The $\triangle SNRs$ at the left ear in the single-noise (a) and the multi-noise (c) conditions; The $\triangle SNRs$ at the right ear in the single-noise (b) and the multi-noise (d) conditions.

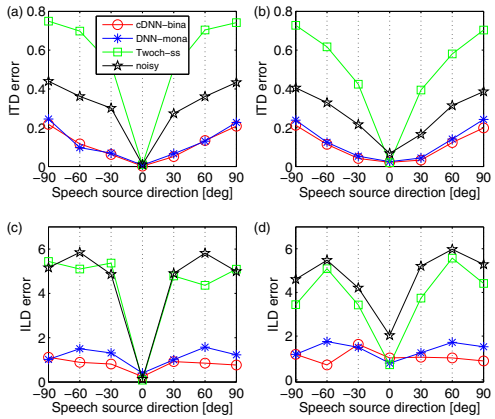

Fig. 3. The ITD errors in the single-noise (a) and the three-noise (b) conditions; The ILD error in the single-noise (c) and the three-noise (d) conditions.

that great $\triangle SNRs$ can be achieved in multi-noise condition and the cDNN based method has the best performance. Similar performance can be find that the left ear achieved greater $\triangle SNRs$ than the right ear which is also because of the input signals in the left ear are more noisy.

4.4. Experimental evaluations for binaural cues preservation

The binaural cue preservation evaluations are performed in single-noise ($S_\theta N_0$) and three-noise ($S_\theta N_{-30,60,90}$) conditions with the target source direction θ varied from -90° to 90° in increments of 30° and the noise at the fixed position(s). The ITD error (E_{ITD}) and the ILD error (E_{ILD}) of the outputs [21] are used in this experiment. The E_{ITD} and E_{ILD} are defined as the difference value of the ITD and ILD between the output enhanced signal and the input noisy signal. The smaller E_{ITD} and E_{ILD} are, the higher the performance of the tested algorithm in binaural cue preservation is.

The results in E_{ITD} and E_{ILD} averaged across all tested utterances processed using the three different approaches and those of the noisy speech are shown respectively in Fig.3. Better performance can be seen for cDNN and DNN based methods compare with the two-input two-output spectral subtraction approach and the cDNN based method has slight advantage than the DNN based method. From Fig.3(a) and (c), symmetry of E_{ITD} and E_{ILD} along with the median plane is observed in the single-noise condition which is mainly because the symmetry of the HRIRs against the median plane. This means that the binaural signals from the sources localized at the median plane involve the equivalent binaural cues. Consequently, the binaural cues of the target signal equal to those of the interference signals in the S_0N_0 scenario, in which our algorithms yield almost no benefit in reducing E_{ITD} and E_{ILD}. In other cases in which the target signal is not on the median plane, the proposed approach shows satisfiable E_{ITD} and E_{ILD} reduction. The results in the three-noise conditions shown in Fig.3(b) and (d) show that the E_{ITD} and E_{ILD} have been decreased in all directions of the target speech include 0°. This may because of the interference signals are not only placed at the 0° as the single-noise condition so as the interferences have different affect on the left and right ears.

5. CONCLUSION

In this study, we proposed a binaural speech enhancement algorithm which aims to reduce the interference noise and preserve the binaural cues. In our approach, the binaural signals are combined into a monaural complex signal, a complex mask is estimated using the complex deep neural network and then applied to the monaural complex signal. The effectiveness of the suggested approach is proved by objective SNR improvement and ITD and ILD errors evaluations in the single-noise and the multi-noise conditions.

6. ACKNOWLEDGMENT

This work is partially supported by the National Key Research and Development Program (Nos. 2017YFB1002803, 2016YFB0801203, 2016YFC0800503), the National Natural Science Foundation of China (Nos. 11590770-4, 61650202, 11722437, U1536117, 61671442, 11674352, 11504406, 61601453) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 2016A03007-1).

7. REFERENCES

[1] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, Alexey Ozerov, Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[2] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[3] Joseph G Desloge, William M Rabinowitz, and Patrick M Zurek, "Microphone-array hearing aids with binaural output. i. fixed-processing systems," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 529–542, 1997.

[4] B Kollmeier, J Peissig, and V Hohmann, "Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain.," *Scandinavian Audiology. Supplementum*, vol. 38, pp. 28–38, 1993.

[5] Thomas J Klasen, Tim Van den Bogaert, Marc Moonen, and Jan Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, 2007.

[6] Junfeng Li, Shuichi Sakamoto, Satoshi Hongo, Masato Akagi, and Yôiti Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[7] Jacob Benesty, Jingdong Chen, and Yiteng Huang, "Binaural noise reduction in the time domain with a stereo setup.," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 8, pp. 2260–2272, 2011.

[8] Nathaniel I Durlach, "Binaural signal detection-equalization and cancellation theory.," 1972.

[9] G. Enzner, M. Azarpour, and J. Siska, "Cue-preserving mmse filter for binaural speech enhancement," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[10] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[11] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483–492, 2016.

[12] Henry Leung and Simon Haykin, "The complex backpropagation algorithm," *IEEE Transactions on signal processing*, vol. 39, no. 9, pp. 2101–2104, 1991.

[13] Yuan-Shan Lee, Chien-Yao Wang, Shu-Fan Wang, Jia-Ching Wang, and Chung-Hsien Wu, "Fully complex deep neural network for phase-incorporating monaural source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 281–285.

[14] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[15] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[16] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[17] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] Matthias Dorbecker and Stefan Ernst, "Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation," in *European Signal Processing Conference, 1996. EUSIPCO 1996. 8th*. IEEE, 1996, pp. 1–4.

[20] Thomas P Barnwell III, MA Clements, and SR Quackenbush, "Objective measures for speech quality testing," 1988.

[21] Tim Van den Bogaert, Jan Wouters, Simon Doclo, and Marc Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel wiener filter," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–565.