inter.noise
27-30 AUGUST HONG KONG 2017

Combining beamforming and deep neural networks for multi-channel speech extraction

Haoran Zhou; Jing Lu

Key Lab of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China

ABSTRACT

Post-filtering is a popular technique for multi-channel speech enhancement system, which is aimed to further improve the output signal to noise ratio (SNR) after beamforming. However, the commonly used post-filtering methods require a considerably cumbersome parameter tuning procedure in order to achieve a reasonable trade-off between noise suppression and speech quality, especially when the speech is corrupted by nonstationary noises, e.g., interfering speech from other speakers. In this paper, a scheme to extract desired speech from interference is proposed based on the combination of beamforming and deep neural networks (DNN). The 'mapping' from the beamformers' outputs to the enhanced signal is learned by DNN. The 'mapping' is divided into two stages heuristically, where features from each beamformer's output are extracted at first, followed by the inference of clean speech from all the features. The experimental system is established in TENSORFLOW, and the results demonstrate the benefits of the proposed method and show the system's generalization ability in acoustic environments which are not presented in the training set.

Keywords: Speech enhancement, DNN, Beamforming

1. INTRODUCTION

Microphone arrays have been widely utilized in hearing aids, distant speech recognition and etc. [1], and post-processing is often a crucial step in such application scenarios to further suppress noise and interference when the array is of limited size with limited number of microphones. The commonly used generalized sidelobe canceller (GSC) [2] and multichannel Wiener filter (MWF) [3] usually requires an effective voice activity detection (VAD) subroutine which is usually unreliable in complex scenarios with interfering speech signals. The multi-channel optimally modified log spectral amplitude (OMLSA) [4] is another feasible post-processing scheme. Although it theoretically works without an explicit VAD, the cumbersome parameter tuning procedure to estimate the speech presence probability makes it hard to guarantee a satisfactory performance in different noisy conditions.

Deep neural networks have been utilized by researchers in acoustical modeling for speech recognition [5], adaptive beamforming [6], ideal ratio mask (IRM) inference [7], and have shown superior results compared to conventional methods without the need to worry about the modeling details. Different from [6], where the generalized cross correlation (GCC) is fed into the DNN to produce the beamforming filters, and [7], where the beamformer is optimized by the feedback of the VAD and IRM information, the speech extraction strategy used in our work is more of a multi-channel post-processing approach and share similar structures like the multi-channel OMLSA algorithm.

In this paper, the combination of beamforming and DNN is investigated with the goal of extracting the desired speech in environments with interfering speech from other speakers. Two beamformers based on a circular array of five microphones are designed, then the convolutional neural network (CNN) [8] and the stacked long short term memory (LSTM) [9] network are trained with a massive generated dataset using a multi-task learning approach to infer the desired signal from the beamformer's output. The features of both beamformer's output are fed into the DNNs respectively and then merged into a single channel feature to make the final inference. Since the rules of enhancement are greatly simplified by the beamforming procedure, the DNNs converge quickly and generalize well in acoustic environments that are not presented in the training set. Both

the simulated test set and the experimental data acquired using an MEMS microphone array demonstrate a promising performance of the proposed system.

2. Multichannel speech enhancement using DNNs

2.1 Array configuration, beamforming and feature extraction

The microphone array used in this paper is shown in Figure 1, where one microphone is located at the center and 4 others are distributed equally on a circle of radius 0.27 m. For simplicity, only two beamformers, each with a 90 degree beamwidth, are designed in the prototype system using the superdirective beamforming method [10]. The first beamformer is steered at 0 degree and the desired speaker is assumed to be in its beamwidth. The second beamformer is steered at 180 degree to pick up possible interference from 135 degree to 225 degree. For practical applications, interference from other directions should also be captured and the networks' structure in the following sections should be expanded accordingly. Other topological structure of array is also feasible and the spatial aliasing are allowed because DNNs are capable of dealing with the aliasing effects using the time frequency context.

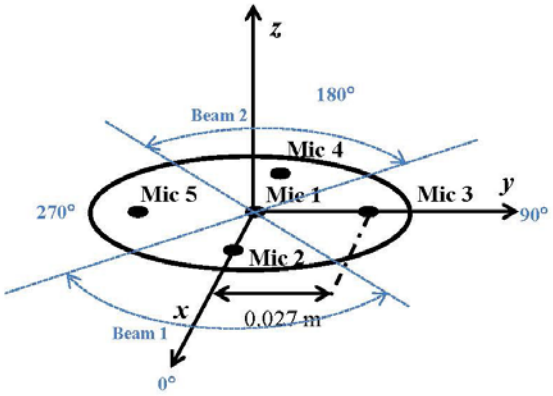

Figure 1 – Array configuration and fixed beamforming.

The outputs of the beamformers are then transformed into the frequency domain by $N_{\mathrm{FFT}} = 256$ points FFT using a hamming window with a frame shift of 64, with an effective feature vector size of $N_{\mathrm{eff}} = 129$. The features are transformed to the logarithmic scale as it is proved to be well related to the human auditory characteristics. Feature normalization is done using the group mean and group variance, namely, the mean and variance of all the data in the training set. Denote $\mathbf{b}_{t,1,}\,\mathbf{b}_{t,2}$ as the preprocessed output of the first beamformer and the second beamformer at time t. For CNN, 8 frames of the data, $\mathbf{b}_{t,1,}\,\mathbf{b}_{t,2,}\,\mathbf{b}_{t-1,1,}\,\mathbf{b}_{t-1,2,}\,\ldots\mathbf{b}_{t-7,1,}\,\mathbf{b}_{t-7,2}$ are packed as the input of the network and for stacked LSTM network, $\mathbf{b}_{t,1,}\,\mathbf{b}_{t,2}$ are presented as the input of the network at a time.

2.2 Training objective

Fed with one or several synchronous pairs of frames from the spectrograms of the beamformers' outputs, the network is expected to output the clean spectrum inference $\mathbf{y}_{\mathrm{i},t}$ and the frequency bin-wise VAD $\mathbf{v}_{\mathrm{i},t}$, where t denotes the time index of the last frame of the input pairs. This kind of multi-task learning strategy is widely used in speech signal processing [11] and incorporating a VAD inference task is found beneficial for the clean spectrum inference. The whole process is causal and will not introduce delay in practical use. The loss function of the clean spectrum inference task is defined as the l_2 loss between the clean spectrum inference and the training target spectrum

$$\mathrm{E}\left[\left\|\mathbf{y}_{\mathrm{r},t}-\mathbf{y}_{\mathrm{i},t}\right\|_2^2\right], \tag{1}$$

where $\mathbf{y}_{\mathrm{r},t}$ is the training target spectrum. The loss function of the VAD inference task is defined as the sum of the cross entropy between VAD inference and VAD target of each frequency bin

$$\mathrm{E}\left[\sum_{j=0}^{N_{\mathrm{eff}}-1} -\mathbf{v}_{\mathrm{r},t}(j)\log\left(\mathbf{v}_{\mathrm{i},t}(j)\right)-\left(1-\mathbf{v}_{\mathrm{r},t}(j)\right)\log\left(1-\mathbf{v}_{\mathrm{i},t}(j)\right)\right], \tag{2}$$

with $\mathbf{v}_{\mathrm{r},t}$ is the training target VAD. The definition of the target spectrum and VAD are described in Section 2.5.

2.3 Multichannel speech extraction using a CNN

The CNN used is similar to that used in [12], where only one dimensional convolution is used because it is found to be more efficient for this task. Concretely, the convolution is performed only in frequency axis rather than in the frequency axes and the other axis at the same time. The structure of this network is shown in Figure 2.

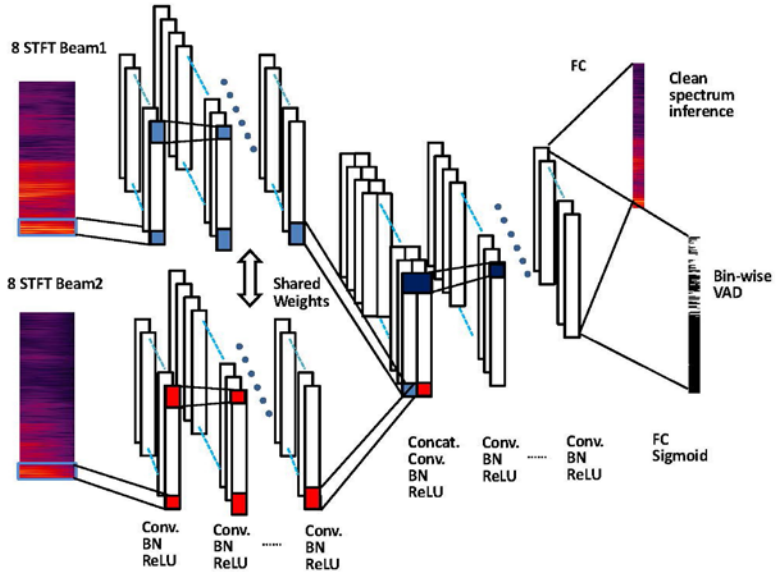

Figure 2 – Multichannel speech extraction using a CNN.

The inputs of the CNN are the synchronous pair of 8 continuous frames of STFT from the two beamformers' output. The first 4 layers of the network operate on each channel separately. By sharing the weights between the two channels, the network is expected to extract features that can apply to all channels. The outputs of the 4th layer are concatenated to form a unified feature as the input of the next 5 convolution layers. The clean spectrum inference is obtained through a fully connected (FC) layer and the bin-wise VAD inference through a logistic regression layer. All the one dimensional convolutions are operated with a ReLU activation function, and the batch normalization (BN) [13] is used in each convolution layer which can greatly facilitate the convergence and generalization of the network. A detailed description of the network configuration can be viewed in Table 1.

Table 1 – Layer configuration of the CNN

Layer num.	**1 - 4**	**5 - 9**	**10**
Layer configuration	**Type:** double Conv. Layer (with ReLU, BN) **Parameters:** (feature maps, filter length) (12, 13) ,(16, 11),(20, 9),(24,7)	**Type:** Conv. Layer (with ReLU, BN) **Parameters:** (feature maps, filter length) (32, 7) ,(24, 7),(20, 9),(16,11),(12, 13)	**Type:** FC layer FC layer (with sigmoid) **Parameters:** output size 129

2.4 Multichannel speech extraction using a stacked LSTM network

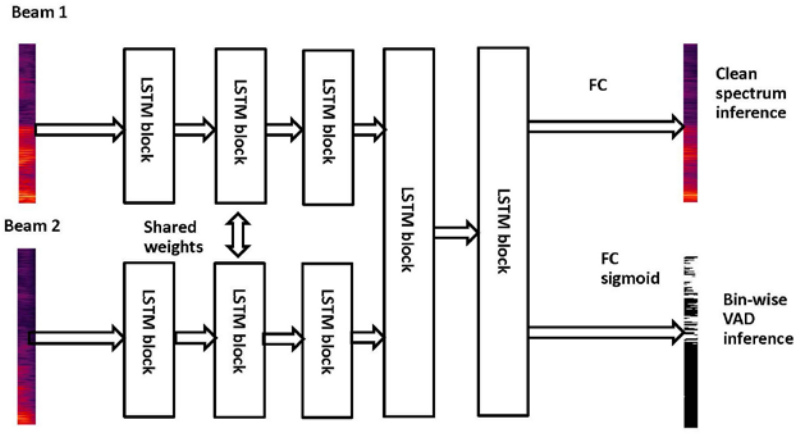

Figure 3 – Multichannel speech extraction using a stacked LSTM network.

Similar to the strategy used in CNN, the stacked LSTM network also merge the features of all channels after extracting them separately on each channel. The weights of the first 3 layers are shared across channels while the memory cells of each channel are independent. Layer normalization (LN) [14] is used in each LSTM layer. Continuous synchronous pairs of log-spectral amplitudes are fed sequentially into the network with only one pair at a time. A detailed description of the network configuration can be viewed in Table 2

Table 2 – Layer configuration of the stacked LSTM network

Layer num.	**1 - 3**	**4 - 5**	**6**
Layer configuration	**Type:** double LSTM cells (with LN) **Parameters:** nodes per layer 256 for each channel	**Type:** LSTM cell (with LN) **Parameters:** nodes per layer 256	**Type:** FC layer, FC layer (with sigmoid) **Parameters:** output size 129

2.5 Dataset, preprocessing and waveform reconstruction

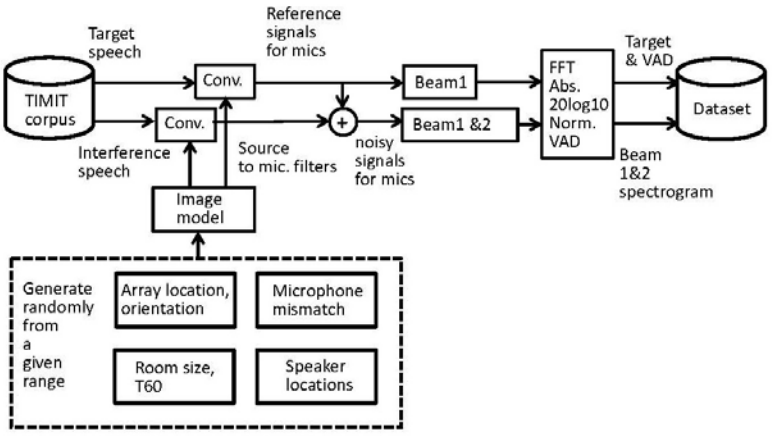

Figure 4 – Generation of the dataset.

The method of generating the dataset is shown in Figure 4. The TIMIT database [15] is used to generate the reverberant data with image mode [16] for each microphone and then the output of each beamformer. Of all the 6318 utterances in the TIMIT, 4620 utterances are assigned to generate the training set, 800 utterances are assigned to generate the cross validation set and the rest 898 utterances are assigned to generate the test set. The procedure of generate the dataset is summarized as follows:

(a) Choose a target utterance and an interference utterance from the corpus, e.g., for the training set, each one of the 4620 utterances is selected as the target speech and the inference signal is chosen randomly from the 4619 utterances left.
(b) Sample the parameters from the given range (Table 3) assuming a uniform distribution of all the variables.
(c) Using the parameters chosen above, the transfer functions from the target and interference speakers to the microphones can be computed using the image model [16].
(d) The beamformer output of the noisy signal and the target signal can be computed using the signal received by each microphone generated in (c). The log-spectral amplitudes of the beamformer output of the target speaker are saved as the clean spectrum targets and the mix spectrums of the target speaker and interference speaker are saved as the input data. The reference bin-wise VAD is considered active if the corresponding time frequency point in clean spectrum targets is greater than 1 / 1000 of the maximum magnitude observed in the whole clean spectrum.
(e) Repeat (a) to (d) to generate the whole dataset.

Table 3 – Range of the parameters for simulating the microphone signals

Parameter	**Room size**	**T60**	**Horizontal distance from the speaker to the array**
Lower bound	4.5 m × 4 m × 2 m	0.3 s	0.5 m
Upper bound	8.5 m × 8 m × 3.5	0.6 s	2.1 m
Parameter	**Height of the speech source**	**Angle of the target speaker relative to the array**	**Angle of the interference speaker relative to the array**
Lower bound	0.8 m	-45 degree	135 degree
Upper bound	1.7 m	45 degree	225 degree
Parameter	**Amplification factor applied to the interference source**	**Height of the array**	**Deviation of the microphone amplitude response**
Lower bound	0.25	0.8 m	-1.5 dB
Upper bound	1.25	1.5 m	1.5 dB
Parameter	**Array position**	**Array orientation**	
Range	All feasible positions and orientations given the parameters set above		

The method of reconstruction of the waveform is illustrated in Figure 5, where phase information remains unchanged during the whole process and the time domain signal is reconstructed using IFFT and adding the overlapping parts.

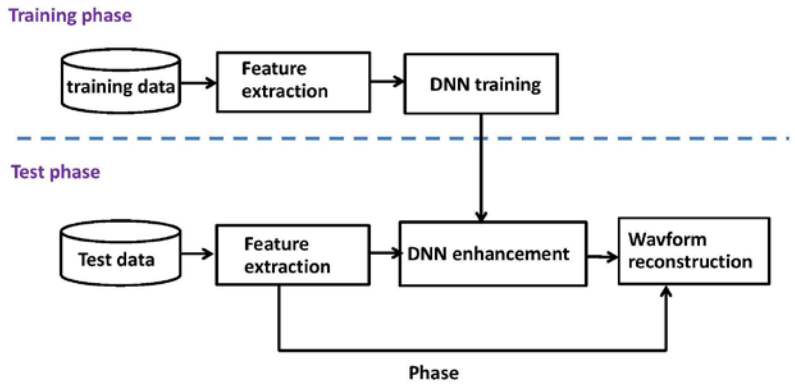

Figure 5 – Waveform reconstruction.

2.6 Learning Parameters

All the weights are initially sampled from a Gaussian distribution with the mean of 0 and the standard deviation of 0.1, and values more than 2 standard deviations from the mean are re-picked. For the CNN, all bias values are initiated as 0.1 and for the LSTM the bias values are initiated as 0. All networks are trained using Adam [17] with the learning parameters: lr = 0.0001, β_1 = 0.9, β_2 = 0.999, and ε = 1e-8. The mini-batch size is set as 128. The stacked LSTM network is unfolded through 20 time steps and the gradients are truncated if their absolute values are greater than 1.25. If the cross validation loss does not drop for 3 continuous epochs, the training will be terminated. The generalization ability of the CNN can be slightly improved with l_2 regularization (λ = 0.01).

3. Objective evaluation

3.1 Evaluation based on the test set

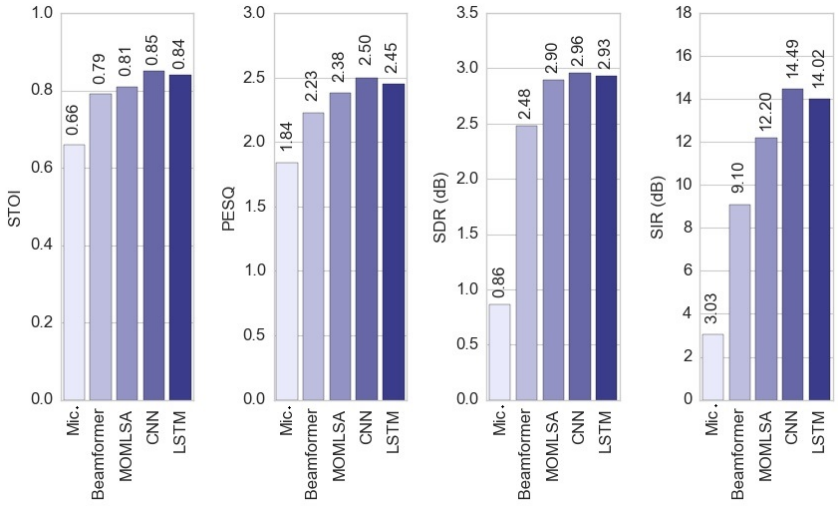

Figure 6 – Objective evaluation based on the test set.

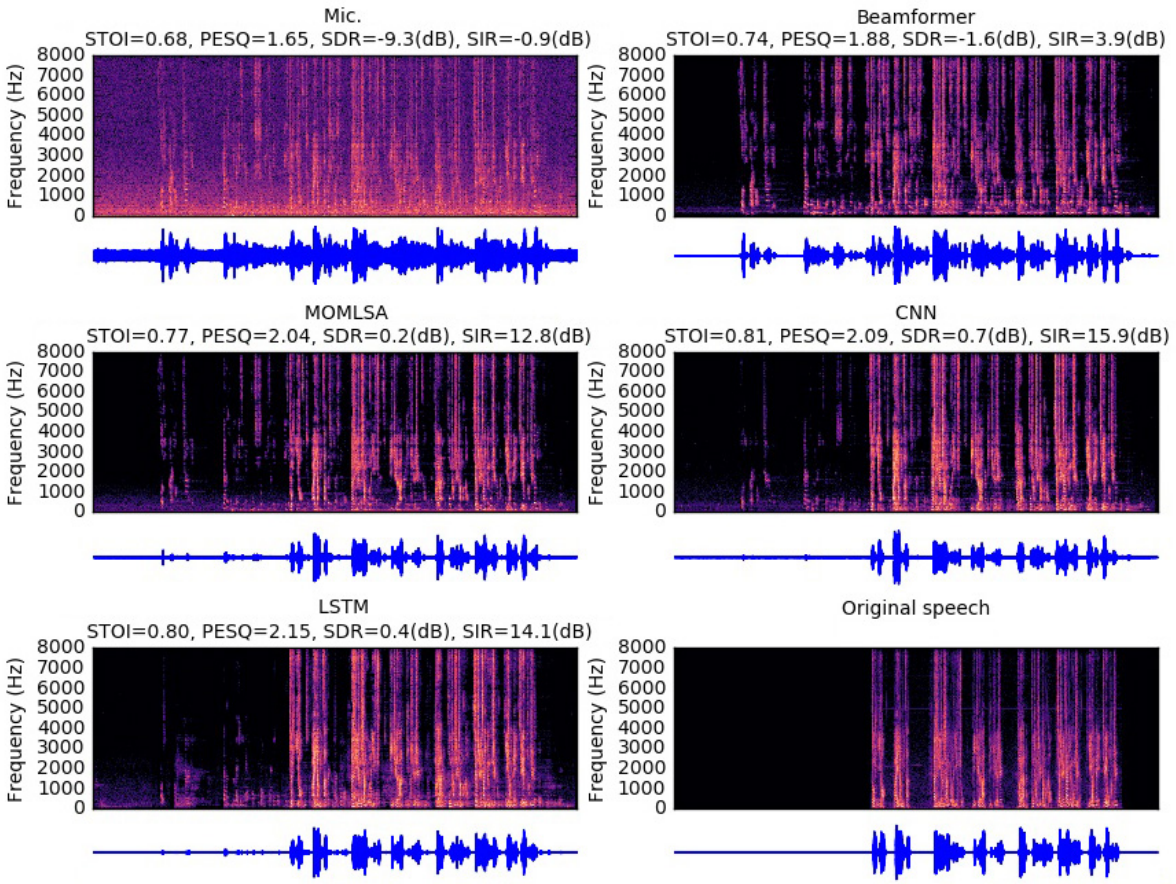

Figure 7 – Spectrograms and objective evaluation of a test sequence. The outputs of the beamformer, MOMLSA, CNN, and LSTM are enhanced using the same single-channel OMLSA algorithm to remove stationary noise.

Several objective metrics, namely the Short time Objective Intelligibility (STOI) [18], Perceptual Evaluation of Speech Distortion (PESQ) [19], Signal to Distortion Ratio (SDR) and Signal to Interference Ratio (SIR) based on the PEASS toolkit [20], are used to evaluate the performance of the proposed method. The signal received by the central microphone, the first beamformer output, the extracted speech of the superdirective beamfomer, the multichannel OMLSA (MOMLSA), the CNN, the stacked LSTM network are evaluated using the hold-out test set and the averaged results are shown in Figure 6. The CNN performs slightly better than the LSTM network, and both outperform the MOMLSA significantly. A possible explanation of the CNN's better performance over the stacked LSTM network is as follows. Speech signals are short-time stationary and capturing the long-time dependency using the LSTM might not benefit the speech extraction process. On the other hand, the deeper structure of CNN helps it to excel at extracting higher level features.

3.2 Evaluation based on the experimental data

Though the networks perform reasonably well on the test set, their performance on real data should also be validated. The experiment is conducted on an MEMS microphone array in a room with a T60 around 0.3 s. The performances of the DNNs are quite consistent during the test and an example is shown in Figure 7. Since the DNNs are not trained to handle stationary noise, e.g. the self-noise of the microphone, the outputs of the DNNs are enhanced using a single channel OMLSA algorithm [21], and the output of the beamformer, the MOMLSA output are also enhanced using the same algorithm for comparison. The DNNs generalize well to the experimental data and achieves greater interference attenuation and less distortion than the MOMLSA algorithm. For frequency bins above 6370 Hz, where spatial aliasing occurs, the MOMLSA is programmed explicitly to handle the spatial aliasing problem using the probability of speech absence below that frequency. On the contrary, the DNNs learn to deal with that automatically, resulting in better interference attenuation

in high frequencies as can be viewed in Figure 7. The CNN outperforms all the other methods with respect to the objective metrics of STOI, SDR and SIR, while the stacked LSTM network achieves the highest PESQ score.

4. Conclusions

In this paper, the DNNs are used to make clean spectrum inference by training with multiple beamformer outputs based on a carefully designed dataset. The DNNs generalize reasonably well to the test set and experimental data, achieving a significant improvement over the commonly used multichannel OMLSA algorithm. The network used in this work is relatively simple and feasible to larger datasets. Furthermore, better performance might be achieved by introducing more advanced structures such as skip connections in ResNet [22] and inception blocks in GoogLeNet [23].

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China with Grant No. 11374156.

REFERENCES

1. Benesty J., Chen J., Huang Y., Microphone Array Signal Processing. Berlin, Gerrmany: Springer-Verlag; 2008.
2. Park J., Kim W., Han D. K., et al., Two-Microphone Generalized Sidelobe Canceller with Post-Filter Based Speech Enhancement in Composite Noise. ETRI Journal. 2016; 38(2): 366-375.
3. Kuklasinski A., Jensen J., Multichannel Wiener Filters in Binaural and Bilateral Hearing Aids—Speech Intelligibility Improvement and Robustness to DoA Errors. Journal of the Audio Engineering Society. 2017; 65(1): 8-16.
4. Cohen, I., Multichannel post-filtering in nonstationary noise environments. IEEE Transactions on Signal Processing. 2004;52(5): 1149-1160.
5. Hinton G., Deng L., Yu D., et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine IEEE. 2012;29(6):82-97.
6. Xiao X., Watanabe S., Erdogan H., et al., Deep beamforming networks for multi-channel speech recognition. Acoustics, Speech and Signal Processing (ICASSP). 2016 IEEE International Conference on; 20-25 March 2016; Shanghai, China 2016. p. 5745-5749.
7. Du J., et al., The USTC–iFlytek System for CHiME-4 Challenge. Proc. CHiME (2016): 36-38.
8. LeCun Y., Bottou L., Bengio Y., et al., Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11): 2278-2324.
9. Hochreiter S., Schmidhuber J., Long short-term memory. Neural computation. 1997; 9(8): 1735-1780.
10. Cox H., Zeskind R., Kooij T.. Practical supergain. IEEE Transactions on Acoustics, Speech, and Signal Processing.1986; 34(3): 393-398.
11. Deng L., Hinton G., Kingsbury B., New types of deep neural network learning for speech recognition and related applications: An overview. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on; 26-31 May 2013; Vancouver, BC, Canada 2013. p. 8599-8603.
12. Park S. R., Lee J., A Fully Convolutional Neural Network for Speech Enhancement. 2016; arXiv preprint, 2016; arXiv:1609.07132.
13. Ioffe S., Szegedy C.. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint, 2015; arXiv: 1502.03167.
14. Ba J., Kiros J., Hinton G., Layer normalization. arXiv preprint, 2016; arXiv: 1607.06450.
15. Garofolo J., Lamel L., Fisher W., et al., DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report n. 1993;93.
16. Allen J., Berkley D., Image method for efficiently simulating small - room acoustics. The Journal of the Acoustical Society of America. 1979;52(5): 65(4): 943-950.
17. Kingma D., Ba J., Adam: A method for stochastic optimization. arXiv preprint, 2014; arXiv: 1412.6980.
18. Taal C. H., Hendriks R. C., Heusdens R., et al., A short-time objective intelligibility measure for time-frequency weighted noisy speech. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on; 14-19 March 2010; Dallas, TX, USA 2010. p. 4214-4217.
19. Rix A. W., Beerends J. G., Hollier M. P., et al., Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Acoustics, Speech, and

Signal Processing, 2001 (ICASSP'01). 2001 IEEE International Conference on; 7-11 May 2001; Salt Lake City, UT, USA 2010. p. 749-752.
20. Emiya V., Vincent E., Harlander N., et al., The PEASS Toolkit-Perceptual Evaluation methods for Audio Source Separation. 9th Int. Conf. on Latent Variable Analysis and Signal Separation; 27-30 Sept. 2010; Saint-Malo, France 2010. inria:00545477.
21. Cohen I., Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Transactions on speech and audio processing; 2003; 11(5): 466-475.
22. He K., Zhang X., Ren S., et al., Deep Residual Learning for Image Recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 26-July 1 2016; Las Vegas, Nevada, USA 2016. p. 770-778.
23. Szegedy C., Liu W., Jia Y., et al., Going Deeper With Convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 7-12 June 2015; Boston, MA, USA 2015. p. 1-9.