Binaural Sound Source Localization Based on Convolutional Neural Network

Lin Zhou[1, *], Kangyu Ma[1], Lijie Wang[1], Ying Chen[1, 2] and Yibin Tang[3]

Abstract: Binaural sound source localization (BSSL) in low signal-to-noise ratio (SNR) and high reverberation environment is still a challenging task. In this paper, a novel BSSL algorithm is proposed by introducing convolutional neural network (CNN). The proposed algorithm first extracts the spatial feature of each sub-band from binaural sound signal, and then combines the features of all sub-bands within one frame to assemble a two-dimensional feature matrix as a grey image. To fully exploit the advantage of the CNN in extracting high-level features from the grey image, the spatial feature matrix of each frame is used as input to train the CNN model. The CNN is then used to predict azimuth of sound source. The experiments show that the proposed algorithm significantly improves the localization performance of BSSL in various acoustic environments, especially to deal with low SNR and high reverberation conditions.

Keywords: Binaural sound source localization, convolutional neural network, two-dimensional spatial feature.

1 Introduction

Sound source localization (SSL) plays an important role in speech signal processing, and has a wide range of applications such as robot navigation, hearing aids, blind source separation and human-machine interface. Microphone-array based SSL methods need more microphones, which seriously increase computational complexity. However, “Cocktail Party Effect” indicates that human can track one sound from multiple sound sources in a noisy environment with limited ‘microphones’ that is two ears. Inspired by this phenomenon, researchers have proposed several methods to extract binaural spatial features from only two microphones for SSL. The BSSL methods do not require a microphone array with large size, which can greatly reduce the computational overhead.

Binaural spatial feature is critical for BSSL. In the early days, Lord Rayleigh [Rayleigh (1907)] proposed the duplex theory based on the assumption of the spherical head. Duplex theory demonstrates that, to locate the source, the human auditory system mainly depends on inter-aural time difference (ITD) and inter-aural intensity difference (IID) of binaural signal. As a consequence, ITD and ILD have been viewed as a useful binaural

[1] School of Information Science and Engineering, Southeast University, Nanjing, 210096, China.

[2] Department of Psychiatry, Columbia University and NYSPI, New York, 10032, USA.

[3] College of Internet of Things Engineering, Hohai University, Changzhou, 213022, China.

[*] Corresponding Author: Lin Zhou. Email: Linzhou@seu.edu.cn.

CMC. doi:10.32604/cmc.2019.05969

feature and extended to more spatial features. Inter-aural envelope difference (IED) [Roman, Wang and Brown (2003)], as a version of spatial features, works well in high frequency. Inter-aural phase difference (IPD) [Nix and Hohmann (2006)] is more commonly used, though it may suffer from a phase ambiguity problem in some case.

Nevertheless, in the real environment, BSSL should be effectively against the noise and reverberation. To deal with this issue, various robust BSSL methods have been presented. For example, Raspaud et al. [Raspaud, Viste and Evangelista (2010)] established a parametric model with ITD and ILD to estimate the sound source location. Pang et al. [Pang, Liu, Zhang et al. (2017)] presented a novel BSSL approach based on reverberation weighting and generalized parametric mapping. Machine learning is also encouraged in BSSL. Chen et al. [Chen and Ser (2009)] introduced a least squares support vector machines (LS-SVMs) approach to improve the localization accuracy. May Tobias et al. [May, Van de Par and Kohlrausch (2011)] divided the acoustic signal into multiple sub-bands and proposed the Gaussian mixed model (GMM) to model the binaural cues. Xiao et al. [Xiao, Zhao; Zhong et al. (2015)] extracted generalized cross correlation (GCC) as the input feature to train a multi-layer perceptron neural network. Roden et al. [Roden, Moritz, Gerlach et al. (2015)] combined ITD, ILD, amplitude and phase spectra to train a deep neural network (DNN) models. Yu et al. [Yu, Wang and Han (2016)] further applied DNN to stereo sound localization. Moreover, DNN can also efficiently map binaural features in each frequency band to the corresponding source azimuth [Ma, Brown and May (2015)]. Later, Ma et al. [Ma, May and Brown (2017)] used DNN and head rotation to achieve multi-sound source localization under reverberation conditions. Ma et al. [Ma, Gonzalez, Brown et al. (2018)] also combined sound spectrum and DNN in the time-frequency (TF) unit to estimate the azimuth. Yiwere et al. [Yiwere and Rhee (2017)] used ILD and CCF as input features to train DNN models.

Although the neural-network-based BSSL algorithms improve the localization accuracy, the performance under the low SNR and high reverberation is still a challenge problem. Previous research models ITD, IID and cross-correlation function (CCF) as one-dimensional feature within one sub-band or one frame, which seldom consider the feature correlation of consecutive sub-bands. Therefore, we attempt to introduce CNN to the existing BSSL framework, where CNN [Cui, McIntosh and Sun (2018)] is well-known and successfully incorporated in image recognition and video analysis. In details, we utilize CNN as a classifier for the two-dimension spatial feature. First, our algorithm calculates the CCF of binaural sound signals in each sub-band. Then, CCF of all sub-bands within one frame is assembled into a two-dimensional feature matrix as a grey map. Sequentially, CNN is trained to establish the relationship between feature matrices and the azimuth of the sound source. The established CNN model is then used to predict the location of binaural testing signals. Experimental results show that the BSSL algorithm based on CNN classification significantly improves localization performance under low SNR and high reverberation environment.

The remainder of the paper is organized as follows. Section 2 presents an overview of our CNN-based BSSL system. Section 3 describes the structure of our CNN network. The simulation results and analysis are provided in Section 4. The conclusion is drawn in Section 5.

2 System overview

The core idea of our algorithm is to exploit the correlation of features among consecutives sub-bands through CNN model. The flowchart of our BSSL system is given in Fig. 1 Binaural signals, including training and testing ones, are used as the system input. Left- and right-ear signals are decomposed into TF units independently by 33-channel Gammatone filters. Then, CCF between left- and right-ear signals are extracted in each TF unit, and assembled to form a two-dimensional spatial feature matrix for each frame. Sequentially, these feature matrices are treated as CNN input. In the training phase, CNN is used to establish the relationship between the spatial feature matrix and the sound azimuth. Afterwards, the estimated azimuth is achieved through trained CNN model with the testing signals in the predict stage.

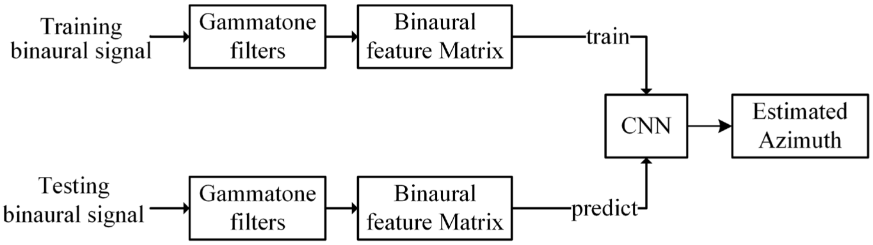


Figure 1: Flowchart of the proposed BSSL system

Here, the clean source is $s(t)$, the left- and right-ear signals, $x_L(t)$ and $x_R(t)$, are defined as binaural signals, with:

$$\begin{aligned} x_L(t) &= h_L * s(t) + n_L(t) \\ x_R(t) &= h_R * s(t) + n_R(t) \end{aligned}, \tag{1}$$

where h_L and h_R represent the binaural room impulse response (BRIR) for left and right ears, '*' denotes linear convolution, $n_L(t)$ and $n_R(t)$ are additive noise for each ear.

We first decompose both left- and right-ear signals into cochleagrams. In detail, the central frequencies of Gammatone filters ranges from 50 Hz to 8000 Hz on the equivalent rectangular bandwidth (ERB): The output of each channel is divided into 20-ms frame length with 10-ms frame shift. Thus, the binaural signals are converted into TF units. In each unit, we extract normalized CCF between the left-ear and right-ear signals. The CCF of a TF unit pair is defined as

$$CCF(i,\tau,d) = \frac{\sum_{m=0}^{N-1} x_L(i,\tau,m) x_R(i,\tau,m+d)}{\sqrt{\left[\sum_{m=0}^{N-1} x_L^2(i,\tau,m)\right]\left[\sum_{m=0}^{N-1} x_R^2(i,\tau,m)\right]}} \quad -L \le d \le L, \tag{2}$$

where $x_L(i,\tau,d)$ and $x_R(i,\tau,d)$ represent the binaural signals of TF unit at channel i and frame τ, m is the sample number in a TF unit, N is the frame length, d is the delay between binaural signals. For the 16 kHz sampling rate, we set the value L as 16, which means that the dimension of CCF is 33.

As we know, CCF in each TF unit is usually treated as the main spatial feature for BSSL. More location information can be provided from the corresponding CCFs of more channels within one frame. An example is given for the CCFs from different channels in Fig. 2, where the sound source is located at -10° azimuth with BRIR and TIMIT data. The upper sub-figure of Fig. 2 describes various CCF curves in each channel, while the bottom one is provides a CCF curve of all channels in one frame. Here, we note that the CCF has a similar peak in low frequency channel, which reflects the source azimuth. However, with the frequency increasing, CCF suffers from the peak fluctuation due to the phase wrapping. The CCFs of each channel can be regarded as the features map, which may provide a robust localization.

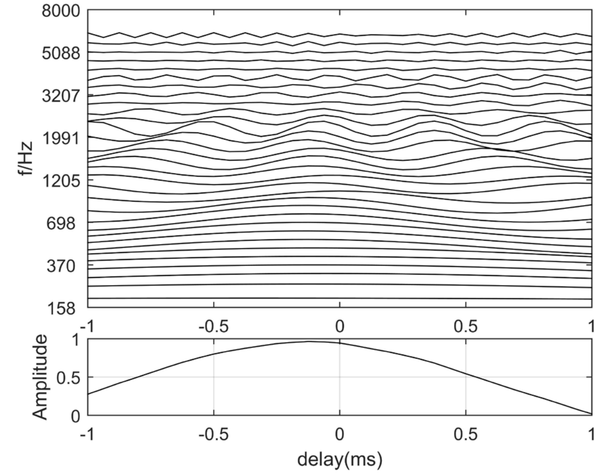


Figure 2: CCF of each channel within a frame

CCF by Eq. (2) is the spatial feature vector of one channel, and then CCF of each channel within a frame are combined into a matrix, which is defined as follows:

$$\boldsymbol{R}(\tau)=\begin{bmatrix} CCF(1,\tau,-L) & CCF(1,\tau,-L+1) & \cdots & CCF(1,\tau,L) \\ CCF(2,\tau,-L) & CCF(2,\tau,-L+1) & \cdots & CCF(2,\tau,L) \\ \vdots & \vdots & \ddots & \vdots \\ CCF(K,\tau,-L) & CCF(K,\tau,-L+1) & \cdots & CCF(K,\tau,L) \end{bmatrix}, \tag{3}$$

where $R(\tau)$ is the feature matrix of frame τ with the channel number K=33.

Since $R(\tau)$ is a matrix of 33×33, it can be regarded as a grey map. Thus we visualize the CCF grey image of Fig. 2 in Fig. 3. The source corresponding to $R(\tau)$ is located at -90° azimuth, where the dark grid represents the corresponding element $R(\tau)$ with large value. The structure of grey image is efficiently related to the source azimuth.

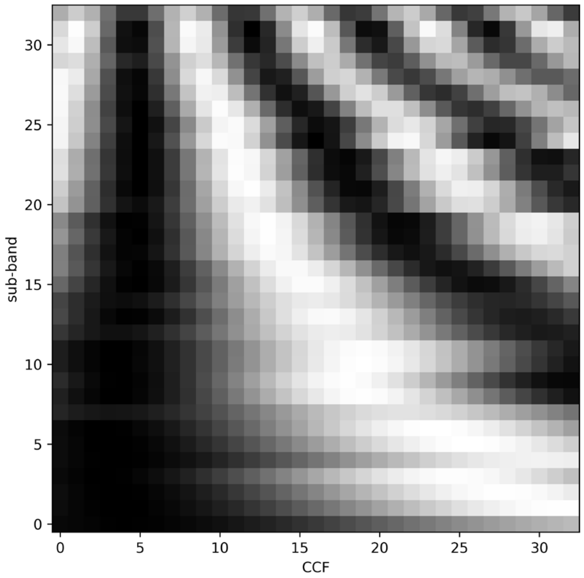


Figure 3: Grey map of $R(\tau)$

3 The Architecture of CNN

We use CNN to train a set of feature matrices $R(\tau)$s in frame level with given azimuths. The azimuth ranges from -90° to 90° with a step of 5°, corresponding to 37 positions.

Since CNN is a multi-layer perceptron neural network seriously depending on parameter setting, more details of our CNN are given follows: one input layer, 4 convolutional-pooling layers and a fully connected layer. For input layer, the input is the 33×33 feature matrix $R(\tau)$, described in Section 2. All convolutional layers use a 2×2 convolution kernel with a step of 1. The output of the previous layer is zero-filled to ensure that the feature size does not shrink. The number of convolution kernels from layer 1 to layer 4 is 18, 36, 72, and 144, respectively. The pooling layer adopts the maximum pooling of 2×2, with the step size of 2. The rectified linear unit (ReLU) activation function is used for the hidden layers.

After four-layer convolution and pooling, the 33×33 $R(\tau)$ becomes a 3×3×144 3-dimensional (3D) feature. The 3D feature is then expanded into 1296×1 one-dimensional and the dropout method is introduced to avoid overfitting. The dropout parameter is 0.5, that is, half of the 1296×1 feature is randomly discarded. After dropout, the fully connected layer converts the feature to probability by softmax function for final estimation of source azimuth. The CNN structure of the proposed algorithm is depicted in Fig. 4.

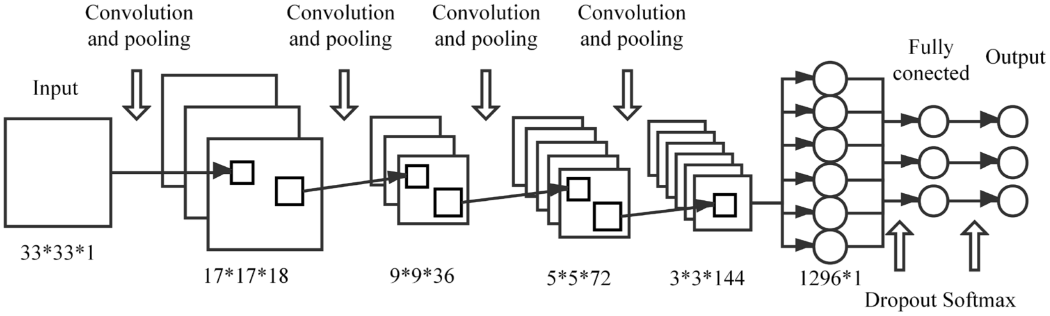


Figure 4: The structure of CNN

4 Simulation and result analysis

4.1 Simulation setting

For both training and testing, the mono source signals taken from the CHAINS Speech Corpus [Cummins, Grimaldi, Leonard et al. (2006)] are convoluted with the BRIR to generate binaural reverberation signals. The CHAINS speech corpus contains 33 sentences spoken by 36 speakers. 9 sentences are selected from the CSLU Speak Identification corpus and 24 sentences are from the TIMIT corpus. In particularly, 20 sentences by 15 speakers are randomly selected and concatenated as training data, and the other sentences from the remaining speakers are regarded as testing data. The sound source is placed with azimuth between -90° and 90° with a step of 5°.

We use two sets of BRIR database to generate the reverberation environment. The first one is obtained by ROOMSIM software [Campbell, Palomaki and Brown (2005)], which uses measured head related impulse responses (HRIR) from the MIT HRIR database in combination with the image method for simulating room acoustics. In the absence of reverberation, the BRIR is degenerated to the HRIR. The reverberation time (RT60) of BRIR is set to 0 s, 0.2 s and 0.6 s. RT60=0 s means the anechoic condition.

The second BRIR dataset [Hummersone (2011)] is recorded using a dummy head and torso in different types of room, named as X, A, B, C, and D at the University of Surrey. Since room X is similar to anechoic environment, only BRIRs recorded in rooms A, B, C, and D is used. The RT60 and the direct-to-reverberant ratio (DRR) of each room are listed in Tab. 1.

Table 1: Room acoustical properties of the Surrey BRIR

Room	A	B	C	D
RT60 (s)	0.32	0.47	0.68	0.89
DRR (dB)	6.09	5.31	8.82	6.12

Those two BRIR databases are utilized in different stage. During the CNN training process, only the first BRIR database is used. In testing, both BRIRs are used to generate the reverberation. The acoustic environment generated by the first BRIR is called simulation environment. The environment generated by the second one is called real environment.

Besides the reverberation, the Gaussian white noise is also added to the binaural signal as the background noise. The noise is uncorrelated with the binaural signals. In addition, binaural noise is uncorrelated with each other. The SNR for training and testing is set to 0, 5 dB, 10 dB, 15 dB and 20 dB.

We measure the reliability of the algorithm by localization accuracy and root-mean-square error (RMSE) of correct localization. The localization accuracy is calculated as:

$$P = n_c / N_{all} \tag{4}$$

where n_c represents the number of correct localization frames; N_{all} is the total number of frames.

Here, the correct localization is defined that the estimation azimuth $\tilde{\theta}_j$ lies within the ±5° of the true azimuth θ.

RMSE only considers the localization error of correct localization frames, which is defined as follows:

$$RMSE = \sqrt{\frac{1}{n_c}\sum_j |\tilde{\theta}_j - \theta|^2} \tag{5}$$

where $\tilde{\theta}_j$ is the estimation azimuth of jth correct localization frame, and θ is the true azimuth.

We compare the performance of the proposed algorithm with several related methods for BSSL. BSSL based on zeros-crossing time difference (ZCTD) [Kim and Kil (2007)] and sub-band SNR estimation (SNRE) [Zhou, Zhao, Cheng et al. (2015)] are selected as the comparison.

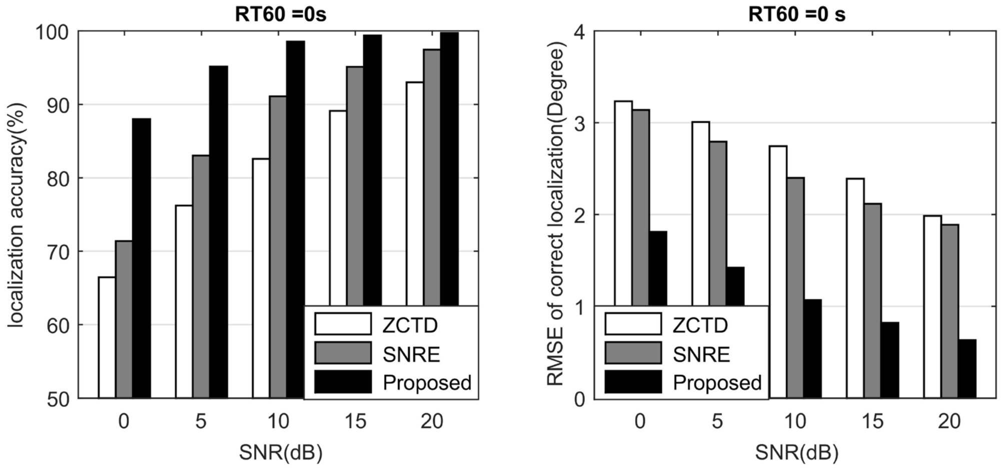
4.2 Evaluation in simulated environment

Figs. 5, 6, 7 depict the performance comparison for ZCTD, SNRE and the proposed algorithm. Fig. 5a, Fig. 6(a) and Fig. 7(a) give the localization accuracy as a function of SNR, while Fig. 5(b), Fig. 6(b) and Fig. 7(b) show the RMSE for the different algorithm. The abscissa represents the SNR, and the ordinate is the performance measurement.

For the localization accuracy, as expected, each of the methods performs well at high SNR. However, the performance of ZCTD and SNRE algorithms become bad as reverberation time increases and SNR declines. The proposed algorithm still performs well even with low SNR and high reverberation. The reason is that CNN regards the spatial feature matrix $R(\tau)$ as a whole grey image, and extracts the efficient spatial features from $R(\tau)$ in diverse environment.

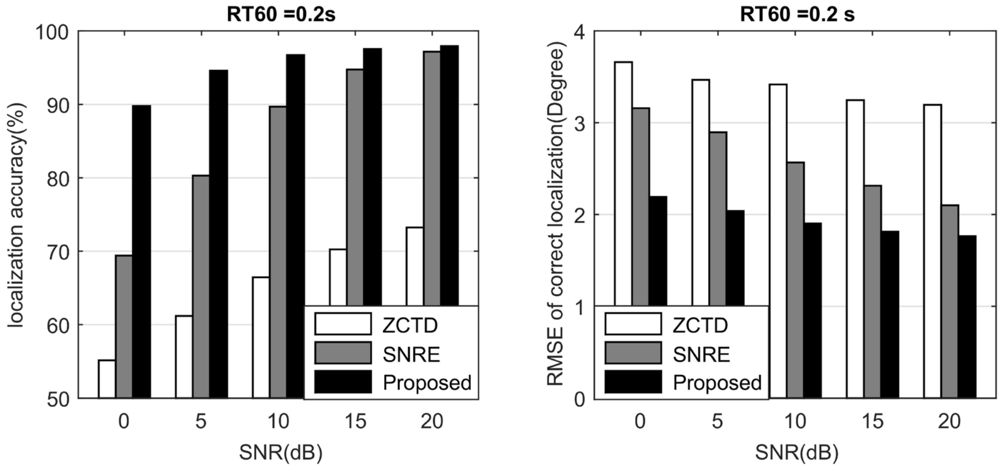
As for the RMSE, the proposed algorithm outperforms ZCTD and SNRE algorithm for each SNR and reverberation condition. RMSE of ZCTD and SNRE is close to 4, which means azimuth estimated by those algorithms is mostly located at the ±5° to the true azimuth. For the proposed algorithm, the value of RMSE is about 2, which indicates the proposed algorithm estimates the azimuth mostly the same as the true azimuth.

In addition, as shown in Figs. 5, Fig. 6 and Fig. 7, the performance of the proposed algorithm changes slowly, reflecting the robustness of the algorithm to the environment.



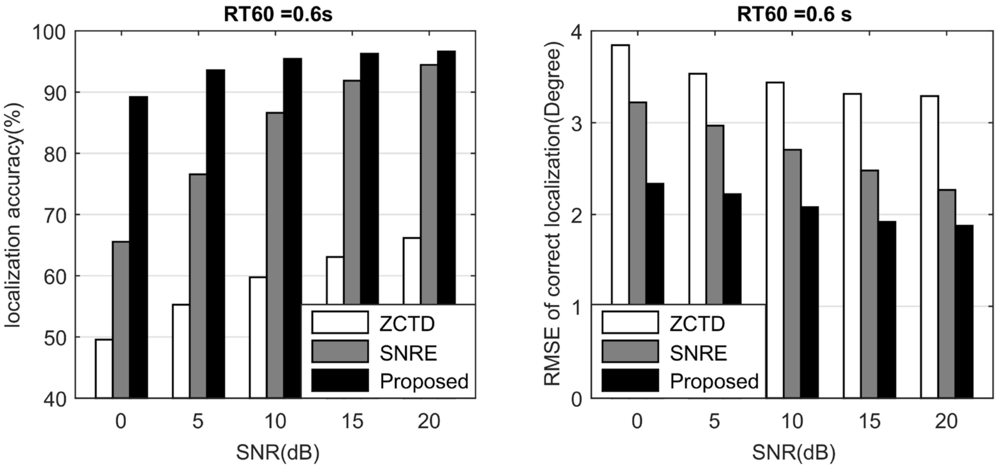
(a) Localization Accuracy (b) RMSE of correct localization

Figure 5: Performance Comparison for ZCTD, SNRE and proposed algorithms with RT60=0 s and variable SNR



(a) Localization Accuracy (b) RMSE of correct localization

Figure 6: Performance comparison for ZCTD, SNRE and proposed algorithms with RT60=0.2 s and variable SNR

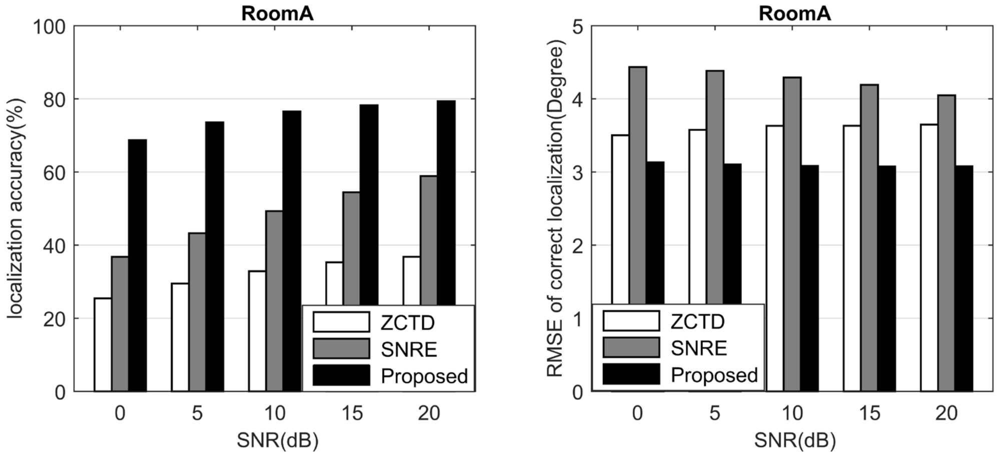


(a) Localization Accuracy (b) RMSE of correct localization

Figure 7: Performance comparison for ZCTD, SNRE and proposed algorithms with RT60=0.6 s and variable SNR

4.3 Evaluation in real environment

Beside the result in the simulated environment, we also evaluate the performance in real environment. The localization accuracy and RMSE of different algorithm in room A, B, C, and D are depicted from Fig. 8 to Fig. 11.



(a) Localization Accuracy (b) RMSE of correct localization

Figure 8: Performance comparison of different algorithm for Room A

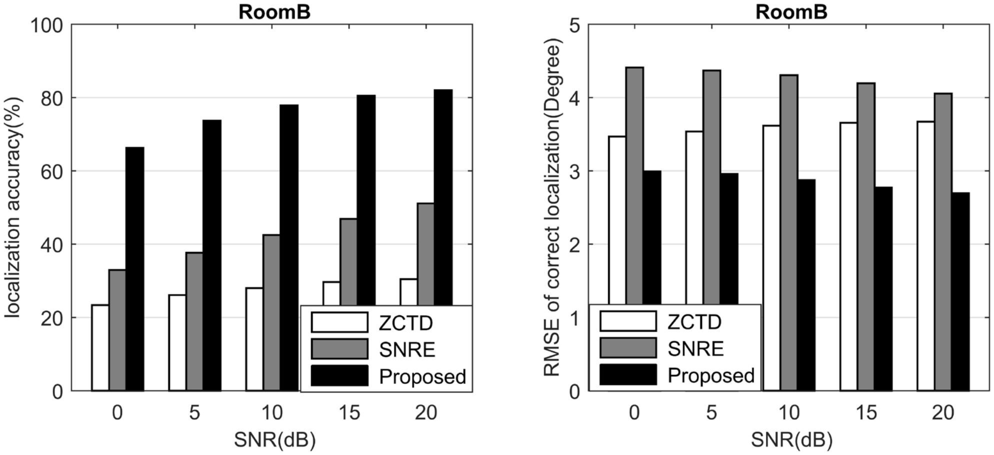


Figure 9: Performance comparison of different algorithm for Room B

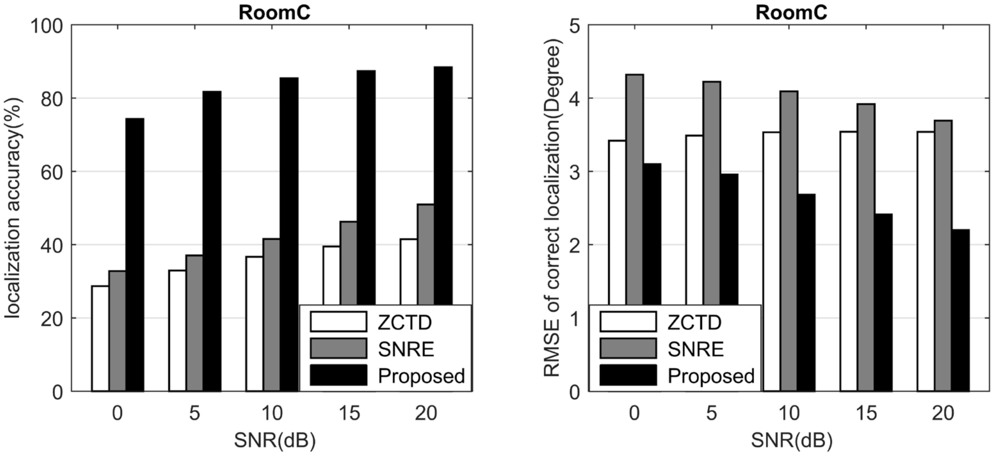


Figure 10: Performance Comparison of different algorithm for Room C

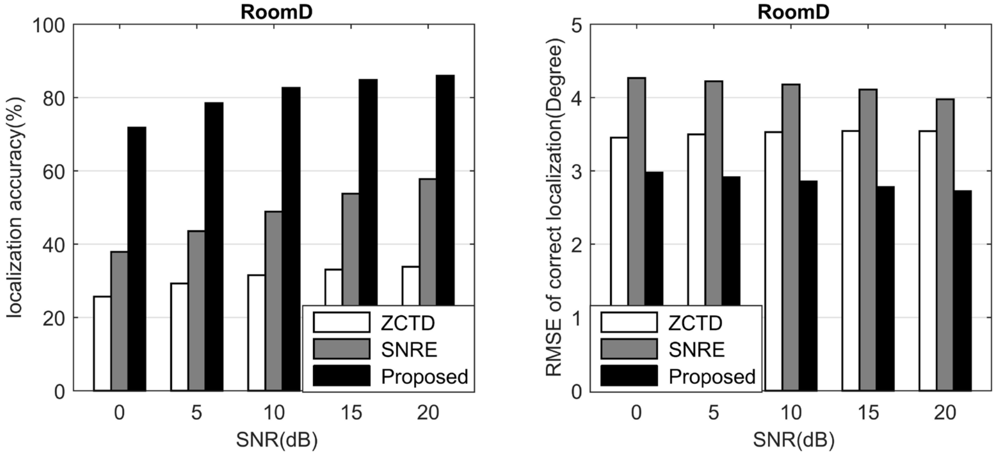


Figure 11: Performance Comparison of different algorithm for Room D

For BRIR database from University of Surrey, the localization performance of the proposed algorithm is significantly better than other algorithms, even for room D with RT60=0.89 s and room C with DRR=8.82 dB.

Since the dispatch of BRIR data in training and testing, the localization results of the proposed algorithm in real environment are not as good as those in simulated environment. But compared with other algorithms, the proposed algorithm still have the better localization performance, which indicates that CNN based algorithm is generalization to untrained conditions.

5 Conclusion

In this work, we have presented a CNN-based binaural sound source localization algorithm. Differing from the previous algorithms, the proposed algorithm extracts the CCF of each Gammatone filter and combines the CCFs of all channels to assemble a two-dimensional feature matrix within one frame. Due to the advantage of CNN on tackling grey image, we treat the two-dimensional feature matrix as a gray image and utilize CNN to establish the relationship between feature matrix and sound azimuth. Experiments show that the CNN-based BSSL algorithm proposed in this paper significantly improves localization performance especially in low SNR and high reverberation conditions.

Acknowledgement: This work is supported by the National Nature Science Foundation of China (NSFC) under Grant No. 61571106, the Fundamental Research Funds for the Central Universities under Grant No. 2242013K30010 and Jiangsu Province key development program under Grant No. BE2017071 and No. BE2017647.

References

Campbell, D.; Palomaki, K.; Brown, G. (2005): A MATLAB simulation of "shoebox" room acoustics for use in research and teaching. *Computing and Information Systems*, vol. 9, no. 3, pp. 48.

Chen, H.; Ser, W. (2009): Acoustic source localization using LS-SVMs without calibration of microphone arrays. *IEEE International Symposium on Circuits and Systems*, pp. 1863-1866.

Cui, Q.; McIntosh, S.; Sun, H. (2018): Identifying materials of photographic images and photorealistic computer generated graphics based on deep CNNs. *Computers, Materials & Continu*a, vol. 55, no. 2, pp. 229-241.

Cummins, F.; Grimaldi, M.; Leonard, T.; Simko, J. (2006): The chains corpus: Characterizing individual speakers. *Proceedings of SPECOM*, vol. 6, pp. 431-435.

Hummersone, C. (2011): *A Psychoacoustic Engineering Approach to Machine Sound Source Separation in Reverberant Environments (Ph.D. Thesis)*. University of Surrey, United Kingdom.

Kim, Y. I.; Kil, R. M. (2007): Estimation of interaural time differences based on zero-crossings in noisy multisource environments. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 734-743.

Ma, N.; Brown, G.; May, T. (2015): Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. *International Speech Communication Association*, vol. 2015, pp. 160-164.

Ma, N.; Gonzalez, J. A.; Brown, G. J. (2018): Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122-2131.

Ma, N.; May, T.; Brown, G. J. (2017): Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 12, pp. 2444-2453.

May, T.; Van de Par, S.; Kohlrausch, A. (2011): A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1-13.

Nix, J.; Hohmann, V. (2006): Sound source localization in real sound fields based on empirical statistics of interaural parameter. *Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 463-479.

Pang, C.; Liu, H.; Zhang, J.; Li, X. (2017): Binaural sound localization based on reverberation weighting and generalized parametric mapping. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1618-1632.

Raspaud, M.; Viste, H.; Evangelista, G. (2010): Binaural source localization by joint estimation of ILD and ITD. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68-77.

Rayleigh, L. (1907): XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214-232.

Roden, R.; Moritz, N.; Gerlach, S.; Weinzierl, S.; Goetze, S. (2015): On sound source localization of speech signals using deep neural networks. *Deutsche Jahrestagung Für Akustik (DAGA)*, pp. 1510-1513.

Roman, N.; Wang, D. L.; Brown, G. J. (2003): Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236-2252.

Xiao, X.; Zhao, S.; Zhong, X.; Jones, D. L.; Chng, E. S. et al. (2015): A learning-based approach to direction of arrival estimation in noisy and reverberant environments. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2814-2818.

Yiwere, M.; Rhee, E. J. (2017): Distance estimation and localization of sound sources in reverberant conditions using deep neural networks. *International Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 12384-12389.

Yu, Y.; Wang, W.; Han, P. (2016): Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. *EURASIP Journal on Audio, Speech, and Music Processing*, no. 7, pp. 1-18.

Zhou, L.; Zhao, X. Y.; Cheng, X.; Wu, Z. Y. (2015): Binaural sound source localization based on sub-band SNR estimation. *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 5, pp. 303-314.