Deep Learning for Binaural Sound Source Localization with Low Signal-to-noise Ratio

Fengnian Zhao, Ruwei Li* and Dongmei Pan
Lab of Acoustical and Optical Information Processing, College of Information and Communication, Beijing University of Technology, Beijing, China
*Email: liruwei@bjut.edu.cn

Abstract. A novel deep learning (DL) method is proposed for binaural sound source localization with low SNR. Firstly, the binaural sound signals are decomposed into several channels by using Gammatone filter. Secondly, the 4 feature parameters of Head-related Transfer Function, interaural time difference (ITD), interaural coherence (IC), interaural level difference (ILD), and interaural phase difference (IPD) are extracted. Thirdly, ITD and IC go through a Deep Belief Network (DBN) to determine the quadrant of the sound source and reduce the positioning range. Then, ITD, IC, ILD, and IPD go through a Deep Neural Network (DNN) to obtain the azimuthal angle within 90 degrees. Experimental results show that the proposed algorithm can solve the front-back confusion, and obtain a superior performance with lower complexity and higher precision under low SNR conditions.

1. Introduction
Binaural Sound Source Localization is to determine the spatial orientation of a sound source using two microphones or two small microphone arrays at both ears. Head-related Transfer Function (HRTF) characterizes how the ear receives sound signals from a spatial point and can be used to synthesize the binaural sound and locate the sound source [1]. Raspaud proposed binaural source localization based on a joint estimation of ITD and ILD [2] which only concerns the frontal azimuth plane. However, in actual applications, the algorithm is unavailable if the sound source comes from the rear half plane. Xiang et al. proposed spatial feature learning for robust binaural sound source localization [3]. This method reduces the amount of input by using the compound eigenvectors. Liu et al. proposed a hierarchical binaural sound source localization method based on an interaural matching filter [4]. Three levels of probability models are used in this method. Li et al. proposed Binaural Sound Source Localization Algorithm based on HRTF and GMM with Gammatone Filter Decomposition [5]. In the illustration of this method, the authors did not discuss the problem about front-back confusion in sound source localization.

Deep Neural Network (DNN) has strong nonlinear mapping and modeling capabilities, which makes it the state-of-the-art technology that can be used in sound source localization [6]. Ma et al. exploited DNN and head movements for binaural localization of multiple speakers in reverberation conditions [7], but the angle of head rotation is limited to the range of [-30° , 30° ], which cannot meet the needs of the actual situation.

In this paper, a method based on Deep Learning (DL) is proposed. In this method, the sound signals are first decomposed into different channels using the Gammatone filter bank and the information sensitive to the human ear is extracted, thereby the amount of data is reduced [8]. Then, the DNN is
used to map the binaural features obtained from the auditory models to the corresponding azimuths of the horizontal plane. Experimental results show that the proposed algorithm can solve the front-back confusion, and obtain a superior performance with lower complexity and higher precision under low SNR conditions.

2. Sound Source Localization Algorithm Based on Two-level Deep Learning

Figure 1 shows a block diagram of the proposed DL-based source localization method. The noisy speech from training data are used to extract features such as ITD, IC, ILD and IPD after T-F decomposition. Then ITD and IC are used to train Deep Belief Network(DBN) for the quadrant of sound source. Then ITD, IC, ILD and IPD are used to train DNN for sound azimuths. Finally, the outputs of the two-level DNN are combined to obtain the azimuth information of the sound source.

2.1 Reduction of Data with Gammatone Filter Decomposition

The Gammatone filter divides the noisy speech into 22 sub-bands. The $i^{th}$ sub-band signal is given by

$$G(n, f_i) = \cos(2\pi f_i T_s + \phi) \cdot \exp(-2\pi B T_s) \cdot B^i \cdot n^{l-1} \cdot U(n) \cdot d(n)$$  \hspace{1cm} (1)

where $n$ is the sample index, $T_s$ is the sampling period, $f_i$ is the center frequency of the $i^{th}$ sub-band, and $\phi$ is the initial phase of the filter. $J=4$ is the order of the filter. $U(n)$ is a unit step function and $d(n)=[d(n), d'(n)]^T$ are noisy speech signals, $B = b(24.7 + 0.108 f_i)$ is the bandwidth, where $b=1.019$ is the attenuation coefficient.

The compressed data $x(n)$ is given by $x(n) = w^T G$, where $w_i$ is the weight parameter of $i^{th}$ sub-band, $w = \text{diag}(w_1, w_2, ..., w_{22})$, $G = [G(n, f_1), G(n, f_2), ..., G(n, f_{22})]$. As the first 8 sub-bands are more important, the weight parameters of $w_1, w_2, ..., w_8$ are set to 1, and the rest are set to 0.

2.2 Extraction of Sound Source Localization Features with HRTF

The ITD, IC, ILD, IPD of Head-related Transfer Function (HRTF) are calculated as follows:

$$ITD_i = \arg \max_d \left\{ \sum_{n=1}^{n_0} x_i(n) x_j(n + d) \sqrt{\sum_{n=1}^{n_0} x_i(n) \sum_{n=1}^{n_0} x_j(n)} \right\}, \quad IC_i = \max_j \left\{ \sum_{n=1}^{n_0} x_i(n) x_j(n + d) \sqrt{\sum_{n=1}^{n_0} x_i(n) \sum_{n=1}^{n_0} x_j(n)} \right\}$$

$$ILD_i = 10 \log_{10} \sum_{n=1}^{n_0} x_i^2(n), \quad IPD_i = E \left( \frac{X_i(f)X_j(f)^*}{|X_i(f)||X_j(f)|} \right)$$

where $i$ is the number of the sub-band, $l$ and $r$ correspond to left and right ear, $n$ is the index of samples, $f_i$ is the total number of sample points, $d$ is the sound source delay, $E()$ means expectation, and $*$ means conjugation.
2.3 Sound Source Localization Based on Deep Learning

Considering that the sound source localization has the problem of front-back confusion, and a deep neural network has poor performance in multiple classifications at a low SNR, we use a two-level DNN to resolve the two problems. The first-level DNN determines the quadrant of the sound source, which narrows the positioning range. The second-level DNN furtherly determines the specific location of 72 orientations divided by 5° of the sound source in the quadrant.

2.3.1 Quadrant Determination of Sound Source Using Deep Belief Network (DBN)

The first DNN used in our work is a Deep Belief Network (DBN). The labels and extracted HRTF features are effective to solve the front-back confusion. First, the feature parameters of ITD and IC extracted from HRTF are used as the feature set of the input of DBN and label those speech features. Then, the feature set is pre-trained through two Restricted Boltzmann Machines (RBM). Back propagation is used to adjust the DBN. At the same time, the distribution of the training set is simulated and the classification model of the test set is formed. Finally, as the DBN output, the horizontal plane is evenly divided into four quadrants, as shown in figure 2. In this way, the range of sound source location is reduced.

The diagram of DBN is shown in figure 3. The DBN consists of two RBMs and a BP network with the structure of “52-158-158-158-4”: the feature set consists of 52-dimensional ITD and IC, the RBMs and the BP are of 158 conditions, and the output data include 4 conditions according to the 4 quadrants. The RBM is a neurological perceptron consisting of a visible layer and a hidden layer. The visible layer is denoted by \( v \) and the hidden layer is denoted by \( h \). The amounts of neurons contained in the visible layer and the hidden layer are \( m \) and \( n \) respectively. The visible and hidden neurons are fully connected in two ways.

The RBM structure is shown in figure 4. \( b \) or \( c \) express the weight of each neuron while \( w \) indicates the strength of each connection. The BP network consists of an input layer, an output layer and a hidden layer, with an output function of softmax. While RBM network is used to classify the four quadrants of the horizontal plane, the BP network is used to adjust the network layer’s parameters and the entire DBN slightly [9] to optimize DBN [10].

2.3.2 Sound Source Localization Based on Deep Neural Network (DNN)

Based on the quadrant reached in DBN, the exact azimuth within 90 degrees of the sound source is further determined in DNN. The DNN has a structure of 104-312-312-312-18, separately the number of conditions of the feature vector, 4 hidden layers and the output data of azimuth. The DNN diagram is shown in figure 5. There are 104 neurons in the input layer. The number of neurons in each hidden layer is set to 312, and the number of hidden layers is 4. There is one neuron in the output layer.

The output of the hidden layer is given by

\[
H_j = g \left( \sum_{i=1}^{n} w_{ij} x_i - a_j \right), \quad j = 1, 2, ..., l
\]  

Figure 2. Division of the horizontal quadrant. Figure 3. Diagram of DBN. Figure 4. RBM Structure Diagram.
where the transfer function of each hidden layer is a Sigmoid function, that is, \( g(x) = \frac{1}{1 + e^{-x}} \). \( n \) is the current level of nodes, \( i \) represents the input layer, \( j \) represents the hidden layer, \( w_{ij} \) is the weight coefficient matrix between the input layer and the hidden layer, \( x \) is the input vector, \( a_j \) is the threshold of the hidden layer, and \( l \) is for the next level of nodes. Similar to (3), the outputs of other hidden layers use the outputs from the previous hidden layers as the inputs to the current hidden layers. \( w_{ij}, a_j \) and the learning efficiency are initialized.

The output \( O_k \) of the output layer is given by

\[
O_k = \sum_{j=1}^{l} H_{jk} w_{jk} - b_k, k = 26 \tag{4}
\]

where \( k \) represents the output layer, \( w_{jk} \) is the weight coefficient matrix between the hidden layer and the output layer, and \( b_k \) is the threshold of the output layer. \( w_{jk} \) and \( b_k \) are initialized.

The prediction error of the DNN is defined as

\[
e_k = Y_k - O_k, k = 26 \tag{5}
\]

where \( Y_k \) is the expected output.

On the one hand, the forward propagation of the DNN takes the four characteristics as inputs and outputs the azimuth of the sound source. On the other hand, through the back propagation, the output error can be allocated to the elements of the hidden layers and the input layer. The weights and thresholds between the layers are updated until the output azimuth of the sound source reaches the optimal result. The weights and the thresholds are updated according to (6):

\[
w_{i} = w_{i} + \eta_{1} H_{i} (1 - H_{i}) x(i) \sum_{k=1}^{26} w_{ik} e_k, \quad i = 1, 2, \ldots, 104
\]

\[w_{jk} = w_{jk} + \eta_{2} H_{j} e_k, \quad j = 1, 2, \ldots, 312,
\]

\[a_j = a_j + \eta_{3} H_{j} (1 - H_{j}) \sum_{k=1}^{26} w_{jk} e_k, \quad j = 1, 2, 3, 4
\]

\[b_k = b_k + e_k, \quad k = 1, 2, \ldots, 26
\]

where \( \eta_1, \eta_2, \eta_3 \) are Learning Rates. Combining the quadrant output from the first DNN and the azimuth (within 90 degrees) output from the second DNN, the azimuth of the sound source in the horizontal plane is obtained.

3. Results and Discussions

3.1 Experimental Data

Sound source localization experiments are conducted to evaluate the performance of the algorithm. We evaluate the performance of the proposed algorithm in the horizontal plane through simulations, using MIT’s HRTF measurement database of the KEMAR dummy-head [11]. In the experiment, the 96 streams of pure speech are from the NTT8k standard voice speech database and the 15 samples of different types of noise are from the SPIB16k noise library. These noises are called buccaneer1, buccaneer2, babble, destroyerengine, destroyerops, f16, factory1, factory2, hfchannel, leopard, m109, machinegun, pink, volvo, and white. For all the speeches, 76 utterances are chosen randomly as training data, and the remaining 20 utterances are used as testing data. Seven SNR conditions used include -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB. The data are recorded with a sampling rate of 16 KHz and 16-bit quantification.
In order to verify the effectiveness of the proposed algorithm, we select three reference algorithms. The 1st reference algorithm is spatial feature learning for robust binaural sound source localization using a composite feature vector [3]. The 2nd reference algorithm is a new hierarchical binaural sound source localization method based on interaural matching filter [4]. The 3rd reference algorithm exploits deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments [7].

### 3.2 Sound Source Localization Performance with Different Types of Noise

![Figure 6. Localization Precision of the Sound Source of Noisy Speech Signals.](image)

The localization precision is indicated by the percentage of correctness (with a deviation within 2.5°) of the estimated spatial orientation of the source. Figure 6 shows the localization precisions of the sound sources of the speech signals with fifteen different types of noise. The blue—, red—, grey—, and yellow— curves respectively represent the proposed algorithm and the 1st, 2nd, and 3rd reference algorithm. It is shown in figure 6 that the localization precision of the proposed method is higher than that of the reference algorithms in most noisy environments. The precision of sound source localization raises along with the increase of SNR. Under low SNR conditions, especially in -10 dB when most of the speech is masked by the noise, the proposed algorithm is of a significant improvement in localization precision compared with the other three reference algorithms. However, when the noise type is destroyerengine or machinegun, the performance of the proposed algorithm is close to one of the reference algorithms. This is because our proposed algorithm is badly affected by the abrupt jump noise of destroyerengine and machinegun, so its positioning performance is not as good as the performance in other types of environments.

RMSE measures the deviation between the estimated and the true source orientations. Figure 7 shows the RMSEs of sound source orientation estimations of speech signals with fifteen different types of noise. The blue—, grey—, yellow— and red— rectangles respectively represent the proposed algorithm and the 1st, 2nd, and 3rd reference algorithm. The RMSE results of the proposed method are the best in most cases with different noise types, especially in a low SNR. Compared with the composite feature model of the 1st reference algorithm, the three-layer probability model of the 2nd reference algorithm and the DNN and head rotation model of the 3rd reference algorithm, the proposed
method can learn the original data features more accurately and distinguish different categories more clearly.

It is revealed in Figure 6 and Figure 7 that the proposed localization method has superior performance in high localization precision and small RMSE of estimation of the orientation. The reason lies in the following 3 points. Firstly, the HRTF and the Gammatone filter keep sufficient information in the speech data for sound source localization. Secondly, DL has a strong ability to learn the intrinsic characteristics. Thirdly, two-level DL provides stability and accuracy for the multi-category classification. As a result, the front-back confusion of sound source localization is avoided, and the positioning accuracy is greatly improved, especially in low SNR environment.

![Figure 7. RMSEs of the Sound Source Orientation Estimation of the Noisy Speech Signals](image)

### 3.3 Sound Source Localization Performance in Different Reverberant Environments

To better evaluate the performance of the proposed method for indoor sound source localization, we test the algorithms in different reverberant environments. For evaluation, the Surrey BRIR database [12] and a BRIR set recorded at TU Berlin [13] were used to reflect different reverberant room conditions. Table 1 shows the precision of sound source localization with different reverberation levels in the speech signals. As the reverberation time increases, the precision of the sound source localization decreases. Nevertheless, at the same level of reverberation in the speech signals, the localization precision with the proposed method was higher than those with the reference algorithms. When the reverberation time was between 0 and 0.3 seconds, the precision of the proposed method exceeded 80%. The reason is that two-level DL has strong learning ability to resist various reverberations and shows better performance of multiple classification.
Table 1. Localization precision with different reverberation times in the speech signals.

| Reverberation Time(s) | Proposed algorithm | Contrast 1 | Contrast 2 | Contrast 3 |
|-----------------------|--------------------|------------|------------|------------|
| 0                     | 98.8               | 98.2       | 97.1       | 97.4       |
| 0.1                   | 96.3               | 95.4       | 92.6       | 93.8       |
| 0.2                   | 90.5               | 82.6       | 70.8       | 87.0       |
| 0.3                   | 83.2               | 68.3       | 64.7       | 78.1       |
| 0.4                   | 76.9               | 61.5       | 52.9       | 69.5       |
| 0.5                   | 69.3               | 55.9       | 40.0       | 63.9       |
| 0.6                   | 60.2               | 48.1       | 34.5       | 55.7       |

4. Conclusion
A binaural sound source localization method based on DNN is proposed. One DNN is used to locate the sound source in one of the four quadrants and the other DNN is used to locate the sound source within 90 degrees. Using the two-level deep learning, the 360-degree horizontal sound source localization is achieved, and the problem of front-back confusion is solved. With low SNR, the proposed algorithm is significantly more accurate than the other algorithms. It is because the two-level deep learning can better learn the characteristics of sound source localization. The experimental results show that the proposed localization method has high precision at a low SNR and strong robustness against noise. Meanwhile, the proposed method can work well in many types of noise environment. This is because DNN has a strong learning ability for each noise feature and two-level DL has pretty strong multiple classification ability.

Acknowledgment
This work was supported by Beijing Municipal Natural Science Foundation Cooperation Beijing Education Committee (No. 201910005007) and the National Natural Science Foundation of China (No. 61971016).

References
[1] Marelli D, Baumgartner R, and Majdak P 2015 Efficient approximation of head-related transfer functions in subbands for accurate sound localization IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(7) 1130–43
[2] Raspaud M, Viste H and Evangelista G 2010 Binaural Source Localization by Joint Estimation of ILD and ITD IEEE Transactions on Audio Speech & Language Processing 18(1) 68-77
[3] Wu X and Talagala D S 2016 Spatial feature learning for robust binaural sound source localization using a composite feature vector IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)2016 6320-4
[4] Zhang J and Liu H 2015 Robust acoustic localization via time-delay compensation and interaural matching filter IEEE Trans. Signal Processing 65(18) 4771-83
[5] Li R W, Pan D M, Zhang S, and Zhang Y Y 2018 Binaural Sound Source Localization Algorithm Based on HRTF and GMM under Gammatone Filter Decomposition Journal of Beijing University of Technology 44(11)1385-90
[6] Bengio Y. 2009 Learning deep architectures for AI Found. Trends Mach. Learn 1(2) 1-127
[7] Ma N and Brown G J 2017 Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberation conditions IEEE Transactions on Audio Speech & Language Processing 25(12) 2444-53
[8] Onizawa N, Koshita S, Sakamoto S, Abe M, Kawamata M, and Hanyu T 2016 Gammatone filter based on stochastic computation *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2016 1036-40

[9] Aierken A, Abudureyimu H, and He Y 2016 A Uyghur spam classification method based on deep belief network *Computer Engineering and Science* 38(10) 2134-9

[10] Chen Y 2014 *Research on Chinese Information Extraction Based on Deep Belief Nets* (Harbin, China: Harbin Institute of Technology) doctoral dissertation

[11] Gardner Band Martin K. 1994 *HRTF Measurements of a KEMAR Dummy-Head Microphone*. Tech. Rep, MIT Media Lab Perceptual Computing

[12] Hummersone C, Mason R, and Brookes T. 2010 Dynamic precedence effect modeling for source separation in reverberant environments *IEEE Trans. Audio Speech Lang. Process* 18(7) 1867-71

[13] Ma N, May T, Wierstorf H, and Brown G J 2015 A machine-hearing system exploiting head movements for binaural sound localisation in reverberant condition *IEEE International Conference on Acoustics, Speech and Signal Processing* 2015 2699-703