International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12384-12389

Distance Estimation and Localization of Sound Sources in Reverberant Conditions using Deep Neural Networks

Mariam Yiwere[1] and Eun Joo Rhee[2]

[1] Department of Computer Engineering, Hanbat National University, Daejeon 34158, Korea.
[2] Professor, Department of Computer Engineering, Hanbat National University, Daejeon 34158, Korea.

[1]Orcid id: 0000-0002-9346-6971
[2]Orcid: 0000-0002-3783-2359, Corresponding Author

Abstract

This paper proposes a method to predict the direction (azimuth) and distance of binaural sound sources simultaneously. With the goal of achieving human-like auditory perception in machines, the method trains a Deep Neural Network to predict both direction and distance by learning from the same set of training features. The training features are the two signal channels' cross correlation series and their interaural level difference values. The proposed method simultaneously predicted the direction and distance of sound sources in the range of 1m to 3m and azimuths of 0°,30° and 60°, with high accuracy values; the values are comparable to previous methods and they are relatively higher in the case of training and testing in separate rooms.

Keywords: Binaural Signals, Distance Estimation, Direction Estimation, Deep Neural Networks.

INTRODUCTION

Background and Objective

Sound source distance estimation and sound source localization have been widely studied by researchers [1-15] in the past few decades. Both direction and distance estimation of sound sources are useful in various fields. For example, in human-robot interaction, where a robot can locate the position of a human speaker, video surveillance, where the surveillance camera rotates and focuses on the position of an event that outside it's field of view, hearing aid systems, smart houses, wearable mobile devices, etc...

Although humans simultaneously perform both sound source localization and distance estimation with little or no difficulty, robots and other machines are far from reaching this level of performing these two tasks in the same or similar manner. The reason is that, researchers mostly study these two problems separately by focusing on either one of the two.

Furthermore, the problem of sound source distance estimation has received a relatively lesser attention in comparison with sound source localization and it is usually tackled with the use of microphone arrays; however, a human-like system should have only two microphones to mimic the biological structure of the human auditory system. Therefore, it is best for proposed research methods to function for binaural systems (less expensive in production) as well, since many systems consist of only two microphones, and they should maintain high prediction accuracies.

Related Work

Researchers have proposed different methods [10-15] to tackle the problem of sound distance estimation in the past few years, as in the case of the sound source localization. In the case of distance estimation, although researchers have done a significant amount of work, binaural distance estimation remains a challenging task, since many of the proposed methods [13-15] use more than two microphones. Other researchers proposed some methods for binaural systems; however, there is room for improvement in terms of their performance accuracies. Some features commonly used in these methods include binaural cues such as Interaural Time Difference (ITD) and Interaural Level Difference (ILD), spectral magnitude cues, Direct to Reverberant Ratio (DRR) and Binaural Signal Magnitude Difference Standard Deviation (BSMD-STD).

To control a mobile robot in terms of azimuth and distance, J. Gontmacher et al. [15] used a spherical microphone array consisting of six microphones in their research. Although microphone arrays such as this one may produce good accuracies, they tend to increase both production and computation costs. S. Vesa [10] used magnitude-squared coherence, a frequency-dependent feature for binaural distance estimation. They trained the model with white noise and then tested it with speech signals. Even though the model was able to classify the speech signals, it was required for them to know the azimuth of the listener in advance, since the training features used were position-dependent.

Using statistical properties of binaural signals, Eleftheria G. et al. [11] proposed a novel feature for learning sound source distances, the Binaural Spectral Magnitude Difference Standard Deviation (BSMD-STD). They used this feature in addition to some other ILD-related features, to estimate sound source distances. Their method performed well in unknown environments; however, it also had lower performance with

fine distance resolution in comparison with the S. Vesa method [10]. In addition, L. Ghamdan et al. [12] used a combination of BSMD-STD features and other binaural cues to estimate the joint direction and distance of binaural sound sources by learning Gaussian Mixture Models for the task. The evaluation of their method produced high performance accuracies in the training room, however, when tested in a different room, the performance significantly deteriorated.

Some researchers have also exploited the power of DNNs for sound source localization, showing the potential of DNNs in this area. For example, Ning Ma et al. [7] applied DNN for the localization of multiple speakers in reverberant conditions and Ryu Takeda et al. [8], also proposed a DNN-based source localization method, which incorporates directional information.

Suggestion

Through a survey of previous research in the areas of sound source localization and sound source distance estimation, we noticed that, the following problems are common to previous methods. First, most of the methods proposed in this field perform only sound source localization or only distance estimation; very few methods exist for the simultaneous performance of the two processes. This implies that, in order to estimate the position (direction and distance) of a sound source like humans do, there is a need for two separate algorithms. In addition, in the case of GMM or SVM learning-based distance estimation methods, the feature extraction steps include too many computations. Lastly and most importantly, the performances of these methods leave room for improvement in the prediction or estimation accuracy.

In order to solve the noted problems, we suggest a joint direction and distance estimation method using a single DNN. Due to its powerful learning capabilities, the DNN can learn to predict both azimuth and distance simultaneously, from the same set of features. This will solve the first problem of using two separate algorithms to predict the position of a sound source. With DNN, training can take place with less complex features, which are easy to extract, thereby, solving the problem of too many computations. Not all, the trained DNN model can also attain much higher prediction accuracies as compared to other machine learning models.

Firstly, we record training data at different azimuths and distances in a room, and we extract features that are relative to both channels of the binaural signals. Next, we supply the extracted feature vectors to a DNN for training, by performing a multiclass classification, predicting both the azimuth and the distances of the recorded training data.

The organization of this paper is as follows: Section 2 introduces the proposed method, section 3 describes our DNN Model Design and Training, section 4 presents experiments and discussion, and section 5 presents our conclusion.

PROPOSED METHOD

In this section, we present our method for learning the direction and distance of a sound source. The method combines the learning of sound source direction and sound source distance into one network using a single set of input features per training sample.

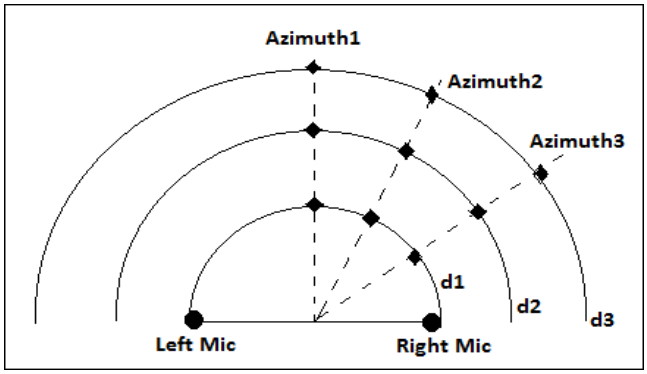

Figure 1: Training Data Recording Positions

Feature Extraction and Preprocessing

In order to learn the prediction of binaural sound source direction or distance with a DNN, we chose a set of suitable features, with a focus on using minimal computations. Our goal was to ensure that the extraction of our chosen features is a simple and less time-consuming process. Moreover, the chosen features had to be relative to the two channels of the binaural signal in order to preserve necessary direction and distance dependent information. For this reason, we chose the cross correlation series of the two channels as our training features. We performed the time-domain cross correlation using equation 1, for only a relevant range of the correlation series. Equation 2 shows the computation of the relevant range, using the sampling frequency (f), velocity of sound (v) and distance between microphones (d). Performing time-domain cross correlation for a short relevant range is computationally less expensive, compared to using the frequency domain cross correlation computation.

$$CrossCorr(l,r)_j(t) = \sum_{k=0}^{N-1} l_{j+k} . r_k \qquad (1)$$

$$\text{Range}[\min \tau, \max \tau] = \left[\frac{-df}{v}, \frac{df}{v}\right] \qquad (2)$$

Instead of selecting the index of the maximum correlation value (argmax), which is the ITD value in this case, we used the entire cross correlation series as input features [7]. The motivation is that, the selection of an ITD value may not always be robust in the presence of noise. In addition, the relationship between the peak value and its side lobes may carry relevant information that the DNN can learn from, for effective classification. Figure 2 shows a graphical representation of a sample of cross correlation series. The relationship between the maximum

value (i.e. the peak) and its side lobes carry rich information that is not obvious to the human eye.

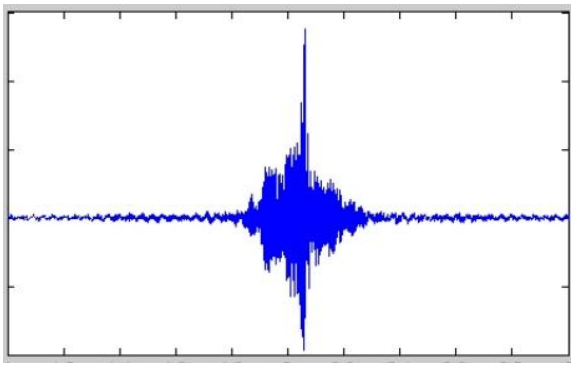

Figure 2: Sample Cross Correlation between left and right channels of a binaural signal

$$ILD_{(r,l)} = 20log_{10}\frac{\sum {l_n}^2}{\sum {r_n}^2} \qquad (3)$$

In addition to the cross correlation series, we also computed the ILD binaural cue using equation 3. The ILD value carries information about the relationship between the intensity levels of the two channels of the binaural signals as received by the microphones. For our experiments, the total number of cross correlation series is 81 values. Adding one ILD value, we get 82 feature values in each feature vector.

Direction Estimation

Firstly, we trained our DNN model to predict the direction of sound sources using the features described in section 2.1. The model was successfully trained to predict one of seven different azimuth values for a given input signal; achieving prediction accuracies above 99%, as shown in Table I. The initial parameters of our network consisted of three hidden layers with ten hidden neurons each. We increased the numbers of the hidden neurons by doubling the previous number in each run, until there was no significant effect on the training and validation accuracy, at eighty neurons per hidden layer. Figure 3 displays the effect of increasing the number of hidden neurons in terms of the validation accuracy.

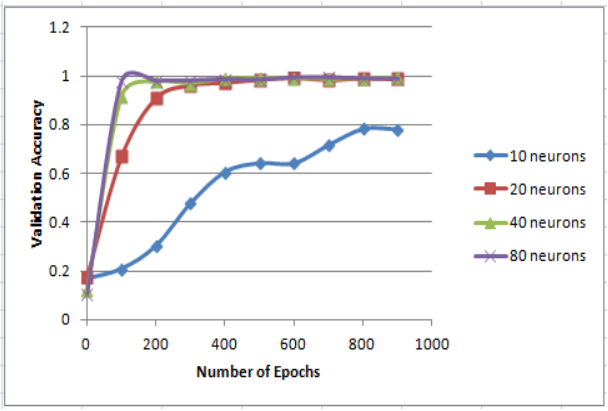

Figure 3: Effects of hidden neurons on validation accuracy

Table I: Performance of the Direction Estimation DNN Model

Training	Testing	Test Accuracy
Room1	Room1	99.7839%
Room1	Room2	99.6549%
Room1&2	Room3	99.4236%

Joint Direction and Distance Estimation

To learn the prediction of the distance between sound sources and the receiving microphones, we implemented a number of new approaches (in terms of input features) without much success. However, we empirically discovered that the same input features used for learning the sound source direction (i.e. the entire cross correlation series of the binaural signals) has information that the DNN uses to learn the distances of the training dataset. We therefore adjusted the parameters of our DNN, specifically the output layer, in order to make simultaneous predictions of both direction and distance for the training dataset, using the same input feature vectors.

The new direction and distance prediction model successfully learned to predict both the direction and distance of sounds similar to those used in the training process. We then performed further experiments with different signals in different rooms to exploit the power of DNNs in the prediction of sound source distances.

DNN MODEL DESIGN AND TRAINING

We designed a Deep Neural Network and trained it to map the 82 dimensional feature vector discussed in the Feature Extraction and Preprocessing section, to their corresponding direction-distance labels The architecture of our DNN is a fully connected neural network with eight hidden layers of hundred neurons each. For each hidden layer, we used a Rectified Linear Unit (Relu) activation function. Since our method is a multiclass classification - classifying datasets into multiple direction and distance classes - our output layer was a softmax classifier. By using the softmax classifier, the DNN model outputs a probability value for each of the possible direction and distance classes.

Figure 4 shows a block diagram of the proposed method. We extract input features from the training dataset and we use them to train the model, which is then used to classify new input signals.

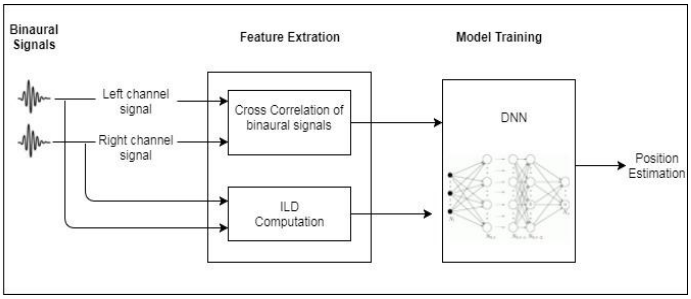

Figure 4: Block diagram of proposed method

EXPERIMENTS AND DISCUSSION

In this section, we discuss the performance of the proposed method based on the evaluation performed in different reverberant environments. We implemented our method on an Intel PC using Visual C++ with Portaudio library, and Python programming. The experimental setup includes two dynamic cardioid microphones connected to a TASCAM US 4x4 audio interface. We position the microphones in the room at a distance of 30cm apart. We collected our training data by recording sounds, including speech signals from the TIMIT database, played from a speaker at different azimuth and distance positions. The sampling rate for recording was 44.1 kHz

We used three different azimuth positions; 0°, 30° and 60° in the training and testing phases along with four different distances from the microphone setup; 1meter, 1.5 meters, 2 meters and 3 meters. In total, we used approximately 12,000 samples for training and validation, and for the evaluation, we used test datasets of approximately 5000 signals recorded in different rooms to test the model. Figure 1 shows the recording positions for preparing the training and test datasets. To evaluate the trained model, we performed experiments in three different rooms. In section 2.2, we showed that the model, which we trained for only direction prediction, was successful in predicting the direction of new sound sources, achieving accuracies above 99%. When it comes to the simultaneous prediction of both direction and distance, we trained the model first in Room 1 and tested it in the same room using test signals that are same as those used to train the model. The model achieved high accuracies above 96%, as shown in Table II. Testing the same model in Rooms 2 and 3, where we did not train the model, the prediction accuracies were much lower.

Again, we preformed the model training in Rooms 2 and 3 and tested them in all the other rooms. The performances of both models reduce to an average of approximately 61.5624%. Finally, we trained the model with combined dataset from Room 2 and 3, and we noticed that, while testing in either of the rooms, the accuracy of prediction increased to above 80%.

Table II: Performance of the Proposed Method in Different Training and Testing Rooms

Training	Testing	Test Accuracy
Room1	Room1	96.0625%
Room1	Room2	67.9459%
Room2	Room1	57.5938%
Room 2&3	Room1	92.8042%
Room2&3	Room2	79.1573%
Room2&3	Room3	89.7075%

We compared the performance of our method to the performance of a previous binaural distance detection method [11], as recorded in their paper. They performed experiments for two different sets of distance classes; course distance classes and fine distance classes, and they recorded maximum accuracies of 62.8% and 61.2% for the fine distance classes and a maximum accuracy of 95.9% for the course distance classes.

In addition, we compared with a previous joint direction and distance estimation method [12]. They recorded accuracies of 60% and below when they evaluated their method in different rooms from the training room. In comparison, our method performs better than the joint direction and distance method [12] in the case of testing the model in a different room, because our method achieved a maximum of 67.9459% accuracy. However, for the previous distance only method [11], they recorded accuracies slightly greater than our method's in some cases and accuracies that are much lower than our proposed method in other cases.

Both of the previous methods used GMM learning algorithms together with multiple computations for the extraction of features, whereas for our proposed method, we computed only a simple cross correlation in addition to the ILD values for the DNN training. Yet the proposed method achieves prediction accuracies that are comparable to those of the previous methods. Therefore, we can conclude that if we use better or richer features in the training of our model, we will see a significant improvement in its performance. Furthermore, extending our training dataset to include different kinds of expected signals taken from rooms with different reverberation values may lead to a better generalization of the model.

CONCLUSION

This paper presents a method for the simultaneous prediction of both direction and distance of binaural sound sources using Deep Neural Networks. The proposed method employs simple and easy-to-extract features such as cross correlation series and Interaural Level Differences for the training of the DNN model. We empirically discovered that the cross correlation series together with the ILD values carry distance-dependent

information based on which we could train the DNN to predict the distance between sound source and receiver. The goal of our study was to achieve a more human-like auditory perception in robots and other machines; hence, we used the same set of features to train our model to predict both direction and distance of binaural sound sources in a simultaneous manner.

We have shown that the proposed method successfully predicts the direction and distances of sound sources with high accuracy (above 95%) when we tested the model in the same room where training took place. When testing was performed in other rooms, the performance of the model slightly reduced, however, it remains comparable with the previous methods [11, 12] we compared with. In the case of training and testing the model in separate rooms, our model outperforms the previous joint direction and distance estimation method [12].

We concluded that the proposed DNN model for simultaneous prediction of direction and distance could generalize to different types of rooms and conditions if we use more training data taken from such rooms in the training of the model. Furthermore, by training the model with better training features, we expect that the proposed method will significantly outperform the previous methods in terms of prediction accuracy. The performance of the model shows that it is possible to train and use it in real world applications to estimate the position of a given sound source.

Our future work includes extending our training dataset to improve the generalization of the model and determining which features will be better at increasing the performance of the system to achieve maximum prediction accuracy. We plan to extend the model to predict the positions of multiple sound sources in a room.

ACKNOWLEDGMENT

This research was supported by the research fund of Hanbat National University in 2017.

REFERENCES

[1] M. Yiwere and E. J. Rhee, “Fast Time Difference of Arrival Estimation using Partial Cross Correlation,” Journal of Information Technology Applications & Management, vol. 22, no. 3, pp.106-114 , September 2015.

[2] T. M. Sreejith, P.K. Joshin, S. Harshavardhan, and T.V. Sreenivas, “TDE Sign Based Homing Algorithm for Sound Source Tracking Using a Y-shaped Microphone Array,” 23rd European Signal Processing Conference(EUSIPCO), pp. 1207-1211, September 2015

[3] J. Gontmacher, P. Havkin, D. Michri and E. Fisher, “DSP-based Audio Processing for Controlling a Mobile Robot using a Spherical Microphone Array,” 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, pp. 1-5, November 2012.

[4] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum Likelihood Sound Source Localization and Beamforming for Directional Microphone Arrays in Distributed Meetings," IEEE Transactions on Multimedia, vol. 10, no. 3, pp. 538 - 548, April 2008

[5] M. Papez and K. Vlcek, "Acoustic Source Localization Based on Beamforming," Recent Advances in Systems Science, pp. 259 – 264, July 2013.

[6] D. Kurc, V. Mach, K. Orlovsky, and H. Khaddour, "Sound Source Localization with DAS Beamforming Method using Small Number of Microphones," International Conference on Telecommunications and Signal Processing, pp. 526 - 532, July 2013.

[7] N. Ma, G. J. Brown and T. May, "Exploiting Deep Neural Networks and Head Movements for Binaural Localization of Multiple Speakers in Reverberant Conditions," Proc. Interspeech, pp. 3302–3306, 2015.

[8] R. Takeda and K. Komatani, "Sound source localization based on Deep Neural Networks with Directional Activate Function Exploiting Phase Information," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 405 - 409, March 2016.

[9] S. Chakrabarty and E. A. P. Habets, "Broadband DOA Estimation using Convolutional Neural Networks Trained with Noise Signals," arXiv: 1705.00919 [cs.SD], May 2017.

[10] S. Vesa, "Binaural Sound Source Distance Learning in Rooms," IEEE Transaction on Audio Speech and Language Processing, vol. 17, no. 8, pp. 1498 - 1507, November 2009.

[11] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos, "Sound Source Distance Estimation in Rooms based on Statistical Properties of Binaural Signals," IEEE Transactions on Audio, Speech and Language Processing, vol. 21, no. 8, pp. 1727 - 1741, August 2013.

[12] L. Ghamdan, M. A. I. Shoman, R. A. Elwahab, and N. A. E. Ghamry, "Position estimation of binaural sound source in reverberant environments," Egyptian Informatics Journal, vol. 18, pp. 87 - 93, 2017.

[13] P. Smaragdis and P. Boufounos, "Position and Trajectory Learning for Microphone Arrays," IEEE Transaction on Audio, Speech and Language Processing, vol. 15, no. 1, pp. 358 - 368, January 2007.

[14] J. K. Nielsen, N. D. Gaubitch, R. Heusdens, J. Martinze, T. L. Jensen, and S. H. Jensen, "Real-time Loudspeaker Distance Estimation with Stereo Audio," Signal Processing Conference (EUSIPCO), 2015 23 European , pp. 250 - 254, September 2015.

[15] J. Gontmacher, A. Yarhi, P. Havkin, D. Michri, and E. Fisher, "DSP-based audio processing for controlling a mobile robot using a spherical microphone array," 2012 IEEE 27th Convention of Electrical and Electronics Engineering in Israel (IEEEI), pp. 1 - 5, November 2012.