engineering proceedings

MDPI

Proceeding Paper

Active Binaural Auditory Perceptual System for a Socially Interactive Humanoid Robot †

Sohaib Siddique Butt, Mahnoor Fatima, Ali Asghar and Wasif Muhammad *

Intelligent Systems Laboratory, Department of Electrical Engineering & Technology, University of Gujrat, Gujrat 50400, Pakistan; sohaibbutt448506@gmail.com (S.S.B.); mahnoorfatima2209@gmail.com (M.F.); ali.asghar@uog.edu.pk (A.A.)

* Correspondence: syed.wasif@uog.edu.pk

† Presented at the 1st International Conference on Energy, Power and Environment, Gujrat, Pakistan, 11–12 November 2021.

Abstract: Sound Source Localization (SSL) and gaze shift to the sound source behavior is an integral part of a socially interactive humanoid robot perception system. In noisy and reverberant environments, it is non-trivial to estimate the location of a sound source and accurately shift gaze in its direction. Previous SSL algorithms are deficient in the optimum approximation of distance to audio sources and to accurately detect, interpret, and differentiate the actual sound from comparable sound sources due to challenging acoustic environments. In this article, a learning-based model is presented to achieve noiseless and reverberation-resistant sound source localization in the real-world scenarios. The proposed system utilizes a multi-layered Gaussian Cross-Correlation with Phase Transform (GCC-PHAT) signal processing technique as a baseline for a Generalized Cross Correlation Convolution Neural Network (GCC-CNN) model. The proposed model is integrated with an efficient rotation algorithm to predict and orient toward the sound source. The performance of the proposed method is compared with the state-of-art deep network-based sound source localization methods. The findings of the proposed method outperform the existing neural network-based approaches by achieving the highest accuracy of 96.21% for an active binaural auditory perceptual system.

Keywords: sound source localization (SSL); gaussian cross-correlation with phase transform (GCC-PHAT); generalized cross-correlation convolution neural network (GCC-CNN)

check for updates

Citation: Butt, S.S.; Fatima, M.; Asghar, A.; Muhammad, W. Active Binaural Auditory Perceptual System for a Socially Interactive Humanoid Robot. *Eng. Proc.* **2021**, *12*, 83. https://doi.org/10.3390/engproc2021012083

Academic Editor: Qasim Awais

Published: 11 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Sound source localization (SSL) is a crucial component of an active human–robot interaction, where the robot needs to precisely detect a speaker and respond appropriately. However, many SSL algorithms have been developed, but they lack in accurately approximating the distance to audio sources, as well as to detect, interpret, and distinguish comparable sound sources due to challenging acoustic environments [1].

The purpose of this research is to develop a generalized biologically plausible model for binaural sound source localization that is both noise and reverberation resistant for a variety of previously unknown scenarios. A critical question is how to actively pay attention to a sound source and how to use a deep learning model to improve sound source localization (SSL). To address these questions, this paper investigated the utilization of deep learning to construct a reliable and generalized model for binaural sound source localization.

To determine the direction of the audio source in the real-world environment, many SSL techniques [2,3] are developed but they include front-back uncertainties and learning of the Head-related Transfer Function (HRTF) for the broader angles than head motions, which necessitates non-linear estimations of the audio-motor map. Although, the need for HRTF was eliminated by an active audition system [4,5], but it is still challenging to cancel the unavoidable motor noises made by the robot itself to interpret sounds in motion, an assessment in loud environments, the front-back challenge, and vertical localization (elevation) without using the HRTF. To address these issues, a new technique that uses arrays of microphones was employed in [6,7] for the identification of a sound source, but it was unable to process sound efficiently such as sound source isolation and automated speaker identification.

With the advancement of sound processing and intelligent systems technology, it is now possible to provide artificial sound source localization capabilities to robots and machines [8–12]. In these methods, sources were assumed to be static and the localization of moving sound sources within the DNN framework was constrained. This paper mainly addresses their robustness and improves it against the reverberation and number of sound sources. Moreover, a suitable configuration of DNNs for SSL is investigated in more depth as it seriously affects performance.

2. Methodology

2.1. Data Synthesis

The dataset was obtained using the Image Source Method (ISM) technique due to its precise control over the SNR and reverb time. At each microphone, an impulse response was created by using the Image Source Method (ISM). The acoustic properties of a room were represented by a room impulse response (RIR). A dataset of spatial audio was built by combining RIRs with an existing collection of audio recordings; then, from any random sound source, a stereo audio file was created by convolving the impulse response of room with an audio recording to introduce the room characteristics in the recording.

2.2. Signal Processing Model

Generalized Cross-Correlation with Phase Transform (GCC PHAT)

For calculation of the Time Delay of Arrival (TDOA) in a binaural sound source localization (SSL), Gaussian cross-correlation with phase transform (GCC-PHAT) is the most classical signal processing model [13]. This signal processing technique was used to calculate the direction of arrival θ, where V_{sound} is the speed of sound ($\approx$343 m/s), f_{sample} is the sample rate in Hz, d is the distance between microphones in meters and τ is the estimated delay between the signals of two ears.

$$\theta = arcsin\left(\frac{V_{sound}\tau}{f_{sample}d}\right) \tag{1}$$

2.3. Generalized Cross-Correlation Convolutional Neural Network (GCC CNN)

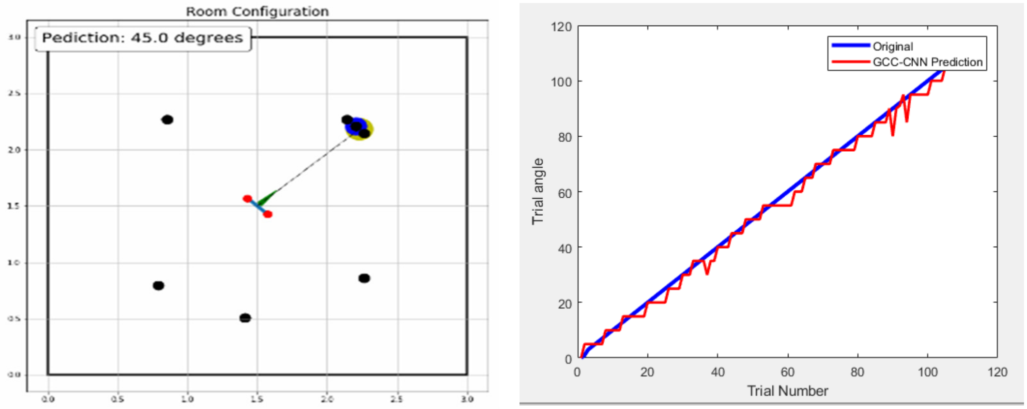
This model uses the interpolated GCC-PHAT vector as the input feature. During the early stages of development, a completely linked layer was inserted before the final softmax layer to enable transfer learning. Batch normalizations were frequently employed to overcome the difficulties of training in very deep models. To lower the network's computational complexity, wide convolutional and max pooling strides were simultaneously employed in the first two layers.

2.4. Rotation Model

A rotation algorithm was proposed in this paper to remove front-end ambiguity. The rotation algorithm was created for labeling the process time by tagging the datasets so that each front-back DOA pair had the same integer label. In this rotation algorithm, six iterations were applied. With the labeling process, each pair of DOAs in a single prediction fell into one of the following two hemifields: left or right in the 1st iteration. As a result, the recording equipment would turn 90 degrees clockwise or counterclockwise on the basis of each pair of DOAs in a single prediction falling into one of two hemifields after which a 2nd prediction was made. If any predicted DOA occurred twice, the model delivered this value as the DOA and the labeling process was discontinued. If no prediction happened twice, the recording equipment would turn 45 degrees to face the quadrant with the highest prediction count. A new prediction was then made, and the prediction list was checked for duplicates. A general rotation loop was entered after two rotations. A count of rotations were recorded for further process, and a rotation-prediction pattern was used. The expected DOA is the mean of the predictions in the quadrant with the highest prediction count if six rotation-prediction combinations occur without a single repeated prediction.

3. Results

The Pyroomacoustics package is used for the implementation of the ISM for this research. An artificial simulation room is created in which the red dots represent the position of microphones, the green arrow is the frontal direction and the yellow dot is the position of a sound source. The setup for the experiment is shown in Figure 1a, where an audio source is placed at 48 degrees. The black dots represent the predictions made by the rotation model, as explained in Section 2. The blue circle represents the final predicted angle, which is 50 degrees close to the actual position, resulting in an accuracy of 96.21%, which is almost 4.21% better accuracy than the previous results of the related research.



(**a**) (**b**)

Figure 1. (**a**) Binaural auditory information-based head orientation using the proposed GCC CNN model. (**b**) Simulation of 120 trials of random sound source locations and predicted location using GCC CNN model.

To generalize the accuracy of the model for the binaural sound source localization system, the trials of one hundred and ten simulation are completed, as shown in Figure 1b. The x-axis represents the number of trials, and the y-axis represents the angle at which trials are conducted, which is shown at the top of the right corner in Figure 1b. The blue color represents the exact angle from which the sound is coming, and the red color represents the predictions of the GCC CNN model. Table 1 illustrates the comparison of this GCC CNN model with different other NN models [14]. The GCC CNN Model outperforms all the other models with 4.21% greater accuracy and a much lower mean error of 1.73 as compared to the other models, resulting in the most accurate model for the Binaural Sound Source Localization System.

Table 1. Comparison of our method with the state-of-art SSL approaches.

Models	Mean Square Error	Accuracy (%)
CNN-GCCFB [1]	4.11	90%
TSN-GCCFB [2]	4.64	91%
MLP-GCC [3]	4.18	92%
GCC-CNN	1.73	96.21%

[1] GCC-PHAT on mel-scale filter bank [14], [2] two-stage neural network with GCCFB [14], [3] multilayer perceptron with GCC-PHAT [14].

4. Conclusions

In this paper, we proposed that a GCC CNN Model is an effective noise and reverberation resistant, deep learning-based sound source localization model. Employing GCC-PHAT input features is a common norm in machine learning SSL systems, which results in machine learning-based SSL using hand-crafted features and is unlikely to require the very deep Convolutional Neural Networks (CNNs), as demonstrated in this paper. The performance of the proposed SSL algorithm is compared with the state-of-art SSL algorithms, and it achieved the highest accuracy of 96.21% and the lowest rms error with a value of 1.73. For developing a more reliable SSL model, the most pressing issue that requires additional investigation is the neural network algorithms used. This research work is extendable to integrate this auditory information with visuals for a better socially interactive humanoid robot. A practical real-world system for paying attention to a single person in a group of more than two people can be considered as additional research.

References

1. Keyrouz, F. Advanced binaural sound localization in 3-D for humanoid robots. *IEEE Trans. Instrum. Meas.* **2014**, *63*, 2098–2107. [CrossRef]
2. Rodemann, T.; Heckmann, M.; Joublin, F.; Goerick, C.; Scholling, B. Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; pp. 860–865.
3. Hornstein, J.; Lopes, M.; Santos-Victor, J.; Lacerda, F. Sound localization for humanoid robots building audio-motor maps based on the HRTF. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; pp. 1170–1176.
4. Nakadai, K.; Okuno, H.G.; Kitano, H. Epipolar geometry-based sound localization and extraction for humanoid audition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Expanding the Societal Role of Robotics in the Next Millennium (Cat. No. 01CH37180), Maui, HI, USA, 29 October–3 November 2001; Volume 3, pp. 1395–1401.
5. Nakadai, K.; Lourens, T.; Okuno, H.G.; Kitano, H. Active audition for humanoid. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00), Austin, TX, USA, 30 July–3 August 2000; pp. 832–839.
6. Sun, H.; Yang, P.; Zu, L.; Xu, Q. An auditory system of robot for sound source localization based on microphone array. In Proceedings of the IEEE International Conference on Robotics and Biomimetics, Tianjin, China, 14–18 December 2010; pp. 629–632.
7. Nakadai, K.; Nakajima, H.; Murase, M.; Okuno, H.; Hasegawa, Y.; Tsujino, H. Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; pp. 852–859.
8. Youssef, K.; Argentieri, S.; Zarader, J.L. A learning-based approach to robust binaural sound localization. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 2927–2932.
9. Ma, N.; Brown, G.J.; May, T. Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. In Proceedings of the INTERSPEECH 2015, Dresden, Germany, 6–10 September 2015; pp. 3302–3306.
10. Xiao, X.; Zhao, S.; Zhong, X.; Jones, D.L.; Chng, E.S.; Li, H. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 2814–2818.
11. Takeda, R.; Komatani, K. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 405–409.
12. Yalta, N.; Nakadai, K.; Ogata, T. Sound Source Localization Using Deep Learning Models. *J. Robot. Mechatron.* **2017**, *29*, 37–48. [CrossRef]
13. Murning, K. *Binaural Sound Localisation Using Machine Learning*; Department of Electrical Engineering at the University of Cape Town: Cape Town, South Africa, 2019.
14. He, W.; Motlicek, P.; Odobez, J.M. Deep neural networks for multiple speaker detection and localization. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 74–79.