Predicting Ensemble Width and Location in Binaural Recordings of Music with Convolutional Neural Networks

Paweł Antoniuk[1,*] and Sławomir Zieliński[1]

[1]Faculty of Computer Science, Białystok University of Technology
[*]Corresponding author: pawel.antoniuk@sd.pb.edu.pl

Abstract

Binaural audio technology has been around for many years, but its popularity has dramatically increased over the past decade due to advancements in virtual reality and streaming technologies. Along with its growing popularity, the quantity of publicly accessible binaural audio materials has also expanded. Consequently, there is now a need for automated and objective measurements of spatial content information, with ensemble width and location being the most important. This study presents a novel method for predicting ensemble width and location in binaural recordings. To this end, 30 head-related transfer functions and 192 binaural music recordings from publicly accessible multi-track recording repositories were used to synthesize 23,040 binaural recordings. The synthesized recordings were then used to train a multi-task convolutional neural network prediction model, the aim of which was to predict the width and location of the ensemble for unseen recordings. The results indicate that models of ensemble breadth and position can be successfully constructed with low prediction errors — 4.76° (±0.10°) for ensemble location and 8.57° (±0.19°) for ensemble width. This approach is the first of its kind to predict both ensemble width and location, offering a more accurate representation of spatial properties. This suggests significant potential for advancing spatial audio applications in virtual reality and streaming technologies, by providing audio engineers with tools that can leverage these methods to enhance spatial audio experiences.

1 Introduction

The human auditory system demonstrates exceptional proficiency in segregating, localizing, and interpreting diverse auditory signals, despite being limited to two ears. This capability arises from the intricate analysis of temporal, amplitude, and spectral disparities, a process termed binaural hearing [1], which enables precise localization of sound sources in complex auditory environments. An important advantage of binaural hearing is demonstrated by the 'cocktail party effect', which showcases the system's capacity to concentrate on individual sounds while suppressing background noise [2]. An understanding of the auditory system is essential for comprehending its limits and creating more immersive binaural experiences for entertainment purposes [3]. It also helps to improve hearing aid devices by enhancing auditory signal reception and improving spatial awareness [4].

The advancement of sophisticated machine learning techniques, particularly deep learning networks, has prompted an intriguing exploration of the extent to which these tools can emulate the human auditory system without relying on advanced spatial audio feature engineering, traditionally employed in audio source localization tasks [5–7]. To investigate this, the study developed an audio localization method based on a multi-task convolutional neural network (CNN) model.

Inspired by the fact, that humans tend to localize sound sources in groups rather than individually [8, 9], the objective of the proposed model is to predict ensemble location and width instead of positions of individual sources. This study is unique as it not only conceptualized the method but also tested it on a vast, real-life music corpus of 23,040 binaural excerpts that were synthesized using 192 multi-track music recordings and 30 publicly available head-related transfer functions (HRTFs) from various sources. The music recordings encompassed various genres, including rock, jazz, pop, and classical music.

The developed method has the potential to be highly beneficial in automated assessment tasks, where a significant number of binaural recordings must be evaluated and labeled in terms of their spatial content information. This may assist audio engineers in objectively assessing and segregating binaural audio recordings with regard to their spatial content. Furthermore, it could

facilitate the development of an autonomous web-crawler bot that will collect binaural recordings from publicly accessible repositories and label them according to the spatial properties of the sound sources, such as the location of the music ensemble or the sparsity of audio source positions.

The findings demonstrate that this method is effective in accurately predicting the spatial characteristics of sound sources in near-real-world scenarios. Furthermore, this paper presents an experiment framework that allows for the objective measurement of a binaural localization technique in an objective manner, utilizing a large-scale dataset synthesized from real-world examples of music signals (for other use cases for this framework, see [10–13]). One of the key advantages of the proposed method is that it does not assume the number of audio sources. However, a significant limitation of this study is the lack of reverberation in the synthesized recordings, which represents a material for future research.

The remainder of this paper is organized as follows. In Section 2, related studies are presented. The description of the method developed as part of this study is provided in Section 3. This section also includes a detailed definition of ensemble width and the description of the experiments that was used to evaluate this method. Section 4 discusses the results of the method-evaluation experiments conducted in this study. Finally, the paper is summarized in Section 5.

2 Related studies

The majority of existing literature on the subject of sound source localization employs techniques that leverage the advantages of microphone arrays with more than two channels [14–19]. In the context of sound source localization in binaural signals, the predominant focus of research is on the identification of individual sound sources, rather than groups of sounds [20–27]. In terms of source direction of arrival (DOA) methods, the majority of research assumes a fixed number of sound sources [7, 23, 26, 28, 29], which limits its practical applications as this information is rarely known in real-life binaural recordings. Moreover, the majority of studies have focused on relatively homogeneous signals, namely speech [5, 20–27, 30–32].

In contrast to the aforementioned studies, the proposed method is not limited by the number of sources. Furthermore, the approach is not limited to speech and has been applied to much more diverse musical datasets, including sound sources such as instruments and vocals. In contrast to previous studies that have focused on individual sources, the proposed method does not aim to separate them, but rather considers them as an ensemble, or in this case, a musical ensemble. This approach is inspired by the way real musical ensembles are arranged on stage. To the authors' knowledge, this is one of the first methods to localize ensemble width (see [10] for the previous ensemble-width-related study), and the first to localize both ensemble position and width simultaneously using a multi-task model.

Sound localization methods can be classified into two primary categories based on the implementation of their underlying algorithms: glass-box and black-box. The glass-box methods, which are more traditional in the literature, rely on manual designing algorithms that mimic the auditory system to explicitly extract key features from the signal for location prediction, such as one of combination of the following: interaural level difference, interaural time difference, interaural coherence or interaural phase difference (see [1] for features description). In contrast, black-box methods typically employ only basic feature engineering and rely heavily on modern machine learning techniques to both extract features and make the final prediction. Consequently, they are less reliant on manual intervention (see [28] for an example). The black-box algorithm extracts features implicitly and internally, without knowledge of their internal purpose or the equivalent of the human auditory system. This renders such methods less transparent and less predictable. This, in conjunction with the fact that they are typically implemented with deep neural networks with a large number of learning parameters, necessitates the development of them with the use of very large datasets, typically comprising hundreds of thousands of examples.

3 Methodology

Experiments in this study were conducted on 23040 binaural recordings of music. The binaural recordings were synthesized semi-automatically using 192 multi-track publicly-available music recordings and 30 HRTF databases. For each multi-track recording and HRTF database pair, four binaural recordings were synthesized for different random ensemble parameters — its location θ and width θ — as defined in Section 3.1. Both parameters were drawn from uniform random distribution.

3.1 Ensemble location and width definition

The primary objective of the model developed in this study is to predict the ensemble location (θ) and width (ω), as illustrated in Figure 1. The ensemble is defined as a group of audio point sources. The ensemble width ω is defined as the angular width between two extreme point sources, while the ensemble location θ represents the middle angle between two extreme sound sources.

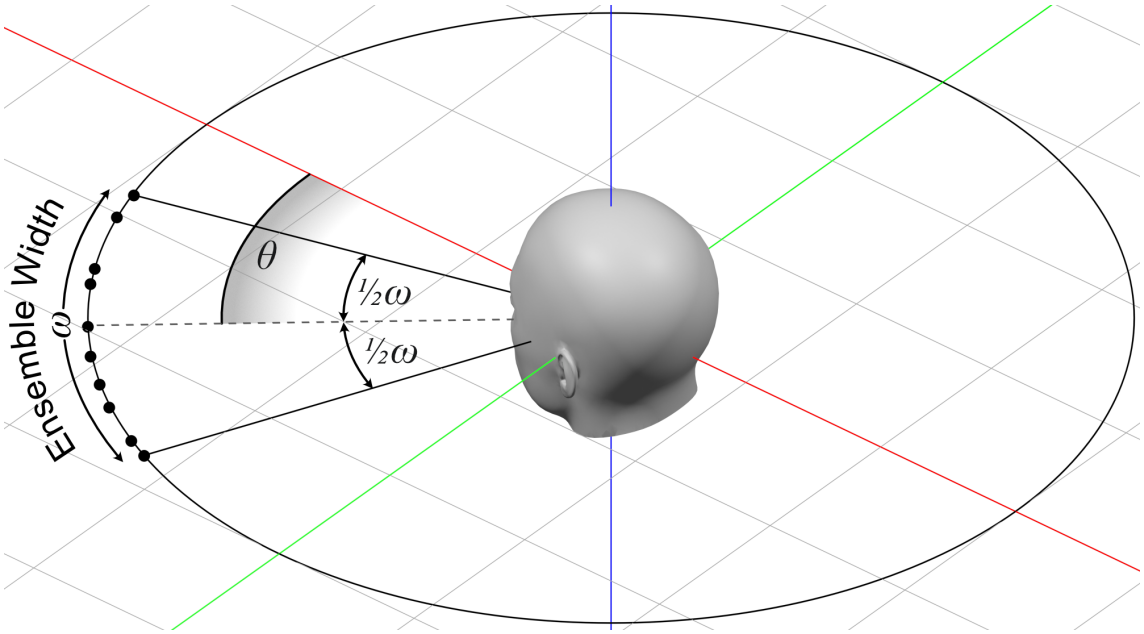

Figure 1: Illustration of ensemble width (ω) and ensemble location (θ) relative to the direction of the head. Black dots represent the positions of audio sources. The ensemble location (θ) is defined as the angular position of the center of the ensemble relative to the direction the head is turned. The ensemble width (ω) is defined as angular distance between two extreme audio sources.

3.2 Synthesis of binaural music recordings

3.3 Spectrogram conversions of binaural music recordings

Prior to being fed into the model, the binaural recordings of music were transformed into magnitude spectrograms. Figure 2 illustrates example spectrograms for four distinct binaural recordings of music. It should be noted that the original spectrograms were stored in a floating-point-precision matrix format.

3.4 Convolutional neural network topology

3.5 Model training and evaluation

4 Results

Figure 5 illustrates the comparison between the actual and the predicted ensemble width. The results demonstrate that the model exhibits better prediction quality for narrower ensemble widths (12.44° for $\omega > 80°$) and that its performance deteriorates with the increase of the ensemble width ($Z°$ for $W°$). Figure 6 further demonstrates that the relationship between the prediction error and the actual width is not linear, exhibiting a depression between 60 and 75 degrees. This suggests, that the ensemble width has a significant impact on the ensemble width estimation.

In contrast to the correlation between ensemble width and its prediction error, there is no significant relationship between the actual location and its prediction error, as illustrated in Figures 3 and 4. This finding indicates that the model's capabilities for localizing the center of the ensemble is robust, unaffected by the actual spatial positioning of the ensemble, including lateral locations.

Figure 7 illustrates the influence of both ensemble width and location on mean absolute error for ensemble location. This figure provides a more detailed view of the data presented in Figure

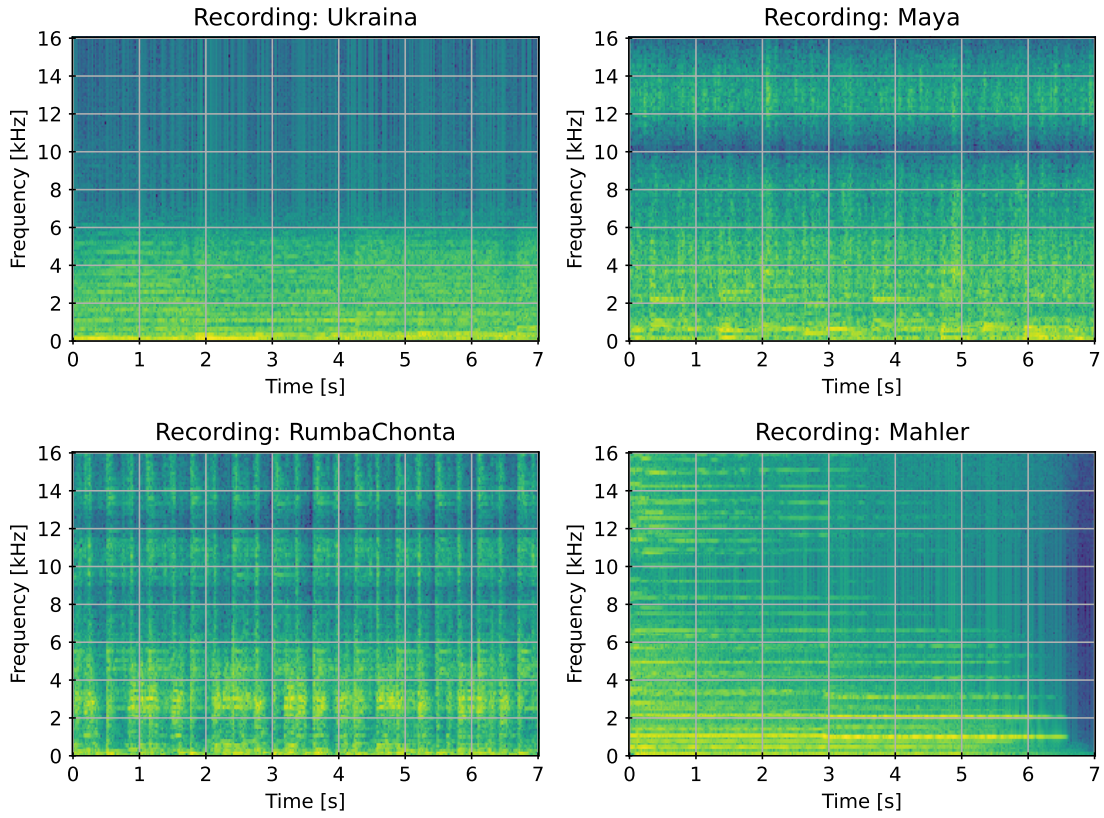

Figure 2: An illustrative example of the magnitude spectrograms generated from synthesized binaural recordings of music. Each spectrogram was derived from a different randomly selected binaural music recording. These spectrograms were subsequently employed as input data for training and assessing the convolutional neural network (CNN) in this study.

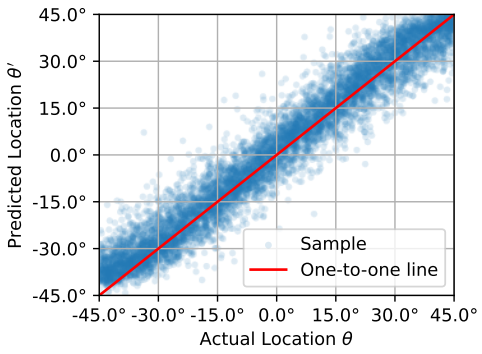

Figure 3: A comparison between the actual ensemble location θ and the predicted ensemble location θ' for a single iteration (of the total five)

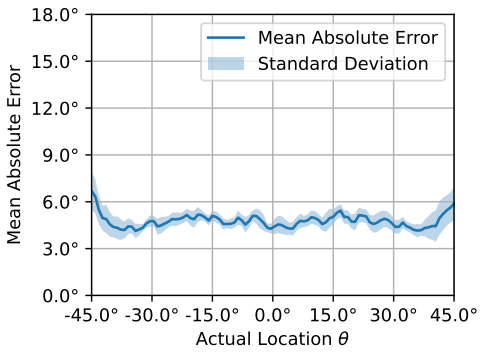

Figure 4: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

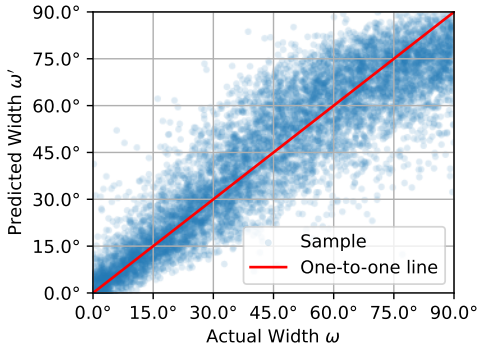

Figure 5: A comparison between the actual ensemble width ω and the predicted ensemble width ω' for a single iteration (of the total five)

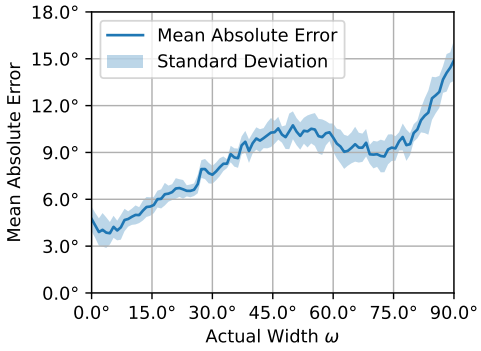

Figure 6: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

4. The figure indicates a slight asymmetry. Similarly, Figure 8 illustrates the influence of both ensemble width and location on mean absolute error for ensemble width.

5 Conclusions

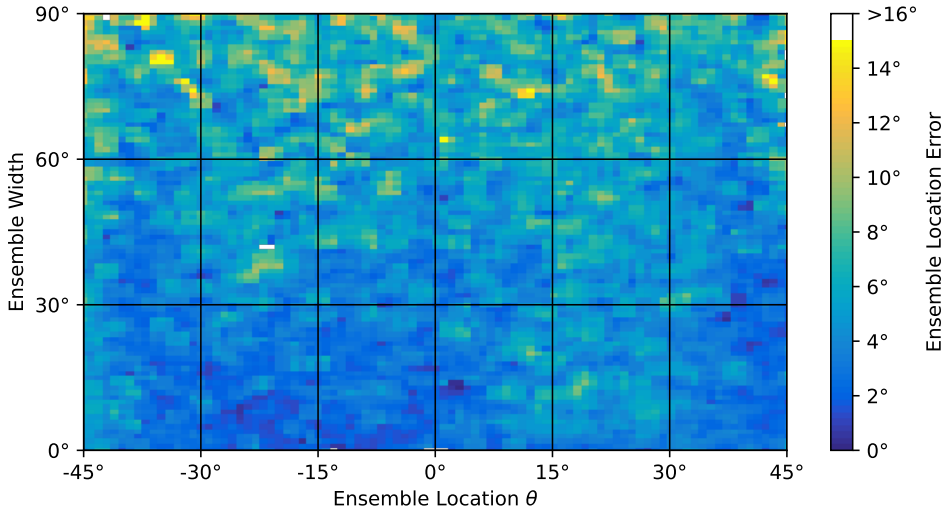

Figure 7: The heatmap that illustrates the mean absolute error (MAE) of ensemble location distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors.

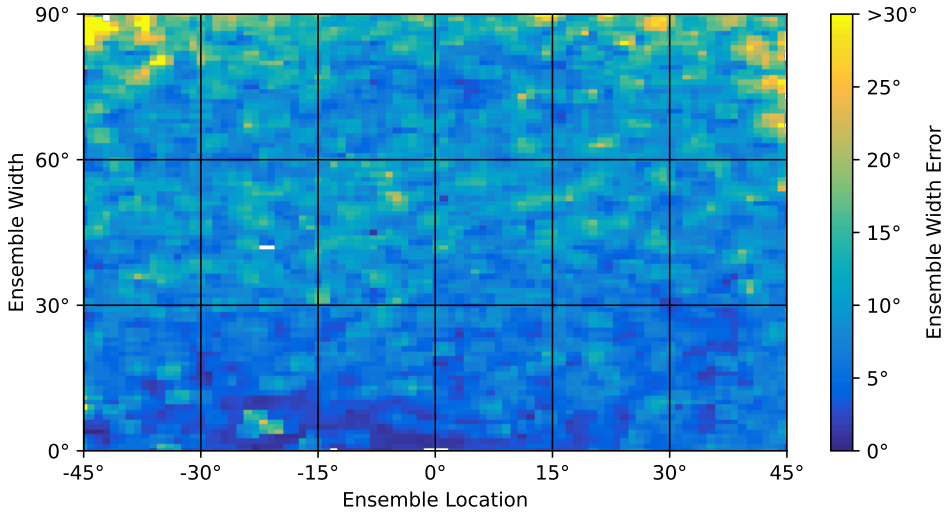

Figure 8: The heatmap that illustrates the mean absolute error (MAE) of ensemble width distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors. Notably, the values between 30° and 60° on the y-axis exhibit unexpectedly higher MAE values region — please see Figure 6 for comparison.

References

1. Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization* ISBN: 978-0-262-26868-4. https://doi.org/10.7551/mitpress/6391.001.0001 (The MIT Press, Oct. 1996).
2. Cherry, E. C. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America* **25.** Place: US Publisher: Acoustical Society of American, 975–979. ISSN: 0001-4966(Print) (1953).
3. Zhang, W., Samarasinghe, P. N., Chen, H. & Abhayapala, T. D. Surround by Sound: A Review of Spatial Audio Recording and Reproduction. en. *Applied Sciences* **7.** Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, 532. ISSN: 2076-3417. https://www.mdpi.com/2076-3417/7/5/532 (2024) (May 2017).
4. Thiemann, J., Müller, M., Marquardt, D., Doclo, S. & van de Par, S. Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene. *EURASIP Journal on Advances in Signal Processing* **2016,** 12. ISSN: 1687-6180. https://doi.org/10.1186/s13634-016-0314-6 (2024) (Feb. 2016).
5. Yang, Q. & Zheng, Y. *DeepEar: Sound Localization with Binaural Microphones* in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications* ISSN: 2641-9874 (May 2022), 960–969. https://ieeexplore.ieee.org/document/9796850 (2024).
6. Vera-Diaz, J. M., Pizarro, D. & Macias-Guarasa, J. Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates. en. *Sensors* **18.** Number: 10 Publisher: Multidisciplinary Digital Publishing Institute, 3418. ISSN: 1424-8220. https://www.mdpi.com/1424-8220/18/10/3418 (2024) (Oct. 2018).
7. Pang, C., Liu, H. & Li, X. Multitask Learning of Time-Frequency CNN for Sound Source Localization. *IEEE Access* **7.** Conference Name: IEEE Access, 40725–40737. ISSN: 2169-3536. https://ieeexplore.ieee.org/document/8668414 (2024) (2019).
8. Bregman, A. in *Journal of The Acoustical Society of America - J ACOUST SOC AMER* Journal Abbreviation: Journal of The Acoustical Society of America - J ACOUST SOC AMER (MIT Press, Jan. 1990).
9. Rumsey, F. Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *Journal of the Audio Engineering Society* **50,** 651–666 (Sept. 1, 2002).
10. Antoniuk pawel & zielinski slawomir krzysztof. blind estimation of ensemble width in binaural music recordings using 'spatiograms' under simulated anechoic conditions. *journal of the audio engineering society* (Apr. 2023).
11. Zieliński, S. K., Antoniuk, P., Lee, H. & Johnson, D. Automatic discrimination between front and back ensemble locations in HRTF-convolved binaural recordings of music. *EURASIP Journal on Audio, Speech, and Music Processing* **2022,** 3. ISSN: 1687-4722. https://doi.org/10.1186/s13636-021-00235-2 (2024) (Jan. 15, 2022).
12. Zieliński, S. K., Antoniuk, P. & Lee, H. Spatial Audio Scene Characterization (SASC): Automatic Localization of Front-, Back-, Up-, and Down-Positioned Music Ensembles in Binaural Recordings. *Applied Sciences* **12.** Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, 1569. ISSN: 2076-3417. https://www.mdpi.com/2076-3417/12/3/1569 (2024) (Jan. 2022).
13. Zieliński, S. K., Lee, H., Antoniuk, P. & Dadan, O. A Comparison of Human against Machine-Classification of Spatial Audio Scenes in Binaural Recordings of Music. *Applied Sciences* **10.** Number: 17 Publisher: Multidisciplinary Digital Publishing Institute, 5956. ISSN: 2076-3417. https://www.mdpi.com/2076-3417/10/17/5956 (2024) (Jan. 2020).
14. Kaveh, M. & Barabell, A. The statistical performance of the MUSIC and the minimum-norm algorithms in resolving plane waves in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **34.** Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing, 331–341. ISSN: 0096-3518. https://ieeexplore.ieee.org/document/1164815 (2024) (Apr. 1986).

15. Pavlidi, D., Puigt, M., Griffin, A. & Mouchtaris, A. *Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures* in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ISSN: 2379-190X (Mar. 2012), 2625–2628. https://ieeexplore.ieee.org/document/6288455 (2024).

16. Pan, Z., Zhang, M., Wu, J., Wang, J. & Li, H. Multi-Tone Phase Coding of Interaural Time Difference for Sound Source Localization With Spiking Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **29.** Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2656–2670. ISSN: 2329-9304. https://ieeexplore.ieee.org/document/9502013 (2024) (2021).

17. Hahmann, M., Fernandez-Grande, E., Gunawan, H. & Gerstoft, P. Sound source localization using multiple ad hoc distributed microphone arrays. *JASA express letters* **2,** 074801. ISSN: 2691-1191 (July 2022).

18. Chung, M.-A., Chou, H.-C. & Lin, C.-W. Sound Localization Based on Acoustic Source Using Multiple Microphone Array in an Indoor Environment. *Electronics* **11.** Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, 890. ISSN: 2079-9292. https://www.mdpi.com/2079-9292/11/6/890 (2024) (Jan. 2022).

19. Liu, M., Hu, J., Zeng, Q., Jian, Z. & Nie, L. Sound Source Localization Based on Multi-Channel Cross-Correlation Weighted Beamforming. *Micromachines* **13.** Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, 1010. ISSN: 2072-666X. https://www.mdpi.com/2072-666X/13/7/1010 (2024) (July 2022).

20. Dietz, M., Ewert, S. D. & Hohmann, V. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication. Perceptual and Statistical Audition* **53,** 592–605. ISSN: 0167-6393. https://www.sciencedirect.com/science/article/pii/S016763931000097X (2024) (May 2011).

21. May, T., van de Par, S. & Kohlrausch, A. A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End. *IEEE Transactions on Audio, Speech, and Language Processing* **19.** Conference Name: IEEE Transactions on Audio, Speech, and Language Processing, 1–13. ISSN: 1558-7924. https://ieeexplore.ieee.org/document/5406118 (2024) (Jan. 2011).

22. May, T., van de Par, S. & Kohlrausch, A. A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation. *IEEE Transactions on Audio, Speech, and Language Processing* **20.** Conference Name: IEEE Transactions on Audio, Speech, and Language Processing, 2016–2030. ISSN: 1558-7924. https://ieeexplore.ieee.org/document/6178270 (2024) (Sept. 2012).

23. Woodruff, J. & Wang, D. Binaural Localization of Multiple Sources in Reverberant and Noisy Environments. *IEEE Transactions on Audio, Speech, and Language Processing* **20.** Conference Name: IEEE Transactions on Audio, Speech, and Language Processing, 1503–1512. ISSN: 1558-7924. https://ieeexplore.ieee.org/document/6129395 (2024) (July 2012).

24. May, T., Ma, N. & Brown, G. J. *Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues* in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* ISSN: 2379-190X (Apr. 2015), 2679–2683. https://ieeexplore.ieee.org/document/7178457 (2024).

25. Ma, N. & Brown, G. J. *Speech Localisation in a Multitalker Mixture by Humans and Machines* in *Proc. Interspeech 2016* (2016), 3359–3363.

26. Ma, N., May, T. & Brown, G. J. Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localisation of Multiple Sources in Reverberant Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **25.** arXiv:1904.03001 [cs, eess], 2444–2453. ISSN: 2329-9290, 2329-9304. http://arxiv.org/abs/1904.03001 (2024) (Dec. 2017).

27. Benaroya, E. L. *et al.* Binaural Localization of Multiple Sound Sources by Non-Negative Tensor Factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **26.** Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1072–1082. ISSN: 2329-9304. https://ieeexplore.ieee.org/document/8294267 (2024) (June 2018).

28. Vecchiotti, P., Ma, N., Squartini, S. & Brown, G. J. End-to-end Binaural Sound Localisation from the Raw Waveform. *CoRR* **abs/1904.01916.** arXiv: 1904.01916. http://arxiv.org/abs/1904.01916 (2019).

29. S, A. & T V, S. *Spatiogram: A phase based directional angular measure and perceptual weighting for ensemble source width* arXiv:2112.07216 [cs, eess]. Dec. 2021. http://arxiv.org/abs/2112.07216 (2024).

30. Wang, J., Wang, J., Qian, K., Xie, X. & Kuang, J. Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. *EURASIP Journal on Audio, Speech, and Music Processing* **2020,** 4. ISSN: 1687-4722. https://doi.org/10.1186/s13636-020-0171-y (2024) (Feb. 10, 2020).

31. Liu, Q., Wang, W., de Campos, T., Jackson, P. J. B. & Hilton, A. Multiple Speaker Tracking in Spatial Audio via PHD Filtering and Depth-Audio Fusion. *IEEE Transactions on Multimedia* **20.** Conference Name: IEEE Transactions on Multimedia, 1767–1780. ISSN: 1941-0077. https://ieeexplore.ieee.org/document/8119824 (2024) (July 2018).

32. Ma, N., Gonzalez, J. A. & Brown, G. J. 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* **26.** Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2122–2131. ISSN: 2329-9304. https://ieeexplore.ieee.org/document/8410799 (2024) (Nov. 2018).