Predicting Ensemble Location and Width 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 in existence for many years, but its popularity has significantly increased over the past decade as a consequence of 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 location and width being the most prominent. This study presents a novel method for predicting ensemble location and width 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 model, the aim of which was to predict the location and width of the ensemble for unseen recordings. The results indicate that models of ensemble location and width can be successfully constructed with low prediction errors — 4.76° (±0.10°) for ensemble location and 8.57° (±0.19°) for ensemble width. The method developed in this study outperforms previous spatiogram-based methods recently published in the literature and holds promise for future development as a part of a novel tool for audio engineers to assess binaural audio.

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, 5].

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 [6–8]. To investigate this, the study developed an audio localization method based on a multi-task convolutional neural network (CNN) model [9, 10].

Inspired by the fact, that humans tend to localize sound sources in groups rather than individually [11, 12], 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 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 [13–16]). 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 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 remainder of this paper is organized as follows: Section 2 presents related studies. The description of the method developed for this study is provided in Section 3, which also includes detailed definitions of ensemble location and width, along with a description of the experiments used to evaluate this method. Section 4 discusses the results of the method-evaluation experiments conducted in this study. Finally, the paper concludes 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 [17–22]. 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 [23–30]. In terms of source direction of arrival (DOA) methods, the majority of research assumes a fixed number of sound sources [8, 26, 29, 31, 32], 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 [6, 23–30, 33–35].

In contrast to the aforementioned studies, the proposed method is not constrained by the number of sources. Furthermore, the approach is not limited to speech and has been applied to a wide range of musical datasets, including instruments and vocals. In contrast to previous studies that primarily 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 similar to how the real musical ensembles are arranged on stage. To the authors' knowledge, this is one of the first methods to localize ensemble width (see [13] 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 categories based on the implementation of their underlying algorithms: glass-box (e.g., [23–29, 35]) and black-box (e.g., [6, 7, 31]). Glass-box methods, more traditional in the literature, rely on manually designed algorithms that mimic the auditory system to explicitly extract key features for location prediction, such as interaural level differences, interaural time differences, interaural coherence, or interaural phase differences (see [1] for feature descriptions). An example of one of the most advanced auditory models that is able to extract such features was developed by the Two!Ears project [36].

In contrast, black-box methods typically employ minimal feature engineering and heavily rely on modern machine learning techniques to extract features and make predictions. These methods implicitly and internally extract features, without disclosing their purposes or mimicking the human auditory system. This lack of transparency and predictability, combined with their reliance on deep neural networks with numerous learning parameters, necessitates the development and evaluation using very large datasets. These datasets often comprise thousands of examples (e.g., 6300 examples in the TIMIT corpus [37] used in [6, 8, 27, 29–31, 33, 35]) or custom corpora consisting of hundreds of thousands of recordings used in [13–16].

3 Methodology

This section presents a detailed description of the main objective of the model developed as part of this study, as outlined in Section 3.1. It also describes the audio dataset used for training and evaluating the model, as detailed in Section 3.2. In Section 3.3, the feature extraction procedure is presented. Section 3.4 describes the model topology, whereas Section 3.5 address model training and evaluation.

3.1 Ensemble location and width definition

The objective of the model developed in this study is to predict the ensemble location (θ) and width (ω), as illustrated in Figure 1. An ensemble is defined as a group of audio point sources positioned on a circle around the listener on a virtual acoustic scene with equal distance to the listener. The location of source i is denoted by θ_i. The ensemble width (ω) is defined as the angular width between two extreme point sources ($max_i(\theta_i) - min_i(\theta_i)$), while the ensemble location, designated by θ, represents the middle angle between two extreme sound sources ($(max_i(\theta_i) + min_i(\theta_i))/2$). For the purposes of this study, the locations of the sources were limited to the frontal hemisphere only, i.e. $\theta \in [-45°, 45°]$, $\omega \in [0°, 90°]$, as this range encompasses the majority of real-world recording scenarios. It should be noted that although humans possess some limited abilities to localize sound sources in the vertical plane, in this study all sources are placed in the horizontal plane, at the height of the listener. This covers the majority of cases for real-world recordings (see [15, 35] for similar studies that cover top-down discrimination).

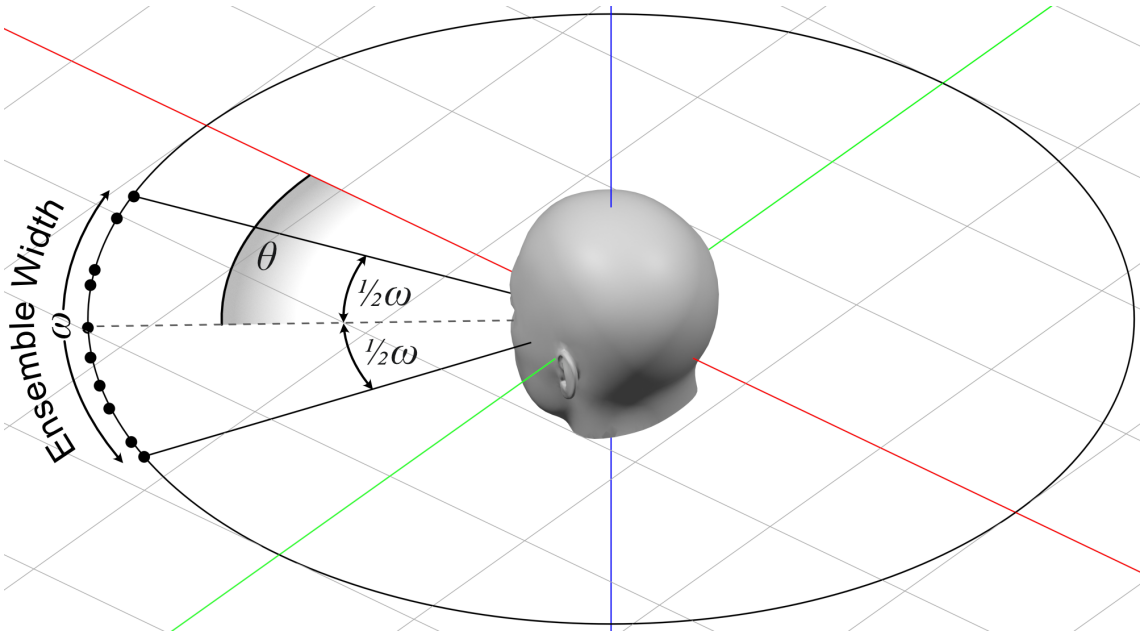

Figure 1: Illustration of ensemble width (ω) and ensemble location (θ) that are relative to the direction of the head. Black dots represent the positions of audio sources θ_i. The ensemble location (θ) is defined as the angular position of the center of the ensemble relative to the direction the head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

3.2 Synthesis of binaural music recordings

The experiments conducted in this study involved 23,040 binaural recordings of music. The binaural recordings were synthesized using 192 multi-track publicly-available music recordings [38] and 30 HRTF databases (see Appendix A for a detailed list of HRTF databases used in this study). The number of tracks in multi-track recording ranged from 5 to 62, with median of 9. For each pair of a multi-track recording and an HRTF database, four binaural recordings were synthesized with different random ensemble parameters, namely location θ and width ω, as defined in Section 3.1. Both parameters were drawn from a uniform random distribution. Furthermore, the tracks of the input multi-track recordings were randomly assigned to sound source positions (θ_i) to enhance the diversity of the final binaural corpora.

The binaural recordings were obtained in this study using the binaural synthesis procedure, known as binauralization, whose aim was to simulate the positions of sound sources within a virtual acoustic environment [1]. This was achieved by convolving multi-track signals with head-related impulse responses from a specified head-related transfer function (HRTF) database. The resulting binaural output signal $y_c[n]$ for each stereo channel c (left or right) at sample n is given by the following equation:

$$y_c[n] = \sum_{i=1}^{N} \sum_{k=0}^{K-1} x_i[k] \times h_{c,\theta_i}[n-k], \tag{1}$$

where x_i represents the signal of an individual sound source i from the input music recording and h_{c,θ_i} denotes the head-related impulse response for channel c at location θ_i of source track i.

Due to copyright restrictions, the music corpus utilized in this study was not published and can be provided upon request to the authors of this paper.

3.3 Feature extraction

Prior to input into the model, the binaural recordings of music were transformed into magnitude spectrograms. It is important to note that the original spectrograms were preserved in a floating-point precision matrix format, which prevented any loss of information due to precision conversion. The spectrograms were calculated using the Fast Fourier Transform (FFT) algorithm, with a limitation of 150 bands, spaced linearly from 100 Hz to 16 kHz. Additionally, a Hamming window of 40 ms was applied to each frame of the signal, resulting in a total of 349 time frames. This procedure was conducted for both the left and right channels, yielding two spectrograms for each binaural sample. Each sample was represented by a matrix of dimensions $2 \times 349 \times 150$. This method parallels the procedure presented in [14].

3.4 Network topology

The network topology used in this study was highly inspired by a AlexNet convolutional neural network [39]. Although AlexNet network was designed for image classification, the binaural recordings in this study were converted into magnitude spectrograms, as described in Section 3.3. The network is consisting of a series of convolutional units followed by a series of classification units.

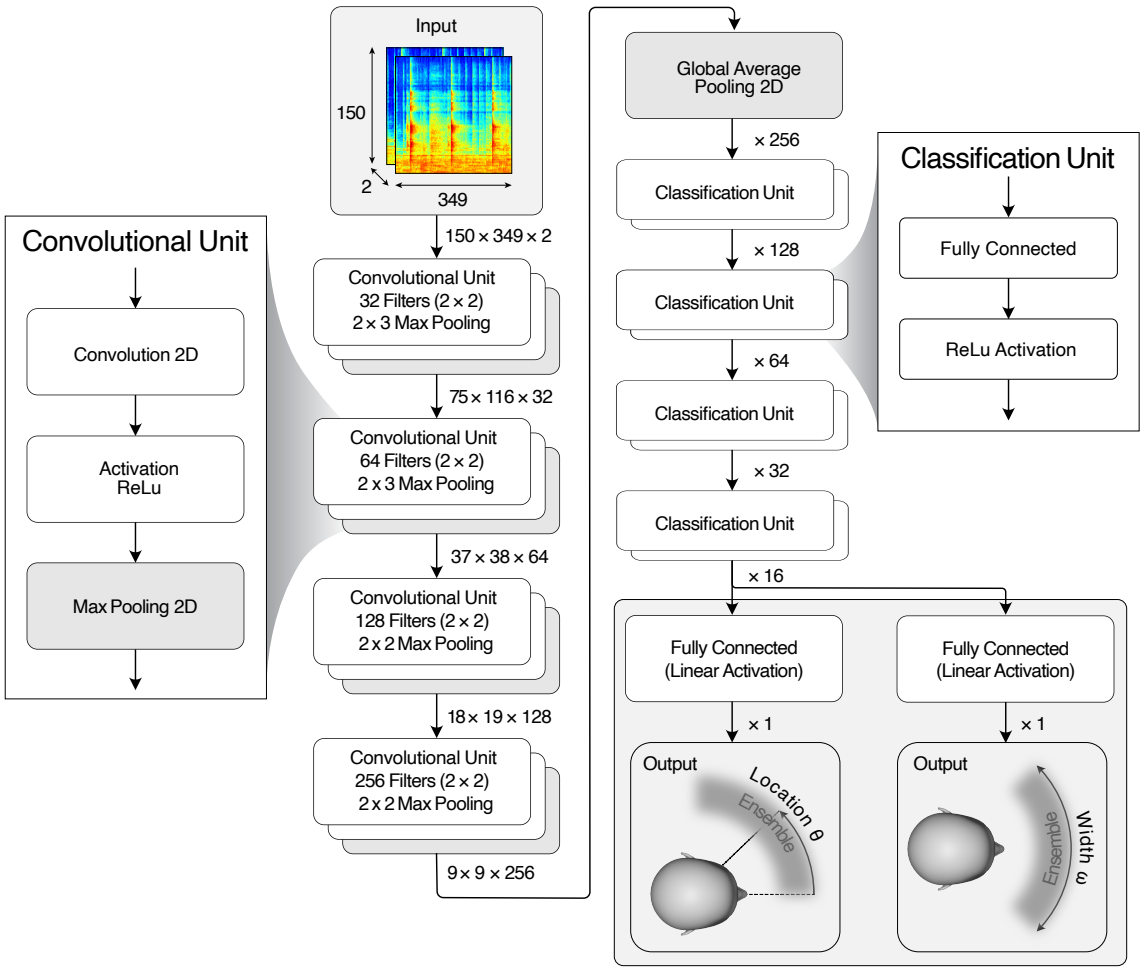

Figure 2: Topology of the Convolutional Neural Network (CNN) used for identifying ensemble location and width

3.5 Model training and evaluation

4 Results

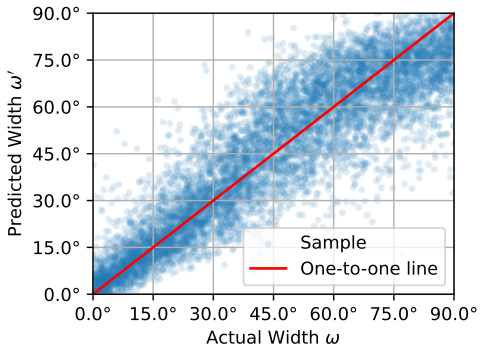

Figure 3: A comparison between the actual ensemble width ω and the predicted ensemble width ω' for a single iteration (of the total five)

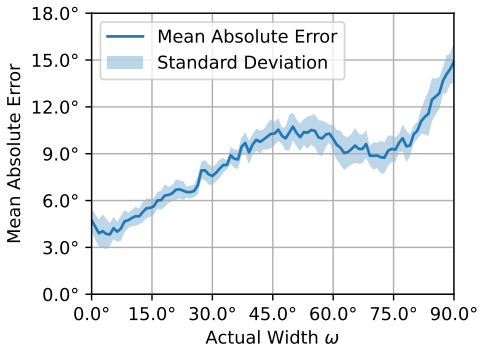

Figure 4: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

Figure 3 illustrates the comparison between the actual and predicted ensemble widths. The results demonstrate that the model exhibits better prediction quality for narrower ensemble widths (5.65° for $\omega < 30°$) and that its performance deteriorates with an increase in ensemble width (12.44° for $\omega > 80°$). This suggests that the actual ensemble width significantly impacts the accuracy of ensemble width estimation, leading to worse predictions as width increases. Figure 4 further demonstrates that the relationship between the error and the width is not linear, exhibiting an interesting depression between 60° and 75°.

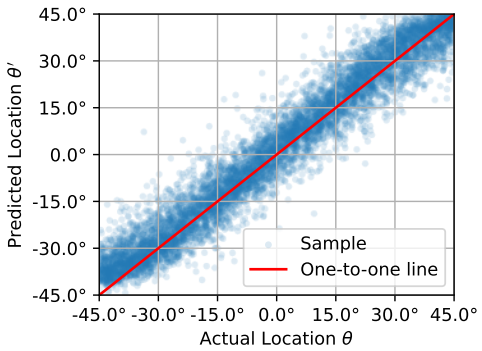

Figure 5: A comparison between the actual ensemble location θ and the predicted ensemble location θ' for a single iteration (of the total five)

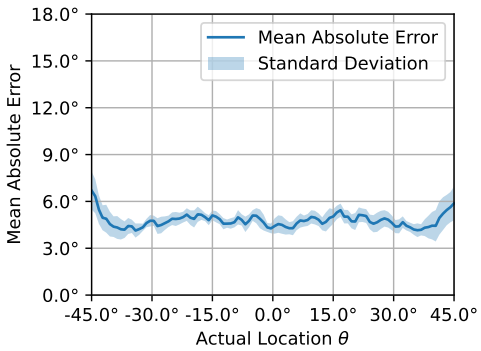

Figure 6: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

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 5 and 6. 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.

The mean absolute prediction error for ensemble width was 8.57° (±0.19°), while the error for ensemble location prediction was 4.76° (±0.10°). Although both ensemble parameters were constrained within the same angular span of 90°, the ensemble location was predicted with significantly better precision — approximately 80% better. This shows that the model prediction quality is much better for ensemble location prediction than width prediction. This difference can be attributed to the fact that the model's performance on ensemble width deteriorates significantly for wider ensembles, whereas location prediction remains largely unaffected. For a visual comparison, please refer to Figures 6 and 4.

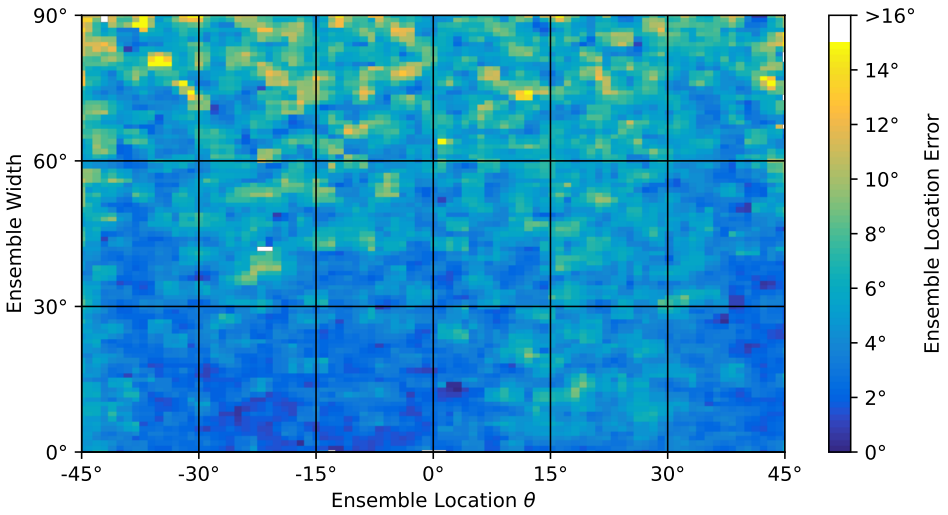

Figure 7: The heatmap that illustrates the mean absolute error (MAE) of ensemble location distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors.

Figure 7 illustrates the influence of both ensemble location and width on the mean absolute error for ensemble location, offering a detailed view complementing the results presented in Figure 6. Some asymmetric anomalies are depicted in this figure, mostly within the $\theta \in [15^\circ, 30^\circ]$ region, which can be attributed to the sparsity of sample result data across this heatmap. Although this figure primarily shows that ensemble location does not significantly influence the model's location prediction accuracy, it does demonstrate that ensemble width has a significant impact. Similarly, Figure 8 reveals a characteristic depression in $\omega \in [30^\circ, 60^\circ]$ previously shown from a different perspective in Figure 4. This heatmap highlights another interesting phenomenon in its upper corners — the error in these areas is considerably higher. This indicates that the model's performance for estimating ensemble width is substantially worse at extreme widths and locations, i.e., when both the width and locations are at their maximum.

5 Conclusions

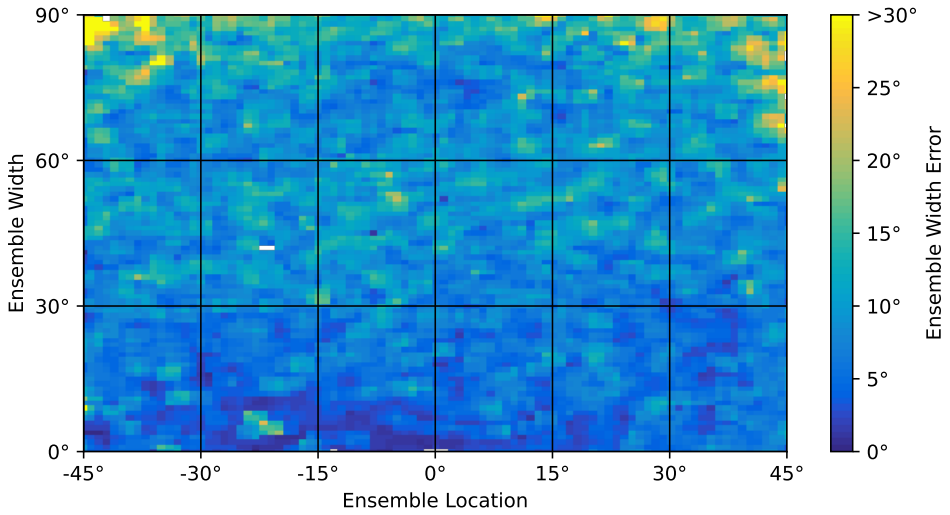

Figure 8: The heatmap that illustrates the mean absolute error (MAE) of ensemble width distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors. Notably, the values between 30° and 60° on the y-axis exhibit unexpectedly higher MAE values region — please see Figure 4 for comparison.

Appendix A

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. Hirsh, I. J. Binaural Hearing Aids: A Review Of Some Experiments. *Journal of Speech and Hearing Disorders* **15,** 114–123. ISSN: 0022-4677, 2163-6184. http://pubs.asha.org/doi/10.1044/jshd.1502.114 (2024) (June 1950).
5. 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).
6. 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).
7. 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).
8. 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).
9. LeCun, Y. *et al. Handwritten Digit Recognition with a Back-Propagation Network* in *Advances in Neural Information Processing Systems* **2** (Morgan-Kaufmann, 1989). https://proceedings.neurips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html (2024).
10. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/nature14539 (2024) (May 28, 2015).
11. 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).
12. 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).
13. 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).
14. 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).
15. 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).

16. 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).

17. 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).

18. 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).

19. 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).

20. 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).

21. 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).

22. 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).

23. 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).

24. 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).

25. 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).

26. 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).

27. 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).

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

29. 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).

30. 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).

31. 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).

32. 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).

33. 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).

34. 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).

35. 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).

36. Raake, A. *A computational framework for modelling active exploratory listening that assigns meaning to auditory scenes—reading the world with two ears* http://twoears.eu/ (2024).

37. Garofolo, John S. *et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus* Artwork Size: 715776 KB Pages: 715776 KB. 1993. https://catalog.ldc.upenn.edu/LDC93S1 (2024).

38. *The 'Mixing Secrets' Free Multitrack Download Library* https://cambridge-mt.com/ms/mtk/ (2024).

39. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks* in *Advances in Neural Information Processing Systems* **25** (Curran Associates, Inc., 2012). https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (2024).