jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2022, VOL. 68, NO. 4, PP. 300–306
Manuscript received September 10, 2022; revised October, 2022. DOI: 10.24425/ijet.2022.139xxx

Assessing Models for Estimation Ensemble Width in Binaural Music Recordings: Robustness to Reverberation and Noise

Paweł Antoniuk and Sławomir Krzysztof Zieliński

***Abstract*—Binaural technology has been known for decades. However, advancements in software and consumer electronics have facilitated its widespread adoption, primarily in the post-millennium era. It is anticipated that with the popularization of binaural sound, the demand for spatial analysis tools will increase. This paper compares three methods for assessing ensemble width in musical binaural recordings: (1) a method based on an auditory model and decision trees, (2) a method using neural networks, and (3) a method leveraging ”spatial spectrograms”. This work extends the research on these methods by analyzing their resilience to reverberation and noise. For this purpose, simulations of virtual acoustic rooms were conducted. The results of the models' noise resilience tests showed that the models are highly resistant to noise, with the model based on the auditory model demonstrating the highest robustness, presenting a mean absolute error of 12.34° for a signal-to-noise ratio (SNR) of 10 dB. The reverberation resilience tests revealed that the models' effectiveness significantly decreases even at a relatively short reverberation time of 0.3 seconds (according to RT60 method). This study may contribute to the improvement of models for estimating musical ensemble width in binaural recordings, which could influence the development of binaural sound analysis tools, with potential applications in audio production.**

***Keywords*—binaural audio, ensemble width, audio perception, localization, reverberation, machine learning**

I. Introduction

Binaural audio has been a cornerstone of immersive headphone listening for decades [1]. Recently, with its integration into virtual and augmented reality, its popularity has surged significantly [2]. By simulating the human auditory system's perception of sound in natural environments, binaural audio plays a crucial role in creating immersive audio-visual experiences for entertainment applications. Owing to its ability to allow listeners to naturally localize audio sources in direct-to-ear playback, binaural audio has also found successful applications in fields such as avionics [3] and hearing aid devices [4]. The utility of binaural hearing for these applications is illustrated by the 'cocktail party effect,' which highlights the human auditory system's ability to focus on foreground sounds while suppressing background noise [5].

The increasing availability of binaural audio highlights the need for advanced spatial analysis methods. These methods could facilitate automated, objective assessments of binaural recordings by analyzing spatial characteristics, such as the position and size of sound sources. Such analysis could support the development of tools to classify recordings based on these features and help assess the fidelity of binaural audio systems through spatial characteristics.

The aim of this study is to compare methods for estimating one of the most prominent spatial features: ensemble width. This feature is based on the observation that humans tend to localize groups of sound sources (ensembles) rather than individual sources [6], [7]. The approach draws from Rumsey's scene-based paradigm [7], which describes ensemble width as the 'overall width of a defined group of sources.' In immersive audio, this feature is particularly important, as wider ensembles enhance the perception of immersion by broadening the spatial distribution of sound sources, creating a more enveloping experience [8]. Notably, one of the methods presented also estimates ensemble location; however, this parameter will be omitted from the study due to limited comparative research.

This paper presents a summary and comparasion of the three methods for ensemble width estimation recently introduced by Sławomir Zieliński and Paweł Antoniuk. It also demonstrates their robustness to reverberation and noise.

II. Related Studies

Estimating ensemble width is a unique approach in binaural audio literature, which typically focuses on estimating locations of individual sound sources [9]–[16]. While estimating individual sound sources might seem more useful due to more precise information, such methods have limitations that prevent their use in practical applications. These include a limited or predetermined number of sound sources and predetermined type of audio signal——typically speech [9]–[11], [13]–[15], [15]–[17]. The ensemble approach can be seen as a workaround for these limitations, as it provides useful spatial information without such constraints.

Many traditional audio localization methods have focused on arrays with more than two microphones [18]–[22]. While adding microphones can enhance precision through additional channel information, they do not utilize binaural hearing

The work was supported by the grants from Bialystok University of Technology (WI/WI-IIT/3/2022 and WZ/WI-IIT/5/2023) and funded with resources for research by the Ministry of Science and Higher Education in Poland.

P. Antoniuk and S. K. Zieliński are with Faculty of Computer Science, Bialystok University of Technology, Poland (e-mail: pawel.antoniuk@sd.pb.edu.pl, s.zielinski@pb.edu.pl).

principles, rendering them ineffective for binaural recording assessment. As demonstrated by [Author et al.], two-microphone systems can achive better localization accuracy when incorporating binaural hearing principles [17].

The majority of studies on sound source localization in binaural recordings concentrate on the localization of individual sources in isolation, typically referred to as Direction of Arrival (DoA). Although this granular approach provides detailed information, the existing methods require a priori knowledge of the number of sources, typically limiting analysis to between one and six sources. These constraints present significant challenges in real-life scenarios, where such advance knowledge is unavailable. Moreover, these methods have been developed primarily for homogeneous signals, especially speech, making them impractical for real-world binaural recordings where signals are often heterogeneous.

In a series of recent studies, Sławomir Zieliński and Paweł Antoniuk introduced an alternative approach, treating sound sources as ensembles that can be characterized by their location and width, as illustrated in Figure 1. This method overcomes the limitations of traditional DoA approaches by focusing on ensemble characteristics rather than precise individual source locations. The approach eliminates the need for *a prior* knowledge of the number of sources and has been validated across diverse musical content, including both instrumental and vocal recordings.

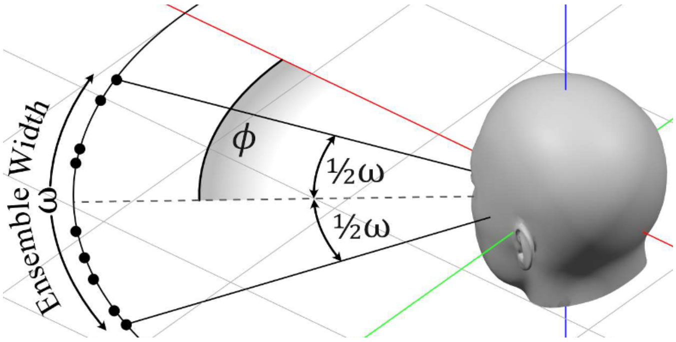

Fig. 1. An example of an ensemble consisting of nine point-like sound sources represented by dots. Ensemble width is signified by ω, whereas ϕ represents location of an ensemble center (counterclockwise).

This approach aligns with the second level of Rumsey's sound source localisation framework, which defines three distinct levels: (1) single source, (2) scene, and (3) environment. Scene-level analysis, as described by Rumsey, better matches the human auditory system's natural source-grouping mechanisms. The approach mirrors real-world musical performance configurations where instruments and vocals occupy adjacent spatial positions. Notably, the methods incorporated in this study specifically measure physical source width rather than apparent source width - two related but distinct parameters whose relationship warrants further investigation.

III. Methodology

This study compares three recent methods for ensemble width estimation:

1) a method based on an auditory model and decision trees (see Section IV),
2) a method using deep neural networks (see Section V),
3) a method leveraging spatial spectrograms (see Section VI).

Initially, these methods were evaluated using anechoic recordings without noise. To test their performance in more ecological conditions, the methods were also tested using signals with predefined signal-to-noise ratios and simulated rooms with different reverberation characteristics (see Section VIII).

The objective of the methods incorporated in this study is to estimate the ensemble width (ω) illustrated in Figure 1. An ensemble is defined as a group of audio point sources positioned equidistantly around the listener on a circular virtual acoustic scene. The location of source i is denoted by θ_i. The ensemble width (ω) represents the angular distance between the two extreme point sources $(\max_i(\theta_i) - \min_i(\theta_i))$, while the ensemble location, represented by ϕ, indicates the midpoint angle between these extreme sources $((\max_i(\theta_i) + \min_i(\theta_i))/2)$. In this study, source locations are restricted to the frontal hemisphere, specifically $\theta \in [-45°, 45°]$ and $\omega \in [0°, 90°]$. It should be noted that although humans have some limited abilities to localize sound sources in the vertical plane, all sources in this study are positioned on the horizontal plane at ear level. These constraints reflect most real-world recording scenarios.

IV. Auditory-Model-Based Method

V. Neural-Networks-Based Method

VI. Spatial-Spectrogram-Based Method

VII. Dataset Preparation

All models were trained and evaluated on the same binaural audio corpus consisting of 23,040 binaural recordings of music. These recordings covered many different genres, including rock, jazz, pop, and classical music. They were synthesized using 192 publicly-available multi-track music recordings and 30 HRTF databases. The number of tracks ranged from 5 to 62, with a median of 9. Four binaural recordings were synthesized for each multi-track recording and HRTF database, using different random ensemble locations ϕ and widths ω. The input tracks were randomly assigned to sound source positions θ_i. Before synthesis, each track was equalized to $-23\,\mathrm{LKFS}$ according to the ITU-R BS.1770-5 recommendation.

The objective was to create a diverse array of binaural recordings with the intention of enhancing the model's generalisability. This is particularly important for HRTFs since the specific HRTF used in real-world binaural synthesis is often unknown. Therefore, a model trained on a single HRTF would have limited practical use. Furthermore, the large dataset size benefits the machine learning models used in this study, particularly the one based on deep neural networks. For further details on the chosen multi-track recordings and HRTFs, see [1].

The binaural recordings were obtained using a binauralization procedure, implemented by convolving multi-track signals with head-related impulse responses from a specified 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 denotes the signal of an individual sound source i from the input music recording, and h_{c,θ_i} represents the head-related impulse response for channel c at location θ_i of source track i. Here, N denotes the number of track sources in the input multi-track recording, and K represents the number of samples in the recording.

The synthesized recordings were truncated to 7 seconds following binauralization, with sine-squared fade-in and fade-out effects of 0.01 seconds applied. Subsequently, the signals were RMS-normalized, scaled by a factor of 0.9, DC-offset corrected, and stored as uncompressed files at 48 kHz with 32-bit resolution.

The binaural recordings were randomly split into training and test sets with a 2:1 ratio. To prevent information leakage, this split was made in such a way that no multi-track recordings used for training were used for testing. To reduce the complexity of the experiment, the HRTFs were shared between both sets, which could be seen as a limitation of this study. However, it is known that the human auditory system operates with HRTFs that undergo only minimal changes throughout life, mainly during infancy. Therefore, this limitation could be considered consistent with how the human auditory system behaves in real life.

The binauralization and split procedures implemented in this study are consistent with those originally described in the reference models, with minor modifications. The primary modification pertains to the spatial-spectrogram-based model, which utilized a single HRTF database and employed a reduced parameter set. This modification had minimal impact on the results, as the method employs a deterministic approach rather than machine learning techniques, requiring the training set only for the optimization of two parameters.

VIII. Environmental Simulation

To enhance ecological validity, the original recording synthesis procedure was modified to enable evaluation under two additional scenarios: recordings with additive noise and recordings in reverberant conditions. In the first scenario, 9 test sets were prepared with different Signal-to-Noise Ratios (SNR) ranging from -10 to 60 dB, specifically at -10, -3, 0, 10, 20, 30, 40, 50, and 60 dB. This was achieved by adding decorrelated white noise signals to the binaural recordings originally used in the testing procedure.

In the reverberation scenario, 6 different rooms were simulated with reverberation times ranging from 0.1 to 3 s, measured using the RT60 metric. The rooms were simulated using MCRoomSim—a multichannel shoebox room acoustic simulator based on image source and diffuse rain algorithms implemented as a MATLAB package. This simulator enabled the creation of faithful reverberation simulations used to generate Binaural Room Impulse Responses (BRIRs) based on provided HRTFs, with the number of virtual speakers matching the spatial density of measurement points in the HRTF database. The virtual listener, modeled as a head with two receivers representing ears, was positioned in the center of the room. The receivers were configured to filter the input signal directionally using head-related impulse responses from the given HRTF database. The distance between each virtual impulse source and the head center matched the measurement radius of the given HRTF database, ranging from 0.9 to 1.95 m. The room reverberation characteristics were controlled by configuring the following parameters: room width and depth (2-5 m), height (2.5-5 m), wall absorption coefficients (0.05-0.95), and wall scattering coefficients (0.01-0.8).

IX. Results

Experimental results revealed that the auditory system-based model (1) performed best with an MAE of 6.63° (±0.12°), followed by the neural network-based model (2) at 8.57° (±0.19°), and the spatial spectrogram model (3) at 13.62° (±0.93°). The differences between results are significant, with $p < 0.01$ for all comparisons. As shown in Figure 2, all models demonstrated limited resilience to noise, with model (2) exhibiting the highest robustness for $\mathrm{SNR} > -3$, model (1) with $\mathrm{SNR}>10$, while model (3) showed lowest resilience for $\mathrm{SNR}>60$. The reverberation experiment indicated that none of the models exhibited significant robustness, with even the best-performing model (1) showing significant degradation at $\mathrm{RT}_{60} = 0.1$. This lack of robustness is attributed to the models not being trained on reverberant signals, suggesting a potential area for future improvement. The improvement could potentially lead to the development of objective assessment tools for audio engineers working with binaural audio.

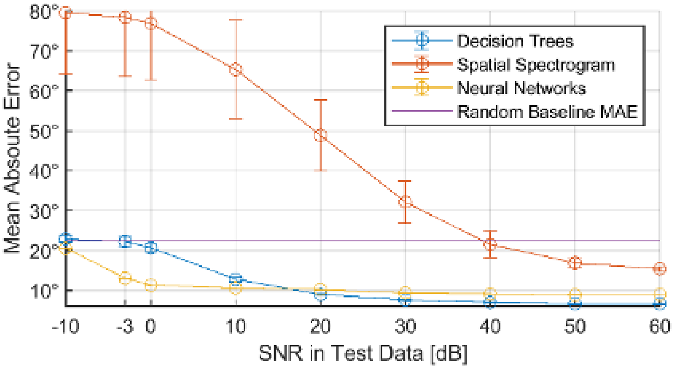

Fig. 2. Robustness to noise of the tested models, illustrating the mean absolute error (MAE) at varying signal-to-noise ratios (SNR)

X. Conclusions

Experimental results revealed that the auditory system-based model performed best with an MAE of 6.63° (±0.12°). All models demonstrated limited resilience to noise, with the neural-network-based model exhibiting the highest robustness for $\mathrm{SNR} > -3$. The reverberation experiment indicated that none of the models exhibited significant robustness, showing significant degradation at $\mathrm{RT}_{60} = 0.1$ s. This lack of robustness is attributed to the models not being trained on reverberant signals, suggesting a potential area for future improvement. The improvement could potentially lead to the development of objective assessment tools for audio engineers working with binaural audio.

References

[1] S. Paul, "Binaural Recording Technology: A Historical Review and Possible Future Developments," *Acta Acustica united with Acustica*, vol. 95, pp. 767–788, Sep. 2009.

[2] l. siegfried, "binaural audio in the era of virtual reality: a digest of research papers presented at recent aes conventions," *journal of the audio engineering society*, vol. 51, no. 11, pp. 1066–1072, Nov. 2003.

[3] D. Begault and E. Wenzel, "Techniques and Applications for Binaural Sound Manipulation," *International Journal of Aviation Psychology - INT J AVIAT PSYCHOL*, vol. 2, pp. 1–22, Feb. 1992.

[4] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 12, Feb. 2016. [Online]. Available: https://doi.org/10.1186/s13634-016-0314-6

[5] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, Sep. 1953, _eprint: https://pubs.aip.org/asa/jasa/article-pdf/25/5/975/18731769/975_1_online.pdf. [Online]. Available: https://doi.org/10.1121/1.1907229

[6] A. Bregman, "Auditory Scene Analysis: The Perceptual Organization of Sound," in *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, Jan. 1990, vol. 95, journal Abbreviation: Journal of The Acoustical Society of America - J ACOUST SOC AMER.

[7] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, p. 16, 2002.

[8] D. Griesinger, "The Psychoacoustics of Apparent Source Width, Spaciousness and Envelopment in Performance Spaces," *Acta Acustica united with Acustica*, vol. 83, pp. 721–731, Jul. 1997.

[9] E. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raumel, and S. Argentieri, "Binaural Localization of Multiple Sound Sources by Non-Negative Tensor Factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1072–1082, Jun. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8294267/

[10] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, May 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016763931000097X

[11] N. Ma and G. J. Brown, "Speech Localisation in a Multitalker Mixture by Humans and Machines," in *Interspeech 2016*. ISCA, Sep. 2016, pp. 3359–3363. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2016/ma16c_interspeech.html

[12] N. Ma, T. May, and G. J. Brown, "Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8086216/

[13] T. May, S. Van De Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5406118/

[14] ——, "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*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6178270/

[15] T. May, N. Ma, and G. J. Brown, "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)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 2679–2683. [Online]. Available: http://ieeexplore.ieee.org/document/7178457/

[16] J. Woodruff and D. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6129395/

[17] Q. Yang and Y. Zheng, "DeepEar: Sound Localization With Binaural Microphones," *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 359–375, Jan. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9954178/

[18] M.-A. Chung, H.-C. Chou, and C.-W. Lin, "Sound Localization Based on Acoustic Source Using Multiple Microphone Array in an Indoor Environment," *Electronics*, vol. 11, no. 6, p. 890, Jan. 2022, number: 6 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9292/11/6/890

[19] M. Hahmann, E. Fernandez-Grande, H. Gunawan, and P. Gerstoft, "Sound source localization using multiple ad hoc distributed microphone arrays," *JASA Express Letters*, vol. 2, no. 7, p. 074801, Jul. 2022.

[20] M. Liu, J. Hu, Q. Zeng, Z. Jian, and L. Nie, "Sound Source Localization Based on Multi-Channel Cross-Correlation Weighted Beamforming," *Micromachines*, vol. 13, no. 7, p. 1010, Jul. 2022, number: 7 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2072-666X/13/7/1010

[21] Z. Pan, M. Zhang, J. Wu, J. Wang, and H. Li, "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*, vol. 29, pp. 2656–2670, 2021, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing. [Online]. Available: https://ieeexplore.ieee.org/document/9502013

[22] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "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)*, Mar. 2012, pp. 2625–2628, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/6288455