INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 4, PP. 1–7
Manuscript received May 28, 2025; revised October 2025. doi: 10.24425/ijet.2025.155458

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. As binaural sound becomes more popular, the demand for spatial analysis tools is expected to grow. This paper evaluates three methods for assessing ensemble width in binaural music recordings: (1) an auditory model with decision trees, (2) a neural network model, and (3) a spatial spectrogram approach. Under ideal, anechoic conditions, the auditory model performed best with a mean absolute error (MAE) of 6.59° (±0.11°), followed by the neural network (8.57° ±0.19°) and the technique based on spatial spectrograms (13.54° ±0.92°). Extending previous work, this study analyzes the methods' robustness to reverberation and noise. Noise resilience tests indicate moderate resistance, with the auditory model yielding an MAE of 12.34° at a 10 dB signal-to-noise ratio. However, reverberation tests show a significant drop in accuracy even at an RT60 reverberation time of 0.1 seconds. The findings may contribute to the improvement of models for estimating ensemble width in binaural recordings of music, 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 exploiting 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 work was supported by 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).

The increasing availability of binaural audio applications 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, while one of the presented methods also estimates ensemble location (see Section VI), this parameter will be omitted from the study because ensemble width is the only parameter comparable across all three methods.

This paper provides a comparative summary of three ensemble width estimation methods. The first two methods were introduced by Antoniuk et al., where the first one was based auditory model and gradient-boosted decision trees [9] and the second one was based on convolutional neural network with very limited feature engineering [10]. The third method employed spatial spectrograms. It was introduced by Arthi and Sreenivas [11] and later refined by Antoniuk and Zieliński [12]. The primary contribution of this study is an evaluation of these methods' robustness under more ecologically valid conditions, focusing on their resilience to noise and reverberation. This evaluation offers insights into their applicability in real-world scenarios.

Deep learning paradigm has become the dominant technique in modern machine learning research and applications. It often shows superiority over traditional feature-engineering-based techniques thanks to its ability to extract unknown features (and thus knowledge) from large sets of data. However, deep learning methods typically require datasets with larger sample sizes and greater variability. This is necessary to effectively

© The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided that the Article is properly cited.

'discover' features needed for accurate prediction [13]. As a result, these models face an increased risk of overfitting when such data requirements are not met. In contrast, traditional machine learning techniques generally possess lower 'capacity' and rely more heavily on feature engineering, making them more robust against overfitting. This robustness stems from the feature engineering process, which transforms data into more informative representations by incorporating domain knowledge already discovered by researchers.

Feature engineering also mitigates the risk of spurious correlations—instances where models learn patterns from incidental correlations in training data that fail to generalize to real-world scenarios [14]. A compelling example of this phenomenon emerges from computer vision research, where Ribeiro et al. demonstrated a neural network that misclassified 'husky' dogs as wolves based primarily on the presence of snow in the background, rather than the distinctive morphological features of the animals themselves [15]. This type of failure, where the model attends to contextual rather than intrinsic features, can be substantially reduced through thoughtful feature engineering that explicitly encodes domain-relevant characteristics.

Given that none of the models in this study were trained on data containing noise and reverberation, we hypothesize that the first model—based on an auditory model combined with gradient-boosted decision trees—will demonstrate the best overall performance and the greatest robustness to noise and reverberation. Conversely, we expect the second model, which utilizes a convolutional neural network architecture, to exhibit greater sensitivity to these adverse conditions. The third model, although leveraging an innovative and promising algorithmic spatial-spectrogram technique, will likely yield the least favorable results, consistent with its relatively low accuracy reported in previous research [12].

II. Related Studies

Estimating ensemble width represents a unique approach within binaural audio literature, which more commonly focuses on identifying the locations of individual sound sources [16]–[20]. While analyzing individual sound sources might seem more useful because it yields more precise information, such methods have limitations that hinder their practical application. These include a limited or predetermined number of sound sources and a predetermined type of audio signal—typically speech [16], [17], [19]–[21]. The ensemble approach serves as a workaround for these limitations by providing useful spatial information without such constraints.

Traditional audio localization methods often rely on arrays with more than two microphones to improve precision through additional channel information [22]–[24]. While adding microphones can enhance precision through additional channel information, they do not utilize binaural hearing principles, rendering them ineffective for binaural recording assessment. By contrast, as Yang et al. demonstrate, systems using only two microphones can achieve superior localization accuracy by integrating binaural cues [21].

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) [16]–[20]. Although this granular approach provides detailed information, the existing methods require *a priori* knowledge of the number of sources, usually 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, Arthi and Sreenivas [11], Antoniuk at al. [9], [10], [12], 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 [9], [10], [12].

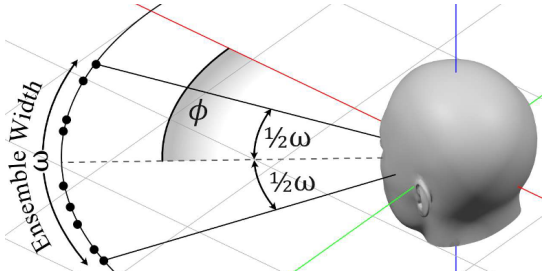

Fig. 1. An ensemble example comprising nine point-like sound sources shown as dots. The ensemble's width is denoted by ω, while ϕ indicates the counterclockwise position of the ensemble's center.

This approach aligns with the second level of Rumsey's spatial audio scene-based framework [7], 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 ensemble width rather than apparent ensemble 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 (Section V),
2) a method using deep neural network (Section VI),
3) a method leveraging spatial spectrograms (Section VII).

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) as 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. Dataset Preparation

The experimental evaluation was conducted using a corpus of 23,040 synthesized binaural music recordings. The source material comprised 192 publicly-available multi-track recordings spanning diverse musical genres including rock, jazz, pop, and classical music. The number of tracks ranged from 5 to 62, with a median of 9.

To ensure robust evaluation across diverse Head-Related Transfer Function (HRTF) characteristics, the synthesis process incorporated 30 HRTF databases (see Table **??** in Appendix for a detailed list). These databases were evenly divided into measurements from human subjects (15 databases) and measurements from artificial heads (15 databases), including industry-standard devices such as the Neumann KU 100 and KEMAR DB-4004. Distances between the head and loudspeaker during HRTF measurements ranged from 0.9 to 1.95 meters, with a median of 1.2 meters.

For each combination of multi-track recording and HRTF database, four unique binaural versions were synthesized by randomly varying two ensemble parameters: location (ϕ) and width (ω). Within these spatial constraints, individual tracks in each recording were randomly assigned to specific source positions (θ_i). Prior to synthesis, all tracks were loudness-normalized to -23 LKFS in accordance with ITU-R BS.1770-5 recommendations [25], ensuring consistent relative levels across the corpus.

Multiple HRTFs and multi-track recordings were selected to create diverse binaural recordings, enhancing the model's generalisability. This diversity is crucial for HRTFs since the specific HRTF used in real-world binaural synthesis is often unknown, making a single-HRTF model impractical for general use. Additionally, the large dataset provides essential training material for all machine learning models used in this study, with particular importance for the deep neural networks, as their performance benefits significantly from extensive data.

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. Additionally, 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 with a sample rate of 48 kHz and 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 [26]. 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 [9], [10], [12], 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.

V. Auditory-Model-Based Method

As shown in Figure 2, the auditory-model-based method for ensemble width estimation consists of two main components: a binaural auditory model that extracts features from the input signals, followed by a gradient-boosted decision tree regressor that predicts the ensemble width [9]. The auditory model processes the binaural signals through a gammatone filterbank and extracts standard binaural cues, including interaural time differences (ITD), interaural level differences (ILD), and interaural cross-correlation (IACC).

The auditory model is based on the work of Søndergaard and Majdak [27], enhanced by May et al. [19], and further refined by Decorsière and May [28] within Two!Ears Project [29]. The model consists of a gammatone filterbank with 64 frequency channels spanning from 100 Hz to 16 kHz. For each frequency channel, the inner hair-cell envelopes are extracted through half-wave rectification followed by low-pass filtering using a second-order Butterworth filter with

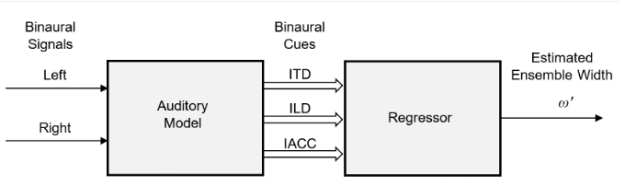

Fig. 2. A flowchart of the auditory-model-based method [9]

a 1 kHz cutoff frequency. This simulates the loss of phase-locking in the auditory nerve at higher frequencies. Rate maps, representing auditory nerve firing rates, are then calculated by smoothing the inner hair-cell signal with a leaky integrator (time constant of 8 ms) and averaging within 20 ms Hann-windowed frames with 10 ms step size. Finally, the rate maps were used to estimate the binaural cues.

The features extracted in the previous component are aggregated across time-frames by computing their mean values and standard deviations, resulting in a total of 384 feature vectors (64 frequency channels $\times$ 3 types of cues $\times$ 2 statistics). These aggregated features are then used as input to the regressor, whose objective is to estimate the ensemble width of the binaural audio signal. The regressor employs gradient-boosted decision trees implemented with LightGBM, known for its computational efficiency and accuracy [30]. The model's hyperparameters—including the number of leaves, tree depth, and learning rate—were optimized using grid search procedure. The final training was conducted using validation set and early-stopping technique based on mean absolute error. For further details, see [9].

VI. Neural-Network-Based Method

While the auditory-model-based method attempts to mimic human hearing mechanisms, the neural network approach takes advantage of Convolutional Neural Networks (CNNs) and their ability to automatically learn relevant features from spectral representations of audio signals. In contrast to the auditory-model-based approach, the neural network method uses a basic feature extraction technique based on magnitude spectrograms [10].

A Hamming window of 40 ms with an overlap of 20 ms is applied, resulting in 349 time frames extracted for each binaural input signal. For each time frame, the Fast Fourier Transform (FFT) is applied. The magnitudes of its output are aggregated into 64 linearly spaced frequency bands ranging from 100 Hz to 16 kHz, effectively creating a spectrogram with dimensions 349×64. This process is performed separately for the left and right channels, producing a pair of spectrograms that are then used in the neural network to simultaneously estimate two ensemble parameters: width and location. However, only ensemble width is considered in this study.

In the next step, the spectrograms are input into a two-dimensional CNN model to estimate the ensemble width, effectively treating the spectrograms as visual data. The network's topology is based on the AlexNet model introduced by Krizhevsky et al. [31]. The input layer is followed by five convolutional units, each consisting of a ReLU-activated 2D convolution layer with a 2×2 filter size, followed by a max pooling layer of size 2×3 or 2×2. The number of convolutional filters in each layer is 32, 64, 128, and 256, respectively. Following these layers, a global average pooling layer is applied to reduce overfitting [32]. The next stage consists of four fully connected layers with ReLU activation, reducing the activation map's dimensions from 256 to 6. Finally, two parallel fully connected layers with linear activation are used to predict the ensemble parameters; one outputs the ensemble width, and the other outputs the ensemble location.

In total, this topology resulted in a model with 216,562 learning parameters. The model was trained using a Monte Carlo cross-validation procedure with 10 repetitions and an early-stopping validation subset. For further details, see [10].

VII. Spatial-Spectrogram-Based Method

The spatial-spectrogram-based method, originally introduced by Arthi and Sreenivas [11], employs a phase-only spatial correlation (POSC) function to estimate ensemble width, treating the binaural signals primarily in the frequency domain. The method consists of three main steps: calculation of generalized cross-correlation functions, generation of spatial spectrograms, and ensemble width estimation.

First, two generalized cross-correlation functions with phase transform (GCC-PHAT) are calculated:

$$\rho(k) \overset{\mathcal{F}}{\longleftrightarrow} \frac{X_r(\omega) \times X_l^*(\omega)}{|X_r(\omega) \times X_l^*(\omega)|}, \tag{2}$$

$$\rho_\theta(k) \overset{\mathcal{F}}{\longleftrightarrow} \frac{H_r^\theta(\omega) \times H_l^{\theta *}(\omega)}{|H_r^\theta(\omega) \times H_l^{\theta *}(\omega)|}, \tag{3}$$

where $\rho(k)$ denotes the GCC-PHAT function for the k-th sample of the binaural signal; $\rho_\theta(k)$ represents the GCC-PHAT function for the k-th sample of the HRIR; $X_l(\omega)$ and $X_r(\omega)$ are Fourier transforms of the left and right channel signals, respectively; H_l^θ and H_r^θ are Fourier transforms of the left and right channels, respectively, for the HRIR at azimuth θ; and * denotes complex conjugate. The phase-only spatial correlation (POSC) function $C_\rho(\theta)$ is then calculated as:

$$C_\rho(\theta) \triangleq \sum \rho(k) \times \rho_\theta(k) \tag{4}$$

To account for the observation that sources closer to 0° have a greater impact on $C_\rho(\theta)$ than more distant sources, a correction is applied:

$$\widetilde{C_\rho(\theta)} = C_\rho(\theta) \times (1 + \theta u), \tag{5}$$

where u is a correction weight determined through optimization.

The ensemble width is estimated using a three-step algorithm:

1) Find $\max \widetilde{C_\rho(\theta)}$ considering all frames in the binaural excerpt.

2) Find the minimal (θ_a) and maximal (θ_b) roots of:

$$\widetilde{C_\rho(\theta)} = t_h \times \max \widetilde{C_\rho(\theta)}, \tag{6}$$

where $t_h \in [0, 1]$ is a threshold coefficient.

3) Calculate ensemble width as $\omega = \theta_b - \theta_a$ averaged over all frames.

The method requires optimization of only two parameters: the correction weight u and threshold coefficient t_h. These parameters are determined using a grid search procedure with $u \in [0, 2]$ and $t_h \in [0, 1]$. Unlike the previous two methods, this approach is deterministic and does not require extensive training data, making it computationally efficient but potentially less accurate. For further details, see [12].

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, nine 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. While the upper range of these SNR values (40-60 dB) approaches imperceptible noise levels for human listeners, this extended testing range was included to test whether models not trained on noisy data would show degraded performance with even minimal signal interference. This wide testing range proved necessary only for the spatial-spectrogram-based model, which showed significant sensitivity to noise levels that would be barely perceptible to human listeners.

In the reverberation scenario, six different rooms were simulated with reverberation times ranging from 0.1 to 3 s, measured using the RT60 metric. The simulations were performed with MCRoomSim—a multichannel 'shoebox' room acoustic simulator based on image source and diffuse rain algorithms implemented as a MATLAB package [33]. This simulator enabled the creation of 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

Under baseline anechoic, noise-free conditions, the method based on an auditory model (1) achieved the highest accuracy, with a Mean Absolute Error (MAE) of $6.59°$ ($\pm 0.11°$). This was followed by the neural network-based method (2) at $8.57°$ ($\pm 0.19°$) and the spatial-spectrogram-based method (3) at $13.54°$ ($\pm 0.92°$). All differences between the methods were statistically significant ($p < 0.01$). Auditory-model-based (1) and neural-network-based methods (2) exhibited varying degrees of noise resilience (Figure 3). The neural-network-based method (2) maintained reasonable performance down to $\mathrm{SNR} = -3\,\mathrm{dB}$, while the method incorporating an auditory model and decision trees (1) required $\mathrm{SNR} > 10\,\mathrm{dB}$ for comparable results. The spatial-spectrogram-based method (3) was the most sensitive to noise, requiring $\mathrm{SNR} \geq 60\,\mathrm{dB}$ to operate reliably. These differences were statistically significant, with $p < 0.01$.

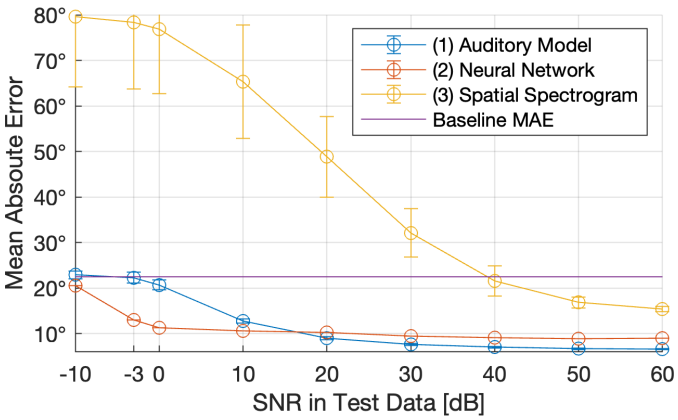

Fig. 3. Robustness to noise of the tested methods illustrating the mean absolute error (MAE) at varying signal-to-noise ratios (SNR). Error bars denote standard deviations. The baseline constant predictor (MAE = 22.5°) is included for comparison.

Under reverberant conditions, all of the tested methods revealed significant limitations in performance (Figure 4). In particular, the auditory-model-based method demonstrated notable difficulties even at minimal investigated reverberation times ($\mathrm{RT60} = 0.1\,\mathrm{s}$). While all the methods outperformed the random baseline MAE at $\mathrm{RT60} = 0.1\,\mathrm{s}$ ($p < 0.01$), they demonstrated notable limitations. The primary cause seems to stem from their exclusive training on anechoic signals, leaving them ill-suited for the added temporal and spectral complexity of room reflections.

X. Conclusion

This study summarizes and compares three approaches to ensemble width estimation in binaural recordings of music. The auditory-system-based method demonstrated superior performance in the baseline test, with an MAE of $6.59° \pm 0.11°$. This suggests that combining auditory modeling expertise with a traditional feature-based machine learning algorithm can be more effective than relying solely on deep learning techniques in this context.

Auditory-model-based (1) and neural-network-based (2) methods demonstrated a moderate robustness to noise, down to $\mathrm{SNR} = 10\,\mathrm{dB}$ and $\mathrm{SNR} = -3\,\mathrm{dB}$, respectively. However the performance of all the tested techniques exposed significant limitations under reverberant conditions, showing that

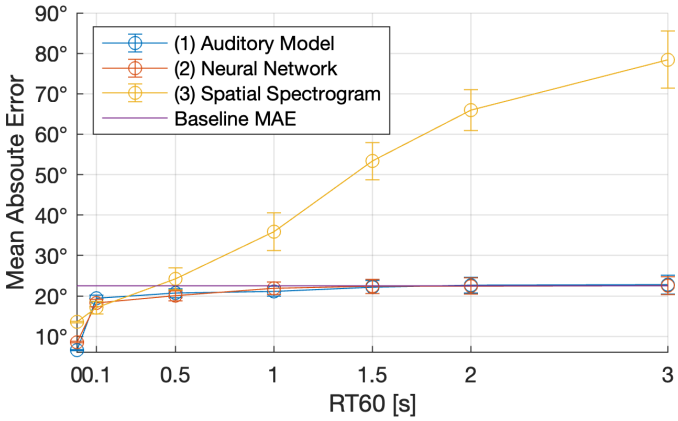

Fig. 4. Robustness to reverberation across tested models, shown as mean absolute error (MAE) for simulated rooms with varying RT60 values. Error bars denote standard deviations. The baseline constant predictor (MAE = 22.5°) is included for comparison.

these methods are not yet ready for real-world applications. This limitation is likely caused by the fact that these models were trained exclusively on anechoic signals, suggesting a direction for future research: developing models that incorporate realistic room acoustics into the training process.

The findings only partially confirm the hypothesis that the auditory model-based approach outperforms the deep learning neural-network method. This holds true only for baseline results without interference. Surprisingly, the neural-network-based model outperformed the auditory model at SNR levels below 20 dB, demonstrating unexpected resilience to noise. This suggests that the neural network successfully extracted robust features from the data, effectively mitigating spurious correlations despite the absence of explicit human-controlled feature engineering. The spatial-spectrogram-based method performed as hypothesized, showing the least robustness to both noise and reverberation.

Addressing the limitations above could lead to more robust binaural audio quality-assessment tools suitable for practical applications in audio production. Additionally, the varying performance characteristics observed across different acoustic conditions suggest potential benefits in hybrid approaches combining strengths of multiple models.

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] S. Linkwitz, "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," *Journal of the Audio Engineering Society*, vol. 50, pp. 651–666, Sep. 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] P. Antoniuk, S. K. Zieliński, and H. Lee, "Ensemble width estimation in HRTF-convolved binaural music recordings using an auditory model and a gradient-boosted decision trees regressor," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 53, Oct. 2024. [Online]. Available: https://doi.org/10.1186/s13636-024-00374-2

[10] P. Antoniuk and S. K. Zielinski, "Estimating Ensemble Location and Width in Binaural Recordings of Music with Convolutional Neural Networks," *Archives of Acoustics*, 2024, Accepted for publication.

[11] S. Arthi and T. V. Sreenivas, "Binaural Spatial Transform for Multi-source Localization determining Angular Extent of Ensemble Source Width," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*. Bangalore, India: IEEE, Jul. 2022, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9840782/

[12] P. Antoniuk and S. K. Zieliński, "Blind estimation of ensemble width in binaural music recordings using 'spatiograms' under simulated anechoic conditions," in *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio*, Aug. 2023. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=22203

[13] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, p. 91, May 2023, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2073-431X/12/5/91

[14] W. Ye, G. Zheng, X. Cao, Y. Ma, and A. Zhang, "Spurious Correlations in Machine Learning: A Survey," May 2024, arXiv:2402.12715 [cs]. [Online]. Available: http://arxiv.org/abs/2402.12715

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939778

[16] 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/

[17] 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

[18] 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/

[19] 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/

[20] 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/

[21] 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/

APPENDIX

TABLE I: List of HRTF sets used to synthesize binaural audio excerpts

No.	Type	Head	Radius [m]	Source	Acronym
1.	Human	Human subject	1.2	RWTH Aachen University	AACHEN
2.	Artificial	GRAS 45BB-4 KEMAR	1		
3.	Human	Subject 2	1.2	Austrian Academy of Sciences	ARI
4.	Human	Subject 4	1.2		
5.	Human	Subject 10	1.2		
6.	Artificial	ARI Printed Head	1.2		
7.	Human	Subject 012	1	CIPIC Interface Laboratory, University of California	CIPIC
8.	Human	Subject 015	1		
9.	Human	Subject 020	1		
10.	Artificial	Neumann KU 100	0.9	NASA (2007)	CLUBFRITZ
11.	Artificial	Neumann KU 100	1.5	Helsinki University of Technology (2009)	
12.	Artificial	FABIAN	1.47	Technical University Berlin, Huawei Technologies, Munich Research Centre, Sennheiser Electronic	HUTUBS
13.	Human	Subject pp2	1.47		
14.	Human	Subject pp3	1.47		
15.	Human	Subject 1003	1.95	IRCAM, AKG	LISTEN
16.	Human	Subject 1002	1.95		
17.	Artificial	KEMAR DB-4004 (DB-061)	1.4	MIT	MIT
18.	Artificial	KEMAR DB-4004 (DB-065)	1.4		
19.	Human	Subject 001	1.5	Tohoku University	RIEC
20.	Human	Subject 002	1.5		
21.	Artificial	Koken SAMRAI	1.5		
22.	Artificial	Neumann KU 100	1.2	University of York	SADIE II
23.	Human	Subject H3	1.2		
24.	Human	Subject H4	1.2		
25.	Artificial	KEMAR	1	South China University of Technology	SSCUT
26.	Artificial	Neumann KU 100	1	TH Köln	STH Köln
27.	Artificial	FABIAN	1.7	TU Berlin	TU Berlin
28.	Artificial	GRAS 45BA KEMAR	1		
29.	Artificial	GRAS 45BB-4 KEMAR - subject A attachment	1	Aalborg University; University of Iceland	VIKING
30.	Artificial	GRAS 45BB-4 KEMAR - subject B attachments	1		

[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

[23] 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.

[24] 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

[25] "ITU-R BS.1770-5: Algorithms to measure audio programme loudness and true-peak audio level," in *International Communications Union*, Geneva, Switzerland, Nov. 2023.

[26] A. J. King, O. Kacelnik, T. D. Mrsic-Flogel, J. W. Schnupp, C. H. Parsons, and D. R. Moore, "How Plastic Is Spatial Hearing?" *Audiology and Neurotology*, vol. 6, no. 4, pp. 182–186, Nov. 2001. [Online]. Available: https://doi.org/10.1159/000046829

[27] J. Blauert, Ed., *The Technology of Binaural Listening*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. [Online]. Available: http://link.springer.com/10.1007/978-3-642-37762-4

[28] R. Decorsière and T. May, "Auditory front-end. Two Ears Project Documentatio," 2016. [Online]. Available: https://docs.twoears.eu/en/latest/afe/

[29] A. Raake, "A computational framework for modelling active exploratory listening that assigns meaning to auditory scenes—reading the world with two ears," 2016. [Online]. Available: http://twoears.eu/

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3065386

[32] M. Lin, Q. Chen, and S. Yan, "Network In Network," *CoRR*, vol. abs/1312.4400, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:16636683

[33] A. Wabnitz, N. Epain, C. T. Jin, and A. v. Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics (ISRA)*, 2010. [Online]. Available: https://www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ISRA2010/Papers/P5d.pdf