EURASIP Journal on Audio,
Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Ensemble width estimation in HRTF-convolved binaural music recordings using an auditory model and a gradient-boosted decision trees regressor

Paweł Antoniuk[1], Sławomir K. Zieliński[1*] and Hyunkook Lee[2]

Abstract

Binaural audio recordings become increasingly popular in multimedia repositories, posing new challenges in indexing, searching, and retrieval of such excerpts in terms of their spatial audio scene characteristics. This paper presents a new method for the automatic estimation of one of the most important spatial attributes of binaural recordings of music, namely "ensemble width." The method has been developed using a repository of 23,040 binaural excerpts synthesized by convolving 192 multi-track music recordings with 30 sets of head-related transfer functions (HRTF). The synthesized excerpts represented various spatial distributions of music sound sources along a frontal semicircle in the horizontal plane. A binaural auditory model was exploited to derive the standard binaural cues from the synthesized excerpts, yielding a dataset representing interaural level and time differences, complemented by interaural cross-correlation coefficients. Subsequently, a regression method, based on gradient-boosted decision trees, was applied to the formerly calculated dataset to estimate ensemble width values. According to the obtained results, the mean absolute error of the ensemble width estimation averaged across experimental conditions amounts to 6.63° (SD 0.12°). The accuracy of the method is the highest for the recordings with ensembles narrower than 30°, yielding the mean absolute error ranging between 0.8° and 10.2°. The performance of the proposed algorithm is relatively uniform regardless of the horizontal position of an ensemble. However, its accuracy deteriorates for wider ensembles, with the error reaching 25.2° for the music ensembles spanning 90°. The developed method exhibits satisfactory generalization properties when evaluated both under music-independent and HRTF-independent conditions. The proposed method outperforms the technique based on "spatiograms" recently introduced in the literature.

Keywords Ensemble width, Binaural recordings, Spatial audio scene characterization

1 Introduction

The increased popularity of binaural technologies seen in the recent decade gave rise to a surge in the number of spatial audio recordings uploaded in the binaural format to publicly available repositories on the Internet, including Freesound, YouTube, and Vimeo. A substantial number of these excerpts incorporate music. Hence, one of the greatest challenges faced nowadays by researchers in the area of music information retrieval (MIR) is to analyze such recordings in terms of their spatial properties—a task referred to as spatial audio scene characterization

*Correspondence:
Sławomir K. Zieliński
s.zielinski@pb.edu.pl
[1] Faculty of Computer Science, Białystok University of Technology, Białystok 15-351, Poland
[2] Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield HD1 3DH, UK

(SASC) [1]. Consequently, there is a growing need for the development of methods dedicated to spatial audio scene analysis in binaural signals. The potential application scope of such techniques is not limited to the semantic search or retrieval of binaural recordings. These methods might also enhance the performance of music genre recognition algorithms as spatial characteristics of music recordings are likely to be genre-specific [2]. For example, in classical or jazz music recordings, music ensembles are typically arranged on the stage in front of the listeners (stage-audience scenario), whereas, for pop, dance, or electronic music recordings, sound sources tend to be distributed by mixing engineers in a much wider way, even to the extent that the listeners are surrounded by musicians (360° stage scenario [3]). Moreover, information derived by spatial binaural analyzers could be used to adaptively "steer" the parameters of the up-mixing algorithms, e.g., converting binaural signals to multichannel loudspeaker-based formats.

Drawing inspiration from Rumsey's scene-based paradigm [4], we propose that the spatial content of music audio recordings could be characterized at the three abstraction levels: low, mid, and high. At the low level of abstraction, spatial scenes might be described in terms of the spatial distribution and characteristics of the individual sound sources, including their widths and depths, as illustrated in Fig. 1. At the mid-level, scenes could be portrayed using geometrical properties of ensembles of sources. For example, a music ensemble can be described in terms of its width and depth. The term "ensemble" is defined in the paper as a group of music sound sources such as an orchestra, string quartet, choir, or pop music band. Finally, at a high level of abstraction, scenes could be characterized globally. For instance, they may be described using the volume of a recording venue (e.g., small, medium, large). The list of the provided examples is not exhaustive.

The aim of the existing techniques in the area of spatial analysis of binaural audio signals is predominantly to localize individual audio sources [5–12] or to estimate the perceived width of a single sound source [13]. Hence, referring to the terminology introduced above, most of the work within the area of spatial analysis of binaural signals so far has been limited to the scene analysis at the low level of abstraction, ignoring mid and high-level spatial attributes. In contrast, this study focuses on the computational analysis of spatial scenes at a mid-level using the attribute called "ensembles width" as described below.

In a scenario where the listener is situated at the center of a 360° stage, an ensemble of musicians may surround them. In this work, however, we considered a traditional scenario, constraining the location and spread of ensembles to a frontal semicircle in the horizontal plane. The rationale for this constraint is to make the task of ensemble width estimation easier, due to the potential for front-back errors inherent to the lack of simulations of head movements in this work. We leave for future research the exploration of more contemporary spatial scenes, with ensembles spanning 360° in the

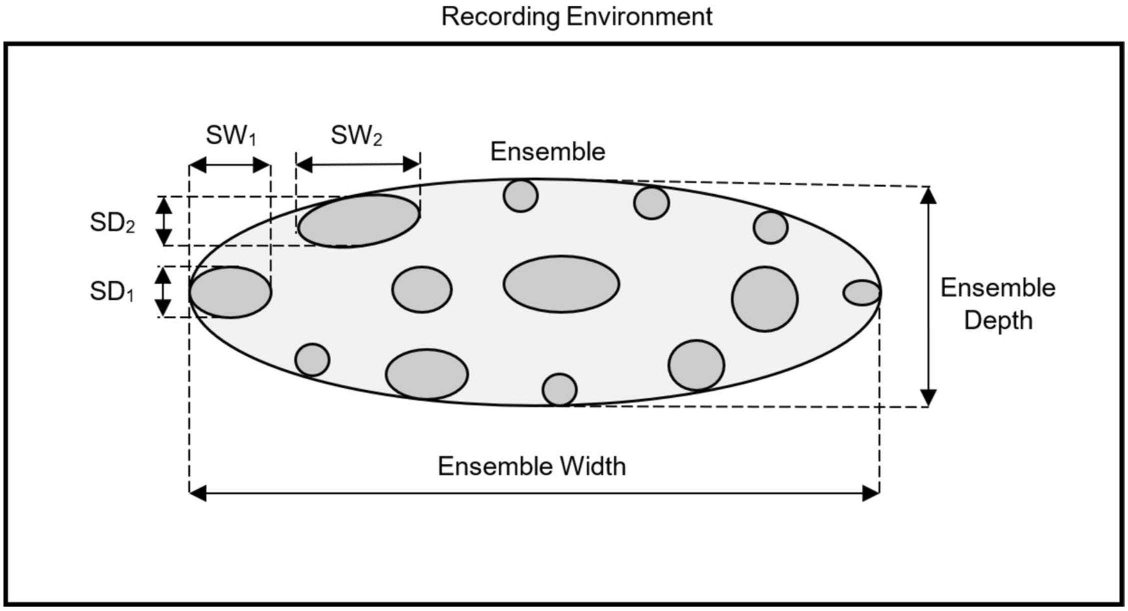


Fig. 1 Example of a spatial audio scene content with a hypothetical music ensemble. Dark-grey ellipses represent individual sound sources. SW_1 and SW_2 denote the width values of the two selected individual sources, whereas SD_1 and SD_2 signify their depths

horizontal plane, or even with ensembles positioned above or below the listener.

1.1 Definition of ensemble width

In this study, we define "ensemble width" (EW) as the physical extent between the leftmost and rightmost sound sources in an ensemble. Since we have adopted the polar coordinate system with a listener at the origin, the ensemble width is measured by an angle ω between the extreme sound sources of an ensemble in the horizontal plane, as shown in Fig. 2. This definition is similar to the one proposed by Rumsey [4], who described EW as the "overall width of a defined group of sources" (macro entity). It is also similar to the definition of "ensemble source width" that was recently introduced by Arthi and Sreenivas [14]. They described the ensemble source width as the "angular difference between the extreme sources". For simplicity, we assume that all individual sound sources are equidistant from the listener, and we treat them as point sources (we consider each music source to be infinitely small). In this work, 0° refers to the listeners' front-facing direction, and azimuth angles are measured counter-clockwise.

In contrast to the work of Arthi and Sreenivas [14, 15], who assumed the ensemble position to be an invariant factor, in our study we also investigated how the ensemble location affects the performance of the method. For this reason, we defined the location of an ensemble as an angle ϕ between the listener's front-facing direction and the direction of the center of an ensemble. The center of an ensemble is a notional mid-point on the arch between the sound sources positioned at the edges of an ensemble, splitting an ensemble into two halves of ½ ω each (see Fig. 2). For $\phi = 0°$, the center of an ensemble is located in front of the listener.

It should be noted that the concept of ensemble width, as estimated in this study, is distinctively different from the term "apparent source width" (ASW), which is commonly used in the context of concert hall acoustics. ASW is defined as the perceived horizontal extent of an individual sound source (performing entity) [16]. It is widely accepted that ASW is attributed to early room reflections [17], which cause the effect of broadening the perceived width of a sound source. Therefore, due to the above effect, for a given sound source placed in a reverberant environment, the ASW tends to be greater than the actual (physical) source width (SW), hence the term "apparent" in its name. ASW of a reproduced sound source, even under anechoic conditions, can also be artificially modified by decorrelating the signals reaching the listener's ears [18]. Unlike ASW, which describes a single source, EW is an attribute describing an ensemble (a macro entity representing a group of sources). Furthermore, whereas ASW is a perceptual parameter, EW is a physical attribute. For illustrative purposes, examples of SW of the two individual sound sources are provided in Fig. 1, where they are signified as SW1 and SW2, respectively. In comparison, EW represents the width of a set of audio sources, as depicted in Fig. 1. ASW is not shown in this figure as it constitutes an attribute in the perceptual

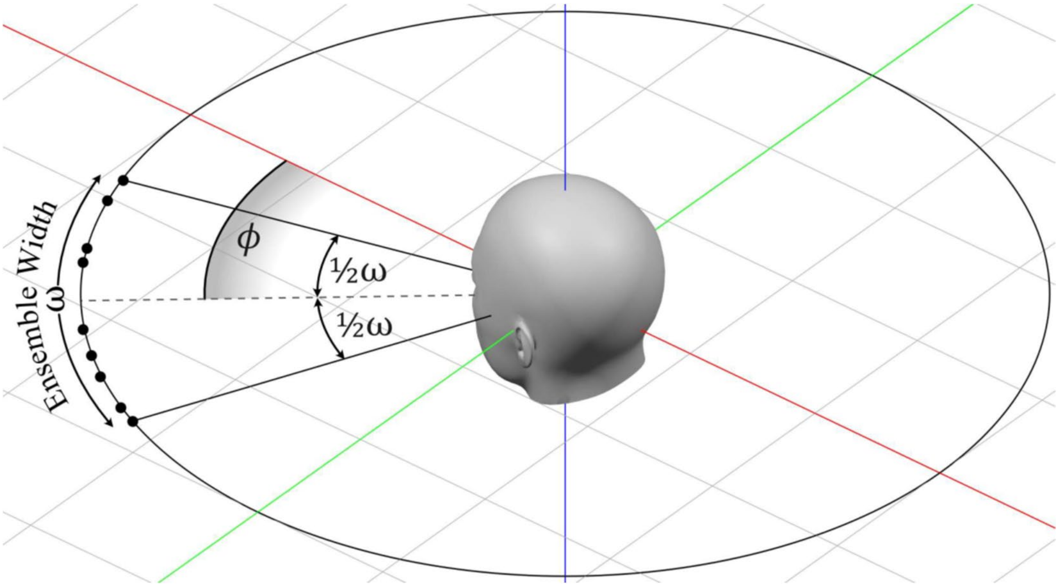


Fig. 2 An example of an ensemble consisting of nine point-like sound sources represented by dots. Ensemble width is signified by ω, whereas φ represents location of an ensemble center (counterclockwise)

domain. For clarity, see the definitions provided in Table 1.

To distinguish between ensemble width in the physical domain and that in the perceptual domain, we introduce the term "apparent ensemble width" (AEW), denoting an extent between the leftmost and rightmost sound sources in an ensemble as perceived by listeners (see Table 1). Given that under reverberant conditions, the perceived width of a single sound source tends to be greater than the physical width [16, 17], it may be postulated that room reflections may also lead to an increase AEW compared to EW. However, at this stage of our research, we limited the scope of the study to the estimation of the EW. The inclusion of reverberant conditions and the incorporation of listening test data representing AEW scores were left for future work.

1.2 Research objectives and overview of developments

The aim of this study was to develop a method for estimating EW in binaural music recordings based on the auditory model combined with the advanced regression algorithm. It was hypothesized that such an approach would lead to more accurate estimates than the deterministic technique using so-called "spatiograms" recently proposed in the literature [15, 19]. The complementary goal of the work was to find out which binaural cues, as extracted from the auditory model, play the most significant role in the estimation of EW.

The proposed technique was designed to estimate ensemble width in a "blind" way, making no assumptions about the number of music sources or their spectral characteristics. Moreover, the method was intended to operate regardless of music genre and irrespective of a head-related transfer function (HRTF) inherent to a binaural audio synthesis or recording process. Given the context of the research introduced above, the method was developed to characterize spatial audio scenes at the mid-level of abstraction.

For practical reasons, in this work, the binaural excerpts were obtained in a simulated physical (anechoic) environment. Namely, the method has been developed using a repository of 23,040 binaural excerpts synthesized by convolving 192 multi-track music recordings with a "diversified" set of HRTFs. For this purpose, 30 groups of publicly available HRTFs were employed, measured using both human and artificial heads, and utilizing different recording and post-processing techniques. The reason for employing a relatively large number of diverse HRTF sets was the above-mentioned goal of improving the robustness of the method to the changes in HRTFs since binaural localization methods are susceptible to variations in HRTF characteristics [20]. The synthesized binaural excerpts represented various spatial distributions of music audio sources along a frontal semicircle in the horizontal plane. Subsequently, a binaural auditory model was exploited to derive the standard binaural cues from the synthesized excerpts comprising interaural time differences (ITD), interaural level differences (ILD), and interaural cross-correlation (IACC) coefficients. Finally, a regression method based on gradient-boosted decision trees was employed to estimate music ensemble width values.

The developed method could be utilized for the search and retrieval of binaural recordings according to the width of ensembles. Additionally, the information acquired by the developed method may enhance the performance of music genre recognition techniques. Furthermore, it is anticipated that such a method, which employs a classical auditory model, will serve as a baseline method for benchmarking methods utilizing deep learning techniques or some simplified algorithms optimized for real-time applications.

1.3 Novelty and contributions of the study

In this work, we employed an auditory model originally developed by Søndergaard and Majdak [21], improved by May et al. [6], and later modified by Decorsière and May [22] within the Two Ears Project [23]. Additionally, we utilized a state-of-the-art machine learning technique, namely gradient-boosted decision tree regression based on LighGBM implementation introduced by Ke et al. [24]. The novelty of this study lies in the combination of these two "building blocks" and their application to the task of EW estimation. The contributions of this work are as follows:

Table 1 The selected attributes describe complex spatial audio scenes

Attributes	Domain	Definition
Source width (SW)	Physical	Physical width of a sound source
Apparent source width (ASW)	Perceptual	Width of a sound source as perceived by listeners [13, 16, 17]
Ensemble width (EW)	Physical	Physical extent between the leftmost and rightmost sound sources in an ensemble [4, 14]
Apparent ensemble width (AEW)	Perceptual	Extent between leftmost and rightmost sound sources in an ensemble as perceived by listeners

(1) We introduce a new method for the estimation of ensemble width in binaural music recordings employing an auditory model and a gradient-boosted decision trees regressor. The proposed method outperforms the technique utilizing "spatiograms" that were recently introduced in the literature [14, 15, 19]. Moreover, the method gives satisfactory results independently of music content and regardless of HRTF used to synthesize the binaural excerpts. In contrast to the binaural multi-source localization methods described in the literature, typically constrained to analyzing up to $2-5$ concurrent sound sources [5–12, 25–28] and/or requiring metadata about the number [7–9, 11, 27] or characteristics of the individual sound sources [5, 10, 12, 29], our method is capable of a "blind analysis" of binaural music recordings and is not restricted by the maximum number of simultaneously sound-emitting audio sources.

(2) We provide some insight into the importance of the selected binaural features derived from an auditory model. The results indicate that interaural level and time differences are the most significant factors, while interaural cross-correlation plays a secondary role.

(3) We discuss the importance of the selected factors affecting the performance of the method, which can guide further developments in this area. The results show that the performance of the model is relatively uniform regardless of the horizontal position of an ensemble but tends to deteriorate for wider ensembles. Furthermore, the performance of the method appears to improve for ensembles consisting of a large number of individual sound sources. However, in contrast to the binaural sound source localization techniques [30], the estimation error of the proposed method is independent of the spectral characteristics of the analyzed binaural excerpts.

The paper is organized as follows. The next section provides an overview of the related studies. Section 2 outlines the methodology employed in this study. The results and their discussion are presented in Sects. 3 and 4, respectively. The conclusions drawn from this work are given in the last section of the paper.

2 Related studies

This section gives an overview of state-of-the-art methods for spatial analysis of binaural signals. Moreover, it discusses the importance of the binaural cues in the context of this research. Finally, it describes existing methods for estimating ensemble width.

2.1 Methods for localizing sound sources using binaural signals

Most of the studies undertaken in the area of spatial analysis of binaural signals have been devoted to the localization of individual sound sources, predominantly speakers (as opposed to music sources) [5–12, 20, 25, 27, 29]. While the term "sound localization" is commonly used in the context of binaural spatial analysis, in fact, almost all the sound localization methods developed so far are limited to the estimation of sound source direction of arrival (DOA), ignoring the distance between the listener and the source. The recent work of Krause et al. [26] as well as Zohourian and Martin [31] constitute the exception in this respect, as the researchers developed methods for speaker distance estimation. Following the early work of Jeffress [32], who proposed a DOA model performing the correlation operation between binaural signals using a network of interconnected delay lines, the traditional DOA algorithms typically employed some form of correlation derivation, e.g., based on the inter-aural cross-correlation (IACC) function [11, 20, 29] or utilizing the generalized cross-correlation function with phase transform (GCC-PHAT) [27, 33, 34]. The latter approach proved to be more robust to reverberation. Interestingly, Jeffress' early idea of modeling human hearing using a neurophysiologically-inspired array of delay lines [32] has been recently exploited by Pan et al. [35] in their sound source localization method employing spiking neural networks.

Conceptually, the DOA binaural methods developed so far could be divided into two categories, resembling either a black-box or a glass-box approach. This distinction is made on the basis of a degree of "transparency" of the mechanisms employed and the understanding of the rules adopted by the machine learning algorithms to perform a given task [36]. In the former case, "raw" binaural signals are applied directly to the inputs of a deep learning model, yielding an estimated DOA value at its output [37], a scenario referred to as an end-to-end approach. In the latter case, dual-stage processors are typically used, employing an auditory model at the front-end, followed by a machine learning algorithm at its back-end. The role of the auditory model is to extract such binaural cues as interaural level differences (ILD), interaural time differences (ITD), and interaural cross-correlation (IACC) coefficients. Subsequently, machine learning algorithms are used at the back-end to estimate DOA. Examples of the employed machine learning techniques include Gaussian mixture models [5–9] and deep neural networks [10, 11, 29].

There is an increasing number of state-of-the-art DOA binaural methods that cannot be neatly classified using the above-mentioned categorization as they follow a hybrid approach. For example, Yang and Zheng [27] recently developed a method using latent features automatically derived by an autoencoder (a black-box approach) augmented by hand-engineered features including inter-aural cross-correlation values (a glass-box approach). Similar to the above approach, the other researchers also tend to use deep neural networks fed with a mixture of binaural audio signals represented by magnitude spectrograms [26, 27, 38–40], phase spectrograms [27, 38–40], as well as manually engineered features such as cepstral coefficients [20, 28], binaural cues [26, 28, 29], cross-correlation coefficients [33], or Mel-frequency spectral features [28]. To improve the performance of the DOA binaural methods, the researchers utilize algorithms mimicking human “attention mechanisms” [39], simulating head movements [9, 11], performing source separation algorithms [41], or taking advantage of the HRTF individualization [33]. Furthermore, there is a growing body of research demonstrating that the accuracy of the DOA methods could be further improved by dividing the horizontal plane into smaller zones and undertaking the task of sound localization for each zone separately [26–28]. Moreover, the recent improvements can also be attributed to better ways of model training, employing multi-conditional training techniques [11], or including ecologically valid training scenarios [30]. Despite the above-mentioned advancements, the state-of-the-art binaural analysis methods exhibit three major limitations, making them unsuitable for the task of music ensemble estimation constituting the focus of this study. They are outlined below.

(1) The state-of-the-art binaural analysis methods have been predominantly developed and optimized for the analysis of speech signals (not music) [20, 27, 39–41]. Consequently, they involve signal processing techniques specific to speech, e.g., utilizing voice activity detectors [6, 9] or deriving fundamental frequency from spoken voice utterances [5]. While a few recent studies extended their scope beyond the domain of speech signals by incorporating environmental sounds [26, 28, 33], they have not considered music audio signals.

(2) Most of the state-of-the-art binaural analysis methods are capable of localizing only a single source at a time [20, 29, 33, 37–41]. Some methods were developed for the simultaneous localization of speakers in multi-talker complex binaural scenes [5–12, 25–28]. However, their localization capabilities are limited to approximately three concurrent speakers, which is insufficient for analyzing music signals. Their localization performance significantly deteriorates when the number of active sound sources increases [27]. It is important to note that the number of concurrent music sources considered in this study ranged from 5 up to 62.

(3) The methods capable of localizing speakers in multi-talker scenarios rely on a priori knowledge about the number the concurrent sources to be localized and/or their characteristics [5, 7–12, 27, 29]. Such knowledge is normally unavailable in the case of the real-life repositories of binaural music recordings.

Considering the above limitations, the existing state-of-the-art binaural localization methods cannot be directly used as building blocks of higher-level scene analyzers of music recordings operating in a “blind” manner. However, they could be either adapted or used as inspiration for the development of such methods. In fact, the method proposed in this paper has been inspired by the localization algorithms exploiting a binaural auditory model as a feature extractor (a glass-box approach), as proposed by such researchers as Dietz et al. [5], May et al. [6, 7, 9], Woodruff and Wang [8], Ma et al. [11, 29], as well as Ma and Brown [10]. The reason for adopting a glass-box approach is that such methods are particularly useful in explaining the importance of employed binaural cues. Since the topic of ensemble width estimation is relatively new, an exploratory approach taken in this study is deemed to be justified. Moreover, the results obtained using such an approach could serve in the future as the baseline for developing more elaborate deep learning-based techniques.

2.2 Importance of binaural cues in DOA algorithms

It is widely accepted that the ILD and ITD cues provide sufficient information to localize sound sources in the horizontal plane [42]. Therefore, they were selected as predictors of EW in this study. However, this does not mean that the interaural cross-correlation (IACC) is unimportant for the localization of sources in the horizontal plane. In real-life applications, under adverse acoustical conditions, ILD and ITD cues are often corrupted by room reflections and background noise. Therefore, to make the DOA algorithms more robust, ILD and ITD cues could be augmented by IACC features, providing auxiliary information about the degree of reliability of ILD and ITD cues. Under adverse acoustical conditions, ILD and ITD cues reliably describe DOA only for those frequency bands and time-frames where IACC coefficients approach unity [43]. To further improve the robustness of the DOA models, onset detection is exploited to select “undistorted” binaural cues based on the precedence effect [5, 8]. In light of the aforementioned considerations, for completeness, in addition to the ILD and ITD cues, the IACC feature was also incorporated into the EW estimation model. However, for the sake of simplicity, onset detection was not included in the current implementation of the model.

While for those DOA algorithms that are limited to the horizontal plane, ILD and ITD constitute the fundamental descriptors, whereas IACC cues provide only auxiliary information enhancing the robustness of the methods [43], these authors hypothesized that the opposite would be true for an ensemble width estimator developed in this work. It was assumed by these authors that the wider the ensemble, the more uncorrelated the signals reaching the listener's ears, potentially resulting in a change in the IACC values. Moreover, the task of EW assessment resembles the procedure of ASW estimation, widely researched in the literature. There is strong evidence that ASW is primarily associated with interaural cross-correlation (IACC) coefficient, and to a lesser extent with ILD, and ITD cues [13, 16–18]. Therefore, we postulated that a regression model estimating EW would also be predominantly "guided" by the IACC cues, and to a lesser extent by the ILD and ITD features. In retrospect, while IACC did not turn out to be the leading cue, it demonstrated a notable prominence, as evidenced by the obtained results (see Sect. 2).

Baumgartner et al. [44] as well as Barumerli et al. [45] developed the DOA models that are capable of localizing a single sound source not only along the horizontal plane but also within the sagittal plane. In addition to the ILD and ITD cues, these models rely heavily on spectral features, as they are deemed instrumental for the localization in the sagittal plane [46]. Since in our work we limited the estimation of EW to the horizontal plane, we considered spectral cues as unimportant and therefore they were not included in the feature extraction algorithm. Our supposition that the model would be insensitive to spectral features was validated through empirical evidence (see Sect. 4.3.5).

2.3 Methods for estimating ensemble width

The method for EW estimation proposed in this paper is not the only one in the literature. In their recent work, Arthi and Sreenivas [14, 15] proposed a technique for estimating "ensemble source width" based on the phase-only spatial correlation (POSC) function. They also introduced the notion of "spatiograms" being two-dimensional images representing POSC function values evolving in time. While they provided proof of the concept illustrating how POSC functions and spatiograms could be used for the estimation of EW for simple binaural recordings, their method has not been quantitatively validated using real-life binaural recordings. More importantly, in their work, Arthi and Sreenivas assumed the number of sound sources to be a known parameter [14]. This requirement prevents their method from being used in a "blind" way and restricts the scope of its practical applications. Recently, this restriction was removed in a modified version of their method proposed by Antoniuk and Zieliński [19], who demonstrated that the spatiograms-based method could be used as a blind estimator of music ensembles in binaural recordings under simulated anechoic conditions. They also showed that the method employing spatiograms yields satisfactory results both under the HRTF-dependent and HRTF-independent conditions, indicating its promising generalization property. Therefore, the last-mentioned method will serve as a baseline algorithm in this study.

3 Methodology

The flowchart of the proposed method is presented in Fig. 3. The model comprises an auditory component in conjunction with a regressor. The function of the auditory model is to extract the binaural cues from the binaural input signals (left and right), whereas the objective of the regressor is to estimate the ensemble width. As previously stated in the Introduction, we employed an auditory model developed by Søndergaard and Majdak [21], enhanced by May et al. [6], and refined by Decorsière and May [22], while a gradient-boosted decision trees algorithm based on the LightGBM implementation proposed by Ke et al. [24] was utilized as the regressor. It is important to reiterate that neither the auditory model nor the gradient-boosted decision tree algorithm were developed by these authors. The novelty of this work lies in the

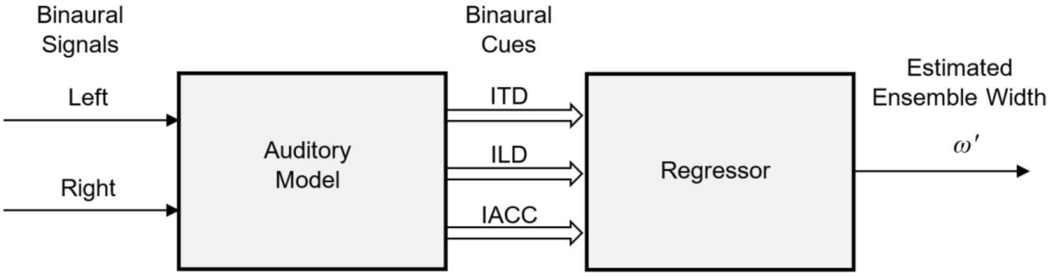


Fig. 3 A flowchart of the algorithm implemented in the study

combination of a well-recognized auditory model with a state-of-the-art regression technique for the purpose of estimating EW. The initial part of this section delineates the methodology employed for the generation of the binaural excerpts utilized in the development of the method. This is followed by a description of the auditory model and the regressor.

3.1 Synthesis of binaural recordings of music

The binaural recordings used in this study were synthesized under simulated anechoic conditions, using the procedure of the convolution of the monophonic sound signals with head-related transfer functions (HRTF) acquired in the anechoic chambers. To this end, 192 multi-track studio recordings were acquired from the publicly available repository [47]. They exemplified a broad range of music genres such as classical music, opera, pop music, jazz, country, electronica, dance, rock, and heavy metal. The monophonic signals from each track represented individual sound sources. The number of tracks varied across the recordings (min. 5, max. 62, median 9). The monophonic signals from each track were loudness equalized to −23 LUFS, according to ITU-R BS.1770 recommendation [48].

Similar to the method employed in our previous work [19], the location of each individual sound source i is defined using azimuth θ_i, within boundaries determined by ensemble location ϕ and width ω (see Fig. 2), so that

$$\theta_{\mathrm{i}} \in \left[\phi - \frac{\omega}{2}, \phi + \frac{\omega}{2}\right], \tag{1}$$

where $i \in [1, N]$; $\phi \in [-45^\circ, 45^\circ]$; and $\omega \in [0^\circ, 90^\circ]$; while N represents the number of tracks for a given music recording. Due to the above constraints, the ensembles were always located in the frontal semicircle. The ensembles were generated only in the horizontal plane (elevation equal to 0°).

In order to synthesize the binaural excerpts, the loudness-equalized monaural signals were convolved with the head-related transfer functions (HRTF), according to the following equation:

$$y_c[n] = \sum_{i=1}^{N} \sum_{k=0}^{K-1} h_{c,\theta_i}[n-k] \times x_i[k], \tag{2}$$

where $y_c[n]$ represents an output binaural signal for an audio channel c (left or right) for a given music recording and sample n; $x_i[k]$ denotes a kth sample of an ith monophonic signal (individual music sound source); $h_{c,\theta_{\mathrm{i}}}[n]$ is a head-related impulse response (HRIR) for an audio channel c and azimuth θ_i for an ith monophonic signal; and N is the number of tracks for a given music recording. The duration of the binaural excerpts was limited to 7 s. Therefore, the upper summation limit K in Eq. (2) was equal to 336×10^3 (sample rate × duration), as it represented the total number of samples within each binaural excerpt, given the sample rate of 48 kHz.

In this work, 30 publicly available HRTF sets were utilized in the procedure of the synthesis of the binaural excerpts. The employment of a relatively large number of diverse HRTF sets served to enhance the robustness of the method in response to changes in HRTFs. Furthermore, it enabled the rigorous testing of the generalization property of the developed method. The selected HRTFs were measured in various anechoic chambers using diverse measurement procedures. The selection comprised 15 human HRTFs and 15 artificial ones. The artificial heads included Knowles Electronics Manikin for Acoustic Research (KEMAR) head and torso simulator (GRAS 45BA, 45BB-4, DB-4004) [49–56], as well as Neumann (KU 100) [57–59], FABIAN [60], SAMRAI (Koken) [61], and ARI Printed Head [62]. The employed HRTFs were measured both in the near and far field, with the measurement radius ranging from 90 cm to 1.95 m. The average horizontal resolution of HRTFs spanned from 0.3° to 15°, with a median value of 2.5°. A detailed description of the selected HRTFs is provided in 3.1: Table 4.

To address the issue of the restricted spatial resolution of the selected HRTF sets, HRTFs were interpolated during the synthesis process. To this end, we employed the bilinear interpolation algorithm, as proposed by Freeland et al. [63], implemented in the “interpolateHRTF” function of the MATLAB Audio Toolbox [64]. This algorithm allows for the interpolation of HRTFs in two dimensions. In retrospect, given that the ensembles were arranged in the horizontal plane, a linear interpolation algorithm could have been sufficient for the majority of the employed HRTF sets, with the exception of HRTF No. 1 (see Table 4), which lacked measurements at an elevation of 0°.

For each multi-track recording and every HRTF, four binaural excerpts were synthesized, representing ensembles whose locations ϕ and widths ω were drawn from a continuous random distribution, constrained by Eq. (1). For each ensemble, the locations of individual sound sources θ_i were also randomly generated. This way, 23,040 binaural excerpts were synthesized (192 multi-track recordings × 30 HRTFs × 4 ensembles). Recall, that according to the constraints given by Eq. (1), all the synthesized music ensembles were located within the frontal arc in the horizontal plane.

The convolved audio signals were trimmed to 7-s excerpts and had fade-in and fade-out effects applied using a 0.01-s sine-square window. Finally, root mean square (RMS) normalization, amplitude scaling by 0.9, and DC equalization were applied to the resulting audio excerpts. They were stored in 23,040 uncompressed audio files at 48 kHz sample rate and with 32-bit resolution. The procedure of the synthesis of the binaural excerpts described in this paper was undertaken using MATLAB. For reproducibility, the code written for this purpose was made publicly available at GitHub [65]. The excerpts were randomly auditioned by these authors to determine the correctness of their spatial characteristics, including perceived ensemble width, degree of externalization, and level of sound quality. More information on the informal listening impressions is provided in Sect. 4.

3.2 Feature extraction using an auditory model

Drawing inspiration from the DOA estimation methods mimicking human hearing mechanism in the initial part of their signal processing chain (a glass-box approach [5–11, 29]), an auditory model was employed in this study as a front-end. It consisted of a standard gammatone filter bank [66] implemented by Søndergaard and Majdak [21]. In line with our previous work regarding the analysis of binaural signals [67], its low cut-off frequency was set to 100 Hz. In the DOA estimation methods intended for speech applications, the number of gammatone filters is typically set to 32 [6, 7, 9, 11, 29, 37, 41]. However, in our work, due to the processing of audio recordings exhibiting broader spectra than those typically observed in speech signals, the number of gammatone filters was increased from 32 to 64. For the same reason, we extended the upper-frequency limit of the gammatone filter bank from the typical values of 5–8 kHz [7, 8, 11] to 16 kHz. We verified that extending the high-frequency limit of the gammatone filter bank from 8 to 16 kHz reduced the mean absolute error of the proposed method by 0.1°. While this improvement could be considered very small, it was statistically significant at a p-level of 4.84×10^{-3}. It is important to note that the decision to extend the upper frequency limit from 5–8 to 16 kHz could be open to challenge on the grounds that human hearing is considered to be insensitive to binaural cues above 8 kHz [68] (see Discussion). To simulate the loss of phase-locking in the auditory nerve at higher frequencies, the inner hair-cell envelopes of the band-pass filtered signals were then extracted by applying half-wave rectification followed by low-pass filtering using a second-order Butterworth filter with a cut-off frequency of 1 kHz [69]. Then, the “rate maps” were calculated from these envelopes. The rate maps constitute a graphical representation of auditory nerve firing rates [70]. Finally, the rate maps were used to estimate the standard binaural features, comprising ILD, ITD, and IACC coefficients. In this study, the IACC coefficients were calculated as the maximum values of the normalized interaural cross-correlation function, restricted to the time lag of $\tau = \pm 1$ ms, as originally proposed by Sato and Ando [16]. The algorithm employed to calculate the ITD cues was based on the technique proposed by May et al. [6], with the refined precision due to the parabolic interpolation method used.

While some authors argue that the human-hearing system undertakes a continuous analysis of binaural signals, without dividing the signals into smaller time-frames [30], in this work a conventional approach was taken whereby the feature extraction procedure was undertaken using a sliding time window. In line with the studies described in [7, 8, 10, 11, 29, 41], the window length was set to 20 ms with a 10-ms overlap. The binaural features (ILD, ITD, IACC) were extracted separately for each time window and then aggregated by calculating their mean and standard deviation statistics. This resulted in the extraction of 384 feature vectors (64 frequency channels × 3 types of cues × 2 statistics). The feature extraction procedure was undertaken using the MATLAB implementation of the Two Ears project’s auditory model [23]. For reproducibility, the code employed to extract the features has been made publicly available on GitHub [65].

3.3 Gradient-boosted decision tree regression algorithm

There are several suitable candidate methods for use as regression algorithms in this work. A classical linear regression algorithm was ruled out due to a strong multicollinearity effect in the data obtained from the auditory model. The overlapping characteristics of the filters in the gammatone filter bank used in the auditory model result in highly correlated binaural cues. Inspection of the data revealed that out of 384 feature vectors extracted from the auditory model, 105 vectors exhibited strong mutual association with a variance inflation factor (VIF) exceeding 100. Other regression techniques were considered, including ridge regression and lasso technique, which are said to be resilient to the multicollinearity effect [71]. However, a gradient-boosted decision tree algorithm was selected for this work due to its resilience to the multicollinearity problem, accuracy, and computational efficiency. Gradient-boosted decision tree algorithms are regarded as highly efficient, accurate, and interpretable machine learning methods [72]. They exhibit outstanding performance in regression, classification, and ranking tasks [73]. Moreover, they are suitable for handling relatively big datasets [74], with some evidence of providing even more reliable results than logistic regression [75].

Several popular implementations of gradient-boosted decision tree algorithms are currently available to researchers, including a highly acclaimed XGBoost technique [76], pGBRT [77], scikit-learn [78], gbm in R [79], and LightGBM [24]. In our work, there were many experimental repetitions, requiring a fast and accurate regression technique. Therefore, we decided to use a LightGBM implementation of a gradient-boosted decision tree algorithm as it is renowned for its computational efficiency and accuracy [24]. Moreover, it has two distinct features that differentiate it from the other implementations. Namely, instead of growing trees depth-wise (Fig. 4a), it grows trees leaf-wise (Fig. 4b). Furthermore, it employs a histogram-based algorithm to find approximate split points, exhibiting faster performance compared to the standard techniques [80].

In our study, we performed the regression computations using a graphical processing unit (NVIDIA RTX 2080). We employed the LightGBM software library developed by Zhang et al. [80]. The next section provides the details regarding the hyperparameter tuning, the training procedure, and the evaluation of the employed regression method.

3.4 Model training and evaluation

The aim of the training and evaluation procedures was to optimize the regression model and assess its performance. The procedure of model training and evaluation was undertaken seven times, with different data splits (see below), in order to evaluate the degree of the repeatability of the results. The selected number of experimental runs represented a reasonable trade-off between the computational load of repeating the experiments and the statistical significance of the obtained results. To characterize the generalizability property of the method, the training and evaluation tasks were undertaken using the data sets with unique music recordings.

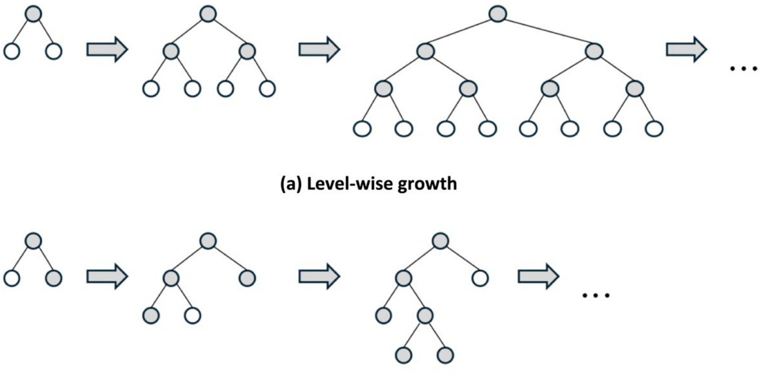
The training and evaluation procedures consisted of the following steps. Namely, 192 music recordings were randomly split into two sets in proportion 128/64 (corresponding to a 2:1 split) seven times, which formed seven pairs of development and test sets. The development sets were then used to fine-tune the hyperparameters during the development procedure, while the test sets were employed to assess the models' performance. For clarity, the data splits between the development and test sets are outlined in Table 2.

The features (binaural cues) extracted by the auditory model were fed to the input of the gradient-boosted decision tree algorithm. They were treated as independent variables in the regression model. The feature vectors were paired with the associated ensemble width values (ranging from 0° to 90°) which were earlier applied in the process of the generation of the binaural excerpts. These ensemble width values were regarded as ground-truth data (the dependent variable in the regression model). All the data were represented as floating-point numbers.

To reduce the risk of overfitting the data, the tree pruning technique was used. For this purpose, parameters such as the maximum number of leaves in a tree and the maximum depth for a tree were adjusted in the standard tenfold cross-validation procedure applied to the validation set (the exact values are given below). During the cross-validation process, the development data were further divided into the training and validation data sets. Note that for each fold, the training and validation sets

Table 2 Development and test data splits in the repository of binaural audio excerpts

Dataset type	No. of music recordings	No. of HRTF sets	No. of ensemble variants (widths and rotations)	No. of binaural excerpts
Development	128	30	4	15,360
Testing	64	30	4	7,680



(a) Level-wise growth

(b) Leaf-wise growth

Fig. 4 Comparison of decision tree growth techniques: **a** level-wise growth, **b** leaf-wise growth [24]. White circles represent leaves, whereas grey circles signify nodes

were mutually exclusive with respect to musical recordings (the trained models were validated on "unseen" recordings). Such diversification of data during the training procedure reduces the risk of "information leakage" between the training and validation sets. Finally, once the model was trained, its performance was evaluated on the test data, which was not used in the training procedure. While the training, validation, and test sets were unique in terms of the music recordings, they still shared some information since the same HRTF sets were used to synthesize the binaural excerpts, potentially leading to an overfitting effect. To investigate this effect in more detail, a separate experiment was conducted in which all the data sets were unique in terms of both the music recordings and the HRTF sets. The outcome of this experiment proved that the method is generalizable with respect to music recordings and HRTF (Sect. 4.4).

During the fine-tuning procedure, the hyperparameters of the gradient-boosted decision tree model were optimized using a grid search algorithm on the development set. The search space was defined as follows: the maximum number of leaves in one tree $n \in [500, 1000, 1500]$, the maximum depth for a tree model $d \in [3, 6, 9]$, and the learning rate $l \in [0.001, 0.01, 0.2]$. This search space was chosen after many empirical trials to ensure that the sought hyperparameter values of the best model do not exceed the search space boundary. The remaining hyperparameters of the selected regression model were fixed. Namely, the maximum number of bins that the feature values were bucketed to was equal to 31, and the number of estimators was set to 500 (see [81] for the description of the hyperparameters of the gradient-boosted decision tree regression method).

In order to evaluate the performance of the model, the mean absolute error (MAE) was selected as the primary metric. The MAE is a widely used metric for appraising the performance of DOA algorithms [12, 27, 33, 35, 40]. The calculations were performed as follows:

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}\left|\omega_i - \omega_i^{'}\right|, \tag{3}$$

where ω_i represents the actual ensemble width, $\omega_i^{'}$ denotes the predicted ensemble width, and M is the number of observations. Moreover, the strength of the association between the predicted and actual EW values was evaluated using Pearson's correlation coefficient, according to the following formula:

$$r = \frac{\sum_{i=1}^{M}(\omega_i - \overline{\omega})(\omega_i' - \overline{\omega'})}{\sqrt{\sum_{i=1}^{M}(\omega_i - \overline{\omega})^2 \sum_{i=1}^{M}\left(\omega_i' - \overline{\omega'}\right)^2}}, \tag{4}$$

where $\overline{\omega}$ and $\overline{\omega'}$ denote the mean values for the actual and predicted ensemble width, respectively. Furthermore, in some analyses we also assessed the goodness of fit of the developed model using the coefficient of determination, defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{M}\left(\omega_i - \omega_i^{'}\right)^2}{\sum_{i=1}^{M}(\omega_i - \overline{\omega})^2}. \tag{5}$$

To estimate the degree of potential bias of the method (an offset between the predicted and the actual EW values), we calculated the mean signed difference (MSD) score using the following equation:

$$\mathrm{MSD} = \frac{1}{M}\sum_{i=1}^{M}\omega_i - \omega_i^{'}. \tag{6}$$

We must reiterate that the experiments were run seven times, with different data splits in each repetition (maintaining the same proportions between the development and testing sets as given in Table 2), giving rise to slightly different results in each experimental run. Therefore, the outcomes of the experiments obtained for all the above-mentioned metrics were summarized using their mean values and standard deviations calculated across all seven experimental runs.

For the purpose of hyperparameter tuning, a single metric, namely MAE, was utilized. For each hyperparameter combination, $k = 10$ evaluations were performed using a standard k-fold cross-validation technique. The tuning procedure described above consisted of $(3 \times 3 \times 3) \times 10$ iterations, yielding 270 iterations in total. It was used to find the best combination of hyperparameters for each of the seven experimental repetitions. The best values of the searched hyperparameters, as identified using this procedure, are presented in 3.4: Table 5.

Finally, for each of the seven experimental repetitions, the best models were tested on the test sets. Recall that the test sets were unique in terms of music recordings compared to the development sets employed for the training and optimization procedures. Thanks to the experimental repetitions, it was possible to measure the repeatability degree of the experiment, by estimating the mean values and standard deviation of the selected metrics, as explained above. The results of the experiments are described in the next section.

4 Results

At the beginning of this section, the overall performance of the method is characterized. Then, the follow-up analyses are presented, revealing the importance of the binaural cues in the task of ensemble width estimation, and showing the influence of the selected experimental factors on the performance of the model. Finally, the results of the generalizability tests are provided at the end of the section. Unless otherwise stated, the presented results are averaged across randomly selected ensemble widths constrained by $\omega \in [0°, 90°]$, and randomly chosen ensemble locations within the limits of $\phi \in [-45°, 45°]$ (see Sect. 3.1). Moreover, the results were averaged across seven experimental runs with different splits between the development and test sets as described above in detail in Sect. 3.4.

4.1 Overall model performance

According to the obtained prediction results using the test set, the MAE of the developed regression models is equal to 6.63° (SD 0.12°). The correlation coefficient between the actual and predicted ensemble width values equals 0.94 (SD 0.003), whereas the coefficient of the determination R^2 of the regression model amounts to 0.88 (SD 0.01). Hence, the obtained averaged results can be regarded as satisfactory. Moreover, the observed relatively small values of standard deviations indicate a high degree of experimental repeatability.

It is worth reiterating that the experiment was conducted seven times, yielding seven models that exhibited slight discrepancies in their respective results. The results obtained using the best-performing model are illustrated in Fig. 5. The figure shows a scatter graph of the actual and predicted ensemble width values. Note that the prediction results presented on the scatterplot were obtained using the test data, not the train data. It can be seen that the predicted ensemble width values match the actual ones well, as most of the data are scattered along the diagonal reference line ($y = x$). Three additional observations can be made from Fig. 5. First, the model performance is the best for the narrow ensembles, as the scatter of the data tends to diminish closer to the coordinates' origin. Second, the residuals are cone-shaped, which implies that heteroskedasticity occurs. Third, the model tends to underestimate the predicted values for wide ensembles ($\omega > 70°$).

To explore the above-mentioned heteroskedasticity effect in more detail, the relationship between the actual ensemble width and the model performance was analyzed using a graph presented in Fig. 6. The plot shows the MAE values as a function of the actual ensemble width ω. The results were computed using the data taken from the scatter graph discussed above (Fig. 5), by applying a moving window technique (window length was equal to 0.5°). It can be seen that the model performs relatively well for narrow ensembles ($\omega < 40°$), yielding MAE being less than 8°. However, for very wide ensembles ($\omega > 80°$), the performance of the model considerably deteriorates. Beyond this threshold, the MAE rapidly increases, reaching up to 14.30° at $\omega = 90°$. It can be concluded that, to a first approximation, the wider the ensemble, the worse the performance of the model. However, it can be seen

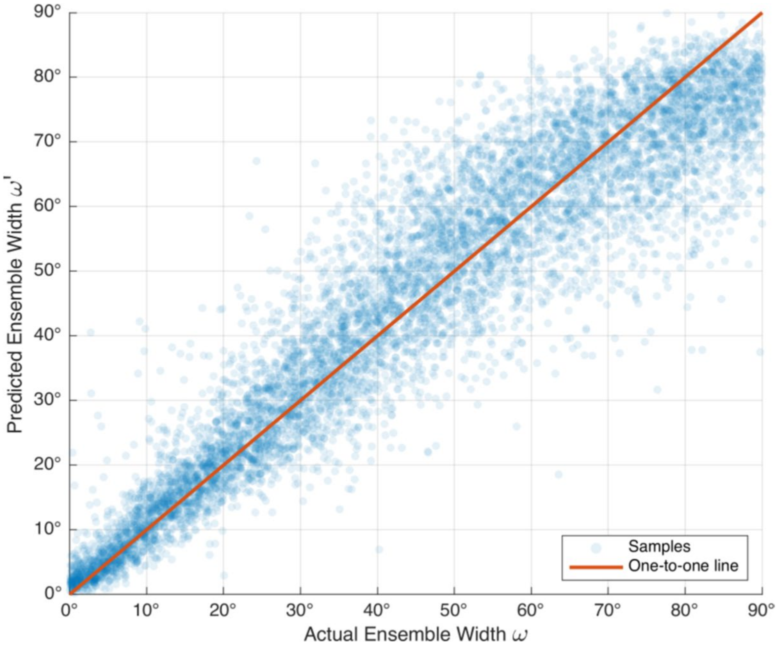


Fig. 5 A scatterplot of the actual vs. predicted ensemble width values obtained using the best-performing model

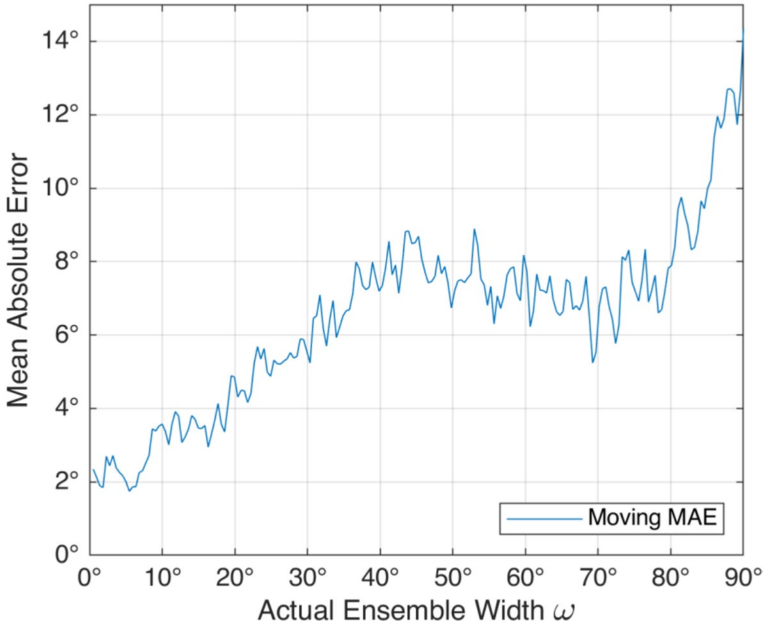


Fig. 6 Relationship between the mean absolute error and the actual ensemble width

that the relationship between MAE and the ensemble width is not monotonic, as a valley occurs for the ensemble width values between 60° and 80°. The exact reason for this phenomenon is at present difficult to identify and would require further experimentation. A possible reason could be linked to the characteristics of the binaural cues. The developed model relies heavily on interaural level differences (see Sect. 4.3.2), which also exhibit a non-monotonic relationship with respect to DOA, with a maximum value reported for approximately 45° [5].

Figure 7 depicts the mean signed difference (MSD) between the actual and predicted EW values as a function of the actual ensemble width. It is evident that for the narrow ensembles ($\omega < 30°$), the model overestimates the EW values by approximately 2°. This effect is even more pronounced for the mid-sized ensembles ($40° < \omega < 50°$),

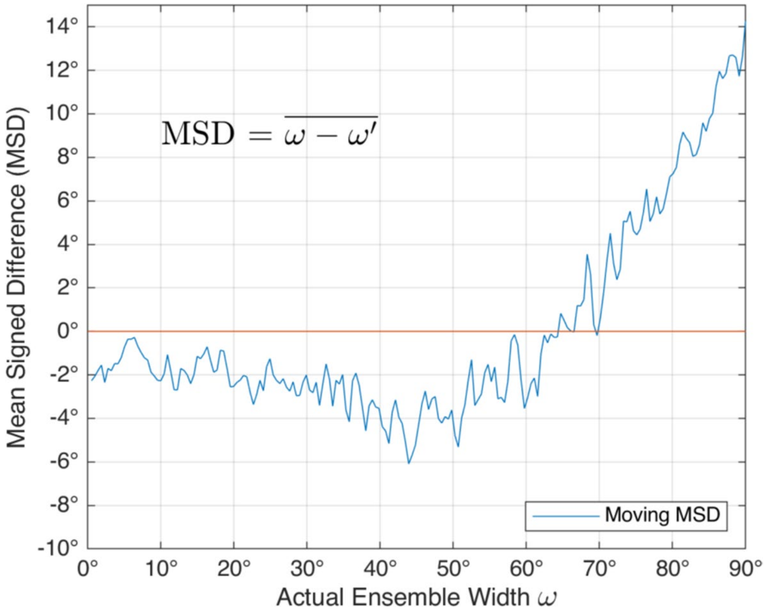


Fig. 7 Relationship between the mean signed difference and the actual ensemble width

with an overestimation level of approximately 6°. Conversely, for wide ensembles ($\omega > 70°$), the model exhibits an underestimation of the results. For an actual ensemble width equal to the maximum value of 90°, the degree of underestimation reaches as much as 14°.

To investigate the potential cause of the overestimation effect, we analyzed how the values of the binaural cues change with respect to EW (the graphs are omitted in the paper due to space constraints). Our findings indicate that most of the cues exhibit pronounced changes for narrow ensembles. However, for wide ensembles, the rate of change decreases, suggesting that they provide the regressor with reduced information about the EW values. Consequently, the predictions of EW are prone to errors.

4.2 Comparison with the state-of-the-art method

The performance of the regression method described in this paper was compared to that of the deterministic algorithm employing spatiograms, recently proposed by Arthi and Sreenivas [14, 15]. To this end, a publicly available implementation of the improved version of the spatiogram-based method was used [19] (see Sect. 2.3 for the details regarding the introduced improvements). For consistency of the comparison, it was trained and tested according to the methodology described in this paper, using the development and test sets employed in this study, with frontally-located ensembles in the horizontal plane (Sect. 3). According to the obtained results, the spatiogram-based technique yielded an MAE amounting to 16.46° (SD 0.69°), compared to an error of 6.63° (SD 0.12°) exhibited by the method introduced in this paper. Hence, the technique proposed in this study proved to be 7.83° better in terms of an MAE. The improvement was statistically significant at $p = 9 \times 10^{-9}$ level.

While the above comparison outcome clearly proves the superiority of the method developed in this study relative to the spatiogram-based technique, it is difficult to compare fairly the two methods in terms of the computational load as neither was optimized in this respect. They also employed external software libraries which have not been computationally optimized. The task of the computational optimization of the two methods was left for future work.

4.3 Influence of the selected experimental factors on the model performance

4.3.1 Model sensitivity to ensemble location

In our study, we also investigated how the ensemble location affects the performance of the method. Recall, that the ensemble location is defined as an angle ϕ between the listener's front-facing direction and the direction of the center of an ensemble (see Fig. 2). According to the obtained results, shown in Fig. 8, the performance of the regression model is relatively uniform across the ensemble locations. However, it exemplifies substantial variations in the prediction error caused by the changes in ensemble width.

The model exhibits the local maxima in the prediction error with the average error ranging up to 15.7° and the maximum error reaching 46.36°, for the ensembles being approximately 30°–60° wide, located centrally or slightly off-center ($|\phi| < 30°$). This effect is represented by the

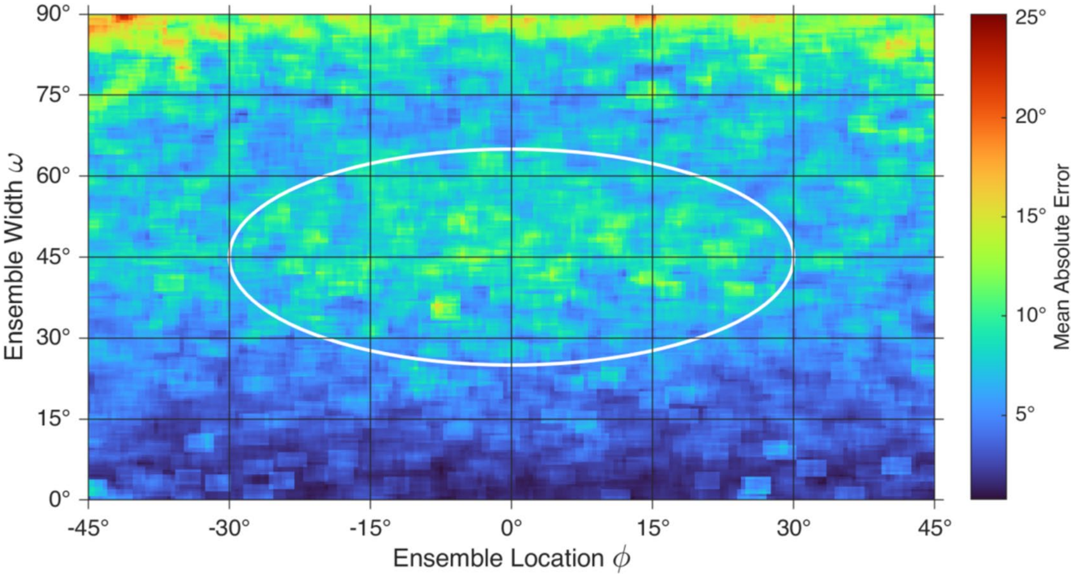


Fig. 8 Influence of ensemble location and ensemble width on the mean absolute error. The ellipse-shaped figure indicates the region of the local maxima in the prediction error

ellipse-shaped "green island" in the middle of the graph in Fig. 8. The exact reason for this phenomenon is currently unknown and requires further scrutiny. Surprisingly, the performance of the method tends to improve for the ensembles located sideways, close to the boundaries of the investigated range of locations ($\phi = \pm 45°$). For narrow ensembles ($\omega < 15°$), the performance of the method is relatively consistent, yielding a small prediction error with MAE being less than 9.2°. On the other hand, for very wide ensembles ($\omega > 80°$), the prediction error is high, extending up to 25.2°. This substantial increase in the error is caused by the underestimation effect illustrated earlier in Fig. 5. Overall, the method proves to be relatively robust to the changes in ensemble location.

4.3.2 Importance of the binaural cues

Figure 9 illustrates the performance of the 21 regression models. Each model was trained using either an individual type of binaural cues (e.g., IACC) or their combinations (e.g., ITD + IACC). Moreover, the models were trained employing only the mean values of the binaural cues calculated across the time-frames (signified by blue bars in Fig. 9), exploiting solely their standard deviations (red bars), or utilizing the combination of the above-mentioned statistics (green bars). The vertical axis in the figure represents the average values of the prediction MAE with associated standard deviation values, estimated using seven experimental runs.

Considering the results obtained for the models trained using the mean values of the features (indicated in blue color in Fig. 9), the model based solely on interaural cross-correlation (IACC) coefficients exhibited relatively good performance, yielding MAE equal to 8.57°. By contrast, the models exploiting either interaural time differences (ITD) or interaural level differences (ILD) showed markedly worse performance, with the MAE values being equal to 12.27° and 11.95°, respectively. Similarly, the model employing the combination of interaural level and time differences (ILD + ITD) performed rather poorly, giving an MAE of 10.65°. However, when the interaural level or time difference cues were used in combination with the interaural cross-correlation coefficients (ITD + IACC, ILD + IACC, or ILD + ITD + IACC), the performance of the models markedly improved, with the MAE values ranging from 7.50° to 8.21°. The above outcomes suggest that the mean values of the interaural time and level differences, used in isolation from interaural cross-correlation, are unsuitable for the task of ensemble width prediction. Surprisingly, the models based solely on the standard deviations of the binaural cues show consistently acceptable performance regardless of whether the cues are used in isolation (e.g., IACC) or in combination with each other (e.g., ITD + IACC), with an MAE varying from approximately 7° to 8° (see red bars in Fig. 9). Moreover, the results could be further improved if the models are trained using the combination of both statistics: mean values and standard deviations (indicated in green in Fig. 9).

Overall, the best prediction result was obtained for the combination of all the binaural cues (ILD + ITD + IACC) using both their mean values and standard deviations, with an MAE of 6.63°. Interestingly, the prediction results obtained using a simpler model, utilizing only interaural level and time differences (ILD + ITD), were only

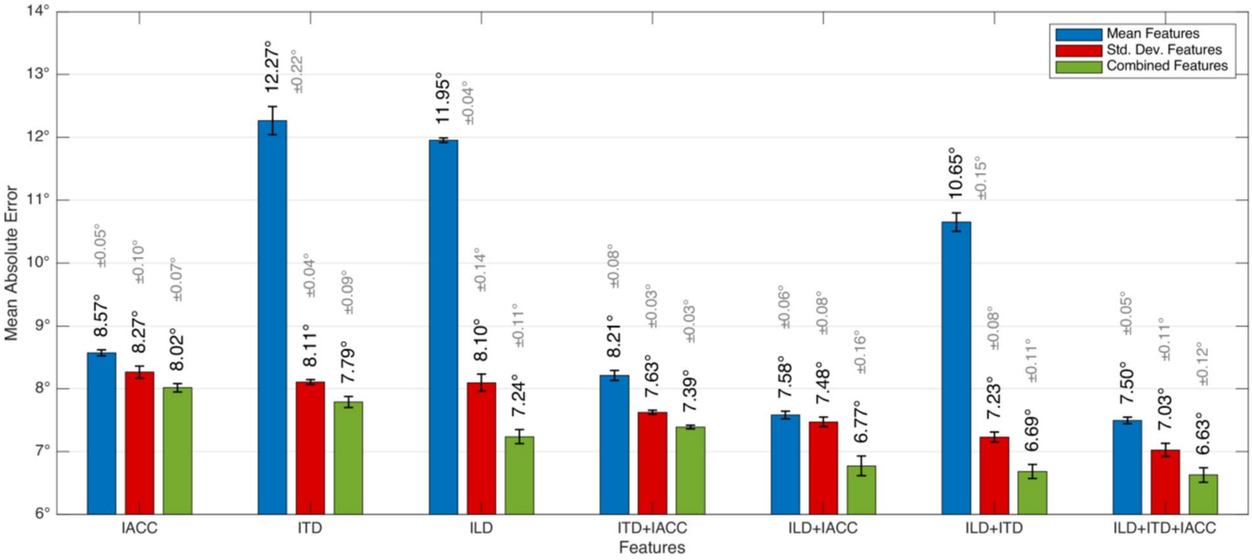


Fig. 9 Influence of binaural cues on the model performance. Error bars denote standard deviations

marginally worse, with an average MAE being equal to 6.69° (a difference of approximately 0.06°). Importantly, this difference was statistically not significant ($p = 0.09$). Nevertheless, due to its best outcome seen in Fig. 9, the most complex model, utilizing all three groups of the binaural features (ILD + ITD + IACC) and both types of statistics, has been used in the experiments reported throughout this paper.

4.3.3 Influence of the ensemble width on binaural cues

The surprisingly good performance of the models based on the standard deviations of the binaural cues combined with the poor performance of some of the models employing their mean values, observed in the previously discussed Fig. 9, prompted these authors to undertake a follow-up analysis of the influence of the ensemble width on the values of the binaural cues. To this end, new binaural excerpts have been generated with ensembles located centrally in front of a listener ($\phi = 0°$), their widths ω varied between 0° and 90°, and with individual sound sources symmetrically distributed between left and right sides of a listener. The ensembles were still "asymmetric" regarding musical instruments. In other words, each sound source was unique (there were no left–right music source duplicates as this would give rise to a phantom mono effect). The reason for generating new excerpts instead of relying on the existing ones was the need to examine the effects of the changes in the binaural cues for spatially symmetric ensembles located in front of the listener. Recall that the existing repository of the binaural excerpts has been generated using randomized positions of the ensembles and random (asymmetric) distribution of sound sources within each ensemble.

Figure 10 illustrates how the changes in the ensemble width affect the mean values and standard deviations of the binaural cues (ILD, ITD, and IACC). For this example, a "rock music" excerpt has been utilized comprising nine individual music sound sources. The figure is divided into four rows, each one depicting the results obtained for the ensemble width ω being equal to 0°, 30°, 60°, and 90°, respectively. In the case of an infinitely narrow ensemble ($\omega = 0°$), illustrated in the top-most row of the figure (all sources located centrally in front of the listener), the mean values of both ILD and ITD equal zero, while the mean values of IACC coefficients amount to unity. The above outcome is in accordance with the expectations, as both binaural signals reaching left and right ears are the same. However, as the width ω is increased, the changes in the mean values of the binaural cues are more prominent. In line with the duplex theory [82], the changes in the mean values of the ILD cues are the most noticeable at high frequencies whereas the variations in the ITD cues are the most pronounced at low frequencies. Interestingly, for all three types of cues (ILD, ITD, and IACC), the standard deviations tend to increase with the broadening of the ensemble.

The examination of similar graphs obtained for the other binaural excerpts revealed that the results are specific to music recordings. Interestingly, for the excerpts containing a relatively large number of music sources, the results tend to be similar to those observed for the ensemble of decorrelated noise sound sources, illustrated in Fig. 11. In contrast to the outcomes depicted earlier for the rock music recording, for the ensemble of uncorrelated noise sources, the mean values of ILD and ITD approach zero, irrespective of the ensemble width ω. However, their standard deviations still constitute reliable indicators of the ensemble width, as their values tend to increase with the broadening of the ensemble. These observations constitute the explanation of the effects discussed in the previous section, regarding the poor performance of the models employing the mean values of the ILD and ITD cues and the satisfactory performance of the models utilizing their standard deviations. More graphs representing the influence of ensemble width on the values of the binaural cues (omitted in the paper due to space limitations) can be accessed on GitHub [65]. In the above repository, we also provided "animated" graphs illustrating the temporal changes of binaural cues as a function of the ensemble width.

4.3.4 Influence of the number of sources

In this work, we also investigated the potential effect of the number of sources within an ensemble on the performance of the regression model. Figure 12 illustrates the relationship between the prediction error (MAE) and the number of sources within ensembles. It can be seen that the prediction error tends to be "inversely" proportional to the number of sources within an ensemble. For example, for an excerpt entitled "A Place for Us" (indie pop/rock recording), consisting of 26 individual sound sources, the model exhibited a relatively small value of an MAE being equal to 5.0°. By contrast, for an excerpt titled "Nostalgic" (pop/electronica), the value of an MAE reached up to 12.2°. In the latter case, the ensemble consisted of only seven sources. The red curve in Fig. 12 represents a nonlinear regression fit using an exponential model, with the coefficient of determination beginning equal to $R^2 = 0.158$.

It should be emphasized that, in contrast to binaural localization techniques, our method is not restricted to an assumed maximum number of audio sources. For example, it can be seen in Fig. 12 that even for the excerpt consisting of 65 sources, titled "Donizetti" (opera), the prediction error was rather small, with an MAE being equal to 6.3°. As mentioned earlier, most

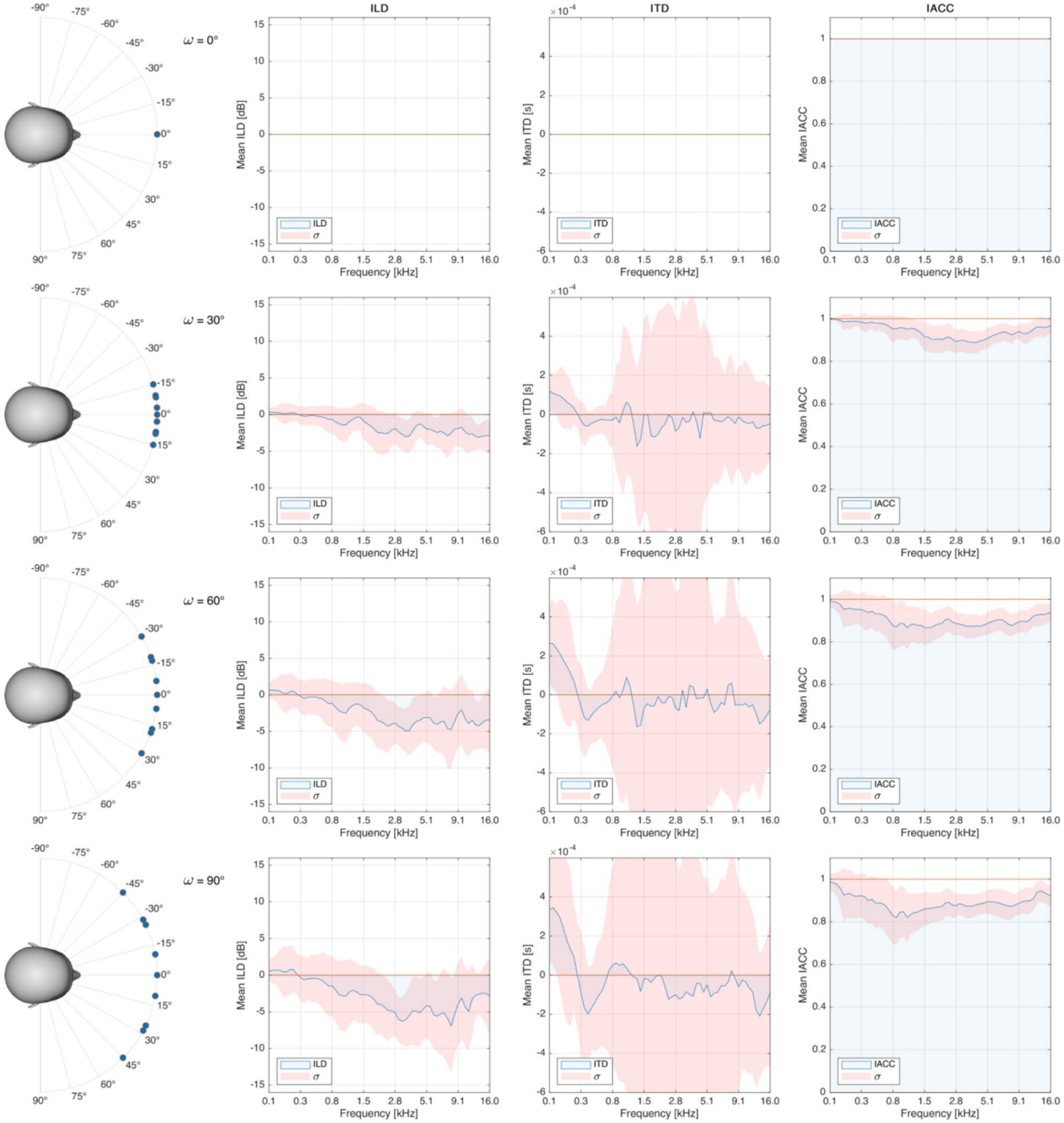


Fig. 10 Influence of ensemble width on binaural cues for a "rock music" excerpt with $N = 9$ music sound sources

of the existing DOA methods are limited to the analysis of a maximum of approximately three concurrently sound-emitting sources [5–12, 25–28]. As illustrated in Fig. 12, the outcomes are in line with expectations. When ensembles comprise a substantial number of individual sources, they tend to form more consistent macro entities, which results in a relatively low mean absolute error (MAE) compared to ensembles with a limited number of sources. Note that ensembles with a larger number of sources exhibit less variation in the MAE values.

4.3.5 Influence of the spectral features

Drawing inspiration from the study of Francl et al. [30], who discovered that the performance of their DOA model depended on the spectral characteristics of the binaural recordings, a separate experiment was

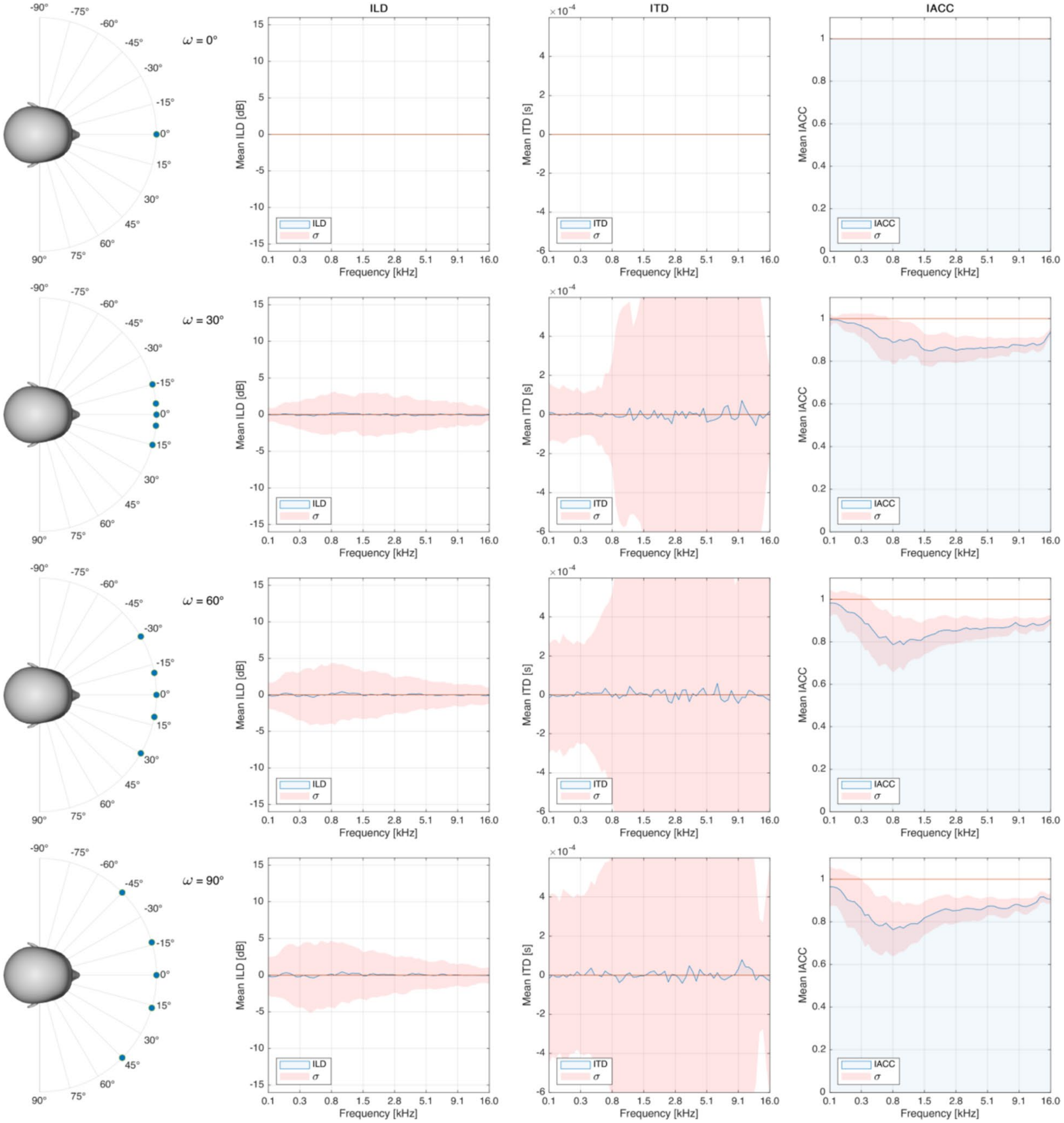


Fig. 11 Influence of ensemble width on binaural cues for $N=5$ uncorrelated noise sources

undertaken during which the influence of the spectral features on the accuracy of the method developed in this study was investigated. In our work, the following standard spectral features were considered: centroid, spread, brightness, high-frequency content, crest, decrease, entropy, flatness, irregularity, kurtosis, skewness, roll-off, flux, and variation. The mathematical definitions of these features are provided in [83]. They were calculated using the MATLAB toolbox developed within the Two Ears project [23]. For these calculations, a Hann window of 0.02 s with an overlap of 50% was used. For every feature, the mean value and standard deviation were estimated across the frames.

According to the obtained results, the skewness feature had the highest correlation with the MAE values. The spectral skewness is a measure of the degree of symmetry

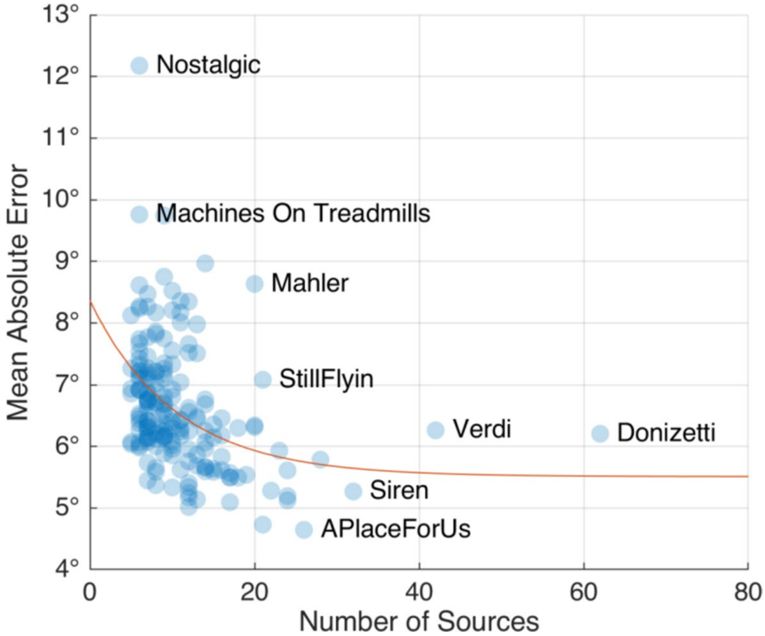


Fig. 12 Influence of the number of individual sound sources within ensembles on the model performance. The red curve represents a nonlinear regression fit using an exponential model ($R^2 = 0.158$)

of the spectrum around the spectral centroid, with zero skewness indicating perfect symmetry. Nevertheless, the observed value of Pearson's correlation coefficient was relatively small, amounting to $r = 0.044$ (SD 0.016). For comparison, the correlation coefficient obtained for spectral flux and spectral kurtosis was even smaller, amounting to 0.042 and 0.039, respectively. These outcomes suggest that the evaluated spectral features exerted a negligibly small influence on the predictive accuracy of the model.

Figure 13 shows the relationship between an error of the method and the above-mentioned spectral skewness

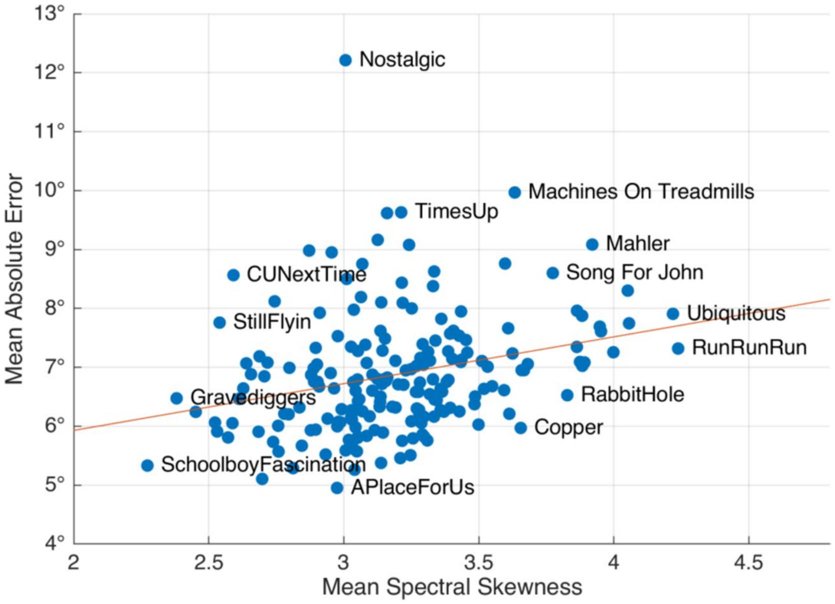


Fig. 13 Relationship between the spectral skewness of the music excerpts and the ensemble width estimation error (MAE). The red line represents a linear regression fit ($R^2 = 0.102$)

feature. The red line represents a linear regression fit with the coefficient of determination being equal to $R^2 = 0.102$. The recording marked as "nostalgic" was considered as an outliner and, hence, it was excluded from the regression analysis. It can be observed that music recordings with a higher degree of spectral symmetry (smaller skewness values) tend to slightly reduce the prediction error. Nevertheless, the figure visually confirms that the spectral skewness had a very small effect on the prediction error. Hence, it can be concluded that in contrast to the study of Francl et al. [30], the performance of the method proposed in this paper does not depend on the spectral characteristics of the music recordings.

4.4 Generalizability tests

Recall that the datasets used for the development and testing procedures contained different music recordings. Hence, all the results presented above were obtained under the "music-independent" condition and, therefore, they already give some insight into the degree of generalizability of the developed method. However, the training and the testing procedures were carried out in an "HRTF-dependent" way since the same 30 HRTF datasets were used to synthesize the binaural excerpts applied both for development and for testing. The potential risk of overtraining the machine learning models using HRTF-convolved sound recordings, limiting their generalization properties, was recently highlighted by Wang et al. [20]. Therefore, to estimate the generalizability of the developed method more rigorously, additional two experiments were performed. They were undertaken not only under the music-independent but also under the HRTF-independent scenario, referred to as the "mismatched HRTF condition" by Wang et al. [20].

In the first experiment, the model was trained using a subset of the development dataset, limited to N randomly selected HRTFs (out of 30 HRTFs used in this study). Then, the trained model was tested on the test set reduced to the excerpts synthesized with the HRTFs "unseen" during the training procedure. This procedure was repeated 9 times for $N = [1, 2, 3, 5, 7, 10, 15, 20, 25]$, with the results presented in Fig. 14. It shows that the more HRTFs are used during the training, the better the model generalizes to the unseen HRTFs. Moreover, it can be seen that as few as $N = 5$ different HRTFs are sufficient to reduce the MAE below 10°.

In the second experiment, another approach was taken to investigate the generalizability property of the method. First, the model was trained on the subset of the excerpts from the development set, synthesized using the HRTFs derived from the "artificial" heads. Subsequently, it was tested on the subset of the test set, with the excerpts generated employing solely the "human" HRTFs. In the second part of the experiment, the procedure was reversed. First, the model was trained on the excerpts obtained using "human" HRTFs, and then it was tested on the recordings synthesized with the "artificial" HRTFs. The obtained results are summarized in Table 3. It can be seen that for the regression model trained on the excerpts generated using the "artificial" HRTFs and then tested on the recordings obtained with the "human" HRTFs, the MAE was equal to 7.84°. When the procedure was reversed, the

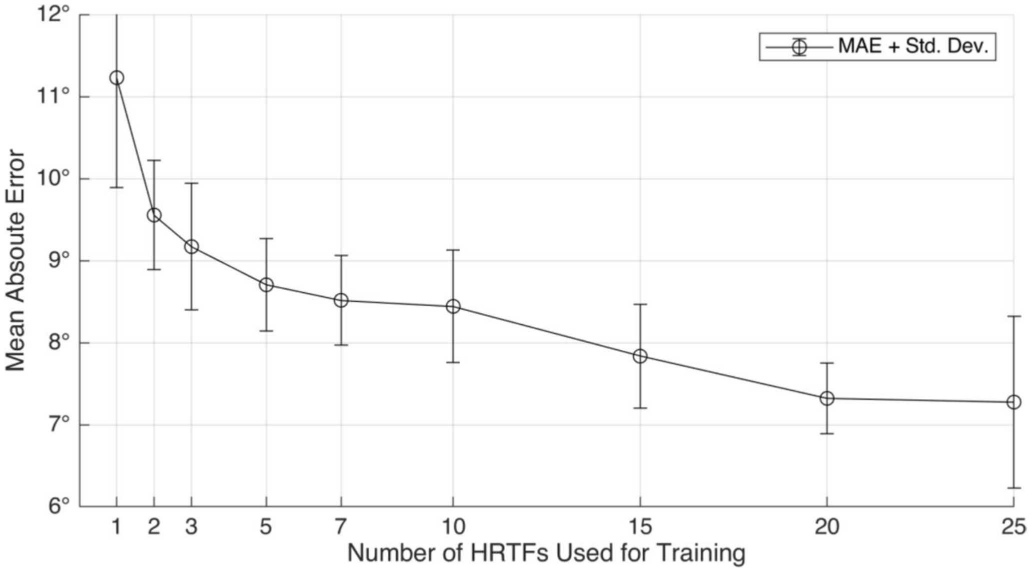


Fig. 14 MAE values obtained under the HRTF-independent test as a function of the total number of HRTFs used for training. Error bars denote standard deviations

Table 3 Influence of the types of heads used in the training and testing procedures

HRTFs used for training	HRTFs used for testing	Mean absolute error	Standard deviation
Artificial	Human	7.84°	0.15°
Human	Artificial	9.43°	0.14°

error increased to 9.43°. This difference was statistically significant at $p = 8.64 \times 10^{-8}$ level, according to the t-test. This outcome can be explained by the observation that artificial heads can be thought of as "averaged" versions of human heads in terms of their dimensions and are therefore more representative. Hence, they are better suited for training models than highly differentiated human heads. Artificial heads also tend to have better-quality microphones, placed in "optimal" locations. Furthermore, HRTFs obtained using human heads may exhibit inferior technical quality relative to those acquired with artificial heads. The involvement of humans in such recordings presents a number of challenges, including the immobilization of their heads during the measurement procedure, the fitting of microphones in ear canals, and the reduction of sound exposure levels and durations (thereby reducing the signal-to-noise ratio).

To summarize, according to the experiments described in this section, the developed method proved to be generalizable concerning the types of HRTFs applied for training. Using artificial HRTFs for training, instead of human ones, may yield slightly better generalization results. A minimum of five different HRTF datasets are required for the training procedure to achieve a reasonable level of generalizability of the method, with MAE being less than 10°.

5 Discussion

In contrast to the state-of-the-art DOA binaural analysis algorithms [5, 7–12, 27, 29], the developed method does not require any a priori knowledge about the number or characteristics of sound sources in analyzed recordings. Moreover, unlike the DOA methods, the performance of the technique proposed in this paper tends to improve with the increased number of audio sources. Hence, the method is capable of the "blind" analysis of the binaural audio signals. However, since it has been developed under simulated anechoic conditions, using HRTF-convolved binaural excerpts, it cannot be reliably applied to analyze "reverberant" music recordings, such as those available on the Internet, which constitutes a significant limitation of the method. Reverberant conditions will be considered in future developments, extending beyond the scope of this study.

The binaural music excerpts employed in this study were randomly auditioned by these authors. The subjectively experienced level of externalization was rather modest, with some ensembles perceived almost "in the head." This effect can be attributed to the anechoic nature of the auditioned recordings, as reverberations play a key role in the externalization of the binaurally reproduced sound sources [84]. Interestingly, for some of the auditioned excerpts, the perceived ensemble width appeared to be wider than intended during their synthesis (convolution) procedure. This effect could be explained by the study of Pulkki et al. [85], who observed that even with a loudspeaker-based reproduction system, the ensemble width is perceived as slightly wider than physically reproduced. Nevertheless, the above observation highlights another limitation of the study. Namely, the developed method predicts the mathematically "intended" rather than subjectively "perceived" ensemble width of binaural music recordings. While this observation does not invalidate the study, formal listening tests are planned in the future to gather more data regarding the above hypothetical disparity.

The rationale for employing the auditory model as a feature extractor in the proposed method was to emulate the human hearing mechanism. Nevertheless, the decision to extend the upper limit of the auditory model from a typical value of 5–8 kHz [7, 8, 11] to 16 kHz exceeds the human capacity to process binaural cues [68], thereby introducing inconsistency to the applied methodology. While Baumgartner et al. [44] and Barumerli et al. [45] have also extended the upper frequency of the gammatone filter bank, even up to 18 kHz, they did so for the purpose of extracting "monaural" spectral cues from the filter bank output signals. This allows for the acquisition of crucial information regarding the positions of sound sources in the sagittal plane. For instance, elevated sound sources often result in a discernible frequency notch in the range of 6–9 kHz, as compared to the magnitude spectra below 6 kHz and above 9 kHz [46]. Given that our study was constrained to the horizontal plane, extending the frequency beyond 8 kHz does not align with the human hearing perspective.

At the outset of this study, we assumed that the wider the ensemble, the more uncorrelated the signals reaching the listener's ears, resulting in a change in the IACC coefficient values. Hence, we anticipated that IACC would serve as a primary indicator in the estimation of EW. However, the results presented in Sect. 4.3.2 have not corroborated the hypothesis that IACC is the most significant cue in predicting EW. Nevertheless, while IACC did not turn out to be the leading cue, it demonstrated a notable prominence, as evidenced by the experimental outcomes. For instance, Fig. 9 demonstrates that the model can estimate EW with a mean absolute error (MAE) of less than 9° using IACC as the sole cue (without interaural time difference or interaural level difference). Moreover, Fig. 10 and Fig. 11 illustrate that IACC coefficients fluctuate in conjunction with EW, exhibiting lower values for wider ensembles. Interestingly, the results provided in Sects. 4.3.2 and 4.3.3 suggest that standard deviations of all three types of binaural cues (ITD, ILD, and IACC) also carry useful information as their values tend to increase with the broadening of ensembles. This observation is in line with the study of Lee and Johnson [86], comparing the 3D audio microphone arrays, who observed that the standard deviations of the binaural cues fluctuations over time constitute prominent metrics differentiating between the performance of the microphone arrays.

6 Conclusions

This study demonstrates that an auditory model coupled with a gradient-boosted decision trees regression algorithm can be successfully used to estimate ensemble width in binaural recordings of music under simulated anechoic conditions. The proposed method outperforms the technique based on spatiograms, recently introduced in the literature [15]. The mean absolute error of the developed method averaged across investigated conditions is equal to 6.63° (SD 0.12°). The method exhibits the best performance for ensembles narrower than 30°, with the error ranging between 0.8° and 10.2°. Its accuracy deteriorates for wider ensembles, with the error reaching 25.2° for the music ensembles spanning 90°. In general, the estimation error tends to increase with the width of the ensembles. However, the relationship between the estimation error and ensemble width is not monotonic, exhibiting a local minimum (signifying relative performance improvements), for ensemble widths encompassing the range of approximately 60°–80°.

A distinct feature of the proposed method is that it does not require any a priori knowledge regarding the number or characteristics of audio sources in analyzed music recordings. Moreover, unlike the DOA estimation algorithms, its performance is not hindered by the increased number of concurrent audio sources. Furthermore, the method is relatively robust to the changes in the location of analyzed ensembles. Additionally, the technique is insensitive to the variations in spectral characteristics of the binaural signals.

The results of the experiments demonstrated that interaural level differences (ILD) and interaural time differences (ITD) are the primary factors influencing the estimation of ensemble width. Moreover, interaural cross-correlation (IACC) coefficients contribute to this estimation process, providing supplementary information that facilitates the task of estimating ensemble width. Furthermore, the results revealed the importance of the standard deviations of all three binaural cues (ILD, ITD, and IACC).

The developed technique exhibits satisfactory generalization properties when evaluated both under music-independent and HRTF-independent conditions. The generalizability tests revealed that a minimum of five different HRTFs, applied to synthesize the training excerpts, are sufficient to reduce the level of the mean absolute error to below 10°. The method was found to be HRTF-independent for the HRTF sets obtained using typical human heads, in terms of their dimensions, or commonly used artificial heads, such as the Neumann KU100, KEMAR, and FABIAN. However, it has yet to be validated with HRTFs that deviate significantly from the physical properties of typical heads.

The method has been developed and tested under simulated anechoic conditions with HRTF-convolved binaural signals, excluding such factors as background noise, room reflections, or different recording conditions. Consequently, it cannot be reliably applied to real-life reverberant recordings with accompanying ambient noise, limiting the scope of its practical applications. More ecologically valid conditions, including reverberant environments, will be considered in future developments.

The proposed method is relatively complex. It employs a sophisticated auditory model, which is computationally demanding. Moreover, it uses highly-dimensional data and advanced machine learning algorithm, requiring significant computational resources. These factors present significant challenges in terms of real-time applications of the developed technique. Currently, it is not feasible to deploy the method in mobile devices or in live settings. Therefore, the next phase of the research will involve the computational optimization of the method. Furthermore, listening tests are scheduled for the future to evaluate the extent to which the developed model is capable of predicting the width of the ensemble as perceived by humans. Moreover, in line with the current trends in machine learning [28, 33, 39], the authors plan to incorporate deep learning techniques for the task of the estimation of ensemble width in binaural recordings. This will facilitate a comparative assessment of their performance with that achieved through more transparent (explainable) methods, such as the one proposed in this paper. The authors hope that this work will prompt other researchers to shift their focus from low to high levels of the analysis of spatial audio scenes, leading to a more comprehensive characterization of spatial audio recordings.

Appendix

Table 4 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 [49]	AACHEN
2.	Artificial	GRAS 45BB-4 KEMAR	1		
3.	Human	Subject 2	1.2	Austrian Academy of Sciences [62]	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 [87]	CIPIC
8.	Human	Subject 015	1		
9.	Human	Subject 020	1		
10.	Artificial	Neumann KU 100	0.9	NASA (2007) [57]	CLUBFRITZ
11.	Artificial	Neumann KU 100	1.5	Helsinki University of Technology (2009) [57]	
12.	Artificial	FABIAN	1.47	Technical University Berlin, Huawei Technologies, Munich Research Centre, Sennheiser Electronic [60]	HUTUBS
13.	Human	Subject pp2	1.47		
14.	Human	Subject pp3	1.47		
15.	Human	Subject 1003	1.95	IRCAM, AKG [51]	LISTEN
16.	Human	Subject 1002	1.95		
17.	Artificial	KEMAR DB-4004 (DB-061)	1.4	MIT [50]	MIT
18.	Artificial	KEMAR DB-4004 (DB-065)	1.4		
19.	Human	Subject 001	1.5	Tohoku University [61]	RIEC
20.	Human	Subject 002	1.5		
21.	Artificial	Koken SAMRAI	1.5		
22.	Artificial	Neumann KU 100	1.2	University of York [58]	SADIE II
23.	Human	Subject H3	1.2		
24.	Human	Subject H4	1.2		
25.	Artificial	KEMAR	1	South China University of Technology [52]	SCUT
26.	Artificial	Neumann KU 100	1	TH Köln [59]	TH Köln
27.	Artificial	FABIAN	1.7	TU Berlin [53, 54]	TU Berlin
28.	Artificial	GRAS 45BA KEMAR	1		
29.	Artificial	GRAS 45BB-4 KEMAR – subject A attachment	1	Aalborg University; University of Iceland [55, 56]	VIKING
30.	Artificial	GRAS 45BB-4 KEMAR – subject B attachments	1		

Table 5 The hyperparameters values identified as the best in the cross-validation procedure

Experiment repetition number	1	2	3	4	5	6	7
Maximum number of leaves *n*	1000	10007	500	1500	500	1500	500
Maximum depth for a tree model *d*	6	6	6	6	6	6	6
Learning rate *l*	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Abbreviations

AEW	Apparent ensemble width
ASW	Apparent source width
DC	Direct current
DOA	Direction of arrival
EW	Ensemble width
GCC-PHAT	Generalized cross-correlation function with phase transform
gbm	Generalized boosted models
HRIR	Head-related impulse response
HRTF	Head-related transfer function
IACC	Inter-aural cross-correlation
KEMAR	Knowles Electronics Manikin for Acoustic Research
ILD	Interaural level difference
ITD	Interaural time difference
LightGBM	Light gradient boosting machine
LUFS	Loudness units (relative to) full scale
MAE	Mean absolute error
MIR	Music information retrieval
MSD	Mean signed difference
pGBRT	Parallel gradient boosted regression trees
POSC	Phase-only spatial correlation
RMS	Root-mean-square

SASC	Spatial audio scene characterization
SD	Standard deviation
SW	Source width
VIF	Variance inflation factor
XGBoost	Extreme gradient boosting

Acknowledgements
Not applicable.

Authors' contributions
PA synthesized the binaural excerpts, adopted the auditory model, applied the gradient-boosted decision trees regression algorithm, and carried out the experiments. SKZ, PA, and HL jointly developed the research methodology. HL helped in the interpretation of the results. SKZ wrote the initial draft of the manuscript. All authors contributed to amending the subsequent versions of the paper and further read and approved the final manuscript.

Funding
The work was supported by grants from Białystok 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.

Availability of data and materials
The code developed within this study is made publicly available at GitHub [65]. The binaural music excerpts, generated and exploited in this work, are not publicly available due to copyright restrictions. However, they are available from the authors upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 21 March 2024 Accepted: 2 October 2024
Published online: 10 October 2024

References

1. S.K. Zieliński, P. Antoniuk, H. Lee, Spatial Audio Scene Characterization (SASC): automatic localization of front, back, up, and down-positioned music ensembles in binaural recordings. Appl. Sci. **12**(3), 1569 (2022). https://doi.org/10.3390/app12031569
2. F. Rumsey, T. McCormick, *Sound and Recording: An Introduction* (Focal Press, London, 2009), pp. 429–468. https://doi.org/10.4324/9780080953960
3. H. Lee, C. Millns, In *Proc. of the 143rd AES Convention.* Microphone array impulse response (MAIR) library for spatial audio research (AES, New York, NY, USA, 2017), Convention e-Brief 356
4. F. Rumsey, Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. J. Audio Eng. Soc. **50**(9), 651–666 (2002)
5. M. Dietz, S.D. Ewert, V. Hohmann, Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Commun. **53**(5), 592–605 (2011). https://doi.org/10.1016/j.specom.2010.05.006
6. T. May, S. van de Par, A. Kohlrausch, A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans. Audio, Speech, Language Process. **19**(1), 1–13 (2011). https://doi.org/10.1109/TASL.2010.2042128
7. T. May, S. van de Par, A. Kohlrausch, A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. IEEE Trans. Audio, Speech, Language Process. **20**(7), 2016–2030 (2012). https://doi.org/10.1109/TASL.2012.2193391
8. J. Woodruff, D. Wang, Binaural localization of multiple sources in reverberant and noisy environments. IEEE Trans. Audio, Speech, Language Process. **20**(5), 1503– 1512 (2012). https://doi.org/10.1109/TASL.2012.2183869
9. T. May, N. Ma, G.J. Brown, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues (South Brisbane, Australia 2015), pp. 2679–2683. https://doi.org/10.1109/ICASSP.2015.7178457
10. N. Ma, G.J. Brown, in *Proc of the INTERSPEECH*. Speech localisation in a multitalker mixture by humans and machines (San Francisco, CA, USA, 2016), pp. 3359–3363. https://doi.org/10.21437/Interspeech.2016-1149
11. N. Ma, T. May, G.J. Brown, Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(12), 2444–2453 (2017). https://doi.org/10.1109/TASLP.2017.2750760
12. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raumel, S. Argentieri, Binaural localization of multiple sound sources by non-negative tensor factorization. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(6), 1072–1082 (2018). https://doi.org/10.1109/TASLP.2018.2806745
13. J. Käsbach, M. Hahmann, T. May, T.Dau, in *Proc. of DAGA (Deutsche Gesellschaft für Akustik)*. Assessing and modeling apparent source width perception. (Aachen, Germany, 2016)
14. S. Arthi, T.V. Sreenivas, Spatiogram: a phase based directional angular measure and perceptual weighting for ensemble source width (2021). https://arxiv.org/abs/2112.07216
15. S. Arthi, T.V. Sreenivas, in *Proceedings of the IEEE International Conference on Signal Processing and Communications (SPCOM).* Binaural Spatial Transform for Multi-source Localization determining Angular Extent of Ensemble Source Width. (Bangalore, India, 2022), pp. 1–5. https://doi.org/10.1109/SPCOM55316.2022.9840782
16. S. Sato, Y. Ando, Apparent Source Width (ASW) of Complex Noises in Relation to the Interaural Cross-correlation Function. J. Temporal Des. Arch. Environ. **2**(1), 29–32 (2002)
17. M. Barron, A.H. Marshall, Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. J. Sound Vib. **77**(2), 211–232 (1981). https://doi.org/10.1016/S0022-460X(81)80020-X
18. E.K. Canfield-Dafilou, J.S. Abel, In *Proc. of the 144th AES Convention.* A Group Delay-Based Method for Signal Decorrelation (AES, Milan, Italy, 2018), Convention Paper 9991
19. P. Antoniuk, S.K. Zieliński, in *Proc. of the AES International Conference on Spatial and Immersive Audio.* Blind estimation of ensemble width in binaural music recordings using 'spatiograms' under simulated anechoic conditions. Paper Number 15 (Huddersfield, UK, 2023)
20. J. Wang, J. Wang, K. Qian, X. Xie, J. Kuang, Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. EURASIP J. Audio, Speech Music Process. **4** (2020). https://doi.org/10.1186/s13636-020-0171-y
21. P.L. Søndergaard, P. Majdak, The Auditory Modeling Toolbox, in *The Technology of Binaural Listening*, ed. by J. Blauert (Springer, Berlin Heidelberg, 2013), pp. 33–56. https://doi.org/10.1007/978-3-642-37762-4
22. R. Decorsière, T. May, Auditory front-end. Two Ears Project Documentation (2016). https://docs.twoears.eu/en/latest/afe/ Accessed 5 June 2021.
23. A. Raake, A computational framework for modelling active exploratory listening that assigns meaning to auditory scenes—reading the world with two ears (2016), http://twoears.eu. Accessed 5 June 2021.
24. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Liu, in *Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. LightGBM: a highly efficient gradient boosting decision tree (Long Beach, CA, USA, 2017). pp. 3149–3157
25. Q. Liu, W. Wang, T. de Campos, P.J.B. Jackson, A. Hilton, Multiple speaker tracking in spatial audio via PHD filtering and depth-audio fusion. IEEE Trans. Multimed. **20**(7), 1767–1780 (2018). https://doi.org/10.1109/TMM.2017.2777671
26. D.A. Krause, A. Politis, A. Mesaros, in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* Joint Direction and Proximity Classification of Overlapping Sound Events from Binaural Audio (New Paltz, NY, USA, 2021), pp. 331–335. https://doi.org/10.1109/WASPAA52581.2021.9632775
27. Q. Yang, Y. Zheng, DeepEar: Sound localization with binaural microphones. IEEE Trans Mob Comput. Early Access (2022). https://doi.org/10.1109/TMC.2022.3222821
28. L. Wang, Z. Jiao, Q. Zhao, J. Zhu, Y. Fu, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Framewise multiple sound source localization and counting using binaural spatial

audio signals (Rhodes Island, Greece, 2023), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096463

29. N. Ma, J. Gonzalez, G.J. Brown, Robust Binaural Localization of a Target Sound Source by Combining Spectral Source Models and Deep Neural Networks. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(11), 2122–2131 (2018). https://doi.org/10.1109/TASLP.2018.2855960
30. A. Francl, J.H. McDermott, Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nat. Hum. Behav. **6**, 111–133 (2022). https://doi.org/10.1038/s41562-021-01244-z
31. M. Zohourian, R. Martin, Binaural direct-to-reverberant energy ratio and speaker distance Estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 92–104 (2020). https://doi.org/10.1109/TASLP.2019.2948730
32. L.A. Jeffress, A place theory of sound localization. J. Comp. Physiol. Psychol. **41**, 35–39 (1948). https://doi.org/10.1037/h0061495
33. A.M. El-Mohandes, N.H. Zandi, R. Zheng, DeepBSL: 3-D personalized deep binaural sound localization on earable devices. IEEE Internet Things J. **10**(21), 19004–19013 (2023). https://doi.org/10.1109/JIOT.2023.3281128
34. R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S.Q. Lee, H.-M. Park, Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments. IEEE Access **8**, 7373–7382 (2020). https://doi.org/10.1109/ACCESS.2019.2963768
35. Z. Pan, M. Zhang, J. Wu, J. Wang, H. Li, Multi-tone phase coding of interaural time difference for sound source localization with spiking neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 2656–2670 (2021). https://doi.org/10.1109/TASLP.2021.3100684
36. M. Kuk, S. Bobek, G.J. Nalepa, Comparing Explanations from Glass-Box and Black-Box Machine-Learning Models, in *Lecture Notes in Computer Science. Computational Science – ICCS 2022, vol 13352*, ed. by D. Groen *et al.* (Springer, Cham, 2022). https://doi.org/10.1007/978-3-031-08757-8_55
37. P. Vecchiotti, N. Ma, S. Squartini, G.J. Brown, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end binaural sound localisation from the raw waveform (Brighton, UK, 2019), pp. 451–455. https://doi.org/10.1109/ICASSP.2019.8683732
38. C. Pang, H. Liu, X. Li, Multitask learning of time-frequency CNN for sound source localization. IEEE Access **7**, 40725–40737 (2019). https://doi.org/10.1109/ACCESS.2019.2905617
39. Q. Hu, N. Ma, G.J. Brown, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Robust Binaural Sound Localisation with Temporal Attention (Rhodes Island, Greece, 2023), pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096640
40. B. Yang, H. Liu, X. Li, Learning deep direct-path relative transfer function for binaural sound source localization. IEEE/ACM Trans. Audio, Speech, Language Process. **29**, 3491–3503 (2021). https://doi.org/10.1109/TASLP.2021.3120641
41. I. Örnolfsson, T. Dau, N. Ma, T. May, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Exploiting non-negative matrix factorization for binaural sound localization in the presence of directional interference (Toronto, Canada, 2021), pp. 221–225. https://doi.org/10.1109/ICASSP39728.2021.9414233
42. J. Blauert, *Spatial hearing. The psychology of human sound localization* (MIT Press, London, 1974), pp. 97–115. https://doi.org/10.7551/mitpress/6391.003.0006
43. C. Faller, J. Merimaa, Source localization in complex listening situations: selection of binaural cues based on interaural coherence. J. Acoust. Soc. Am. **116**(5), 3075–3089 (2004). https://doi.org/10.1121/1.1791872
44. R. Baumgartner, P. Majdak, B. Laback, Modeling sound-source localization in sagittal planes for human listeners. J. Acoust. Soc. Am. **136**(2), 791–802 (2014). https://doi.org/10.1121/1.4887447
45. R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, R. Baumgartner, A Bayesian model for human directional localization of broadband static sound sources. Acta Acustica **7**(12), 1–13 (2023). https://doi.org/10.1051/aacus/2023006
46. B. Zonooz, E. Arani, K.P. Körding et al., Spectral Weighting Underlies Perceived Sound Elevation. Sci. Rep. **9**, 1642 (2019). https://doi.org/10.1038/s41598-018-37537-z
47. M. Senior, The 'mixing secrets' free multitrack download library. Cambridge Music Technology. Music recording repository (2023). https://cambridge-mt.com/ms/mtk/. Accessed 15 Jan 2023
48. ITU-R Rec. BS.1770–5, Algorithms to measure audio programme loudness and true-peak audio level. International Communications Union (Geneva, Switzerland, 2023).
49. H.S. Braren, J. Fels, in *A high-resolution individual 3D adult head and torso model for HRTF simulation and validation*. 3D data. Technical Report. Institute of Technical Acoustics (RWTH Aachen University, 2020). https://doi.org/10.18154/RWTH-2020-06761
50. B. Gardner, K. Martin, HRTF measurements of a KEMAR dummy-head microphone. MIT Media Lab (1994), https://sound.media.mit.edu/resources/KEMAR.html. Accessed 15 June 2021
51. LISTEN HRTF Database (2003), http://recherche.ircam.fr/equipes/salles/listen. Accessed 15 June 2021
52. G. Yu, R. Wu, Y. Liu, B. Xie, Near-field head-related transfer-function measurement and database of human subjects. J. Acoust. Soc. Am. **143**(3), EL194 (2018). https://doi.org/10.1121/1.5027019
53. F. Brinkmann, A. Lindau, S.S. van de Par, M. Müller-Trapet, R. Opdam, M. Vorländer, A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations. J. Audio Eng. Soc. **65**(10), 841–848 (2017). https://doi.org/10.17743/jaes.2017.0033
54. H. Wierstorf, M. Geier, A. Raake, S. Spors, in *Proc. of the 130th AES Convention*. A free database of head-related impulse response measurements in the horizontal plane with multiple distances (AES, London, UK, 2011) e-Brief 6
55. S. Spagnol, K.B. Purkhus, S.K. Björnsson, R. Unnthórsson, in *Proc. of the 16th Sound & Music Computing Conference (SMC 2019)*. The Viking HRTF dataset (Malaga, Spain, 2019)
56. S. Spagnol, R. Miccini, R. Unnthórsson, *The Viking HRTF dataset v2* (2020), https://zenodo.org. Accessed 15 June 2021. https://doi.org/10.5281/zenodo.4160401
57. A. Andreopoulou, D.R. Begault, B.F.G. Katz, Inter-Laboratory Round Robin HRTF Measurement Comparison. IEEE J. Sel. Topics Sig. Process. **9**(5), 895–906 (2015). https://doi.org/10.1109/JSTSP.2015.2400417
58. C. Armstrong, L. Thresh, D. Murphy, G. Kearney, A perceptual evaluation of individual and non-individual HRTFs: a case study of the SADIE II database. Appl. Sci. **8**(2029) (2018). https://doi.org/10.3390/app8112029
59. C. Pörschmann, J.M. Arend, A. Neidhardt, in *Proc. of the 142nd AES Convention*. A spherical near-field HRTF set for auralization and psychoacoustic research (AES, Berlin, Germany, 2017) e-Brief 322
60. F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, S. Weinzier, A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses. J. Audio Eng. Soc. **67**(9), 705–718 (2019). https://doi.org/10.17743/jaes.2019.0024
61. K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, S. Sato, Dataset of head-related transfer functions measured with a circular loudspeaker array. Acoust. Sci. Tech. **35**(3), 159–165 (2014). https://doi.org/10.1250/ast.35.159
62. HRTF-Database. Acoustic Research Institute. Austrian Academy of Sciences (2014), https://www.oeaw.ac.at/en/isf/das-institut/software/hrtf-database. Accessed 15 June 2021
63. F.P. Freeland, L.W.P. Biscainho, P.S.R. Diniz, In *Proc. of the 12th European Signal Processing Conference*, Interpolation of Head-Related Transfer Functions (HRTFS): A multi-source approach (Vienna, 2004), pp. 1761–1764
64. The MathWorks Inc. (2022). MATLAB 9.12.0.2529717 (R2022a), Audio Toolbox. Natick, Massachusetts, United State. Retrieved from https://www.mathworks.com/help/audio/
65. P. Antoniuk, Ensemble width estimation in HRTF-convolved binaural music recordings using an auditory model and a gradient-boosted decision trees regressor. Software Repository (2024). https://github.com/pawel-antoniuk/appendix-paper-eurasip-2024. Accessed 29 January 2024.
66. B.R. Glasberg, B.C.J. Moore, Derivation of auditory filter shapes from notched-noise data. Hear. Res. **47**(1–2), 103–138 (1990). https://doi.org/10.1016/0378-5955(90)90170-T
67. S.K. Zieliński, P. Antoniuk, H. Lee, D. Johnson, Automatic discrimination between front and back ensemble locations in HRTF-convolved binaural recordings of music. EURASIP J Audio Speech Music Process. **2022**, 3 (2022). https://doi.org/10.1186/s13636-021-00235-2
68. E. Verschooten, S. Shamma, A.J. Oxenham, B.C.J. Moore, P.X. Joris, M.G. Heinz, C.J. Plack, The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints.

Hear. Res. **377**, 109–121 (2019). https://doi.org/10.1016/j.heares.2019.03.011

69. A.J. Peterson, P. Heil, Phase Locking of Auditory Nerve Fibers: The Role of Lowpass Filtering by Hair Cells. J. Neurosci. **40**(24), 4700–4714 (2020). https://doi.org/10.1523/JNEUROSCI.2269-19.2020
70. G.J. Brown, M. Cooke, Computational auditory scene analysis. Comput. Speech Lang. **8**(4), 297–336 (1994). https://doi.org/10.1006/csla.1994.1016
71. D. Schreiber-Gregory, K. Bader, in *Proc. of the SAS Conference Proceedings: Western Users of SAS Software*, Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets (Denver, CO, USA. 2018), Paper 248
72. J.H. Friedman, Greedy function approximation: a gradient boosting machine. Ann. Statist. **29**(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
73. P. Li, C.J.C. Burges, Q. Wu, in *Proc. of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*. McRank: learning to rank using multiple classification and gradient boosting (Red Hook, NY, USA, 2007), pp. 897–904
74. S. Neelakandan, D. Paulraj, A gradient boosted decision tree-based sentiment classification of twitter data. Int. J. Wavelets Multiresolut. Inf. Process. **18**(04), 2050027 (2020). https://doi.org/10.1142/S0219691320500277
75. H. Seto, A. Oyama, S. Kitora, H. Toki, R. Yamamoto, J. Kotoku, A. Haga, M. Shinzawa, M. Yamakawa, S. Fukui, T. Moriyama, Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci. Rep. **12**, 15889 (2022). https://doi.org/10.1038/s41598-022-20149-z
76. T. Chen, C. Guestrin, in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. XGBoost: a scalable tree boosting system (ACM, 2016). https://doi.org/10.1145/2939672.2939785
77. S. Tyree, K.Q. Weinberger, K. Agrawal, J. Paykin, in *Proc. of the 20th international conference on World wide web (WWW '11)*. Parallel boosted regression trees for web search ranking (ACM, 2011), pp. 387–396. https://doi.org/10.1145/1963405.1963461
78. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
79. G. Ridgeway. Generalized boosted models: a guide to the gbm package. Update, **1**(1), (2007)
80. H. Zhang, S. Si, C.-J. Hsieh, in *Proc. of the SysML Conference*. GPU Acceleration for Large-scale Tree Boosting (Standford, CA, USA, 2018). https://arxiv.org/abs/1706.08359
81. LightGBM: Light Gradient Boosting Machine. Software Repository. Microsoft (2021) https://github.com/microsoft/LightGBM. Accessed 13 Jan 2024
82. W.M. Hartmann, B. Rakerd, Z.D. Crawford, P.X. Zhang, Transaural experiments and a revised duplex theory for the localization of low-frequency tones. J. Acoust. Soc. Am. **139**(2), 968–985 (2016). https://doi.org/10.1121/1.4941915
83. G. Peeters, B. Giordano, P. Susini, N. Misdariis, S. McAdams, Extracting audio descriptors from musical signals. J. Acoust. Soc. Am. **130**(2902), 2902–2916 (2011). https://doi.org/10.1121/1.3642604
84. V. Best, R. Baumgartner, M. Lavandier, P. Majdak, N. Kopčo, Sound externalization: a review of recent research. Trends. Hear. (2020). https://doi.org/10.1177/2331216520948390
85. V. Pulkki, H. Pöntynen, O. Santala, Spatial perception of sound source distribution in the median plane. J. Audio Eng. Soc. **67**(11), 855–870 (2019). https://doi.org/10.17743/jaes.2019.0033
86. H. Lee, D. Johnson, 3D microphone array comparison: objective measurements. J. Audio Eng. Soc. **69**(11), 871–887 (2021). https://doi.org/10.17743/jaes.2021.0038
87. V.R. Algazi, R.O. Duda, D.M. Thompson, C. Avendano, in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*. The CIPIC HRTF Database (IEEE, Mohonk Mountain House, New Paltz, NY, USA, 2001). https://doi.org/10.1109/ASPAA.2001.969552

Publisher's Note