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

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

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

Abstract

Binaural audio technology has been in existence for many years, but its popularity has significantly increased over the past decade as a consequence of advancements in virtual reality and streaming technologies. Along with its growing popularity, the quantity of publicly accessible binaural audio materials has also expanded. Consequently, there is now a need for automated and objective measurements of spatial content information, with ensemble location and width being the most prominent. This study presents a novel method for predicting ensemble location and width in binaural recordings of music. To this end, 30 head-related transfer functions and 192 binaural music recordings from publicly accessible multi-track recording repositories were used to synthesize 23,040 binaural recordings. The synthesized recordings were then used to train a multi-task convolutional neural network model, the aim of which was to predict the location and width of the ensemble for unseen recordings. The results indicate that models of ensemble location and width can be successfully constructed with low prediction errors — 4.76° (±0.10°) for ensemble location and 8.57° (±0.19°) for ensemble width. The method developed in this study outperforms previous spatiogram-based methods recently published in the literature and holds promise for future development as a part of a novel tool for audio engineers to assess binaural audio.

1 Introduction

The human auditory system demonstrates exceptional proficiency in segregating, localizing, and interpreting diverse auditory signals, despite being limited to two ears. This is possible by internal examination of interaural differences in time, loudness, and frequency, known as binaural hearing (Blauert, 1996), which enables precise localization of sound sources in complex auditory environments. A notable advantage of binaural hearing is exemplified by the ‘cocktail party effect’, highlighting the system’s capacity to concentrate on foreground sound sources while suppressing background noise (Cherry, 1953).

Understanding of the auditory system is essential for comprehending its limits and creating more immersive binaural experiences for entertainment purposes (W. Zhang et al., 2017). It is also important for enhancing auditory signal reception in hearing aid devices (Hirsh, 1950; Thiemann et al., 2016).

The advancement of sophisticated machine learning techniques, particularly deep learning networks, has prompted an intriguing exploration into the extent to which these tools can emulate the human auditory system. This exploration proceeds without reliance on the advanced spatial audio feature engineering traditionally used in audio source localization, as evidenced by some prominent recent examples (Pang et al., 2019; Vera-Diaz et al., 2018; Yang & Zheng, 2022). While the application of convolutional neural networks (CNNs) (LeCun et al., 1989) — traditionally employed in the visual domain — to audio signals, particularly in conjunction with spectrograms (Espi et al., 2015; Han et al., 2017; Thomas et al., 2014) or other feature engineering techniques (Abdel-Hamid et al., 2012; Sainath et al., 2013), is not novel. However, these approaches are continually being refined and adapted to the audio realm. Building on these foundations, this study developed an audio localization method by constructing a spectrogram-based multi-task CNN model.

Inspired by the fact, that humans tend to localize sound sources in groups rather than individually (Bregman, 1990; Rumsey, 2002), the objective of the proposed model is to predict ensemble location and width instead of positions of individual sources. This study is unique as it not only conceptualized the method but also tested it on a vast, real-life music corpus of 23,040 binaural excerpts that were synthesized using 192 multi-track music recordings (Senior, 2023) and 30 publicly available head-related transfer functions (HRTFs) from various sources (see Table 1 in 5). The music recordings covered many different types of genres, including rock, jazz, pop, and classical music.

The findings demonstrate that this method is effective in accurately predicting the spatial characteristics of sound sources in near-real-world scenarios. Furthermore, this paper demonstrates an experimental framework that facilitates the objective measurement of a binaural localization technique, employing a large-scale dataset synthesized from real-world music signal (for applications of similar frameworks, see: Antoniuk and Zieliński, 2023; Zieliński et al., 2020, 2022a, 2022b). One of the key advantages of the proposed method is that it does not assume the number of audio sources. However, significant limitations of this study include the absence of reverberation in the synthesized recordings and the method's inapplicability to real-time scenarios — both are critical areas for future research.

The developed method has the potential to be highly beneficial in automated assessment tasks, where a significant number of binaural recordings must be measured or labeled in terms of their spatial content information. This could be utilized in the development of a hypothetical autonomous web-crawler bot that will collect binaural recordings from publicly accessible repositories and label them according to the spatial properties of the sound sources, such as the location of the music ensemble or the sparsity of audio source positions. This method may also assist audio engineers in objectively assessing and segregating binaural audio recordings with regard to their spatial content.

The structure of this paper is organized as follows. Section 2 presents related studies. The description of the method developed for this study is provided in Section 3, which also includes detailed definitions of ensemble location and width, along with a description of the experiments used to evaluate this method. Section 4 presents and discusses the performance of the presented method and the results of the experiments conducted in this study. The paper concludes in Section 5.

2 Related studies

The majority of existing literature on the subject of sound source localization employs techniques that leverage the advantages of microphone arrays with more than two channels (Chung et al., 2022; Hahmann et al., 2022; Kaveh & Barabell, 1986; M. Liu et al., 2022; Pan et al., 2021; Pavlidi et al., 2012). These method simplify the task and can improve the localization performance; however, they do not utilize binaural hearing, rendering them ineffective for binaural recordings.

In the context of sound source localization in binaural signals, the focus of research is on the identification of individual sound sources, rather than groups of sounds (Benaroya et al., 2018; Dietz et al., 2011; Ma & Brown, 2016; Ma et al., 2017; May et al., 2011, 2012, 2015; Woodruff & Wang, 2012). In terms of source direction of arrival (DOA) methods, the majority of research assumes a fixed number of sound sources (Arthi & Sreenivas, 2021; Ma et al., 2017; Pang et al., 2019; Vera-Diaz et al., 2018; Woodruff & Wang, 2012), which limits its practical applications as this information is rarely known in real-life binaural recordings. Moreover, the majority of studies have focused on relatively homogeneous signals, namely speech (Benaroya et al., 2018; Dietz et al., 2011; Q. Liu et al., 2018; Ma & Brown, 2016; Ma et al., 2017, 2018; May et al., 2011, 2012, 2015; Wang et al., 2020; Woodruff & Wang, 2012; Yang & Zheng, 2022).

In contrast to the aforementioned studies, the proposed method is not constrained by the number of sources. Furthermore, the approach is not limited to speech and has been applied to a wide range of musical datasets, including instruments and vocals. In contrast to previous studies that primarily focused on individual sources, the proposed method does not aim to separate them, but rather considers them as a group, or in this case, a musical ensemble. This approach is similar to how the real musical ensembles are arranged on stage, thereby emphasizing practical applications in live settings. To the authors' knowledge, this is one of the first methods to localize ensemble width (see Antoniuk and Zieliński, 2023 for the previous ensemble-width-related study), and the first to localize both ensemble position and width simultaneously using a multi-task model.

Sound localization methods can be classified into two categories based on the implementation of their underlying algorithms: glass-box (e.g., (Dietz et al., 2011; Ma & Brown, 2016; Ma et al., 2017, 2018; May et al., 2011, 2012, 2015; Woodruff & Wang, 2012)) and black-box (e.g., (Vera-Diaz et al., 2018; Yang & Zheng, 2022)). Glass-box methods, more traditional in the literature, rely on manually designed algorithms that mimic the auditory system to explicitly extract key features for location prediction, such as interaural level differences, interaural time differences, interaural coherence, or interaural phase differences (see Blauert, 1996 for feature descriptions). An example of one of the most advanced auditory models that is able to extract such features was developed by the Two!Ears project (Raake, n.d.).

In contrast, black-box methods generally employ minimal feature engineering, relying instead on advanced machine learning techniques to autonomously extract features and make predictions. These methods, while effective, do not explicitly reveal the features they extract nor necessarily mimic the human auditory system. This opacity and the unpredictable nature of the outcomes, coupled with their dependency on deep neural networks that have extensive learning parameters, require the use of very large datasets for development and evaluation. Such datasets often include thousands of examples, as seen in the TIMIT corpus with 6300 examples (Garofolo et al., 1993) used in multiple studies (Benaroya et al., 2018; Ma et al., 2017, 2018; May et al., 2015; Pang et al., 2019; Vera-Diaz et al., 2018; Wang et al., 2020; Yang & Zheng, 2022), or custom corpora comprising hundreds of thousands of recordings (Antoniuk & Zieliński, 2023; Zieliński et al., 2020, 2022a, 2022b). This poses a significant challenge in assembling a sufficiently large and diverse collection of labeled binaural recordings. However, this challenge can be addressed through the synthesis of binaural sounds, as demonstrated in various studies (Antoniuk & Zieliński, 2023; Ma et al., 2018; Yang & Zheng, 2022; Zieliński et al., 2020, 2022a, 2022b) and further elaborated upon in Section 3.2 of this paper.

3 Methodology

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

3.1 Ensemble location and width definition

The objective of the model developed in this study is to predict the ensemble location (θ) and width (ω), as illustrated in Figure 1. An ensemble is defined as a group of audio point sources positioned on a circle around the listener on a virtual acoustic scene with equal distance to the listener. The location of source i is denoted by θ_i. The ensemble width (ω) is defined as the angular width between two extreme point sources ($max_i(\theta_i) - min_i(\theta_i)$), while the ensemble location, designated by θ, represents the middle angle between two extreme sound sources ($(max_i(\theta_i) + min_i(\theta_i))/2$). For the purposes of this study, the locations of the sources were limited to the frontal hemisphere only, i.e. $\theta \in [-45°, 45°]$, $\omega \in [0°, 90°]$, as this range encompasses the majority of real-world recording scenarios. It should be noted that although humans possess some limited abilities to localize sound sources in the vertical plane, in this study all sources are placed in the horizontal plane, at the height of the listener. This covers the majority of cases for real-world recordings (see Ma et al., 2018; Zieliński et al., 2022a for related studies that cover top-down discrimination).

3.2 Synthesis of binaural music recordings

The experiments conducted in this study involved 23,040 binaural recordings of music. The binaural recordings were synthesized using 192 multi-track publicly-available music recordings (Senior, 2023) and 30 HRTF databases (see Table 1 in Appendix A for a detailed list of HRTF databases used in this study). The number of tracks in multi-track recording ranged from 5 to 62, with median of 9. For each pair of a multi-track recording and an HRTF database, four binaural recordings were synthesized with different random ensemble parameters, namely location θ and width ω, as defined in Section 3.1. Both parameters were drawn from a uniform random distribution. Furthermore, the tracks of the input multi-track recordings were randomly assigned to sound source positions (θ_i) to enhance the diversity of the final binaural corpora. Before the synthesis, the tracks were equalized to -23 LUFS, in accordance with the ("ITU-R BS.1770-5", 2023) recommendation.

The binaural recordings were obtained in this study using the binaural synthesis procedure, known as binauralization, whose aim was to simulate the positions of sound sources within a virtual acoustic environment (Blauert, 1996). This was achieved by convolving multi-track signals with head-related impulse responses from a specified head-related transfer function (HRTF) database. The resulting binaural output signal $y_c[n]$ for each stereo channel c (left or right) at sample n is given by the following equation:

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

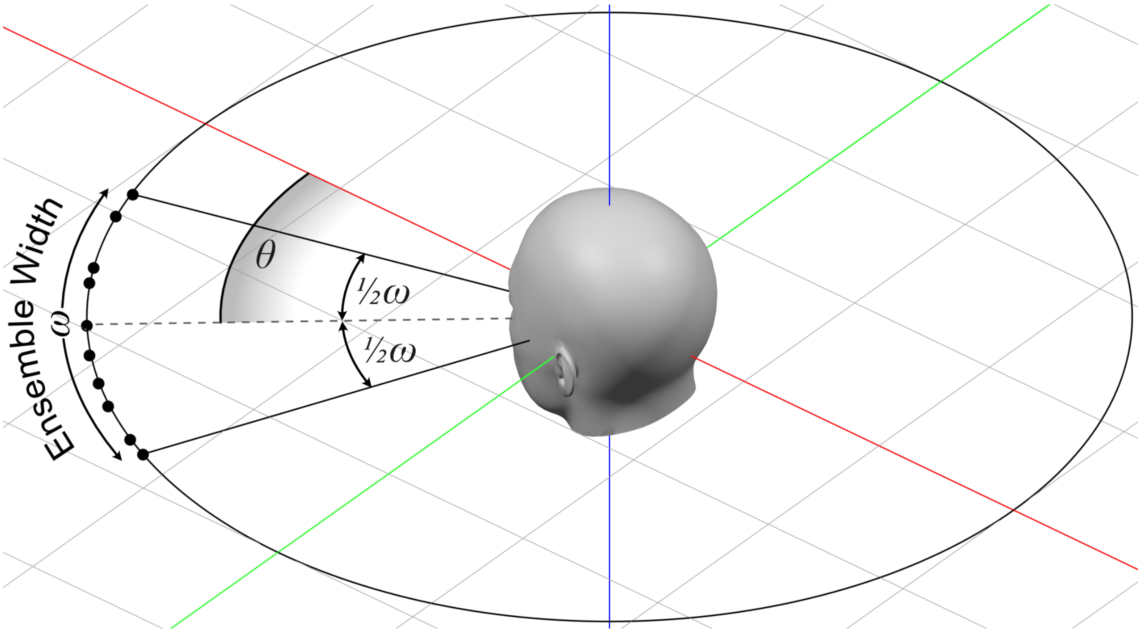

Figure 1: Illustration of ensemble width (ω) and ensemble location (θ) that are relative to the direction of the head. Black dots represent the positions of audio sources θ_i. The ensemble location (θ) is defined as the angular position of the center of the ensemble relative to the direction the head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

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

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

3.3 Feature extraction

Prior to input into the model, the binaural recordings of music were transformed into magnitude spectrograms. Although spectrograms do not directly provide information that can be translated into ensemble features, especially ensemble width, the goal of this task was to reduce the number of independent variables compared to the raw audio signal by extracting more compressed and informative data in the frequency domain. This step was necessary to decrease the likelihood of overfitting, reduce the number of examples needed to train the model, and thereby lower the overall computational power requirements. It is worth mentioning, however, that recent methods have shown that CNNs can be suitable for end-to-end audio localization without the spectrogram extraction step (Vecchiotti et al., 2019; Vera-Diaz et al., 2018).

To prepare the input for the model, a Hamming window of 40 ms was applied to each frame of the signal, resulting in a total of 349 time frames. From each frame, spectrograms were extracted using the Fast Fourier Transform (FFT) algorithm, with 150 frequency bands spaced linearly from 100 Hz to 16 kHz. This procedure was conducted for both the left and right channels, yielding two spectrograms for each binaural sample. Consequently, each sample was represented by a floating-point precision matrix of dimensions $2 \times 349 \times 150$. This method parallels the procedure presented by Zieliński et al., 2022b.

3.4 Network topology

The network topology employed in this study was heavily influenced by the AlexNet convolutional neural network introduced by Krizhevsky et al., 2012. While AlexNet was originally designed for image classification, it was adapted for the audio prediction task in this study by converting binaural recordings into magnitude spectrograms, as described in Section 3.3. This conversion allowed the spectrograms to be treated like images, allowing them to be used in a standard image-recognition-like task.

As illustrated in Figure 2, the network architecture consists of an input layer accepting a pair of spectrograms, followed by a series of convolutional units and classification units, culminating in

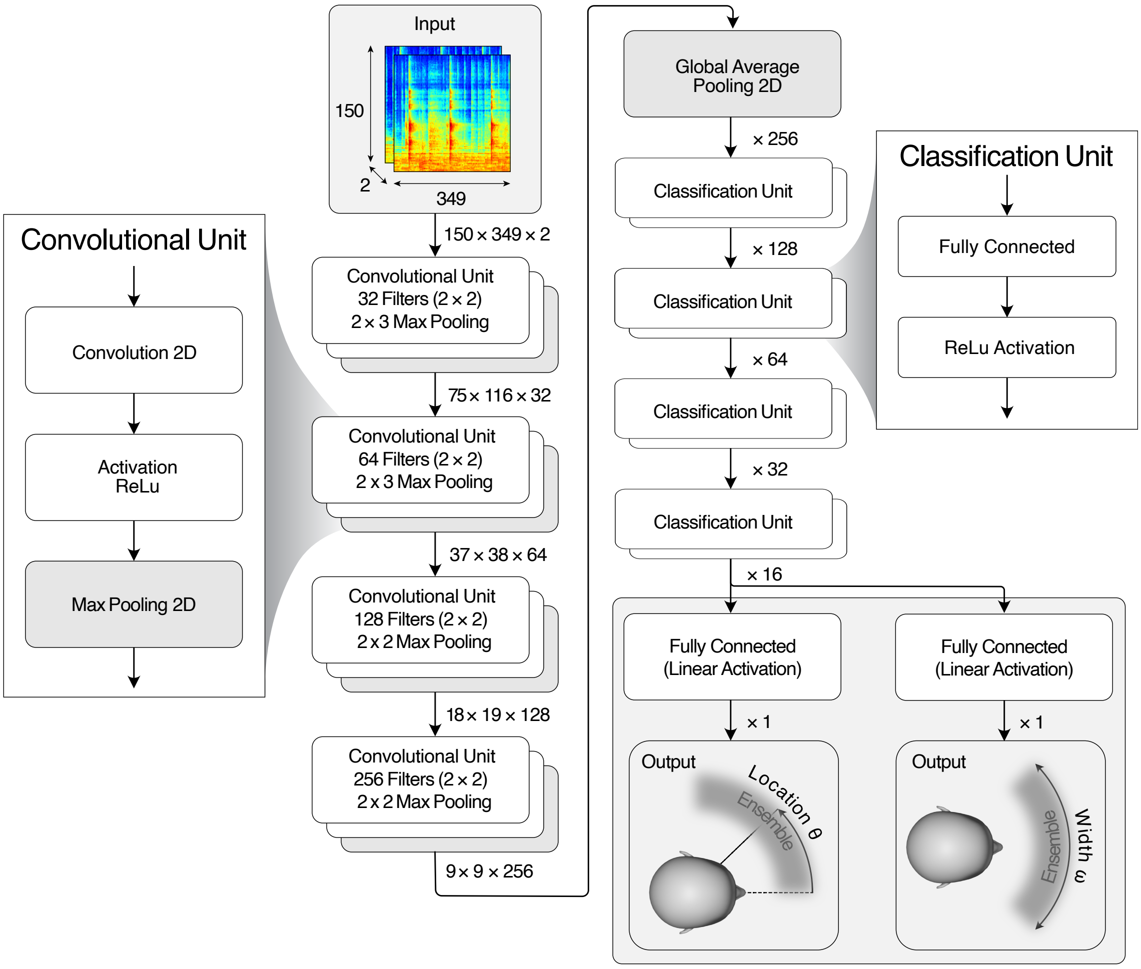

Figure 2: Topology of the Convolutional Neural Network (CNN) used for identifying ensemble location and width

two output heads for predicting ensemble width and location. This design employs a multi-task approach, enabling a single network to predict both ensemble parameters simultaneously.

The topology finalized in this study was constructed by selecting from many alternative architectures, choosing the one with the highest prediction quality observed on the validation dataset. Despite the existence of numerous algorithms for automatic topology selection (Branke, 1995; Miikkulainen et al., 2017; Shafiee et al., 2016; Stanley & Miikkulainen, 2002; H. Zhang et al., 2018), the final topology was determined manually, primarily due to the high computational demands relative to the available resources.

Various architectural configurations were assessed, with key parameters being varied such as the number of convolutional layers (from 1 to 5), the number of classification units (from 1 to 5), the inclusion or exclusion of Max Pooling Layers after each Convolution Layer, the number of filters within the Convolution Layers, the dimensions of these filters (2 × 2, 2 × 3, or 3 × 3), the stride size, and the dimensions of the Max Pooling Layers (2 × 2, 2 × 3, or 3 × 3). Based on this, it was concluded that the model is robust against variations in the assessed topologies. The differences in mean prediction error among the configurations were minimal, typically less than 1° for most configurations. Among the many tested topologies that yielded similar errors, the simplest one was selected to optimize both performance efficiency and model simplicity.

Despite the availability of widely used techniques for addressing overfitting, such as the Dropout Layer (Srivastava et al., 2014), and for accelerating training, such as Batch Normalization (Ioffe & Szegedy, 2015), neither technique was employed in this study as they were observed to be ineffective for the specific prediction task being undertaken. Instead, a Global Average Pooling Layer was utilized, known for its capabilities in reducing overfitting (Lin et al., 2013). This was confirmed in this particular task, as the inclusion of this layer significantly reduced overfitting, lowering the final mean absolute error score by 0.83° (average across 10 trials) compared to configurations using a Flattening Layer.

3.5 Model training and evaluation

The topology described in the previous section resulted in a model with 216,562 learning parameters. This model training procedure was repeated 10 times, employing the Monte Carlo cross-validation method (Kuhn & Johnson, 2013). For each repetition, the entire dataset was randomly divided into two parts: a development set containing two-thirds of the dataset (15,360 recordings) for model construction, and a test set consisting of the remaining one-third (7,680 recordings) for evaluation.

To ensure that the evaluation process was unbiased, this division was done in such a way that no original multi-track recordings used for synthesis were included in both the development and test sets simultaneously. However, this rule was not applied to HRTFs databases, allowing for the possibility of HRTF information leaking between the development and test sets. This could be seen as a significant limitation of the study. However, it is known that a human auditory system uses a single HRTF represented by ears and head, only slightly changing throughout the entire life, mainly during infancy (Clifton et al., 1988; King et al., 2001), so this limitation could be considered in pair how the human auditory system behaves in the real life. Conversely, there are studies that do not implement such a limitation and test binaural localization models in an HRTF-independent manner (Antoniuk & Zieliński, 2023; Zieliński et al., 2022a, 2022b).

The development set was divided into training and validation subsets at a 7:1 ratio, with 13,440 recordings in the training subset and 1,920 recordings in the validation subset. The training subset was used to update the model's learning parameters, while the validation subset was solely used for early stopping (Morgan & Bourlard, 1989; Pocock & Hughes, 1989) and model checkpointing (Eisenman et al., 2020), which were used for selecting the model with the best generalization capabilities and preventing overfitting. The test subset, which included data not seen during the training or validation phases, was used solely for performance assessment once per a repetition. This divide-train-and-evaluate process was repeated to collect 10 mean absolute errors, from which the final model error was determined.

For each sample, the model received two spectrograms as input: one for the left channel and one for the right channel. The rationale behind the application of Convolutional Neural Networks (CNNs) to this task was to automatically extract local features from the spectrograms and use these features to predict two contiguous ensemble parameters: ensemble location and width, both measured in degrees. The prediction errors, calculated as the difference between the actual ensemble parameters (known a priori from the binaural synthesis described in Section 3.2) and the predicted values, were used in the Adam optimization algorithm (Kingma & Ba, 2014). During training, the losses for both outputs were combined by summing them, thereby treating both ensemble features with equal importance.

The computational work for this study was conducted on a workstation equipped with an RTX Nvidia GeForce 4090 GPU and a 48-core AMD Ryzen ThreadRipper processor. On the software side, MATLAB (The MathWorks Inc., 2022b) with the Audio Toolbox (The MathWorks Inc., 2022a) was used for the binaural recording synthesis, while Python (Van Rossum & Drake, 2009) with the SciPy package (Virtanen et al., 2020) was used for feature extraction and Keras (Chollet et al., 2015) for training the CNN model. The complete source code for all stages is publicly available on the GitHub repository (Antoniuk, 2024). The feature extraction phase required 21 minutes, data partitioning took 34 minutes, and the total training time for all iterations was 40 minutes, making the entire training and evaluation process 95 minutes long.

4 Results & discussion

The overall model performance measured across 10 experiment iterations, expressed as mean absolute error (MAE), was equal to 8.57° (±0.19°) for ensemble width and 4.76° (±0.10°) ensemble location. As both ensemble parameters were constrained within the same range of 90°, the results demonstrate that the model exhibits a 44% higher accuracy for ensemble location estimation than for ensemble width. This outcome is not unexpected, given that ensemble location is a less complex parameter. Essentially, it represents the average location of all sources. Therefore, it is more resistant to temporal fluctuations in individual audio sources than ensemble width, which is dependent on the two most extreme sound sources that vary over time. Furthermore, predicting ensemble width necessitates the identification of these two extreme sources, a process that is inherently more complex than estimating a single average location.

Figures 3 and 4 compare the actual and predicted ensemble widths. The results indicate that the model provides more accurate predictions for narrower ensemble widths, with an average MAE of 5.65° for $\omega < 30°$. However, performance deteriorates as the ensemble width increases, resulting in

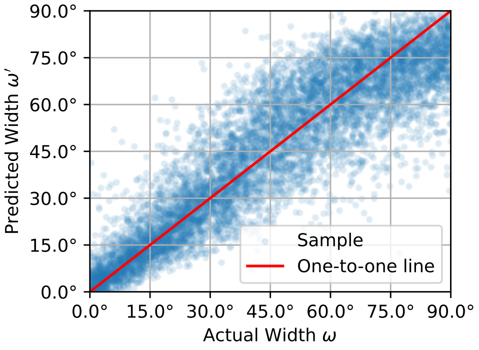

Figure 3: A comparison between the actual ensemble width ω and the predicted ensemble width ω' for a single iteration (of the total five)

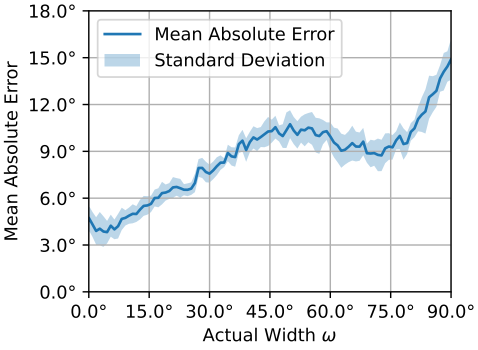

Figure 4: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

a MAE of 12.44° for $\omega > 80°$. This deterioration highlights the significant impact of ensemble width on the accuracy of width estimation, particularly for wider ranges. In contrast, when comparing predicted locations to actual ones, the error remains relatively stable, as shown in Figures 5 and 6. The deterioration in the width prediction can be attributed to the sparse distribution of audio sources in wider ensembles, which amplifies the influence of extreme sound sources on prediction errors, resulting in greater inaccuracies as the ensemble width increases. Moreover, Figure 4 reveals that the relationship between the ensemble width and the error is nonlinear, displaying a notable decrease in error between 60° and 75°. The reason for this nonlinearity is currently unclear and requires further investigation.

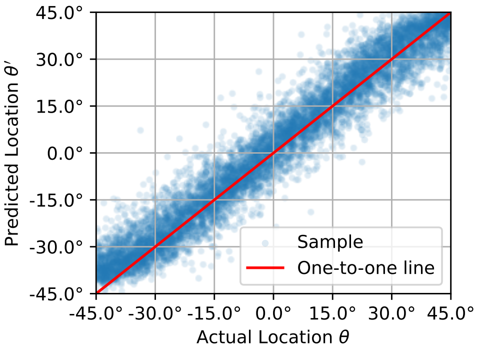

Figure 5: A comparison between the actual ensemble location θ and the predicted ensemble location θ' for a single iteration (of the total five)

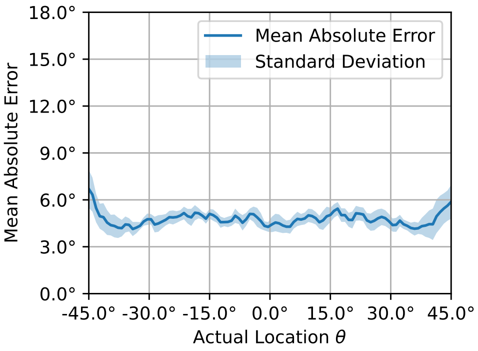

Figure 6: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

In contrast to the correlation between ensemble width and its prediction error, there is no significant relationship between the actual location and its prediction error, as illustrated in Figures 5 and 6. This finding indicates that the model's capabilities for localizing the center of the ensemble is more robust, unaffected by the actual spatial positioning of the ensemble, including lateral locations.

Figure 7 illustrates the influence of both ensemble location and width on the mean absolute error for ensemble location, providing a detailed perspective complementing the results presented in Figure 6. Notably, the figure highlights some asymmetric anomalies, especially within the $\theta \in [15°, 30°]$ range compared to the $\theta \in [-30°, -15°]$ range. These anomalies can be attributed to the sparsity of sample result data across this heatmap. While the figure indicates that ensemble location does not significantly affect the model's location prediction accuracy, it clearly shows that ensemble width has a substantial impact. Similarly, Figure 8 reveals a characteristic depression in $\omega \in [30°, 60°]$ previously shown from a different perspective in Figure 4. This heatmap highlights another interesting phenomenon in its upper corners —– the error in these areas is considerably

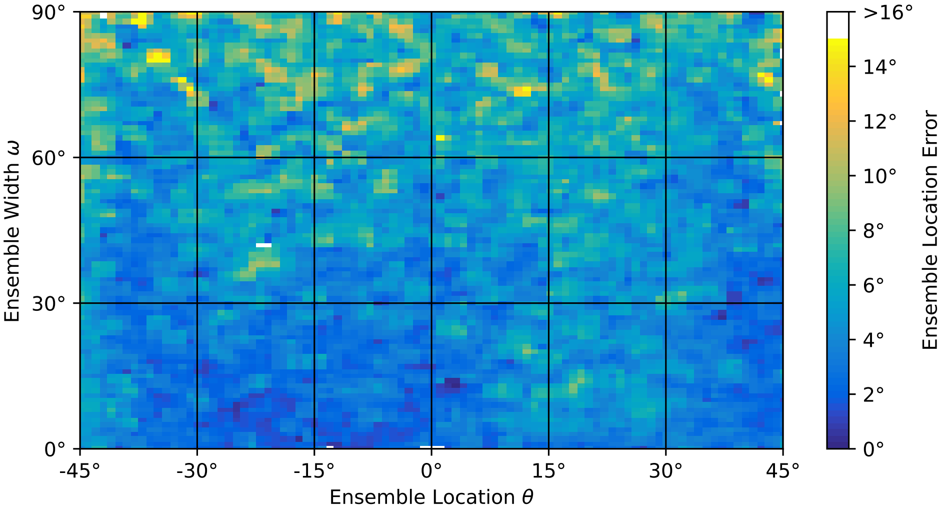

Figure 7: The heatmap that illustrates the mean absolute error (MAE) of ensemble location distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors.

higher. This indicates that the model's performance for estimating ensemble width is substantially worse at extreme widths and locations, i.e., when both the width and locations are at their maximum.

The achieved model's performance was better to the model based on Spatiograms (Arthi & Sreenivas, 2021), which performance was further investigated by Antoniuk and Zieliński, 2023 within the same evaluation scenario, using the same dataset, thus making it possible to compare these two results directly. The Spatiogram-based model achieved result of 13.62°, making the model presented in this study better by 5.05°, suggesting a significant improvement in the prediction of this ensemble parameter. Furthermore, the model presented in the previous study did not predict the ensemble location, which further enhances the effectiveness of this method.

5 Conclusions

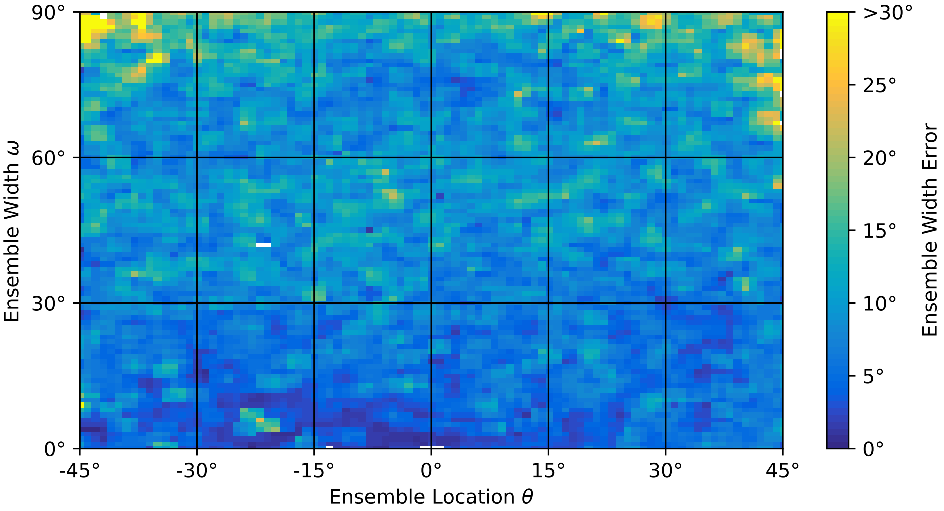

Figure 8: The heatmap that illustrates the mean absolute error (MAE) of ensemble width distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors. Notably, the values between 30°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

Appendix A

Table 1: 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 (Braren & Fels, 2020)	AACHEN
2.	Artificial	GRAS 45BB-4 KEMAR	1		
3.	Human	Subject 2	1.2	Austrian Academy of Sciences (“HRTF-Database”, n.d.)	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 (Algazi et al., 2001)	CIPIC
8.	Human	Subject 015	1		
9.	Human	Subject 020	1		
10.	Artificial	Neumann KU 100	0.9	NASA (2007) (Andreopoulou et al., 2015)	CLUBFRITZ
11.	Artificial	Neumann KU 100	1.5	Helsinki University of Technology (2009) (Andreopoulou et al., 2015)	
12.	Artificial	FABIAN	1.47	Technical University Berlin, Huawei Technologies, Munich Research Centre, Sennheiser Electronic (Brinkmann et al., 2019)	HUTUBS
13.	Human	Subject pp2	1.47		
14.	Human	Subject pp3	1.47		
15.	Human	Subject 1003	1.95	IRCAM, AKG (“LISTEN HRTF DATABASE”, 2023)	LISTEN
16.	Human	Subject 1002	1.95		
17.	Artificial	KEMAR DB-4004 (DB-061)	1.4	MIT (Gardne & Martin, 1994)	MIT
18.	Artificial	KEMAR DB-4004 (DB-065)	1.4		
19.	Human	Subject 001	1.5	Tohoku University (Watanabe et al., 2014)	RIEC
20.	Human	Subject 002	1.5		
21.	Artificial	Koken SAMRAI	1.5		
22.	Artificial	Neumann KU 100	1.2	University of York (Armstrong et al., 2018)	SADIE II
23.	Human	Subject H3	1.2		
24.	Human	Subject H4	1.2		
25.	Artificial	KEMAR	1	South China University of Technology (Yu et al., 2018)	SSCUT

Continued on next page

Table 1: List of HRTF sets used to synthesize binaural audio excerpts (Continued)

26.	Artificial	Neumann KU 100	1	TH Köln (Pörschmann et al., 2017)	STH Köln
27.	Artificial	FABIAN	1.7	TU Berlin (Brinkmann et al., 2017; Wierstorf et al., 2011)	TU Berlin
28.	Artificial	GRAS 45BA KEMAR	1		
29.	Artificial	GRAS 45BB-4 KEMAR - subject A attachment	1	Aalborg University; University of Iceland (Spagnol, 2019; Spagnol et al., 2020)	VIKING
30.	Artificial	GRAS 45BB-4 KEMAR - subject B attachments	1		

References

1. Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4277–4280. https://doi.org/10.1109/ICASSP.2012.6288864

2. Algazi, V., Duda, R., Thompson, D., & Avendano, C. (2001). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, 99–102. https://doi.org/10.1109/ASPAA.2001.969552

3. Andreopoulou, A., Begault, D. R., & Katz, B. F. G. (2015). Inter-laboratory round robin HRTF measurement comparison. *IEEE Journal of Selected Topics in Signal Processing*, *9*(5), 895–906. https://doi.org/10.1109/JSTSP.2015.2400417

4. Antoniuk, P. (2024). *Software repository: Predicting ensemble location and width in binaural recordings of music with convolutional neural networks.*

5. Antoniuk, P., & Zieliński, S. K. (2023). Blind estimation of ensemble width in binaural music recordings using 'spatiograms' under simulated anechoic conditions. *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio.*

6. Armstrong, C., Thresh, L., Murphy, D., & Kearney, G. (2018). A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database. *Applied Sciences*, *8*(11), 2029. https://doi.org/10.3390/app8112029

7. Arthi, S., & Sreenivas, T. V. (2021). Spatiogram: A phase based directional angular measure and perceptual weighting for ensemble source width. *ArXiv*, *abs/2112.07216*.

8. Benaroya, E. L., Obin, N., Liuni, M., Roebel, A., Raumel, W., & Argentieri, S. (2018). Binaural Localization of Multiple Sound Sources by Non-Negative Tensor Factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(6), 1072–1082. https://doi.org/10.1109/TASLP.2018.2806745

9. Blauert, J. (1996, October). *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press. https://doi.org/10.7551/mitpress/6391.001.0001

10. Branke, J. (1995). Evolutionary algorithms for neural network design and training. *Proceedings of the First Nordic Workshop on Genetic Algorithms and its Application*, 145–163.

11. Braren, H. S., & Fels, J. (2020). A high-resolution individual 3d adult head and torso model for HRTF simulation and validation: HRTF measurement.

12. Bregman, A. (1990, January). Auditory Scene Analysis: The Perceptual Organization of Sound. In *Journal of The Acoustical Society of America - J ACOUST SOC AMER* (Vol. 95). MIT Press. https://doi.org/10.1121/1.408434

13. Brinkmann, F., Dinakaran, M., Pelzer, R., Grosche, P., Voss, D., & Weinzierl, S. (2019). A cross-evaluated database of measured and simulated HRTFs including 3d head meshes, anthropometric features, and headphone impulse responses. *Journal of the Audio Engineering Society*, *67*(9), 705–718. https://doi.org/10.17743/jaes.2019.0024

14. Brinkmann, F., Lindau, A., Weinzerl, S., Van De Par, S., Müller-Trapet, M., Opdam, R., & Vorländer, M. (2017). A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations. *Journal of the Audio Engineering Society*, *65*(10), 841–848. https://doi.org/10.17743/jaes.2017.0033

15. Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*, 975–979. https://doi.org/10.1121/1.1907229

16. Chollet, F., et al. (2015). Keras.

17. Chung, M.-A., Chou, H.-C., & Lin, C.-W. (2022). Sound localization based on acoustic source using multiple microphone array in an indoor environment. *Electronics*, *11*(6), 890. https://doi.org/10.3390/electronics11060890

18. Clifton, R. K., Gwiazda, J., Bauer, J. A., Clarkson, M. G., & Held, R. M. (1988). Growth in head size during infancy: Implications for sound localization. *Developmental Psychology*, *24*(4), 477–483. https://doi.org/10.1037/0012-1649.24.4.477

19. Dietz, M., Ewert, S. D., & Hohmann, V. (2011). Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, *53*(5), 592–605. https://doi.org/10.1016/j.specom.2010.05.006

20. Eisenman, A., Matam, K. K., Ingram, S., Mudigere, D., Krishnamoorthi, R., Annavaram, M., Nair, K., & Smelyanskiy, M. (2020). Check-n-run: A checkpointing system for training recommendation models. *ArXiv*, *abs/2010.08679*.

21. Espi, M., Fujimoto, M., Kinoshita, K., & Nakatani, T. (2015). Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1), 26. https://doi.org/10.1186/s13636-015-0069-2

22. Gardne, B., & Martin, K. (1994). *HRTF measurements of a KEMAR dummy-head microphone*. Retrieved June 19, 2024, from https://sound.media.mit.edu/resources/KEMAR.html

23. Garofolo, J. S., Lamel, L., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). DARPA TIMIT:: Acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1.

24. Hahmann, M., Fernandez-Grande, E., Gunawan, H., & Gerstoft, P. (2022). Sound source localization using multiple ad hoc distributed microphone arrays. *JASA express letters*, *2*(7), 074801. https://doi.org/10.1121/10.0011811

25. Han, Y., Park, J., & Lee, K. (2017). Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification. *Workshop on Detection and Classification of Acoustic Scenes and Events*.

26. Hirsh, I. J. (1950). Binaural hearing aids: A review of some experiments. *Journal of Speech and Hearing Disorders*, *15*(2), 114–123. https://doi.org/10.1044/jshd.1502.114

27. *HRTF-database* [Austrian academy of sciences]. (n.d.). Retrieved June 19, 2024, from https://www.oeaw.ac.at/en/ari/das-institut/software/hrtf-database

28. Ioffe, S., & Szegedy, C. (2015, July 7). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 448–456, Vol. 37). PMLR.

29. ITU-R BS.1770-5: Algorithms to measure audio programme loudness and true-peak audio level. (2023).

30. Kaveh, M., & Barabell, A. (1986). The statistical performance of the MUSIC and the minimum-norm algorithms in resolving plane waves in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *34*(2), 331–341. https://doi.org/10.1109/TASSP.1986.1164815

31. King, A. J., Kacelnik, O., Mrsic-Flogel, T. D., Schnupp, J. W., Parsons, C. H., & Moore, D. R. (2001). How plastic is spatial hearing? *Audiology and Neurotology*, *6*(4), 182–186. https://doi.org/10.1159/000046829

32. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

33. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

34. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

35. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, *2*.

36. Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *CoRR*, *abs/1312.4400*.

37. *LISTEN HRTF DATABASE*. (2023). Retrieved June 19, 2024, from http://recherche.ircam.fr/equipes/salles/listen/

38. Liu, M., Hu, J., Zeng, Q., Jian, Z., & Nie, L. (2022). Sound source localization based on multi-channel cross-correlation weighted beamforming. *Micromachines*, *13*(7), 1010. https://doi.org/10.3390/mi13071010

39. Liu, Q., Wang, W., de Campos, T., Jackson, P. J. B., & Hilton, A. (2018). Multiple speaker tracking in spatial audio via PHD filtering and depth-audio fusion. *IEEE Transactions on Multimedia*, *20*(7), 1767–1780. https://doi.org/10.1109/TMM.2017.2777671

40. Ma, N., & Brown, G. J. (2016). Speech Localisation in a Multitalker Mixture by Humans and Machines. *Proc. Interspeech 2016*, 3359–3363. https://doi.org/10.21437/Interspeech.2016-1149

41. Ma, N., Gonzalez, J. A., & Brown, G. J. (2018). Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(11), 2122–2131. https://doi.org/10.1109/TASLP.2018.2855960

42. Ma, N., May, T., & Brown, G. J. (2017). Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localisation of Multiple Sources in Reverberant Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2444–2453. https://doi.org/10.1109/TASLP.2017.2750760
Comment: 10 pages.

43. May, T., Ma, N., & Brown, G. J. (2015). Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2679–2683. https://doi.org/10.1109/ICASSP.2015.7178457

44. May, T., van de Par, S., & Kohlrausch, A. (2011). A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 1–13. https://doi.org/10.1109/TASL.2010.2042128

45. May, T., van de Par, S., & Kohlrausch, A. (2012). A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(7), 2016–2030. https://doi.org/10.1109/TASL.2012.2193391

46. Miikkulainen, R., Liang, J. Z., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B. E., Shahrzad, H., Navruzyan, A., Duffy, N. P., & Hodjat, B. (2017). Evolving deep neural networks. *ArXiv*, *abs/1703.00548*.

47. Morgan, N., & Bourlard, H. (1989). Generalization and parameter estimation in feedforward nets: Some experiments. *Advances in Neural Information Processing Systems*, *2*.

48. Pan, Z., Zhang, M., Wu, J., Wang, J., & Li, H. (2021). Multi-tone phase coding of interaural time difference for sound source localization with spiking neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2656–2670. https://doi.org/10.1109/TASLP.2021.3100684

49. Pang, C., Liu, H., & Li, X. (2019). Multitask Learning of Time-Frequency CNN for Sound Source Localization. *IEEE Access*, *7*, 40725–40737. https://doi.org/10.1109/ACCESS.2019.2905617

50. Pavlidi, D., Puigt, M., Griffin, A., & Mouchtaris, A. (2012). Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2625–2628. https://doi.org/10.1109/ICASSP.2012.6288455

51. Pocock, S. J., & Hughes, M. D. (1989). Practical problems in interim analyses, with particular regard to estimation. *Controlled Clinical Trials*, *10*(4), 209–221. https://doi.org/10.1016/0197-2456(89)90059-7

52. Pörschmann, C., Arend, J., & Neidhardt, A. (2017). A spherical near-field HRTF set for auralization and psychoacoustic research.

53. Raake, A. (n.d.). *A computational framework for modelling active exploratory listening that assigns meaning to auditory scenes—reading the world with two ears*. Retrieved June 11, 2024, from http://twoears.eu/

54. Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, *50*, 651–666.

55. Sainath, T. N., Mohamed, A.-r., Kingsbury, B., & Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8614–8618. https://doi.org/10.1109/ICASSP.2013.6639347

56. Senior, M. (2023). *The 'mixing secrets' free multitrack download library*. Retrieved June 10, 2024, from https://cambridge-mt.com/ms/mtk/

57. Shafiee, M. J., Mishra, A. K., & Wong, A. (2016). Deep learning with darwin: Evolutionary synthesis of deep neural networks. *Neural Processing Letters*, *48*, 603–613.

58. Spagnol, S. (2019). THE VIKING HRTF DATASET.

59. Spagnol, S., Miccini, R., & Unnthorsson, R. (2020). The viking HRTF dataset v2.

60. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

61. Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, *10*(2), 99–127. https://doi.org/10.1162/106365602320169811

62. The MathWorks Inc. (2022a). *Audio toolbox version: 9.13.0 (r2022b)*. Natick, Massachusetts, United States.

63. The MathWorks Inc. (2022b). *Matlab version: 9.13.0 (r2022b)*. Natick, Massachusetts, United States.

64. Thiemann, J., Müller, M., Marquardt, D., Doclo, S., & van de Par, S. (2016). Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 12. https://doi.org/10.1186/s13634-016-0314-6

65. Thomas, S., Ganapathy, S., Saon, G., & Soltau, H. (2014). Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2519–2523. https://doi.org/10.1109/ICASSP.2014.6854054

66. Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

67. Vecchiotti, P., Ma, N., Squartini, S., & Brown, G. J. (2019). End-to-end binaural sound localisation from the raw waveform. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 451–455.

68. Vera-Diaz, J. M., Pizarro, D., & Macias-Guarasa, J. (2018). Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors*, *18*(10), 3418. https://doi.org/10.3390/s18103418

69. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

70. Wang, J., Wang, J., Qian, K., Xie, X., & Kuang, J. (2020). Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. *EURASIP Journal on Audio, Speech, and Music Processing*, *2020*(1), 4. https://doi.org/10.1186/s13636-020-0171-y

71. Watanabe, K., Iwaya, Y., Suzuki, Y., Takane, S., & Sato, S. (2014). Dataset of head-related transfer functions measured with a circular loudspeaker array. *Acoustical Science and Technology*, *35*(3), 159–165. https://doi.org/10.1250/ast.35.159

72. Wierstorf, H., Geier, M., Raake, A., & Spors, S. (2011). A free database of head-related impulse response measurements in the horizontal plane with multiple distances.

73. Woodruff, J., & Wang, D. (2012). Binaural Localization of Multiple Sources in Reverberant and Noisy Environments. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(5), 1503–1512. https://doi.org/10.1109/TASL.2012.2183869

74. Yang, Q., & Zheng, Y. (2022). DeepEar: Sound Localization with Binaural Microphones. *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 960–969. https://doi.org/10.1109/INFOCOM48880.2022.9796850

75. Yu, G., Wu, R., Liu, Y., & Xie, B. (2018). Near-field head-related transfer-function measurement and database of human subjects. *The Journal of the Acoustical Society of America*, *143*(3), EL194–EL198. https://doi.org/10.1121/1.5027019

76. Zhang, H., Kiranyaz, S., & Gabbouj, M. (2018). Finding better topologies for deep convolutional neural networks by evolution. *ArXiv*, *abs/1809.03242*.

77. Zhang, W., Samarasinghe, P. N., Chen, H., & Abhayapala, T. D. (2017). Surround by Sound: A Review of Spatial Audio Recording and Reproduction. *Applied Sciences*, *7*(5), 532. https://doi.org/10.3390/app7050532

78. Zieliński, S. K., Antoniuk, P., & Lee, H. (2022a). Spatial audio scene characterization (SASC): Automatic localization of front-, back-, up-, and down-positioned music ensembles in binaural recordings. *Applied Sciences*, *12*(3), 1569. https://doi.org/10.3390/app12031569

79. Zieliński, S. K., Antoniuk, P., Lee, H., & Johnson, D. (2022b). Automatic discrimination between front and back ensemble locations in HRTF-convolved binaural recordings of music. *EURASIP Journal on Audio, Speech, and Music Processing*, *2022*(1), 3. https://doi.org/10.1186/s13636-021-00235-2

80. Zieliński, S. K., Lee, H., Antoniuk, P., & Dadan, O. (2020). A comparison of human against machine-classification of spatial audio scenes in binaural recordings of music. *Applied Sciences*, *10*(17), 5956. https://doi.org/10.3390/app10175956