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

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

[1] *Faculty of Computer Science, Białystok University of Technology*

Wiejska 45A, 15-351 Białystok, Poland

[*] *Corresponding Author Email: pawel.antoniuk@sd.pb.edu.pl*

Binaural audio technology has been in existence for many years. However, its popularity has significantly increased over the past decade as a consequence of advancements in virtual reality and streaming techniques. Along with its growing popularity, the quantity of publicly accessible binaural audio recordings has also expanded. Consequently, there is now a need for automated and objective retrieval of spatial content information, with ensemble location and width being the most prominent. This study presents a novel method for estimating these ensemble parameters in binaural recordings of music. For this purpose, a dataset of 23,040 binaural recordings was synthesized from 192 publicly-available music recordings using 30 head-related transfer functions. The synthesized excerpts were then used to train a multi-task spectrogram-based convolutional neural network model, aiming to estimate the ensemble location and width for unseen recordings. The results indicate that a model for estimating ensemble parameters 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 techniques recently published in the literature and shows promise for future development as part of a novel tool for binaural audio recordings analysis.

Keywords: ensemble width, ensemble location, binaural, spatial audio, localization, convolutional neural network, head-related transfer function, angle of arrival

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, among other factors, 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 humans' capability to concentrate on foreground sound sources while suppressing background noise (Cherry, 1953). Understanding of the auditory system is essential for comprehending its limits but also for leveraging these insights to create more immersive binaural experiences for entertainment purposes (Zhang *et al.*, 2017). It is also important for enhancing auditory signal reception in hearing aid devices (Hirsh, 1950; Thiemann *et al.*, 2016).

The advance of sophisticated machine learning techniques, especially deep learning networks, has initiated an interesting exploration of their potential to emulate the human auditory system. Recently emerged studies have demonstrated that relying on the advanced spatial audio feature engineering is not necessary in computational audio source localization (Pang *et al.*, 2019; Vera-Diaz *et al.*, 2018; Yang, Zheng, 2022). While applying convolutional neural networks (CNNs) (LeCun *et al.*, 1989) to audio signals is well-established, often 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), these approaches continue to be refined and adapted for audio processing. Building on these foundations, this study develops an audio localization method using a spectrogram-based multi-task CNN model.

Humans tend to localize groups of sound sources rather than individual ones (Bregman, 1990; Rumsey, 2002). Inspired by this fact, the objective of the proposed model is to estimate the location and width of these groups, termed "ensembles," instead of the positions of individual sources. This study is unique as it not only developed the method but also tested it on a relatively large, realistic music corpus. The corpus comprised 23,040 binaural excerpts synthesized using 192 multi-track music recordings (from a repository provided by Senior (2023)) and 30 sets of publicly available head-related transfer functions (HRTFs) acquired from various sources (see Table 1 in Appendix for a detailed list). The music recordings covered many different genres, including rock, jazz, pop, and classical music.

The findings demonstrate that this method is effective in accurately estimating the spatial characteristics of groups of sound sources in near-real-world scenarios. This paper also demonstrates an experimental framework that facilitates the objective measurement of a binaural localization technique, employing a large-scale dataset synthesized from real-world music signals (for applications of similar frameworks, see studies conducted by Antoniuk and Zieliński (2023) and 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 information retrieval tasks, where a significant number of binaural recordings must be analyzed 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.

This paper is structured 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 proposed method as well as the results of the experiments conducted in this study. Finally, Section 5 offers concluding remarks and suggestions for future research.

2 Related studies

Most existing literature on computational sound source localization reports techniques that take advantage of multiple microphone arrays with more than two channels (Chung *et al.*, 2022; Hahmann *et al.*, 2022; Kaveh, Barabell, 1986; Liu *et al.*, 2022; Pan *et al.*, 2021; Pavlidi *et al.*, 2012). Although these methods can improve localization precision by providing additional spatial information, 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 put 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).

Considering 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; 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. Moreover, the approach is not narrowed to speech and has been applied to a wide range of

musical datasets, including instruments and vocals. In contrast to 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 — similar to how real musical ensembles are arranged on stage. 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, termed as glass-box and black-box techniques. Glass-box methods could be considered as more traditional in the literature. They rely on manually designed algorithms that mimic the auditory system to explicitly extract key features for the localization estimation, such as interaural level differences, interaural time differences, interaural coherence, or interaural phase differences (Blauert (1996) provides detailed descriptions of these features). Examples of glass-box methods can be found in numerous studies, including those conducted by Dietz *et al.* (2011), Ma and Brown (2016), Ma *et al.* (2017, 2018), May *et al.* (2011, 2012, 2015), Woodruff and Wang (2012), and Zieliński *et al.* (2022b). These features are typically extracted using an auditory model. An advanced implementation capable of extracting these features was developed as part of the Two!Ears project (Raake, 2016).

Black-box methods use a minimal degree of feature engineering, depending on deep neural networks to both extract features and make estimations. While effective, these methods do not necessarily consistently mimic human hearing, rendering them less suitable for objective measurement tasks (e.g., Vera-Diaz *et al.* (2018) and Yang and Zheng (2022)). Additionally, it is challenging to reveal their internally extracted features. Due to their opacity, unpredictable results, and numerous learning parameters, these methods should be treated more carefully. Moreover, they require large datasets for their development and evaluation. These datasets often contain thousands of examples, such as the TIMIT corpus (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). Some researchers have even created custom corpora with hundreds of thousands of recordings (Antoniuk, Zieliński, 2023; Zieliński *et al.*, 2020, 2022a, 2022b).

The necessity of having a large corpus to train deep learning models poses a significant challenge in gathering 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 discussed further in Section 3.2.

3 Methodology

This part of the paper presents a detailed description of the model developed in 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 spectrograms calculation procedure is presented. Section 3.4 describes the model topology, whereas Section 3.5 addresses model training and evaluation.

3.1 Ensemble location and width definition

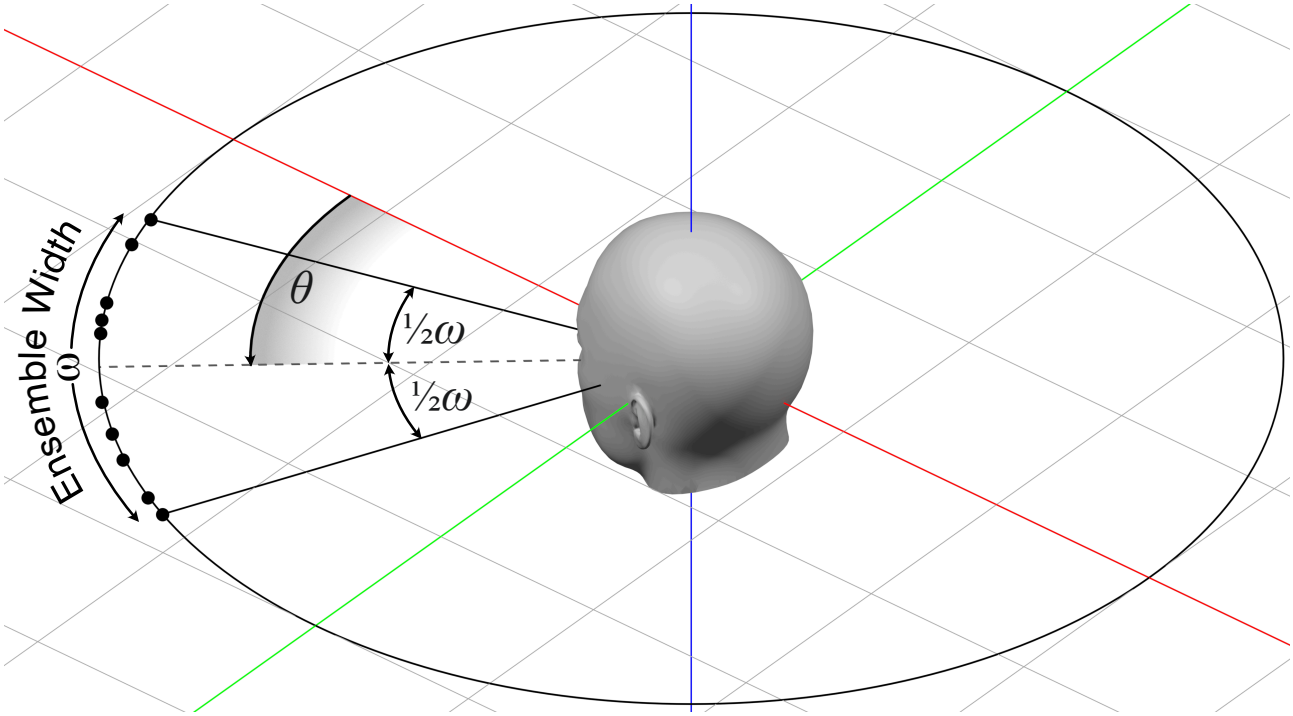

Figure 1: Illustration of ensemble width (ω) and ensemble location (θ) relative to the direction of the head orientation. Black dots represent the positions of audio sources θ_i. The ensemble location (θ) is the angular position of the center of the ensemble relative to the direction the head is facing. The ensemble width (ω) is the angular distance between the two most extreme audio sources in the ensemble.

The objective of the model developed in this study is to estimate 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 distance 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 ear-level of the listener. This covers the majority of cases for real-world recordings (see Ma *et al.* (2018) and 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. These recordings were synthesized using 192 publicly-available multi-track music recordings (Senior, 2023) and 30 HRTF databases (see Table 1 in Appendix for a detailed list). The large number of HRTF databases was necessary to make the model as generalizable as possible. In real-world scenarios, the HRTF used for binaural synthesis is often unknown, so constructing a model for a single HRTF would have limited practical utility. The aim was to predict ensemble parameters regardless of the specific HRTF function used. Additionally, the large number of HRTF functions increased the amount of data available for model training, which is particularly beneficial in the context of deep neural networks. The number of HRTF databases (30) was determined using heuristics from previous study conducted by Zieliński *et al.* (2022b), which suggested this number should be sufficient for the task.

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 signals in each track were equalized to -23 LKFS, 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:

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

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.

After the binauralization procedure, the synthesized recordings were truncated to a duration of seven seconds, with sine-squared fade-in and fade-out effects of 0.01 seconds applied. The recordings

were then RMS-normalized, scaled by a factor of 0.9, and DC-offset corrected. They were stored as uncompressed files at a 48 kHz sampling rate and with a 32-bit resolution.

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

3.3 Calculation of spectrograms

Prior being 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 also necessary to decrease the likelihood of overfitting, reducing the number of examples needed to train the model, and thereby lower the overall computational power requirements. It is worth mentioning, however, that recently published studies have shown that CNNs are suitable for end-to-end audio localization without the spectrogram extraction step, as demonstrated by Vecchiotti *et al.* (2019) and Vera-Diaz *et al.* (2018).

To prepare the input for the model, a Hamming window of 40 ms with an overlap of 20 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 32-bit 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 strongly influenced by the AlexNet convolutional neural network introduced by Krizhevsky *et al.* (2012). While AlexNet was originally designed for image classification, in this study it was adapted for the audio analysis task by converting binaural recordings into magnitude spectrograms, as described in Section 3.3. This conversion allowed the spectrograms to be treated like visual data, enabling them to be used in an 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 two outputs predicting ensemble location and width, respectively. This design employs a multi-task

approach, enabling a single network to estimate both ensemble parameters simultaneously.

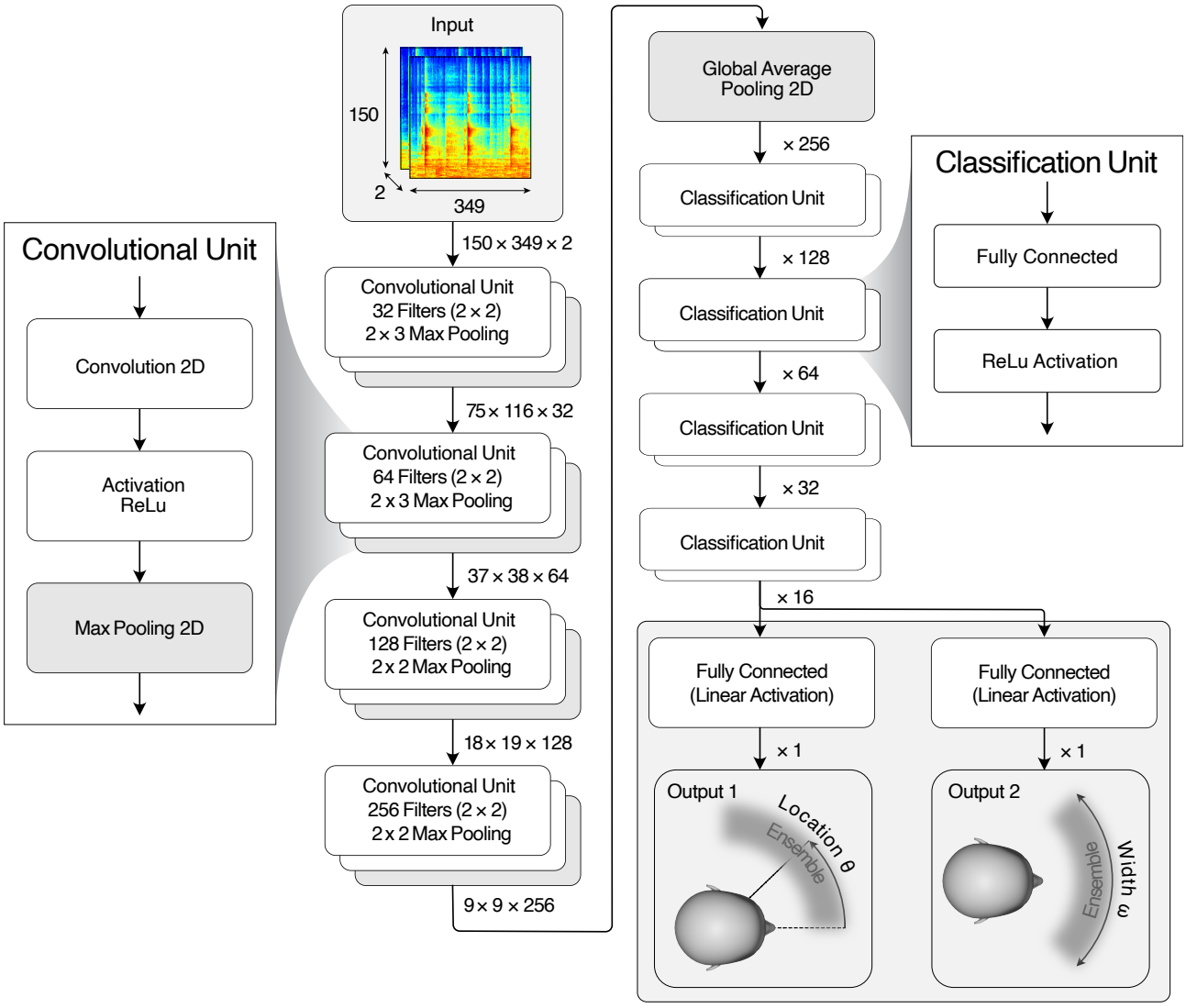

Figure 2: Topology of the Convolutional Neural Network (CNN) used for estimating ensemble location and width, illustrating the layers (grouped in "convolutional" and "classification" units) and connections of the network architecture.

The topology finalized in this study was chosen, among many alternative architectures, based on 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; 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 units (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 convolutional 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 procedure, 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 an overfitting effect, 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 estimation 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 (MAE) score by 0.83° (average across 10 trials) compared to configurations where a simple flattening layer was used instead.

3.5 Model training and evaluation

The topology described in the previous section resulted in a model with 216,562 learning parameters. The model training procedure was repeated 10 times, employing the Monte Carlo cross-validation method, as described by Kuhn and 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 of the dataset (7,680 recordings) for its evaluation. This repetition procedure was employed to ensure more reliable and generalizable results by assessing the model's performance across different subsets of the data. Additionally, it helped to account for the inherent variability in neural network training, where slight changes in initial conditions or optimization paths can lead to different model outcomes. While a large and diverse dataset could mitigate this issue, the binaural excerpts used in this study were generated from only 196 multi-track music recordings. This limited source material raised concerns by these authors about potential significant variations between the development and test sets in each repetition. In hindsight, these concerns were valid, as the maximum observed difference in MAE between repetitions reached up to 0.85° for ensemble width.

To ensure that the evaluation process was unbiased, the data split 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, head, and torso, only slightly changing throughout the entire life, mainly during infancy (Clifton *et al.*, 1988; King *et al.*, 2001). Therefore, this limitation could be considered in pair how the human auditory system behaves in real life. Nevertheless, it is worth noting that some studies implement HRTF-independent testing for binaural localization models, as demonstrated by Antoniuk and Zieliński (2023) and 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). These techniques were employed to select the model with the best generalization capabilities and prevent 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 estimate two contiguous ensemble parameters: ensemble location and width, both measured in degrees. For model training, the Adam algorithm (Kingma, Ba, 2014) was used. The algorithm minimized 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.

The optimizer was configured with the following hyperparameters: an initial learning rate of 10^{-3}, a decay rate of 10^{-6}, and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Training was conducted using a batch size of 8, with a maximum of 256 epochs set. An early stopping technique was implemented to prevent overfitting, terminating the process if no improvement was observed on the validation set for 20 consecutive epochs. Consequently, the maximum number of epochs was never reached; instead, training concluded after 25 to 36 epochs, with a median of 27.5 epochs. During the training process, the losses for both outputs were combined additively, ensuring equal weighting of both ensemble features.

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 (up to 4.5 GHz). 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 the experimental stages is publicly available on the GitHub repository (Antoniuk, 2024). The spectrogram calculation phase required 21 minutes, data partitioning took 34 minutes, and the total training time for all iterations amounted to 40 minutes, making the entire training and evaluation process 95 minutes long.

4 Results and 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°) for ensemble location. As both ensemble parameters were constrained within the same range of 90°, the results demonstrate that the model exhibits a 44% lower error for ensemble location compared to 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, estimating ensemble width necessitates the identification of these two extreme sources, a process that is inherently more complex than estimating a single average location.

Figure 3 compares the actual and predicted ensemble widths for each sample, showing a heteroscedastic relationship between them, with a slight bias towards predicting lower ensemble width values for higher actual widths. This relationship exhibits a strong positive correlation, with a Pearson coefficient r of 0.90. Additionally, the results indicate that the model provides more precise 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 an MAE of 12.44° for $\omega > 80°$. This effect is more visible in Figure 4, which highlights the impact of the actual ensemble width on the precision of prediction. The correlation between the actual ensemble width and prediction error shows a weak positive relationship, with a Pearson coefficient r of 0.27.

The reduced accuracy 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 lower precision 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.

The correlation between the actual and predicted ensemble location values exhibits a very high

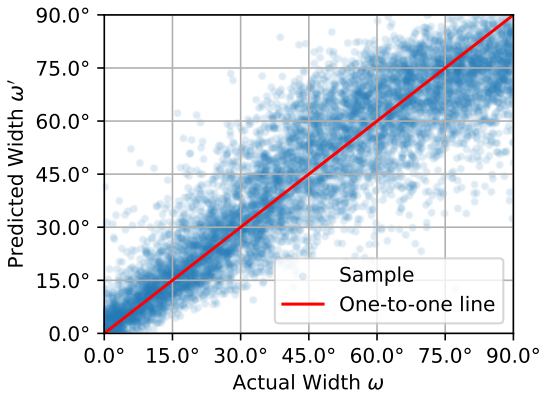

Figure 3: A comparison between the actual ensemble width ω and the predicted ensemble width ω' for a single iteration (of the total ten).

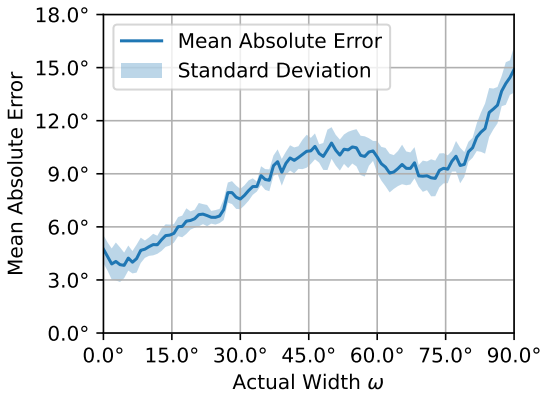

Figure 4: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all ten iterations, with indicated standard deviation.

degree of correlation, as illustrated in Figure 5. In this case, Pearson correlation coefficient r is equal to as much as 0.97. In contrast to the ensemble width, no significant relationship is observed between actual location and its prediction error. This finding suggests that the model's ability to localize the center of the ensemble is robust, unaffected by the actual spatial positioning of the ensemble, including lateral locations. Figure 6 corroborates this observation, demonstrating relatively consistent location error across most positions, with minor increases at extreme locations. The negligible correlation ($r = -0.03$) between the absolute location value and prediction error further supports the model's spatial invariance in its performance.

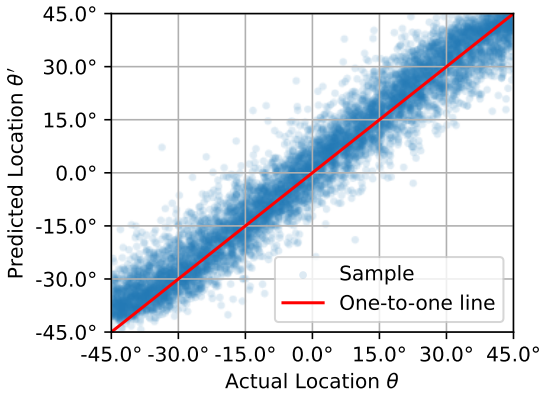

Figure 5: A comparison between the actual ensemble location θ and the predicted ensemble location θ' for a single iteration (of the total ten)

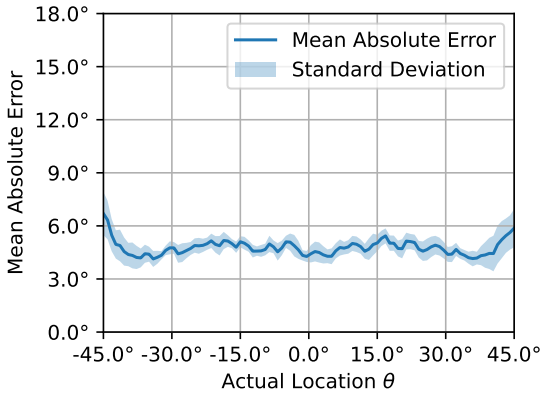

Figure 6: The impact of the actual ensemble width ω on the mean absolute prediction error, averaged across all ten iterations, with indicated standard deviation.

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 asymmetric anomalies, particularly within the $\theta \in [15°, 30°]$ range compared to the $\theta \in [-30°, -15°]$ range, which can be attributed to the sparsity of sample result data across specific regions of this heatmap. While the figure suggests that ensemble location does not significantly affect the model's precision in predicting location, it clearly demonstrates that ensemble width has a substantial impact. Specifically, there is a positive correlation between the width of the ensemble and the error in its location prediction, with error magnitude increasing as the width expands.

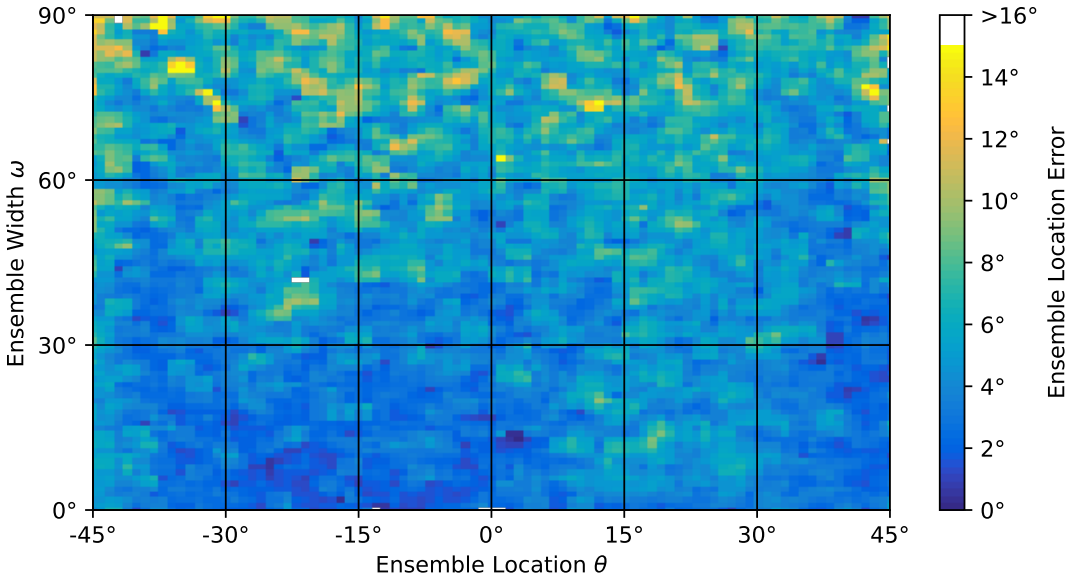

Figure 7: The heatmap illustrating the mean absolute error (MAE) of ensemble location distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors.

Figure 8 reveals a characteristic performance 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 as the error in these areas is considerably higher. This indicates that the model's performance for estimating ensemble width is substantially worse at extreme widths and locations, i.e., when both the width and locations are near their maximum investigated values ($|\theta| \approx 45°, \omega \approx 90°$).

The model presented in this study demonstrates a significant improvement in ensemble-width performance compared to the Spatiogram-based model, first introduced by Arthi and Sreenivas (2021) and further investigated by Antoniuk and Zieliński (2023), under similar evaluation conditions. While the dataset used in this study was expanded with 40 additional multi-track recordings and 10 HRTF databases, Antoniuk and Zieliński (2023) showed that the Spatiogram model's performance does not improve with further increases in dataset size. This finding enables a direct comparison of results between the two models in terms of the precision of the ensemble width estimation, despite the differences

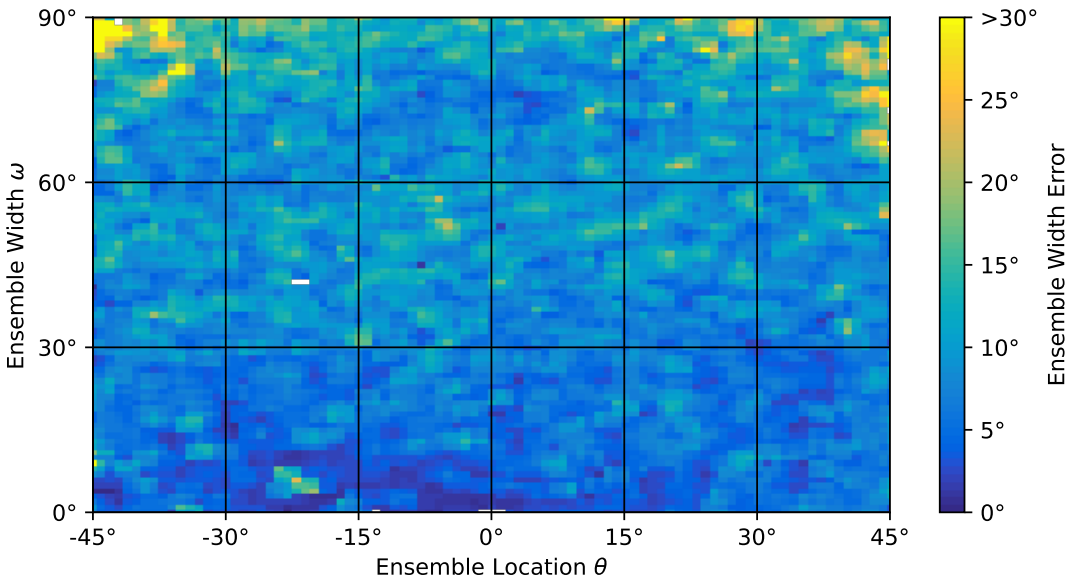

Figure 8: The heatmap illustrating the mean absolute error (MAE) of ensemble width distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors.

in dataset composition. Our model achieved a MAE of 8.57°, outperforming the Spatiogram-based model's result of 13.62° by 5.05°. This substantial improvement is further enhanced by the current model's ability to estimate ensemble location, a feature absent in the previous model.

In terms of ensemble location prediction, the novelty of the proposed method makes direct comparison with existing literature challenging. However, its efficacy can be only evaluated indirectly against state-of-the-art individual-source localization techniques. The ensemble location prediction precision (MAE = 4.76°) of the proposed method can be contextualized with the leading-edge binaural localization DeepEar model introduced by Yang and Zheng (2022). Their model reported MAEs of 7.4° and 2.3° for multi-source and single-source Angle of Arrival (AoA) estimation, respectively. As another promising example, the WaveLoc-CONV model developed by Vecchiotti *et al.* (2019) demonstrated errors of 0° in anechoic conditions and $1.7° - 2.4°$ in multi-condition scenarios. However, these results are limited to single-source speech localization, a substantially less complex task than the ensemble location prediction addressed by the proposed method. These experiments, while differing in objectives and datasets, provide valuable context for the proposed method's performance within current DoA and AoA estimation research.

5 Conclusions

This paper introduces a novel approach to locating audio sources in binaural recordings. Unlike traditional methods that predict the locations of individual audio sources, this study focuses on estimating "ensemble parameters" of audio sources, thus allowing the audio scene to be described using two parameters only: ensemble location and width. This approach makes it possible to avoid making restrictive assumptions about the number of audio sources, rendering the proposed method more suitable for real-world applications. The study also explores the use of convolutional neural networks (CNN) in conjunction with spectrograms applied to their inputs. According to the obtained results, the networks show exceptionally good performance, demonstrating their suitability for the investigated scenario.

The method was developed using 23,040 synthesized binaural excerpts intended to mimic real-world music recordings. The results show its outstanding performance, with the model achieving mean absolute error (MAE) of $4.76^\circ \pm 0.10^\circ$ and $8.57^\circ \pm 0.19^\circ$ for the estimation of ensemble location and width, respectively. While the model is resilient to lateral ensemble locations, it is sensitive to the actual ensemble width, lowering the model accuracy as the width increases. The proposed method demonstrates a significant improvement over the previous technique based on spatiograms (Antoniuk, Zieliński, 2023), lowering the mean absolute error by 5.05°.

Despite its high precision, the method exhibits certain limitations. Since it has been developed using the binaural excerpts synthesized with the head-related impulse responses being inherently anechoic in their characteristics, the method's performance under reverberant conditions has not been validated. Moreover, the proposed method is incapable of operating in real-time scenarios. Validating the method under reverberant conditions as well as optimizing its architecture for practical real-time scenarios constitute the topics for future research. Other minor limitations include the lack of head-related transfer-function (HRTF) independence between the development and test sets, and the absence of vertical variations in audio source placement, as all sources were positioned on the horizontal plane. Additionally, the proposed approach requires substantial computational resources, particularly GPU usage, which was not necessary for the previously used spatiogram-based method.

These limitations, however, present opportunities for future research. Despite the current constraints, this study introduces a novel method for characterizing acoustic scenes in binaural recordings of music, demonstrating substantial potential for advancing binaural audio analysis. The method offers promising prospects for developing innovative tools that can objectively analyze large repositories of binaural audio recordings, focusing on spatial content.

Acknowledgments

The work was supported by the 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.

Appendix

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", 2014)	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)	

Continued on next page

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

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 (Gardner, 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
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		

Continued on next page

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

29.	Artificial	GRAS 45BB-4 KEMAR - subject A attachment	1	Aalborg University; University of Iceland (Spagnol *et al.*, 2019, 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. doi: 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. doi: 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. doi: 10.1109/JSTSP.2015.2400417.

4. Antoniuk, P. (2024). *Software repository: Estimating ensemble location and width in binaural recordings of music with convolutional neural networks* [GitHub]. Retrieved July 1, 2024, from https://github.com/pawel-antoniuk/ensemble-width-cnn

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. doi: 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. doi: 10.1109/TASLP.2018.2806745.

9. Blauert, J. (1996, October). *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1121/1.1907229.

16. Chollet, F., *et al.* (2015). *Keras* [GitHub]. Retrieved July 1, 2024, from https://github.com/fchollet/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. doi: 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. doi: 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. doi: 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. doi: 10.1186/s13636-015-0069-2.

22. Gardner, 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. doi: 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. doi: 10.1044/jshd.1502.114.

27. *HRTF-database* [Austrian academy of sciences]. (2014). 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). *International Communications Union*.

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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. (2016). *A computational framework for modelling active exploratory listening that assigns meaning to auditory scenes—reading the world with two ears.*

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. doi: 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., Miccini, R., Unnthorsson, R. (2020). The viking HRTF dataset v2.

59. Spagnol, S., Purkhús, K. B., Unnthórsson, R., Björnsson, S. K. (2019). THE VIKING HRTF DATASET.

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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.3390/app10175956.