Zieliński et al. EURASIP Journal on Audio, Speech, and Music Processing (2022) 2022:3
https://doi.org/10.1186/s13636-021-00235-2

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH Open Access

Automatic discrimination between front and back ensemble locations in HRTF-convolved binaural recordings of music

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

Abstract

One of the greatest challenges in the development of binaural machine audition systems is the disambiguation between front and back audio sources, particularly in complex spatial audio scenes. The goal of this work was to develop a method for discriminating between front and back located ensembles in binaural recordings of music. To this end, 22, 496 binaural excerpts, representing either front or back located ensembles, were synthesized by convolving multi-track music recordings with 74 sets of head-related transfer functions (HRTF). The discrimination method was developed based on the traditional approach, involving hand-engineering of features, as well as using a deep learning technique incorporating the convolutional neural network (CNN). According to the results obtained under HRTF-dependent test conditions, CNN showed a very high discrimination accuracy (99.4%), slightly outperforming the traditional method. However, under the HRTF-independent test scenario, CNN performed worse than the traditional algorithm, highlighting the importance of testing the algorithms under HRTF-independent conditions and indicating that the traditional method might be more generalizable than CNN. A minimum of 20 HRTFs are required to achieve a satisfactory generalization performance for the traditional algorithm and 30 HRTFs for CNN. The minimum duration of audio excerpts required by both the traditional and CNN-based methods was assessed as 3 s. Feature importance analysis, based on a gradient attribution mapping technique, revealed that for both the traditional and the deep learning methods, a frequency band between 5 and 6 kHz is particularly important in terms of the discrimination between front and back ensemble locations. Linear-frequency cepstral coefficients, interaural level differences, and audio bandwidth were identified as the key descriptors facilitating the discrimination process using the traditional approach.

Keywords: Spatial audio information retrieval, Feature engineering, Binaural recordings, HRTF

1 Introduction

The renewed and still increasing popularity of binaural technologies, seen over the past decade, promotes the creation of large repositories of binaural audio or audio-visual recordings. This tendency might give rise to currently unknown challenges in semantic search and retrieval of such recordings in terms of their "spatial information."

Most of the studies in the area of spatial audio information retrieval intended for binaural signals aim to localize "individual" audio sources, while the attempts to characterize the location, depth, or width of the "ensembles" of sources are scarce. Although the precise localization of virtual audio sources might be beneficial in the design of speech communication systems, e.g., to control beamforming algorithms, in the case of music applications this aspect might be considered of secondary importance, giving precedence to higher-level spatial audio scene descriptors, such as ensemble width, depth, or its location [1].

* Correspondence: s.zielinski@pb.edu.pl
[1]Faculty of Computer Science, Białystok University of Technology, 15-351 Białystok, Poland
Full list of author information is available at the end of the article

The aim of this study was to (1) develop a method discriminating between front and back located music ensembles in binaural recordings and to (2) quantify the influence of the selected parameters on its performance. While the above discrimination task may appear to be simple, it is very challenging to undertake since machines, like humans, suffer from a well-known front-back confusion phenomenon [2]. Even the state-of-the-art algorithms intended for the localization of single talkers in binaural speech signals struggle to disambiguate front speakers from their back constituents, when used without simulated head movements [3, 4]. The task of front-back discrimination is even more challenging in the case of complex spatial audio scenes considered in this study, with many simultaneous sound-emitting sources (such as in music ensembles), due to confounding of binaural cues reaching the artificial ears.

The development of the method for the automatic discrimination between front and back located music ensembles in binaural recordings is important for several reasons. Such a method could be used in future systems for search and retrieval of binaural recordings (while music ensembles are traditionally located in front of a listener, modern binaural audio recordings increasingly include unusual music ensemble settings). Moreover, it could improve the performance of binaural localization methods since an initial "crude" front-back disambiguation procedure may help in subsequent "fine" estimation of the direction of sound incidence [5]. Besides, it could aid music genre recognition algorithms since spatial properties of musical ensembles are likely to be genre specific. In addition, the method might be used as a building block for spatial audio up-mixing algorithms to disambiguate groups of front sources from the back ones. Furthermore, it might be employed in algorithms for objective quality assessment of binaural audio signals, as there is a strong interaction between the spatial distribution of foreground audio content around a listener and the audio quality [6].

This study builds on the prior work of the present authors. While some of our former studies also, among other factors, considered the automatic localization of music ensembles in binaural signals [7–9], they were either focused on achieving the task of classification of spatial audio scenes without too much effort paid to understanding the mechanisms underlying the classification process [7], or they aimed to compare human against machine-classification of spatial audio scenes [9]. The limitation of the work described in [8] is that it only studied the method performance under the HRTF-dependent scenario, ignoring an important aspect of HRTF-independent testing. In this work, much more effort was put into getting a better insight as to how the algorithms undertake their discrimination, and to find which features play the most important role in this process. For example, the performance of CNN was visualized using a gradient attribution mapping technique [10] (visualization of deep learning algorithms is very rare in the literature concerning binaural modeling of hearing).

Another key aspect differentiating this work from the previous ones is that our former studies were based on the binaural recordings synthesized using a set of thirteen binaural-room-impulse-responses (BRIR). These BRIR sets were captured in real-life acoustical venues such as recording studios or control rooms in order to increase the ecological validity of these studies. However, this imposed some problems that potentially affected the results. The first challenge was related to spatial sparsity of the BRIR used in that they were measured at a limited number of selected angles, which made it difficult to accurately pan virtual sources and might have introduced spatial aliasing effects. The second challenge considered a relatively small number of BRIR sets used (13 in total), making it difficult to test the generalizability of the development methods, particularly in the context of deep learning techniques. To overcome these limitations, in this study a relatively large number of HRTF sets was used (74 in total) allowing the authors to thoroughly validate the generalizability of the developed method. Moreover, the employed HRTF sets were of much higher spatial resolution compared to the previously used BRIR sets.

The paper is organized as follows. Section 2 introduces the background to this study. The binaural audio repository is described in Section 3. Section 4 describes the common methodology employed in all the experiments. In addition, it gives a concise overview (outline) of the experimental work. Section 5 presents the experiments in separate sections in chronological order. The discussion, including the limitations of the study, and the conclusions are presented in Sections 6 and 7, respectively.

2 Background

2.1 Spatial audio scene-based paradigm

Complex spatial audio scenes could be described at the three following hierarchical levels: (1) low-level of individual audio sources, (2) mid-level of ensembles of sources, and (3) high-level of acoustical environments, as pointed out by Rumsey [1] in his spatial audio scene-based paradigm. However, most of the studies concerned with spatial audio scene analysis in binaural signals are limited to the first level of spatial scenes characterization, predominantly to the localization of individual sources (see [11] for the review of the binaural localization models), whereas the studies aiming to develop methods for the automatic characterization of higher-level properties of spatial audio scenes are still very rare [12]. This study is concerned with the automatic characterization of spatial scenes at the mid-level, using the above hierarchy.

2.2 Perception of front and back audio sources

Perception of complex spatial scenes within the horizontal plane is predominantly governed by the three types

of binaural cues, namely interaural level differences (ILD), interaural time differences (ITD), and interaural coherence (IC), whereas the perception of scenes in the vertical plane mainly depends on the spectral cues conveyed by the head-related transfer functions (HRTF) [13]. The main challenge hindering the developments of the systems for the binaural spatial audio analysis is the so-called front-back confusion effect, originating from the ambiguity of the binaural cues conveyed by the signals reaching the ears. In order to disambiguate front sources from the back ones, humans rely on visual cues or subconsciously employ micro head movements [4, 14]. In the absence of the visual cues or with lack of the dynamic cues induced by the head movement, humans must rely on HRTF spectral cues, similar to the localization of sources in the vertical plane. However, the literature is not conclusive as to which spectral cues in particular play the major role in the front-back disambiguation. Some studies support the view that there exist some universal macroscopic spectral regions responsible for the front-back disambiguation [15–19], whereas some others conclude that there are no generic spectral cues but it is the listener-specific spectral cues that help to discriminate front and back sources [20, 21].

2.3 The state-of-the-art binaural localization models

The state-of-the-art algorithms are capable of localizing several sources emitting sounds simultaneously, both in anechoic and reverberant environments [3, 4, 22–24]. Traditionally, the two-stage topology was employed, consisting of the audio feature extractor followed by the machine learning classifier [3, 25, 26]. The features fed to the input of such classifiers normally included a selection of binaural cues (ILD, ITD) [25]. Due to the problems with the accurate estimation of ITD descriptor under noisy or reverberant conditions, some of the recent models also exploit raw values of the inter-aural cross-correlation function as a feature vector [3, 23].

In line with the general trend in artificial intelligence, the recently developed binaural localization models utilize deep learning techniques [3, 24, 27–29]. Binaural signals can either be directly used at the input of the deep neural networks [27] or indirectly after some form of signal processing, typically involving a traditional feature extraction [4], estimation of the inter-aural cross-correlation function [3, 23, 24, 29], or calculation of spectrograms [28]. The last mentioned solution was also applied in this study. The fundamental weakness of the state-of-the-art methods is that, for complex scenes, such as those used in this work, they require an a priori knowledge of the number and characteristics of the individual sources in an analyzed scene [3, 13, 24–26]. Such information is normally unavailable in real-life binaural audio repositories with music recordings, rendering these methods impractical for the discrimination of ensemble locations. Drawing inspiration from human hearing systems, some state-of-the-art binaural audio localization methods were designed to mimic the head movements, demonstrating significant improvements in discrimination between front and back sources [3, 4]. However, such an approach limits the scope of its practical applications, since the system must be either equipped with a robotic head (such as in [30]) or it should be capable of adaptive synthesis of the binaural signals [3]. The state-of-the-art binaural localization models are typically developed and tested under HRTF-dependent conditions, that is in a setting where sound recordings used for training and testing are generated using the same HRTF sets. This approach tends to give inflated results (due to an over-fitting effect) and is insufficient to test the generalization property of the developed models [29].

The key features of the developed model presented in this paper is that (1) it incorporates a static-head approach, (2) the method is assumption free with regard to the number and type of music audio sources in analyzed recordings, and (3) the developed method was thoroughly tested both under the HRTF-dependent and HRTF-independent conditions, allowing the authors for a comprehensive characterization of the method's performance.

3 Repository of binaural music recordings

In total, 22,496 binaural music recordings were generated by convolving 152 multitrack music recordings with 74 HRTF sets. Each binaural excerpt had a duration of 7 s and represented either a front or back located music ensemble (see Fig. 1 for an example of the spatial distribution of audio sources). The ensemble width was restricted to fixed azimuth limits of ± 30° during the convolution procedure (this limit was adopted for consistency with our previous studies [7, 9]), with the angles randomly selected for individual sources. Despite the relatively narrow intended ensemble width, informal listening tests undertaken by these authors showed that the perceived width for some stimuli was much wider than intended, spanning almost all semicircles (± 90°), which is consistent with the observation made by Pulkki et al. [31]. HRTFs were not interpolated during the convolution procedure. For each intended direction of an individual source, the HRTFs for the nearest available angle were selected.

The music recordings represented a broad range of genre types, including classical music orchestral excerpts, opera, pop music, jazz, country, electronica, dance, rock, and heavy metal. The number of individual sound sources within each recording ranged from 5 to 62 with a median of 10. All individual multitrack signals

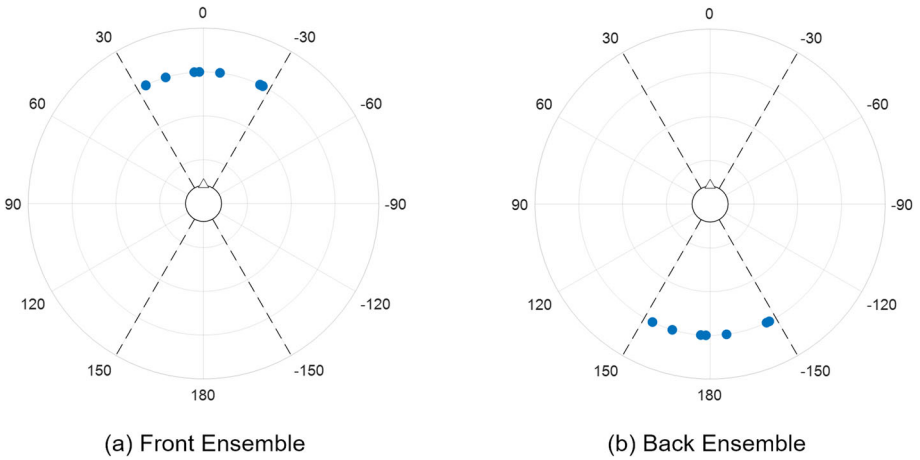

Fig. 1 Examples of spatial audio sources distribution considered in the study: **a** front ensemble, **b** back ensemble. Dots represent $N = 7$ binaurally synthesized virtual audio sources for the selected music recording

were RMS-normalized prior to the convolution. The binaural excerpts generated were stored as uncompressed two-channel files, sampled at a rate of 48 kHz with a 24-bit resolution.

The reason for including a relatively large number of HRTFs (74 in total) was the need to thoroughly test the generalization property of the developed method. The employed HRTFs could be divided into thirteen groups based on their origin, e.g., the institution where the measurements were undertaken [32–46] (see Table 3 in the Appendix for a detailed description of the HRTFs used in this study). The selection included both human HRTFs as well the artificial ones such as KEMAR (GRAS 45BA, 45BB-4, 45BC, DB-4004), Neumann (KU 100), Head Acoustics (HMSII), FABIAN, SAMRAI (Koken), and ARI Printed Head. The employed HRTFs were measured in near or far field, with the measurement radius ranging from 50 cm to 2 m. Some of the HRTFs were selected based on the study of So et al. [47], who within the set of 196 non-individualized HRTFs identified a subset of 12 HRTFs, representative of the clusters for forward and backward directional sound (HRTFs No. 10–16, 31–33 in Appendix: Table 3).

The repository of binaural renderings was divided into two sets intended for model training and testing, respectively. Table 1 illustrates the data distribution between the sets. Note, that while the training and testing sets were mutually exclusive in terms of music recordings, they shared the same spatial characteristics as the identical 74 HRTFs were used to generate the recordings in both sets. This may potentially cause a problem regarding the model "generalizability" testing, however, this was addressed in the final experiment by using three different data filtering strategies "within" the original data splits (Sec. 5.8), ensuring that "unique" HRTFs were employed in the train and test repositories.

4 Methodology

This section describes the common methodology shared across the experiments. Differences in the experimental protocol pertinent to the individual experiments will be described separately in Section 5.

4.1 Experimental outline

For clarity, the empirical work will be presented in separate sections, with the specific aims, method, and results of each experiment described individually. The outline of these experiments is shown in Fig. 2.

The first three experiments, highlighted at the top left-hand side of Fig. 2, concerned the process of "feature engineering" pertinent to the traditional machine learning approach. Their aim was to evaluate the usefulness of the literature-inspired front-back cues, the cepstral coefficients, and the binaural cues, respectively. These experiments are presented in Sections 5.1–5.3. In the fourth experiment, described in Section 5.4, the above groups of features were combined and compared. The assumption was that combining the individual groups of

Table 1 Train and test data split in the repository of binaural audio excerpts

Dataset type	No. of music recordings	No. of HRTF sets	No. of ensemble locations	No. of binaural excerpts
Training	112	74	2	16576
Testing	40	74	2	5920

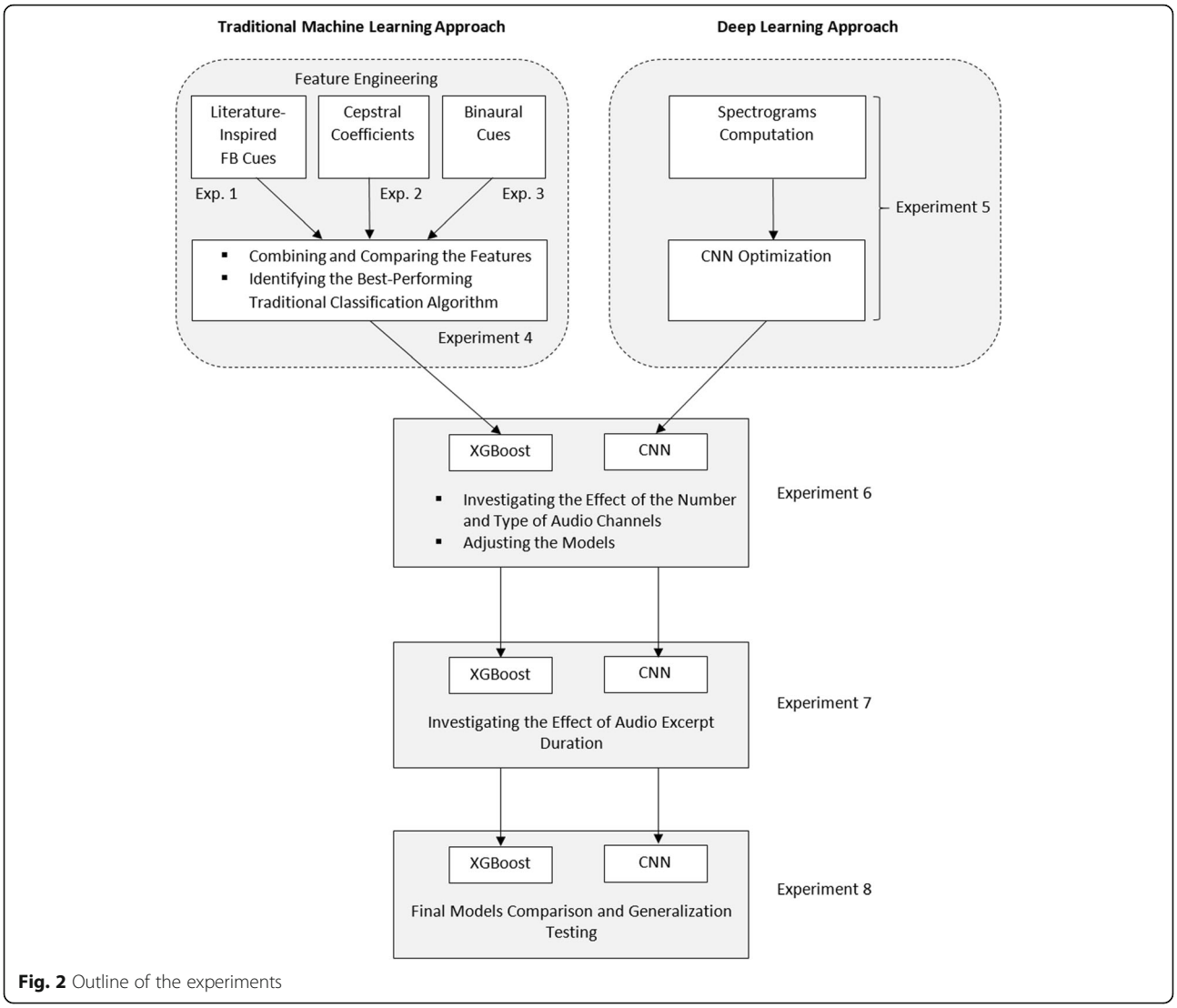

Fig. 2 Outline of the experiments

features would improve the discrimination accuracy. Moreover, the performance of the three traditional classification algorithms was evaluated and the best-performing classifier was identified (XGBoost was found to be the "winning" algorithm). The fifth experiment, described in Section 5.5, concerned the development of the deep learning-based method, as indicated at the top right-hand side of Fig. 2. It involved the spectrograms computation and the optimization of CNN, respectively.

The remaining three experiments outlined in Fig. 2 allowed for further "fine-tuning" of the best-performing traditional method (XGBoost) and the deep learning method (CNN). In the sixth experiment described in Section 5.6, the effect of audio channel selection on the performance of both methods was investigated and, consequently, the methods were adjusted according to the obtained results. In the seventh experiment (Sec. 5.7), the effect of the audio excerpt duration was quantified. In the concluding experiment, the final models were compared and tested for their "generalizability" (Sec. 5.8).

Section 5.9 presents the results of an additional "follow-up" experiment. Its purpose was to get a better understanding of how the algorithms are undertaking their discrimination tasks obtained during the study. For clarity, this experiment was omitted in Fig. 2.

4.2 Discrimination methods and their implementation

Two approaches to the discrimination of music ensemble locations were taken: traditional machine learning and deep learning. The traditional approach typically involves a feature extraction procedure followed by a classification routine. By contrast, in the deep learning approach, the audio signals either are fed directly to the input of the classification algorithm [27, 48], thus making the feature extraction part obsolete, or undergo some initial transformations, most notably conversion to spectrograms [28, 49]. Another

distinct feature of the deep learning algorithms is their large number of "trainable" parameters. In the computer vision classification task deep learning methods markedly outperformed the traditional machine learning algorithms [50]. However, in our former work [9] concerning the automatic classification of complex audio scenes in the BRIR-convolved binaural recordings, the performance of the deep learning algorithm was similar to that of the traditional machine learning methods. Therefore, in this study, both the traditional machine learning methods and the deep learning technique were incorporated, and their performance compared.

The performance of the following three traditional classification algorithms was compared: Logistic regression (Logit), support vector machines (SVM), and an extreme gradient boosting (XGBoost) technique. Logit method was selected due to its inherent property of feature selection, when used with $L1$ regularization [51], and because of its robustness to data multicollinearity (the feature vectors extracted in this study exhibited a high multicollinearity level). Moreover, in one of our previous studies regarding the localization of complex spatial scenes in binaural signals the Logit algorithm showed a similar generalization performance compared to CNN [9]. The SVM-based classification algorithm was employed in this work due to its popularity and applications in acoustic scene classification [52]. The rationale for selecting the XGBoost technique was related to its exceptionally good performance, for some applications even outperforming deep learning techniques [53].

In the initial three experiments described in this paper (Sections 5.1–5.3), only the Logit method was used, since the experiments focused on the feature evaluation rather than the comparison of the classification methods. The Logit method was employed for this task since, as already mentioned above, it showed a similar generalization performance compared to CNN in one of our previous studies [9]. Then, the performance of the three aforementioned algorithms was compared in the 4th experiment described in Section 5.4. According to obtained results, XGBoost was selected as the best traditional machine learning algorithm and it was later used for the comparison against CNN.

As far as the deep learning method approach is concerned, it was decided to employ the technique based on CNN, as it is widely used in audio classification applications [49]. While raw waveform-based deep learning methods gain in popularity [27, 48], the best-performing deep learning algorithms in the area of acoustic scene classification or audio event recognition still exploit intermediate "spectrogram" signal representation rather than learnable signal transformations [49]. Consequently, in this work, the binaural signals were also converted to spectrograms, before being fed to the input of CNN.

Logit, SVM, and XGBoost algorithms were implemented in the Python ecosystem using *NumPy*, *scikit-learn*, and *XGBoost* libraries. While the first two algorithms were deployed on a single CPU (Intel i7-7820X, 4.3 GHz, 8 cores, 16 threads), XGBoost calculations were GPU-accelerated using an NVIDIA graphics card (RTX 2080 Ti). CNN was implemented in MATLAB using the Deep Learning Toolbox. The computations involving CNN were undertaken employing Microsoft Azure cloud computing service. They were accelerated with 4 GPU units (NVIDIA Tesla T4).

4.3 Feature extraction

All the binaural audio excerpts were RMS-normalized prior to feature extraction. In line with our former work [7, 9], all the features considered in this study, with the exception of the binaural cues, were extracted from the left and right signals (x and y), respectively, as well as from their sum and difference signals (m and s).

The features were extracted using time frames of 20 ms in duration, overlapping by 50%. A Hamming window was applied to the signal in each frame. The features extracted for each frame were subsequently "aggregated" by calculating two statistics: mean value and standard deviation. Only two statistics (mean and standard deviation) were used due to the results of the pilot tests, demonstrating that adding extra statistics, such as kurtosis or skewness, brings no improvements in the performance of the algorithms. In contrast to our former work [7], no delta features were computed, as according to the pilot tests, they had a negligibly small effect on the discrimination results.

The extracted features were standardized before they were applied to the input of the classification algorithms (mean value of each feature was equalized to zero, whereas their variance was normalized to unity). While some of the machine learning methods may benefit from the automatic feature selection, according to our pilot tests, the advantage of using such techniques was negligibly small. Consequently, the automatic feature selection was not used in this work. However, the features were selected "manually," based on the comparative analysis of their contribution to the performance of the method, as explained in detail in Section 5.6.

Since the left (x), right (y), and sum (m) signals were very similar, the extracted features showed a high correlation level. While a transform employing a principal component analysis (or a similar technique) could be used as a preprocessing stage to "de-correlate" the features, this would introduce additional difficulty in interpretation of the results (explaining which features are the most important). Therefore, such technique was deliberately avoided in this work.

4.4 Performance evaluation

The traditional classification algorithms were trained using the features extracted from the training set of the

repository of the audio excerpts and subsequently tested with the use of the features extracted from the testing set of the above repository. The same principle applied to CNN, but with the spectrograms used instead of the features. The data between train and test sets were divided in proportions of 74 to 26%, as illustrated above in Table 1.

The performance of each algorithm was evaluated using the discrimination "accuracy" metric, being the ratio of the correctly discriminated excerpts in the test set to their total number. In addition, for some of the models, the confusion matrices were inspected. While in many machine learning studies, such metrics as precision, recall, and F scores are also included (along with analysis of receiver operating characteristic curves) [51], in this work they were omitted, as the single "accuracy" metric combined with the inspection of the confusion matrices was deemed by these authors sufficient to evaluate the performance of the developed models.

All the data, consisting in total of 22,496 synthesized excerpts, were divided into the train and test sets in proportions of 74 to 26%, as already illustrated above in Table 1 (16,576 data records were allocated for training whereas the remaining 5920 records were assigned for testing purposes). The standard 10-fold cross-validation technique combined with a grid-search optimization technique was used to train and optimize the models using the "train" data set. Then, the best model was tested on the held-out "test" set. In order to compare the results statistically, each model was repeatedly trained and tested seven times. In each repetition, the split between the train and test sets was changed randomly, maintaining the same proportion of "data records" between the sets (74 to 26%) and ensuring that different music recordings are allocated to both sets. The mean discrimination accuracy scores and associated standard deviation values, calculated across repetitions, are reported throughout the paper.

4.5 HRTF-dependent and HRTF-independent tests

Note that for each data split, the train and test feature sets were music recording independent —in that the sets contained the features that were extracted from the unique music recordings. However, the train and test sets were "HRTF dependent," since the same HRTFs were used to generate the audio excerpts in both sets. Unless otherwise stated, the models described in this paper were developed and subsequently tested under HRTF-dependent conditions.

The disadvantage of HRTF-dependent testing is that it may give "inflated" results due to an overfitting effect. Moreover, such an approach does not allow for checking how the model would perform under unknown HRTF conditions. Therefore, to examine the "generalization" property of the final models developed within this study, in the concluding experiment described in Section 5.8, HRTF-independent tests were performed. In the latter case, the train and test audio repositories were mutually exclusive (independent) in terms of the HRTF sets used to generate the binaural excerpts. In other words, under the HRTF-independent test conditions, the train and test repositories were "unique" in terms of the HRTF sets.

The reason for employing only HRTF-dependent tests during the development phase was related to the computational time required to train, test, and optimize the models. Under the HRTF-dependent conditions, only 7 models had to be developed and tested (a separate model for each data split). By contrast, depending on the strategy of HRTF-independent testing, the number of models required for testing was much greater. For example, when the technique of dividing 74 HRTF sets to 13 independent HRTF corpora was adopted, described in Section 5.8, the number of models to be evaluated was equal to 91 (7 data splits × 13 independent HRTF corpora), which posed a considerably increased computational load.

It must be stressed that HRTF-independent tests, also referred to as tests under HRTF "mismatched" conditions, are still rare in the literature, as pointed out by Wang et al. [29]. In this study, three different techniques of HRTF-independent tests were carried out (Section 5.8).

5 Experiments

5.1 Experiment 1: Evaluating the usefulness of literature-inspired front-back cues

The first and novel group of features included in this study was inspired by the literature in the area of spatial hearing. The aim of the experiment described in this section was to ascertain whether such features could be exploited in an algorithm for front and back ensemble discrimination in binaural recordings.

5.1.1 Front-back cues calculation

According to the pioneering work of Blauert [16], there are five distinct frequency regions, referred to as "boosted bands," indicating local spectral maxima responsible for the perceived front or back direction of broadband signals (such as music). They are labeled as $B1 \div B5$ at the top of Fig. 3. For narrowband noise signals, Blauert [16] identified slightly different frequency regions, naming them as "directional bands." They are signified as $D1 \div D4$ in Fig. 3.

To capture information present in Blauert's boosted bands, in this work a simple but effective method was employed based on RMS level estimation. Signals x, y, m, and s were first band-pass filtered using the cut-off

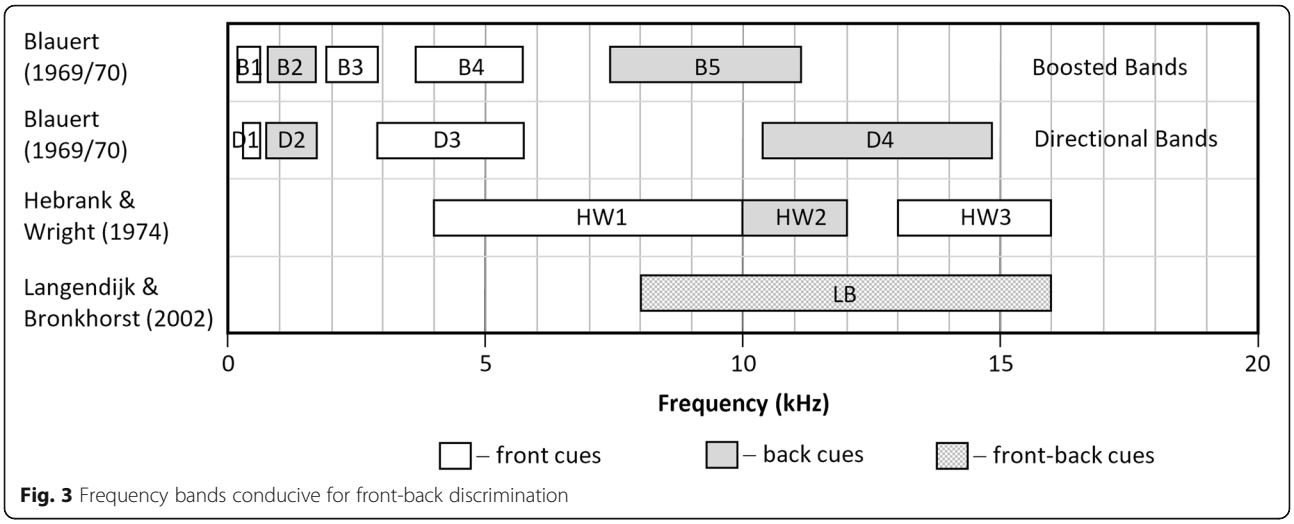

Fig. 3 Frequency bands conducive for front-back discrimination

frequencies given in Table 4 in the Appendix. To this end, a finite-impulse response filter of the order of 512 was applied. The filter was designed using the "window" method and implemented in MATLAB employing the "filtfilt" function, allowing for a zero-phase filtering (zero phase distortions). For reproducibility, the code has been made publicly available at GitHub [54]. Then, RMS levels of the filtered signals were estimated and used as the features. As a result, 40 feature vectors were calculated (4 signals × 5 bands × 2 statistics). In a similar manner, the features based on Blauert's four directional bands were also calculated (see Fig. 3), giving additional 32 vectors (4 signals × 4 bands × 2 statistics).

According to Hebrank and Wright [17], there are three frequency regions conveying information about front or back located sound sources, labeled as *HW*1, *HW*2, and *HW*3 in Fig. 3. For these three bands, a similar procedure was also employed as before, extending the feature matrix by further 24 vectors (4 signals × 3 bands × 2 statistics).

In contrast to the previously outlined studies, Langendijk and Bronkhorst [20] asserted that listener-specific "spectral variations" at high frequencies, in the range of 8–16 kHz (illustrated at the bottom of Fig. 3), constitute important front-back localization cues. To account for these cues, signals x, y, m, and s were initially band-pass filtered (8–16 kHz) and then the standard spectral descriptors [55] were extracted from the filtered signals using MATLAB's Audio Toolbox. The following features were extracted: spectral centroid, spectral crest, spectral decrease, spectral entropy, spectral flatness, spectral flux, spectral kurtosis, spectral roll-off point, spectral skewness, spectral slope, and spectral spread. As a result, additional 88 feature vectors were generated (4 signals × 11 descriptors × 2 statistics). In total, 184 feature vectors were extracted using the literature-inspired front-back cues described in this section. The frequency limits of the frequency bands used in this study are provided in Table 4 in the Appendix.

5.1.2 Discrimination method and its optimization

Since the aim of the experiment was to evaluate the usefulness of the front-back cues, not to compare the classifiers, a single classification method was used, namely Logit. This method was selected since, according to our previous work [9], its generalization performance was similar to that of CNN. The parameter C of the selected classification algorithm (being an inverse of the regularization coefficient, as used in the *scikit-learn* package [56]) was "tuned" using a standard grid-search procedure undertaken with a 10-fold cross-validation technique. The values were selected from the set $C \in \{0.01, 0.1, 1\}$, with its elements established during the pilot tests. The Logit algorithm was used with $L1$ regularization. Selecting $L1$ regularization, as opposed to $L2$ one, takes advantage of the method's native property of feature selection [51].

5.1.3 Results

Figure 4 shows the results obtained using various combinations of the literature-inspired front-back cues. Unsurprisingly, the features based on Blauert's boosted bands (BB) gave better results than those obtained employing his directional bands (DB), as the latter were established using narrowband signals rather than broadband ones such as music [16]. It can also be noted that combining the front-back cues in groups improves the results. When all four front-back cues were used together (BB, DB, HW, LB), Logit classifier was able to discriminate front and back located music ensembles with the accuracy of reaching almost 90%.

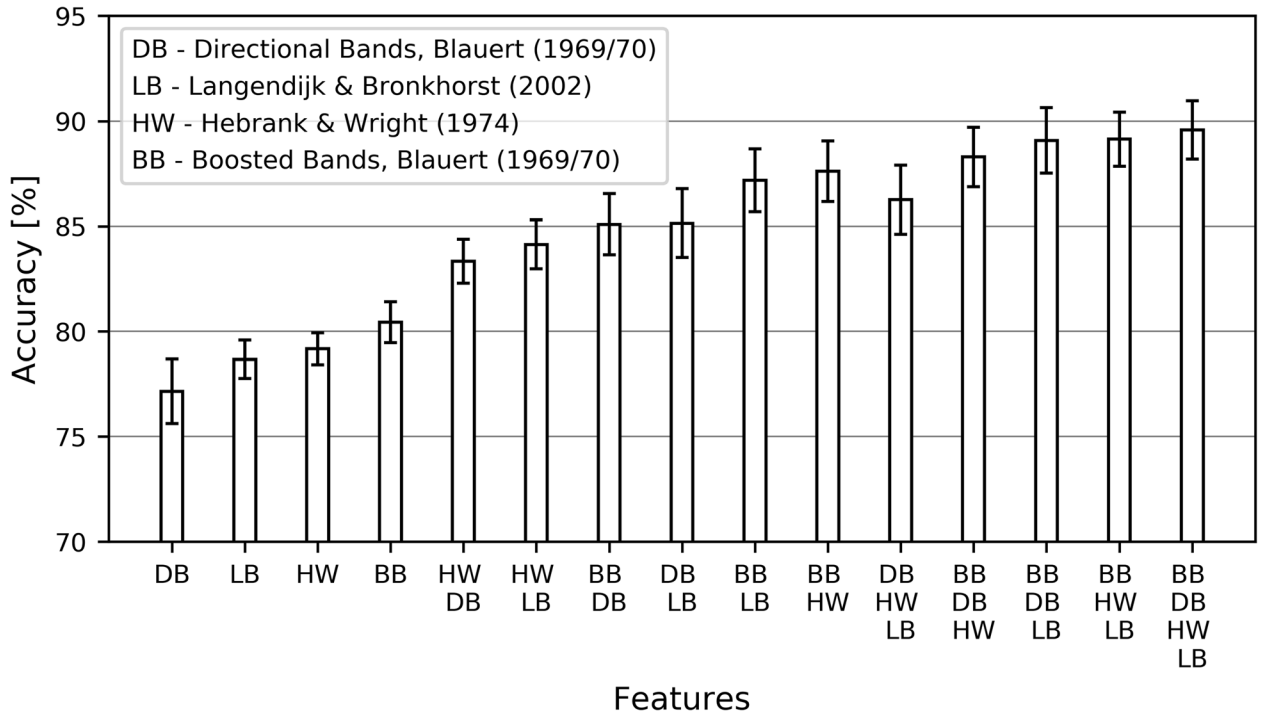

Fig. 4 Mean classification accuracy scores obtained for the literature-inspired front-back cues using Logit classifier. Error bars denote standard deviations

An example confusion matrix obtained using a combination of all four groups of front-back cues (DB, LB, HW, BB) and Logit classifier is presented in Fig. 5a (For consistency, all the confusion matrices presented in Fig. 5 were obtained using Logit classifier). It can be seen that out of the 5920 binaural excerpts included in the test set, 2621 recordings were discriminated correctly as representing frontally located music ensemble and 2707 excerpts were identified correctly as representing back located ensemble. It can also be seen that 339 of the recordings with frontally located ensemble were incorrectly classified as representing back-located ensemble, whereas 253 excerpts with an ensemble located at the back were misclassified as the excerpts with a frontally located ensemble.

A follow-up analysis of the classification errors revealed that they were predominantly related to HRTFs used for the convolution of the binaural excerpts rather than to the characteristics of the music recordings. Out of 592 misclassified locations, 101 were attributed to the excerpts obtained using TH KÖLN [42] HRTF corpus. This corpus was unique in our repository. It was the only corpus containing the measurements obtained with the artificial head by Head Acoustic, involving some "unusual" scenarios (placing a baseball cap or a virtual reality headset on top of the head).

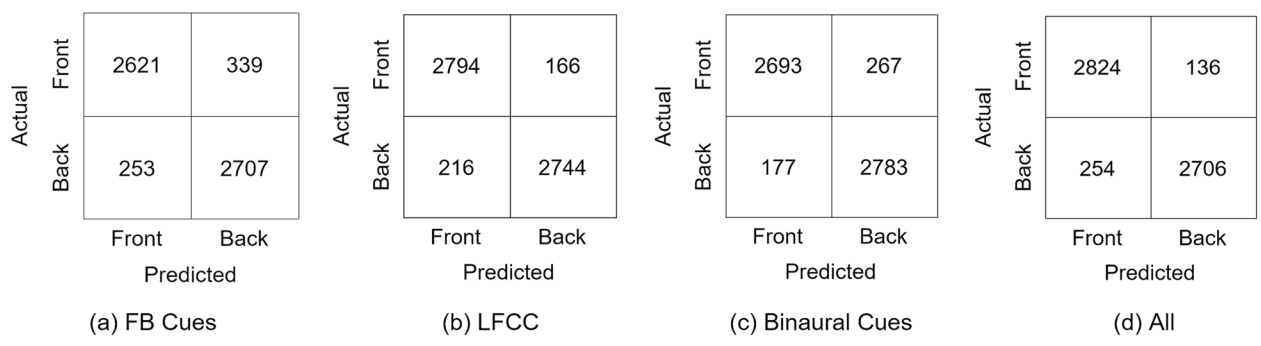

Fig. 5 Examples of confusion matrices obtained for **a** literature-inspired front-back cues, **b** linear-frequency cepstral coefficients, **c** binaural cues, and **d** combination of all features. The results were achieved using Logit classifier

The main observation that can be drawn from the results is that even simple RMS estimators of bandlimited signals, designed based on the spatial audio domain expert knowledge, provide the discrimination accuracy ranging from 75 to almost 90%, considerably exceeding a no-information rate which in this study was equal to 50%. The main outcome of this experiment is that all the investigated front-back cues (BB, DB, HW, LB) should be included in experiment 4 aiming to combine and compare different groups of features, described later in Section 5.4.

5.2 Experiment 2: Evaluating the usefulness of cepstral coefficients

The experiment described in the previous section demonstrated that the literature-inspired front-back cues could be used for the discrimination between front and back ensemble locations. The aim of the experiment described in this section is to examine whether the cepstral coefficients could also be exploited for the same task and, if so, to check which type of the cepstral coefficients (linear-frequency or Mel-frequency) gives better results.

Depending on the frequency scale used in calculations, it is possible to distinguish between two types of cepstral coefficients, namely Mel-frequency cepstral coefficients (MFCCs) and linear-frequency cepstral coefficients (LFCCs). MFCCs are known to be efficient descriptors of spectral envelopes, and for this reason, they are commonly applied in various machine audition tasks including acoustic scene classification [52]. Therefore, one can hypothesize that they may also well account for any spectral cues responsible for the discrimination of front and back located music ensembles, which was, to an extent, confirmed by one of our initial studies [9]. However, for some applications, linear-frequency cepstral coefficients (LFCCs) could produce better classification results when distinct spectral characteristics are distributed across higher frequencies [57]. Therefore, for comparative purposes, both MFCCs and LFCCs were included in the study. MFCCs and LFCCs were used in the experiments interchangeably (they were not merged together).

5.2.1 Cepstral coefficients calculation

In the literature, the number of cepstral coefficients used for speech recognition is commonly limited to 13 [58]. However, the optimum number of coefficients intended for ensemble location discrimination was unknown at the outset of this experiment. Therefore, its value was selected from the set of $n_{coeff} \in (13, 20, 30, 40)$. The set of considered values n_{coeff} was developed heuristically during the pilot test. Both MFCCs and LFCCs were calculated in MATLAB using the Audio Toolbox.

Similar to the experiment described above, only a single classification method was used, namely Logit. Its optimization procedure was the same as before.

5.2.2 Results

The results of the discrimination of the binaural excerpts using either MFCCs or LFCCs are illustrated in Fig. 6. Both MFCCs and LFCCs appear to constitute useful features allowing for the discrimination between front and back located ensembles with the accuracy ranging from 89 to 93%, slightly outperforming the cues discussed in the previous section. Most importantly, note that LFCCs give better results than MFCCs. For example, when the number of selected coefficients is set to 30, the mean discrimination score obtained for LFCCs is by 2.7 percentage points greater than that achieved for MFCCs. According to the statistical test of proportions, this difference is significant at $p = 7.3 \times 10^{-8}$ level. Therefore, in the remainder of this study, only LFCCs were retained. The recordings were discriminated with the accuracy of approximately 93% when the number of LFCCs was set to 30, as illustrated in the figure. Consequently, the above number of coefficients was selected in the subsequent experiments that employ the cepstral coefficients reported in this paper. As a result of this choice, the cepstral coefficients were represented by 240 feature vectors (30 coefficients × 4 channels × 2 metrics).

An example of the confusion matrix obtained using 30 LFCCs and Logit classifier is presented in Fig. 5b. In contrast to the outcomes of the error analysis performed earlier in experiment 1, exploration of the misclassified examples revealed that there was no single HRTF corpus predominantly responsible for the observed errors. However, out of 382 observed errors, as much as 65 could be attributed to the same music recording. Namely, it was an opera recording with a relatively narrow spectral content compared to the remaining recordings. Considering that LFCCs are good descriptors of spectral envelopes, it could be tentatively hypothesized that in this case, they introduced some bias towards the spectral characteristics of the music recordings.

The main outcome of the experiment is that only LFCCs should be included in experiment 4 aiming to combine and compare different groups of features, described later in Section 5.4. The number of coefficients should be set to 30.

5.3 Experiment 3: Evaluating the usefulness of binaural cues

The aim of the experiment described in this section was to examine whether binaural cues can also be used in the task of discrimination between the front and back located music ensembles. The experimental details related to the discrimination method (Logit) and its

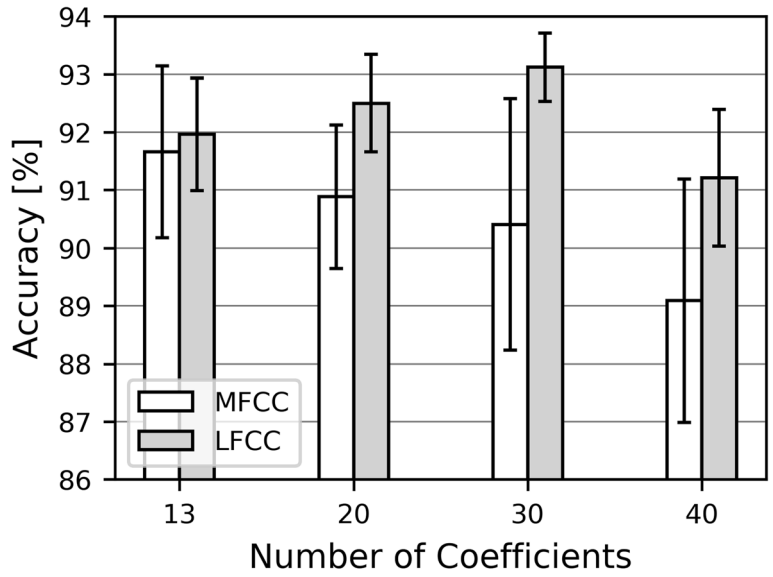

Fig. 6 Mean discrimination accuracy scores obtained using MFCCs and LFCCs (Logit classifier). Error bars denote standard deviations

optimization were omitted here as they were identical to those employed in the previous two experiments.

5.3.1 Binaural cues calculation

In this study, three groups of binaural cues were extracted: ILD, ITD, and IC. To this end, left and right channel signals (x and y) were band-pass filtered using a 42-channel gammatone filter bank, with the lowest and highest frequencies set to 100 Hz and 16 kHz, respectively. Then, the inner hair-cell envelope of the bandpass filtered signals was extracted by half-wave rectification of the signals followed by their low-pass filtering with a cut-off frequency being equal to 1000 Hz, as proposed by Dau [59]. Subsequently, the rate maps [60] were calculated, which were then used to estimate the binaural cues (ILD, ITD, and IC). In this way, 252 feature vectors were extracted (42 channels × 3 types of coefficients × 2 statistics). They were calculated using the MATLAB implementation of the auditory model developed within the Two Ears project [30].

5.3.2 Results

Figure 7 shows the discrimination results obtained using the binaural cues. Observe that ILD features give the best results, approaching 90% (when used in isolation from the other types of binaural cues), outperforming IC cues, and ITD cues. Combining the cues improves the results. When all three types of binaural cues were combined (ILD, ITD, IC), Logit classifier managed to discriminate the excerpts with an accuracy of 92.6%. An example of the confusion matrix obtained using Logit classifier for the combined binaural cues is presented in Fig. 5c. The main outcome of the experiment is that all three types of binaural cues (ILD, ITD, and IC) should be employed in the experiment combining and comparing different groups of features (experiment 4), which is described in the next section.

5.4 Experiment 4: Comparing the features and identifying the best-performing traditional classification algorithm

The goal of this experiment was twofold. First, it aimed at combining and comparing the features identified in the three previous experiments. Second, its purpose was to select the best-performing traditional classifier, out of the three classifiers considered in the study. Instead of pursuing these two aims in separate experiments, it was decided to undertake a single study looking into these two aspects simultaneously. The reason for this decision was the "interactions" between features and classifiers observed in the pilot experiment (relative importance of features, in terms of their discrimination accuracy, was classifier specific).

According to the main outcomes of the previous experiments, the following three groups of features were "combined" (concatenated) in the model: 184 features representing front-back cues from the first experiment (Sec. 5.1), 240 features representing LFCCs from the second experiment (Sec. 5.2), and 252 features based on binaural cues from the third experiment (Sec. 5.3).

5.4.1 Discrimination methods and their optimization

As mentioned earlier, the three traditional classification algorithms were considered in the study: Logit, SVM, and XGBoost. The hyper-parameters of the employed classification algorithms were "tuned" using a standard grid-search procedure undertaken with a 10-fold cross-

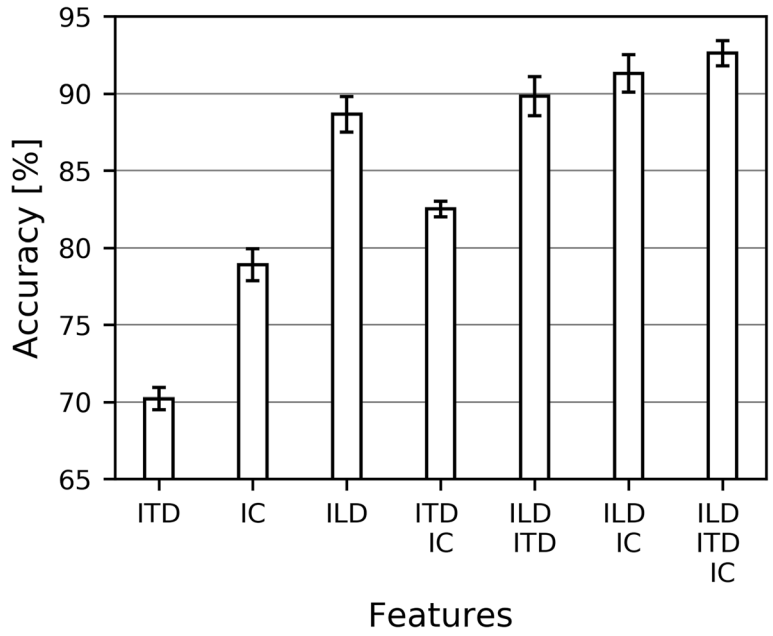

Fig. 7 Mean classification accuracy scores obtained for binaural cues (Logit classifier). Error bars denote standard deviations

validation technique. For the Logit method, the C value (an inverse of the regularization coefficient [56]) was selected from the set $C \in \{0.01, 0.1, 1\}$. The Logit algorithm was used with $L1$ regularization. Two kernels of the SVM algorithm were compared: linear and radial basis function (RBF). When the RBF was applied as the kernel, the following hyper-parameter values were considered during the grid search procedure: $C \in \{0.1, 1, 100\}$ (regularization parameter [56]) and $\gamma = 1/n_f$ (positive constant in RBF kernel [51]), where n_f was the total number of feature vectors used in a given experiment. Due to its computational efficiency, SVM method with a linear kernel was implemented using a stochastic gradient descent (SGD) learning technique with the "hinge" loss function. During the grid-search procedure its regularization coefficient α was selected from the following set: $\alpha \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. The number of estimators for the XGBoost algorithm was fixed, being set to 200 (a value adjusted during the pilot tests). Since after adjusting the number of estimators to 200 XGBoost algorithm already yielded a very high discrimination accuracy with its default hyper-parameter values, no further optimization of this algorithm was performed. Consequently, the remaining hyper-parameters of the above algorithm were set to their default values in the respective software library [61].

5.4.2 Results

Figure 8 shows the results for all three groups of features explored above, separately for all the classifiers. The numbers in brackets placed under the labels along the horizontal axis indicate the total number of feature vectors within each feature group. For clarity, literature-inspired front-back cues are abbreviated as FB in that figure. They represent the combined set of BB, DB, HW, and LB features analyzed earlier in experiment 1.

It can be seen in Fig. 8 that regardless of the classification algorithm, FB features give the worst results. The results obtained for LFCCs and binaural features depend on the classification algorithm used. Logit and SVM-Lin both yield similar results ranging from approximately 91 to 93%. However, when SVM-RBF classifier is used, the results obtained for the binaural cues are markedly better than those obtained for the LFCCs, reaching almost 98%. Combining the features tends to improve the results. However, when all the features are combined, the results deteriorate slightly for the Logit and SVM-Lin classification algorithms. An example of the confusion matrix obtained for the combination of all features (All) is presented in Fig. 5d (Logit classifier was used in this example).

Overall, the results obtained using XGBoost and SVM-RBF classifiers were superior compared to those obtained with the Logit or SVM-Lin classification algorithm, except for FB features. For the last mentioned condition, Logit outperformed XGBoost classifier ($p = 7 \times 10^{-10}$).

Out of all conditions presented in Fig. 8, the best results were obtained using XGBoost classifier with LFCCs combined with binaural cues (LFCC + Binaural) along with the combination of all features (All). For these two "winning" conditions, the discrimination accuracy was similar, being equal to 98.39% and 98.34%, respectively. According to the statistical test

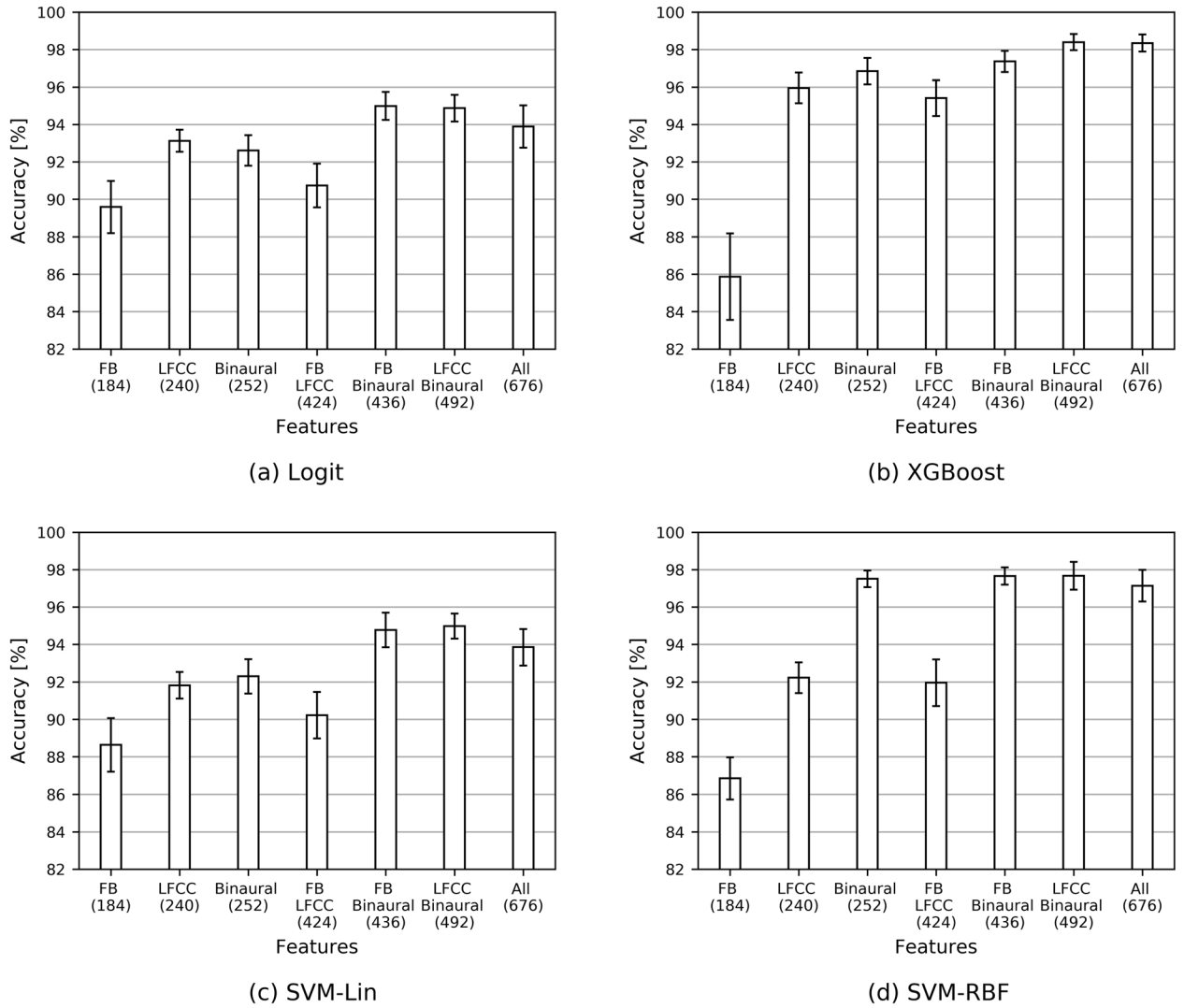

Fig. 8 Mean accuracy scores obtained for combination of features using **a** Logit, **b** XGBoost, **c** SVM-Lin, and **d** SVM-RBF. Numbers in brackets indicate the number of features. Error bars denote standard deviations

of proportions, this difference was not statistically significant (p = 0.83). However, for these two conditions, the mean scores attained for the XGBoost classifier were greater than those obtained for the SVM-RBF classification algorithm. The observed differences were statistically significant, with p values being equal to 4.7×10^{-3} and 9×10^{-3}, respectively. These results show that the XGBoost method outperformed the SVM-RBF algorithm.

The main outcome of this experiment is that all feature groups (FB, LFCC, and Binaural) should be used in the next experiment (experiment 6), with XGBoost employed as the classification algorithm. The discrimination accuracy obtained using the combination of the feature groups is greater than the one obtained with the feature groups used in isolation.

5.5 Experiment 5: Discrimination results using a deep learning approach

The aim of this experiment was to check whether the deep learning-based method, employing CNN, can be used for the discrimination of the front and back located ensembles in binaural recordings of music.

5.5.1 Spectrogram calculation

Similar to the work of Han et al. [28] in the area of the automatic acoustic scene classification, four spectrograms were extracted from each binaural audio excerpt. The spectrograms were calculated for x, y, m, and s signals, respectively. Since the sum (m) and difference (s) signals are linear combinations of the left (x) and right (y) signals, it might be argued that they are redundant, however, they may improve the performance of the

model, as demonstrated in our previous work [8]. Moreover, this approach is consistent with the strategy of feature extraction in the case of the traditional algorithms, described above, as the features were also calculated from the left, right, sum, and difference signals.

Mel-frequency spectrograms are commonly exploited in the deep learning algorithms used for speech recognition, acoustic scene classification, or audio event recognition [28, 49, 62, 63]. While Mel-frequency spectrograms offer a better resolution at low frequencies, it is unknown whether such a type of spectrogram is still superior in terms of discrimination between front and back ensemble locations in HRTF-convolved binaural recordings of music. Therefore, for comparative purposes, both Mel-frequency and linear-frequency spectrograms were employed in this study. Note, that the proposed comparison between Mel-frequency and linear-frequency spectrograms resembles the comparison between Mel-frequency and linear-frequency cepstral coefficients calculation described earlier in the case of the traditional machine learning approach.

The spectrograms were calculated using a frame length of 40 ms with 50% overlap. Since the duration of each audio excerpt was equal to 7 s, each spectrogram represented 349 time frames. A Hamming window was applied to the signal in each frame. Similar to our previous study regarding the spatial scene classification in binaural signals [9], 150 frequency bands were used to calculate the spectrograms. This choice was verified in the first experiment employing CNN described below in Section 5.5.4. The low- and high-frequency limits of the spectrograms were set to 100 Hz and 16 kHz, respectively. The dynamic range of each spectrogram was limited (clipped) to 90 dB relative to its peak value. It was assumed by the authors that spectrogram components below 90 dB relative to their peak level contained noise or spurious artifacts introduced by the music recording equipment or by the HRTF measurement procedures [35]. In theory, these undesired components could be detected and exploited by deep learning algorithms [64]. Therefore, they were filtered out (clipped). In accordance with the typical practice in the area of machine learning [28, 63], the spectrograms were standardized (mean value equalized to zero and the variance normalized to unity) prior to their use in CNN. The spectrograms were calculated in MATLAB using a VOICEBOX toolbox [65].

5.5.2 Topology of the convolutional neural network

A well-proven AlexNet [50] topology was adapted to our purposes. The layout of the network is depicted in Fig. 9. Four spectrograms with an image resolution of 150 × 349 (number of frequency bands × number of time frames) were fed to the input of the network. In the experiment described in Section 5.6, the number of spectrograms N_{ch} was reduced and the results compared.

The network consisted of five convolutional units and three classification units, as illustrated in Fig. 9. Its topology was designed heuristically. Each convolutional unit contained a 2D convolutional layer, followed by an activation layer, a batch normalization layer, and a max pooling layer. The parameters of the network, including the size of the convolutional kernels along with their stride, the number of filters, and a max pooling size, are presented in Fig. 9. As a result of the processing undertaken in the four convolutional layers, the image resolution of the spectrograms was reduced from 150 × 340 to 5 × 5 pixels. The shrunken "images," after flattening to a vector of a length of 3 200, were then directed to the input of the cascade of the classification units. Each classification unit consisted of a dropout layer (to prevent the network from overfitting), fully connected layer, and the activation layer. The dropout rate was adjusted to 50%. The number of nodes in the fully connected layers was set to 64, 32, 16, and 2, respectively, as shown in Fig. 9. A rectified linear unit (ReLU) was used in all the convolutional and dense layers, whereas a Softmax function was applied to the output layer.

5.5.3 Optimization and performance evaluation

The initial version of the network, with the number of trainable parameters exceeding 8 million, was trained and optimized using a grid-search procedure with an incomplete 10-fold cross-validation technique (out 10 folds only five ones were exploited due to long computational time). In the above procedure, the training dataset was shuffled randomly. Then, it was split into 10 folds. Subsequently, a single fold was taken for validation whereas the remaining nine folds were employed for training. The last mentioned step was repeated five times, each time with a different fold selected for validation (in the standard 10-fold cross-validation procedure the last mentioned step is repeated 10 times). For each fold, the data between the train and validation sets were divided in the HRTF-independent manner (90% of HRTFs were used for training and 10% for validation). The learning rate values were selected from the set $l_r \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. The batch size values considered during the grid-search procedure were selected from the set $b \in \{16, 32, 64, 128, 256\}$. In combination with the above grid-search technique, the network topology was adjusted heuristically by trial and error. The reduction of the network size from 8 million to approximately 400 thousand trainable parameters, accomplished by reducing the number of layers, substantially improved its performance. The final topology of the network, with the best hyper-parameters, was already presented in Fig. 9 above.

After adjusting the topology and the hyper-parameters of the network, described above, it was trained on the whole training set and then tested using the test set. In this procedure, the initial value of the learning rate was set to 5×10^{-3}. The learning rate was reduced adaptively

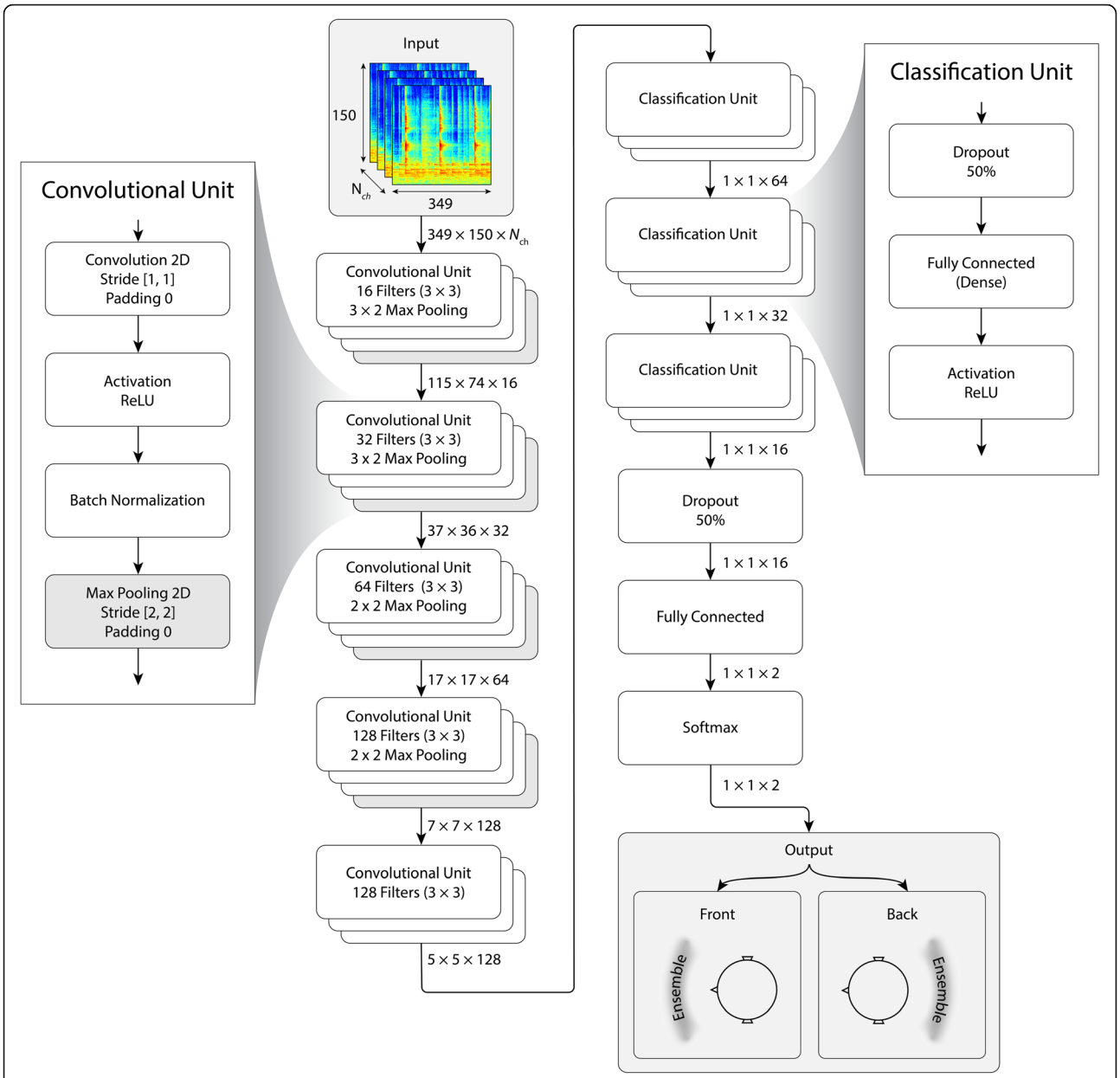

Fig. 9 CNN topology used for the discrimination of front and back ensemble locations

with a rate drop period of 5 epochs and a rate drop factor of 0.5 (the learning rate was halved in value every 5 epochs). Batch size was set to 128. Cross-entropy was used as a loss function. The training procedure was terminated after 30 epochs.

In contrast to the traditional classification algorithms, which were iteratively trained and tested seven times, unless otherwise stated, CNN was trained and tested three times, due to much longer computational time. In order to undertake a "like-for-like" comparison, for both methods in each iteration the same seed was applied in the routine undertaking the train and test data splits. The mean discrimination accuracy scores calculated across the iterations are presented in the remainder of the paper. For each iteration, the same splits between the train and test sets were used for both the traditional algorithms and CNN.

5.5.4 Results

Figure 10 shows the results obtained using CNN with two types of spectrograms at its input: linear-frequency spectrograms (solid and dotted lines) and Mel-frequency spectrograms (dashed line). To undertake this experiment, some adjustments to the original CNN topology (shown in Fig. 9) had to be made. For the frequency

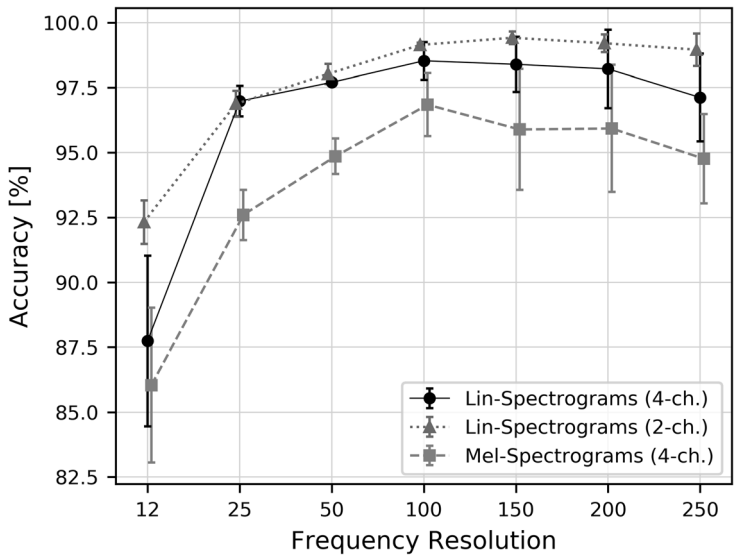

Fig. 10 Influence of the frequency resolution and type of the spectrograms on the discrimination results using CNN. Error bars denote standard deviations

resolution reduced from 150 to 12 frequency bins, the network was simplified by removing three convolutional units. For the conditions where the number of frequency bins was set to 25 or 50, the number of removed convolutional units was equal to 2 and 1, respectively.

It can be seen in Fig. 10 that, similarly to the outcomes of the experiment investigating the difference between MFCCs and LFCCs described earlier in Section 5.2.2, linear spectrograms provide better results compared to their Mel-scale counterparts. Interestingly, the relationship between the frequency resolution and the accuracy does not appear monotonic. For example, for linear spectrograms, the maximum discrimination accuracy is obtained for the frequency resolution ranging from 100 to 200 frequency bins (the effect of the diminishing accuracy for the frequency resolution exceeding 200 frequency bins could probably be attributed to the overfitting effect). Therefore, linear frequency scale spectrograms with a frequency resolution set to 150 frequency bins were exploited in the subsequent experiments described in the remainder of this paper.

In addition to demonstrating the influence of the spectrogram types on the obtained results, Fig. 10 also shows that the results depend on the number of channels (spectrograms) employed at the network input. The dotted line represents the case when the number of spectrograms was reduced to 2 (which were derived from the left and right channels). The remaining two lines illustrate the results for the original case with all 4 spectrograms utilized at the network input. Surprisingly, the accuracy levels obtained for the reduced 2-channel scenario tended to exceed those obtained for the original 4-channel spectrograms for most of the examined frequency resolutions.

The following conclusions can be drawn from this experiment: (1) deep learning method, employing CNN, can also be used for the discrimination of the front and back located ensembles in binaural recordings of music; (2) linear-frequency spectrograms should be exploited at the CNN's input, rather than the Mel-frequency ones, with their frequency resolution set to 150 frequency bins; (3) the effects of the number and types of spectrograms should be further investigated, which was the topic of the subsequent experiment, described in the next section.

5.6 Experiment 6: Investigating the effect of the number and type of audio channels

The previous experiment, as a side outcome, indicated that the reduction of input channels from 4 to 2 could improve the results, which could be considered an intriguing effect. The aim of this experiment, therefore, was to investigate the influence of the number and type of audio channels used at the input of both the best-performing traditional algorithm (XGBoost) and the CNN algorithm.

The methodologies of the hyper-parameter optimization, training, and testing the models will not be reiterated here, as they were the same as before in the case of the XGBoost algorithm and CNN, respectively. They were described earlier in Sections 5.4.1 and 5.5.3.

Figure 11 illustrates the effect of selecting audio channels on the obtained results for two best-performing methods: XGBoost and CNN. Observe that for the traditional algorithm (XGBoost) reducing the number of channels from 4 to 3 (LRM and LRS), or even down to 2 (LR), has almost no effect on the results. This means that the complexity of the model may be reduced without compromising the results. Such modification should enhance the generalization

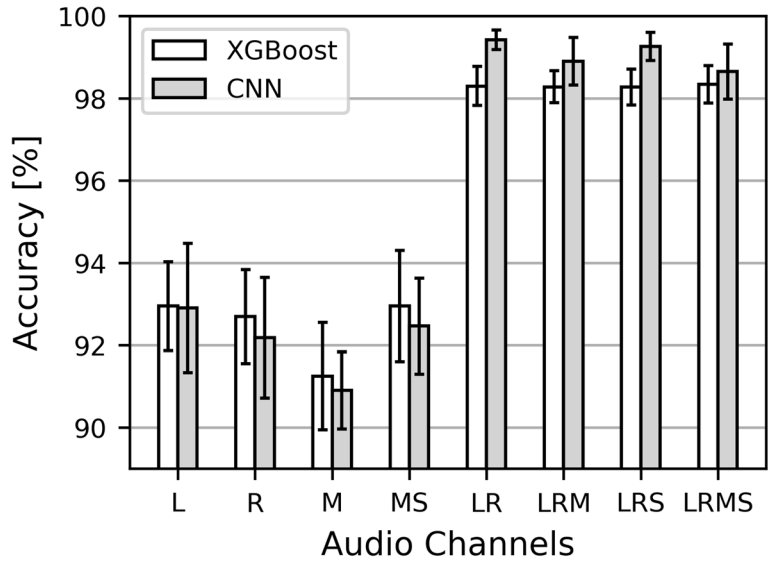

Fig. 11 The effect of audio channels on the discrimination results. Error bars denote standard deviations.

property of the model, as simpler models tend to be more generalizable [51]. For this reason, in the subsequent experiments involving XGBoost classifier, it was decided to retain only the features extracted from the left and right channels (ignoring the features calculated from *m* and *s* signals). Note, that in the original dataset (before a feature set reduction) there were 464 features extracted from the left and right channels and 212 features obtained from the sum and difference channels. Consequently, the number of features used by XGBoost classifier was reduced from 676 to 464, maintaining very good performance, with the accuracy scores exceeding 98%.

In contrast to XGBoost classifier, the results obtained for CNN show that the reduction of the number of channels from 4 to 3, or even down to 2, improves the results. For example, the mean accuracy score obtained for the two-channel scenario (LR) was exceptionally good, with an accuracy of 99.42%, outperforming the XGBoost algorithm by 1.12%. The observed difference was significant at $p = 8.6 \times 10^{-9}$, according to the statistical test of proportions. Therefore, in the subsequent experiments involving CNN, described below, only spectrograms obtained from the left and right channels were retained and used by the network.

Note, that using only the sum and difference channels (MS) results in a relatively high discrimination accuracy, ranging from 92 to 93%. Interestingly, for both XGBoost and CNN classification algorithms, the reduction of the number of channels to a single channel (left or right) still results in the mean accuracy scores exceeding 90% (see Fig. 11). This means that the proposed method is still capable of undertaking a discrimination task at the acceptable accuracy level even when "listening" to the binaural recordings with a single "ear."

An example of a confusion matrix obtained for CNN with the number of spectrograms reduced to 2 (LR) is presented in Fig. 12a. The matrix demonstrates its remarkable performance with only 6 misclassified recordings for frontally located ensembles and 10 erroneously made classifications for back located ensembles, relative to the total of 5920 recordings exploited in the test set. For reference, an example confusion matrix for the XGBoost classification algorithm is presented in Fig. 12b.

The main outcome of this experiment is that the previously developed models (XGBoost and CNN), exploiting four audio channels (left, right, sum, and difference) can be simplified by utilizing only the two audio channels (left and right) without compromising their performance (XGBoost) or even with improving the discrimination accuracy (CNN). Therefore, the models used in the next two experiments were adjusted accordingly.

5.7 Experiment 7: Quantification of the effect of audio recording duration

The aim of this experiment was to investigate the effect of audio recording duration on the discrimination accuracy. The duration of the audio excerpts used in all the previously described experiments was constant, being equal to 7 s. In this experiment, the duration was progressively decreased, while examining its impact on the discrimination accuracy examined. In line with the outcome of the previous experiment, the features (for the traditional algorithms) and spectrograms (for CNN) were calculated based only on the left and right channel signals. The remaining experimental protocol was intact.

Figure 13 illustrates the influence of the duration of the binaural audio recordings on the discrimination

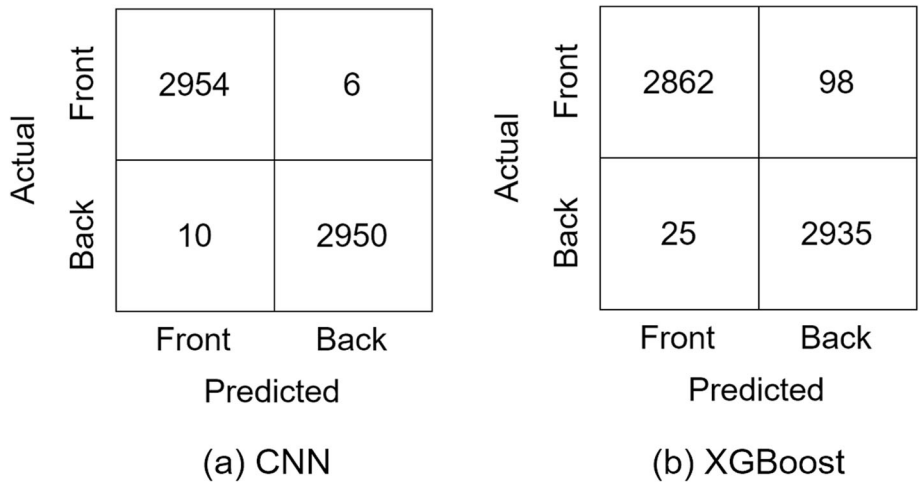

Fig. 12 Examples of confusion matrices obtained for the best-performing methods: **a** CNN (LR), **b** XGBoost (LR)

results. For clarity, the results obtained with Logit classifier were omitted here, as they were similar to the results achieved using the SVM-Lin classification algorithm. According to the results, signals with a minimum duration of 3 s are required before the accuracy curves flatten. Extending the duration beyond this value results in diminishing improvements in accuracy. While reducing the duration of the audio excerpts to 200 ms caused expected deterioration in the discrimination accuracy, for XGBoost and CNN classifiers its level was still high, amounting to approximately 95%, which could be considered to be a surprising outcome.

The main outcome of this experiment is confirmation that a 7-s duration of the binaural audio excerpts selected at the outset of the study is adequate for the discrimination method developed in this work. A minimum duration of 3 s is required for the above task. Moreover, CNN showed a very high discrimination accuracy (99.42%), outperforming the traditional method (XGBoost).

5.8 Experiment 8: Final model testing

The exceptionally good performance exhibited by the CNN method, with a discrimination accuracy equal to 99.42%, may cast some doubts on its generalization property, as this may be an indicator of the over-fitting effect. For this reason, it is imperative to test its performance not only under HRTF-dependent but also under HRTF-independent conditions. The aim of the

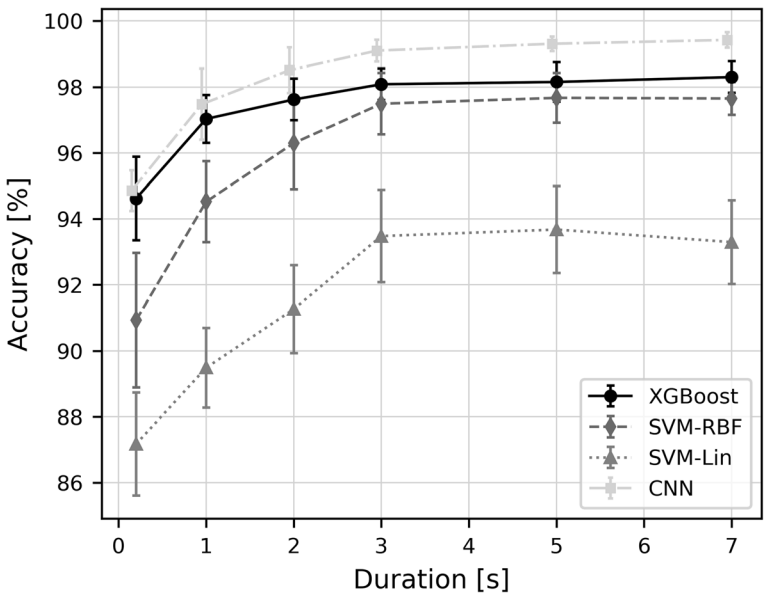

Fig. 13 The effect of duration of the audio excerpts on the discrimination results. Error bars denote standard deviations

experiment is to compare the performance of the "final" models under HRTF-independent conditions.

5.8.1 Method

Under the HRTF-dependent test scenario, adopted in all the above experiments, the original data set was split into the train and test sets in such a way the tests were music recording independent (different recordings were used in the train and test sets). The procedure of the data splitting was as already illustrated in Table 1 and described in detail in Section 4.4 above. In the experiments described in this section, the same procedure was taken in order to ensure that the train and test sets are music recording independent. However, in contrast to the above experiments, the additional "data filtering" procedures were taken, to ensure that the train and test sets are not only music recording-independent but also independent in terms of the HRTF sets used to generate such recordings. Note, that the abovementioned data filtering procedures were performed "within" the train and test sets. In other words, first, the original data set was split into the train and test sets according to the same procedure which was taken in the all above experiments, and, second, the train and test sets were filtered in such a way that they were HRTF-unique, which is specific to the experiments described in this section. The reason for filtering the train and test data, warranting that they were both music recording-independent and HRTF-independent, was to enhance the validity of the generalization performance testing. The three different procedures for the data filtering were taken. They are described below.

The first proposed technique of HRTF-independent test is based on the assumption that each "corpus" of HRTFs measured in a given institution is unique in terms of their characteristics (HRTFs bear an individual electro-acoustical "fingerprint," specific to the head used, the distance between the head and the loudspeakers, the frequency response of the microphones and the loudspeaker, spatial resolution, acoustical properties at low frequencies, a type of low-frequency extension, just to mention a few factors). This method was undertaken iteratively 13 times. In each iteration, a single corpus of HRTFs was "filtered out" from the training set. For example, in the first iteration, HRTFs No. 1 and 2, measured at AACHEN, were excluded from the training set, while the test set included solely AACHEN HRTFs. In the next iteration, HRTFs No. 3–9 measured at ARI were excluded from the training set, whereas the test set included solely ARI HRTFs, etc. In total, 13 corpora were used in this method. They are outlined in Table 3 in alphabetical order according to their acronym.

In the second technique of the HRTF-independent testing, a given number of HRTFs, randomly selected from the original set of 74 HRTFs, was retained in the training set, while the test set was reduced to HRTFs that were taken out of from the training set. For example, if only the two following HRTFs were retained in the training set, namely HRTFs No. 16 and 35 (effectively reducing the number of HRTFs in the training set from 74 to only 2), then the original test set was reduced from 74 to 72 HRTFs (HRTFs No. 16 and 35 were "filtered out" from the test set). In this experiment the total number of HRTFs n_r retained in the training data set was selected from the set of $n_r \in (2, 10, 20, 30, 50, 70)$. The HRTF sets were selected using a random sampling technique without replacement. In order to ensure the randomness of the selection, the testing procedure was repeated seven times for all the examined algorithms. For each repetition, the seed in the pseudorandom generator was the same across the algorithms, to maintain the consistency of the comparison.

In the third technique, the train set was reduced to the human HRTFs, whereas the test set was limited to the artificial ones. In other words, the models were trained on the human HRTFs and tested on the artificial ones. Then, the procedure was reversed, as the models were trained on the artificial and tested on the human HRTFs, with the results compared. The total number of human and artificial HRTFs used in the original set of 74 HRTFs was relatively balanced, being equal to 36 and 38, respectively.

5.8.2 Results

Figure 14 shows the results obtained under the HRTF-independent test using the first technique that was described above. While for some HRTF corpora, such as ARI, SADIE, and TU-BERLIN, both discrimination methods (XGBoost and CNN) yielded high discrimination accuracy levels (exceeding 95%), for TH-KÖLN corpus, the obtained accuracy was much lower, ranging between 80 and 90%. A possible reason for such low accuracy results obtained for TH-KÖLN corpus is that it included the "unusual" HRTFs measured using the artificial head manufactured by Head Acoustics equipped with a baseball cap or the headset by Oculus Rift.

The weighted mean accuracy score, compensated for an imbalanced number of HRTFs across corpora, calculated for the HRTF-independent test for the CNN and XGBoost methods was equal to 93.65% and 94.51%, respectively. According to the statistical test of proportions, the difference was significant ($p = 0.047$). Thus, this outcome indicates that the XGBoost technique exhibits a slightly better generalization performance compared to CNN.

For the CIPIC corpus, the XGBoost method gave markedly better results compared to those obtained using CNN, while the opposite outcome could be observed for the TH-KÖLN corpus. Identifying the reasons for such an interaction would require undertaking a separate experiment, which is beyond the scope of this study. However, it is reasonable to suggest that there is a difference in information captured by the hand-engineered features (used by XGBoost) and by spectrograms (utilized by CNN). This difference could become evident when the models are tested

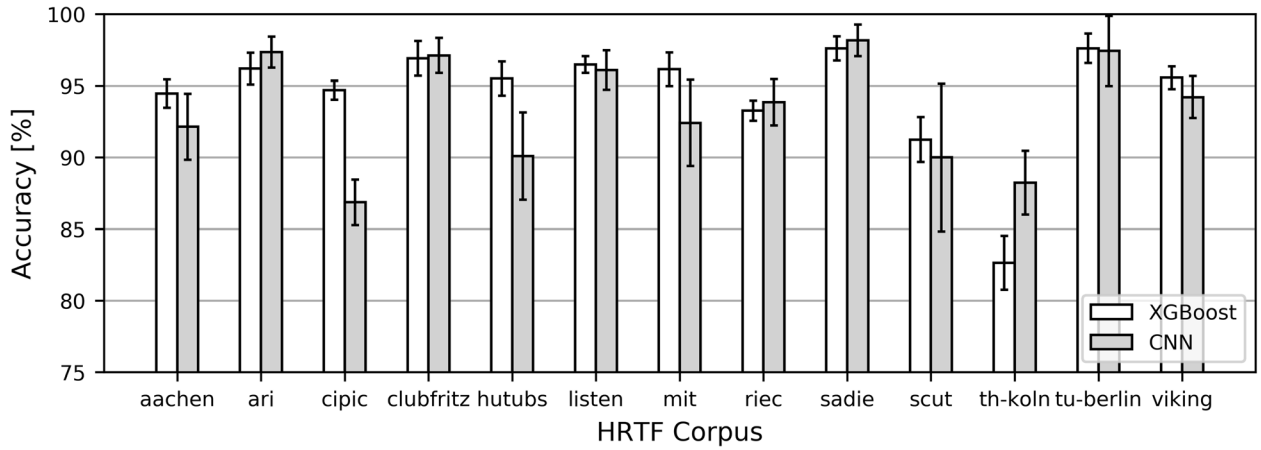

Fig. 14 HRTF-independent test results for separate corpora of HRTFs obtained using the best performing traditional discrimination method (XGBoost) and CNN. Graph shows mean discrimination accuracy scores with associated standard deviations. Acronyms of HRTF corpora are explained in Table 3 in the Appendix

under new or unusual HRTFs. Therefore, it might be hypothesized that the observed effects could be attributed to the way information is extracted from the binaural signals and fed to the classifiers, rather than due to the differences between the classification algorithms themselves.

The influence of the total number of randomly included HRTFs in the training set on the discrimination results is illustrated in Fig. 15. Note that this procedure also constitutes a form of another HRTF-independent test, as the HRTFs selected for training were at the same time excluded from testing. For clarity, the results obtained for Logit classifier were omitted, since they were very similar to those obtained using the SVM algorithm with the linear kernel (SVM-Lin). According to the results, for the two best-performing traditional methods (XGBoost and SVM-RBF), at least 20 HRTFs must be included in the training set in order to attain an accuracy level exceeding 90%. However, for CNN, the minimum number of HRTFs to achieve a similar level of accuracy is higher, being equal to 30. This discrepancy could be explained by the property of CNN. Compared to the traditional classification technique, CNN's performance might be substantially degraded when relatively small data sets are used for training (for 20 HRTFs the number of training excerpts is equal to 1600, whereas for 10 HRTFs the above number is halved to 800).

The third HRTF-independent test applied in this study aimed to verify what happens when the methods are trained solely on human HRTFs and tested on the artificial ones (and vice versa). The outcomes of the tests are presented in Table 2. For both scenarios XGBoost classifier with hand-engineered features produces the best results, with the mean accuracy scores being approximately equal or slightly exceeding 90% (proving its relatively good generalization property). When the models were trained on the human HRTFs and tested on the artificial ones, the XGBoost method exhibited a slightly better performance compared to CNN with a difference of only 2 percentage points. However, a more noticeable difference (almost 7 percentage points) was observed between the performance of XGBoost and CNN when the models were trained on the artificial HRTFs and tested on the human ones.

5.9 Follow-up data exploration

The aim of the additional analysis and tests described in this section was to get a better understanding of how the discrimination methods worked.

5.9.1 Exploring feature importance

In order to better understand the importance of the features, as assessed by the best-performing traditional method (XGBoost), a feature importance analysis was made, with the outcomes illustrated in Fig. 16. The figure shows the twelve most important features ranked according to their "gain" (estimated by the XGBoost algorithm), representing the fractional contribution of each feature to the model based on the total gain of a given feature's split [66]. The mathematical formula for XGBoost gain can be found in [67]. Observe that the mean value of the 4^{th} LFCC coefficient for the left channel signal was ranked as the most important feature. The next two most important features are standard deviations of ILD measured at frequency bands 30 and 31 of the gammatone filter bank. These two channels have center frequencies equal to approximately 5.5 kHz and 6 kHz, respectively. This means that fluctuations of ILDs within those frequency bands play an important role in the discrimination between front and back located ensembles. The 4th important feature ranked in Fig. 16 is the mean value of the spectral roll-off point of the left channel, accounting for the audio bandwidth of the binaural excerpts.

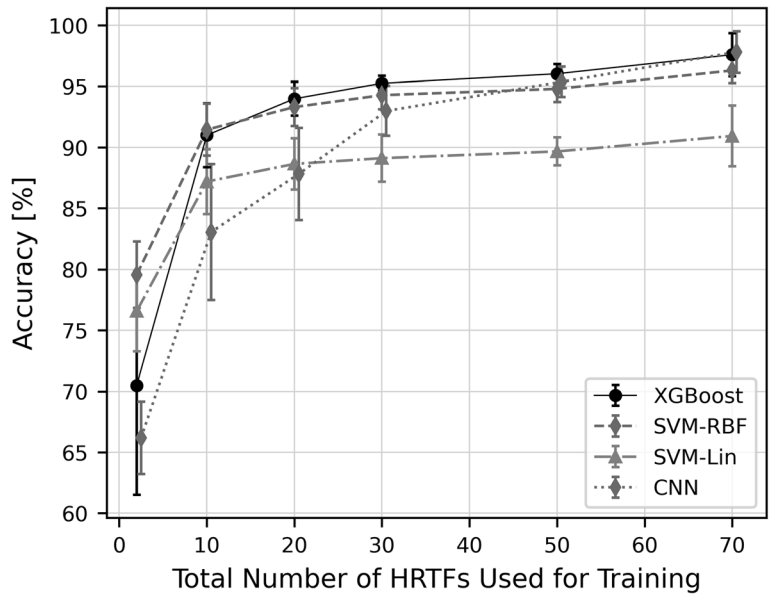

Fig. 15 Discrimination mean accuracy scores obtained under HRTF-independent test as a function of the total number of HRTFs used for training. Error bars denote standard deviations

The next two most important features are associated with the 10th and 25th cepstral coefficients.

5.9.2 CNN visualization

In order to identify which parts of the spectrograms are treated as the most important by CNN, a gradient attribution maps (GAM) [10] technique was performed. It was selected due to a better frequency resolution compared to a more popular gradient-weighted class activation mapping (Grad-CAM) technique [68]. The results are presented in Fig. 17 for four experimental repetitions (in each repetition CNN was trained from its random initial conditions for different train-test data splits). The graphs represent the activation maps averaged across all the audio excerpts in both training and testing sets. The interpretation of the obtained results proved to be difficult due to many very small differences between the spectrograms. The visualization maps also changed considerably between repetitions. However, after averaging the maps across repetitions, interesting pictures emerged (see Fig. 18). Namely, for both the front and back located ensembles, a frequency region between 5 kHz and 6 kHz was identified as the most important. This observation supports a view that there exists a universal frequency band that is used by CNN to undertake the discrimination task. The latter supposition is also supported by the observations made in the previous section, concerning the importance of features, whereby frequency bands centered at approximately 5.5 kHz and 6 kHz were identified to be important in the discrimination process.

6 Discussion

The distinct features of the developed method are that it incorporates a static-head approach and is assumption free with regard to the number and type of music audio sources in analyzed scenes, making it suitable for a wide range of spatial audio applications. The developed method was thoroughly tested using a broad range of music recordings with the number of individual sources ranging from 5 to 62, both under the HRTF-dependent and HRTF-independent conditions. The results presented in this paper cannot be compared directly to those obtained in our previous work [8]. While in the present and former studies a similar deep learning technique was used (CNN), there were fundamental methodological differences preventing the authors from

Table 2 Test results obtained under HRTF-independent test according to the type of HRTFs used (human versus artificial heads). The table shows mean discrimination accuracy scores with associated standard deviations. Numbers in bold type represent the maximum mean accuracy scores

HRTFs used for training	HRTFs used for testing	Discrimination scores [%]				
		XGBoost	SVM-RBF	SVM-Lin	Logit	CNN
Human	Artificial	**89.7** (1.0)	85.7 (1.1)	85.2 (1.5)	83.9 (1.7)	87.9 (1.8)
Artificial	Human	**92.3** (0.7)	91.1 (0.9)	84.4 (1.0)	84.7 (1.0)	85.6 (1.1)

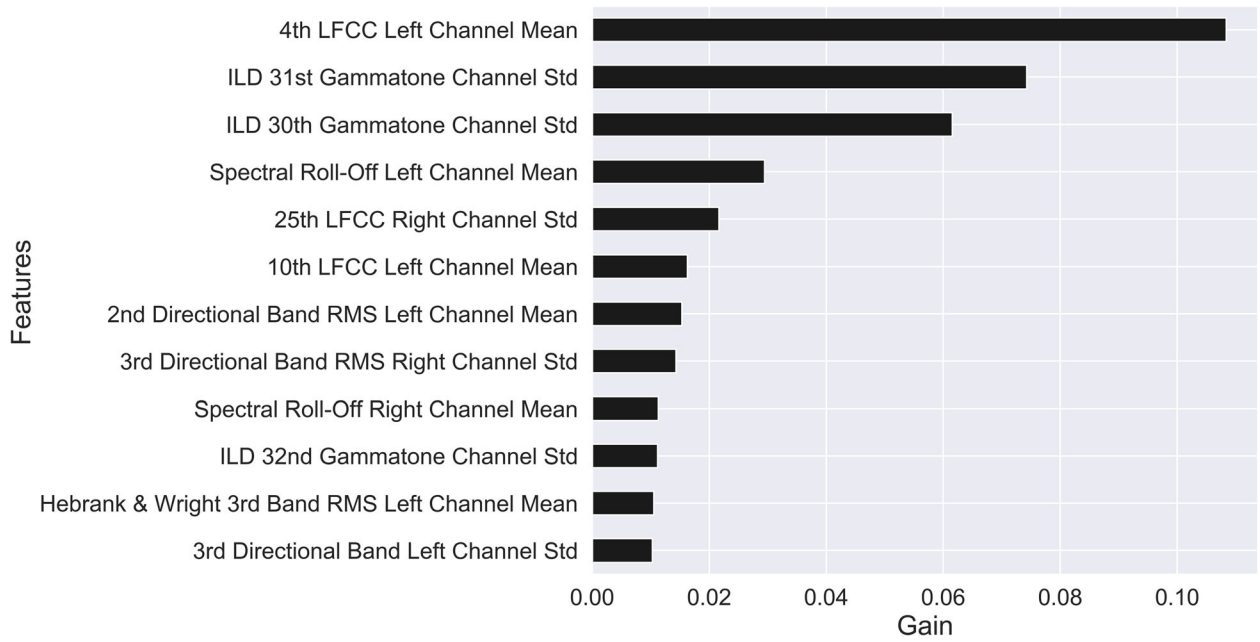

Fig. 16 Feature importance for the twelve most prominent features as identified by XGBoost classification algorithm ("Std" denotes standard deviation)

undertaking a consistent comparison. For example, in this work, only two ensemble locations were discriminated (front or back), while in the previous study three types of spatial audio scenes were classified (ensemble in front, ensemble at the back, and ensembles located concurrently in front and at the back of a listener). Moreover, in contrast to this study which was based on the "anechoic" head-related impulse responses (HRIRs), in our previous work, we used recordings generated using BRIRs. Employing HRIR sets as opposed to BRIR sets could be a confounding factor in terms of the localization of audio sources under binaurally rendered conditions. While Mel-frequency spectrograms were employed in our former work, in this research project we proved that using linear-frequency spectrograms at the input of CNN yields better results. To understand how CNN works, in our previous work we visually inspected filter activations at the output of the convolutional layers, with rather inconclusive results. In this paper, we explored the model using the GAM technique, providing some evidence that the frequency region between 5 and 6 kHz is most important in terms of the "decision making" by CNN. Finally, while the CNN model used in our previous experiment was tested only under HRTF-dependent conditions, the model developed in this work was examined both under HRTF-dependent and HRTF-independent conditions, which constitutes an added value of the study and demonstrates the importance of "generalizability" testing.

While machine learning algorithms do not have to mimic human auditory system to discriminate between recordings, as in principle they can exploit a different approach compared to humans, according to the results obtained using both the traditional and deep learning methods, there is some evidence that machine learning algorithms tend to predominantly exploit cues at a frequency band centered at approximately 5 kHz when undertaking the discrimination task. The above frequency band is similar to the fourth boosted band and the third directional band identified by Blauert [16] in his studies that investigate spatial hearing in humans.

The ability of the model to disambiguate front and back located ensembles with only one simulated ear may appear to be intriguing. However, this outcome is in accordance with psychoacoustics as, to some extent, humans are also able to disambiguate front and back audio sources exploiting only one ear [69].

There are five limitations of this study which should be acknowledged. First, the spatial distribution of individual audio sources within each ensemble was restricted to the horizontal plane only. Second, the ensemble widths were restricted to ± 30°, with the symmetric boundaries between front and back hemispheres. Inclusion of elevated audio sources with varying limits of the ensembles within the experimental protocol are left for future experiments. Third, the experimental progression shown in the outline in Fig. 2 is not the only one that could be adapted for such a study, and, consequently, may not be the optimal one. For example, investigating the effect of audio excerpt duration "before" quantifying the effect of the number and type of audio channels (swapping the order of the experiments) could lead to slightly different results. Fourth, the effect of the temporal resolution (the frame length) set during the feature extraction and during the spectrograms calculations

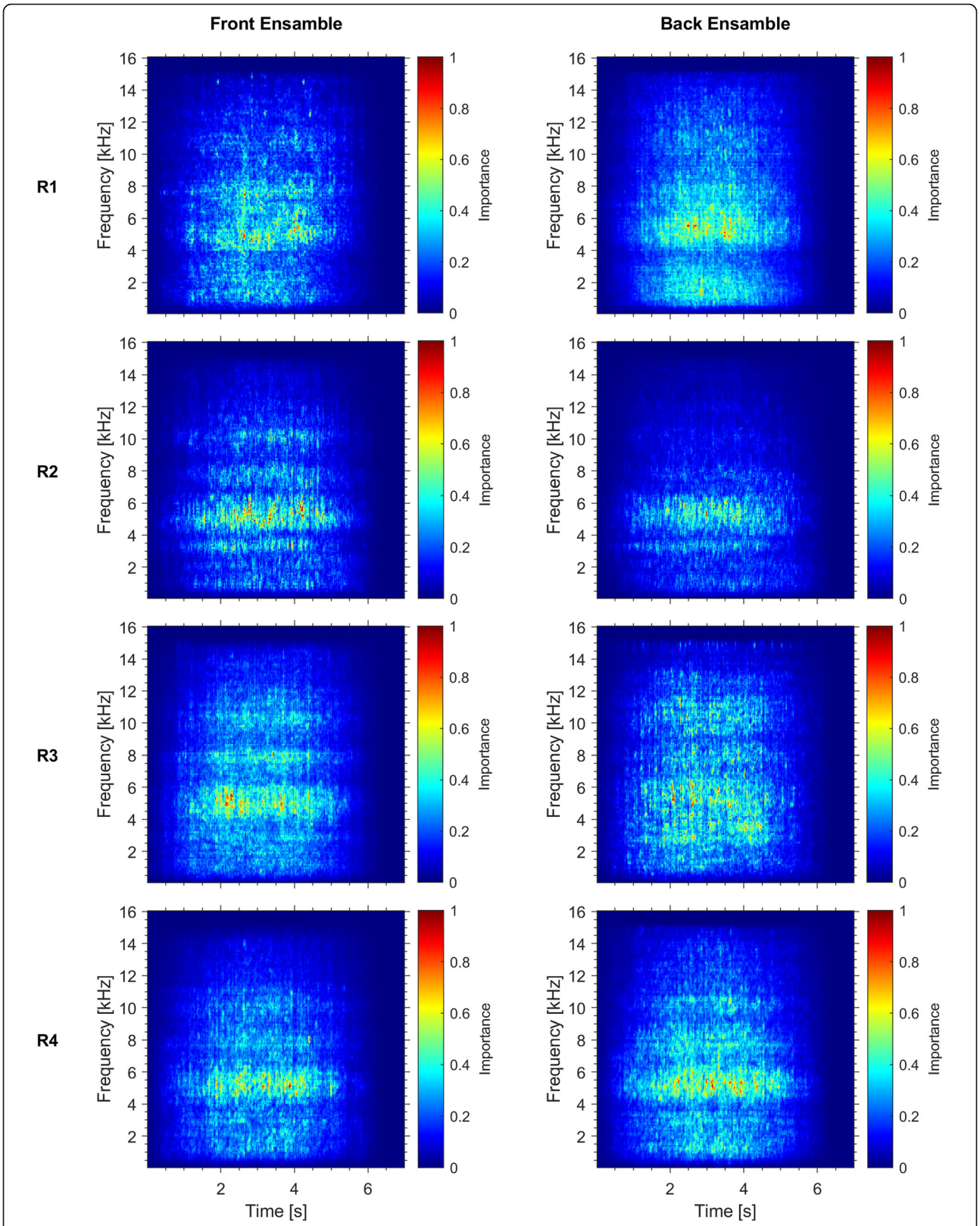

Fig. 17 Examples of the most important spectrogram regions identified using the GAM technique. The left and right columns represent the average importance maps for the front and back ensembles, respectively. The rows illustrate the results obtained with the same CNN for four separate experiment repetitions (*R*1, *R*2, *R*3, and *R*4)

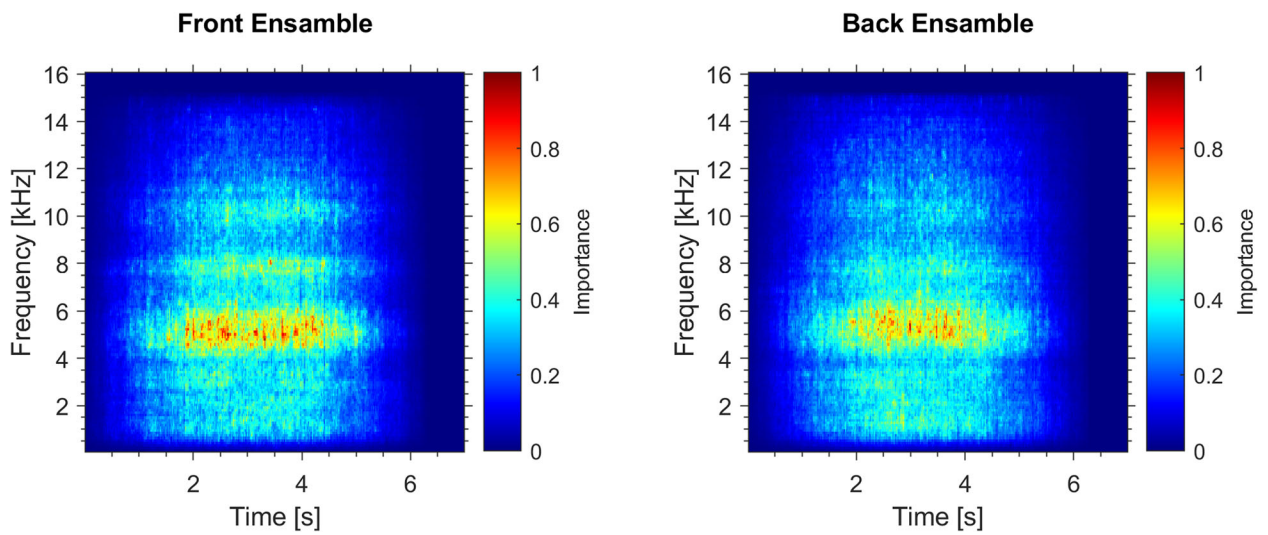

Fig. 18 Gradient attribution maps averaged across experiment repetitions: **a** front ensemble, **b** back ensemble

was not quantified in this study (this factor is left for future experimentation). Fifth, some similarities between 74 HRTF sets used in this study might exist, e.g., due to employing the same types of the artificial heads, reducing the validity of "HRTF-independent" tests. Since there are no widely accepted metrics of independence between the HRTFs, it is left to the reader to judge how similar the HRTFs were, based on the detailed description of the HRTFs used in the study provided in Table 3 in the Appendix. Nevertheless, it must be stressed that, to the best of the authors' knowledge, this study is one of the most comprehensive in terms of the HRTF-independent tests.

7 Conclusions

The aim of this study was to develop a method of discriminating between front and back located music ensembles in binaural recordings and to quantify the influence of the selected parameters on its performance. According to the results obtained under HRTF-dependent test conditions, CNN showed a very high discrimination accuracy (99.4%), slightly outperforming the traditional method (XGBoost). However, under the HRTF-independent test scenario, CNN performed worse than the traditional algorithm, highlighting the importance of testing the algorithms under HRTF-independent conditions and indicating that the traditional methods may exhibit a better generalization property compared to CNN. Even simple RMS estimators of bandlimited signals, designed based on psychoacoustic literature, provide a discrimination accuracy approaching almost 90%. The proposed method is still capable of undertaking a discrimination task at the level of approximately 92% even when utilizing only one simulated ear (left or right).

Linear-frequency spectrograms give better results when used at the CNN input than Mel-frequency ones. Likewise, linear-frequency cepstral coefficients provide better discrimination results compared to Mel-frequency cepstral coefficients when applied to the traditional classification algorithms. Audio excerpts with a minimum duration of 3 s are adequate for the automatic discrimination of ensemble locations. Out of the selection of 74 HRTFs considered in this study, a minimum of 20 HRTFs were required during the development of the traditional algorithms in order for them to achieve satisfactory generalization performance. For CNN, the above number is greater, amounting to a minimum of 30 HRTFs.

Linear-frequency cepstral coefficients, interaural level differences, and audio bandwidth were identified as the key descriptors facilitating the discrimination process using the traditional approach. There is some evidence that machine learning algorithms (both the traditional ones and CNN) tend to predominantly exploit cues at a frequency band centered at approximately 5 kHz while undertaking the discrimination task.

The current study was limited to the anechoic conditions. Adapting the methodology to reverberant environments constitutes a future research progression. Prospective experiments may also investigate whether the developed method could be integrated with the binaural localization models to help reduce their localization errors caused by the front-back confusion effect. Moreover, upcoming work may involve extending the method to encompass the quantification of ensemble width, depth, and height.

8 Appendix

Table 3 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 [32]	AACHEN
2.	Artificial	GRAS 45BB-4 KEMAR	1		
3.	Human	Subject 2	1.2	Austrian Academy of Sciences [33]	ARI
4.	Human	Subject 4	1.2		
5.	Human	Subject 5	1.2		
6.	Human	Subject 8	1.2		
7.	Human	Subject 10	1.2		
8.	Artificial	ARI Printed Head	1.2		
9.	Artificial	Neumann KU 100	1.2		
10.	Human	Subject 012	1	CIPIC Interface Laboratory, University of California [34]	CIPIC
11.	Human	Subject 015	1		
12.	Human	Subject 020	1		
13.	Human	Subject 028	1		
14.	Human	Subject 051	1		
15.	Human	Subject 147	1		
16.	Human	Subject 148	1		
17.	Artificial	Neumann KU 100	1.95	IRCAM (2004) [35]	CLUBFRITZ
18.	Artificial	Neumann KU 100	0.9	NASA (2007) [35]	
19.	Artificial	Neumann KU 100	2	IRCAM (2007) [35]	
20.	Artificial	Neumann KU 100	1.5	Helsinki University of Technology (2009) [35]	
21.	Artificial	Neumann KU 100	1.3	NHK (2009) [35]	
22.	Artificial	Neumann KU 100	1.3	NICT (2009) [35]	
23.	Artificial	Neumann KU 100	1	Nagoya University (2009) [35]	
24.	Artificial	FABIAN	1.47	Technical University Berlin, Huawei Technologies, Munich Research Centre, Sennheiser Electronic [36]	HUTUBS
25.	Human	Subject pp2	1.47		
26.	Human	Subject pp3	1.47		
27.	Human	Subject pp4	1.47		
28.	Human	Subject pp5	1.47		
29.	Human	Subject pp6	1.47		
30.	Human	Subject pp7	1.47		
31.	Human	Subject 1003	1.95	IRCAM, AKG [37]	LISTEN
32.	Human	Subject 1046	1.95		
33.	Human	Subject 1054	1.95		
34.	Human	Subject 1006	1.95		
35.	Human	Subject 1002	1.95		
36.	Human	Subject 1004	1.95		
37.	Human	Subject 1005	1.95		
38.	Artificial	KEMAR DB-4004 (DB-061)	1.4	MIT [38]	MIT
39.	Artificial	KEMAR DB-4004 (DB-065)	1.4		
40.	Human	Subject 001	1.5	Tohoku University [39]	RIEC
41.	Human	Subject 002	1.5		
42.	Human	Subject 003	1.5		
43.	Human	Subject 004	1.5		

Table 3 List of HRTF sets used to synthesize binaural audio excerpts *(Continued)*

No.	Type	Head	Radius [m]	Source	Acronym
44.	Human	Subject 005	1.5		
45.	Artificial	Koken SAMRAI	1.5		
46.	Artificial	KEMAR	1.5		
47.	Artificial	Neumann KU 100	1.2	University of York [40]	SADIE II
48.	Artificial	GRAS 45BC KEMAR	1.2		
49.	Human	Subject H3	1.2		
50.	Human	Subject H4	1.2		
51.	Human	Subject H5	1.2		
52.	Human	Subject H6	1.2		
53.	Human	Subject H7	1.2		
54.	Artificial	KEMAR	1	South China University of Technology [41]	SCUT
55.	Artificial	KEMAR	0.5		
56.	Artificial	Neumann KU 100	0.5	TH Köln [42]	TH Köln
57.	Artificial	Neumann KU 100	0.75		
58.	Artificial	Neumann KU 100	1		
59.	Artificial	Neumann KU 100	1.5		
60.	Artificial	Head Acoustics HMSII	2		
61.	Artificial	Head Acoustics HMSII + baseball cap	2		
62.	Artificial	Head Acoustics HMSII + Oculus Rift	2		
63.	Artificial	FABIAN	1.7	TU Berlin [43, 44]	TU Berlin
64.	Artificial	GRAS 45BA KEMAR	0.5		
65.	Artificial	GRAS 45BA KEMAR	1		
66.	Artificial	GRAS 45BA KEMAR	2		
67.	Artificial	GRAS 45BA KEMAR	3		
68.	Artificial	GRAS 45BB-4 KEMAR – subject A attachment	1	Aalborg University; University of Iceland [45, 46]	VIKING
69.	Artificial	GRAS 45BB-4 KEMAR – subject B attachments	1		
70.	Artificial	GRAS 45BB-4 KEMAR – subject C attachments	1		
71.	Artificial	GRAS 45BB-4 KEMAR – subject D attachments	1		
72.	Artificial	GRAS 45BB-4 KEMAR – subject E attachments	1		
73.	Artificial	GRAS 45BB-4 KEMAR – subject F attachments	1		
74.	Artificial	GRAS 45BB-4 KEMAR – subject G attachments	1		

Table 4 Boundaries of the frequency bands conducive for front-back discrimination (see Fig. 3)

Study	Acronym	Direction	Frequency limits
Blauert (1969/70) [16] (Boosted Bands)	B1	Front	150–540 Hz
	B2	Back	0.72–1.7 kHz
	B3	Front	1.9–2.9 kHz
	B4	Front	3.6–5.8 kHz
	B5	Back	7.4–11.1 kHz
Blauert (1969/70) [16] (Directional Bands)	D1	Front	0.28–0.56 kHz
	D2	Back	0.72–1.8 kHz
	D3	Front	2.9–5.8 kHz
	D4	Back	10.3–14.9 kHz
Hebrank and Wright (1974) [17]	HW1	Front	4–10 kHz
	HW2	Back	10–12 kHz
	HW3	Front	13–16 kHz
Langendijk and Bronkhorst (2002) [20]	LB	Front/back	8–16 kHz

Abbreviations

BRIR: Binaural-room-impulse-response; AACHEN: Aachen University; ARI: Acoustics Research Institute; BB: Boosted bands; CIPIC: Center for Image Processing and Integrated Computing; CNN: Convolutional neural network; CPU: Central processing unit; DB: Directional bands; FB: Front-back cues; GAM: Gradient attribution maps; Grad-CAM: Gradient-weighted class activation mapping; GPU: Graphics processing unit; HRIR: Head-related impulse response; HRTF: Head-related transfer function; HW: Hebrank and Wright [17]; IC: Interaural coherence; ILD: Interaural level difference; ITD: Interaural time difference; LB: Langendijk and Bronkhorst [20]; LFCC: Linear-frequency cepstral coefficient; LRM: Left, right, and mid channels; LRS: Left, right, and side channels; LR: Left and right channels; Logit: Logistic regression; MFCC: Mel-frequency cepstral coefficient; ReLU: Rectified linear unit; RMS: Root-mean-square; SGD: Stochastic gradient descent; SVM: Support vector machines; SVM-RBF: Support vector machines with radial basis function; SVM-Lin: Support vector machines with linear function; XGBoost: Extreme gradient boosting

Acknowledgements

Not applicable

Authors' contributions

SKZ selected HRTFs, generated the binaural excerpts, undertook the feature engineering task, developed and tested the traditional discrimination algorithms. PA implemented, optimized, and tested CNN. In addition, he was responsible for the CNN visualization. SKZ, HL, DJ, and PA jointly developed the research methodology. HL and DJ helped in interpretation of the results. All authors contributed in writing the manuscript and further read and approved the final manuscript.

Funding

The work was supported by the grant from Białystok University of Technology (WZ/WI-IIT/4/2020) and funded with resources for research by the Ministry of Science and Higher Education in Poland.

Availability of data and materials

The binaural music excerpts, generated and exploited in this study, are not publicly available due copy right restrictions but are available from the authors upon reasonable request. The trained CNN model along with the source code of the final models used in the study has been made publicly available at GitHub [54].

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Computer Science, Białystok University of Technology, 15-351 Białystok, Poland. [2]Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield HD1 3DH, UK.

Received: 16 June 2021 Accepted: 23 December 2021
Published online: 15 January 2022

References

1. F. Rumsey, Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. J. Audio Eng. Soc. **50**(9), 651–666 (2002)
2. J. Blauert, *Spatial hearing. The psychology of human sound localization* (MIT Press, London, 1974), pp. 179–180
3. N. Ma, T. May, G.J. Brown, Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(12), 2444–2453 (2017). https://doi.org/10.1109/TASLP.2017.2750760
4. T. May, N. Ma, G.J. Brown, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues IEEE, Brisbane, 2015, pp. 2679–2683
5. T. Usagawa, A. Saho, K. Imamura, Y. Chisaki, in *2011 IEEE Region 10 Conference TENCON*. A solution of front-back confusion within binaural processing by an estimation method of sound source direction on sagittal coordinate (Bali, Indonesia, 2011), pp. 1–4. https://doi.org/10.1109/TENCON.2011.6129051
6. S.K. Zieliński, F. Rumsey, R. Kassier, S. Bech, Development and initial validation of a multichannel audio quality expert system. J. Audio Eng. Soc. **53**(1/2), 4–21 (2002)
7. S.K. Zieliński, H. Lee, Automatic spatial audio scene classification in binaural recordings of music. Appl. Sci. **9**(1724) (2019). https://doi.org/10.3390/app9091724
8. S.K. Zieliński, in *Proc. of the International Conference on Computer Information Systems and Industrial Management*. Improving classification of basic spatial audio scenes in binaural recordings of music by deep learning approach Białystok, Poland, 2020, pp. 291–303. https://doi.org/10.1007/978-3-030-47679-3_25
9. S.K. Zieliński, H. Lee, P. Antoniuk, O. Dadan, A Comparison of Human against Machine-Classification of Spatial Audio Scenes in Binaural Recordings of Music. MDPI. Appl. Sci. **10**(5956) (2020). https://doi.org/10.3390/app10175956
10. M. Ancona, E. Ceolini, C. Öztireli, M. Gross, in *Proc. of the Sixth International Conference on Learning Representations (ICLR)*. Towards better understanding of gradient-based attribution methods for deep neural networks (ICLR, Vancouver, 2018)
11. T. May, S. van de Par, A. Kohlrausch, in *Binaural localization and detection of speakers in complex acoustic scenes, in the technology of binaural listening, modern acoustics and signal processing*, ed. by J. Blauert. (Springer, London, 2013), pp. 397–425
12. J. Nowak, Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations. J. Acoust. Soc. Am. **142**(1634) (2017). https://doi.org/10.1121/1.5003917
13. C.I. Cheng, G.H. Wakefield, Introduction to head-related transfer functions (HRTFs): representations of HRTFs in time, frequency, and space. J. Audio Eng. Soc. **49**(4), 231–249 (2001)
14. F.L. Wightman, D.J. Kistler, Resolution of front–back ambiguity in spatial hearing by listener and source movement. J. Acoust. Soc. Am. **105**(2841) (1999). https://doi.org/10.1121/1.426899
15. A. Kulkarni, H.S. Colburn, Role of spectral detail in sound-source localization. Nature **397**, 747–749 (1998)
16. J. Blauert, Sound localization in the median plane. Acustica **22**, 205-213 (1969/70)
17. J. Hebrank, D. Wright, Spectral cues used in the localization of sound sources on the median plane. J. Acoust. Soc. Am. **56**(1829) (1974). https://doi.org/10.1121/1.1903520

18. F. Asano, Y. Suzuki, T. Sone, Role of spectral cues in median plane localization. J. Acoust. Soc. Am. **88**(159) (1990). https://doi.org/10.1121/1.399963
19. M. Morimoto, M. Yairi, K. Iida, M. It, The role of low frequency components in median plane localization. Acoust. Sci. & Tech. **24**(2), 76–82 (2003). https://doi.org/10.1250/ast.24.76
20. E.H.A. Langendijk, A.W. Bronkhorst, Contribution of spectral cues to human sound localization. J. Acoust. Soc. Am. **112**(1583) (2002). https://doi.org/10.1121/1.1501901
21. P.X. Zhang, W.M. Hartmann, On the ability of human listeners to distinguish between front and back. Hear Res. **260**(1-2), 30–46 (2010). https://doi.org/10.1016/j.heares.2009.11.001
22. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raumel, S. Argentieri, Binaural localization of multiple sound sources by non-negative tensor factorization. IEEE/ACMTrans. Audio Speech Lang. Process. **26**(6), 1072–1082 (2018). https://doi.org/10.1109/TASLP.2018.2806745
23. N. Ma, G.J. Brown, *in Proc of the INTERSPEECH, Speech localisation in a multitalker mixture by humans and machines* ((San Francisco, CA, USA, 2016), pp. 3359–3363 10.21437/Interspeech.2016–1149
24. N. Ma, J.A. Gonzalez, G.J. Brown, Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(11), 2122–2131 (2018). https://doi.org/10.1109/TASLP.2018.2855960
25. T. May, S. van de Par, A. Kohlrausch, A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. IEEE/ACM Trans. Audio, Speech, Language Process. **20**(7), 2016–2030 (2012). https://doi.org/10.1109/TASL.2012.2193391
26. M. Dietz, S.D. Ewert, V. Hohmann, Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Commun. **53**(5), 592–605 (2011). https://doi.org/10.1016/j.specom.2010.05.006
27. P. Vecchiotti, N. Ma, S. Squartini, G.J. Brown, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end binaural sound localisation from the raw waveform (Brighton, UK, 2019), pp. 451–455. https://doi.org/10.1109/ICASSP.2019.8683732
28. Y. Han, J. Park, K. Lee, in *Proc. of the Conference on Detection and Classification of Acoustic Scenes and Events*. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification (Munich, Germany, 2017), pp. 1–5
29. J. Wang, J. Wang, K. Qian, X. Xie, J. Kuang, Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. EURASIP J. Audio, Speech Music Process. **4** (**2020**). https://doi.org/10.1186/s13636-020-0171-y
30. A. Raake, A computational framework for modelling active exploratory listening that assigns meaning to auditory scenes—reading the world with two ears (2016), http://twoears.eu. Accessed 5 June 2021.
31. V. Pulkki, H. Pöntynen, O. Santala, Spatial perception of sound source distribution in the median plane. J. Audio Eng. Soc. **67**(11), 855–870 (2019). https://doi.org/10.17743/jaes.2019.0033
32. H.S. Braren, J. Fels, in *A high-resolution individual 3D adult head and torso model for HRTF simulation and validation*. 3D data. Technical Report. Institute of Technical Acoustics (RWTH Aachen University, 2020). https://doi.org/10.18154/RWTH-2020-06761
33. HRTF-Database. Acoustic Research Institute. Austrian Academy of Sciences (2014), https://www.oeaw.ac.at/en/isf/das-institut/software/hrtf-database. Accessed 15 June 2021.
34. V.R. Algazi, R.O. Duda, D.M. Thompson, C. Avendano, in *Prof. of the IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*. The CIPIC HRTF Database (IEEE, Mohonk Mountain House, New Paltz, NY, USA, 2001). https://doi.org/10.1109/ASPAA.2001.969552
35. A. Andreopoulou, D.R. Begault, B.F.G. Katz, Inter-Laboratory Round Robin HRTF Measurement Comparison. IEEE J. Sel. Topics Sig. Process. **9**(5), 895–906 (2015). https://doi.org/10.1109/JSTSP.2015.2400417
36. F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, S. Weinzier, A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses. J. Audio Eng. Soc. **67**(9), 705–718 (2019). https://doi.org/10.17743/jaes.2019.0024
37. LISTEN HRTF Database (2003), http://recherche.ircam.fr/equipes/salles/listen. Accessed 15 June 2021.
38. B. Gardner, K. Martin, HRTF measurements of a KEMAR dummy-head microphone. MIT Media Lab (1994), https://sound.media.mit.edu/resources/KEMAR.html. Accessed 15 June 2021.
39. K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, S. Sato, Dataset of head-related transfer functions measured with a circular loudspeaker array. Acoust. Sci. Tech. **35**(3), 159–165 (2014). https://doi.org/10.1250/ast.35.159
40. C. Armstrong, L. Thresh, D. Murphy, G. Kearney, A perceptual evaluation of individual and non-individual HRTFs: a case study of the SADIE II database. Appl. Sci. **8**(2029) (2018). https://doi.org/10.3390/app8112029
41. G. Yu, R. Wu, Y. Liu, B. Xie, Near-field head-related transfer-function measurement and database of human subjects. J. Acoust. Soc. Am. **143**(3), EL194 (2018). https://doi.org/10.1121/1.5027019
42. C. Pörschmann, J.M. Arend, A. Neidhardt, in *Proc. of the 142nd AES Convention. A spherical near-field HRTF set for auralization and psychoacoustic research*. A spherical near-field HRTF set for auralization and psychoacoustic research (AES, Berlin, Germany, 2017) e-Brief 322
43. F. Brinkmann, A. Lindau, S.S. van de Par, M. Müller-Trapet, R. Opdam, M. Vorländer, A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations. J. Audio Eng. Soc. **65**(10), 841–848 (2017). https://doi.org/10.17743/jaes.2017.0033
44. H. Wierstorf, M. Geier, A. Raake, S. Spors, in *Proc. of the 130th AES Convention*. A free database of head-related impulse response measurements in the horizontal plane with multiple distances (AES, London, UK) e-Brief 6
45. S. Spagnol, K.B. Purkhus, S.K. Björnsson, R. Unnthórsson, in *Proc. of the 16th Sound & Music Computing Conference (SMC 2019)*. The Viking HRTF dataset (Malaga, Spain, 2019)
46. S. Spagnol, R. Miccini, R. Unnthórsson, *The Viking HRTF dataset v2* (2020), https://zenodo.org. Accessed 15 June 2021). https://doi.org/10.5281/zenodo.4160401
47. R.H.Y. So, B. Ngan, A. Horner, J. Braasch, J. Blauert, K.L. Leung, Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: cluster analysis and an experimental study. Ergonomics **53**(6), 767–781 (2010). https://doi.org/10.1080/00140131003675117
48. T. Kim, J. Lee, J. Nam, Comparison and analysis of SampleCNN architectures for audio classification. IEEE J. Sel. Topics Signal Process. **13**(2), 285–297 (2019). https://doi.org/10.1109/JSTSP.2019.2909479
49. J. Abeßer, A review of deep learning based methods for acoustic scene classification. Appl. Sci. **10**(6) (2020). https://doi.org/10.3390/app10062020
50. A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017). https://doi.org/10.1145/3065386
51. G. James, D. Witten, T. Hastie, R. Tibshirani, *An introduction to statistical learning with applications in R* (Springer, London, UK, 2017), pp. 148–149 219, 280
52. D. Barchiesi, D. Giannoulis, D. Stowell, M.D. Plumbley, Acoustic scene classification: classifying environments from the sounds they produce. IEEE Signal. Process. Mag. **32**(3), 16–34 (2015). https://doi.org/10.1109/MSP.2014.2326181
53. Q.-T. Phan, Y.-K. Wu, Q.-D. Phan, in *Proc. of the IEEE International Symposium on Computer, Consumer and Control (IS3C)*. A comparative analysis of XGBoost and temporal convolutional network models for wind power forecasting (IEEE, Taichung City, 2020), pp. 416–419. https://doi.org/10.1109/IS3C50286.2020.00113
54. P. Antoniuk, Automatic discrimination between front and back ensemble locations in HRTF-convolved binaural recordings of music. Software Repository (2021). https://github.com/pawel-antoniuk/appendix-paper-front-back-eurasip-2021. Accessed 2 Oct 2021.
55. G. Peeters, B. Giordano, P. Susini, N. Misdariis, S. McAdams, Extracting audio descriptors from musical signals. J. Acoust. Soc. Am. **130**(2902), 2902–2916 (2011). https://doi.org/10.1121/1.3642604
56. F. Pedragoza et al., Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
57. X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, in *Proc. of the IEEE Workshop on Automatic Speech Recognition & Understanding*. Linear versus mel frequency cepstral coefficients for speaker recognition (IEEE, Waikoloa, 2011), pp. 559–564. https://doi.org/10.1109/ASRU.2011.6163888
58. A.K.H. Al-Ali, D. Dean, B. Senadji, V. Chandran, G.R. Naik, Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. IEEE Access **5**, 15400–15413 (2017). https://doi.org/10.1109/ACCESS.2017.2728801
59. T. Dau, D. Püschel, A. Kohlrausch, A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. J. Acoust. Soc. Am. **99**(6), 3615–3622 (1996). https://doi.org/10.1121/1.414959

60. G.J. Brown, M. Cooke, Computational auditory scene analysis. Comput. Speech Lang. **8**(4), 297–336 (1994). https://doi.org/10.1006/csla.1994.1016
61. T. Chen, C. Guestrin, in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. XGBoost: a scalable tree boosting system (ACM, 2016). https://doi.org/10.1145/2939672.2939785
62. Y. Wu, T. Lee, *In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Enhancing sound texture in CNN-based acoustic scene classification* ((IEEE, Brighton, UK, 2019), pp. 815–819. https://doi.org/10.1109/ICASSP.2019.8683490
63. A. Rakotomamonjy, Supervised representation learning for audio scene classification. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(6), 1253–1265 (2017). https://doi.org/10.1109/TASLP.2017.2690561
64. R. Geirhos, J.H. Jacobsen, C. Michaelis, et al., Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**, 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z
65. M. Brookes, VOICEBOX: speech processing toolbox for MATLAB, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. Accessed 11 June 2020.
66. T. Chen, T. He, M. Benesty, et al., Extreme gradient boosting. CRAN Repository https://github.com/dmlc/xgboost. Accessed 7 Oct 2021
67. M. Chen, Q. Liu, S. Chen, Y. Liu, C. Zhang, XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access **7**, 13149–13158 (2019). https://doi.org/10.1109/ACCESS.2019.2893448
68. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. Grad-CAM: visual explanations from deep networks via gradient-based localization (IEEE, Venice, 2017), pp. 618–626. https://doi.org/10.1109/ICCV.2017.74
69. S. Irving, D.R. Moore, Training sound localization in normal hearing listeners with and without a unilateral ear plug. Hear. Res. **280**(1-2), 100–108 (2011). https://doi.org/10.1016/j.heares.2011.04.020

Publisher's Note