DOI: 10.61782/fa.2023.1179

forum acusticum 2023

IDENTIFICATION OF BEST-MATCHING HRTFs FROM BINAURAL SELFIES AND MACHINE LEARNING

Olivier Warusfel
Sciences et Techniques de la Musique et du Son
IRCAM, CNRS, Sorbonne Université, Ministère de la Culture
1, Place Igor Stravinsky, 75004 Paris, France

ABSTRACT

Augmented reality applications consist in embedding synthetic sound events into the real world of the listener. The accuracy of the spatial processing applied to the virtual sound objects is essential for the overall quality of experience. It requires means for automatically identifying the acoustic properties of the environment, including the head-related transfer functions (HRTFs) of the listener. The long-term aim of the study is automatic selection of the best matching HRTFs set within a database, given binaural selfies, i.e. signals recorded in arbitrary environments by the listener equipped with in-ear microphones. The approach builds upon prior machine learning methods capable of end-to-end estimation of the direction of incidence of a sound source from binaural signals. Extracted features are then exploited by an additional model to estimate the best matching set of HRTFs among available databases. The mobility of the listener during the recording is an asset to accumulate knowledge about these features, enhancing the confidence when estimating the HRTFs set likelihood. As a proof of concept, the method is first performed with synthesized as well as real binaural selfies of listeners whose HRTFs belong to the database to verify that they are actually elected as best-matching.

Keywords: *HRTF individualisation, binaural selfie, machine learning.*

**Corresponding author*: *warusfel@ircam.fr.*

1. INTRODUCTION

Audition is a key modality to interact with our spatial environment and plays a major role in Augmented Reality (AR) applications. Embedding computer-generated or pre-recorded auditory content into a user's real acoustic environment creates an engaging and interactive experience that can be applied to video games, museum guides or radio drama. Such interactive augmented 3D audio scenes are typically rendered through binaural signals and played back over headphones. It is well known that convincing binaural sound reproduction requires individual Head Related Transfer Functions (HRTFs). Ideally, individual HRTFs should be measured in an anechoic chamber using highly calibrated signals and apparatus that allow to replicate the measurement along a dense spatial grid of sampling points [1]. Obviously, such a complex procedure is not accessible to the general public.

Various studies have been dedicated to the design of alternative HRTFs individualisation methods, less costly in time and compatible with real-world conditions. Several studies have shown that HRTFs can be computed from 3D head scans with sufficient accuracy [2] applying numerical simulation such as Finite Element Method, Boundary Element Method, or Fast Multipole Boundary Element Method [3]. Multiple Linear Regression (MLP) and Neural Networks [4–6] have been applied to learn high-level relationships between measured HRTFs and morphological parameters. Other methods propose to guide listeners into a subjective comparison of different HRTFs sets through perceptual tests. This becomes rapidly tedious when the number of HRTF sets to be compared increases. Generative neural networks have been proposed to spatially up-sample HRTFs sets from sparse

measurements [7, 8].

Machine learning (ML) has been extensively used in binaural listening for source separation and source localisation [9, 10]. In [11] a machine-hearing system is proposed to reduce front-back confusions by combining head movements with deep neural networks (DNN) that learn the relationship between azimuth and binaural cues. In [12] an end-to-end method is introduced for sound source localisation from the raw binaural waveform. In [13], the performance of DNN based source localisation methods is shown to decline in mismatched HRTFs condition, i.e. when the HRTF set used for testing differs from the HRTF set used during the training.

The ultimate aim of the proposed approach is the automatic selection of the best matching HRTFs set among a database, applying ML techniques to binaural selfies, i.e. signals that are captured in real and unsupervised acoustic environments (i.e. not knowing the source signals nor the source positions) by a listener equipped with nowadays available in-ear microphones or hear-through earphones. In [14] a method was described to select a best matching HRTF set on the basis of such binaural recordings. Its principle was to extract interaural cues from each time frame based on the equalisation-cancellation (EC) auditory model [15]. These interaural cues were compared to that of different HRTFs in a given database in order to elect the best matching direction, whichever subject it would belong to. Repeating this operation for successive time frames of the binaural recording was expecting to reveal the best matching HRTF set, i.e. the subject of the database which HRTFs were most often selected during the binaural selfie analysis. In the present study, the approach builds upon a similar principle but differs in two ways. First, the underlying frame by frame source localisation task is driven through a ML approach inspired by [12] instead of an explicit interaural and monaural cues extraction. Second, the mobility of the listener during the binaural selfie recording is exploited to gain confidence on the estimated best matching HRTFs set. Indeed, the spatial consistency of the estimated trajectory is exploited to evaluate the HRTFs set likelihood. Indeed, in real-world conditions, the listener and source movements are generally continuous. Hence, it is expected that the estimated source-listener relative trajectory will present smooth spatial behavior if based from a matching HRTFs set. In contrast, trajectory estimation from mismatched HRTFs sets should present erratic movements such as front-back confusions or elevation instabilities.

The paper is organised as follows. The overall principle and ML architecture are presented in section 2. Section 3 presents the virtual data sets used to train the models. As a proof of concept, the method is then applied on binaural selfies corresponding to listeners whose HRTFs belong to the database in order to verify that their HRTFs would actually be elected as best matching. These tests are first conducted with synthesized binaural selfies (section 4) in order to study the influence of different ML architectures and parameters. Finally, a preliminary test is conducted with real-world binaural selfies (section 5).

2. SYSTEM DESCRIPTION

The proposed best matching HRTF set selection is based on a two-steps process, illustrated in Figure 1. The binaural selfie recorded by the listener s_i is segmented into successive two-channels time frames p_j each associated to an instantaneous direction of incidence d_j.

A first step performs a trajectory estimation by selecting the best direction matching within a given subject HRTFs set for each time frame of the binaural selfie. It uses a Convolutional Neural Network (CNN) to extract suitable features for sound localisation, combined with a Deep Neural Network (DNN) made of dense layers that performs localisation in polar coordinates. A second step uses a DNN to estimate the likelihood of the resulting trajectory, and can be interpreted as the likelihood that the binaural selfie has been recorded on a the corresponding subject's head, or at least a "compatible" one. Training is performed starting with the first step, then freezing the model to train the second step.

2.1 Trajectory estimation

Input of the first step is the raw binaural signal with left (L) and right (R) channels sampled at 44.1kHz and divided into 20ms frames with no overlap. Resulting input frames are matrices of size 2x882. The architecture of this step is a reproduction of the WaveLoc-CONV system from the work of [12], with only a few changes. This system was shown to provide good azimuth direction estimation, including in reverberant conditions, a matter of critical importance in real-world conditions.

The CNN has three stages. First stage is composed of 64 kernels of size 1x256. Vecchiotti & al. show that it can be closely related to a frequency band filtering. The second stage performs an interaural cross-correlation with 64 kernels of size 2x18. The third stage uses 64 kernels of size 1x6 to generate latent features later used for local-

isation task. All stages are followed by 1x2 Max Pooling and rectified linear unit activations (*reLU*).

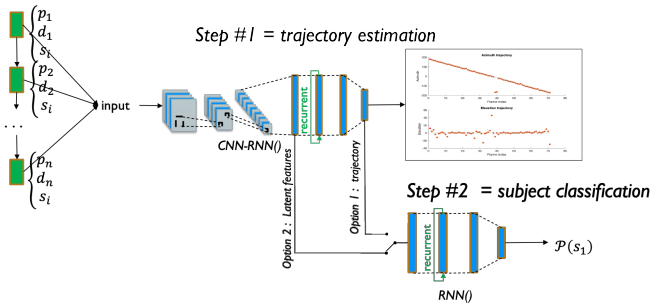

Figure 1. Architecture overview

The DNN solves a multinomial classification problem. It aims at properly recovering the instantaneous source direction of incidence from a set of possible directions in the HRTF spatial grid. To address this task, it uses two consecutive dense layers of 1024 cells each, then a final dense layer makes a projection onto the desired output space. *Softmax* function is applied to normalize the output to a probability distribution, thus generating a probability vector. In contrast with [12], the task is not limited to the horizontal plane, but generalised to the whole sphere, performing multitask learning as in [16]. Two versions of the network are run in parallel, one estimating azimuth and the other estimating elevation. In a variant of the architecture, the DNN is replaced by a Recurrent Neural Network (RNN) using bidirectional Long Short-Term Memory (biLSTM) layers of 512 cells. It allows taking advantage of adjacent information for all time frames of a given audio clip, and conforms with the assumption of continuous movements of the sound source or rotation of the listener head.

2.2 Subject likelihood

The subject likelihood system is a DNN addressing a binary classification task. Its configuration depends on the features used to feed the network. A first option uses the successive direction probability vectors provided by step 1 as inputs. The model integrates vectors from all time frames in the signal and uses two biLSTM layers of 16 cells to estimate the trajectory likelihood. Following the multitask learning paradigm, this network is duplicated and ran in parallel for both azimuth and elevation probability vectors. Resulting vectors are concatenated and fed into a final dense layer producing a single output normalized to a probability with *sigmoid* activation function. A second option uses latent parameters extracted from the last stage of the first-step CNN network. The model integrates latent features from all time frames and uses two biLSTM layers of 512 cells each. This configuration aims at extending the use of convolutional parameters generated from localisation task to a subject classification task.

3. VIRTUAL TRAINING SETUP

3.1 Data

A recurrent difficulty when applying ML approach to sound spatialisation problems is to get access to large enough and well structured 3D sound databases. To overcome this problem it is often proposed to train the model through virtual sound scenes that span a larger variety of situations and to verify that it generalises to real situations. The training and testing datasets are synthesized using single sound source signals. Speech signals are first chosen since they are common sound sources. Second, the motivation is to assert the system's resilience to sounds which are non-stationnary and sparse in frequency, as such properties may challenge the trajectory estimation. Samples forming the datasets are created by spatialising 44.1kHz speech audio clips taken from the HiFi database [17]. A set of 10 000 randomly selected speech clips from 9 speakers (5 female, 4 male) is used to construct the training datasets, for a total of 11 hours of speech. 10 000 files from speakers unseen during training, are used to construct the unique validation dataset, for a total of 9 hours.

Performing on-the-fly spatialisation is too much time-consuming, hence training datasets are generated in-between training iterations. Each audio file is spatialised with a randomly generated trajectory so as to simulate head or source movements. The trajectory is specified with a spatial resolution 6° in both dimensions (azimuth and elevation) and a time resolution of 100ms. During the rendering, the actual trajectory is interpolated with a 30ms time granularity. The azimuth and elevation distributions over the entire dataset are almost uniform besides slightly decreasing towards elevation extrema. Azimuth directions range from $-180°$ to $+174°$. Elevation directions range from $-30°$ to $30°$. Trajectory speed varies from 2 to 20°/s azimuth-wise, and from 4 to 6°/s elevation-wise. Although being concatenated into large audio files for practical reasons, the segmentation information between its constitutive audio clips is kept. Hence, each forward inference is performed on a single audio clip, lasting

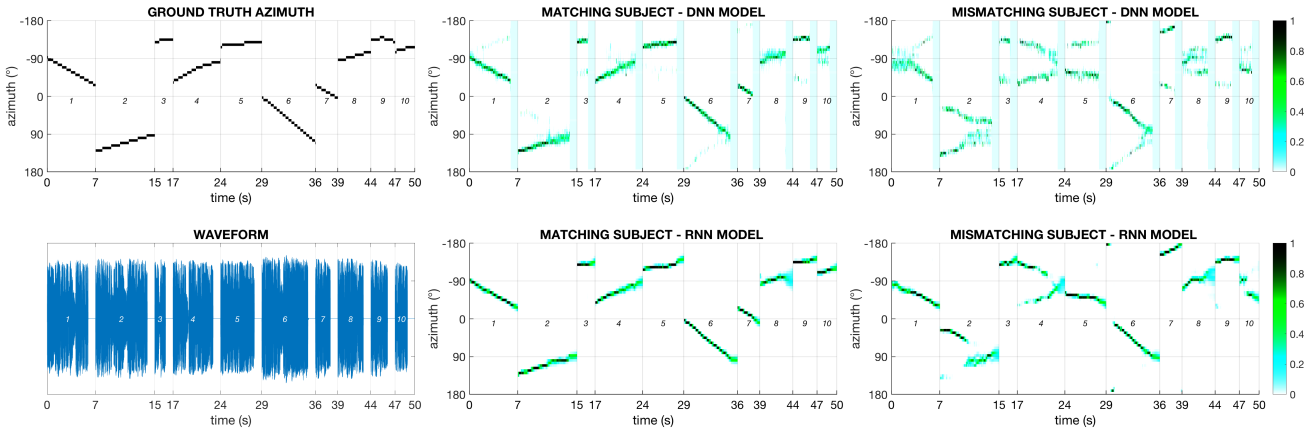

Figure 2. Azimuth probability maps (APMs). Left: waveform of 10 audio clips and associated azimuth trajectories. Middle: APMs for the matching subject using DNN (top) or RNN (bottom) models. Right: APMs for a mismatching subject.

on average a few seconds (typically 2 to 10s). Its average azimuth coverage is $42°$, and average elevation coverage is $18°$. Between each training iteration, a new trajectory is generated and embedded as metadata into the audio files, using the object oriented audio description model (ADM). The rendering of the audio files uses the `adm.renderer` module of `spat5` library [18]. The direct sound is rendered with Head-Related Impulse Response (HRIR) sets taken from BiLi database [1] that satisfies SOFA conventions [19]. This database contains 56 sets of HRIRs for 54 human subjects and 2 dummy heads measured in anechoic conditions. HRIRs were measured originally with a $6°$ resolution in azimuth and elevation and have been interpolated in the spherical harmonics domain to reach a spatial resolution of $2°$. Reverberant conditions consist in adding a room effect to the direct sound. As a first step, and for sake of genericity, the room effect is approximated by a diffuse reverberation tail synthetised using a feedback delay network (FDN) module from the `spat5` library allowing for flexible control of the reverberation time (RT). The reverberation tail and direct sound are mixed together with control of the short pre-delay ($[10 - 20]ms$) and direct sound over reverberation level ratio (D/R).

3.2 Training process

Training of the trajectory estimation step and subject likelihood step are performed separately. Training processes use gradient descent with *Adam* optimizer and initial learning rate of 10^{-3} to train the parameters of the networks. An iteration consists in generating a new training dataset and runing a forward-backward propagation over all files. Each inference feeds the model with batches of 24 files split in 20ms time frames before back-propagating. Validating process occurs in-between iterations and applies loss function and binary accuracy metric over all the validation dataset. Training process stops when the validation error value has not improved for 10 consecutive iterations. Models used for the experiments are the ones generating minimal validation error.

In the trajectory estimation process, training and validation datasets are generated with the HRIR set of a given conditioning subject from the BiLi database. *Categorical-crossentropy* loss function is used to generate the direction error. In the subject likelihood process, training and validation datasets are generated half-time with the conditioning subject and the other half with a randomly selected HRIR set from the rest of the database. *Binary-crossentropy* loss function is used to teach the model to discriminate between matching and mismatching conditions, i.e. whether or not the subject corresponds to the conditioning one.

3.3 Testing conditions

Different testing conditions are established in order to evaluate three main parameters' influence. Trajectory estimation is evaluated by displaying Azimuth and Elevation

Probability Maps (APMs and EPMs, respectively) of two different models: the DNN model, that operates on each time frame individually, and the RNN model that enforces the continuity across time frames of a given audio clip.

Subject likelihood is evaluated according to its performance under reverberant conditions. In addition to anechoic condition, five rooms are simulated, with two different RT: 500ms (rooms A and B) or 1000ms (rooms C, D and E) and three D/R ratios: +15dB (rooms A and C), +10dB (rooms B and D) or +12.5dB (room E). In the anechoic-training setup (ANE), only the anechoic signals are used to train the models. In the multiconditional-training setup (MCT), Rooms A, B, C and D are used during training, together with anechoic condition. Room E is used for testing both ANE and MCT setups.

Finally, for all setups, the two architecture options described in 2.2 are evaluated, i.e. either exploiting the direction probability maps (in azimuth and elevation) or the latent parameters generated by the CNN stage.

4. EVALUATION OF THE MODELS

4.1 Step 1: trajectory estimation analysis

Figure 2 shows estimated APMs for 10 independent audio clips separated with vertical lines. True azimuth trajectory is represented at the top-left, with associated waveform below. Middle column displays generated APMs for each clip under matching subject condition. Right column presents an example of a mismatching subject condition. Middle and right columns compare probability maps from DNN (top) and RNN (bottom) models. In the matching condition, both estimations are closely correlated to the ground truth. RNN model infers trajectory continuity during ending silences, while DNN model presents uniformly low probability distribution. In the mismatching condition, the DNN model displays almost systematic front-back confusions revealed by mirroring patterns around the interaural axis ($\pm 90°$). The RNN resolves most of the front-back instabilities except for audio clips 2, 4 and 6. At first glance, APMs generated by the DNN model ought to be given more credit for this matching vs non-matching subject discrimination, yet subject likelihood analysis does not confirm the hypothesis.

4.2 Step 2: subject likelihood analysis

Figure 3 displays average binary decision, i-e proportion of audio clips labeled by the estimator as "subject matching" (likelihood greater than 0.5), on a set of 20 subjects from BiLi database. All models were conditioned on subject 1100, and tested on all the validation data with matching condition (subject 1100) and mismatching conditions (subjects 1101 to 1119). First value from the left can be read as the proportion of true matching classifications, whereas others can be interpreted as the proportion of false matching classifications for different subjects.

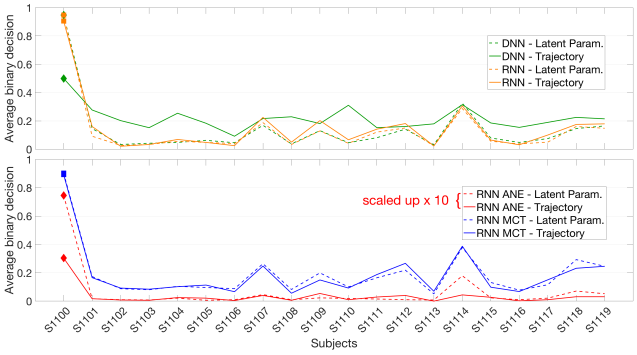

Figure 3. Average binary decision for 20 BiLi subjects. Top: RNN vs DNN models trained with subject 1100. Bottom: MCT vs ANE training setups.

Upper graphic compares RNN and DNN models in anechoic conditions, when fed with either latent parameters or trajectories (direction probability maps). All performances are similar, except for the DNN model fed by trajectories. Training for this specific setup appeared unstable and led to collapsing performances.

Lower graphic displays average binary decisions provided by RNN model trained with MCT and ANE setups when tested on a reverberant condition (here room E). Values for ANE setup were scaled up by a factor of 10 in order to ease the comparison. MCT setup proves to be generalising much better in reverberant conditions than ANE, confirming observations of Vecchiotti & al. [12], since it produces 90% of true matching classifications versus less than 10% for ANE setups. However, although the MCT setup provides better average scores, the ANE setup still preserves the structure of the results, keeping a significant contrast between matching and mismatching subjects.

Anticipating the analysis of the next section dedicated to real-world recordings, figure 4 shows the evaluation of theoretical subject likelihood estimation for two models of the HRTFs database trained with heads 1130 (dummy head) and 1154 (human), respectively. Both models were trained with the RNN based architecture for step 1 and used the predicted azimuth and elevation probability vec-

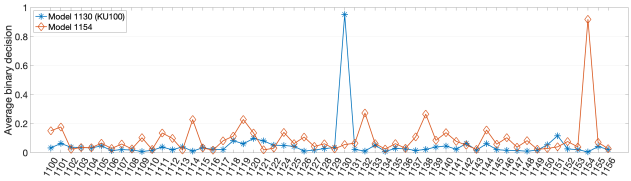

Figure 4. theoretical binary decision scores of the two models 1130 (dummy head) and 1154 (human) used for the real-world recordings

tors as input for the second step. They were trained through five virtual environments (one anechoic and the 4 virtual rooms A, B, C and D). Both models are shown to provide consistent estimation of their respective best matching subject (resp. 1130 and 1154), and to reject all other subjects of the database. The binary accuracy performance of the dummy head model 1130 shows higher contrast with a lower average decision score (mean 0.03 +/- 0.02) for the non-matching subjects compared to the rejection of non-matching subjects by the human head model 1154 (mean 0.08 +/- 0.06). This behavior may reveal the specificity of the dummy head compared to the human heads of the database. In contrast, it is noticeable for the model trained with human head 1154 that some mismatching subjects stand out consistently from the rest of the database (especially subjects 1114, 1119, 1132, 1138). Further studies are needed to determine if this behavior is linked to an actual similarity between these HRTFs sets and that of subject 1154. Such an investigation could be based for instance on the probabilistic auditory model for sound localisation described in [20].

5. REAL-WORLD RECORDINGS

5.1 Experimental conditions

This section presents a preliminary evaluation of the method with binaural selfies recorded in real conditions on human subject 1154 and dummy head 1130. Both belong to the HRTF database. The theoretical performance of their respective model trained in virtual conditions have been described in the previous section (cf. figure 4). For the real recordings they were equipped with a binaural microphone headset DPA4560. Although light, this headset differ from the in-ear microphones that were used to measure the HRTFs database. In particular, they do not fulfill the blocked ear condition recommended for HRTF measurements. The experiment was thus repeated with the dummy head either using its native internal microphones or equipped with this microphone headset to reveal its possible influence.

The sound stimuli were speech signals extracted from the same HiFi speech database. The binaural selfies were recorded under various acoustic conditions. The speech signals were played successively through eight loudspeakers positioned at different heights and distances around the dummy-head/listener to cover different elevation angles of incidence as well as different direct over reverberant (D/R) ratios. The experiment was conducted in the variable acoustic hall of IRCAM, a parallelepipedic room ($24\text{x}14\text{x}11m^3$) which walls and ceiling consist of three sided prisms that can be oriented to present their absorbing, reflecting or diffusing side, thus offering a large range of reverberation times (RT). The whole experiment was repeated under three acoustic configurations (referred to as Room1, Room2 and Room3) which RT, measured at 1kHz, were 1.2s, 1.7s and 2.7s, respectively. The resulting D/R ratio measured at the listener position for each loudspeaker is reported in table 5.1, together with the distance and the elevation angle relative to the subject's head. According to the loudspeaker and room configuration, the D/R ratio is seen to range between 3dB to 17dB. Both the experimental ranges of the RT and of the D/R ratio significantly exceed that of the training virtual situations. The maximum trained RT was 1.0s, whereas Room3 reaches 2.7s. The lowest trained D/R ratio was 10dB, whereas several experimental values corresponding to distant loudspeakers are significantly lower, especially in Room3 condition. To provide better and reproducible control of the relative source to listener movements, the subject (or dummy head) was seating on a turntable (see Figure 5). For each loudspeaker, the same rotation profile was applied to the turntable covering a back and forth excursion of circa 310° with a rotation speed ranging from 2°/s to 16°/s. The overall rotation profile duration, was about 2m30s for each loudspeaker. The position and orientation of the subject was tracked and recorded simultaneously with the binaural selfie to allow for comparison between the estimated and actual relative source to listener direction.

5.2 Observations in real conditions

Figure 6 presents the average subject likelihood provided by the two models when fed with the binaural selfies. Several observations can be made. Each model is able to discriminate the binaural selfies corresponding to the head it was conditioned with (matching subject). However, the

Figure 5. Schematic view of the recording situation in the variable acoustic hall at IRCAM

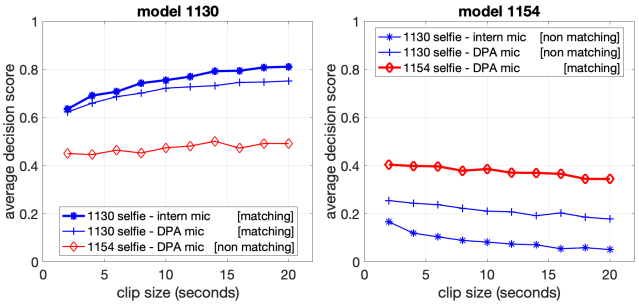

Figure 6. Evolution of decision score with clip size

Table 1. Acoustical and geometrical conditions

	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8
elevation(°)	26	0	-15	11	-5	-24	0	7
distance(m)	3.95	2.6	2.75	5.4	3.6	2.6	4.75	5.5
Room1 D/R (dB)	8.0	17.3	12.1	5.8	13.4	13.7	9.6	9.6
Room2 D/R (dB)	8.0	16.6	10.6	5.2	13.0	12.0	8.6	8.0
Room3 D/R (dB)	6.9	14.7	9.8	3.1	11.3	10.8	7.8	7.1

Table 2. Binary decision for each loudspeaker

Model 1130 (KU100)	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8
selfies 1130 intern mic	**0.81**	**0.31**	**0.79**	**0.70**	**0.85**	**0.81**	**0.62**	**0.78**
selfies 1130 DPA mic	**0.84**	**0.31**	**0.69**	**0.76**	**0.70**	**0.80**	**0.68**	**0.71**
selfies 1154 DPA mic	0.62	0.20	0.46	0.67	0.58	0.09	0.47	0.71
Model 1154 (human)	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8
selfies 1130 intern mic	0.04	0.06	0.17	0.07	0.17	0.11	0.14	0.07
selfies 1130 DPA mic	0.10	0.17	0.51	0.14	0.37	0.18	0.30	0.14
selfies 1154 DPA mic	**0.32**	**0.35**	**0.56**	**0.44**	**0.51**	**0.23**	**0.53**	**0.32**

contrast with non-matching selfies is much lower than in virtual conditions. Several factors may have contributed to this behavior. The real-world binaural selfies were not segmented according to its constitutive audio clips in contrast with the training situation. Instead, a regular segmentation was applied disregarding of sound events. This parameter is important since it drives the time horizon of the recurrent neural network (RNN) of step 1. The contrast is slightly increasing with this clip size. The binaural selfies captured by the dummy head with its internal microphones or with the binaural microphone headset were hardly discriminated by the model conditioned with the dummy-head, whereas they were consistently discriminated by the model conditioned with the human head. The influence of the room was shown marginal. In contrast, the influence of the loudspeaker was significant (table 5.2). From such a preliminary test, it is not possible to infer any conclusion about the possible influence of their height or associated D/R ratio. It is hypothesised that this is linked to the presence of first reflections, which were neglected in the simplified room effect simulations. Future work includes extensive tests to evaluate the performance among listeners not anymore belonging to the database and to investigate if the HRTFs set likelihood provided by the model is corroborated by perceptual metrics or tests.

6. CONCLUSION

This article proposes a two-steps ML architecture to drive the selection of the best-matching HRTFs set from binaural selfies recorded by a listener. The system uses prior ML method addressing source localisation, further extended by a classification task. Models trained and evaluated on synthesised sound scenes succeed in discriminating between matching or non-matching HRTFs sets. Preliminary tests conducted on real-world binaural recordings suggest that the discrimination is still possible. Further tests are needed to evaluate if this discrimination could generalise and to investigate the relationship between the subject likelihood generated by the model and actual HRTFs objective and perceptual similarity.

7. ACKNOWLEDGMENTS

This work was funded by the HAIKUS project ANR-19-CE23-0023. The author wishes to express his gratitude to Anatole Moreau and Raphaël Penot for the development of the initial versions of this work.

8. REFERENCES

[1] T. Carpentier, H. Bahu, M. Noisternig, and O. Warusfel, "Measurement of a head-related transfer function database with high spatial resolution," in *proc. of 7th Forum Acusticum (EAA)*, Sept. 2014.

[2] W. Kreuzer, P. Majdak, and Z. Chen, "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *The Journal of the Acoustical Society of America*, vol. 126, pp. 1280–1290, 09 2009.

[3] B. Katz, "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2440–2448, 2001.

[4] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, 2008.

[5] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3707–3710, 05 2013.

[6] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing hrtfs from anthropometric features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 559–570, 2016.

[7] N. H. Zandi, A. M. El-Mohandes, and R. Zheng, "Individualizing head-related transfer functions for binaural acoustic applications," in *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 105–117, 2022.

[8] P. Siripornpitak, I. Engel, I. Squires, S. J. Cooper, and L. Picinali, "Spatial up-sampling of hrtf sets using generative adversarial networks: A pilot study," *Frontiers in Signal Processing*, vol. 2, 2022.

[9] M. Lovedee-Turner and D. Murphy, "Application of Machine Learning for the Spatial Analysis of Binaural Room Impulse Responses," *Applied Sciences*, vol. 8, 2018.

[10] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds," *International Journal of Neural Systems*, vol. 25, no. 01, 2015.

[11] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

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

[13] J. Wang, J. Wang, K. Qian, X. Xie, and J. Kuang, "Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched hrtf condition," *EURASIP Journal on Audio, Speech, and Music Processing*, 2020.

[14] M. Maazaoui and O. Warusfel, "Estimation of individualized hrtf in unsupervised conditions," in *Proc. of 140th $AESConvention, May$*2016.

[15] N. I. Durlach, "End-to-end binaural sound localisation from the raw waveform," *Journal of the Acoustical Society of America*, vol. 35, pp. 1206–1218, 1963.

[16] C. Pang, H. Liu, and X. Li, "Multitask Learning of Time-Frequency CNN for Sound Source Localization," *IEEE Access*, vol. 7, pp. 40725–40737, 2019.

[17] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Z. Zhang, "Hi-Fi Multi-Speaker English TTS Dataset," *Interspeech 2021*, Septembre 2021.

[18] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty Years of Ircam Spat: Looking Back, Looking Forward," in *Proc. of 41st ICMC*, (Denton, TX, United States), pp. 270 – 277, Sept. 2015.

[19] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, and M. Noisternig, "spatially oriented format for acoustics: a data exchange format representing head-related transfer functions," in *Proc. of the Audio Engineering Society Convention 134*, (Roma, Italy), may 2013.

[20] R. Barumerli, P. Majdak, J. Reijniers, R. Baumgartner, M. Geronazzo, and F. Avanzini, "Predicting directional sound-localization of human listeners in both horizontal and vertical dimensions," in *Proc. of Audio Engineering Society Convention 148*, May 2020.