Towards all-purpose full-sphere binaural localization

1st Shoken Kaneko*
Department of Computer Science
University of Maryland, College Park
College Park, MD, USA
kaneko60@umd.edu

2nd Hannes Gamper
Audio and Acoustics Research Group
Microsoft Research
Redmond, WA, USA
hannes.gamper@microsoft.com

***Abstract*—Sound source localization from binaural signals has important applications ranging from machine listening to psychoacoustics, yet challenges including generalization and robustness to noise and reverberation remain. Here we propose a binaural localizer (BL) framework that produces a full-sphere spatial activity map for every audio input frame. The framework enables individual-agnostic training of a convolutional neural network using head-related impulse response (HRIR) sets with arbitrary measurement grids and is shown to perform well on unseen HRIRs and binaural recordings. Unlike BLs trained with the HRIRs of a specific known subject or dummy head, the proposed individual-agnostic BL is intended to perform robustly without any a priori knowledge about the process creating the binaural signals. Localization tests with binaural speech renderings and recordings show that the proposed BL performs well in the presence of noise and reverberation and compares favorably to individual-specific BLs. Furthermore, preliminary results indicate that the proposed BL is applicable to the localization of multiple simultaneous and moving sources.**

***Index Terms*—Binaural sound source localization, binaural hearing, head-related transfer functions, spatial audio**

I. INTRODUCTION

Humans localize sound by learning to map sound source locations to features embedded into the two ear input signals. A machine able to localize sound from binaural audio has important applications including robotics, sound scene analysis, as well as psychoacoustics and binaural audio evaluation. However, challenges remain for existing binaural localizers (BLs), including generalization across individuals and robustness under realistic noisy and reverberant conditions. While BLs can be classified into functional models [1] and machine learning-based models, here we focus on the latter class which date back to the early 90s [2]–[4]. Neti et al. created a neural network (NN) that learns the mapping from the HRIRs to the sound source direction using HRIRs of a cat [4]. Jin et al. reported reasonable agreement in terms of the localization characteristics between a human subject and a NN-based BL trained with the same subject's HRIR set using band noise sources [5]. Jiang et al. developed a deep NN (DNN) for binaural sound source separation of speech based on time-frequency bin-wise classification [6]. Ma et al. used a DNN for binaural localization of multiple sources in the horizontal plane, which also incorporated active head movements [7]. Thuillier et al. studied saliency maps of a convolutional NN (CNN)-based BL for median plane localization trained with an individual-agnostic setup [8]. Wu et al. developed a random forest-based BL trained on a single subject's HRIR set for localization of both azimuthal and elevation angle and reported mean angular errors of about 10 degrees on binaural signals recorded in a laboratory. Wang et al. studied DNN-based BLs in the mismatched HRIR condition where the HRIR set used at test time is different from the set used for training, and proposed a method for clustering HRIR sets based on the similarity of the localization performance of BLs [9]. Yang et al. developed a multi-task CNN-based BL for lateral and polar angle classification [10]. Francl et al. studied a massive CNN-based BL trained on dummy head HRIRs and reported various similarities between human spatial hearing and model behaviour, e.g., the emergence of the precedence effect, sensitivity to spatial cues, and the challenges arising in localization of concurrent sources [11]. Their results also indicated that an individual-specific BL may have limited generalization ability across individuals in elevation localization.

Most existing BLs are either individual-specific or limited to dataset-dependent directions on a subset of the sphere. While an individual-specific BL may be useful for applications targeting a specific user, training such a BL requires individual HRIRs which may not be available in practice. This motivates the development of a general-purpose individual-agnostic BL which can provide localization estimates without a priori knowledge of the process or HRIRs underlying the binaurally spatialized audio. This would allow localizing sounds in a broad variety of binaural media (games, music, movies, video conferencing calls), based on the sole assumption that the spatialized sound is intended for binaural playback to a human listener. Another potential use case for a general-purpose BL is to provide an estimate for how an *average* listener might localize a certain binaural rendering.

Here, we propose a general-purpose individual-agnostic BL using a CNN that outputs a spatial activation map covering the entire sphere for each processed audio frame, allowing extension to multiple/moving sources. The contributions of the present work are: a) design of a full-sphere, HRIR dataset-independent model output format, b) design of an individual-agnostic training scheme with noise and reverberation augmentation that uses soft targets rather than hard binary targets, and c) a general-purpose BL with robust localization performance on binaural speech recordings with unknown HRIRs, noise,

*This work was done while Shoken Kaneko was an intern at Microsoft Research Labs in Redmond, WA, USA.

and reverberation, thus supporting any binaural audio input. Results are shown for a single model and an ensemble of the proposed CNN architectures trained with different conditions.

II. Proposed method

The proposed BL takes a pair of binaural audio signals as input and produces directional activation maps (Fig. 2) for each frame. A short-time Fourier transform (STFT) converts both channels into a 2 × F × T log-magnitude spectrogram as well as a 1 × F × T spectrogram of interaural phase differences (IPDs), where T and F are the number of time frames and frequency bins, respectively. The log-magnitude spectrogram and the IPD spectrogram are fed into a "two-legged" CNN, inspired by [10]. The network architecture and hyperparameters are shown in Fig. 1 and Table I, respectively. Each leg of the CNN has six convolution blocks, each consisting of convolution, batch normalization, max-pooling, and the nonlinearity. Max-pooling was applied with a pooling factor of two in the frequency axis and no pooling was applied in the time axis. The outputs of the two CNNs are concatenated and fed into a fully-connected (FC) network with one hidden layer. The leaky-ReLU activation function was used for all layers except for the output layer which has the sigmoid activation function. The output layer forms the directional activation map, where each output neuron is assigned to a specific direction on the sphere. The output of the network is a D × T matrix where D is the number of bins in the output direction map. Here, the 2048-point spherical Fibonacci grid [12] was used for the output direction map. This dataset-independent output format was motivated by the fact that most available public HRIR datasets do not have a common angular grid. The proposed output format allows combining datasets with arbitrary angular grids, thus increasing the size of the available training data. Furthermore, this output format is source-number independent and facilitates the application of the model to multi-source localization. The network is trained using the AdaMod optimizer [13] with the multivariate binary cross-entropy (BCE) loss between the network output and the ground truth target as the minimization objective. Rather than using binary classification targets of "source present" or "source absent" for each output grid point, soft targets based on the von Mises-Fisher probability density function are used for a source with ground truth direction $\boldsymbol{\mu} \in \mathbb{R}^d$:

$$\boldsymbol{f}(\mathbf{x};\boldsymbol{\mu},\kappa) = C(\kappa)\exp\left(\kappa\mathbf{x}\boldsymbol{\mu}\right) \in \mathbb{R}^{\mathrm{D}}, \qquad (1)$$

where $\mathbf{x}$ is the $\mathrm{D} \times d$ matrix of output directions, $d = 3$ for the (x, y, z) coordinates, $C(\kappa)$ is a coefficient normalizing the L^{∞} norm of the vector $\boldsymbol{f}(\mathbf{x};\boldsymbol{\mu},\kappa)$, and $\exp$ is the element-wise exponential function. The concentration parameter κ is initialized with 2, and is doubled every 100 iterations until reaching 512, to gradually sharpen the soft targets. The use of soft targets was inspired by the success of "fuzzy" targets in musical onset detection [14].

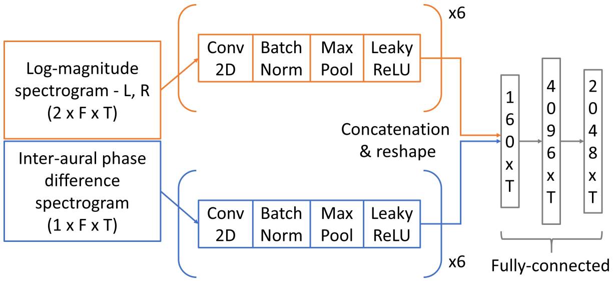

Fig. 1. The network architecture of the proposed model.

TABLE I
Model hyperparameters

Sampling rate	32kHz
FFT size	512 taps
STFT hop size	120 taps
STFT window	Kaiser ($\beta = 4$)
Kernel sizes (frequency-axis)	(5, 5, 5, 3, 3, 3)
Kernel sizes (time-axis)	(5, 5, 5, 3, 3, 3)
# Feature maps in each conv. layer	(5, 10, 20, 40, 80, 80)
# Hidden units in FC layer	4096
# Output units	2048

III. Experimental evaluation

The proposed method was evaluated in terms of the localization error angle of a single static source, with preliminary results for multiple or moving sources. For the single static case, three different test tasks were used. The first two tasks are localization given binaural audio synthesized by convolving speech with public BRIR datasets, namely a dataset of horizontal plane BRIRs for rooms with various reverberation times (*IoSR Rooms*) [15], and a dataset of 22.2 channel BRIRs (*IoSR 22.2ch*) [16]. The third test task is localization of static sound sources present in binaural recordings; LOCATA challenge corpus - evaluation set - task 1 (*LOCATA Task1*) [17].

A. Model training and validation

The input binaural signal for training was generated by convolving HRIRs from public datasets with clean speech signals. The CIPIC [18], ARI [19], RIEC [20], ITA [21], Viking [22], and CHEDAR [23] HRIR datasets were used for training, resulting in 1680 subjects in total. The HRIRs captured using the KEMAR manikin were excluded from the CIPIC and Viking datasets. From the CHEDAR dataset, the HRIRs with source distances of 0.5 m, 1 m, and 2 m for the first 1240 shapes were used for training. The HRIRs associated with the first thirteen shapes and source distance of 2 m from the CHEDAR dataset were used to evaluate the model on seen HRIR data (*Seen*). Acoustically measured HRIRs from the last thirteen subjects excluding dummy heads from the HUTUBS dataset [24] were used for validating the model on unseen HRIR data (*Unseen*). The GRID corpus [25] was used as the speech source. The set of speakers was split into 90%, 5%, and 5% and the first two subsets were used for training and validation, respectively. It is known that noise and room reverberation degrade the performance of a BL [7], [10], [11], [26]. To emulate challenging real-world conditions, stereo white noise and BRIRs simulated by the image source method [27] were added to the input binaural signal. During training, the signal-to-noise ratio (SNR) was randomized by uniformly

TABLE II
BRIR SIMULATION PARAMETERS AND THEIR RANGES.

Room width, depth	(3m, 12m)
Room height	(3m, 10m)
Source/receiver position	At least 1 m from the walls
Source/receiver height	(1m, 2m)
Reflection order	20 (fixed)
Wall impedance ratio	(5, 19)
Receiver's sight direction	Yaw angle in $(0, 2\pi)$

sampling the realized SNR from the range $[L_{\text{Noise}}^{\max}, 40]$ dB_A. The noise was normalized accordingly to realize this SNR. $L_{\text{Noise}}^{\max}$ is a hyperparameter corresponding to the minimum SNR which was chosen from $\{20, 25, 30, 40, \textit{None}\}$, where *None* represents the clean condition without additive noise. An ensemble of the proposed models was formed from five models trained with these five different maximum noise level conditions. BRIRs excluding the direct path were precomputed using HRIRs from the train and validation set. The direct path signal is prepared separately on-the-fly during training. The room simulation parameters were uniformly sampled from predefined ranges of room dimensions and impedance of the walls, as shown in Table II. Four different sets of synthesized binaural signals were used to monitor the performance of the model during training, which includes binaural signals synthesized using a subset of the seen training set HRIRs and the unseen validation set HRIRs, with or without additive stereo noise and binaural reverberation. These four datasets are denoted in the following as *Seen clean* (same HRIRs in training and validation), *Seen N+R* (additive noise and reverberation), *Unseen clean*, and *Unseen N+R*, respectively. 64 directions based on the spherical Fibonacci grid were used to sample HRIR directions for the validation runs. During validation, the SNR was set to 30 dB_A for the *N+R* cases. The direction with maximum activation in the output direction map is considered as the estimated sound source direction in the static single source case. The direction maps were accumulated by taking the maximum over all time frames before making the decision about the sound source direction. Analyses based on lateral and polar angle error have revealed that the error is dominated by the polar angle error while the lateral angle localization can be highly precise. The polar angle alone, however, is problematic as an evaluation metric since it has singularities at the left and right pole. Hence, following prior works [5], [26], the models were evaluated by the total mean angular error (MAE) where the angular error is the angle between the estimated and ground truth source direction. The models were trained for 200k iterations and the model with the best total MAE averaged over *Unseen clean* and *Unseen N+R* was chosen as the model to evaluate the performance on the separate test sets. The model hyperparameters were manually tuned while monitoring the performance on the validation sets.

B. Single static source localization

Visualizations of the output direction maps of the proposed model on the LOCATA-task1 test clips are shown in Fig. 2. It is interesting to observe the emergence of cones of confusion, i.e., rings of constant lateral angle on the spherical map, even though the proposed model does not explicitly estimate lateral or polar angles. Similarly, human listeners tend to have worse polar angle localization than lateral angle localization, especially when listening to non-individual HRIRs [28].

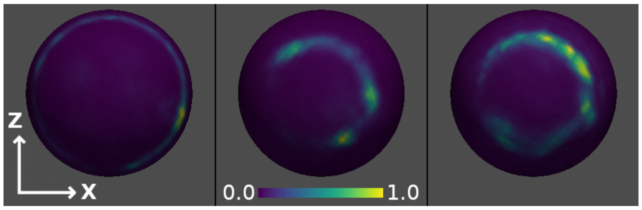

Fig. 2. Typical output activation maps taken from LOCATA-task1 test clips. The right hemisphere of the map is shown with the L^∞ norm of the post-sigmoid output vector normalized after taking the maximum over time frames.

The localization results are summarized in Table III. The proposed model or its training condition was modified by not adding white noise and/or BRIR at training time, by band-limiting the input speech using a low-pass filter with cutoff frequency of 8 kHz, by using white noise as the source signal, or by using hard targets instead of the soft targets for supervision. Modifications to the sources, i.e., band-limiting or using white noise, were applied to both the *Seen* and *Unseen* sets. In the case of hard targets, the target direction bin closest to the ground truth was set to one and all other bins were set to zero. In Table III we can see that the addition of BRIRs, the use of speech as the source signal, and the use of soft targets substantially improve localization performance.

The proposed model was compared with reference models which use either the categorical cross entropy (CCE) or a multi-task CCE loss [10] which is the mean of two CCE losses of two classifiers estimating the lateral and polar angle separately. The CCE loss has been commonly used in previous DNN-based BLs [7]–[9], [11]. Here, the CCE loss model uses the same 2048 direction bins as the direction classes. The multi-task model uses 37 lateral and 72 polar angle classes with 5 degrees step, resulting in 2522 different directions. Except for the output layer, the network architecture and training conditions of the reference models are identical to the proposed model. From Table III it can be observed that the proposed model outperforms these reference models.

In order to study the localization performance of individual-specific (IS) BLs, 13 subjects each from the CHEDAR, simulated HUTUBS, and measured HUTUBS dataset were used to train IS models. An ensemble of IS models was also formed for each HRIR dataset. This ensemble was formed from multiple IS models whose hyperparameters were optimized separately, except the hyperparameters in Table I which were fixed across all models. It was observed that some IS models from the CHEDAR dataset outperform the proposed model in LOCATA-task1. However, as can be seen in the MAE on the other test sets, the maximum MAE for the IS models tend to be larger, indicating that the individual-agnostic model operates more robustly. The results also imply that for a IS model

TABLE III

MEAN ANGULAR ERROR (DEGREES) FOR VARIOUS MODELS.

Model	Seen Clean	Seen N+R	Unseen Clean	Unseen N+R	IoSR Rooms	IoSR 22.2ch	LOCATA Task1	Test Avg.	Test Max.
Proposed (BCE + soft targets)	37.8	48.0	43.6	50.0	16.4	28.6	27.9	24.3	28.6
w/o Noise	31.3	47.6	43.5	50.5	30.7	27.7	24.9	27.8	30.7
w/o Reverb	6.3	53.6	30.5	56.5	35.1	33.2	71.1	46.4	71.1
w/o Noise & Reverb	5.4	69.8	30.4	69.7	38.7	42.9	60.3	47.3	60.3
Band-limited speech[1]	35.9	47.1	39.3	47.0	31.7	30.0	26.4	29.4	31.7
White noise source[2]	20.1	38.5	31.6	49.9	49.7	57.1	65.7	57.5	65.7
BCE + hard targets	46.2	55.0	48.9	51.7	24.9	30.3	49.9	35.0	49.9
CCE loss	49.8	52.6	46.0	52.5	36.9	34.7	38.3	36.6	38.3
Multi-task loss [10]	46.4	51.0	43.7	50.3	46.4	35.8	41.3	41.2	46.4
Proposed ensemble	34.3	45.0	40.9	47.9	**14.4**	**25.0**	26.0	**21.8**	**26.0**
CHEDAR (IS-average)	27.2	39.2	58.3	59.8	28.6	49.6	23.4	33.9	49.6
CHEDAR (IS-ensemble)	-	-	-	-	23.6	47.2	**17.4**	29.4	47.2
HUTUBS-sim (IS-avg.)	1.3	7.1	40.7	40.3	43.4	34.5	64.7	47.5	64.7
HUTUBS-sim (IS-ens.)	-	-	-	-	39.2	26.5	58.3	41.4	58.3
HUTUBS-meas (IS-avg.)	0.3	2.6	38.4	35.5	56.8	43.2	70.9	57.0	70.9
HUTUBS-meas (IS-ens.)	-	-	-	-	53.7	35.5	66.3	51.8	66.3

to be effective, an individual which delivers high expected localization performance needs to be empirically sought.

C. Multiple or moving source localization

To test multiple source localization, binaural recordings from the evaluation set of LOCATA-task2 were used. Example network output activation maps, projected onto the lateral-polar grid, for input audio containing up to four sound sources are shown in Fig. 3. The number of feature maps in the l-th convolution layer was set to 2^{l+2} in this experiment and the model was trained for 1 million weight updates using training data with two sound sources. It can be observed that the model output has multiple vertical lines of high activation in the lateral-polar projection which correspond to the rings of confusion observed in the spherical output map. The proposed model can be used in a sliding window fashion to produce a sequence of directional activation maps. To demonstrate moving source localization, the binaural recordings from the LOCATA-task3 were processed with the proposed model and the sequences of output activation maps were recorded. Animated visualizations of the model output maps are available online[3].

D. Response to panned stereo audio

The interaural time difference (ITD) and the interaural level difference (ILD) are important acoustic cues that affect binaural localization. Artificial modification of the ILD or the ITD is known as amplitude- or delay panning and has been established as techniques to create sound images in stereophonic sound. We use panned stereo to test whether the model generalizes to binaural-like signals not seen during training. Fig. 4 shows the resulting lateral and polar angle estimates made by the proposed model given panned stereo input. The results for delay panning suggest that the model has learned to associate the ITD with the lateral angle. As opposed to the delay panning case exhibiting a smooth curve for the estimated lateral angle, amplitude panning resulted in a noisy step-like profile.

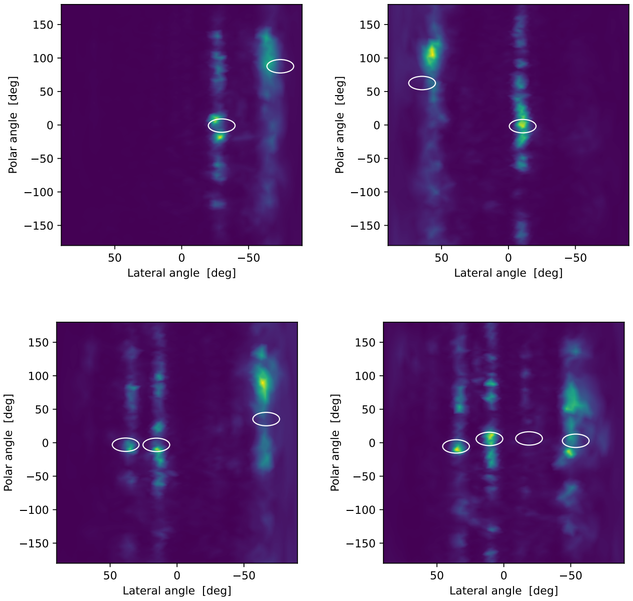

Fig. 3. Output maps shown in the lateral-polar grid, for LOCATA-task2 test clips with two to four sources. The ovals denote the true source directions.

IV. CONCLUSION AND DISCUSSION

We proposed a general-purpose individual-agnostic binaural localizer that does not require a priori knowledge about the generation process of the binaural audio and covers sound sources located anywhere on the 4π sphere. The proposed model output format allows training on HRIR datasets with arbitrary angular grids and in an individual-agnostic manner, and together with the proposed training scheme resulted in a robust binaural localizer which theoretically generalizes to alternative spatialization methods and multiple/moving sources. We have empirically shown the benefits of individual-agnostic training, data augmentation by noise and reverberation, and the

[3] https://github.com/microsoft/Binaural-localizer-demos

[1] Band-limited speech was used for training, seen and unseen evaluation

[2] White noise source was used for training, seen and unseen evaluation

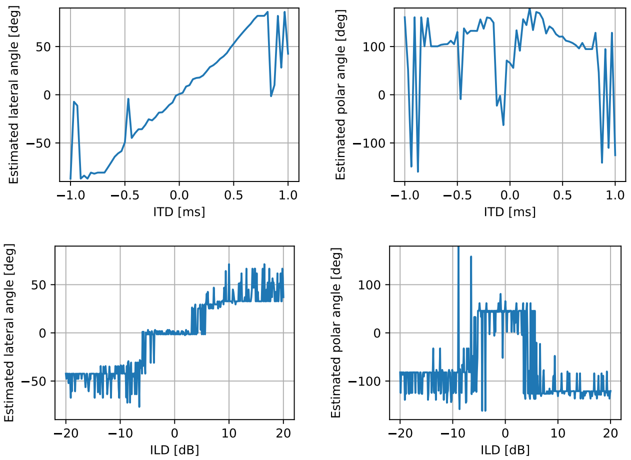

Fig. 4. The estimated lateral (left column) and polar (right column) angle as function of ITD (top row) or ILD (bottom row) of the panned stereo signal.

use of soft targets. The proposed method was tested with real BRIRs as well as binaural recordings which contain noise and reverberation and exhibited superior localization performance compared to individual-specific BLs or BLs trained with CCE or multi-task losses or hard targets. While some single-subject models performed well on one of the test sets, the results indicate that the individual-agnostic scheme may be more robust and generalize better.

Future work may include further refinement of the proposed model for multiple/moving sources and training on non-speech audio as well as alternative spatialization methods, including simple ITD and ILD panning. Finally, a comparison with humans' subjective localization tests may indicate whether the proposed model could predict subjective localization performance in psychoacoustic studies, as indicated by [11].

References

[1] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 791–802, 2014.

[2] F. Palmieri, M. Datum, A. Shah, and A. Moiseff, "Learning binaural sound localization through a neural network," in *Proceedings of the 1991 IEEE Seventeenth Annual Northeast Bioengineering Conference*. IEEE, 1991, pp. 13–14.

[3] A. Moiseff, F. Palmieri, M. Datum, and A. Shah, "An artificial neural network for studying binaural sound localization," in *Proceedings of the 1991 IEEE Seventeenth Annual Northeast Bioengineering Conference*. IEEE, 1991, pp. 1–2.

[4] C. Neti, E. D. Young, and M. H. Schneider, "Neural network models of sound localization based on directional filtering by the pinna," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3140–3156, 1992.

[5] C. Jin, M. Schenkel, and S. Carlile, "Neural system identification model of human sound localization," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1215–1235, 2000.

[6] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[7] 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.

[8] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial audio feature discovery with convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6797–6801.

[9] 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*, vol. 2020, no. 1, pp. 1–16, 2020.

[10] Y. Yang, J. Xi, W. Zhang, and L. Zhang, "Full-Sphere Binaural Sound Source Localization Using Multi-task Neural Network," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 432–436.

[11] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature Human Behaviour*, vol. 6, no. 1, pp. 111–133, 2022.

[12] Á. González, "Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices," *Mathematical Geosciences*, vol. 42, no. 1, pp. 49–64, 2010.

[13] J. Ding, X. Ren, R. Luo, and X. Sun, "An adaptive and momental bound method for stochastic learning," *arXiv preprint arXiv:1910.12249*, 2019.

[14] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 6979–6983.

[15] C. Hummersone, "Binaural Room Impulse Response Measurements," https://github.com/IoSR-Surrey/RealRoomBRIRs, 2011.

[16] J. Francombe, "IoSR Listening Room Multichannel BRIR dataset'," https://github.com/IoSR-Surrey/IoSR_ListeningRoom_BRIRs, 2017.

[17] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. IEEE, 2018, pp. 410–414.

[18] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 99–102.

[19] P. Majdak, "ARI HRTF database," https://www.oeaw.ac.at/isf/das-institut/software/hrtf-database, 2017.

[20] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoustical science and technology*, vol. 35, no. 3, pp. 159–165, 2014.

[21] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in *Proceedings of Meetings on Acoustics 172ASA*, vol. 29, no. 1. Acoustical Society of America, 2016, p. 050002.

[22] S. Spagnol, K. B. Purkhús, R. Unnthórsson, and S. K. Björnsson, "The Viking HRTF dataset," in *16th Sound and music computing conference*. Sound and Music Computing Network, 2019, pp. 55–60.

[23] S. Ghorbal, X. Bonjour, and R. Séguier, "Computed hrirs and ears database for acoustic research," in *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

[24] B. Fabian, D. Manoj, P. Robert, W. Jan Joschka, S. Fabian, V. Daniel, G. Peter, and W. Stefan, "The HUTUBS head-related transfer function (HRTF) database," 2019.

[25] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[26] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Individualized interaural feature learning and personalized binaural localization model," *Applied Sciences*, vol. 9, no. 13, p. 2682, 2019.

[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[28] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.