Binaural source localization using deep learning and head rotation information

Guillermo García-Barrios[1], Daniel Aleksander Krause[2], Archontis Politis[2], Annamaria Mesaros[2], Juana M. Gutiérrez-Arriola[1], Rubén Fraile[1]
[1]*Group on Acoustics and MultiMedia Applications*, *Universidad Politécnica de Madrid*, Madrid, Spain
{guillermo.garcia.barrios, juana.gutierrez.arriola, r.fraile}@upm.es
[2]*Computing Sciences*, *Tampere University*, Tampere, Finland
{daniel.krause, archontis.politis, annamaria.mesaros}@tuni.fi

***Abstract*—This work studies learning-based binaural sound source localization, under the influence of head rotation in reverberant conditions. Emphasis is on whether knowledge of head rotation can improve localization performance over the non-rotating case for the same acoustic scene. Simulations of binaural head signals of a static and rotating head were conducted, for 5 different rotation speeds and a wide range of reverberant conditions. Several convolutional recurrent neural network models were evaluated including a static head scenario, a model without rotation information, and distinct models differentiated on the way of manipulating the quaternions. The results were analyzed based on the direction-of-arrival error, and they show the importance of using quaternions as additional features, with the best localization accuracy obtained when using an additional convolutional branch that merges the features through addition or concatenation. Nevertheless, raw quaternion features presented lower performance than the static baseline model. Additionally, the study shows the importance of the analysis time window length when using information about head rotation.**

***Index Terms*—binaural sound source localization, head rotation, deep neural networks, quaternions**

I. Introduction

Binaural sound source localization (SSL) consists of estimating the position of a sound source using the signals captured by the two ears. SSL has multiple applications in teleconferencing [1], speech separation [2], hearing aids [3], and human-robot interaction [4]. Additionally, there has been a recent increase in the use of microphones in wearable devices, which opens a new field for innovative applications. Based on these applications and the idea that head movement can resolve localization ambiguities in human localization, we analyze how head rotation can contribute to artificial binaural SSL.

Different approaches have been proposed for binaural SSL so far. In [5] the authors estimated the azimuth angle through the joint evaluation of interaural time differences (ITDs) and interaural level differences (ILDs). After that, Zannini *et al*. extended the previous method probing the adverse effect of reverberation [6]. The ITDs were also used for the azimuth estimation in an application for a mobile robot, where they incorporated head-related transfer functions (HRTFs) for improving localization precision [7]. Other research exploited the frequency-domain diversity of HRTFs for single and multiple source scenarios estimating azimuth and elevation angles [8].

Recently, learning-based SSL employing deep neural networks (DNNs) has seen widespread research interest; the reader is referred to a recent review for more information [9]. Regarding binaural approaches, Youseff *et al*. [10] estimated azimuth and elevation angles using video pixel coordinates and binaural cues as the input of a DNN for robotic SSL. Another DNN approach including clustering for evaluating the mismatched HRTF condition was introduced in [11]. More recently, a complex time-frequency mask-guided binaural SSL technique for preserving the direct path of the HRTF and extracting robust binaural cues was proposed in [12].

Only a few articles have studied the influence of head movement in binaural SSL. In [13], the authors tested a machine hearing algorithm in two reverberant rooms, rotating the head between -90º and 90º in the azimuth plane. The evaluation was made for 3 source positions following three strategies: rotation to the exact source positions, rotation towards the source, and random rotation. The same scenario was used in [14] and [15], but limiting the head rotation to the $\pm$30º range and evaluating multiple sources at a full 360º range in 4 different rooms. A similar approach was presented in [16], with rotation limited to 90º and a constant rotation speed of 45º per second for only 5 possible azimuth sound source positions placed at 3 meters from the head. May *et al*. [17] trained their model with a $\pm$60º azimuth range and tested it with a $\pm$30º range for 4 different rooms and 3 speech source positions. Another experiment used several rotation speeds for localizing a speech source in one fixed position in one reverberant room [18]. The most complete work known to the authors was carried out by Lu and Cooke [19]. They estimated azimuth and distance evaluating 8 listener motion strategies including random walking and head rotation for one anechoic and 3 reverberant scenarios, demonstrating the importance of motion-based cues. Despite that, the head rotation was limited to $\pm$90º.

The main contribution of this work is a study of the influence of using head rotation information for binaural SSL. The proposed method mimics human audition, in which we ac-

This work was supported by the Universidad Poltécnica de Madrid through its Programa Propio de I+D+I, specifically the Predoctoral Call, and computing resources on Magerit Supercomputer and the CSC – IT Center for Science of Finland.

tively turn our head to resolve ambiguous localization. Analogously, the method tries to obtain better localization estimates by using rotation information while imposing less restrictions than previous studies, for instance regarding the range for head rotation. The reported experiments investigate the improvement obtained in estimating the direction of arrival (DOA) of one static speaker in a wide range of acoustic scenarios when including head rotation in the azimuth plane as an input for several DNN models.

The rest of the paper is structured as follows: section II presents the proposed method for simulating the head rotation in a DNN-based system; section III describes the dataset used for training the models; section IV shows and analyzes the experimental results; last, section V presents conclusions and future work.

II. Method

A. Model

To perform sound source localization, we employ a typical convolutional recurrent neural network (CRNN) model as a baseline. This kind of DNN has been shown in many studies to perform efficiently in DOA estimation tasks [20]–[22]. The architecture of the utilized model is depicted in Fig. 1.

For the baseline architecture, a $CH \times T \times 256$ feature matrix is fed to the input of the network, where CH stands for the number of feature channels and T denotes the number of frames in a modeled sequence. During training and testing, each file is split into clips of length T. Across all experiments, we use a specific set of audio features. A complex spectrogram of the signal is obtained using a Short-Time Fourier Transform (STFT) with a Hamming window of length 40 ms and 50% overlap. Next, we extract the mean magnitude spectrogram from both binaural channels. To represent spatial information about the signal, we extract sines and cosines of Interaural Phase Differences (IPD), which provide a smooth representation of phase values and avoid phase wrapping. This feature has been shown to outperform traditional phase differences in speech separation [23] and DOA estimation [24]. On top of that, we use ILDs, which constitute another major binaural cue that becomes important above 1.5 kHz. This set of features results in $CH = 4$ feature channels that are fed to the model.

The features are initially processed by three convolutional blocks. Each block consists of a 2D convolutional layer containing 128 3×3 filters, followed by a max-pooling operation across the frequency dimension. The pooling rate of each layer is described by the vector MP mentioned later. Next, the processed feature maps are passed to a single bi-directional gated recurrent unit (GRU), which allows for temporal modeling of a sound signal and enables more efficient use of inter-frame relations, which we consider useful when utilizing head rotation information. Finally, the temporal model features are passed to two fully-connected (FC) layers. The first layer consists of 128 linear neurons, whereas the latter one outputs three Cartesian coordinates of the source direction, these processed by a hyperbolic tangent activation function.

For a number of experiments that involve head rotation, we introduce an additional input convolutional branch in order to process rotational information. For this purpose, we use a block of two convolutional layers, each consisting of 128 filter kernels. This block is fed with a matrix $1 \times T \times 4$, in which for each time frame 4 quaternion values are used to express head rotation. The values of pooling rates in the main branch depend on the use of the rotational branch. For the baseline single-branch model, values $MP = [2, 2, 2]$ are used, whereas for a DNN including the rotational branch we increase the pooling rate to $MP = [4, 4, 4]$ to balance the outputs of both convolutional blocks so that they have identical dimensions.

B. Investigated methods

We propose several methods for including the rotational information in the final model and compare results with a baseline model that does not use head rotation. Hereby we describe all experimental scenarios:

- **Static:** as a baseline scenario, we perform DOA estimation with no head rotation.
- **No rotational features:** we investigate a scenario with a rotating head without the use of rotation features. In this method, the network is supposed to learn head movement patterns from spectral features alone by utilizing the temporal context of many frames.
- **Single matrix:** we include the rotational features without the additional convolutional branch by stacking the quaternion values with the spectral feature matrix. For this purpose we repeat all four quaternion values 64 times to create an additional feature channel layer fitting the dimension $M = 256$, hence $CH = 5$.
- **Cat-raw:** the matrix containing raw quaternions is being concatenated with the output of the convolutional block before feeding it to the recurrent layer.
- **Processed:** finally, we utilize the convolutional branch to process the quaternion values before merging them with the spectral features. Three different methods of merging the two types of features are used: elementwise multiplication (**mul-proc**), addition (**add-proc**) and concatenation (**cat-proc**).

III. Dataset

The dataset used for training and testing the models was generated using a room simulator for calculating the room impulse responses (RIRs) in the Ambisonics format [25], which is the industry standard in 360° spatial audio coding and reproduction. Ambisonics is chosen as an intermediate representation because it allows rotations of the whole sound field in an instantaneous frequency-independent manner. The process is split into a) computing ambisonic RIRs and the respective reverberant speech signals in an unrotated frame of reference, b) applying time-variant rotation matrices to the ambisonic signals in the time-domain, and c) applying time-invariant binaural decoding filters to convert the rotating ambisonic signals to binaural ones. This process avoids the complexity

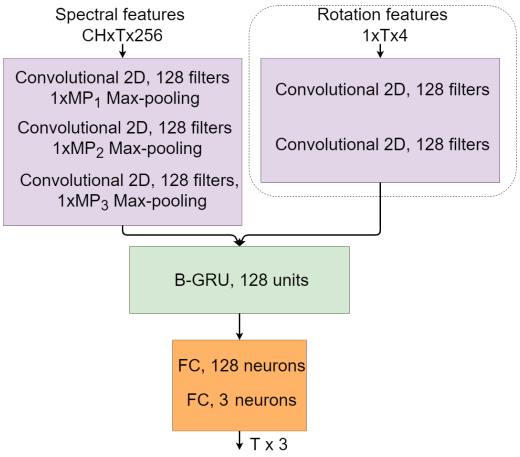

Fig. 1. Architecture of the utilized deep neural network. The dotted line denotes the optional convolutional branch to process rotation features.

of producing directly rotating reverberant binaural signals, which would typically require time-variant filter switching and interpolation [26].

A. Room simulation

The spatial RIRs were calculated using an image-source room simulator for shoebox geometries [27]. The simulator allows frequency-dependent wall absorption and directional encoding of the image sources in Ambisonics format of arbitrary order. In total, 5000 different rooms were simulated for the 5^{th} order of Ambisonics. Tab. I shows the randomized parameters for performing the simulations. The elevation range between the source and the receiver was [-35° 35°].

A list of materials and respective absorption coefficients for each surface type (ceiling, floor and wall) was compiled from the most common ones specified in acoustical engineering tables [28], [29]. For each simulated room instance, one material was randomly assigned to each surface, for a total of 2912 possible combinations. By randomizing materials instead of reverberation times directly, as is commonly done, we avoided matching unnatural reverberation times to certain room volumes (e.g. very long RT60 for a small room) and we ended up with a natural distribution of them. The final RT60 distribution presented a median, 10^{th} and 90^{th} percentile of 0.83 s, 0.42 s and 2.38 s, respectively. Additionally, the source positions were uniformly distributed in terms of the azimuth angle with respect to the receiver.

Male and female anechoic speech recordings obtained from the TIMIT database [30] were convolved with the simulated ambisonic RIRs. Multiple recordings of the same speaker were concatenated to obtain audio files with a length of 10 s. The experiments included 250 such audio files of 10 s duration with a sampling rate of 24 kHz. At the end, 5000 convolved audios of 10 seconds were obtained, where just 2500 were used for the experiments. More in detail, the test and validation splits consisted of 500 files, whereas the training split included 1500 files.

TABLE I
RANDOMIZATION OF PARAMETERS FOR DATA GENERATION

Parameter	Random range
Room width and length	[3.0 15.0] m
Room height	[2.0 7.0] m
Num. of materials (wall, floor, ceiling)	13, 7, 8
Source / receiver height	[1.5 2.2] m
Source-to-surface distance	$>$ 0.5 m
Source-to-receiver distance	$>$ 1.0 m
Source-to-receiver angle	$>$ 30º
Azimuth angle	[-180.0º 180.0º]
Head rotation speed	10, 20, 30, 40, 50 deg/s

B. Rotation integration

The initial 5^{th} order ambisonic reverberant speech $\mathbf{a}_{\text{stat}}(t) = [a_1(t), ..., a_{36}(t)]^{\text{T}}$, represent the directional information of the scene at the receiver for a static listener orientation aligned with the global coordinate frame. To emulate listener head movements, the listener's frame is rotated according to a rotation trajectory defined in terms of quaternions $\mathbf{q}(t_r)$. Quaternions are hypercomplex numbers $\mathbf{q} = ix + jy + kz + w$ formed by one real and three imaginary components, with $x, y, z, w \in \mathbb{R}$, and they are the preferred representation in rotational tracking devices due to being compact while avoiding rotational ambiguities [31]. Spherical linear interpolation (SLERP) was used to produce quaternion values at the desired temporal resolution from reference values defining the trajectory. Quaternions were converted into spherical harmonic rotation matrices [32] for ambisonic signals $\mathbf{M}(\mathbf{q}(t_r))$. Rotation matrices were computed at a temporal resolution of 10 ms and applied to 20 ms ambisonic signal frames with 50% overlap using a Hanning window. In a single windowed frame, the rotating ambisonic signals $\mathbf{a}_{\text{rot}}$ were

$$\mathbf{a}_{\text{rot}}(t) = \mathbf{M}(\mathbf{q}(t_r))\,\mathbf{a}_{\text{stat}}(t). \quad (1)$$

While this operation assumes a fixed rotation for each short windowed segment, it was empirically found suitable to give a continuous representation of rotation for all target speeds.

After the rotating ambisonic signals were produced, an ambisonic-to-binaural decoding was performed, expressed as a 2×36 matrix of filters $\mathbf{D}(t)$. The binaural decoding filters were derived from a densely measured set of HRTFs [33], using the magnitude-least-squares approach [25]. The final rotating binaural signals $\mathbf{y}_{\text{rot}}(t) = [y_1(t), y_2(t)]^{\text{T}}$ were

$$\mathbf{y}(t) = \mathbf{D}(t) * \mathbf{a}_{\text{rot}}(t), \quad (2)$$

where the matrix convolution operation denotes multichannel filter-and-summing across input channels, such that $y_i(t) = \sum_j D_{ij}(t) * a_j(t)$. It is noted that the simulated spatial resolution of 5^{th} order Ambisonics was chosen in order to strike a balance between the computational complexity of encoding reverberation at high orders and modeling the most important aspects of binaural localization: accurate inter-aural time differences at low frequencies, and accurate inter-aural level differences at mid-high frequencies. With

TABLE II
EXPERIMENTAL RESULTS OBTAINED WITH DIFFERENT MODEL TYPES.

Model type	T	DOA error [°]
Static	250	17.62 ± 1.24
No rot. features	250	18.01 ± 1.12
Single matrix	250	16.28 ± 1.09
Cat-raw	250	19.04 ± 1.15
Mul-proc	250	17.04 ± 1.10
Add-proc	250	16.02 ± 1.12
Cat-proc	250	**16.01** ± 1.08
Cat-proc	500	**15.06** ± 0.98
Cat-proc	999	**15.03** ± 0.99

the binaural decoding scheme used herein, it has been shown that those binaural cues are represented adequately [25].

Regarding rotation trajectories, we considered anti-clockwise azimuth head rotation for 5 different angular speeds, as seen in Tab. I. Each speed was assigned to 1000 rooms in the dataset. As the length of the audios is fixed to 10 s, the angle travelled by the head depends on the angular speed.

IV. EXPERIMENTAL RESULTS

To evaluate our models, we use the DOA error as a performance measure, defined for the entire dataset as follows:

$$E_{\text{DOA}} = \sum_{n=0}^{N-1} \sigma(\mathbf{x}_R(n), \mathbf{x}_E(n)), \tag{3}$$

where n denotes the index or segment over which DOAs are estimated by the model. σ stands for the angular distance between two vectors:

$$\sigma = \arccos\left(\frac{\mathbf{x}_E^{\mathrm{T}}\mathbf{x}_R}{||\mathbf{x}_E|| \cdot ||\mathbf{x}_R||}\right) \tag{4}$$

Here, $\mathbf{x}_E$ and $\mathbf{x}_R$ denote the estimated and reference DOA vectors. The final DOA error is averaged over a 5-fold cross-validation split.

Table II shows the results obtained for the different methods described in II-B. The baseline model operating in a static scenario scored a DOA error of $17.62°$, which is a satisfactory performance for binaural localization under reverberant conditions. For a model with a rotating head but omitting rotation features we observe a slight increase of error by $0.39°$, which shows that the network is not able to learn rotation information from spectrum-based features alone. However, adding the quaternion values as additional features can noticeably improve the results, which is shown for most of our proposed methods. Utilizing a single feature matrix with an additional quaternion channel already reduces the error to $16.28°$.

Best results are obtained when using the *cat-proc* and *add-proc* methods, for which the DOA error remains at around $16°$. The *mul-proc* method achieves a slightly lesser improvement by around $1°$. The worst result is observed for the *cat-raw* technique, for which the DOA error is equal to $19.04°$, an even lower performance than the baseline static scenario. This suggests that raw quaternion features might not be enough

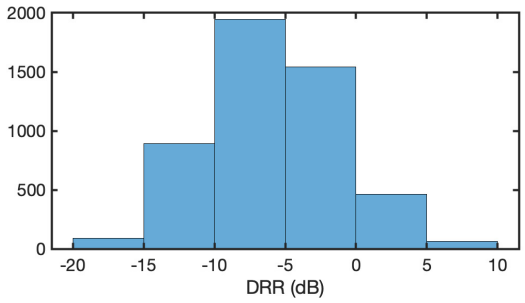

Fig. 2. Histogram of the DRR for the simulated dataset.

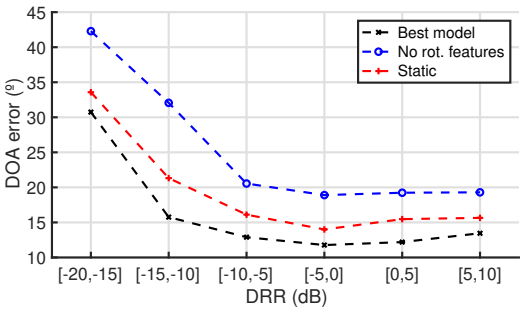

Fig. 3. Median error of the DOA for different DRR ranges.

to teach the model how to follow different head rotation paths, and that the additional processing by a set of convolutional filters allows the DNN to learn better representations.

Most models are trained and tested for data split into sequences with a fixed length of $T = 250$ frames, which is roughly equal to 2.5 s. However, most scenarios included rotation paths longer than that, hence additional experiments based on the *cat-proc* setup were conducted for longer sequence lengths. As can be seen in the lowest rows of Tab. II, increasing the modeled context to 500 frames lowers the DOA error to $15.06°$, which is by far the best performance in our study. Using a sequence length of 999 frames (the whole file) achieves an error of $15.03°$, which shows a negligible difference with its former counterpart. We notice that these results might be very much data-dependent and a different sweet spot might be found for another dataset. However, these outcomes show that when using information about head rotation, the size of the analyzed time context can significantly influence the final localization performance.

Additionally, the direct-to-reverberant ratio (DRR) has been analyzed to determine how it affects the DOA error. As presented in Fig. 2, most of the DRR values are comprised between -15 and 5 dB. In Fig. 3 we have represented the DOA error for the same DRR ranges. It can be seen that the differences in performance for the three models remain the same for all the DRR intervals.

To analyze the significance of the error reduction achieved when including rotational features as inputs for the detector, we have performed a Wilcoxon test comparing the best model and the static one. The obtained p-values for all the interme-

diate DDR ranges (from -15 dB to 5 dB) is less than 0.006, which means that the DOA error improvement is relevant at the 99.4% level. Nevertheless, p-values of the extreme ranges, DDR $<$ -15 dB, and DDR $>$ 15 dB, are 0.27 and 0.34, thus implying that error reduction is not significant in those scenarios. This is due to the small number of simulations.

V. Conclusion

Different CRNN models were proposed to evaluate theimprovement of the DOA in binaural SSL when rotating the head. The baseline model without head rotation presented acceptable accuracy for a large number of scenarios, even with a DRR of -15 dB. The best performance improvement over the baseline model was obtained when using quaternions as inputs for a convolutional branch whose outputs were concatenated with the spectral features. In addition, a slight further improvement was achieved when extending the modeled temporal context to include more information about the rotation path. The experiments show information about head rotation, which is helpful for a human listener, can be transferred to artificial SSL based on a DNN model as well.

In future work, we plan to include a more detailed analysis of how rotation speeds influence the DOA error, the differences between the estimated DOA and the real one, and how the relative position of the listener with respect to the sound source affects the localization accuracy.

References

[1] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[2] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *IEEE Int. Conf. on Acoustics, Speech, Sig. Proc. (ICASSP)*, 2009, pp. 233–236.

[3] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 611–623, 2017.

[4] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2012, pp. 431–438.

[5] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[6] C. M. Zannini, R. Parisi, and A. Uncini, "Binaural sound source localization in the presence of reverberation," in *Int. Conf. on Digital Sig. Proc. (DSP)*, 2011, pp. 1–6.

[7] X. Wan and J. Liang, "Robust and low complexity localization algorithm based on head-related impulse responses and interaural time difference," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. EL40–EL46, 2013.

[8] D. S. Talagala, W. Zhang, T. D. Abhayapala, and A. Kamineni, "Binaural sound source localization using the frequency diversity of the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1207–1217, 2014.

[9] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A review of sound source localization with deep learning methods," *arXiv e-prints*, 2021.

[10] K. Youssef, S. Argentieri, and J.-L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2012, pp. 217–220.

[11] 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 J. Audio Speech Music Process.*, vol. 2020, no. 1, feb 2020.

[12] H. Liu, P. Yuan, B. Yang, G. Yang, and Y. Chen, "Head-related transfer function–reserved time-frequency masking for robust binaural sound source localization," *CAAI Trans. on Intelligence Technology*, 2021.

[13] N. Ma, T. May, H. Wierstorf, and G. J. Brown, "A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2015, pp. 2699–2703.

[14] 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.

[15] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, p. 2122–2131, 2018.

[16] C. Schymura, F. Winter, D. Kolossa, and S. Spors, "Binaural sound source localisation and tracking using a dynamic spherical head model," in *Interspeech 2015*, 2015.

[17] T. May, N. Ma, and G. J. Brown, "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2015, pp. 2679–2683.

[18] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2016, pp. 430–434.

[19] Y.-C. Lu and M. Cooke, "Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners," *Speech Communication*, vol. 53, no. 5, pp. 622–642, 2011.

[20] D. Krause, A. Politis, and K. Kowalczyk, "Data diversity for improving DNN-based localization of concurrent sound events," in *29th European Sig. Proc. Conf. (EUSIPCO)*, 2021, pp. 236–240.

[21] S. Adavanne, A. Politis, and T. Virtanen, "Differentiable tracking-based training of deep learning sound source localizers," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 211–215.

[22] D. Krause, A. Politis, and K. Kowalczyk, "Comparison of convolution types in CNN-based feature extraction for sound source localization," in *28th European Sig. Proc. Conf. (EUSIPCO)*, 2020, pp. 820–824.

[23] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2018, pp. 1–5.

[24] D. Krause, A. Politis, and K. Kowalczyk, "Feature overview for joint modeling of sound event detection and localization using a microphone array," in *28th European Sig. Proc. Conf. (EUSIPCO)*, 2020, pp. 31–35.

[25] F. Zotter and M. Frank, *Ambisonics*. Springer Nature, 2019.

[26] T. Wendt, S. Van De Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *Journal of the Audio Engineering Society*, vol. 62, no. 11, pp. 748–766, 2014.

[27] A. Politis, "Microphone array processing for parametric spatial audio techniques," *Doctoral Thesis, Aalto University*, 2016, https://github.com/polarch/shoebox-roomsim.

[28] D. Adler, *Metric handbook: planning and design data*. Routledge, 2015.

[29] "Sound absorption coefficient chart: JCW acoustic supplies." [Online]. Available: https://www.acoustic-supplies.com/absorption-coefficient-chart/

[30] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[31] S. B. Choe and J. J. Faraway, "Modeling head and hand orientation during motion using quaternions," *SAE Transactions*, vol. 113, pp. 186–192, 2004. [Online]. Available: http://www.jstor.org/stable/44737869

[32] J. Ivanic and K. Ruedenberg, "Rotation matrices for real spherical harmonics. direct determination by recursion," *The Journal of Physical Chemistry*, vol. 100, no. 15, pp. 6342–6347, 1996.

[33] J. G. Bolaños and V. Pulkki, "HRIR database with measured actual source direction data," in *Audio Engineering Society Convention 133*. Audio Engineering Society, 2012.