Binaural Sound Source Distance Estimation and Localization for a Moving Listener

Daniel Aleksander Krause, Guillermo García-Barrios, Archontis Politis, and Annamaria Mesaros, *Senior Member, IEEE*

***Abstract*—In this paper, we investigate the tasks of binaural source distance estimation (SDE) and direction-of-arrival estimation (DOAE) using motion-based cues in a scenario with a walking listener. On top of performing both tasks as separate problems, we study two methods of solving the joint task of simultaneous source distance estimation and localization (SDEL), with a single model. Experiments are conducted for three different scenarios: a static receiver; a static receiver with a rotating head; and a freely moving listener inside a room. The study proposes rotation and translation features to include information about the receiver's motion during model training and studies the effects of these on the final performance. The work includes extended simulation of three datasets containing numerous testing scenarios for sound sources, covering a wide range of DOAs and a source-to-receiver distance up to 15 m. Results are further analyzed with respect to room reverberation, walking speed, as well as source-to-receiver distance. The presented outcomes show large improvements in both DOA and distance estimation for a model that uses motion-based cues as compared with a static scenario. These include a decrease of 9.50° in DOA and 1.56 m in distance errors for a joint model, followed by 16.17° and 0.17 m for separate models.**

***Index Terms*—Sound source localization, sound distance estimation, binaural audio.**

I. Introduction

SOURCE Distance Estimation (SDE) and Direction of Arrival Estimation (DOAE) constitute an important part of the Computational Acoustic Scene Analysis (CASA) research field and have numerous practical applications: autonomous robots can benefit from information about the location of surrounding objects while moving in space [1], [2], [3]; knowledge on Direction of Arrival (DOA) and distance can help enhance the robustness of speech recognition and separation systems [4], [5]; surveillance systems in public spaces and smart homes utilize this information for detailed descriptions of dangerous events [6], [7]; teleconferencing systems can use such information to beamform and enhance speech [8], [9].

Recent research on CASA shows interest in merged tasks utilizing spatial recordings, such as sound event localization and detection (SELD) [10] or sound event detection and acoustic scene classification [11]. Although DOA and source distance are both estimated from multi-microphone recordings, the latter has been investigated to a much lesser extent [12], for two reasons: successful DOAE offers enough information for many downstream spatial filtering tasks, and it is an easier task than SDE; cues that contribute to SDE vanish or become ambiguous very quickly as the distance from the array grows. The shared information between DOA and distance make a joint approach worth investigating, to obtain estimation of the full position of an object. This can be done in a single task approach, where the position is explicitly estimated, or as a multi-task, where the DOA and the distance are separately estimated. We refer to the joint task as Source Distance Estimation and Localization (SDEL).

Most methods for DOA and distance estimation rely on microphone arrays with more than two microphones [13]. However, two-channel binaural recordings are an important format for acoustic scene analysis, being based on the same spatial cues as the human auditory system, therefore matching the perspective of a human or human-like recorder such as anthropomorphic robots [14]. In particular, in recent years, there has been a notable increase in the use of microphones in wearable on-head devices, which opens a whole new space for innovations [15].

Existing research on SDEL, SDE, or DOAE from binaural signals including a dynamic receiver is fairly limited. Most studies focus solely on DOAE and are mostly limited to a certain range of both azimuth and elevation angles, or just the azimuth [16], [17]. On the other hand, binaural source distance estimation based on deep neural networks (DNNs) is a highly under-researched task, with only a few studies investigating it for a limited set of static distances in a near range up to 5 m [18], [19]. The existing literature offers no insight if binaural SDE is possible at further distances in realistic scenarios. Research on joint SDE & DOAE using binaural audio is even more limited, with only one study investigating joint modeling via a DNN architecture just for a few positions in the close frontal plane [20]. Finally, the effect of motion-based cues has been investigated only for DOAE systems in a limited testing scenario of azimuth localization in the range of ±90° [21]. Therefore, there is no research studying binaural

Manuscript received 2 March 2023; revised 20 October 2023 and 8 December 2023; accepted 15 December 2023. Date of publication 22 December 2023; date of current version 5 January 2024. The work of Annamaria Mesaros was supported by the Academy of Finland under Grant 332063 Teaching machines to listen. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniele Salvati. *(Corresponding author: Daniel Aleksander Krause.)*

Daniel Aleksander Krause, Archontis Politis, and Annamaria Mesaros are with the Computing Sciences, Tampere University, 33720 Tampere, Finland (e-mail: daniel.krause@tuni.fi; archontis.politis@tuni.fi; annamaria.mesaros@tuni.fi).

Guillermo García-Barrios is with the Group on Acoustics and MultiMedia Applications, Universidad Politécnica de Madrid, 33014 Madrid, Spain (e-mail: guillermo.garcia.barrios@upm.es).

This article has supplementary downloadable material available at https://doi.org/10.1109/TASLP.2023.3346297, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2023.3346297

DOAE and SDE as a joint task for a more realistic scenario, including tests on a full sphere of DOAs, a wider range of distances, and a moving listener.

This study investigates the scenario of a moving listener for the tasks of DOAE and SDE, performed both separately and jointly. Based on the shortcomings of the current literature, we aim at answering the following research questions: (1) Does the movement of the listener, be it its position or head orientation, improve the performance of SDE and DOAE systems? (2) Are translation and rotation features an efficient way of expressing the listener's motion? (3) Does a joint SDEL system perform comparably or even better than separate SDE and DOAE models?; and finally (4) Does a joint task achieve better performance as single or multi-task trained systems?

Answering the above questions, the contributions of this study are as follows:

- An investigation of two methods of performing binaural SDEL using a single task and a multi-task approach.
- Exhaustive testing scenarios including the use of 2500 simulated rooms and a continuous natural distribution of source DOAs in both azimuth and elevation for a wide range of distances up to 15 m.
- Performance comparison of a joint SDEL task with separate DOAE and SDE systems for multiple scenarios.
- Exhaustive experiments investigating the influence of motion-based cues on the performance of both tasks, including a rotating head and a moving listener.
- In addition, we provide a study on the DOA and distance error dependence on reverberation, source-to-receiver distance, as well as walking speed.
- Moreover, in order to allow for research reproducibility, three datasets containing all testing scenarios[1] and a code repository including the training framework[2] have been made publicly available.

The rest of the paper is organized as follows. Section II presents an overview of the existing literature regarding binaural DOAE and SDE. In Section III, we describe the proposed SDEL method, including the utilized DNN architectures, feature extraction and training procedures. Section IV explains the method evaluation procedure, including a detailed description of the datasets and evaluation metrics used. In Section V, the experimental scenarios are explained and the results of all the experiments are presented along with a detailed analysis of their content. Finally, Section VI summarizes all the conclusions emerging from this study.

II. Related Work

A. Sound Source Localization

Sound source localization (SSL) is the task of determining the location of a sound source with respect to the receiver. The task can be defined as estimating a certain position of the source or its direction of arrival. In this paper, we focus on DOA estimation, since the remaining information is handled by the source distance estimation task. Studies on DOAE vary greatly in the utilized array geometry and spatial audio format, with popular choices being distributed microphones [22], [23], tetrahedral arrays [24], [25], Ambisonics [26], [27], and binaural recordings [16], [17]. Classical DOA techniques include parametric methods such as the generalized cross-correlation with phase transform (GCC-PHAT) [28], Steered Response Power (SRP) [29], and subspace-based approaches such as MUltiple SIgnal Classification (MUSIC) [30] and Estimation of Signal Parameters via Rational Invariance Techniques (ESPRIT) [31]. These methods perform efficiently in certain scenarios, but are limited by various constraints on the acoustic array geometry and the assumptions of the model. Common problems include the knowledge about the number of estimated sound sources or low robustness against reverberation and difficult noise conditions [32]. Regarding binaural localization, joint evaluation of interaural time differences (ITDs) and interaural level differences (ILDs) was used to estimate the azimuth angle in [33]. Zannini et al. improved this method by probing the adverse effect of reverberation [34], while Wan et al. incorporated head-related transfer functions (HRTFs) for improving the precision of azimuth angle estimation [35]. Subsequent studies exploited the frequency-domain diversity of HRTFs for multi-source scenarios, extending the estimation to both, azimuth and elevation planes [36].

The recent methods based on machine learning (ML) overcome many limitations of the classical model-based solutions. ML models trained via supervised learning adapt to different acoustic conditions when provided with sufficient amount of training data. While early solutions used simple models such as Gaussian Mixture Models [37], Support Vector Machines [38] or kernel estimators [39], state-of-the-art solutions use deep learning techniques [27], [40], [41]. Deep Neural Networks provide a higher model complexity, which enables modeling more compound phenomena by utilizing large amounts of data. A systematic review of deep learning for DOAE is available in [42].

Regarding binaural approaches, Youseff et al. [16] estimated azimuth and elevation angles using video pixel coordinates and binaural cues as the input of a deep model. In [17] the authors used affinity propagation clustering for evaluating the mismatched HRTF condition to improve SSL. More recently, Liu et al. introduced a complex time-frequency mask for extracting robust binaural cues by preserving the direct path of the HRTF [43]. To the authors' best knowledge, no study has been done on regression-based binaural DOA estimation that would be tested for a large number of rooms and a full range of azimuth angles. We investigated binaural localization in a scenario with a rotating head, showing that head rotation information significantly improves the estimation precision [44]. This work is a direct continuation of that study.

B. Source Distance Estimation

Source Distance Estimation (SDE) is the task of estimating the line-of-sight distance of a sound source from the receiver. Compared with DOA estimation, SDE is a very under-researched topic, and widely considered a more difficult task. The main

[1] DOI: 10.5281/zenodo.7689063

[2] [Online]. Available: https://github.com/danielkrause/Moving-Binaural-SDEL

reason is that the estimation accuracy degrades rapidly for small-sized arrays typically found in practice (e.g., from a few centimeters to a few tens of centimeters) for even fairly small distances from the array center (e.g., a few meters). Reasons for that can include: a) diminishing inter-channel level differences and constant inter-channel time-differences as the source transitions from a spherical wave to a plane wave captured by the array, b) a dropping signal-to-noise ratio (SNR) and a dropping direct-to-reverberant ratio (DRR) with increasing source distance.

Most work in SDE is based on parametric methods and handcrafted features, for example using information about the direct-to-reverberant ratio [45], the room impulse response (RIR) [46] or mean of the signal and binaural cues such as the interaural intensity difference (IID) [47]. Georganti et al. proposed binaural signal magnitude difference standard deviation (BSMD-STD) to train Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) [48]; Vesa used GMMs trained with magnitude squared coherence (MSC) features to include information about the correlation between channels [49]; while Brendel et al. estimated the coherent-to-diffuse power ratio to find the source-microphone distance via GMMs [12]. Most of the aforementioned methods rely on tunable complex algorithms, which tend to lack robustness against differing acoustic conditions.

Research on SDE using DNNs has been very limited so far. Yiwere et al. took an image classification approach based on convolutional recurrent neural networks (CRNNs) trained with log-mel spectrograms to classify between three different distances across three rooms [18]. The models showed good results on data from the same room, but performed significantly worse for recordings coming from a different environment. Sobghdel et al. proposed relation networks to tackle this problem using few-shot learning, showing improvements over traditional convolutional neural networks (CNNs) [19]. Both studies provided tests for a limited set of finite distances in a close area up to 3–4 meters. This research is the first to present results for a wider range of distances up to 14 meters and tested for more than 5 rooms.

C. Binaural Source Distance and DOA Estimation

Joint SDE and DOAE or SDEL is usually defined as *position estimation*, referring to either a continuous position coordinate in 2D/3D space (e.g., expressed in Cartesian x, y, z coordinates) or to pre-defined position classes corresponding to a spatial “binning” of the region of interest. This topic has been widely researched for multiple types of acoustic systems employing typically distributed microphone arrays, including spherical microphones [50], triangular [51], smart loudspeakers [52] or acoustic sensor arrays [53]. However, only a few studies aimed at position estimation from binaural recordings.

In [47], a binaural distance estimation system was proposed with localization-dependent correction improving the overall distance error. Although the study provided directional azimuth information as explicitly given values, without joint modeling, it showed a strong interdependence between distance and localization cues. Ghamdan et al. [54] used GMMs to classify the sound source position defined as azimuth and distance, from a finite set of classes. The study was limited to [−90°, 90°] degrees in the azimuth plane and a set of three distances in the close area up to 2 meters. The method showed good performance when tested in the training room, but significantly worse under unknown conditions. Finally, Yiwere et al. [20] proposed a generic DNN model to perform joint SDEL using ILD and cross-correlation features. The proposed method performed well under both training and testing conditions; however, experiments were limited to a set of three azimuth angles in the frontal plane and four distance positions up to 3 meters. In [55] we studied joint binaural localization and distance estimation, with both tasks defined as coarse classification tasks, showing they can be solved efficiently by a merged model. Here, we extend some ideas from the previous study, presenting the first approach to full regression-based joint binaural source localization and distance estimation.

D. Binaural DOAE With Head Rotation or a Moving Listener

The most prominent source localization cues of the human auditory system are: the interaural level difference, the interaural time difference, and monaural spectral cues mainly due to direction-dependent filtering effects of the pinna [36], [56]. ITD and ILD cues can be ambiguous for source directions across conical regions around the interaural axis, known as *cones of confusion*, and result in inaccurate elevation estimation and front-back confusions. The auditory system takes advantage of spectral cues to solve these ambiguities, but in adverse conditions (i.e., long reverberation times or presence of noise), the information obtained from these three cues is not enough to achieve optimal localization accuracy. To solve this problem, humans rotate their head or move around the space [57], resolving ambiguities by modulating dynamically the localization cues.

A small number of research works focused on head rotation to improve binaural DOAE. Ma et al. proposed a method that considers maximum three simultaneous speakers in two different reverberant rooms, and three azimuth head rotation strategies limited to the ±90° range [58]. A similar approach studied the same head rotation interval with sound sources placed at 3 m from the receiver and five possible azimuth angles [59]. Another study increased the number of rooms to 4, but reduced the rotation interval to ±60° for training and ±30° for testing [60]. An extension of this was made in [61] and [62], facing the localization problem as a 72-azimuth angle classification problem, where the authors considered multiple sound sources in the full 360° azimuth range, but limiting the rotation of the head to ±30°. Finally, in [63], a minimum mean square error-based localization method was proposed using a Behind-The-Ear (BTE) system. Although they evaluated four head rotation speeds (7.5 deg/s, 15 deg/s, 30 deg/s, and 45 deg/s), the source-to-receiver distance was fixed and only one room was studied.

Only a few studies exist that take advantage of a moving receiver to improve the binaural DOAE results. In this case, all research works also benefit from the rotation of the head. Portello et al. presented a particle filtering method for active speaker localization using two microphones mounted on a spherical

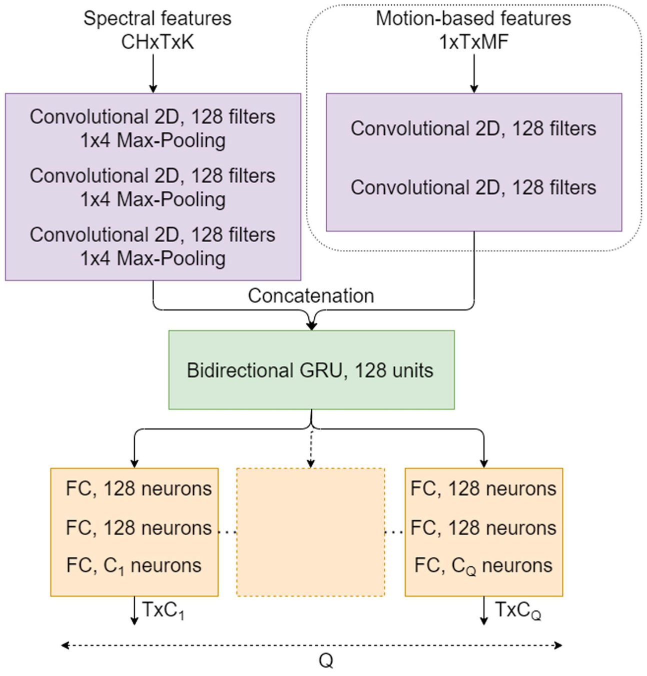


Fig. 1. Architecture of the deep model. The dotted line denotes the optional branch to process motion-based features.

head [64], [65]. Both studies were limited to the azimuth plane, and there is a lack of information about the moving receiver. Also, the recordings consisted of two microphones spaced at a certain distance, which are not truly binaural signals. The same problem was found in [66], [67].

To our knowledge, the only relevant research work that studied how to benefit from the rotation and motion of the head of the listener was proposed in [21]. The authors evaluated 8 different motion strategies in anechoic and reverberant rooms using a particle-filter framework to estimate the azimuth incidence angle and the distance. In more detail, they considered a static receiver with and without head rotation, a moving receiver modeled by a random walk, and different combinations of those. Although they demonstrated the benefit of using motion-based cues in binaural DOAE systems, the investigation was limited to azimuth rotation angles comprised in the range of ±90°. In [44], we first proposed a DNN system that takes advantage of head movement information and explicitly estimates the DOA for an unlimited range of azimuth and elevation angles. In this study, we extend this approach to scenarios with a moving listener.

III. Method

In our experiments, we use a general model architecture scheme, which is depicted in Fig. 1. The proposed block diagram allows for describing models performing DOAE and SDE as separate tasks, as well as in a joint manner. The models take a sequence of spectral features as an input, which is further processed to calculate the desired output. For models utilizing motion-based cues, there is an additional input branch marked with a dotted line, which is responsible for processing the features related to the rotating head and the listener's movement. Depending on which of the tasks the model is supposed to perform, the number of output branches and their respective output layers will vary - separate models performing either SDE or DOAE will have a single output branch, whereas a multi-task joint model will contain two branches. A detailed description of the feature extraction process and the proposed DNN methods is explained in the following sections.

A. Input Feature Extraction

The utilized features may be divided into two main groups: spectral and motion-based. Spectral features are time-frequency representations obtained from the complex spectrogram of the signals, whereas motion-based features directly describe the movement of the listener.

Models take a spectral feature matrix of shape $CH \times T \times K$ as input, where CH is the number of spectral feature channels, T is the length of the temporal sequence and K is the total number of frequency bins. T is set to 250 frames, which represents a 2.5 s slice of the signal. To obtain the spectral features, the complex spectrogram is calculated using a Short-Time Fourier Transform (STFT) with a Hamming window of 20 ms length and 50% overlap between frames. The size of the FFT was 512 samples resulting in $K = 256$ frequency bins up to Nyquist frequency and excluding the DC bin. Then, the **mean magnitude spectrogram** is calculated from both binaural channels to include energy-related information about the signal:

$$|\mathrm{X}|_{mean}[n,k] = \frac{|\mathrm{X}_l[n,k]| + |\mathrm{X}_r[n,k]|}{2}, \qquad (1)$$

where $X[n,k]$ denotes the complex spectrogram value for the n-th frame and k-th frequency bin, whilst l and r stand for the left and right channels, respectively.

Next, we utilize two binaural cues to provide spatial information about the sound sources. Firstly, we extract Interchannel Phase Differences (IPDs) defined as:

$$\mathrm{IPD}[n,k] = \arg(X_l[n,k]) - \arg(X_r[n,k]). \qquad (2)$$

The IPDs are further processed by taking the sines and cosines (**sin&cos**) of the phase values.

$$\mathrm{SI}[n,k] = \sin(\mathrm{IPD}[n,k]), \qquad (3)$$

$$\mathrm{CI}[n,k] = \cos(\mathrm{IPD}[n,k]). \qquad (4)$$

Utilizing the sine and cosine values of phase differences produces a smoother representation compared with raw values and avoids phase wrapping. These features have been firstly proposed for multichannel DNN-based speech separation [68] and further investigated for localization in [44], [69]. On top of that, we utilize the ILDs, which constitute another major binaural cue that becomes important above 1.5 kHz due to the diminishing effect of IPDs related with the physical distance between the ears [70]. The ILDs are defined as follows:

$$\mathrm{ILD}[n,k] = \frac{|\mathrm{X}_l[n,k]|}{|\mathrm{X}_r[n,k]|}. \qquad (5)$$

The mean magnitude spectrogram, sines and cosines of IPDs, and ILDs sum up to a total of $CH = 4$ spectral feature channels.

TABLE I
INPUT PARAMETERS FOR THE MODELS

Motion-based features	**CH**	**T**	**K**	**MF**
Rotation	4	250	256	4
Rotation & Translation	4	250	256	7

For scenarios including movements of the listener, we propose two novel feature types describing the motion: the rotation of the head with respect to the initial head orientation (hereafter referred as **rotation features**), and the Cartesian position of the listener with respect to their initial position (referred to as **translation features**). The motion-based cues are passed to a separate input DNN branch as a single-channel matrix of shape $1 \times T \times MF$, where MF denotes the total number of features. The rotation features are expressed by means of the four quaternion values ($a, b, c, d \in \mathbb{R}$), where a quaternion is a hypercomplex number defined as $\mathbf{q} = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ modeling rotational states or changes. Compared to simpler representations, such as Euler angles, quaternions avoid phase wrapping and rotational ambiguities [71]. Head rotations are relative to an unrotated head leveled on the horizontal plane and with look direction towards the positive x-axis.

Translation features are defined as the difference between the initial and current position of the listener:

$$\mathrm{P}_{\text{translation}}[n] = P_{x,y,z}[n] - P_{x,y,z}[0], \tag{6}$$

where $P_{x,y,z}[n]$ denotes the position of the listener in the n-th frame, expressed in the Cartesian x, y, z coordinates. In total, the number of motion-based features equals to $MF = 4$ for rotation features and $MF = 7$ when using rotation and translation features jointly. All the model input parameters are summarized in Table I.

B. Model Architecture

The basic DNN architecture, depicted in Fig. 1, is based on a model from our previous study [44], which was proven to perform efficiently for an DOAE system utilizing information about head rotation. Here, we introduce a few changes to tackle the new problems under investigation.

The spectral features are initially processed by three convolutional blocks. Each block consists of a 2D convolutional layer containing 128 3 × 3 filters, followed by batch normalization (BN) and a 1 × 4 max-pooling operation across the frequency dimension. The time-frequency filters obtain local frequency dependencies, while also exploiting frame-to-frame changes in the time dimension, allowing for representing changes due to head orientation and listener position. The max-pooling operation is used to compress frequency-specific information and reduce the dimension size for further processing in the following layers. For scenarios with a rotating head and a moving listener, we introduce a special second branch to process motion-based features. The branch consists of two convolutional layers, each consisting of 128 3 × 3 filters and BN. The output of the second branch is then concatenated with the output of the first one along the last dimension.

Next, the processed feature maps are passed to a single bidirectional Gated Recurrent Unit (GRU) layer, consisting of 128 layers and tanh activations. The use of a recurrent layer allows for stronger temporal modeling of the signal, which again allows for exploiting important inter-frame information, emphasizing the extracted motion-based cues. Finally, the unit outputs are further passed to Q parallel output blocks, each consisting of three fully-connected (FC) layers. In each of the output blocks, the first two layers contain $G = 128$ neurons and a a rectified linear unit (ReLU) activation function. The final layer outputs C_q values for each frame in the sequence, resulting in a $T \times C_q$ output matrix. The values of MF, Q and $C_q,\ q = 1, .., Q$, are dependent on the performed task and the investigated scenario. MF stands for the total number of utilized motion-based cues, Q is the number of individual performed tasks and C_q describes the number of outputs values in each q-th output branch. The details about these parameters are further explained in Section III-C.

In most scenarios, we treat both DOAE and SDE as regression tasks, in which the corresponding outputs are trained using the Mean Squared Error (MSE) loss. In some cases, we reduce distance estimation to a binary classification task (see Section III-C), for which the binary cross-entropy (BCE) loss is utilized. All models are trained for a maximum of 200 epochs using Adam optimizer and early stopping after 30 epochs of no improvement in training. The networks are implemented using the PyTorch library [72].

C. Model Outputs

In this study, we investigate three different tasks - separate SDE and DOAE, as well as joint SDEL. In general, we treat all tasks as regression problems, where both the direction of arrival and the distance can be mutually defined by the explicit position of the sound source with respect to the receiver. In this chapter, we describe the definitions of the model outputs to tackle the problems under investigation.

When the receiver is moving, the origin is moving along with it. Given the position defined as a vector in Cartesian coordinates $P_{x,y,z}$, the source-to-receiver distance can be obtained by calculating the norm of the vector:

$$\mathrm{d} = \sqrt{P_x^2 + P_y^2 + P_z^2}, \tag{7}$$

whereas the DOA is defined as the normalized position vector:

$$\mathrm{DOA}_{x,y,z} = \frac{P_{x,y,z}}{\sqrt{P_x^2 + P_y^2 + P_z^2}}. \tag{8}$$

In some cases, we provide additional experiments with a simplified version of source distance estimation, compressing the task to a coarse classification problem, where each sound source is classified as “near” (up to 5 m) or “far” (above 5 m). They serve as additional comparison points to the regression-based SDE, since a coarse classification approach has been proven to be very effective in one of our previous studies [55], whilst accurate distance estimation from binaural recordings is a more challenging task.

TABLE II
OUTPUT PARAMETERS FOR DIFFERENT MODELS

Task	Q	C_q	Output activation
Joint SDEL:			
Single task SDEL	1	[3]	linear
Multi task DOAE+SDE (regression)	2	[3, 1]	[tanh, ReLU]
Multi task DOAE+SDE (binary)	2	[3, 1]	[tanh, sigmoid]
Separate DOAE/SDE:			
SDE (regression)	1	[1]	ReLU
SDE (binary)	1	[1]	sigmoid
DOAE	1	[3]	tanh

Depending on the performed task and utilized method, the model outputs have varying parameters Q and C_q as shown in Fig. 1. Below, we describe the problem definition for each of the investigated tasks and the details about the model outputs, which are also summarized in Table II.

1) Separate DOAE & SDE Models: To benchmark the performance of DNNs performing the joint task of SDEL, experiments for separate SDE and DOAE models are provided. These include:

- *SDE:* to perform distance estimation alone, we utilize a single output branch ($Q = 1$) with a ReLU output (for the regression approach) or a sigmoid output (for the binary classification approach). The model outputs a single distance value, hence $C_q = 1$.
- *DOAE:* separate DOA estimation is obtained by utilizing a single output branch ($Q = 1$) with $C_q = 3$ output neurons, followed by a tanh activation function to obtain the normalized DOA$_{x,y,z}$ vector.

2) Joint SDEL Models: Estimation of the sound source distance and direction, is studied in a single task and a multi task setting, following these two representations:

- *Single task:* In this method a single output branch ($Q = 1$) attempts to model directly the exact position of the sound source $P_{x,y,z}$. The model output contains $C_q = 3$ linear neurons, altogether resembling the x, y, z coordinates of the source.
- *Multi task:* This approach attempts to model both SDE and DOAE tasks by a single DNN with two separate output branches ($Q = 2$), one for each task. Here, the DOA output models the normalized vector DOA$_{x,y,z}$ via a tanh activation function, whereas the SDE branch explicitly outputs the source-to-receiver distance d using a ReLU activation. Moreover, for additional comparison, we investigate an alternative model with a binary output for SDE, classifying each sound source as "near" or "far" as described earlier. In this case, the sigmoid activation function is used. In both cases we obtain three DOA values and a single distance value from separate output branches, resulting in $C_q = [3, 1]$. We calculate the final training loss as a linear combination of both output losses with weights offsetting the uneven initial loss values for both tasks. When training a model with regression SDE, we apply an DOAE/SDE ratio of 3:1, whereas for binary SDE the ratio is 1:11, intended to provide a similar training loss for both tasks during the starting epochs.

IV. Experimental Setup

A. Experimental Scenarios

To investigate the joint performance of DOAE and SDE and the effect of motion cues on the performance of both tasks, we split our experiments into two stages. Firstly, we perform experiments with separate models for each task. Next, we investigate both methods of performing the joint SDEL task and study their performance as compared to the separate models. For both stages, the models are trained and tested for three basic scenarios allowing to investigate the effect of motion-based cues:

- *Static:* Both the source and the receiver are static, hence the models are trained with spectral features only ($MF = 0$).
- *Rotation:* here, we consider a static receiver with a rotating head, defined by a full azimuth-range rotation range. Rotation features are used to exploit the rotation information explicitly ($MF = 4$). The rotation scenario without motion-based cues has been investigated in our previous study [44] and the use of features has been proven to be more effective.
- *Walking:* Finally, we investigate a scenario in which the listener is moving freely inside the room, hence the distance and DOA are dynamically changing from frame to frame. To investigate the extent to which motion-based cues are affecting the model's performance, we perform experiments without additional features ($MF = 0$), with rotation features ($MF = 4$), and using a joint set of rotation and translation features ($MF = 7$).

B. Dataset

The dataset used to evaluate the models comprises three different scenarios: a static receiver (later denoted as "Static"), a static receiver with a rotating head (referred as "Rotation") and a receiver moving in space ("Walking"). For all of them, a single speech sound source is considered. All audio files contain speech signals only, without silence between speakers, to avoid the necessity for an additional detection system. As the source signal, male and female anechoic speech audios of 10 s length from the TIMIT dataset [73] were reverberated in each simulation.

Since the SDEL models in this study are trained and evaluated solely through simulations, the simulations need to be both fast, generating a large number of diverse reverberant scenarios, and accurate, delivering the appropriate spatial and motion features in the simulated audio. More specifically, we target the following simulation requirements: a) realistic acoustic conditions with frequency-dependent absorption, shown to improve generalization to real test conditions [74], b) directional receivers with measured directional responses, such as Head Related Transfer Functions (HRTFs) for binaural simulations, c) moving receivers at various speeds, d) rotating directional receivers at various speeds, and f) the ability to generate thousands of such scenarios with signal output of several seconds in reasonable time. The last requirement is met currently only by simulators using the image-source method (ISM) for shoebox geometries. Due to the lack of publicly available shoebox ISM simulators (e.g. [75],

TABLE III
PARAMETER VALUES FOR DATA GENERATION

Parameter	Value
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 elevation angle	[-35° 35°]
Azimuth angle	[-180° 180°]
Head rotation speed	10, 20, 30, 40, 50 deg/s
Min. walking segment length	2.0 m
Walking speed	0.8, 1.0, 1.2, 1.4, 1.6 m/s

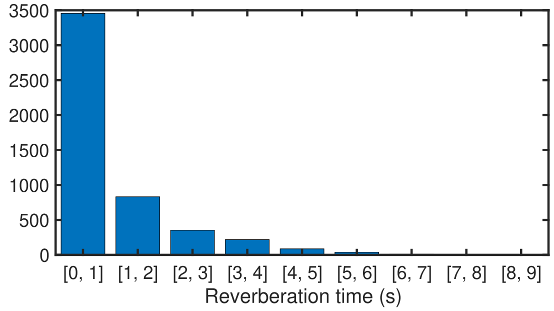


Fig. 2. Distribution of the reverberation time in simulations. The number of rooms with a RT comprised between 5 s and 8 s is 56.

[76]) that fulfilled these requirements, we repurposed a custom one developed earlier by the authors for spatial audio coding studies [77], extended to handle dynamic binaural rendering scenarios.

1) Room and Reverberation Parameters: Room dimensions, source, and receiver positions were randomized in each simulated scenario according to the ranges and constraints shown in Table III. Since the same randomization procedure was used in the earlier rotation-only study of [44], the same 2500 random room configurations generated in that study were re-used here for an equitable comparison with those previous results. Regarding acoustic conditions, instead of targeting a uniform distribution of reverberation times and estimating absorption coefficient values in order to achieve those, as is commonly done, we opted for a more realistic simulation setting. Absorption profiles from tables of typical acoustic materials [78] for wall, ceiling, and floor surfaces were randomly assigned to individual surfaces for each simulation scenario. In this manner we get a more realistic distribution of reverberation times, shown in Fig. 2, and we avoid unnatural cases, e.g., very long reverberation times in very small rooms.

2) Random Head Rotation and Random Walk Simulation: The random receiver positions indicated the listener position for static and rotation-only scenarios, or the initial position for walking scenarios. The sources were static across all three scenarios. Regarding the rotation-only scenario, 5 angular speeds of yaw rotation were simulated as shown in Table III, uniformly distributed across the 2500 simulated rooms (500 per rotation speed). Regarding the walking scenario, a random walk trajectory composed from line segments was drawn in each simulated room. A random horizontal walking direction was computed to form the first walking segment; a new line segment at a new random direction was added every time the trajectory encountered a wall.[3] The orientation of the head corresponded to the direction of the walker. In addition, the head orientation was smoothly rotated each time the walking direction changed. Based on studies on average walking speeds [79], 5 different velocities were generated: 0.8, 1.0, 1.2, 1.4, 1.6 m/s, uniformly distributed across rooms (500 per walking speed). Finally, the distribution of the source-to-receiver distances for all the trajectory points (100 positions) is shown in Fig. 3, and their average distance per trajectory is presented jointly with the source-to-receiver distance of the static scenario in Fig. 4.

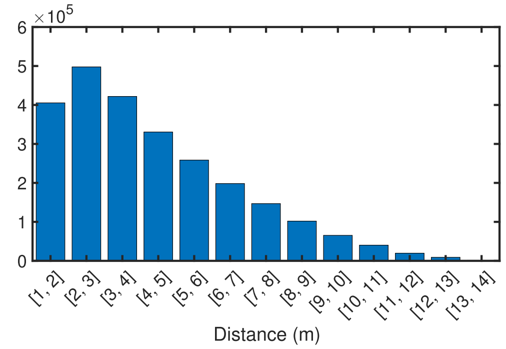


Fig. 3. Distribution of the source-to-receiver distances for all the positions of the simulated trajectories.

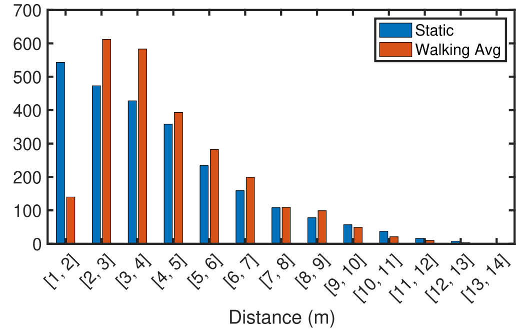


Fig. 4. Distribution of the source-to-receiver average distances for all the positions of the simulated trajectories and the source-to-receiver distances of the static scenario.

C. Evaluation Metrics

The investigated models are evaluated using separate metrics for SDE and DOA estimation. For DOA estimation, we use the standard DOA error, which is defined as follows:

$$E_{\text{DOA}} = \frac{1}{N} \sum_{n=0}^{N-1} \sigma(\mathbf{x}_R[n], \mathbf{x}_E[n]), \tag{9}$$

[3] The complete details of the listener motion simulation are available in the provided supplementary material which presents the entire dataset simulation.

TABLE IV
RESULTS OBTAINED FOR SEPARATE MODELS FOR VARIOUS SCENARIOS

Scenario	Motion cues	DOAE model	SDE model (regression)		SDE model (binary)	DOA ref. [20]	SDE ref. [18]
		$E_{\text{DOA}}[°]$	$E_{\text{Dist}}[m]$	$RelAcc[\%]$	$Acc[\%]$	$Acc[\%]$	$Acc[\%]$
Static	None	17.83 ± 1.24	1.51 ± 0.06	16.93 ± 1.22	75.78 ± 0.67	24.38 ± 1.16	22.91 ± 0.97
Rotation	Rotation	13.14 ± 1.00	1.34 ± 0.27	18.77 ± 3.23	74.38 ± 1.27	-	-
Walking	None	2.06 ± 0.13	1.50 ± 0.05	17.75 ± 1.10	71.42 ± 1.30	8.80 ± 0.60	27.73 ± 0.36
	Rot.	1.67 ± 0.22	1.46 ± 0.06	18.54 ± 0.48	71.71 ± 1.26	-	-
	Rot.&Trans.	1.66 ± 0.23	1.46 ± 0.08	19.14 ± 1.14	73.26 ± 0.68	-	-

The last two columns contain results obtained for reference models (as described in Section V-A).

where n denotes the frame index over which the DOAs are estimated by the model. The error is averaged over all N frames, since in our experiments the sources are active throughout the whole file. σ stands for the angular distance between two vectors:

$$\sigma = \arccos\left(\frac{\mathbf{x}_E^{\mathrm{T}}\mathbf{x}_R}{||\mathbf{x}_E|| \cdot ||\mathbf{x}_R||}\right) \tag{10}$$

where $\mathbf{x}_E$ and $\mathbf{x}_R$ stand for the estimated and reference DOAs.

For source distance estimation, we use two metrics to analyze the model's performance. Firstly, we compute the absolute distance error:

$$E_{\text{Dist}} = \frac{1}{N}\sum_{n=0}^{N-1} |\mathbf{d}_R[n] - \mathbf{d}_E[n]|, \tag{11}$$

Here, $\mathbf{d}_R$ and $\mathbf{d}_E$ denote the reference and estimated distance, respectively. As an additional measure, we calculate the relative accuracy of the model. The relative accuracy measures the number of frames in which the estimated distance falls in an accepted threshold area, deviating from the target distance. This measure might be more appropriate for practical applications, where the performance is expected to be more accurate for closer distances than for further ones. Here, we use a threshold of 10% with respect to the reference distance:

$$RelAcc = \frac{100\%}{N}\sum_{n=0}^{N-1} \mathbb{1}(|\mathbf{d}_R[n] - \mathbf{d}_E[n]| < 0.1\mathbf{d}_R[n]). \tag{12}$$

In some scenarios, we reduce the SDE problem to a binary classification problem, in which sources appearing closer than 5 m are considered "near" (positive one), whilst sources further than 5 m are considered "far" (negative zero). In these cases, we use the standard accuracy measure:

$$Acc = \frac{100\%}{N}\sum_{n=0}^{N-1} \mathbb{1}(\mathbf{d}_{bin,R}[n] == \mathbf{d}_{bin,E}[n]), \tag{13}$$

where $\mathbf{d}_{bin,R}$ and $\mathbf{d}_{bin,E}$ stand for the reference and estimated binary distance labels.

V. Results and Discussion

A. Results for Separate Models

Results for all evaluated scenarios performed by separate DOAE and SDE models are presented in Table IV. The presented numbers are average and standard deviation values, calculated over a 5-fold cross-validation split. Generally, similar differences between scenarios can be seen for both tasks, with a few notable exceptions. As a reference, we use two models from the literature, one for each task. As explained in Section II, no study has been found that would investigate a regression-based model for data with a moving receiver and a wide range of azimuth and elevation angles. Therefore, we selected two studies that were as close to this one as possible, whilst accepting major differences. For DOA estimation, we use the classification-based method presented in [20] with two important changes: a) we change the last layer of the model to a framewise output to allow for evaluating the moving scenario; b) the output layer is adjusted to the full range of azimuth angles $[0°, 360°]$ in the utilized data, resulting in 12 possible classes spaced every 30°. For SDE, we utilize the model proposed in [18]. Again, we adjust the DNN to a framewise output and a wider range of distances. Here, the final set consists of 6 classes, corresponding to the following ranges of distances: $[1, 2), [2, 3), [3, 4), [4, 5), [5, 7)$ and [7,14].

For DOA estimation, the highest error of 17.83° is observed for the static scenario. For the dataset including head rotation, the error drops by over 4° to a level of 13.14°. Both results are in line with our previous study [44], which showed significant improvements from using rotation features. The positive effect of motion-based cues in this case comes from the fact that, by utilizing information about rotation, the model is able to use a wider range of azimuth angles to avoid the cone-of-confusion effect. Still, much larger improvements can be seen for a scenario with a walking listener. Even for training without the use of motion-based cues, the DOA error decreases to a value of 2.06°, which is an over 8-fold improvement over the static scenario. The use of rotation features lowers the error further to 1.67°, whereas the addition of translation features does not seem to significantly change the results.

The far-reaching improvements from the walking scenario alone might be simultaneously linked to two different factors. On the one hand, the walking listener might cause improved performance by entering an area from which localization becomes easier, which in result can lead to a correct DOA estimation. On the other hand, by exploiting a dynamic receiver trajectory, the model receives information about the sound source from a varying range of angles and distances. This information might be beneficial at the temporal modeling stage in the recurrent layers, where the inter-frame context might improve the results for less advantageous receiver positions.

Analogously to the localization task, the worst performance for sound distance estimation is observed for the static scenario, with a distance error of 1.51 m and relative accuracy of 16.93%. This is a considerably high error, showing that achieving good source distance estimation performance for a wide range of distances and angles is a highly challenging task for binaural recordings. As expected, using motion-based features for a rotating head improves the error significantly by 0.17 m and 1.84 p.p. Similarly to DOAE, rotating the head benefits distance estimation by yielding more angles from the receiver, hence providing better distance dependent features such as echo patterns or direct-to-reverberant ratio cues. Contrary to source localization though, the walking scenario does not appear to decrease the error when compared with the rotation scenario. For a model without the use of motion-based cues, the distance error goes back up to 1.50 m. Rotation and translation features reduce this value to 1.46 m, which however is still a higher number than for the rotating head case. The noticeable difference might be caused by the fact that for the walking scenario the models are supposed to estimate dynamically changing distances, which is a harder task compared with a static distance throughout the whole clip. In this case, the motion-based cues partially offset the difficulty by providing explicit information about the movement. However, a static receiver with a changing angle seems to perform better, since a constant distance can be more easily exploited by covering a wide azimuth range.

In the following column, we present the additional model which performs SDE in a binary classification manner. Here, we observe that despite the high distance error for regression-based estimation, satisfactory results can be achieved for a simplified, coarse classification task. For a static receiver, we obtain an accuracy of 74.38% and 75.38% for a scenario with and without a rotating head, respectively. Interestingly, results for a walking scenario, whilst still good, are noticeably worse than for a static receiver, showing an accuracy of around 71%, with an additional increase to over 73% when using translation features. The observed drop in performance might be caused by the nature of coarse classification, in which the label of each sound source does not change throughout the clip as long as the source does not cross the border of 5 m, causing a possible mismatch with the tracking of the dynamic listener's position.

The last two columns present results obtained for the reference models. The DOA model achieves an accuracy of 24.38% for the static scenario, which compared with the accuracy of over 99% [20] reported in the original study shows the the model does not generalize well for a large numbers of rooms and a wider set of azimuth angles. For a moving scenario the accuracy drops to 8.80%, further indicating the method's low robustness to dynamically changing conditions. The SDE model shows an accuracy of 22.91%, which similarly to the DOA model is a significantly lower score compared with the original paper [18], in which the method achieved results between 44.82% and 98.47% depending on the testing conditions. Interestingly, distance estimation improves slightly for the moving scenario, increasing the performance to 27.73%. That might be due to the use of temporal information via RNNs, which analogously to our proposed method, takes advantage of the inter-frame dependencies to model the receiver's movement. While the reference results do not compare with the proposed method because of largely different approaches and evaluation methods, the reported scores show that binaural source localization and distance estimation for a wide range of angles and distances is a challenging task that needs further investigation.

B. Results for Joint Models

Next, we perform the same experiments with the joint SDEL task models, with the final results presented in Table V. Most noticeably, the performance of both tasks seems to decline in most cases when compared with the separate models, an expected effect resulting from modeling two tasks with a single DNN of a similar complexity. This might be partially offset by further fine-tuning of the model parameters, which is out of the scope of this study.

For all joint model types, the differences between the investigated scenarios are generally similar to the observations made in the previous stage of the study - a static scenario shows worst performance, with a notable boost when using a rotating head and an even higher improvement when switching to a walking listener. Most importantly, while translation features seem to have a negligible effect on the general performance of separate DOAE and SDE models, they appear to have a higher influence on the SDEL task. For both the single task and the multi task approach, adding rotation and translation features increases the DOA error when compared with rotation features alone - from 6.91° to 8.83° for the single task method, and from 6.48° to 10.05° for the multi task model.

On the contrary, translation features seem to improve the distance error for both cases - from 1.90 m to 1.79 m for the first model, and from 1.52 m to 1.45 m for the second. In both cases, the total distance error improves over the model utilizing a rotating head, which is another distinction from the experiments performed with separate models. The positive impact of translation features is further demonstrated for the multi-task model with binary SDE, where the accuracy improves by 6.13 p.p. over the rotation features alone. On top of that, the distance estimation performance for the multi-task approach with regression-based SDE is on par with the separate SDE model evaluated for the same scenario (1.45 m vs. 1.46 m). These results show that when opting for joint SDEL DNN training, both motion-based features provide valuable information about the receiver's position, allowing especially for a more accurate prediction of the source-to-receiver distance. However, it is not clear why translation features might in some cases affect the DOAE performance in a negative way.

Comparing the three proposed methods for performing joint SDEL, the multi-task approach with binary SDE seems to be the worst performing one for all scenarios. While in most cases the observed accuracy is comparable with the separate model (with only the walking listener with and without rotation features having considerably lower performance), the DOA error remains the highest for all performed experiments. The low performance of the SSL part might be again caused by the different nature of coarse classification, in which binary SDE labels do not relate

TABLE V
RESULTS OBTAINED FOR JOINT SDEL MODELS - USING A SINGLE TASK APPROACH AND A MULTI-TASK APPROACH IN TWO VERSIONS - WITH A REGRESSION-BASED SDE BRANCH (REGRESSION SDE) AND A BINARY SDE CLASSIFIER (BINARY SDE)

Scenario	Motion cues	Single task			Multi task (regression SDE)			Multi task (binary SDE)	
		$E_{\mathrm{DOA}}[°]$	$E_{\mathrm{Dist}}[m]$	$RelAcc[\%]$	$E_{\mathrm{DOA}}[°]$	$E_{\mathrm{Dist}}[m]$	$RelAcc[\%]$	$E_{\mathrm{DOA}}[°]$	$Acc[\%]$
Static	None	16.41 ± 0.57	2.75 ± 0.07	5.68 ± 0.81	15.31 ± 1.29	3.01 ± 0.05	2.34 ± 0.47	18.70 ± 3.17	75.38 ± 1.48
Rotation	Rot.	15.02 ± 1.13	1.83 ± 0.03	12.55 ± 0.43	15.17 ± 0.37	1.67 ± 0.14	14.25 ± 1.55	16.17 ± 2.86	72.67 ± 1.10
Walking	None	12.09 ± 4.06	3.08 ± 0.08	3.95 ± 0.26	7.94 ± 0.19	3.36 ± 0.07	0.92 ± 0.06	88.74 ± 2.54	69.64 ± 3.12
	Rot.	6.91 ± 0.63	1.90 ± 0.04	13.19 ± 0.60	6.48 ± 0.33	1.52 ± 0.05	17.01 ± 0.55	12.12 ± 2.60	66.79 ± 2.31
	Rot.&Trans.	8.83 ± 1.17	1.79 ± 0.06	14.26 ± 0.06	10.05 ± 2.18	1.45 ± 0.05	17.65 ± 0.21	10.14 ± 1.69	72.92 ± 0.77

to direction-of-arrival vectors as directly as explicit distance values. We also note, that in this case both tasks are trained with different loss types, which might increase the inter-task mismatch even further.

Amongst the fully regression-based models, the multi-task approach seems to outperform the single task method in most cases. While the single task model achieves a lower DOA error of 8.83° for a walking listener with a full set of motion-based cues, the multi-task shows better DOAE performance in all other scenarios. Most notably, for the "Walking" case with rotation features, the model achieves a DOA error of 6.48°, the lowest value amongst all SDEL models, and a very good general performance on the overall joint task. Analogously for the SDE part, the single task approach achieves a slightly lower distance error for a static scenario and a walking scenario without additional features. However, in both these cases the general error is very high, almost double the value of other models. In all remaining cases, the multi-task approach significantly outperforms the single task method, particularly for the combination of rotation and translation features. In this case, the multi-task model achieves distance error lower by 0.34 m, which is the best outcome amongst all SDEL experiments. These results suggest that the multi-task approach is generally more efficient at tackling the tasks of DOAE and SDE jointly, which is also in line with the conclusions from our previous study on joint proximity and direction classification [55]. While the single task method offers a physically intuitive definition of the problem, where the direction-of-arrival and distance are combined into a single x, y, z position, the multi-task model allows for separate optimization of two distinct output branches, whilst still exploiting shared inter-task information in the first layers of the DNN.

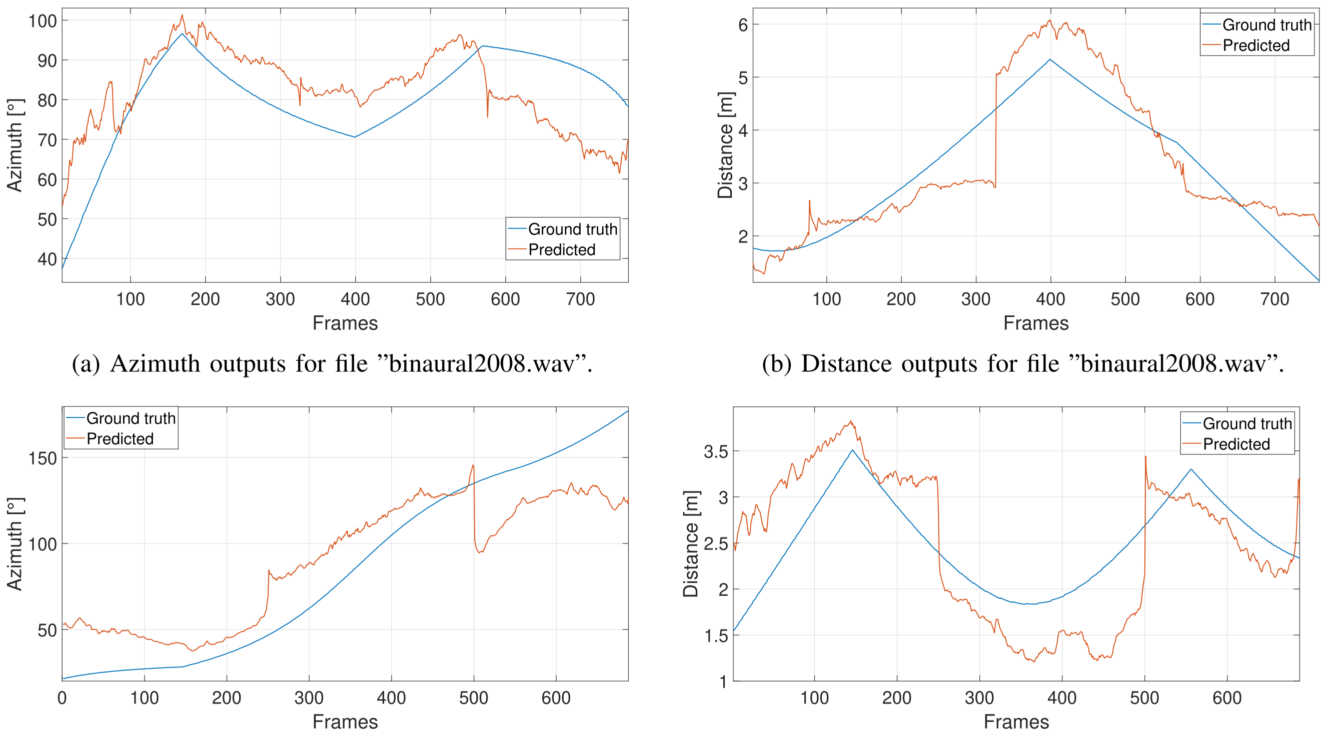
In Fig. 5 we show two examples of differences between ground truth values and model predictions in a moving scenario. The results are shown for calculated azimuth and distance outputs. As can be seen in Fig. 5(a), the azimuth predictions more or less follow the ground truth values. Most notable differences are shown from the 570th frame onwards, where the slope for model predictions is much steeper than for ground truth values. Larger differences can be seen for distance outputs, as in Fig. 5(b). Even though the model predictions seem to follow a similar pattern of increasing and decreasing values, the shape of the slopes are largely different from the reference curve. The most striking difference is a sudden jump of 2 m in the distance value, located around the 310th frame. These jumps in values are related to transitions between separate sequences modeled by the DNN. The model analyses each sequence separately, without temporal knowledge about the neighbouring time spans, potentially leading to inconsistent predictions between different sequences. This effect can be sometimes seen also for azimuth values, as presented in 5(c). Even though the predictions follow the ground truth values, two abrupt value changes can be seen around the 250th and 500th frame. In both cases the jumps create a larger deviation from the reference curve, since frames following the first value in a sequence are biased by the former. The outcome of this error is even more visible in Fig. 5(d), in which we observe large jumps in distance values of over 1 m for the same frames. This problem might be generally mitigated by modeling longer time sequences inside the DNN model, as well as by the utilization of smoothing techniques in the post-processing phase. A technique allowing for transferring temporal information between sequences might help to solve this issue in a more generalized way.

C. Effects of Acoustic Conditions

Fig. 6 depicts the DOA and distance errors with respect to walking speed (for the "Walking" scenario), source-to-receiver distance, direct-to-reverberant ratio and room reverberation time. The presented results are calculated for models performing SDE and SDEL separately and for the following scenarios: static, head rotation, and walking listener with rotation features and with a full set of motion-based cues.

The relation between the distance error and walking speed for a moving receiver is displayed in Fig. 6(a). The error decreases for the speed of 1 m/s as compared with 0.8 m/s by 0.04 m for both analyzed models, however for higher speeds the performance seems to steadily deteriorate. One exception occurs for the highest speed of 1.6 m/s when utilizing all motion-based cues, which in contrast to other cases, seem to outperform the use of rotation features alone. As for DOA estimation (Fig. 6(b)), the relation seems to be less obvious. For a model trained with rotation features only, the error stays at the same level of 2° for all walking speeds, with an exceptional drop to 1° for 1 m/s. When using both rotation and translation features, the performance generally increases from 3° for 0.8 m/s to 4° for 1.6 m/s. On the whole, the walking speed of 1 m/s shows best results for most investigated cases.

Subfigures 6(c) and 6(d) show the distance and DOA errors depending on the source-to-receiver distance in the file. As



(a) Azimuth outputs for file ”binaural2008.wav”.

(b) Distance outputs for file ”binaural2008.wav”.

(c) Azimuth outputs for file ”binaural1452.wav”.

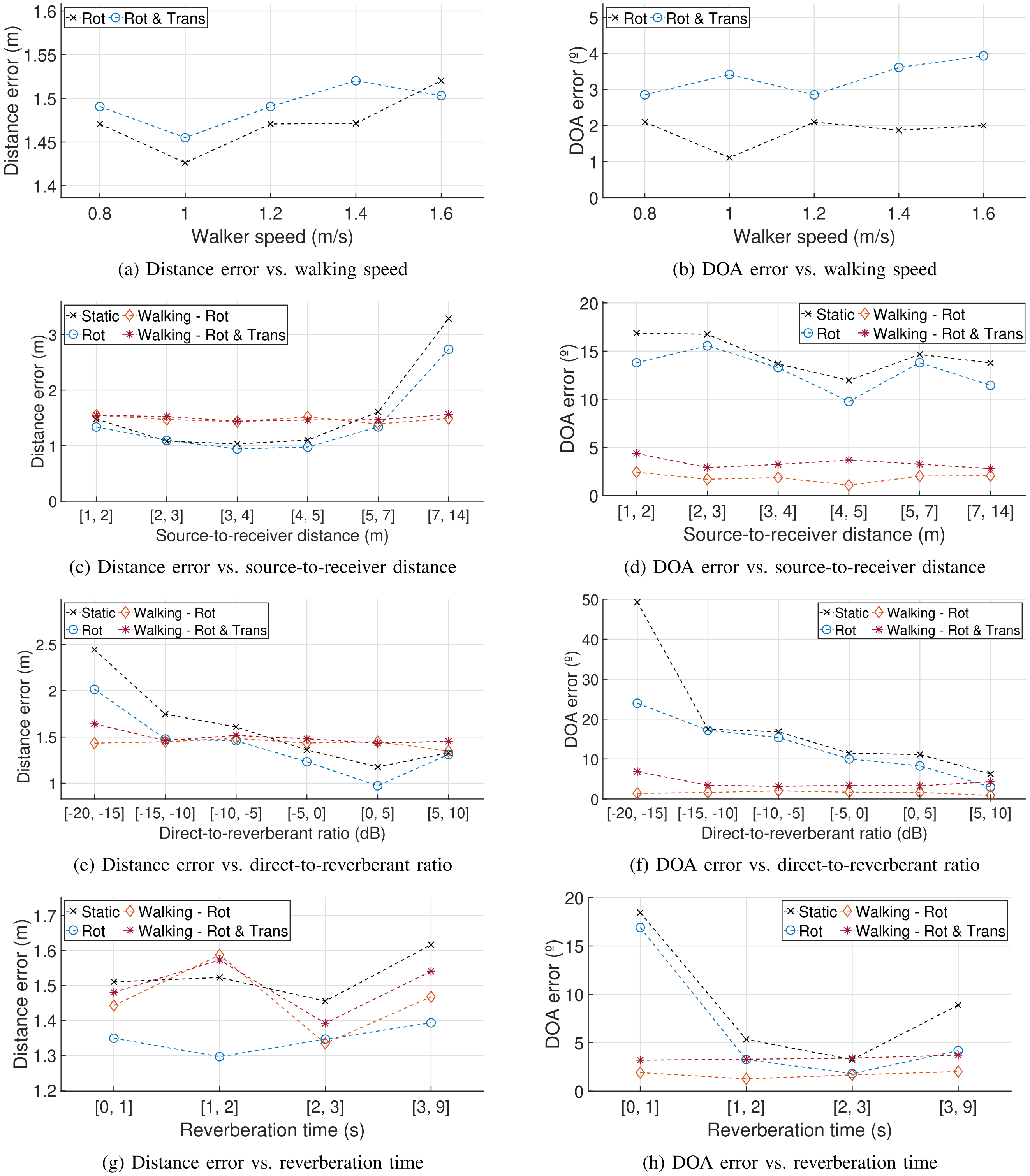
(d) Distance outputs for file ”binaural1452.wav”.

Fig. 5. Plots depicting exemplary differences between ground truth values and model predictions for a joint multi-task SDEL model in a moving scenario. Comparisons are shown for azimuth angles and distances.

for distance estimation, we observe a very similar pattern for the static and head rotation scenario, in which the error starts at around 1.4 m for the near range of 1-2 m and then drops to a level of around 1 m for 2-5 m. For a source-to-receiver distance above 5 m, a steep increase can be observed, with the error going up to 2.73 m and 3.29 m for the rotation and static scenario, respectively. This drastic drop of performance can be explained by a much lower direct-to-reverberant ratio for longer distances, which can significantly hinder distance estimation, which hugely relies on energy-based cues. In contrast to both static scenarios, the walking scenario presents an almost straight line at around 1.5 m for the whole distance range. This shows that by utilizing a dynamic receiver, the model is able to average the information from different positions, effectively flattening its dependence on the source-to-receiver distance. As for the DOAE model, an analogous pattern for both static scenarios can be again observed, with a consequent drop of DOA error for distances in the range of 2-5 m and a slight increase above 5 m. Similarly to SDE, for the walking receiver we observe a fairly straight line, with an exception of a minor drop of error above 2 m, which further proves the aforementioned conclusions. We also note that for all distance ranges the DOA error stays below 5°, which is generally a very good outcome.

In subfigures 6(e) and 6(f) the errors’ dependence on the direct-to-reverberant ratio is shown. As for SDE, there is a consequent decline of distance error with the increase of DRR from -20 dB up to 5 dB for the “Static” and “Rot” scenarios. The steady improvement of the models’ performance is linked with the source energy being more pronounced in the signal, making distance estimation less demanding. Interestingly, for both scenarios we observe a slight increase of distance error for the range of [5, 10] dB. This effect might occur due to the loss of some reverberant information, which might be helpful in certain cases. Contrary to that, both models trained for a scenario with a walking listener show a nearly flat characteristic, which is beneficial when dealing with low values of DRR. However, in the range of $[-5, 5]$ dB the static models achieve better results. DOAE performance is following an analogous pattern for static scenarios, where the DOA error is consistently dropping, whilst the DRR is increasing. Worth noting is the difference of 25.30° between the “Static” and “Rot” scenarios in the range of $[-20, -15]$ dB, showing that especially for low DRRs the effects of a rotating head help to capture more information from different angles, significantly improving the estimation performance. Yet again, models trained for the “Walking” scenario present mostly equal performance across all DRR ranges, largely outperfoming their static counterparts in all cases.

Finally, the effect of room reverberation is shown in subfigures 6(g) and 6(h). For both static scenarios and SDE, the curves depict a relatively flat characteristic for the range up to 3 s with a notable increase of distance error in the range between 3 s and 9 s, in which the reverb might have a higher impact on the results by lowering the DRR. Results for the walking scenario seem to follow a more complicated pattern - for a model utilizing only rotation features, the highest distance error of 1.59 m is observed for the range of 1-2 s, whereas the best performance is shown



(a) Distance error vs. walking speed

(b) DOA error vs. walking speed

(c) Distance error vs. source-to-receiver distance

(d) DOA error vs. source-to-receiver distance

(e) Distance error vs. direct-to-reverberant ratio

(f) DOA error vs. direct-to-reverberant ratio

(g) Distance error vs. reverberation time

(h) DOA error vs. reverberation time

Fig. 6. Plots analysing the DOA and distance errors with respect to source-to-receiver distance, reverberation time, direct-to-reverberant ratio and walking speed. The presented results are shown for tasks performed by separate models.

for RTs between 2 s and 3 s, with an error of 1.33 m. Although, similarly to the static cases, the performance decreases again for higher RT values, the performance seems to be still better than for the middle range of 1-2 s. The model trained with all motion-based features follows an identical pattern, with a slightly higher distance error for most cases. The curves show a particularly different dependence for the DOA error, which reaches the highest values for low RT values in the range 0-1 s when analyzed in a static scenario. For a model trained with rotation features (which analogously to other cases, performs overly better than a completely static model), the error drops by a large number of 15.16° when moving to the 1-2 s range, achieving the lowest error of 1.81° for RT between 2 and 3 seconds. Such a massive difference shows the importance of reverberant information for correct binaural source localization, in which low RT disables the use of wall, floor and ceiling reflections to improve DOA estimation precision. However, we note a decrease of performance for the range of 3-9 s, in which the amount of reverberant might dominate the direct sound to a higher extent. The decrease is less severe for the model utilizing rotation features, showing that a rotating head might partially offset this issue. Interestingly, for both models that utilize a walking listener, no significant differences in DOAE performance can be observed. Similarly to the previous plots, regardless of the set of features utilized, the DOA error seems to stay at a relatively stable level of 1-3° for all RT values. It is worth noting that similar values are obtained for a static scenario in the RT range of 1 s to 3 s. However, for other reverberation time values we observe much higher DOA errors due to a lack of reverberation information or a very low DRR. The use of a moving receiver effectively counters this problem by showing equally good performance for all RT values.

D. Effects of Background Noise

The present work focuses on studying motion-related effects on reverberant source localization with simulations in order to assess their importance independently from other factors encountered in real life scenarios, such as background ambient noise, or errors and biases in the provided listener rotation or translation information. However, to give a basic perspective on more realistic conditions, we present some results including different levels of background noise. To test the models in noisy conditions, we utilize real ambient noise recordings from the TAU-SRIR DB database.[4] The recordings are converted from their original first-order Ambisonic format to binaural using the magnitude-least squares decoding method [80]. Four cases are tested based on the SNR between the simulated binaural speech signals and the ambient noise level, with target SNRs of 5, 10, 15 or 20 dB. The noise signals are selected randomly for each simulation scenario of binaural speech signals for the moving listener according to each specific SNR scenario.

Table VI presents the results for the joint multi-task regression model. As can be seen in the first column, both DOA and distance errors increase significantly after the addition of background noise to the testing set only. Minor differences between SNR ranging from 10 dB to 20 dB can be observed, with the DOA and distance errors increasing by at least 9.26° and 0.33 m respectively. An even more significant drop in performance is seen for an SNR equal to 5 dB: the DOA error increases to 23.06°, whereas the distance error achieves the highest value of 1.89 m. These results can be explained by the total exclusion of background noise from the training process and lack of training techniques targeted to generalize the models to noise. However, as shown in the second column, the performance of the model increases significantly when noise is also present in the training set. For the DOA estimation part, the error drops to the range between 12.55° and 14.32°, which is closer to the original results without noise. Improvements for distance estimation appear to be smaller, with an error ranging from 1.65 m for 20 dB to 1.71 m for 5 dB. The reason for the poorer performance of the SDE branch might be related to noise affecting important distance cues such as the DRR, since the binaural ambience has mostly diffuse properties not unlike the late reverberant speech. Further work is required to assess which are the most important features being affected by background noise the most. These results might be further improved by utilizing specific techniques to increase robustness to noise in the data such as dropout or prior denoising.

TABLE VI
RESULTS FOR A JOINT MULTI-TASK MODEL AS TESTED ON DATA WITH ADDED BACKGROUND NOISE

SNR [dB]	Trained without noise		Trained with noise	
	$E_{\mathrm{DOA}}[^\circ]$	$E_{\mathrm{Dist}}[m]$	$E_{\mathrm{DOA}}[^\circ]$	$E_{\mathrm{Dist}}[m]$
-	10.05 ± 2.18	1.45 ± 0.05	-	-
20	19.31 ± 2.00	1.78 ± 0.08	12.55 ± 1.70	1.65 ± 0.06
15	19.62 ± 1.03	1.83 ± 0.09	12.90 ± 1.44	1.64 ± 0.06
10	19.92 ± 1.02	1.84 ± 0.09	13.09 ± 1.33	1.69 ± 0.05
5	23.06 ± 1.52	1.89 ± 0.08	14.32 ± 1.50	1.71 ± 0.07

The scores represent a model trained without noise in the training data in the first column, and with the inclusion of noise in the second column.

VI. CONCLUSION

In this paper, we propose a method for improving binaural sound distance estimation and sound source localization by utilizing a scenario with a walking listener. Our method is tested for a large dataset of realistic scenarios, including 2500 different rooms and a wide continuous range of source-to-receiver distances and directions-of-arrival of sound sources. On top of traditional spectral features, we propose two motion-based cues, namely rotation and translation features, which are proven in our experiments to have a significant impact on the performance of both tasks. Especially for the DOAE task, the inclusion of these cues decreased the error to 1.66° as compared with 17.83° for a standard static scenario. In addition to improving the general scores, a more detailed analysis shows that a walking listener allows for flattening the error's dependence on source-to-receiver distance and room reverberation time, making both tasks more robust to most acoustic condition changes.

On top of investigating SDE and DOAE separately, we propose two methods of performing them by a joint model. Comparing a single-task and multi-task approach, we show that the latter method outperforms the state-of-the-art technique of position estimation in most testing scenarios. The joint modeling of two tasks leads to an expected decrease of performance on both sides, which might be offset by parameter fine-tuning and different architectural choices. Moreover, we note that despite of notable improvements for distance estimation, it is nevertheless challenging to obtain satisfying results due to the constraints of binaural audio. Finally, the described results should be further proven by experiments performed on real-life data, which is

[4] DOI: 10.5281/zenodo.6408611

difficult to obtain for a large variety of testing scenarios. Hence, these problems shall be explored in our future works.

References

[1] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*2016, pp. 405–409.

[2] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 74–79.

[3] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *J. Robot. Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[4] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*, Hoboken, NJ, USA: Wiley, 2012.

[5] M. Wölfel and J. W. McDonough, *Distant Speech Recognition*, Hoboken, NJ, USA: Wiley, 2009.

[6] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–46, Feb. 2016, doi: 10.1145/2871183.

[7] C. J. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, "Sound based localization and identification in industrial environments," in *Proc. IEEE 43rd Annu. Conf. Ind. Electron. Soc.*, 2017, pp. 6119–6124.

[8] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1120–1124, Sep. 2014.

[9] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *Proc. IEEE 19th Eur. Signal Process. Conf.*, 2011, pp. 1317–1321.

[10] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019, doi: 10.1109/JSTSP.2018.2885636.

[11] D. A. Krause and A. Mesaros, "Binaural signal representations for joint sound event detection and acoustic scene classification," in *Proc. 30th Eur. Signal Process. Conf.*, 2022, pp. 399–403.

[12] A. Brendel and W. Kellermann, "Distance estimation of acoustic sources using the coherent-to-diffuse power ratio based on distributed training," in *Proc. IEEE 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 1–5.

[13] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. Providence, RI, USA: Brown Univ. Providence, 2000.

[14] I. Trowitzsch, J. Mohr, Y. Kashef, and K. Obermayer, "Robust detection of environmental sounds in binaural auditory scenes," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1344–1356, Jun. 2017.

[15] R. Gupta et al., "Augmented/mixed reality audio for hearables: Sensing, control, and rendering," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 63–89, May 2022.

[16] K. Youssef, S. Argentieri, and J.-L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Proc. IEEE Int. Conf. Acoust. Speech Signal. Process.*, 2012, pp. 217–220.

[17] 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, pp. 1–16, Feb. 2020.

[18] M. Yiwere and E. J. Rhee, "Sound source distance estimation using deep learning: An image classification approach," *Sensors*, vol. 20, no. 1, 2020, Art. no. 172. [Online]. Available: https://www.mdpi.com/1424-8220/20/1/172

[19] A. Sobhdel, R. Razavi-Far, and S. Shahrivari, "Few-shot sound source distance estimation using relation networks," 2021, *arXiv: 2109.10561*.

[20] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," *Int. J. Appl. Eng. Res.*, vol. 12, no. 22, pp. 12384–12389, 2017.

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

[22] M. U. Liaquat, H. S. Munawar, A. Rahman, Z. Qadir, A. Z. Kouzani, and M. A. P. Mahmud, "Sound localization for ad-HOC microphone arrays," *Energies*, vol. 14, no. 12, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/12/3446

[23] M. Hahmann, E. Fernandez-Grande, H. Gunawan, and P. Gerstoft, "Sound source localization using multiple ad HOC distributed microphone arrays," *JASA Exp. Lett.*, vol. 2, no. 7, 2022, Art. no. 074801, doi: 10.1121/10.0011811.

[24] D. Krause, A. Politis, and K. Kowalczyk, "Comparison of convolution types in CNN-based feature extraction for sound source localization," in *Proc. IEEE 28th Eur. Signal Process. Conf.*, 2021, pp. 820–824.

[25] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP J. Adv. Signal Process*, vol. 2010, pp. 1–13, Feb. 2010, doi: 10.1155/2010/415840.

[26] T. Huisman, A. Ahrens, and E. MacDonald, "Ambisonics sound source localization with varying amount of visual information in virtual reality," *Front. Virtual Reality*, vol. 2, 2021, Art. no. 722321, doi: 10.3389/frvir.2021.722321.

[27] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2019, pp. 20–24.

[28] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[29] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*1997, pp. 375–378.

[30] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[31] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[32] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust Localization in Reverberant Rooms*. Berlin, Heidelberg:Springer, 2001, pp. 157–180.

[33] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[34] C. M. Zannini, R. Parisi, and A. Uncini, "Binaural sound source localization in the presence of reverberation," in *Proc. IEEE Int. Conf. Digit. Signal Process*, 2011, pp. 1–6.

[35] X. Wan and J. Liang, "Robust and low complexity localization algorithm based on head-related impulse responses and interaural time difference," *J. Acoust. Soc. Amer.*, vol. 133, no. 1, pp. EL40–EL46, 2013.

[36] 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," *J. Acoust. Soc. Amer.*, vol. 135, no. 3, pp. 1207–1217, 2014.

[37] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[38] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. IEEE 14th Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 99–103.

[39] N. Roman, D. Wang, and G. J. Brown, "A classification-based cocktail-party processor," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1425–1432.

[40] S. Chakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 136–140.

[41] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for ambisonics recordings," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 22–33, Mar. 2019.

[42] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A review of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, pp. 107–151, 2022.

[43] H. Liu, P. Yuan, B. Yang, G. Yang, and Y. Chen, "Head-related transfer functionreserved time-frequency masking for robust binaural sound source localization," *CAAI Trans. Intell. Technol.*, vol. 7, no. 1, pp. 26–33, 2021.

[44] G. García-Barrios, D. A. Krause, A. Politis, A. Mesaros, J. M. Gutiérrez-Arriola, and R. Fraile, "Binaural source localization using deep learning and head rotation information," in *Proc. IEEE 30th Eur. Signal Process. Conf.*, 2022, pp. 36–40.

[45] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1793–1805, Sep. 2010.

[46] P. N. Samarasinghe, T. D. Abhayapala, M. Polettfi, and T. Betlehem, “On room impulse response between arbitrary points: An efficient parameterization,” in *Proc. IEEE 6th Int. Symp. Commun. Control Signal Process.*2014, pp. 153–156.

[47] T. Rodemann, “A study on distance estimation in binaural sound localization,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 425–430.

[48] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos, “Sound source distance estimation in rooms based on statistical properties of binaural signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1727–1741, Aug. 2013.

[49] S. Vesa, “Binaural sound source distance learning in rooms,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1498–1507, Nov. 2009.

[50] J. Gontmacher, A. Yarhi, P. Havkin, D. Michri, and E. Fisher, “DSP-based audio processing for controlling a mobile robot using a spherical microphone array,” in *Proc. IEEE 27th Conv. Elect. Electron. Eng. Isr.*, 2012, pp. 1–5.

[51] D. Gabriel, R. Kojima, K. Hoshiba, K. Itoyama, K. Nishida, and K. Nakadai, “2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system,” *Adv. Robot.*, vol. 33, no. 7/8, pp. 403–414, 2019.

[52] J. K. Nielsen, “Loudspeaker and listening position estimation using smart speakers,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 81–85.

[53] J. hwan Hwang, S. Seon, and C.-S. Park, “Position estimation of sound source using three optical mach-zehnder acoustic sensor array,” *Curr. Opt. Photon.*, vol. 1, no. 6, pp. 573–578, Dec. 2017.

[54] L. Ghamdan, M. A. Ismail Shoman, R. A. Elwahab, and N. A. El-Hadid Ghamry, “Position estimation of binaural sound source in reverberant environments,” *Egyptian Inform. J.*, vol. 18, no. 2, pp. 87–93, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866516300378

[55] D. A. Krause, A. Politis, and A. Mesaros, “Joint direction and proximity classification of overlapping sound events from binaural audio,” in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, 2021, pp. 331–335.

[56] L. Rayleigh, “XII, on our perception of sound direction,” *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 13, no. 74, pp. 214–232, 1907.

[57] J. Blauert and S. Hearing, “The psychophysics of human sound localization,” in *Spatial Hearing*, Cambridge, MA, USA: MIT Press, 1997.

[58] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 2699–2703.

[59] C. Schymura, F. Winter, D. Kolossa, and S. Spors, “Binaural sound source localisation and tracking using a dynamic spherical head model,” in *Proc. Interspeech*, 2015, pp. 165–169.

[60] T. May, N. Ma, and G. J. Brown, “Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 2679–2683.

[61] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[62] 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 Trans. Audio, Speech. Lang. Process.*, vol. 26, no. 11, pp. 2122–2131, Nov. 2018.

[63] M. Zohourian and R. Martin, “Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 430–434.

[64] A. Portello, P. Danès, and S. Argentieri, “Active binaural localization of intermittent moving sources in the presence of false measurements,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 3294–3299.

[65] A. Portello, G. Bustamante, P. Danès, J. Piat, and J. Manhes, “Active localization of an intermittent sound source from a moving binaural sensor,” in *Eur. Acoust. Assoc. Forum Acusticum*, Krakov, Poland, Sep. 2014, pp. 12–25. [Online]. Available:https://hal.laas.fr/hal-01969308

[66] L. Kneip and C. Baumann, “Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis,” *J. Acoustical Soc. Amer.*, vol. 124, no. 5, pp. 3108–3119, 2008.

[67] D. Gala, N. Lindsay, and L. Sun, “Realtime active sound source localization for unmanned ground robots using a self-rotational bi-microphone array,” *J. Intell. Robot. Syst.*, vol. 95, no. 3, pp. 935–954, Sep. 2019.

[68] 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 *Proc. IEEE 43rd Int. Conf. Acoust. Speech Signal Process.*2018, pp. 1–5.

[69] D. Krause, A. Politis, and K. Kowalczyk, “Feature overview for joint modeling of sound event detection and localization using a microphone array,” in *Proc. 28th Eur. Signal Process. Conf.*, 2020, pp. 31–35.

[70] R. G. Klumpp and H. R. Eady, “Some measurements of interaural time difference thresholds,” *J. Acoust. Soc. Amer.*, vol. 28, pp. 859–860, 1956. [Online]. Available: https://api.semanticscholar.org/CorpusID:121232144

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

[72] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” *NeurIPS*, pp. 8024–8035, 2019.

[73] J. S. Garofolo, “TIMIT acoustic phonetic continuous speech corpus,” in *Linguistic Data Consortium*, 1993.

[74] P. Srivastava, A. Deleforge, A. Politis, and E. Vincent, “How to (virtually) train your sound source localizer,” 2022, *arXiv:2211.16958*.

[75] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 351–355.

[76] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, “gpuRIR: A python library for room impulse response simulation with GPU acceleration,” *Multimedia Tools Appl.*, vol. 80, pp. 5653–5671, 2021.

[77] A. Politis, “Microphone array processing for parametric spatial audio techniques,” Ph.D. dissertation, Aalto Univ., Espoo, Finland, 2016, [Online]. Available: https://github.com/polarch/shoebox-roomsim

[78] “Sound absorption coefficient chart: JCW acoustic supplies,” [Online]. Available: https://www.acoustic-supplies.com/absorption-coefficient-chart/

[79] R. W. Bohannon and A. Williams Andrews, “Normal walking speed: A descriptive meta-analysis,” *Physiotherapy*, vol. 97, no. 3, pp. 182–189, 2011.

[80] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Berlin, Germany: Springer, 2019.

Daniel Aleksander Krause received the bachelor’s and master’s degree in acoustical engineering from the AGH University of Science and Technology, Kraków, Poland, in 2018 and 2019, respectively. He is currently a Doctoral Researcher with Tampere University, Tampere, Finland. His professional journey includes a tenure as a Data Scientist with Fitech from 2017 to 2018, followed by a role in the Signal Processing Group within the Department of Electronics with AGH from 2019 to 2020. Since 2020, he has been working towards the doctoral studies, contributing as a Member of both the Audio Research Group and the Machine Listening group with Tampere University. Starting in 2021, he assumed the role of Co-organizer for the DCASE scientific challenge. His research interests include data science, signal processing, machine learning, spatial audio, and acoustic scene analysis.

Guillermo García-Barrios was born in Madrid, Spain, in 1994. He received the B.Sc. degree in sound and image engineering, the M.Sc. degree in systems and services engineering for the information society, and the Ph.D. degree in acoustic signal processing from the Universidad Politécnica de Madrid, Madrid, Spain, in 2017, 2018, and 2023, respectively. He is curently with Fivecomm as an R&D Engineer, an IoT company focused on the digitalization of the industry. His research interests include mobile communications, signal processing, cell-free massive MIMO, simulations, machine learning, sound source localization, acoustics, and wireless acoustic sensor networks.

Archontis Politis received the M.Sc. degree in sound and vibration studies from the Institute of Sound and Vibration Research, University of Southampton, Southampton, U.K., in 2008. He received the Doctor of Science degree on spatial audio processing from Aalto University, Espoo, Finland, in 2016. He is currently an Assistant Professor with Tampere University, Tampere, Finland. From 2008 to 2009, he was a Researcher in a joint collaboration between the Glasgow school of Arts and Arup Acoustics, Glasgow, U.K., performing research on virtual acoustics. In 2015 he was a visiting Researcher with the University of Maryland Institute for Advanced Computer Studies, College Park, MA, USA and in the same year he completed a research internship with Microsoft Research, Redmond, WA, USA, on spatial audio technologies. He was an Editor of a book on Parametric Spatial Audio Processing, Organizer with the DCASE scientific challenge, and has chaired various special sessions in international conferences. His research interests include spatial audio technologies, virtual acoustics, array signal processing, and acoustic scene analysis.

Annamaria Mesaros (Senior Member, IEEE) received the Ph.D. degree in signal processing from the Tampere University of Technology, Tampere, Finland, in 2012. She is currently an Associate Professor with Tampere University, Tampere. Her research interests include sound event detection in real-world multisource environments and includes more than 40 scientific publications and many open datasets. She is a Member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society. She is a Co-ordinator of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Vice-chair of the DCASE Steering Group, and currently an Academy of Finland Research Fellow for Teaching Machines to Listen.