Speaker Distance Estimation in Enclosures From Single-Channel Audio

Michael Neri, *Graduate Student Member, IEEE*, Archontis Politis, *Member, IEEE*, Daniel Aleksander Krause, *Graduate Student Member, IEEE*, Marco Carli, *Senior Member, IEEE*, and Tuomas Virtanen, *Fellow, IEEE*

***Abstract*—Distance estimation from audio plays a crucial role in various applications, such as acoustic scene analysis, sound source localization, and room modeling. Most studies predominantly center on employing a classification approach, where distances are discretized into distinct categories, enabling smoother model training and achieving higher accuracy but imposing restrictions on the precision of the obtained sound source position. Towards this direction, in this paper we propose a novel approach for continuous distance estimation from audio signals using a convolutional recurrent neural network with an attention module. The attention mechanism enables the model to focus on relevant temporal and spectral features, enhancing its ability to capture fine-grained distance-related information. To evaluate the effectiveness of our proposed method, we conduct extensive experiments using audio recordings in controlled environments with three levels of realism (synthetic room impulse response, measured response with convolved speech, and real recordings) on four datasets (our synthetic dataset, QMULTIMIT, VoiceHome-2, and STARSS23). Experimental results show that the model achieves an absolute error of 0.11 meters in a noiseless synthetic scenario. Moreover, the results showed an absolute error of about 1.30 meters in the hybrid scenario. The algorithm's performance in the real scenario, where unpredictable environmental factors and noise are prevalent, yields an absolute error of approximately 0.50 meters.**

***Index Terms*—Distance estimation, single-channel, deep learning, reverberation, explainability, attention.**

I. Introduction

SOURCE distance estimation (SDE) refers to the task of estimating the interspace between a microphone and a sound source. It is very often performed in conjunction with direction of arrival (DoA) estimation, in which only the direction information about the source position is obtained. Both tasks are useful in many practical applications, including increasing the robustness of automatic speech recognition [1] by enhancing the performance of acoustic echo cancellers [2] and autonomous robotics [3], [4]. Despite both DoA and source distance being estimated using multi-channel audio in most practical scenarios, the latter has been largely under-researched [5]. Firstly, source distance estimation is widely regarded a more difficult task due to distance cues vanishing with the increased space between the sound source and the receiver. Secondly, DoA offers sufficient information in many downstream spatial filtering tasks. However, many applications such as source separation, acoustic monitoring, and context-aware devices, would still benefit from full information about the sound source position, hence the need for further investigations on source distance estimation (SDE).

Most methods for both DOA and distance estimation rely on arrays with more than two microphones [6]. Multichannel data allows for exploiting spatial cues such as interchannel time differences (ITDs) and interchannel level differences (ILDs) to provide information for efficient DoA estimation, positively affecting distance estimation as well [4]. However, using multiple microphones poses certain limitations in terms of budget and physical portability. To tackle this problem, some studies investigated using binaural recordings for that purpose, allowing for decreasing the number of channels to two by exploiting the human hearing cues [7], [8]. However, the simplest scenario of estimating distance from a single microphone has been largely under-researched [9]. Moreover, the vast majority of studies focus on a classification approach, in which the distance is discretized into a set of disjunctive categories, e.g., "far" and "near", allowing for easier model training and a higher accuracy [10], [11]. However, using pre-defined categories does not allow for continuous estimation, which puts limits on the precision of the obtained sound source position.

Towards this direction, in this work, we propose several novel solutions to tackle the problem of source distance estimation. Firstly, we define the task as a regression problem, differently from most state-of-the-art works that focus on classification-based methods. We propose a novel approach to distance estimation from single-channel audio signals in reverberant environments, overcoming the need for complex microphone arrays. In more detail, the proposed model is a convolutional recurrent neural network (CRNN) with an attention module, which is responsible for learning a time-frequency attention map. By doing so, it is possible to emphasize magnitude- and phase-related features that are the most informative for sound source distance

Manuscript received 13 September 2023; revised 16 February 2024 and 21 March 2024; accepted 23 March 2024. Date of publication 27 March 2024; date of current version 10 April 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nobutaka Ito. *(Corresponding author: Michael Neri.)*

Michael Neri and Marco Carli are with the Department of Industrial, Electronic, and Mechanical Engineering, Roma Tre University, 00154 Rome, Italy (e-mail: michael.neri@uniroma3.it; marco.carli@uniroma3.it).

Archontis Politis, Daniel Aleksander Krause, and Tuomas Virtanen are with the Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland (e-mail: archontis.politis@tuni.fi; daniel.krause@tuni.fi; tuomas.virtanen@tuni.fi).

For reproducible research purposes we make model, code, and synthetic datasets available at https://github.com/michaelneri/audio-distance-estimation.

Digital Object Identifier 10.1109/TASLP.2024.3382504

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

estimation. The effectiveness of our approach is extensively tested for numerous acoustic scenarios, obtained by simulations with randomized configurations of room shapes, materials, and locations of the microphone and the speaker. In addition, tests have been carried out on real reverberant speech recordings, captured directly or emulated with real room impulse responses (RIRs).

The remainder of the manuscript is organized as follows. Section II provides a summary of the state-of-the-art. Section III describes the proposed methods, whereas the performance evaluations are in Section IV. Section V details the experimental results of the proposed approach on three acoustic scenarios. Finally, Section VI includes an overall discussion regarding the work, and Section VII draws the conclusions.

II. Related Works

SDE involves determining the distance between a sound source and the receiver. When compared to DoA estimation, SDE is an area that has received significantly less attention and is generally considered more challenging. This is primarily due to the fact that the accuracy of distance estimation declines rapidly for small-sized arrays commonly used in practice even for relatively short distances from the center of the array (up to 3-4 m). Several factors contribute to this phenomenon, including: a) the decrease in direct-to-reverberant energy ratio (DRR) and signal-to-noise ratios (SNRs) as the source distance increases, b) the reduction in 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.

The majority of studies related to SDE show results in conjunction with the DoA estimation task. Extensive research has been conducted on this subject for various acoustic systems that commonly use distributed microphone arrays. These systems encompass a range of setups, such as intelligent loudspeakers [12], spherical microphones [13], triangular configurations [14], and arrays of acoustic sensors [15]. Simpler audio formats including binaural recordings have been investigated to a much lesser extent, including few studies with classical machine learning methods [4], [16] and very limited research related to deep learning [7], [8].

Regarding SDE modeling in isolation, most of the research has been focused on parametric approaches and manually crafted features. These methods often utilize information such as the DRR [17], RIR [18], or signal statistics and binaural cues such as the interchannel intensity difference (IID) [4]. In some cases, classical machine learning techniques have been employed to leverage statistical features. For instance, a study by Brendel et al. estimated the coherent-to-diffuse power ratio to determine the source-microphone distance via Gaussian mixture models (GMMs) [5]. Vesa utilized GMMs trained with magnitude squared coherence (MSC) features to incorporate information about channel correlation [19], [20]. In [21], the authors used MSC on top of other features to train classifiers with methods such as K-nearest neighbours (KNN) or linear discriminative analysis (LDA). Georganti et al. introduced the binaural signal magnitude difference standard deviation (BSMD-STD) and trained GMMs and support vector machines (SVMs) using this feature [10]. Most of these methods rely on compound algorithms that require careful tuning to adapt to varying acoustic conditions.

Until now, the exploration of source distance estimation using deep neural networks (DNNs) has been quite limited. Yiwere et al. employed an approach inspired by image classification, utilizing CRNNs trained on log-mel spectrograms to classify three different distances in three distinct rooms [22]. Although the models demonstrated promising outcomes for data within the same environment, their performance significantly deteriorated when dealing with recordings from different rooms. In another endeavor, Sobhdel et al. introduced relation networks to address this challenge through few-shot learning, which exhibited enhancements over conventional convolutional neural networks (CNNs) [23]. Both studies conducted tests within a limited range of specific distances, encompassing a close proximity of up to 3-4 meters at most. In [8], the authors conducted experiments for data covering distances for up to 8 m, however the model was classifying them into two binary classes denoted as “far” and “near”.

Additionally, only a few works have addressed the topic of speaker distance estimation using single-channel audio. One of the first works employed low-level features such as linear predictive coding (LPC), skewness, and kurtosis of the spectrum to classify the distance of a speaker [11]. Venkatesan et al. proposed both monaural and binaural features to train GMMs and SVMs [24]. Regarding DNN approaches, Patterson et al. classified “far” and “near” speech in order to perform sound source separation from single-channel audios [9].

To the best of our knowledge, single-channel source distance estimation has been scarcely addressed as a regression problem, prioritizing classification approaches to ease model training. In addition, there are very few studies investigating the use of DNNs in this task. For these reasons, a learning-based approach for continuous estimation of the distance of the speaker is proposed. A first step towards continuous sound source distance estimation occurs in our preliminary study [25] where a CRNN was defined for estimating static speaker distance in simulated reverberant environments from a single omnidirectional microphone. However, that study was evaluated only on simulations, while in this work various degrees of realism are investigated, from simulated RIRs, to synthetic data with measured RIRs, to fully real recordings with distance-annotated sources. Hence, the potential of the method in a real-world scenario is demonstrated. In addition, the preliminary study was based on a simpler architecture without investigation on what architectural components contributed the most to the SDE, while here the architecture is refined and enhanced, with better overall performance, and specific choices investigated in an ablation study.

To cope with these limitations, the contributions of this work are as follows

- a major improvement of the results of the learning-based approach, i.e., a CRNN, proposed in our preliminary study [25] that simultaneously provides temporal frame-wise and utterance-wise distance estimation of the static

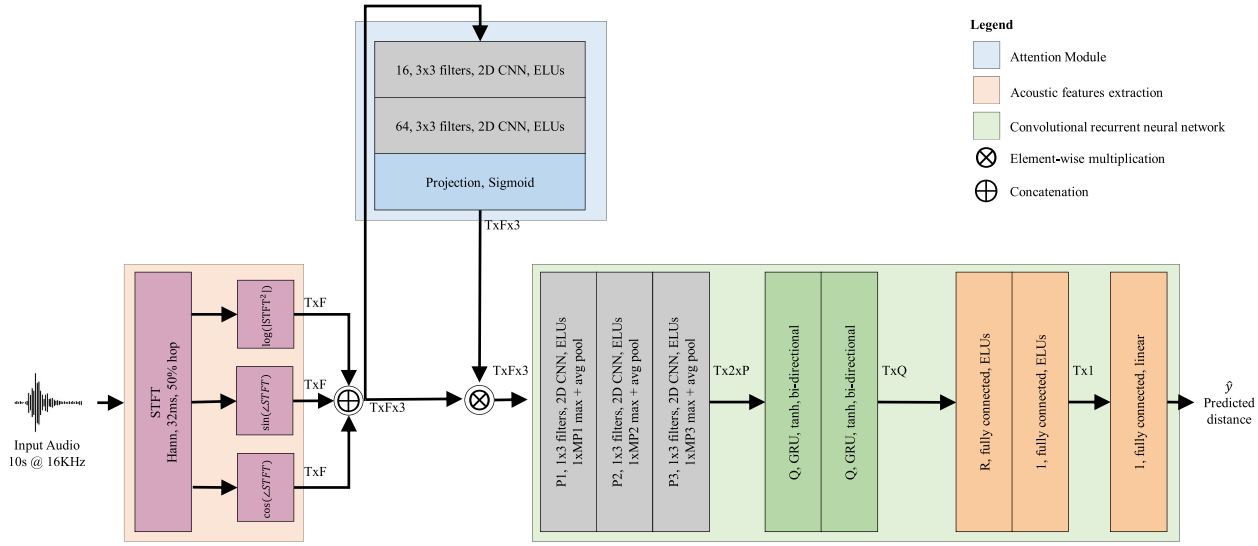


Fig. 1. Proposed architecture for speaker distance estimation. First, acoustic features are extracted from the single-channel audio. In more detail, 3 maps (magnitude of the short-time Fourier transform (STFT), sinus, and cosinus of the STFT phase) are obtained with shape $T \times F$, where T and F are the time and frequency bins, respectively. Then, the maps are stacked along the channel dimension resulting in a feature tensor of size $T \times F \times 3$. To highlight the feature regions that are most informative for distance estimation, an attention map is learned from the three-channel tensor, which is then element-wise multiplied with the input feature tensor. The output is further processed by the convolutional layers with P_i 1×3 kernels, also denoted as *frequency kernels*, yielding a $T \times 2 \times P$ tensor that is arranged in a $T \times Q$ matrix, where $Q = 2P$. Subsequently, the resulting matrix is analyzed by two gated recurrent unit (GRU) layers with Q neurons to model temporal patterns. Finally, the output from recurrent layers $T \times Q$ is fed to three fully connected layers with R, 1, and 1 neurons respectively to map the features to the predicted distance $\hat{y}$.

audio source. In addition, an in-depth study regarding the model architecture is detailed;

- definition of an attention module that estimates the most significant time-frequency patterns from the input features for speaker distance estimation;
- experiments have been conducted on synthetic data, both in noiseless and noisy scenarios, to analyze the response of the proposed approach in controlled environments. Further tests on the CRNN have been conducted on a constructed hybrid dataset, i.e., measured RIRs convolved with anechoic speeches, and two real recording datasets, demonstrating the generalization capabilities of the proposed approach.

III. Proposed Method

In this section, a description of the acoustic features for the source distance estimation is provided. To process temporal, spatial, and spectral characteristics of these features, a CRNN has been employed for the experiments. This type of model has shown good results in many studies for sound event localization and detection (SELD) tasks [26], [27]. In addition, an attention module is introduced to learn an attention map on the time-frequency audio representation. The overall architecture is depicted in Fig. 1.

A. Acoustic Features Extraction

All the operations on the audio files are performed at 16 kHz. The selection of this sampling frequency is because the speech spectrum is mostly contained in the range 0-8 kHz [28]. In addition, a lower frequency yields a lower number of samples, reducing the computational complexity of feature extraction and distance estimation. Initially, a pre-processing stage is employed to extract the complex STFT $\text{STFT}\{\mathbf{x}\} \in \mathbb{C}^{T \times F}$ from the single-channel audio signal $\mathbf{x} \in \mathbb{R}^{1 \times L}$, where T is the number of time frames, F the number of frequency bins, and L the number of samples. This transformation is computed using a Hann window of length 32 ms with 50% overlap. Subsequently, the magnitude ($|\text{STFT}\{\mathbf{x}\}| \in \mathbb{R}^{T \times F}$) and phase ($\angle\text{STFT}\{\mathbf{x}\} \in \mathbb{R}^{T \times F}$) components of the STFT are computed from the complex matrix.

Sinus and cosinus maps of the phase spectrogram are computed by applying $\sin(\cdot)$ and $\cos(\cdot)$ functions element-wise, since the features provide a smoother continuous representation of the raw phase information. The concept of utilizing the phase spectrogram has been adopted from contemporary research on multichannel source separation [29], learning-based localization [30], and speech enhancement [31] as phase information contains cues regarding the acoustic properties of the environment in which the sound propagates [32]. Tests conducted using the raw complex spectrogram in our scenario, i.e., two separate branches that processed real and imaginary parts, yielded unsatisfactory training performance.

Finally, the magnitude of the STFT and the sinus and cosinus maps are stacked into a $T \times F \times 3$ tensor. This representation is then fed into the attention module and the convolutional layers for further processing and analysis.

B. Attention Module

One of the main contributions of this work is the definition of an attention module which computes an attention map $H \in$

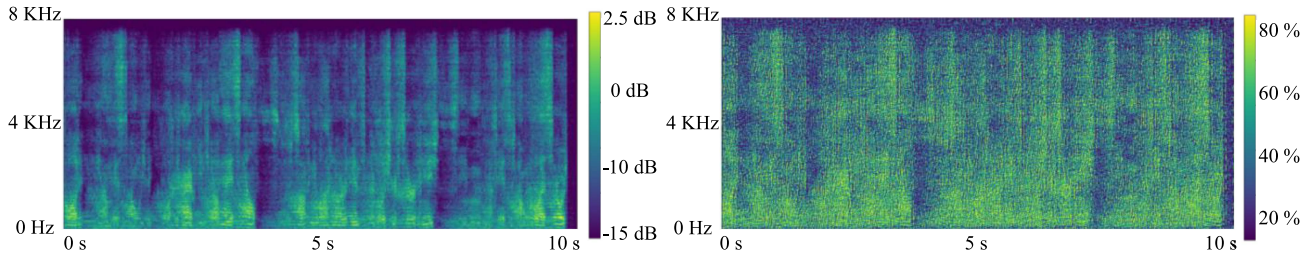


Fig. 2. Example of spectrogram and attention map on a noiseless sample of the synthetic dataset with a speaker talking at 10 meters.

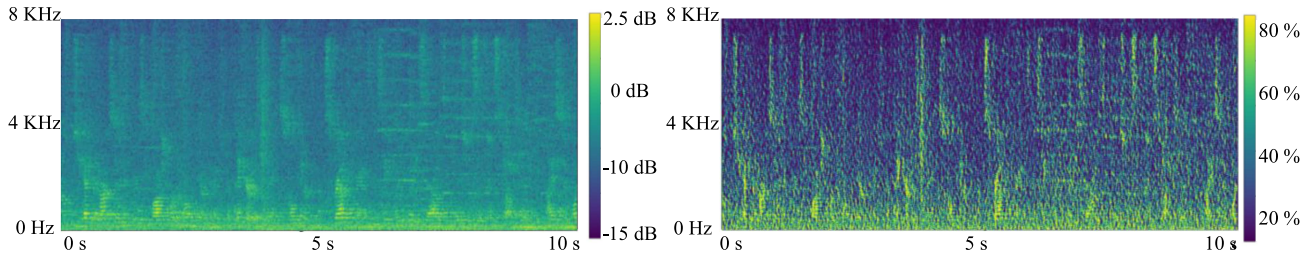


Fig. 3. Example of spectrogram and attention map on a noisy sample (SNR $= 0$ dB) of the synthetic dataset with a speaker talking at 10 meters.

$\mathbb{R}^{+T\times F\times 3}$ from the audio features. The objective of this learned matrix is to emphasize the regions of the features that are most informative for the estimation of the distance. Specifically, this module is the function $\mathrm{f}_{\mathrm{ATT}} : \mathbb{R}^{T\times F\times 3} \rightarrow \mathbb{R}^{+T\times F\times 3}$. Its structure is composed of 2 convolutional blocks, having 16 and 64 3×3 filters, respectively. Then, a 1×1 convolutional layer with three filters, followed by a sigmoid activation, is used to map the features to yield the $T \times F \times 3$ attention map. Finally, the output acoustic features $\tilde{X} \in \mathbb{R}^{T\times F\times 3}$ are obtained by element-wise multiplication ($\otimes$) between the input acoustic features and the attention map as

$$\tilde{X} = f_{\mathrm{ATT}}(X) \otimes X. \tag{1}$$

Examples of noiseless and noisy spectrograms and attention maps are depicted in Figs. 2 and in 3, respectively. It is worth highlighting how the attention module differently focuses on the parts of the signal where the speech is most likely to stand out from the noise, or where the characteristics of the speech are still recognizable. In fact, the attention map in a noiseless case is evenly distributed across the entire frequency range since there is no noise that interferes.

C. Convolutional Layers

The architecture employs three convolutional blocks for feature extraction. In more detail, the structure of each block involves a 2D convolutional layer comprising P_i 1×3 filters, i.e., along the frequency axis with values of 8, 32, and 128 assigned to the respective layers. We denote these filters as *frequency kernels* whereas 3×1 filters are named *time kernels*. Square kernels, known for their capability to capture time-frequency patterns, are commonly used in convolutional layers applied to spectrograms due to their effectiveness in capturing local patterns and structures along the frequency axis. In this work, the proposed model adopts rectangular filters, and temporal information is modeled by recurrent layers at the end of the model. In fact, rectangular filters can be more parameter-efficient compared to square kernels. Since the former has fewer parameters than square kernels of the same receptive field size, they can lead to a more compact model, making training and inference more computationally efficient and potentially reducing the risk of overfitting, especially when working with limited data.

Following this layer, a batch normalization [33] step is applied, along with max and average pooling operations along the frequency dimension. Then, the results of which are summed.

The activation function utilized after each convolutional layer is the exponential linear unit (ELU) [34], which is denoted as

$$\mathrm{ELU}(x) = \begin{cases} x, & x \geq 0 \\ \alpha(e^x - 1), & x < 0 \end{cases} \tag{2}$$

where α is a coefficient that regularizes the saturation of negative values. Notably, each layer employs a specific pooling rate denoted by MP_i, with values of 8, 8, and 2 assigned to the respective layers.

D. Recurrent Layers

To process the feature maps from the convolutional layers, two bi-directional GRU layers are utilized with $\tanh(\cdot)$ as the activation function. These layers have exhibited promising results in audio and speech processing tasks, demonstrating parameter efficiency compared to long short-term memory (LSTM) networks [35].

The output of the CNN with shape $T \times 2 \times P$ is stacked along the channel dimension to produce a $T \times Q$ matrix to be fed to the recurrent layers. Then, in the proposed configuration, the extraction of reverberation-related information primarily relies on integrating information over time with the recurrent layers. Within this implementation, two bi-directional GRUs with $Q = 2P = 128$ neurons each for every time frame are employed.

Then, to predict the distance, three fully connected layers are employed, where an independent mapping between each time frame is performed in each layer. Firstly, the initial linear layer projects time-wise features from the last GRU onto a matrix of dimensions $T \times R$, where $R = 128$. Subsequently, the second linear layer independently maps each time frame of the $T \times R$ matrix onto a vector of size $T \times 1$, denoted as the time-wise distance estimation $\hat{\mathbf{y}}$. Specifically, this vector represents the distance estimation for each time frame. Finally, the last fully connected layer is employed to perform regression and thus estimate the predicted distance, denoted as $\hat{y} \in \mathbb{R}$.

E. Loss Function

The mean squared error (MSE) loss is used to train the DNN system. Let $y \in \mathbb{R}$ be the true distance of a static sound source. In addition, let $\mathbf{y} \in \mathbb{R}^{T\times 1}$ be the vector consisting of frame-wise ground truth distances. Then, the loss used in the training phase for a single sample is

$$\mathcal{L}(y, \hat{y}, \mathbf{y}_t, \hat{\mathbf{y}}_t) = (y - \hat{y})^2 + ||\mathbf{y}_t - \hat{\mathbf{y}}_t||^2, \tag{3}$$

where the loss is averaged across the batch dimension to be exploited by the backpropagation algorithm. Thanks to the imposition of the loss, the model predicts a distance for each time bin and, from this information, a single-valued distance. Having two losses in a static source scenario operates as a regularization term since it forces the proposed approach to return coherently both time-wise and single-distance estimations. However, in the context of dynamic sound sources, it is important to highlight that only frame-wise loss is required.

F. Metrics

The performance evaluation of our approach utilizes the mean absolute error (MAE) ($\mathcal{L}_1$) as the performance measure for the entire test dataset

$$\mathcal{L}_1(y, \hat{y}) = |y - \hat{y}|, \tag{4}$$

where the ground truth $y \in \mathbb{R}$ and the prediction $\hat{y} \in \mathbb{R}$ are considered. Additionally, the performance is assessed by calculating the MAE within different distance ranges. This analysis allows us to quantify the relative error of our model concerning source distance. We define the relative MAE ($r\mathcal{L}_1$), which includes the real speaker distance in the evaluation, as follows:

$$\mathrm{r}\mathcal{L}_1(y, \hat{y}) = \frac{\mathcal{L}_1(y, \hat{y})}{y} = \frac{|y - \hat{y}|}{y}. \tag{5}$$

For the sake of clarity and brevity, MSE has not been considered in the performance evaluation.

IV. Performance Assessment

This section describes how the performance assessment of the proposed approach has been carried out. To validate the work, three levels of realism have been addressed in the scope of speaker distance estimation:

- *Synthetic:* simulated RIRs of an image-source room simulator are convolved with anechoic speech;
- *Hybrid:* measured RIRs are convolved with anechoic speech;
- *Real:* on-field reverberant speech recordings.

Fig. 4 depicts the histograms of distances in each dataset employed in the experimental results.

A. Synthetic Dataset

The dataset used for experiments follows the same setup as in [36]. Briefly, anechoic speech recordings obtained from the TIMIT dataset [37] are convolved with the simulated omnidirectional RIRs from an image-source room simulator for shoebox geometries [38].

This simulator allows for frequency-dependent wall absorption and directional encoding of image sources in 5^{th} order Ambisonics format. The elevation range between the source and the receiver spanned from $-35°$ to $35°$. To compile a list of materials and their respective absorption coefficients for each surface type (ceiling, floor, and wall), we refer to widely used acoustical engineering tables [39]. For each unique simulated room with its room-source-distance configuration, a random material is assigned to each surface, resulting in 2912 possible material combinations. Compared to randomizing directly the target RT60 for each simulated room, this randomization approach allows us to avoid matching unnatural reverberation times to specific room volumes (e.g., a very long RT60 for a small room) and ensure a more natural distribution of reverberation times.

The final distribution of reverberation times exhibits a median, 10^{th} percentile, and 90^{th} percentile of 0.83 s, 0.42 s, and 2.38 s, respectively. Furthermore, the positions of the sound sources are uniformly distributed in terms of the azimuth angle relative to the receiver.

The experiments include 2500 audio files of 10 s duration at 16 kHz in compliance with the speech dataset. In the evaluation, 5-fold cross validation is used where 1500, 500, and 500 files are assigned to training, validation, and testing in each fold.

To assess the performance of the proposed approach under different noise levels, real background noise is added into the synthetic dataset. Specifically, environmental noise recordings from the WHAM! [40] dataset, captured in various urban settings such as restaurants, cafes, and bars, are employed. Random segments of the same length as the simulated speech recordings are injected, mirroring the same split as the WHAM! dataset, with several SNRs levels ([50, 40, 30, 20, 10, 5, 0] dB).

In addition to estimating the mean absolute distance estimation error, the errors are calculated separately for separate distance intervals that are $\{[1, 2), [2, 4), [4, 8), [8, 14)\}$ meters. The MAE errors are averaged using a 5-fold cross-validation split, and the 95% mean confidence intervals are evaluated.

B. Hybrid Dataset - QMULTIMIT

The RIRs used in the hybrid dataset, contained in the C4DM RIR database [41], were measured in three rooms located at Queen Mary, University of London, London, U.K.. A Genelec 8250 A loudspeaker was employed as the source for measuring

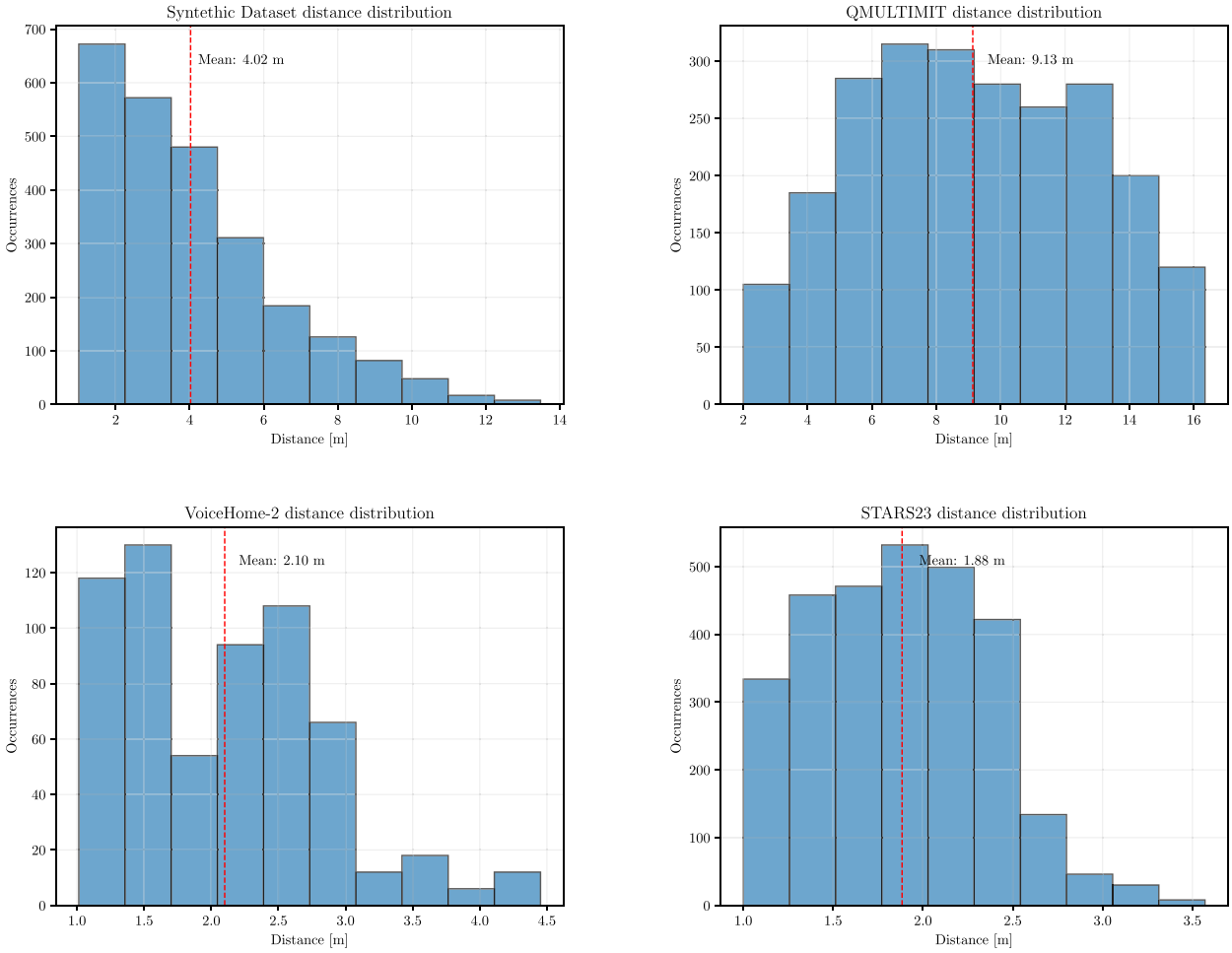


Fig. 4. Distributions of distances in each dataset.

all IRs, while each receiver position was measured using both an omnidirectional DPA 4006 and a B-format Soundfield SPS422B.

A collection of 130 RIRs was captured in a classroom with dimensions $7.5 \times 9 \times 3.5$ m (236 m^3) and consist of reflective surfaces such as a linoleum floor, painted plaster walls, ceiling, and a sizable whiteboard.

The second room, denoted as the Octagon, is a Victorian structure that was finalized in 1888. Presently serving as a conference venue, the walls of this building still showcase book-lined interiors, complemented by a wooden floor and plaster ceiling. As the name implies, this room features eight walls, each measuring 7.5 m in length, and a domed ceiling towering 21 m above the floor, resulting in an estimated volume of 9500 m^3. In the center of the room, a total of 169 RIRs were measured.

The third room is The Great Hall which possesses a seating capacity of approximately 800. It encompasses a stage and seating sections both on the floor and a balcony. To capture the audio, the microphones were positioned within the cleared seating area on the floor, spanning an area of approximately 23×16 m. The microphone placements mirror the layout used for the Octagon, encompassing 169 RIRs over a 12×12 m region.

Following the same setup of the synthetic dataset, anechoic speech recordings are convolved from TIMIT [37] and real background noises from WHAM! [40] are added with the measured RIRs, generating the hybrid QMULTIMIT dataset. For each RIR, 5 random speech recordings are selected from the TIMIT dataset, yielding 2340 audio files. RIRs are randomly divided into training, validation, and testing splits following a percentage ratio of 70-10-20. Finally, the MAE errors averaged across all the distance bins are provided.

C. Real Dataset

- *VoiceHome - 2 [42]:* This dataset is specifically made for distant speech processing applications in domestic environments. It consists of short commands for smart home devices in French, collected in reverberant conditions and uttered by twelve native French speakers facing the microphone. The data is recorded in twelve different rooms corresponding to four houses, with fully annotated geometry, under quiet or noisy conditions. More precisely, VoiceHome - 2 includes everyday noise sources (with no annotations regarding their SNRs) such as competing talkers, TV/radio, footsteps, doors, kitchenware, and electrical appliances. Five speaker positions per room, comprising standing and sitting postures, are selected to encompass

a broad range of angles and distances concerning the microphone array, which maintains a single, fixed position throughout all the room recordings. The sound is then captured by a microphone array consisting of eight micro-electromechanical systems (MEMS) placed near the corner of a cubic baffle. For this study, only the first channel has been extracted. In total, VoiceHome - 2 encompasses 752 audio recordings, lasting approximately 10 seconds for all the twelve rooms and the five noise scenes. The dataset is then randomly split using a percentage ratio of 70-10-20 training, validation, and testing splits, respectively, for the experiments.

- *STARSS22 [43]:* The dataset includes recordings of human interaction scenes with spatio-temporal event annotations for thirteen target classes, primarily focusing on speech. It is part of the DCASE Challenge 2022 Task 3 development set. The recordings were made at two sites, Tampere University in Finland and Sony headquarters in Japan, in a total of eleven rooms maintaining a consistent organization and procedure regarding equipment, recording, and annotations. The dataset utilizes the Eigenmike spherical microphone array, offering two spatial formats. One format involves a tetrahedral sub-array of omnidirectional microphones mounted on a rigid spherical baffle. The corpus is more challenging compared to the other datasets due to the natural movement and orientation of multiple speakers during discussions, as well as the presence of intentional and unintentional sound events other than speech. It also contains diffuse and directional ambient noise at significant levels. Finally, audio data from a single microphone of the Eigenmike array has been processed, extracting 2934 two-second single-speech excerpts that do not overlap with other annotated directional sources. As done before with the other datasets, STARSS22 is split using a percentage ratio of 70-10-20 training, validation, and testing splits, respectively.

It is worth noticing that, as can be inspected in Fig. 4, real dataset distances are differently distributed with respect to the synthetic and hybrid ones. The motivations of this behavior are as follows:

- in many real-world scenarios, as in STARSS23 [44], sound sources are not always at a fixed distance from the recording device;
- different recording environments can introduce variations in the speaker distance distribution. For example, in a controlled studio setting, speakers may be positioned at specific distances from the microphone to achieve desired sound characteristics. In contrast, field recordings or recordings made in everyday settings can have a wider range of distances due to the uncontrollable nature of the environment. Indeed, in this context, VoiceHome-2 [42] has been recorded in a domestic environment whereas STARSS23 [43] has been collected in office-like environments;
- audio datasets are often curated to suit specific applications or scenarios. For instance, a dataset focused on speaker recognition in far-field scenarios may deliberately include more examples with distant speakers to simulate real-world challenges. On the other hand, a dataset for speech enhancement in close-proximity situations may prioritize examples with close speaker distances. VoiceHome - 2 has been curately designed for enhancing distant-microphone speech whereas STARSS23 focuses on SELD, yielding dissimilar distance distributions.

Accordingly with the distributions of distances in real scenarios, the distance bins used are $\{[1,2),[2,3),[3,4.5)\}$ and $\{[1,2),[2,2.5),[2.5,3)\}$ meters for VoiceHome - 2 and STARSS22, respectively. The final MAE errors are averaged using a percentage ratio of 70-10-20 training, validation, and testing splits, respectively.

V. Experimental Results

In this section, the experimental results are shown for each realistic scenario, as detailed in Section IV. First, the proposed architecture is tested on the synthetic dataset, both in noiseless and noisy scenarios, for the selection of hyperparameters. Next, the performance of the approach is evaluated on hybrid and real recordings by comparing the selected solution with different hyperparameters. Finally, an ablation study is provided to demonstrate the effectiveness of the attention module in all scenarios.

A. Implementation Details

For both training and fine-tuning procedures on all scenarios, the model is trained for 60 epochs at a learning rate of 0.001 with batch size of 16 samples. A scheduled reduction (80%) of the learning rate is performed every 5 epochs when the MSE of the validation set does not improve. In this work, fine-tuning is carried out by training again the model, hence without the random initialization of the weights.

B. Results on Noiseless Synthetic Data

The proposed approach efficiently estimates speaker distance with an average error of 11 cm in a noiseless scenario, as it can be inspected from Table I. Since there is no other published method that attempts regression-based SDE with a single microphone, for comparison purposes we present results on binaural SDE following the recently published work of [45]. The binaural estimation model is similar to the CRNN model used herein; however, we modify it to include the attention operation proposed in this work for better comparison purposes. A similar simulator, range of acoustic conditions, and number of rooms was used in [45] as herein. The same spectrogram and binaural features are also used as in the original work. The binaural estimation results (86 cm) we obtain are, on average, better than the ones in [45] (151 cm), with the improvement most likely attributed to the use of the attention layers. However, the most striking difference is that of the monophonic omnidirectional results (11 cm) versus the binaural ones (86 cm). It seems that the complex frequency-, direction-, and orientation-dependent effects imposed by head-related transfer functions (HRTFs) make it harder for the model to associate spectrotemporal reverberation

TABLE I
HYPERPARAMETERS SELECTION ON THE SYNTHETIC DATASET WITH CLEAN SPEECH

Kernels	# params	# GRUs	Average		$[1,2)$		$[2,4)$		$[4,8)$		$[8,14)$	
			$\mathcal{L}_1$	r$\mathcal{L}_1$	$\mathcal{L}_1$	r$\mathcal{L}_1$	$\mathcal{L}_1$	r$\mathcal{L}_1$	$\mathcal{L}_1$	r$\mathcal{L}_1$	$\mathcal{L}_1$	r$\mathcal{L}_1$
Binaural [46]	650 k	2	0.86 ± 0.10	0.29 ± 0.05	1.06 ± 0.35	0.72 ± 0.22	0.70 ± 0.13	0.25 ± 0.05	0.81 ± 0.10	0.15 ± 0.02	1.34 ± 0.61	0.13 ± 0.05
Time	123 k	0	0.55 ± 0.02	0.18 ± 0.01	0.50 ± 0.04	0.35 ± 0.03	0.50 ± 0.03	0.18 ± 0.01	0.57 ± 0.03	0.11 ± 0.01	0.79 ± 0.09	0.08 ± 0.01
Squared	149 k	0	0.70 ± 0.02	0.23 ± 0.01	0.59 ± 0.04	0.42 ± 0.03	0.68 ± 0.04	0.24 ± 0.01	0.71 ± 0.04	0.13 ± 0.01	1.03 ± 0.12	0.11 ± 0.01
Frequency	123 k	0	0.86 ± 0.03	0.30 ± 0.01	0.83 ± 0.05	0.60 ± 0.04	0.80 ± 0.04	0.28 ± 0.02	0.86 ± 0.04	0.16 ± 0.01	1.17 ± 0.14	0.12 ± 0.01
Time	353 k	1	0.16 ± 0.01	0.05 ± 0.00	0.15 ± 0.01	0.11 ± 0.01	0.13 ± 0.01	0.05 ± 0.00	0.19 ± 0.01	0.03 ± 0.00	0.27 ± 0.03	0.03 ± 0.00
Squared	379 k	1	0.15 ± 0.01	0.05 ± 0.00	0.13 ± 0.01	0.09 ± 0.01	0.11 ± 0.01	0.04 ± 0.00	0.16 ± 0.01	0.03 ± 0.00	0.27 ± 0.04	0.03 ± 0.00
Frequency	353 k	1	0.13 ± 0.01	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.10 ± 0.01	0.04 ± 0.00	0.13 ± 0.01	$\mathbf{0.02 \pm 0.00}$	0.24 ± 0.04	$\mathbf{0.02 \pm 0.00}$
Time	650 k	2	0.13 ± 0.01	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.12 \pm 0.01}$	0.09 ± 0.01	0.10 ± 0.01	0.04 ± 0.00	0.13 ± 0.01	$\mathbf{0.02 \pm 0.00}$	0.24 ± 0.07	$\mathbf{0.02 \pm 0.01}$
Squared	676 k	2	$\mathbf{0.11 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.09 \pm 0.00}$	$\mathbf{0.03 \pm 0.00}$	0.12 ± 0.01	$\mathbf{0.02 \pm 0.00}$	0.18 ± 0.03	$\mathbf{0.02 \pm 0.00}$
Frequency	650 k	2	$\mathbf{0.11 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.10 ± 0.00	$\mathbf{0.03 \pm 0.00}$	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.16 \pm 0.02}$	$\mathbf{0.02 \pm 0.00}$

The gray row highlights the proposed approach.

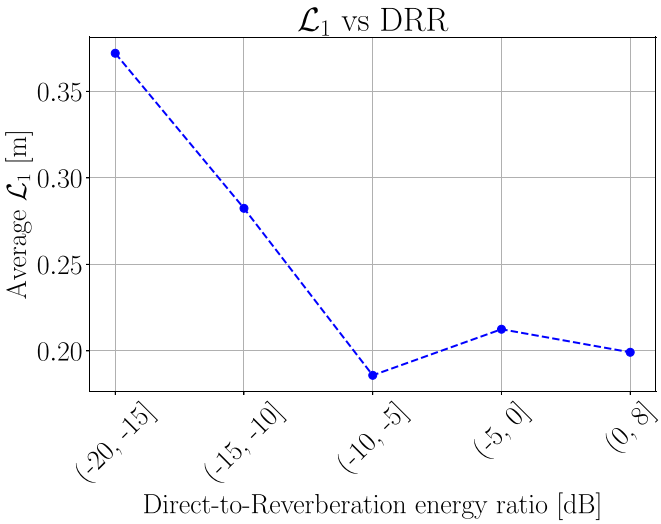


Fig. 5. Relation between DRR and $\mathcal{L}_1$.

patterns with the source distance. However, a definite conclusion on differences between single-channel omnidirectional versus binaural SDE requires further study.

An increasing trend of the errors with respect to the distance is notable. This behavior is expected due to the dominant influence of the late reverberant component compared to the direct and early reflection components of the signal at long distances. These late reverberation cues exhibit statistical diffusion [46], meaning that short-term magnitudes and phases resemble noise-like characteristics. Consequently, extracting meaningful information from these dominant late reverberation cues may pose challenges for the model in effectively estimating speaker distance.

Such behaviour is demonstrated in Fig. 5. Considering that the balance between direct speech energy versus early and late reverberant energy is exemplified in the DRR, measured from the simulated RIRs, it is clear that dominance of the reverberation at low DRRs impacts negatively distance estimation. There seems to be an optimum balance where both direct sound and reverberation contribute to estimation, after which direct sound can start to mask reverberation-related cues for higher DRRs, with a subsequent small drop in performance. A closer investigation of distance estimation at very high DRRs or very small distances at the near-field of the microphone is left for future work.

Moreover, the results of the study demonstrate that the GRU layers play a crucial role in the model's performance. The GRU layers likely contribute to the model's ability to capture sequential patterns and dependencies effectively. Additionally, the study found that using rectangular kernels, as opposed to square kernels, in combination with GRU layers improves the model's efficiency. In this scenario, the rectangular kernels are better at capturing different types of patterns and features in the data, leading to more effective and efficient information processing within the model. This statement, however, does not hold when no GRU layers are present.

In addition, it is worth noting that using a single GRU layer slightly impacts the overall performance of the proposed approach, approximately halving the number of learnable parameters.

TABLE II
EXPERIMENTAL RESULTS ON NOISY SYNTHETIC DATA WITH FIXED SNR AND FREQUENCY KERNELS

SNR	Feature set	$\mathcal{L}_1$	r$\mathcal{L}_1$
50 dB	w/|STFT|	0.48 ± 0.02	0.14 ± 0.01
	w/sinus and cosinus	$\mathbf{0.37 \pm 0.02}$	$\mathbf{0.11 \pm 0.01}$
	|STFT| + sinus and cosinus	0.41 ± 0.02	0.12 ± 0.00
40 dB	w/|STFT|	0.77 ± 0.03	$\mathbf{0.21 \pm 0.01}$
	w/sinus and cosinus	$\mathbf{0.71 \pm 0.03}$	$\mathbf{0.21 \pm 0.01}$
	|STFT| + sinus and cosinus	0.87 ± 0.04	0.24 ± 0.01
30 dB	w/|STFT|	$\mathbf{1.11 \pm 0.04}$	$\mathbf{0.30 \pm 0.01}$
	w/sinus and cosinus	1.51 ± 0.06	0.45 ± 0.02
	|STFT| + sinus and cosinus	1.14 ± 0.04	0.31 ± 0.01
20 dB	w/|STFT|	$\mathbf{1.20 \pm 0.04}$	$\mathbf{0.33 \pm 0.01}$
	w/sinus and cosinus	1.76 ± 0.06	0.56 ± 0.02
	|STFT| + sinus and cosinus	1.21 ± 0.05	$\mathbf{0.33 \pm 0.01}$
10 dB	w/|STFT|	1.30 ± 0.05	0.36 ± 0.01
	w/sinus and cosinus	1.70 ± 0.06	0.56 ± 0.02
	|STFT| + sinus and cosinus	$\mathbf{1.27 \pm 0.05}$	$\mathbf{0.35 \pm 0.01}$
5 dB	w/|STFT|	1.34 ± 0.05	0.38 ± 0.01
	w/sinus and cosinus	1.73 ± 0.06	0.58 ± 0.02
	|STFT| + sinus and cosinus	$\mathbf{1.26 \pm 0.05}$	$\mathbf{0.34 \pm 0.01}$
0 dB	w/|STFT|	1.47 ± 0.05	0.44 ± 0.02
	w/sinus and cosinus	1.77 ± 0.06	0.61 ± 0.02
	|STFT| + sinus and cosinus	$\mathbf{1.39 \pm 0.05}$	$\mathbf{0.42 \pm 0.02}$

The gray row highlights the proposed approach.

C. Analysis of the Impact of Noise on Synthetic Data

To assess the quality of the predictions in relation to noise strength, seven SNR values have been specifically chosen during training. More precisely, a separate model is trained from scratch for each SNR level. Table II depicts the results where a notable discrepancy between the noiseless and noisy scenarios becomes evident. This divergence is primarily attributed to the disruptive influence of background noise on the phase information [25],

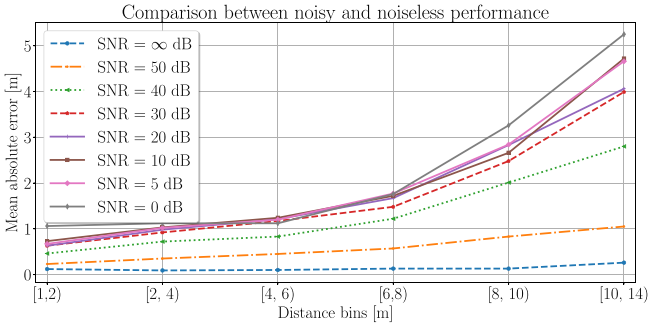


Fig. 6. Comparison between noisy and noiseless performance of the proposed approach on the synthetic dataset.

which has been also demonstrated in speech enhancement studies [47]. It is worth noting from Fig. 6 that the performance of the proposed method remains consistent across all SNR levels for distances up to 6 meters.

However, beyond this distance, the error increases rapidly. This behavior can be attributed to the quadratic inverse relationship between distance and sound intensity, i.e., $I_s \propto \frac{1}{d^2}$. Due to this physical behavior, the direct sound and early distinct echoes exhibit similar energy levels compared to the late reverberant cues, hindering long-distance information.

D. Results on Hybrid Data

As done with the synthetic dataset, five SNR values have been selected to assess the performance of the proposed architecture by training a separate model from scratch for each SNR level. Table III shows the experimental results, highlighting the superiority of the chosen configuration. The notation $[30, +\infty)$ dB denotes the results of the model both in noiseless case and with at most 30 dB of SNR. It is worth noting that, differently from the synthetic scenario, the impact of background noise is smaller even at low SNR. In fact, comparing Table II with Table III, it is evident how synthetic RIRs are more affected by noise at higher SNR with respect to measured ones.

Interestingly, the use of only sinus and cosinus maps yields poor performance at all SNRs levels whereas the STFT magnitude is essential for the task. This result agrees with the previous study [25] where the use of only sinus and cosinus features in noisy audio recordings is ineffective.

E. Results on Real Data

Tables IV and V depict the results on VoiceHome - 2 [42] and STARSS23 [43], respectively. Following the same rationale of the synthetic and hybrid scenarios, the selected configuration outperforms the other models. The results obtained from the analysis of real data demonstrate the clear superiority of the proposed model in accurately estimating distances. Across both datasets, the proposed model consistently outperforms different configurations of the models, showcasing its robustness and effectiveness. However, it is worth noting that a few outliers surfaced in the results, particularly within the VoiceHome - 2 dataset where large confidence intervals are present. This occurrence can be attributed to the limited size of the datasets as the model overfits the training dataset. With a larger dataset, these outliers are expected to be mitigated, and the model's performance is likely to become even more reliable and precise. This observation underscores the potential for further advancement in distance estimation when working with more extensive datasets.

TABLE III
DISTANCE ESTIMATION ERRORS FOR THE QMULTIMIT HYBRID DATASET. GRAY ROW HIGHLIGHTS THE PROPOSED APPROACH

SNR	Hyperparameters	# GRUs	$\mathcal{L}_1$	r$\mathcal{L}_1$
$[30, +\infty)$ dB	*Time*	0	2.49 ± 0.16	0.28 ± 0.02
	Squared	0	2.38 ± 0.15	0.25 ± 0.02
	Frequency	0	2.97 ± 0.17	0.33 ± 0.03
	Time	1	1.58 ± 0.12	0.16 ± 0.01
	Squared	1	1.52 ± 0.12	0.15 ± 0.01
	Frequency	1	1.68 ± 0.12	0.17 ± 0.01
	Time	2	1.70 ± 0.12	0.17 ± 0.01
	Squared	2	$\mathbf{1.48 \pm 0.13}$	$\mathbf{0.14 \pm 0.01}$
	Freq. w/|STFT|	2	1.67 ± 0.13	0.17 ± 0.01
	Freq. w/sinus and cosinus	2	2.17 ± 0.14	0.23 ± 0.02
	Frequency	2	1.52 ± 0.12	0.15 ± 0.01
20 dB	*Time*	0	2.22 ± 0.15	0.24 ± 0.02
	Squared	0	2.36 ± 0.15	0.25 ± 0.02
	Frequency	0	2.88 ± 0.17	0.32 ± 0.02
	Time	1	1.67 ± 0.12	0.16 ± 0.01
	Squared	1	$\mathbf{1.46 \pm 0.12}$	$\mathbf{0.14 \pm 0.01}$
	Frequency	1	1.71 ± 0.12	0.17 ± 0.01
	Time	2	1.66 ± 0.13	0.16 ± 0.01
	Squared	2	1.60 ± 0.13	0.16 ± 0.01
	Freq. w/|STFT|	2	1.64 ± 0.13	0.16 ± 0.01
	Freq. w/sinus and cosinus	2	1.98 ± 0.13	0.21 ± 0.02
	Frequency	2	1.48 ± 0.11	$\mathbf{0.14 \pm 0.01}$
10 dB	*Time*	0	2.23 ± 0.14	0.24 ± 0.02
	Squared	0	2.20 ± 0.14	0.24 ± 0.02
	Frequency	0	2.55 ± 0.14	0.28 ± 0.02
	Time	1	1.71 ± 0.12	0.17 ± 0.01
	Squared	1	1.58 ± 0.13	0.16 ± 0.01
	Frequency	1	1.60 ± 0.12	0.16 ± 0.01
	Time	2	1.65 ± 0.12	0.16 ± 0.01
	Squared	2	1.56 ± 0.13	$\mathbf{0.15 \pm 0.01}$
	Freq. w/|STFT|	2	$\mathbf{1.55 \pm 0.12}$	$\mathbf{0.15 \pm 0.01}$
	Freq. w/sinus and cosinus	2	1.97 ± 0.12	0.21 ± 0.01
	Frequency	2	1.65 ± 0.13	0.17 ± 0.01
0 dB	*Time*	0	2.54 ± 0.14	0.28 ± 0.02
	Squared	0	2.74 ± 0.15	0.30 ± 0.02
	Frequency	0	3.01 ± 0.15	0.33 ± 0.02
	Time	1	1.75 ± 0.12	0.18 ± 0.01
	Squared	1	1.83 ± 0.12	0.19 ± 0.01
	Frequency	1	1.82 ± 0.13	0.19 ± 0.01
	Time	2	2.46 ± 0.15	0.23 ± 0.01
	Squared	2	1.98 ± 0.12	0.21 ± 0.02
	Freq. w/|STFT|	2	$\mathbf{1.63 \pm 0.13}$	$\mathbf{0.17 \pm 0.01}$
	Freq. w/sinus and cosinus	2	2.24 ± 0.13	0.25 ± 0.02
	Frequency	2	1.66 ± 0.13	$\mathbf{0.17 \pm 0.01}$
-10 dB	*Time*	0	3.03 ± 0.14	0.34 ± 0.03
	Squared	0	3.03 ± 0.14	0.33 ± 0.02
	Frequency	0	3.04 ± 0.14	0.33 ± 0.02
	Time	1	3.02 ± 0.14	0.33 ± 0.02
	Squared	1	3.01 ± 0.14	0.33 ± 0.03
	Frequency	1	3.00 ± 0.14	0.33 ± 0.03
	Time	2	3.06 ± 0.14	0.34 ± 0.03
	Squared	2	2.57 ± 0.13	0.28 ± 0.02
	Freq. w/|STFT|	2	$\mathbf{2.28 \pm 0.13}$	$\mathbf{0.25 \pm 0.02}$
	Freq. w/sinus and cosinus	2	3.01 ± 0.14	0.33 ± 0.03
	Frequency	2	2.34 ± 0.13	$\mathbf{0.25 \pm 0.02}$

All features are used if not mentioned.

F. Ablation Study of the Attention Module

To demonstrate the effectiveness of the attention module, an ablation study is performed on all the scenarios. First, performance assessment is carried out without the module. Then, instead of returning a $T \times F \times 3$ matrix, a spectrogram attention map, i.e., $T \times F$, is learned by a module. Then, an element-wise multiplication is performed between the magnitude of the STFT and the attention map.

TABLE IV
DISTANCE ESTIMATION ERRORS FOR THE VOICEHOME - 2 DATASET

Hyperparameters	# GRUs	Average		$[1,2)$		$[2,3)$		$[3,4.5)$	
		$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$
Time	0	0.95 ± 0.10	0.49 ± 0.06	1.00 ± 0.14	0.73 ± 0.11	0.69 ± 0.12	0.28 ± 0.05	1.20 ± 0.25	0.32 ± 0.06
Squared	0	0.90 ± 0.11	0.46 ± 0.07	0.90 ± 0.16	0.69 ± 0.14	0.57 ± 0.11	0.23 ± 0.04	1.34 ± 0.26	0.35 ± 0.07
Frequency	0	0.83 ± 0.09	0.43 ± 0.06	0.85 ± 0.13	0.63 ± 0.11	0.67 ± 0.13	0.27 ± 0.05	1.02 ± 0.20	0.27 ± 0.05
Time	1	0.76 ± 0.09	0.38 ± 0.05	0.73 ± 0.12	0.55 ± 0.10	0.47 ± 0.10	0.19 ± 0.04	1.19 ± 0.23	0.32 ± 0.06
Squared	1	0.74 ± 0.09	0.40 ± 0.07	0.85 ± 0.15	0.65 ± 0.13	0.43 ± 0.09	$\mathbf{0.17 \pm 0.04}$	0.96 ± 0.20	0.26 ± 0.05
Frequency	1	0.74 ± 0.08	0.37 ± 0.05	0.73 ± 0.12	0.54 ± 0.10	0.53 ± 0.10	0.21 ± 0.04	1.06 ± 0.21	0.28 ± 0.05
Time	2	0.64 ± 0.08	$\mathbf{0.31 \pm 0.05}$	$\mathbf{0.59 \pm 0.12}$	$\mathbf{0.44 \pm 0.10}$	0.49 ± 0.09	0.20 ± 0.03	0.94 ± 0.21	0.25 ± 0.05
Squared	2	0.70 ± 0.10	0.35 ± 0.06	0.67 ± 0.14	0.51 ± 0.12	$\mathbf{0.43 \pm 0.12}$	$\mathbf{0.17 \pm 0.05}$	1.11 ± 0.21	0.29 ± 0.05
Freq w /|STFT|	2	0.66 ± 0.08	0.33 ± 0.05	0.63 ± 0.13	0.48 ± 0.11	0.47 ± 0.10	0.19 ± 0.04	0.98 ± 0.17	0.27 ± 0.05
Freq w /sinus and cosinus	2	0.91 ± 0.11	0.46 ± 0.07	0.88 ± 0.14	0.68 ± 0.13	0.52 ± 0.11	0.21 ± 0.04	1.49 ± 0.21	0.40 ± 0.05
Frequency	2	$\mathbf{0.63 \pm 0.08}$	0.32 ± 0.05	0.64 ± 0.11	0.48 ± 0.10	0.48 ± 0.11	0.19 ± 0.04	$\mathbf{0.80 \pm 0.20}$	$\mathbf{0.21 \pm 0.05}$

Gray row highlights the proposed approach. All features are used if not mentioned.

TABLE V
DISTANCE ESTIMATION ERRORS FOR THE STARSS23 DATASET

Hyperparameters	# GRUs	Average		$[1,2)$		$[2,2.5)$		$[2.5,3)$	
		$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$
Time	0	0.51 ± 0.03	0.23 ± 0.01	0.30 ± 0.04	0.16 ± 0.02	0.55 ± 0.03	0.24 ± 0.01	0.76 ± 0.10	0.29 ± 0.04
Square	0	0.50 ± 0.03	0.22 ± 0.01	0.29 ± 0.04	0.16 ± 0.02	0.53 ± 0.03	0.23 ± 0.01	0.85 ± 0.09	0.33 ± 0.03
Frequency	0	0.51 ± 0.03	0.23 ± 0.01	0.35 ± 0.05	0.19 ± 0.03	0.54 ± 0.03	0.24 ± 0.01	0.76 ± 0.10	0.29 ± 0.04
Time	1	0.45 ± 0.02	0.20 ± 0.01	0.26 ± 0.03	$\mathbf{0.14 \pm 0.02}$	0.49 ± 0.03	0.21 ± 0.01	0.70 ± 0.08	0.27 ± 0.03
Square	1	$\mathbf{0.42 \pm 0.02}$	$\mathbf{0.19 \pm 0.01}$	0.33 ± 0.04	0.18 ± 0.02	$\mathbf{0.42 \pm 0.03}$	$\mathbf{0.18 \pm 0.01}$	0.62 ± 0.09	0.24 ± 0.03
Frequency	1	0.46 ± 0.02	0.20 ± 0.01	0.30 ± 0.04	0.16 ± 0.02	0.48 ± 0.03	0.21 ± 0.01	0.69 ± 0.08	0.26 ± 0.03
Time	2	0.46 ± 0.02	0.21 ± 0.01	$\mathbf{0.27 \pm 0.03}$	0.15 ± 0.02	0.49 ± 0.03	0.22 ± 0.01	0.69 ± 0.09	0.26 ± 0.03
Square	2	0.50 ± 0.02	0.22 ± 0.01	0.34 ± 0.04	0.19 ± 0.02	0.51 ± 0.03	0.23 ± 0.01	0.79 ± 0.09	0.30 ± 0.03
Freq w /|STFT|	2	0.46 ± 0.02	0.21 ± 0.01	0.28 ± 0.03	0.15 ± 0.02	0.49 ± 0.03	0.21 ± 0.01	0.71 ± 0.09	0.27 ± 0.03
Freq w /sinus and cosinus	2	0.46 ± 0.02	0.20 ± 0.01	0.28 ± 0.03	0.16 ± 0.02	0.48 ± 0.03	0.21 ± 0.01	0.74 ± 0.09	0.28 ± 0.03
Frequency	2	$\mathbf{0.42 \pm 0.02}$	$\mathbf{0.19 \pm 0.01}$	0.33 ± 0.05	0.18 ± 0.03	0.43 ± 0.03	0.19 ± 0.01	$\mathbf{0.55 \pm 0.09}$	$\mathbf{0.21 \pm 0.04}$

Gray row highlights the proposed approach. All features are used if not mentioned.

TABLE VI
ABLATION STUDY OF ATTENTION MAP USING FREQUENCY KERNELS ON SYNTHETIC DATA WITH CLEAN SPEECH

Attention	Average		$[1,2)$		$[2,4)$		$[4,8)$		$[8,14)$	
	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$
None	0.14 ± 0.01	0.05 ± 0.00	0.13 ± 0.01	0.09 ± 0.01	0.12 ± 0.01	0.04 ± 0.00	0.15 ± 0.01	0.03 ± 0.00	0.28 ± 0.05	0.03 ± 0.00
on spectrogram	0.12 ± 0.00	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	0.04 ± 0.00	0.13 ± 0.01	$\mathbf{0.02 \pm 0.00}$	0.22 ± 0.03	$\mathbf{0.02 \pm 0.00}$
on everything	$\mathbf{0.11 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.10 \pm 0.00}$	$\mathbf{0.03 \pm 0.00}$	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.16 \pm 0.02}$	$\mathbf{0.02 \pm 0.00}$

Gray row highlights the proposed approach.

TABLE VII
ABLATION STUDY OF ATTENTION MAP USING FREQUENCY KERNELS ON HYBRID AND REAL DATA

Attention	QMULTIMIT		VoiceHome - 2		STARSS22	
	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$	$\mathcal{L}_1$	$\mathrm{r}\mathcal{L}_1$
None	2.01 ± 0.06	0.21 ± 0.01	0.78 ± 0.09	0.40 ± 0.06	0.45 ± 0.02	0.20 ± 0.01
on spectrogram	$\mathbf{1.87 \pm 0.06}$	$\mathbf{0.19 \pm 0.01}$	0.73 ± 0.10	0.36 ± 0.06	0.45 ± 0.02	0.20 ± 0.01
on everything	1.90 ± 0.06	0.20 ± 0.01	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.32 \pm 0.05}$	$\mathbf{0.42 \pm 0.02}$	$\mathbf{0.19 \pm 0.01}$

Gray row highlights the proposed approach.

TABLE VIII
CROSS-DATASET GENERALIZATION TESTS WITHOUT FINETUNING

		Test w/o finetuning		
		Synthetic	**Hybrid**	**Real**
Training	**Synthetic**	0.11 ± 0.00	4.28 ± 0.45	4.14 ± 0.08
	Hybrid	6.80 ± 0.59	1.52 ± 0.12	3.76 ± 0.56
	Real	2.26 ± 0.38	8.22 ± 0.54	0.42 ± 0.02

These three modalities are analyzed in Table VI, depicting the errors for each bin with their confidence intervals. Predicting an attention map for each feature provides better distance estimation on average. Moreover, the results demonstrate that all the approaches perform similarly in the short range, up to 8 meters. Conversely, applying the attention map on each of the feature maps in the feature set produces better outcomes in the long range with respect to the other two cases. When the speaker is far from the microphone, the learned attention maps enhance the features set, facilitating the extraction of features of the convolutional layers. Indeed, as the distance between the speaker and the microphone increases, detecting these patterns becomes more challenging due to their reduced salience [46].

Moreover, an ablation study has been carried out also on the hybrid and real data, as it can be inspected in Table VII. The attention map yields the best performance in the hybrid case when it is only applied to the STFT magnitude channel. This fact highlights the ineffectiveness of phase features in this specific use case. Instead, the results demonstrate the superiority of the attention map applied on all the channels in the real scenario.

G. Cross-Corpus Generalization

Tests have been carried out in a cross-corpus training-testing setup, e.g., synthetic-hybrid, synthetic-real, hybrid-real, VoiceHome-STARSS. The model yields very large errors in case no finetuning is performed, as it can be inspected in Table VIII. This behavior highlights the discrepancy of feature patterns among different acoustic scenarios, levels of acoustical realism, and different distance distributions. If the model is fine-tuned to a different realistic scenario, the performance is slightly worse

TABLE IX
CROSS-DATASET GENERALIZATION TESTS WITH FINETUNING

Training	Test w/ finetuning: Synthetic	Hybrid	Real
Synthetic	0.11 ± 0.00	1.57 ± 0.23	0.47 ± 0.05
Hybrid	0.18 ± 0.04	1.52 ± 0.12	0.45 ± 0.05
Real	0.11 ± 0.02	1.54 ± 0.22	0.42 ± 0.02

that the case when the model starts with random weights. The results of this situation is shown in Table IX.

VI. DISCUSSION

From the results of the noisy scenario in the synthetic dataset, it is important to highlight that even a minimal amount of noise severely corrupts phase-based features, which have been identified as the most critical information in our analysis of clean speech. For instance, the presence of direct sound and echo patterns, characterized by transients in the clean signal, becomes blurred over time due to the presence of noise and late reverberation, resulting in a loss of phase coherence across frequencies. This behavior, however, does not occur in the hybrid dataset where the effect of high SNR in the recordings does not correspond to a similar increase in estimation performance. That may be due to the recordings of the RIRs having a level of inherent measurement noise, which limits the effective SNR that we can achieve in the hybrid simulations.

The imposition of the loss in (3) is required for predicting a time-wise distance vector. Due to the lack of baselines and datasets in the literature, only a single value of distance of the sound source is assigned for each time bin to ease the distance tracking task. Generally, this characteristic in audio datasets is referred as *weak labels* [48]. Without time-wise distance references, denoted as *strong labels*, the model encounters challenges in fine-tuning its predictions, decreasing its overall performance. This scenario has been studied in literature for tasks that require a fine temporal resolution output, such as sound event detection (SED) [49] and SELD [50].

Furthermore, it is important to acknowledge that certain portions of the audio data encompass segments where speech information is absent or indiscernible. Consequently, this scarcity of informative speech content can considerably undermine the effectiveness and reliability of the predictors.

In this direction, the proposed attention module can improve the ability of the model (Table VII) to identify the speech information that is relevant for the estimation of the distance. However, it is important to note that the attention module is learned by the model itself, without any direct supervision.

To address these limitations, a potential avenue for improvement emerges, centering around the generation of more comprehensive and fine-grained labels. By augmenting the dataset with *strong labels* that introduces both speech activity and speaker distance estimation, the model may acquire a better understanding of the room acoustics. In addition, this augmentation enables the model to leverage additional contextual cues and refine its predictions, enhancing its performance in accurately estimating speaker distances and capturing the dynamics of speech activity.

Moreover, one of the key areas for improvement is the availability of larger datasets of real recordings with a greater number of rooms and various speaker-microphone configurations. A larger dataset would enable the model to learn more diverse and representative acoustic characteristics, leading to improved performance in distance estimation tasks. Moreover, it could also improve the generalization ability of the approach, as it has been demonstrated how the performance of the proposed model is dependent on the nature of the audio recording (synthetic, hybrid or real). Additionally, by including different room types and microphone placements, the model can better generalize across various real-world scenarios. Furthermore, the use of a transformer-based [51] approach could be explored, leveraging a larger amount of data. Transformer models have shown remarkable success in various natural language processing tasks and have the potential to capture complex patterns and dependencies in acoustic data. Exploiting transformer architectures could enhance the model's ability to estimate distances accurately.

Another possibility for future research is the integration of time-wise distance ground truth, as previously mentioned in the discussion section. By considering temporal information in addition to spatial cues, the model could potentially estimate the distance of a sound source more accurately. This would provide valuable insights in scenarios where multiple sound sources are present. Estimating and tracking the distance of a moving source is an application of interest that is scarcely explored in the literature.

VII. CONCLUSION

This work has explored the task of speaker distance estimation in noisy and reverberant environments. Multiple configurations, in terms of kernel size and recurrent layers of the model, have been provided, motivating the proposed architecture. In fact, the use of rectangular filters across the frequency dimension and the presence of GRUs layers yields the best performance in terms of distance errors. The experimental results obtained from the proposed model have demonstrated remarkable precision in scenarios where several types of RIRs are employed. In a noiseless synthetic scenario where RIRs have been generated with a room-source simulator, the model has achieved an absolute error of only 0.11 meters. With recorded RIRs, an absolute error of about 1.30 meters has been obtained. In the real scenario with on-field recordings, where unpredictable environmental factors and noise were prevalent, the model yielded an absolute error of approximately 0.50 meters. These results underscore the model's resilience and its capacity to effectively manage various realistic scenarios. Variations in performance across these scenarios can be attributed to differences in the distribution of acoustic parameters, such as the distance from the sound source. Analysis on moving sound sources in single-channel recordings will be carried out as a future work.

REFERENCES

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

[2] M. Bekrani, A. W. H. Khong, and M. Lotfizad, "A linear neural network-based approach to stereophonic acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1743–1753, Aug. 2011.

[3] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 653–658.

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

[5] 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 Acoustic Signal Enhancement*, 2018, pp. 1–5.

[6] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Division Eng., Brown Univ., Providence, RI, USA, 2000.

[7] 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.

[8] 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. Audio Acoust.*, 2021, pp. 331–335.

[9] K. Patterson, K. Wilson, S. Wisdom, and J. R. Hershey, "Distance-based sound separation," in *Proc. Interspeech*, 2022, pp. 901–905.

[10] 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.

[11] E. Georganti, T. May, S. van de Par, A. Harma, and J. Mourjopoulos, "Speaker distance detection using a single microphone," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1949–1961, Sep. 2011.

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

[13] 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. Engineers Isr.*, 2012, pp. 1–5.

[14] 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.

[15] J. Hwang, S. Seon, and C. Park, "Position estimation of sound source using three optical Mach-Zehnder acoustic sensor array," *Curr. Opt. Photon.*, vol. 1, no. 6, pp. 573–578, 2017.

[16] L. Ghamdan, M. A. I. Shoman, R. Abd Elwahab, and N. A. E. Ghamry, "Position estimation of binaural sound source in reverberant environments," *Egyptian Inform. J.*, vol. 18, no. 2, pp. 87–93, 2017.

[17] 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.

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

[19] S. Vesa, "Sound source distance learning based on binaural signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2007, pp. 271–274.

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

[21] K. Zhagyparova, R. Zhagypar, A. Zollanvari, and M. T. Akhtar, "Supervised learning-based sound source distance estimation using multivariate features," in *Proc. IEEE Region 10 Symp.*, 2021, pp. 1–5.

[22] M. Yiwere and E. J. Rhee, "Sound source distance estimation using deep learning: An image classification approach," *Sensors*, vol. 20, no. 1, 2019, Art. no. 172.

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

[24] R. Venkatesan and A. B. Ganesh, "Analysis of monaural and binaural statistical properties for the estimation of distance of a target speaker," *Circuits, Syst., Signal Process.*, vol. 39, pp. 3626–3651, 2020.

[25] M. Neri, A. Politis, D. Krause, M. Carli, and T. Virtanen, "Speaker distance estimation from single channel audio in reverberant environments," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

[26] 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.

[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. Workshop Detection Classification Acoustic Scenes Events*, 2019, pp. 20–24.

[28] D. Byrne, "The speech spectrum-some aspects of its significance for hearing aid selection and evaluation," *Brit. J. Audiol.*, vol. 11, no. 2, pp. 40–46, 1977.

[29] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Multi-channel environmental sound segmentation utilizing sound source localization and separation U-net," in *Proc. IEEE/SICE Int. Symp. Syst. Integration*, 2021, pp. 382–387.

[30] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 508–519, 2022.

[31] Z. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[32] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6885–6889.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2015, pp. 448–456.

[34] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–14.

[35] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 92–102, Apr. 2018.

[36] 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. 30th Eur. Signal Process. Conf.*, 2022, pp. 36–40.

[37] J. Garofolo, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[38] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Sch. Elect. Eng., Aalto Univ., Espoo, Finland, 2016.

[39] "Sound absorption coefficient chart: JCW acoustic supplies," Accessed: Jun. 17, 2023. [Online]. Available: https://www.acoustic-supplies.com/absorption-coefficient-chart

[40] G. Wichern et al., "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[41] R. Stewart and M. Sandler, "Database of omnidirectional and b-format room impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 165–168.

[42] N. Bertin et al., "VoiceHome-2, an extended corpus for multichannel speech processing in real homes," *Speech Commun.*, vol. 106, pp. 68–78, 2019.

[43] A. Politis et al., "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. 7th Detection Classification Acoustic Scenes Events Workshop*, 2022, pp. 1–5.

[44] K. Shimada et al., "STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances Neural Inf. Process. Syst.*, vol. 36, 2024.

[45] D. A. Krause, G. García-Barrios, A. Politis, and A. Mesaros, "Binaural sound source distance estimation and localization for a moving listener," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 996–1011, 2024.

[46] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 204–210, 2000.

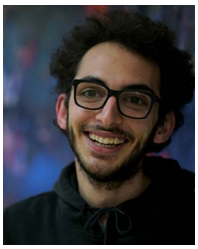
[47] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[48] A. Kumar and B. Raj, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2017, pp. 3475–3482.

[49] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound event detection for human safety and security in noisy environments," *IEEE Access*, vol. 10, pp. 134230–134240, 2022.

[50] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 902–914, 2023.

[51] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.



Michael Neri (Graduate Student Member, IEEE) received the Laurea (B.Sc.) and Laurea Magistrale (M.Sc.) degrees in ICT for Internet and multimedia from the University of Padova, Padua, Italy, in 2019 and 2021, respectively. In 2023, he was a visiting Ph.D. student with the Tampere University of Technology, Tampere, Finland, under the supervision of Prof. Virtanen. He is currently working toward the Ph.D. degree in applied electronics with the Department of Industrial, Electronic and Mechanical Engineering, Roma Tre University, Rome, Italy. His main research interests include computer vision, deep learning, and audio processing.

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

Daniel Aleksander Krause (Graduate Student Member, IEEE) received the bachelor's and master's degrees in acoustical engineering from the AGH University of Science and Technology, Kraków, Poland, in 2018 and 2019, respectively. From 2017 to 2018, he was a Data Scientist with Fitech. From 2019 to 2020, he was with Signal Processing Group, the Department of Electronics, AGH University of Science and Technology. Since 2020, he has been pursuing his doctoral studies, contributing as a Member for both Audio Research Group and Machine Listening group. In 2021, he was the co-organizer for DCASE scientific challenge. He is currently a Doctoral Researcher with Tampere University. His research interests include data science, signal processing, machine learning, spatial audio, and acoustic scene analysis.

Marco Carli (Senior Member, IEEE) received the Laurea degree in telecommunication engineering from the Università degli Studi di Roma 'La Sapienza', Rome, Italy, and the Ph.D. degree from the Tampere University of Technology, Tampere, Finland. He is currently a Full Professor with the Department of Industrial, Electronic, and Mechanical Engineering, Università degli Studi 'Roma TRE', Rome. His research interests include digital signal and image processing with applications to multimedia communications. He is an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and Area Editor for Elsevier *Signal Processing: Image Communication*.

Tuomas Virtanen (Fellow, IEEE) received the M.Sc. and Doctor of Science degrees in information technology from the Tampere University of Technology, Tampere, Finland, in 2001 and 2006, respectively. He is currently a Professor with Tampere University, where he is Leading the Audio Research Group. He is also a Research Associate with the Department of Engineering, Cambridge University, Cambridge, U.K. He is known for his pioneering work on computational acoustic scene analysis and sound source separation. He has authored or coauthored more than 200 scientific publications in his research areas, which have been cited more than 19 000 times. His research interests include machine listening, computational content analysis of audio, and machine learning for audio. He is an IEEE Signal Processing Society Distinguished Lecturer during 2024–2025, and Member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society. He was the recipient of ERC 2014 Starting Grant, IEEE Signal Processing Society Best Paper Awards multiple times, and many other best paper awards.

Open Access funding provided by 'Università degli Studi Roma Tre' within the CRUI CARE Agreement