POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Direction of Arrival Estimation using Convolutional Recurrent Neural Network with Relative Harmonic Coefficients and Triplet Loss in Noisy and Reverberating Environments

Tesi di Laurea Magistrale in
Music and Acoustic Engineering

Luca Cattaneo, 964214

Abstract: The problem of source localization in noisy and reverberating environments is still an open and challenging problem in the signal processing field. Typically, the identification of the so-called direction of arrival (DOA) concerns the estimation of the position of acoustic sources from a multichannel recording. The localization of a sound source can be fundamental in various applications, such as speech and speaker recognition, audio surveillance, and virtual and augmented reality. Recent model-based approaches try to overcome this problem using a spherical harmonics domain (SHD) source feature named relative harmonics coefficients (RHC). Other solutions, use deep learning techniques to address DOA estimation by learning features through artificial networks. In this work, we propose a new method for DOA classification exploring the convolutional recurrent neural network (CRNN) with RHC as input features. In order to classify simultaneously the azimuth and the elevation, the final section of the proposed CRNN is composed of two independent fully connected (FC) networks. Then, we present a siamese neural network trained with the technique known as triplet loss. The main advantage of the proposed training technique is that the network learns a structured feature representation that organizes samples from the same class closer to each other while keeping samples from different classes apart. We demonstrated that the use of triplet loss training to obtain feature embeddings results in a good DOA estimation performance on simulations at various signal-to-noise (SNR) ratios and reverberation time RT_{60}. For the evaluation of the proposed method, we considered the gross error (GE), the localization error (LR), and the mean absolute estimated error (MAEE/°). Experiments confirm that the triplet loss approach produces a more structured and meaningful features embedding, implying superior features space interpretability. Finally, the DOA estimation performance of the proposed approach is compared with conventional subspace methods, demonstrating a more robust performance in noisy and reverberant acoustic scenarios, and higher localization accuracy.

Advisor:
Prof. Fabio Antonacci

Co-advisors:
Mirco Pezzoli

Academic year:
2021-2022

Key-words: direction of arrival estimation, deep learning, relative harmonic coefficients, triplet loss

1. Introduction

Nowadays the number of applications that are based on complex audio systems is increasingly spreading and the capacity to provide high audio quality is more and more significant. The interaction with a device with internal computing capabilities, such as computers or smart devices, is moving towards higher quality. In this context, the ability to extract the direction of arrival (DOA) from multi-channel recordings can be useful information for enhancing the user experience and improving the performance of audio tools [10, 32, 54].
For example, in recent years smart speakers and voice assistants like Amazon Alexa and Google Home have become increasingly popular. These applications exploit natural language processing to recognize spoken commands, allowing the users to interact with the device through their voice. In this scenario, the DOA estimation can be useful to enable the system to identify the direction from which a user is speaking to the device. This information can be used to improve the accuracy of automatic speech recognition [12, 84], improving the ability of the assistant to interpret the commands.
As smart working and virtual meetings continue to gain popularity, teleconferencing systems have become a vital aspect of both personal and professional communication, particularly in the realm of audio technology. In this field, the ability of the system to reproduce audio and video with good quality is crucial for communication. In the aforementioned fields, a microphone array can easily be employed to capture the voice of the user. Then, the system typically exploits the recorded data to recognize the spatial location of a speaker and use this information to improve speech enhancement [30, 80] and speech separation [16, 28] techniques. Moreover, the DOA information can be exploited to facilitate efficient noise reduction techniques [73], aimed at improving the clarity of voice and audio quality. These techniques improve the performance of the algorithms, resulting in efficient communication between the user and the system. Furthermore, the detection of the DOA of a signal source can be crucial also in surveillance systems for identifying the location of a target [15] and tracking their movement, exploiting the multi-channel audio signal. The extraction of the location from audio signals can be performed also on targets that are out of sight, which is a major advantage with respect to image-based techniques [35, 55].

The DOA estimation is a fundamental problem in acoustic signal processing and although it is a long-standing and widely researched topic, it remains a challenging problem to solve. Conventional approaches rely on signal processing techniques making assumptions about the statistics of the signal. Early localization methods, such as the generalized cross-correlation phase transform (GCC-PHAT) [51], rely on determining the time difference of arrival (TDOA) between pairs of sensors. The time delay can then be used to calculate the difference in the distance between the sound source and each microphone and extract the location of the sound source. Another class of popular approaches for source localization are the beamformer-like techniques, such as the steered response power (SRP) [82], and its variant SRP-phase transformed (SRP-PHAT) [20, 23], where the output power of a beamformer is scanned in all possible directions to find out when it reaches the maximum, which corresponds to the source location. One of the most popular solutions due to their simple implementation and reasonable performance are subspace methods [48, 69], whose most popular approach is multiple signal classification (MUSIC) [67, 69, 75].
In recent years, the availability of different array geometries with a high number of microphones raised the adoption of different signal transformations [57, 79]. Spherical harmonics decomposition [63] is one of the most popular sound field representations. As a matter of fact, signals transformed into the spherical harmonics domain (SHD) have been applied to source localization techniques, such as SHD-MUSIC [5]. However, these techniques are susceptible to degraded performance in scenarios with low signal-to-noise (SNR) ratios and reverberation. In [41], the relative harmonic coefficients (RHC) were proposed, which are based on the idea of the relative transfer function (RTF) to provide DOA information that is not influenced by the source signal and can effectively handle noise interference. Inspired by the MUSIC algorithm, the authors of [39] introduced RHC in subspace methods (SHD-RMUSIC) showing improved performance with respect to traditional methods.
Over the past decade, researches have increasingly turned to machine learning to solve a wide range of practical problems [22, 25, 36], including DOA estimation [13, 26, 61]. Data-driven approaches have the advantage of being able to be trained over different acoustic environments and source distributions. In [70], Takeda *et al.* employed the eigenvectors of a MUSIC inspired spatial correlation matrix to train a deep neural network for localizing acoustic sources. A method based on the convolutional recurrent neural network (CRNN) was proposed in [8]. The authors exploit both the magnitude and phase information of the STFT coefficients to train the model and perform joint sound event detection and localization. In [29], Fahim *et al.* proposed a convolutional neural network (CNN) based algorithm which learns the modal coherence patterns from measured spherical harmonics coefficients (SHC). The development of deep learning (DL) techniques to address the source localization problem has followed the broader trend in the DL and signal processing communities towards increasingly complex architectures and novel efficient models [38, 77].

In this work, we propose a DL framework based on the RHC feature, which has demonstrated to be effec-

tive for DOA estimation [41, 42, 44]. To train the models, we employ features based on measured SHC. The input features are composed by SHC in the STFT domain and the estimated RHC, represented by their real and imaginary part. Regarding the DL model, we propose a CRNN-based framework to estimate the DOAs of sound sources in different acoustic scenarios. The convolutional layer provides improved classification by better pattern recognition, while the recurrent layer learns long-temporal information in the input audio signal. This motivated to use the CRNN structure, which combines the advantages of CNN and recurrent neural network (RNN) for efficient DOA estimation. For azimuth and elevation classification, two independent fully connected (FC) networks compose the last part of the proposed architecture. Then, we employ a CNN-based siamese network trained with triplet loss. The siamese network consists of three identical CNNs that take three different input signals and generate a feature vector for each input. Triplet loss is a technique commonly used in computer vision [76] to improve the accuracy of image recognition tasks [17, 68] by optimizing the distance between images in a learned features space. More recently, this approach has also been applied to audio processing tasks [46, 74]. In the context of DOA estimation, the use of triplet loss can be advantageous because it encourages the neural network to learn a structured feature representation where samples of different DOAs are separated and instances of similar DOAs are close, being beneficial for DOA estimation and feature embedding coherence with the spatial domain.
We designed the algorithm to perform DOA estimation while being trained on a fixed room and tested on rooms with different reverberation times and dimensions. We show that the proposed framework is able to generalize the information contained in the input features. Consequently, the network can estimate the source DOA correctly in unseen rooms with different acoustic characteristics. Furthermore, we proved that with few samples for each class, the siamese network can design an embedding where DOA classes are clustered and more separated compared to the other models, demonstrating a beneficial effect on the intelligibility of the features space. Finally, we employ the trained CNN structure of the siamese framework as the pre-trained model in the CRNN-based joint estimation showing that from the embeddings designed with triplet loss, we still obtain good performance in DOA estimation. Therefore, the proposed method consists of three main steps: (1) training of the proposed CRNN architecture with free embeddings design; (2) siamese network training with triplet loss; (3) the transfer learning from the framework trained in step (2) is performed.
For the evaluation of the triplet loss effectiveness, we conducted a comparative analysis of features embedding generated by a free design embedding network, a siamese network trained with triplet loss, and a triplet loss pre-trained network. The analysis of the learned features embedding revealed that the triplet loss training produces a features space that is more easily interpretable, without compromising the DOA estimation performance. Then, we compared the performance of the proposed method with the conventional subspace methods, namely MUSIC and SHD-MUSIC. Results show a more robust performance in high reverberation and low SNR environments, outperforming MUSIC and SHD-MUSIC.

The thesis is organized as follows. In Sec. 2 we provide an overview of the literature for DOA estimation techniques, presenting separately model-based methods and data-driven techniques. Moreover, we include a more detailed description of subspace methods MUSIC, RMUSIC, SHD-MUSIC, and SHD-RMUSIC. Instead, for data-driven methods we provide an exhaustive explanation of the CRNN architecture, which is the main framework of this work, and triplet loss. Furthermore, we provide a comprehensive description of the work proposed by Fahim *et al.* in [29]. In Sec. 3 we present the main methods employed in this work. We start by presenting the RHC estimator proposed in [42]. We continue presenting the proposed architecture exploited for the training of the free embedding design network and the pre-trained model. Furthermore, the siamese network architecture is described together with the triplet loss training procedure. In Sec. 4 we provide the implementation details of the proposed method. We present the details of the simulated dataset generated for this work. Then, we provide an overview of the training configuration and of the metrics employed in order to evaluate the performance of the frameworks. We comment on the results that corroborate the proposed method. Our evaluation of the results supports the effectiveness of the proposed method, with a specific emphasis on features embedding. Our study demonstrates that the triplet loss is effective in clustering data within the same class and separating it from data in other classes. Additionally, we analyzed the features embedding created by the three proposed models. Later, we discuss the localization performance of the employed frameworks. The last section of this work is devoted to the conclusions and future works.

2. State Of the Art and Background

In this chapter, in Sec. 2.1 we will first provide a review of the current state-of-the-art of model-based methods, focusing on subspace methods in Sec. 2.1.1. Then in Sec. 2.2 we will present also a brief review of data-driven methods. Afterwards, in Sec. 2.3 we will present the concept of deep learning, which is the main framework of this thesis. For the same reason, we present the CRNN architecture in Sec. 2.3.1. Finally, in Sec. 2.3.2 we will define the siamese network architecture to introduce triplet loss, which are the main paradigms we used for training.

2.1. Model-based methods

Popular approaches for determining the DOA of a microphone array are TDOA based methods, which exploit the differences in the arrival delays of a signal at different sensors in a microphone array. The basic idea behind TDOA-based methods is that the time delay between the arrival of a signal at two different sensors can be related to the angle of the arrival of the signal with respect to the array. One of the most used techniques to estimate the TDOA is the generalized cross-correlation phase transform (GCC-PHAT) [51]. The TDOA estimate is obtained by finding the time delay between the microphone signals that maximizes the GCC-PHAT function.
Another class of solutions are the steered response power (SRP) [82] based strategies, which are based on pointing beamformers towards each direction in a two-dimensional map and measuring the energy that comes from these directions. The PHAT version (SRP-PHAT) is the most popular of these beamformer-based techniques, which has been shown to be very robust under difficult acoustic conditions [20, 23]. In SRP-PHAT, the power coming from each position of the power-map can be derived as the average of the GCC-PHAT between each microphone pair of the array. However, the direction of arrival estimation for SRP algorithms becomes more challenging when multiple sound sources are present in the acoustic scene.
A different kind of strategies are subspace methods [11, 48, 69], which use covariance or correlation matrix of the signal acquisitions to compute the so-called pseudo-spectrum matrix. The pseudo-spectrum matrix shows peaks in a two-dimensional grid where each point correspond to a DOA and each peak coincides with a sound source. Subspace methods are applied on raw signals, as in multiple signal classification (MUSIC) [67, 75]. Subspace methods can be applied to different types of microphone arrays, such as higher-order microphone (HOA) arrays. The availability of HOA arrays made possible new sound field representations [57, 79], such as spherical harmonics (SH) decomposition. SH are a set of orthogonal basis functions [79], which highlight the frequency-dependent and direction-dependent components of the acoustic field. The spherical harmonics domain decomposition has been applied to MUSIC (SHD-MUSIC) in [5, 50], which employ the spherical harmonics coefficients for the computation of the covariance matrix. However, these subspace approaches only consider a free-field propagation. As a result, their localization accuracy degrades in reverberant environments as the acquired signals are contaminated by multi-path acoustic reverberations. Localizing multiple speakers simultaneously active in reverberant environments remains a challenging task. To address this challenge, in [39] the authors employed the relative sound pressure, which is inspired by the relative transfer function (RTF) that contains DOA information while being independent from the source signal and robust to noise. Relative sound pressure applied MUSIC (RMUSIC) and its counterpart in SHD domain (SHD-RMUSIC) showed improved performance in complex environments. Recent researches [42–44] exploit the properties of the SHD derived feature of relative harmonics coefficients. RHCs contain relevant DOA information while being independent from the source signal. For this reason, RHC-based methods demonstrated an improved robustness to reverberant and noisy environments. Different approaches exploiting RHC were published in recent years, such as [40, 44], where Y. Hu *et al.* employed a pre-defined feature set, which is composed by the theoretical values of the RHC. Then, exploiting the property of RHC of being independent from the time-varying source signal, the estimated RHC from microphones array signals are compared to the analytical set in order to recover the source's DOA. This approach requires an exhaustive search over the pre-computed set using a distance-based metric [40].

2.1.1 Subspace Methods

Subspace methods are state-of-art signal processing-based techniques for DOA estimation. Subspace methods rely on a mathematical approach that from the received signals extracts the covariance matrix, whose columns are the vectors of the noise and signal subspace. Then, the subspace decomposition is exploited to compute the pseudo-spectrum, which shows maxima for each active sound source in the estimated DOA. One of the key advantages of subspace methods is that they are robust in the presence of noise and can estimate the DOA of multiple sources simultaneously [67]. This approach can be applied to different microphone array setups. Therefore, subspace methods can be adapted to sound field representations derivative of HOA [5, 39].
In this section, we will revise the subspace methods: MUSIC, SHD-MUSIC, RMUSIC, and SHD-RMUSIC. In

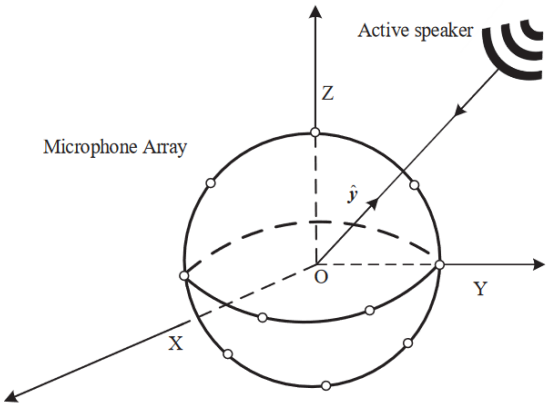


Figure 1: Higher-order microphone array (spherical) setup

Alg. 1- 4, we present the general steps of the aforementioned methods.

MUSIC

We consider the situation where L source signals are impinging on a microphone array of M sensors. As shown in Fig. 1, we consider microphones disposed on the surface of a sphere, whose polar coordinates are $\boldsymbol{x}_j = (r, \theta_j, \phi_j)$, $j = 1, \ldots, M$, with respect to the array origin O, where r is the radius of the sphere, θ_j is the elevation and ϕ_j is the azimuth of the j-th microphone. Assume far-field conditions and L simultaneously active sound sources located at angles $\Psi_l = (\theta_l, \phi_l)$, $l = 1, \ldots, L$, from the array origin, with elevation θ_l and azimuth ϕ_l. The received sound pressure at the multichannel array for each time frame is usually modeled as in [39]:

$$\mathbf{P}(k) = \mathbf{V}(k)\boldsymbol{s}(k) + \boldsymbol{e}(k), \tag{1}$$

where $k = 2\pi f/c$ is the wave number, f is the temporal frequency and c is the speed of sound. Furthermore, $\mathbf{P}(k) = [P(\boldsymbol{x}_1, k), P(\boldsymbol{x}_2, k), \ldots, P(\boldsymbol{x}_M, k)]^T \in \mathbb{C}^{M\times 1}$ where $P(\boldsymbol{x}_j, k)$ corresponds to the sound pressure at microphone in position $\boldsymbol{x}_j$. $\boldsymbol{e}(k) = [e(\boldsymbol{x}_1, k), e(\boldsymbol{x}_2, k), \ldots, e(\boldsymbol{x}_M, k)]^T \in \mathbf{C}^{M\times 1}$ is the noise vector where $e(\boldsymbol{x}_j, k)$ denotes the additive noise signal at the j-th microphone and $\boldsymbol{s}(k) = [s_1(k), s_2(k), \ldots, s_L(k)]^T \in \mathbb{C}^{L\times 1}$ is the source signal vector where $s_l(k)$ denotes the l-th source signal as observed at the origin. In (1), we omitted the time frame index for brevity. The matrix $\mathbf{V}(k)$ in (1) denotes the steering matrix,

$$\mathbf{V}(k) = [\boldsymbol{v}_1(k), \boldsymbol{v}_2(k), \ldots, \boldsymbol{v}_M(k)]^T \in \mathbb{C}^{M\times L}, \tag{2}$$

where $\boldsymbol{v}_j(k) = [e^{i\boldsymbol{k}_1^T\boldsymbol{x}_j}, e^{i\boldsymbol{k}_2^T\boldsymbol{x}_j}, \ldots, e^{i\boldsymbol{k}_L^T\boldsymbol{x}_j}]^T$ represents the steering vector for the j-th microphone, and $\boldsymbol{k}_l = [k\cos\phi_l\sin\theta_l, k\sin\phi_l\sin\theta_l, k\cos\theta_l]^T$ is the wavenumber vector.
The MUSIC approach [67, 69, 75] exploits the $M \times M$ covariance matrix of the observation. Under the basic assumption that the incident signals and the noise are uncorrelated, the covariance matrix is defined as:

$$\mathbf{R}_P(k) \triangleq \mathbb{E}\{\mathbf{P}(k)\mathbf{P}^H(k)\} = \mathbf{V}(k)\mathbf{R}_{\boldsymbol{s}}(k)\mathbf{V}^H(k) + \mathbf{R}_{\boldsymbol{e}}(k), \tag{3}$$

where $[\cdot]^H$ denotes the hermitian operation and $\mathbb{E}$ is the expectation operator.

$$\mathbf{R}_s(k) = \mathbb{E}\{\boldsymbol{s}(k)\boldsymbol{s}^H(k)\}, \tag{4}$$

$$\mathbf{R}_e(k) = \mathbb{E}\{\boldsymbol{e}(k)\boldsymbol{e}^H(k)\}. \tag{5}$$

The covariance matrix consists into M eigenvectors. The eigenvectors can be divided into L vectors related to sound sources and $M - L$ noise eigenvectors. Describing $\mathbf{U}_e$ as the $M \times (M - L)$ matrix whose columns are the $M - L$ noise eigenvectors, we can compute the pseudo spectrum as in [67]:

$$\boldsymbol{\Gamma}_{MUSIC}(k, y_s) = \frac{1}{\boldsymbol{a}(k, y_s)\mathbf{U}_e\mathbf{U}_e^*\boldsymbol{a}^*(k, y_s)}, \tag{6}$$

where $[\cdot]^*$ is the conjugate operation and $\boldsymbol{a}(k, y_s)$ is the steering vector for wavenumber k and direction $y_s = (\theta_s, \phi_s)$. The steering vectors $\boldsymbol{a}(k, y_s)$ are orthogonal to the noise subspace for the directions y_s that point to the sources' DOA. Therefore, the denominator of the pseudo-spectrum has minima around the DOA of the L

Algorithm 1 MUSIC

Data: Time-domain recordings

1: Transfer the recordings into STFT domain
2: **for** $k = 1, 2, \ldots, K$ **do**
3: Calculate covariance matrix $\mathbf{R}_P(k)$ (3)
4: Calculate the subspace $\mathbf{U}_e$
5: Calculate the pseudo spectrum $\mathbf{\Gamma}_{MUSIC}(k, y_s)$ (6)
6: **end for**
7: Average the spectrum over a wide band $\mathbf{\Gamma}_{MUSIC}(y_s)$ (7)
8: Search for L peaks

Algorithm 2 SHD-MUSIC

Data: Time-domain recordings

1: Transfer the recordings into STFT domain
2: Calculate the spherical harmonics coefficients $\alpha(k)$ (8)
3: **for** $k = 1, 2, \ldots, K$ **do**
4: Calculate the covariance matrix $\mathbf{R}_\alpha$ (13)
5: **end for**
6: Calculate the smoothed covariance matrix $\tilde{\mathbf{R}}_\alpha$ (14)
7: Calculate the subspace $\mathbf{U}_e$
8: Calculate the pseudo spectrum $\mathbf{\Gamma}_{MUSIC}(\Psi)$ (15)
9: Search for L peaks

sources, which corresponds to L peaks in the pseudo-spectrum. To obtain a single direction of arrival (DOA) estimation for each time frame, the space pseudo-spectrum described in equation (6) is averaged across a broad frequency range,

$$\mathbf{\Gamma}_{MUSIC}(y_s) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{\Gamma}_{MUSIC}(k, y_s)\ . \tag{7}$$

SHD-MUSIC

The MUSIC approach can be applied also in the SHD. The so-called SHD-MUSIC exploits the transformation in SH of the acoustic pressure. As in [50], the expression for the spherical harmonics decomposition of the sound pressure is:

$$\boldsymbol{P}(k) = \mathbf{B}(k)\mathbf{Y}^H(\Psi_l)\boldsymbol{s}(k) + \boldsymbol{e}(k), \tag{8}$$

matrix $\mathbf{B}(k)$ is defined as $\mathbf{B} = \text{diag}(b_0, b_1, b_1, b_1, \ldots, b_N)$, where $b_n(\cdot)$ is the n-th order spherical Bessel function of the first kind. Furthermore, $\mathbf{Y}(\Psi)$ is the spherical harmonics' matrix of order $n = 0, \ldots, N$ and degree $m = -n, \ldots, n$, where N is the array order of $N \geq kr$. Therefore, $\mathbf{Y}(\Psi_l)$ is composed of L row vectors of length $(N+1)^2$. The elements of the vector in the l-th row are:

$$\boldsymbol{y}_l = [Y_{00}(\Psi_l), Y_{1-1}(\Psi_l), Y_{10}(\Psi_l), Y_{11}(\Psi_l), \ldots, Y_{NN}(\Psi_l)], \tag{9}$$

where

$$Y_{nm}(\theta_l, \phi_l) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}}\mathcal{P}_{nm}(cos\theta_l)e^{im\phi_l}, \tag{10}$$

denotes the spherical harmonics function, and $\mathcal{P}_{nm}(\cdot)$ is the associated Legendre function. In order to estimate the spherical harmonics coefficients, it is assumed that the configuration of the spherical array is known. Considering all the orders up to the N-th order and multiplying (8) from the left by $\mathbf{B}^{-1}$, we can write the matrix form of the spherical harmonics coefficient as:

$$\boldsymbol{\alpha}(k) = \mathbf{Y}^H(\Psi)\boldsymbol{s}(k) + \bar{\boldsymbol{e}}(k), \tag{11}$$

where

$$\bar{\boldsymbol{e}}(k) = \mathbf{B}^{-1}(k)\boldsymbol{e}(k). \tag{12}$$

Similarly to (3), we can derive the modal cross-spectrum as:

$$\mathbf{R}_{\alpha}(k) = \mathbb{E}\{\boldsymbol{\alpha}(k)\boldsymbol{\alpha}^{H}(k)\} = \mathbf{Y}^{H}(\Psi)\mathbf{R}_{\boldsymbol{s}}(k)\mathbf{Y}(\Psi) + \mathbf{R}_{\bar{\boldsymbol{e}}}(k). \tag{13}$$

As in [50], the signal covariance matrix $\mathbf{R}_{\boldsymbol{s}}(k)$ and noise covariance matrix $\mathbf{R}_{\bar{\boldsymbol{e}}}(k)$ are averaged over the frequency-domain to obtain a single covariance matrix for each time frame.

$$\tilde{\mathbf{R}}_{\alpha} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{R}_{\alpha}(k)\ . \tag{14}$$

Finally, if we consider U_e as the noise subspace of $\mathbf{R}_{\alpha}$ and $\boldsymbol{y}_l$ as steering vector, the MUSIC pseudo-spectrum derived in (6) is equivalent to:

$$\mathbf{\Gamma}_{MUSIC}(\Psi) = \frac{1}{\boldsymbol{y}_l(\Psi)\mathbf{U}_e\mathbf{U}_e^{*}\boldsymbol{y}_l^{*}(\Psi)}. \tag{15}$$

RMUSIC

Relative sound pressure has been proposed in [39], defined as the ratio between the pressure captured by the j-th microphone and the pressure at the origin of the array. Therefore, the relative sound pressure definition is

$$Q(\boldsymbol{x}_j, k) = \frac{P(\boldsymbol{x}_j, k)}{P(\boldsymbol{x}_0, k)} = \frac{P(\boldsymbol{x}_j, k)P^{*}(\boldsymbol{x}_j, k)}{|P(\boldsymbol{x}_0, k)|^2}, \quad j = 1, \ldots, M\,, \tag{16}$$

where $\boldsymbol{x}_0 = (0, 0, 0)$ is the origin of the array. However, certain structured arrays, such as the spherical arrays, only have microphones on the array surface. In that case, the pressure at the origin is approximated as the average of the ones on the surface of the array. Assuming the source signal is stationary over a short time period, the relative sound pressure is represented as

$$Q(\boldsymbol{x}_j, k) = \frac{S_{p_j p_0}(k)}{S_{p_0 p_0}(k)}\,, \tag{17}$$

where $S_{p_0 p_0}(k)$ and $S_{p_j p_0}(k)$ denote the power spectral density (PSD) of $P(\boldsymbol{x}_0, k)$ and the cross-PSD (CPSD) between $P(\boldsymbol{x}_j, k)$ and $P(\boldsymbol{x}_0, k)$, respectively. In noisy environments, the PSD of the sound pressure at the origin contains also a noise component. Assuming the source signal is stationary over a short period and the source signal and the noise signal are uncorrelated, the noisy relative sound pressure follows,

$$\bar{Q}(\boldsymbol{x}_j, k) = \frac{S_{p_j p_0}(k)}{S_{p_0 p_0}(k) + S_{n_0 n_0}(k)}\,, \tag{18}$$

where $S_{n_0 n_0}(k)$ is PSD of the noisy component. Dividing (19) by (17), the relation between the noisy and noiseless sound pressure can be derived as:

$$\bar{Q}(\boldsymbol{x}_j, k) = Q(\boldsymbol{x}_j, k)\rho(k), \tag{19}$$

where

$$\rho(k) = \frac{T(\boldsymbol{x}_0, k)}{T(\boldsymbol{x}_0, k) + 1}, \tag{20}$$

only depends on the signal-to-noise ratio (SNR) at the origin of the array. i.e., $T(\boldsymbol{x}_0, k) = S_{p_0 p_0}(k)/S_{e_0 e_0}(k)$. The relative sound pressure represented using the PSD between two microphones, is also robust to the noise. The $M \times M$ covariance matrix of the noisy relative sound pressure can be calculated as:

$$\mathbf{S}_{\bar{\mathbf{Q}}}(k) = \mathbb{E}\{\bar{\mathbf{Q}}(k)\bar{\mathbf{Q}}^{H}(k)\} = \mathbf{V}(k)\mathbf{R}_{\mathbf{S}}(k)\mathbf{V}^{H}(k), \tag{21}$$

where $\mathbf{V}(k)$ is the steering vector defined in (1), and

$$\mathbf{R}_{\mathbf{S}}(k) = \{\bar{\boldsymbol{s}}(k)\rho(k)\bar{\boldsymbol{s}}^{H}(k)\rho^{*}(k)\}, \tag{22}$$

where $\bar{s}(k)$ is the vector of the noisy signal captured at the microphones. From the covariance matrix, we can derive the matrices $\bar{\mathbf{U}}_s(k)$ and $\bar{\mathbf{U}}_e(k)$, which are the subspaces corresponding to the source and noise eigenvectors. The pseudo-spectrum over space can be calculated as:

$$\mathbf{\Gamma}_{MUSIC}(k, y_s) = \frac{1}{||\bar{\mathbf{U}}_e^{H}(k)\boldsymbol{a}(k, y_s)||^2}\ . \tag{23}$$

Finally, to obtain a single estimate of the direction of arrival (DOA) for each time frame, the pseudo-spectrum is obtained as in (7).

Algorithm 3 RMUSIC

Data: Time-domain recordings

1: Transfer the recordings into STFT domain
2: Calculate the relative sound pressure $\bar{Q}(\boldsymbol{x}_j, k)$ (18)
3: **for** $k = 1, 2, \ldots, K$ **do**
4: Calculate covariance matrix $\mathbf{S}_{\bar{\mathbf{Q}}}(k)$ (21)
5: Calculate the subspace $\mathbf{U}_e$
6: Calculate the pseudo spectrum $\mathbf{\Gamma}_{MUSIC}(k, y_s)$ (23)
7: **end for**
8: Average the spectrum over a wide band $\mathbf{\Gamma}_{MUSIC}(y_s)$ (7)
9: Search for L peaks

Algorithm 4 SHD-RMUSIC

Data: Time-domain recordings

1: Transfer the recordings into STFT domain
2: Calculate the relative sound pressure $\bar{Q}(\boldsymbol{x}_j, k)$ (18)
3: Calculate its spherical harmonics coefficients $\bar{\boldsymbol{\beta}}(k)$ (25)
4: **for** $k = 1, 2, \ldots, K$ **do**
5: Calculate the covariance matrix $\mathbf{S}_{\bar{\boldsymbol{\beta}}}(k)$ (26)
6: **end for**
7: Calculate the smoothed covariance matrix $\tilde{\mathbf{S}}_{\bar{\boldsymbol{\beta}}}(k)$ (27)
8: Calculate the subspace $\mathbf{U}_e$
9: Calculate the pseudo spectrum $\mathbf{\Gamma}_{MUSIC}(\Psi)$ (15)
10: Search for L peaks

SHD-RMUSIC

SHD-RMUSIC applies the proposed RMUSIC approach to the spherical harmonics domain. The relative sound pressure measured over the microphone array can be decomposed into the spherical harmonics domain. The spherical harmonic decomposition of the measured relative sound pressure in (19) can be expressed as

$$\bar{\beta}_{nm}(k) = \frac{1}{b_n(kr)} \sum_{j=1}^{M} a_j \bar{Q}(\boldsymbol{x}_j, k) Y_{nm}^*(\theta_j, \phi_j)\,, \tag{24}$$

where a_j denotes the weights of each microphone to ensure the orthogonality of the spherical functions $Y_{nm}(\cdot)$ defined in (10) and $\bar{Q}(\boldsymbol{x}_j, k)$ is the noisy relative sound pressure defined in (19). Traditional spherical harmonics decomposition suffers from the "Bessel zero problem" [63], due to the zero crossing of the Bessel function. For this reason, the noise component in the measured spherical harmonic coefficients is amplified. This issue is contrasted by the relative sound pressure that is less sensitive to noise, as demonstrated in [39]. From (19), assuming plane wave modeling we can rewrite the spherical harmonics coefficients for the noisy relative sound pressure in (24) as:

$$\bar{\boldsymbol{\beta}}(k) = \mathbf{Y}^H(k)\bar{\boldsymbol{s}}(k)\rho(k), \tag{25}$$

where $\mathbf{Y}(k)$ is the $(N+1)^2 \times L$ steering matrix in the SHD. The correlation matrix of $\bar{\boldsymbol{\beta}}(k)$ over the time-varying source signal is:

$$\begin{aligned} \mathbf{S}_{\bar{\boldsymbol{\beta}}}(k) &= \mathbb{E}\{\bar{\boldsymbol{\beta}}(k)\bar{\boldsymbol{\beta}}^H(k)\} \\ &= \mathbf{Y}^H(k)\mathbf{R}_{\mathbf{S}}(k)\mathbf{Y}(k). \end{aligned} \tag{26}$$

In [39], the noise subspace $\mathbf{U}_e$ is computed from the frequency-smoothed covariance matrix, which is implemented as the average of the covariance matrices at different frequency bins [50],

$$\tilde{\mathbf{S}}_{\bar{\boldsymbol{\beta}}}(k) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_{\bar{\boldsymbol{\beta}}}(k) = \mathbf{Y}^H(k)\tilde{\mathbf{R}}_{\mathbf{S}}(k)\mathbf{Y}(k)\,, \tag{27}$$

where

$$\tilde{\mathbf{R}}_{\mathbf{S}}(k) = \frac{1}{k} \sum_{k=1}^{K} \mathbf{R}_{\mathbf{S}}(k)\,, \tag{28}$$

where K frequency bins are exploited. Finally, the noise subspace extracted from the smoothed covariance matrix is exploited for the pseudo-spectrum as defined in (23).

2.2. Data-driven methods

To address the challenging task of localizing a sound source active in reverberant environments a large number of algorithms have been developed. In recent years, the research interest has focused on deep learning techniques. As a result, an increasing number of methods based on deep neural networks (DNNs) have been proposed, exhibiting improved accuracy in complex environments. Most of the reported works have indicated the superiority of DNN-based methods over conventional model-based methods. In [61], the authors were able to address the multiple source localization problem using first-order Ambisonics (FOA) signals with a convolutional recurrent neural network (CRNN). Adavanne *et al.* [9] improved DOA and the number of sources estimation accuracy by exploiting the spectrograms' magnitudes and phases of multichannel audio signals. The phases and the magnitudes of the spectrograms are sequentially mapped using a CRNN with two different output branches. The first output, the spatial pseudo-spectrum (SPS) is generated as a regression task, followed by the DOA estimates as a classification task. More recently, DL-based methods have been proposed for joint sound event localization and detection (SELD) [8], which is a combination of sound event detection (SED) and sound source localization SSL. In particular, the majority of the works are presented in challenges, such as DCASE Challenge Task 3 [1] and L3DAS Challenge [2]. DL-based algorithms for localizing single or multiple sources typically employ a classification process to categorize the DOA. In addition, some algorithms employ regression networks that are better suited for continuous position localization. A comprehensive survey of DL methods can be found in [32].

2.3. Deep Learning Background

Deep learning is a subfield of machine learning that typically processes input data through a series of layers. Each layer applies a mathematical operation on the received data and sends its output to the next layer. A sequence of layers is called neural network, also known as deep neural network (DNN). DNNs enable machines to learn from large volumes of complex data, including images, sound signals, and text. DL is built around the concept of layers of interconnected nodes, known as artificial neurons or perceptrons. These layers allow the model to extract increasingly abstract and complex features. Research on DL-based techniques is constantly proposing new architectures and methods that provide better results in a wide range of applications, as shown in [24].
In the next sections, we will provide an overview of the CRNN model, which is the reference model for this work. Finally, we will define the siamese network structure in order to review the triplet loss function, which is the framework used in this work in order to improve localization performance.

2.3.1 Convolutional Recurrent Neural Network (CRNN)

In recent years, Convolutional Neural Networks (CNNs) [26, 53] and Recurrent Neural Networks (RNNs) [58] have made significant contributions to various fields [14, 31, 59]. However, these types of neural networks have different strengths and limitations. CNNs are excellent at capturing spatial features, while RNNs can model sequences and capture temporal information. The CRNN model is a hybrid neural network architecture that combines CNNs and RNNs, in order to exploit their strengths. For this reason, CRNN has been applied in different applications that involve temporal data, such as speech enhancement [71], text classification [78], music classification [19] and video classification [85].
As shown in Fig. 2, the CRNN model is a sequential composition of the aforementioned types of networks. The input data are fed to the CNN which retrieves important spatial information. Then, the output of the CNN is used as input for the RNN, which extract information on temporal patterns from data. Afterwards, the output of the recurrent block is flattened to represent the data as a one-dimensional vector. In the end, the flattened data are fed to the FC network, which produces the final output vector. Finally, the output vector is usually transformed by applying an activation function, which type can vary depending on the application.

CNN A CNN is usually composed of the concatenation of convolutional layers and pooling layers. The convolutional layer applies the convolution operation to the input and passes the result to the next layer. It is made of a set of filters (kernels). The number of kernels in a convolutional layer is known as the depth of the layer. Each kernel is characterized by its filter size and stride, which defines the dimensions of the mask that is convolved with the input feature. The stride indicates how much the filter is shifted along the spatial

dimensions. Therefore, the convolutional layer applies a set of convolutional filters to the input features to extract a new set of features, which contain relevant spatial information.
Typically, a pooling layer is inserted between consecutive convolutional layers in deep learning architectures. This layer performs non-linear subsampling, effectively reducing the dimensionality of the data and decreasing computational complexity. As in the case of the convolutional layer, the pooling layer is also defined by two parameters: the filter size and the stride. This layer operates independently on every slice of the input, dividing it into smaller patches with filter size. Then, it outputs a value for each patch depending on a specific function, as a result, the input is resized. The function utilized defines the pooling layer. One of the most common forms is max pooling, which outputs the maximum values contained in each patch. Another of the historically most used pooling layers is average pooling, which extracts the average of the patch.

RNN RNNs sequentially process the features and capture the temporal information. Recurrent layers have memory cells that allow them to store and use information from previous time steps. Recurrent layers can be implemented using various types of layers, such as long short-term memory (LSTM) [37] or gated recurrent units (GRU) [18]. Each of these types of layers has a slightly different mechanism for storing and updating the memory cell but they all share the same basic idea to use past information to compute the current predictions. The LSTM layer is a type of layer that can maintain memory over time. A LSTM layer is composed of a memory cell and three multiplicative units, the input, output, and forget gates. The input gate controls the flow of input data, the output gate limits which elements of the memory cell are propagated through the recurrent updates, and forget gate decides to store or reset the information from previous states. Instead, GRU layers have only two gates: the update and reset gates. The reset gate determines how much of the previous information consider in the new state and the update gate controls the balance between the previous state and the new one. By incorporating information from previous steps, recurrent layers can capture complex patterns and dependencies in sequential data that would be difficult for traditional models to capture.

Fully Connected Network Fully connected (FC) layers are a type of layer in deep learning where every neuron in a given layer is connected to every neuron in the following layer. For this reason, FC layers are also called dense layers. In a dense layer, each neuron in the previous layer is connected to each neuron in the current layer with a weight. These weights are learned during the training process and determine the strength of the connection between the neurons. For the input to be processed by a dense layer, it is typically necessary to flatten or reshape it into a one-dimensional vector representation. The input data are transformed by applying a set of weights and adding a bias. The generated output of a dense layer is also a one-dimensional vector, which can then be passed to subsequent layers.

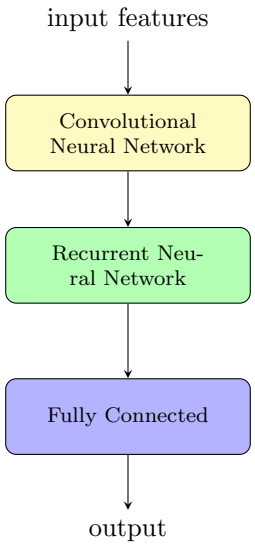


Figure 2: A generic CRNN

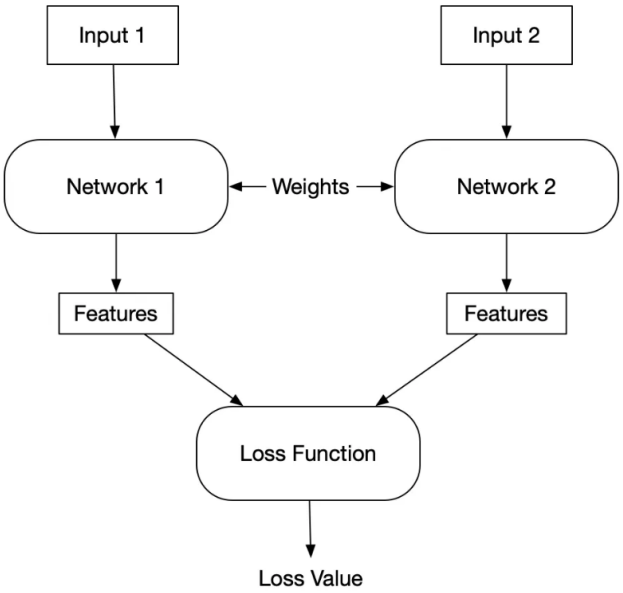


Figure 3: A generic siamese model. A siamese network is often shown as multiple different encoding network that share weights.

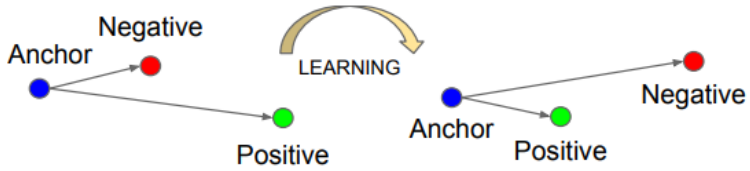


Figure 4: The basic idea behind triplet loss is to train a neural network to produce embeddings, where the anchor sample and the positive sample are closer together than the anchor sample and the negative sample.

2.3.2 Siamese Network

Siamese networks were first introduced in the early 1990s by Bromley and LeCun to solve signature verification as an image-matching problem [45]. Siamese networks are a class of neural networks in deep learning that are designed to compare inputs and determine their similarity. Therefore, two or more identical neural networks are employed to process the same input.
A siamese network is often shown as multiple encoding networks with the same architecture and configuration, as shown in Fig. 3, and they are usually employed in classification tasks [52, 64, 68]. The output of siamese networks is the measure of the loss, which depends on the similarity of the output vectors. The typical loss functions used to train siamese networks are triplet loss [68] and contrastive loss [34].
Triplet loss is a function that compares the output vectors of an anchor, a positive, and a negative input, evaluating their similarity. Therefore, three siamese subnets are employed with triplet loss, one for each element of the triplet. Triplet loss will be explained in detail in this section. Instead, the contrastive loss is a distance-based loss, and it is exploited to learn embeddings in which similar features have a low distance and two dissimilar instances have a large distance.

Triplet Loss

The training process of a siamese network involves minimizing a loss function that measures the dissimilarity between the inputs. The basic idea behind triplet loss is to train a neural network to produce embeddings, which are low-dimensional representations of input data. As illustrated in Figure 4, the embeddings are trained to position features that belong to the same class closer together in the embedding space, while arranging samples that belong to different classes further apart. The triplet loss is based on the generation of triplets of input data: an anchor, a positive, and a negative. The anchor is the sample for which we want to learn the representation. The positive sample is extracted from the same class of the anchor. Instead, the negative

sample comes from a different class. The aim is to learn embeddings where the distance between the anchor and the positive is smaller than the distance between the anchor and the negative. In other words, we want to maximize the similarity between the anchor and the positive, while minimizing the similarity between the anchor and the negative. The hard margin version of triplet loss encourages the similarity between the anchor and a positive sample to be larger than the similarity between the same anchor and the negative instances as below:

$$\mathcal{L}_{triplet} = [\delta + \mathcal{S}(\boldsymbol{\sigma}_a, \boldsymbol{\sigma}^-) - \mathcal{S}(\boldsymbol{\sigma}_a, \boldsymbol{\sigma}^+)]_+ , \tag{29}$$

where δ is the hard margin and $\boldsymbol{\sigma}_a$,$\boldsymbol{\sigma}^+$, and $\boldsymbol{\sigma}^-$ denote an anchor sample, a positive sample, and a negative sample. $\mathcal{S}$ represents a generic similarity function and the $[\cdot]_+$ operator is the hinge function $\max(\cdot, 0)$. The soft margin variant of triplet loss has been demonstrated to be more effective than the hard margin version in applications such as face recognition [68] and person re-identification [49]. The soft margin version replaces the hinge function with the softplus function, which decays exponentially instead of having a hard cut-off. Soft margin triplet loss is defined as:

$$\mathcal{L}_{triplet} = \log(1 + \exp\left(\mathcal{S}(\boldsymbol{\sigma}_a, \boldsymbol{\sigma}^-) - \mathcal{S}(\boldsymbol{\sigma}_a, \boldsymbol{\sigma}^+)\right)) . \tag{30}$$

In conclusion, triplet loss is a powerful loss function for learning embeddings. For this reason, it has been widely used in computer vision applications and has been shown to be effective for various tasks, for example, face recognition [68] and acoustic scene classification [62].

2.4. Deep Learning Methods

In this section, we will describe the work proposed by Fahim *et al.* in [29]. In this work, A DOA estimation technique is proposed using a convolutional neural network algorithm that learns modal coherence patterns of an incident sound field through measured spherical harmonic coefficients. We will describe a few concepts in the SHD that define the modal framework for the proposed DOA estimation technique. Then we will present the deep learning model employed for estimating simultaneously active multiple sound sources on a 3D space using a single-source training scheme.

2.4.1 Multi-Source DOA Estimation through Pattern Recognition of the Modal Coherence of Reverberant Soundfield

In this work [29], Fahim *et al.* propose a multi-source DOA estimation technique based on a convolutional neural network, which learns the modal coherence patterns employing SHD-based input features. This method is capable of estimating simultaneously active multiple sound sources using a single-source training scheme. The training is conducted on a single-source case and the same model is tested in various acoustic environments with multiple active sources. The method is evaluated in various simulated and practical noisy and reverberant environments.

Similarly to the methods described in Sec. 2.1.1, this method considers L sound sources concurrently emitting sound. The sound pressure observed by an omnidirectional microphone placed at a coordinate $\boldsymbol{x}'_j \equiv (r'_j, \theta'_j, \phi'_j)$ inside the room, is modeled as

$$p(\boldsymbol{x}'_j, t) = \sum_{l=1}^{L} h_l(\boldsymbol{x}'_j) * s_l(t), \tag{31}$$

where $h_l(\boldsymbol{x}'_j)$ is the room impulse response (RIR) between the l-th source position and $\boldsymbol{x}'_j$ and $*$ denotes the convolution operation. The corresponding frequency domain representation of (31) in the STFT domain can be obtained using the multiplicative model of convolution and is formulated as

$$P(\boldsymbol{x}'_j, k) = \sum_{l=1}^{L} S_l(k) H_l(\boldsymbol{x}'_j, k), \tag{32}$$

where $\{P, S, H\}$ represent the corresponding signals of $\{p, s, h\}$ in the STFT domain. In (32), the timeframe index is omitted for brevity.

The method intends to estimate the individual DOAs $\hat{\boldsymbol{x}}_l \equiv (\hat{\theta}_l, \hat{\phi}_l), l = 1, \ldots, L$ of the multiple concurrent sound sources, given a set of measured sound pressure $p(\boldsymbol{x}'_j, t), j = 1, \ldots, M$.

Modal Framework The sound field captured on a sphere can be decomposed using the spherical harmonic basis functions as [79]

$$P_{x'_j}(k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k) b_n(kr) Y_{nm}(\theta'_j, \phi'_j), \tag{33}$$

we remind that α_{nm} are the spherical harmonics coefficients, $b_m(\cdot)$, and $Y(\cdot)$ are the n-th order spherical Bessel function of the first kind and the spherical harmonics matrix, respectively. On the assumption that a spherical microphone array is employed to capture the sound pressure, the spherical harmonics coefficients can be calculated from (33) as [7]

$$\alpha_{nm}(k) \approx \frac{1}{b_n(kr)} \sum_{j=1}^{M} w_j P(x_j, k) Y^*_{nm}(\theta'_j, \phi'_j), \tag{34}$$

where r is the array radius and w_j are suitable microphone weights that ensure the validity of the orthonormal property of the spherical harmonics with a limited number of sampling points. Alternative array geometries and formulations can be exploited to achieve the same spherical harmonic decomposition [6, 7, 66].
In reverberant environments, the room transfer function can be decomposed into

$$H_l(\boldsymbol{x}'_j, k) = H_l^{\text{dir}}(\boldsymbol{x}'_j, k) + H_l^{\text{rev}}(\boldsymbol{x}'_j, k), \tag{35}$$

where $H_l^{\text{dir}}(\boldsymbol{x}'_j, k)$ and $H_l^{\text{rev}}(\boldsymbol{x}'_j, k)$ are the corresponding direct and reverberant components of the room transfer function. The room transfer function components can be modeled in the spatial domain. Therefore, we can obtain the spatial domain equivalent of (32) as

$$P(\boldsymbol{x}'_j, k) = \sum_{l=1}^{L} S_l(k) \Big(G_l^{\text{dir}}(k) e^{ik\hat{\boldsymbol{x}}_l \cdot \boldsymbol{x}'_j} + \int_{\mathbb{S}^2} G_l^{\text{rev}}(k, \hat{\boldsymbol{x}}) e^{ik\hat{\boldsymbol{x}} \cdot \boldsymbol{x}'_j} d\hat{\boldsymbol{x}} \Big), \tag{36}$$

where $\hat{\boldsymbol{x}}$ denotes an arbitrary direction on the spherical shell $\mathbb{S}^2$ $G_l^{\text{dir}}(k)$ represents the direct path gain between the origin and the l-th source and $G_l^{\text{rev}}(k, \hat{\boldsymbol{x}})$ is the reflection gain at the origin along the direction of $\hat{\boldsymbol{x}}$ for the l-th source. From (36) we can infer that the sound pressure at the j-th microphone is a combination of the direct signal and the reverberation version of the signal from different directions. Under the assumption of free-field conditions, the spherical harmonic expansion of Green's function is given by [21]

$$e^{ik\hat{\boldsymbol{x}}_l \cdot \boldsymbol{x}'_j} = \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi i^n Y^*_{nm}(\hat{\boldsymbol{x}}_l) b_n(kr) Y_{nm}(\hat{\boldsymbol{x}}'_j), \tag{37}$$

where we consider the unit vector $\hat{\boldsymbol{x}}'_j \equiv (\theta'_j, \phi'_j)$. Substituting (37) into (36) and then comparing it with (33), we obtain an analytical expression for α_{nm} in a reverberant room as [28]

$$\alpha_{nm}(k) = 4\pi i^n \sum_{l=1}^{L} S_l(k) \Big(G_l^{\text{dir}}(k) Y^*_{nm}(\hat{\boldsymbol{x}}_l) + \int_{\mathbb{S}^2} G_l^{\text{rev}}(k, \hat{\boldsymbol{x}}) Y^*_{nm}(\hat{\boldsymbol{x}}) d\hat{\boldsymbol{x}} \Big). \tag{38}$$

The modal coherence in the spherical harmonic domain can be interpreted as the degree of similarity between the modal coefficients of two or more signals decomposed in their respective SHC. High model coherence suggests that the signals have similar spatial structures and are likely to be related, while low modal coherence indicates that the signals have different spatial patterns and are unlikely to be related. Therefore, modal coherence is defined as

$$\mathbb{E}\Big\{ \alpha_{nm}(k) \alpha^*_{n'm'}(k) \Big\}. \tag{39}$$

Considering the independent behaviour of the reflective surfaces in a room, meaning that the reflection gains from the reflective surfaces are independent, and under the assumption of uncorrelated sources, in [28] has been established a closed form expression for the modal coherence. Furthermore, for temporal processing, a common method in temporal processing for estimating the expected value involves applying the exponential moving average technique on the instantaneous measurements. Hence, the modal coherence is estimated as

$$\mathbb{E}\Big\{ \alpha_{nm}(k,t) \alpha^*_{n'm'}(k,t) \Big\} = (1-\mu)\ \alpha_{nm}(k,t) \times \alpha^*_{n'm'}(k,t) + \mu\ \mathbb{E}\Big\{ \alpha_{nm}(k,t-1) \alpha^*_{n'm'}(k,t-1) \Big\}, \tag{40}$$

where $\mu \in [0, 1]$ is the smoothing factor.

Algorithm 5 Algorithm for DOA estimation - training stage

1: Compute spatial coherence $\mathbb{E}\{\alpha_{nm}(k)\alpha^*_{n'm'}\}$ in each time-frequency bin using (40)
2: Calculate $\hat{\mathcal{F}}_{\text{mc}}$ $\forall\theta$, $\forall\phi$ using (41)
3: Apply energy-based pre-selection to filter out low energy time-frequency bins
4: Use pre-selected features to train the model

Algorithm 6 Algorithm for DOA estimation - evaluation stage
Data: α_{nm} $\forall nm$

1: Compute spatial coherence $\mathbb{E}\{\alpha_{nm}(k)\alpha^*_{n'm'}\}$ in each time-frequency bin using (40)
2: Get $\hat{\mathcal{F}}_{\text{mc}}$ using (41)
3: Apply energy-based pre-selection to filter out low energy time-frequency bins
4: Calculate the probability of each class for the time-frequency bins using the trained model
5: Select time-frequency bins where the prediction has a prominent probability
6: Apply (42) to form the multiset $\mathcal{X}$
7: **if** $L == 1$ **then**
8: Select larger peak in $\mathcal{X}$
9: **else**
10: Using a suitable clustering algorithm, divide $\mathcal{X}$ into L clusters
11: Select larger peak from each cluster
12: **end if**

CNN-based DOA estimation In the proposed CNN-based DOA estimator, the authors of [29] pose the DOA estimation problem as an image-classification problem where the input image represents the modal coherence of the sound field. The sound field $\alpha_{n,m}$ can be thought as beamformers in the modal domain due to the inherent properties of the spherical harmonic functions. Hence, the energy distribution of the SHC among different modes can be used as a clue for understanding the source directionality. For this method, the modal coherence model is exploited to construct the input features. For a multi-source scenario, it is common to assume W-disjoint orthogonality [83] in the STFT domain, i.e., only a single sound source remains active in each TF bin of the STFT spectrum. The DOA estimation problem is posed as an image-classification problem for CNN, where the feature snapshot is defined as the modal coherence. Hence, for each time-frequency bin of the STFT spectrum, the features are defined as

$$\hat{\mathcal{F}}_{\text{mc}}(k) = \left\{ \mathbb{E}\Big\{\alpha_{nm}(k)\alpha^*_{n'm'}(k)\Big\} : n \in [0, N], m \in [-n, n], n' \in [0, N], m' \in [-n', n'] \right\}, \tag{41}$$

where $\hat{\mathcal{F}}_{\text{mc}}$ is considered as an image of $[\mathcal{N} \times \mathcal{N}]$ complex-valued pixels with $\mathcal{N} = (N+1)^2$. In order to enable the model to capture the frequency variations of the feature for a specific source position, the $\hat{\mathcal{F}}_{\text{mc}}$ feature is collected from various frequency bands.
Speech signals are typically sparse in both time and frequency domains. Therefore, a significant number of time-frequency bins tend to have lower energy, which can mislead the CNN. In the time domain, the problem can be addressed with a voice activity detector. In order to deal with the sparsity of the speech signals in the frequency domain, an energy-based pre-selection of time-frequency bins is applied, which consists in dropping all the bins with average energy below a given threshold.
Source DOA estimation is performed on the proposed input features by a CNN. As described in Sec. 2.3.1, a CNN is composed of multiple convolution layers followed by a FC network. For this method, a multi-output multi-class classification is performed. Similar the work proposed in this thesis, the convolution layer structure is shared to predict both azimuth and elevation using separate fully connected heads at the last stage. Each fully connected head is responsible for predicting either azimuth or elevation. Ideally, due to the W-disjoint orthogonality assumption, each time-frequency bin is designated with a single DOA. However, in the realistic case is possible to find time-frequency bins whose energy is the sum of the contribution from multiple sound sources. Hence, the authors employed cross entropy loss in order to independently predict the probability of each class in every TF bin. Then, the predictions are only considered when the CNN model predicts a single DOA with a high confidence level, i.e when the predicted probability is greater than a certain threshold.
In multi-source environments, the simplest way of multi-source DOA estimation is to pick L largest peaks from the joint prediction of azimuth and elevation for each time-frequency bin to create the prediction multiset $\mathcal{X}$

$$\mathcal{X} = \Big\{ \text{argmax}(\theta)\{f : \theta \mapsto \mathcal{P}_\kappa(\theta)\}, \text{argmax}(\phi)\{f : \phi \mapsto \mathcal{P}_\kappa(\phi)\} \Big\}, \tag{42}$$

where $\mathcal{P}_\kappa(\cdot)$ is the probability in the κ-th bin between the pre-selected bins. However, this technique can cause errors in a multi-source environment with noisy predictions. In order to apply a more robust technique, Fahim *et al.* employed a suitable clustering algorithm to divide $\mathcal{X}$ into L clusters and pick the prominent peak in each cluster. The steps of the algorithm are outlined in Alg. 5 and 6.

3. Method and Model

In this chapter, we provide a brief description of the RHC feature and estimator. Then, the CRNN baseline is presented. Furthermore, we will describe the triplet loss application to the baseline network.

3.1. Feature Extraction

In this section, we provide some background information on RHC, illustrating their theoretical expression in free and reverberant environments. Subsequently, we detail the framework for RHC estimation from multi-channel microphone array signals in noisy environments. Finally, we illustrate how the input features of the models are composed.

3.1.1 Relative Harmonic Coefficients (RHC)

As shown in Fig. 1, let us assume a single sound source propagating from an unknown position, denoted as $\boldsymbol{x} = (r_s, \theta_s, \phi_s)$, where r_s is the radial distance and (θ_s, ϕ_s) is the elevation and azimuth direction of the source with respect to the origin of the microphone array. We consider a higher-order microphone (HOM) array having M microphones positioned at $\boldsymbol{x_j} = (r, \theta_j, \phi_j)$. As in (33), we remind the reader that the sound pressure captured by the HOM can be decomposed into the spherical harmonics domain [79] as

$$P_{x_j}(t,k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(t,k) b_n(kr) Y_{nm}(\theta_j, \phi_j) \ . \tag{43}$$

Assuming a far-field scenario, the spherical harmonic coefficients for the direct sound source in equation (43) can be derived as in (34)

$$\alpha_{nm}(k) = \frac{1}{b_n(kr)} \sum_{j=1}^{M} w_j P(x_j, k) Y^*_{nm}(\theta_j, \phi_j \ . \tag{44}$$

Preliminary research in [41, 42, 44] define the RHC as the ratio between the spherical harmonic coefficient $\alpha_{nm}(t,k)$ and $\alpha_{00}(t,k)$. This definition allows us to express RHC representation in terms of the order n and mode m as:

$$\beta_{nm}(t,k) = \frac{\alpha_{nm}(t,k)}{\alpha_{00}(t,k)}, \tag{45}$$

The coefficient β_{00}, from definition (45), has a value of 1, which means that it is not dependent on DOA, for this reason, we can discard β_{00} and consider only [β_{1-1} ,β_{10} ,β_{11}] for DOA estimation.

3.1.2 Free-Field Scenario

Under the assumption of free field (anechoic) conditions, we consider the source $\boldsymbol{x}$. The spherical harmonic coefficients due to the incoming direct-path recordings are given by [72]:

$$\alpha^{\text{dir}}_{nm}(t,k) = S_x(t,k) ikh_n(kr_s) Y^*_{nm}(\theta_s, \phi_s), \tag{46}$$

where $S_x(t,k)$ is the source signal at time-frame t and $h_n(\cdot)$ is the n-th order spherical Hankel function of the first kind. Following the definition in (45), we can obtain the RHC of order n and mode m as:

$$\beta^{\text{dir}}_{nm}(t,k) = \frac{2\sqrt{\pi} h_n(kr_s) Y^*_{nm}(\theta_s, \phi_s)}{h_0(kr_s)}, \tag{47}$$

which only depends on the source position (r_s, θ_s, ϕ_s).

3.1.3 Reverberant-Field Scenario

Assuming the case of a reverberant soundfield generated by the sound source $\boldsymbol{x}$, we can express its spherical harmonic as follows

$$\alpha^{\text{rev}}_{nm}(t,k) = \alpha^{\text{dir}}_{nm}(t,k) + \underbrace{\sum_{v=0}^{N} \sum_{u=-v}^{v} \hat{\alpha}^{vu}_{nm}(k) S_x(t,k) ikb_v(kr_s) Y^*_{vu}(\theta_s, \phi_s)}_{\text{Reverberant-path}}, \tag{48}$$

where $\hat{\alpha}_{nm}^{vu}(k)$ is the coupling coefficient that is independent of the time-varying source signal [65]. Note that differently from (46), in (48) we have an additive component due to the reverberant sound field. The reverberant-path component makes the SHC diverge from the theoretical value. From the RHC definition definition (45), we derive the formulation in a reverberant environment as

$$\beta_{nm}^{\text{rev}}(t,k) = \frac{h_n(kr_s)Y_{nm}^*(\theta_s,\phi_s) + \Sigma_{v=0}^{N}\Sigma_{u=-v}^{v}\hat{\alpha}_{nm}^{vu}(k)b_v(kr_s)Y_{vu}^*(\theta_s,\phi_s)}{h_0(kr_s)Y_{00}^*(\theta_s,\phi_s) + \Sigma_{v=0}^{N}\Sigma_{u=-v}^{v}\hat{\alpha}_{00}^{vu}(k)b_v(kr_s)Y_{vu}^*(\theta_s,\phi_s)}, \tag{49}$$

which also only depends on the position of the source $\boldsymbol{x}$ in a static acoustic environment, i.e., it is assumed that the environment parameters, the configuration of the microphone array, and the source position remains constant during the acquisition.

3.1.4 RHC Estimation

In this work, we consider acoustic environments where reverberation is present. In the RHC estimation, since we want to evaluate as realistic as possible scenarios, we take into account also the noise influence on the recorded signals. Therefore, for decomposition in RHC of the multichannel sound pressure, we adopt the biased estimator of RHC used in [42, 44, 81], which exploits the power spectral density (PSD) and cross PSD (CPSD) of the measured signals to alleviate the negative effects caused by the noise. The RHC estimator is defined as follows:

$$\tilde{\beta}_{nm}(t,k) \approx \frac{S_{\alpha_{nm}\alpha_{00}}(t,k)}{S_{\alpha_{00}\alpha_{00}}(t,k)}, \tag{50}$$

$$\begin{aligned} S_{\alpha_{nm}\alpha_{00}}(t,k) &= \frac{1}{T_{\text{est}}}\sum_{t_1=t-t_0}^{t+t_0}\{\alpha_{nm}(t_1,k)\alpha_{00}^*(t_1,k)\}, \\ S_{\alpha_{00}\alpha_{00}}(t,k) &= \frac{1}{T_{\text{est}}}\sum_{t_1=t-t_0}^{t+t_0}\{\alpha_{00}(t_1,k)\alpha_{00}^*(t_1,k)\}, \end{aligned} \tag{51}$$

where T_{est}=$2t_0$+1 refers to the number of time-varying frames, approximating a statistical expectation over $[t-t_0,t+t_0]$ time frames at the k-th frequency bin. Essentially, we assume speech stationarity over T_{est} consecutive time frames, similarly to [40].

3.2. Proposed Model

The proposed work employs a CRNN-based model for DOA estimation. As stated in Sec. 2.3.1, the convolutional network is able to extract a data representation at a higher level. The extracted data representation contain spatial information, thus providing an improved classification of the features. The recurrent layer processes datasets in a sequential manner and learns from present and previous time steps. Hence, we can extract important information from temporal relations that can be useful to detect the signal source. For the aforementioned characteristics, the combination of both networks is widely exploited in recent works [9, 27] and used as the baseline for major challenges [1, 2].

The convolutional network and the recurrent network employed in the baseline are inspired by [8]. The network is fed with the spatial and spectro-temporal features extracted by the output of the feature extraction process. The dimension of the input feature is $T_f \times F \times C$, where T_f is the temporal dimension of the input feature, F is the number of mel bins, and C is the number of channels. In the proposed architecture, the spatial patterns are learned using multiple layers of 2D CNN, as we can see in Fig. 5 . Each convolutional layer has B filters of 3×3 dimensional receptive fields that operate along the time-frequency axis with a rectified linear unit (ReLU) activation. The time and frequency dimensions of the kernel allow the network to acquire intra-channel features suitable for DOA estimation. After each convolutional layer, the output activations are normalized using a 2D batch normalization layer, and the dimensionality is reduced using max-pooling along both the time and frequency axis. Time pooling is applied in order to have T samples equal to T_{lab} label time samples at the output of the network. Therefore, the output after the final layer of the CNN with B filters is of dimension $T \times 2 \times B$, where the reduced frequency dimension of 2 is the result of max-pooling across frequency dimension. The output from CNN is further reshaped to a T frame sequence of length $2B$ and fed to a bidirectional RNN, which is employed to learn temporal information from the CNN output features. The bidirectional RNN consists of GRU with I nodes each and tanh activation.

The RNN produces an output that is fed to the input of two separated FC networks, one for azimuth and one for elevation predictions. The number of output nodes in the final layer corresponds to the total number of azimuth and elevation classes, respectively.

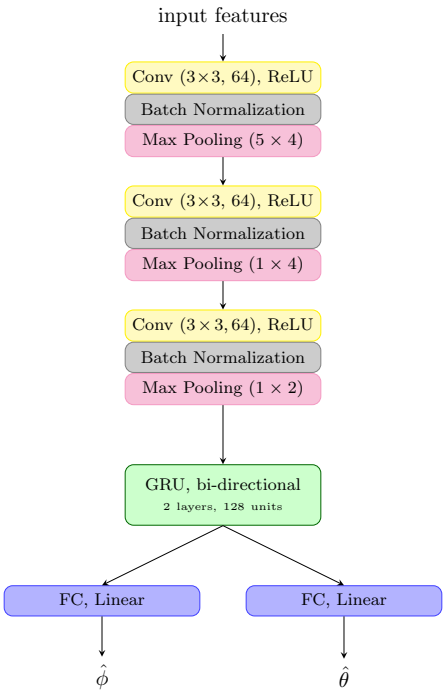


Figure 5: Block diagram of the proposed model architecture

3.3. Proposed Siamese Network

For triplet loss, we employ a siamese network composed of three identical networks, one for each element of the triplet. We employed the same CNN structure present in the model proposed in Sec. 3.2, which corresponds to the first three blocks in Fig. 5. As mentioned in Sec. 3.2, the input feature of the CNN has size $T_f \times F \times C$ and output of size $T \times 2B$ as a result of the time and frequency pooling. Then, the three output vectors are employed to compute triplet loss. In this case, we employed the hard margin triplet loss defined in (29) with margin $\delta = 2$ and cosine similarity. Then, the loss is back-propagated to the siamese network. Since the subnets share the weights we propagate the loss to one subnet and in the next iteration we will use that subnet to compute the output vectors. For greater understanding, in Fig. 6 we provide the architecture scheme of the siamese network employed in triplet loss training.

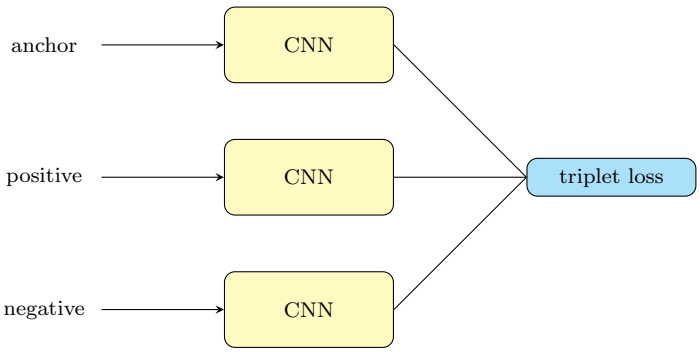


Figure 6: The siamese architecture employed in triplet loss training, where anchor, positive and negative features has size $T_f \times F \times C$ and the output features of the CNNs has size $T \times 2B$. The output features are utilized to evaluate triplet loss.

4. Performance Evaluation

4.1. Implementation Details

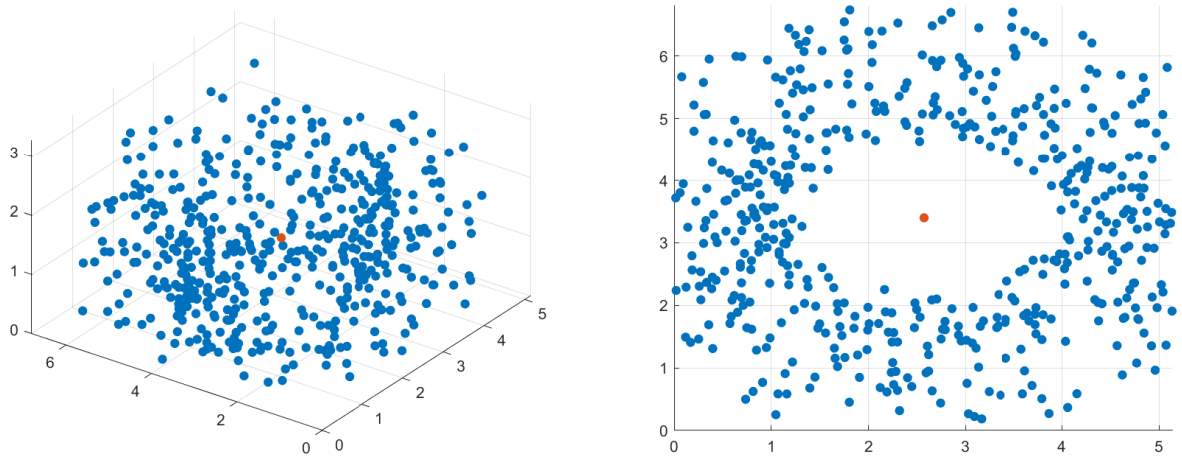
4.1.1 Dataset

In order to evaluate the model on different acoustic environments, we created a synthetic dataset. The dataset contains approximately 165 hours time-domain spherical harmonics up to the 4-th order. We generated the data by convolving simulated room impulse responses obtained from SMIR generator[1] [47] with speech signals. Speech signals have been taken randomly from a Librispeech [60] subset generated in the L3DAS23 dataset for Task 1 [33]. The considered dataset subset is composed of clean speech signals with a duration up to 12 s. The considered subset of Librispeech is divided into approximately 53% male and 47% female speech. We considered rooms with size randomly selected in the range $[4, 8] \times [5, 10] \times [3, 5]$m with uniform distribution. For room impulse response generation, we employed a simulated Eigenmike with 32 microphones and 4.2 cm radius. The microphone is positioned at the center of the room at height 1.3 m, which is the average ear height of a seated person. In order to simulate a realistic scenario, the SNR ranges from 5 to 60 dB randomly selected with a uniform distribution. Several sources are randomly positioned around the spherical array in order to obtain around 500 locations for each room. The DOA of the locations is in the range $\phi \in [0°, 360°]$ and $\theta \in [60°, 130°]$, with distance from the center of the microphone randomly chosen in the interval $[1.5, 3.5]$m with uniform distribution. As far as the reverberation time (RT_{60}) is concerned, we considered 16 values in the interval $[0.25, 1.0]$s with a step of 0.05 s. For spherical harmonic decomposition we used the tool [4] from the STFT of the simulated data. The STFTs of the simulated signals are obtained using a Hamming window of length 512 samples with 16 kHz of sampling frequency. For the training process, we selected samples from one of the simulated rooms. As shown in Fig. 7, the selected room has size $5.1 \times 6.8 \times 3.3$m and $RT_{60} = 0.5$ s. The simulated dataset parameters are summarized in Tab. 1.
In order to compute the input features, we apply a preprocessing stage to the generated signals. We consider the first 4 channels of the generated signals, which correspond to FOA signals. Therefore, the order of the RHC is limited to $N = 1$. The window length of the STFT is set so that as a result, the linear spectrum has one sample every 0.02 seconds. Then, similarly to [8, 9], we apply a log-mel transformation to the linear spectrum, considering 64 mel frequency bins.For the time average required by the RHC estimator (50), we consider a time window $T_{\text{est}} = 0.1$ s. Afterwards, we convert the RHC frequency axis into mel frequency bins, in order to have the same frequency dimensions as the log mel-spectrograms. Since a neural network is best suited to work with real data, we convert the 4-channel complex-valued RHC into an 8-channel real-valued feature. The log mel-spectrograms and the RHC features have the same frequency dimension. Hence, we can combine the extracted features by concatenating them, producing the input feature that is exploited during the training process. This data has a shape of $C \times T \times F$, with C=10 channels, 4 for log mel-spectrograms and 6 for RHC (3 relative harmonic coefficients represented through their real and imaginary part), T frames and F=64 mel bins.

[1] `http://github.com/ehabets/SMIR-Generator`

Simulated dataset	
Input Speech	Audio up to 12 s duration
Azimuth(ϕ)	$[0°, 360°]$ (randomly chosen)
Elevation(θ)	$[60°, 130°]$ (randomly chosen)
Size of room	$[4, 8] \times [5, 10] \times [3, 5]$ (randomly chosen)
Distance of source from microphone	$[1.5, 3.5]$ (randomly chosen)
RT_{60}	$[0.25, 1.0]$ (randomly chosen)
SNR	$[5, 60]$dB (randomly chosen)

Table 1: Simulating data parameters



(a) 3D source positions.

(b) Source distribution on xy plane.

Figure 7: (a) Distribution of the source positions in the space. In blue the sound source positions, in red the spherical array position. (b) Top view of the spherical array position (red dot) and the sound source positions (blue dots).

Since the magnitude of the features is sensible to the source signals, it may cause different orders of variance between the different types of extracted features and this may prevent the model from convergence. Therefore, after the feature extraction, we applied standardization in order to remove the mean and scale them to unit variance.
Since we do not have frame-by-frame annotations on the source activity, we extracted the labels exploited for the training process. In this work, we only considered speech sources. Therefore, to identify the source activity, we employ a voice activity detector (VAD) [3], which operates with a sampling time of 0.1 s. The VAD algorithm detects whether a speech signal is present or not during each time interval. For this task, we computed the clean mono version of each sample of the dataset to facilitate voice detection. Moreover, for each time interval where the voice activity is detected, we label the corresponding frame with the cartesian coordinates of the sound source related to the spherical array. These labels are then employed during the training phase to supervise the learning process of the model.

4.1.2 Training

Here, we introduce the configurations of the architectures employed in our experiments. For the 10 channel input features, we set the time sequence length $T_f = 10$ samples, considering 64 mel frequency bins. For the training of the baseline model, the batch size is set to 256. Therefore, one batch of data has size $256 \times 50 \times 64 \times 10$. The main framework of this work is the CRNN explained in Sec. 3.2. For our experiments, the configuration of the CRNN is similar to [9]. In detail, we added a dropout layer with a dropout rate of 0.2 to each convolutional layer of the CNN to prevent overfitting. In order to reduce the size of the features, we employ pooling layers after the convolutional layer. The kernel size in the temporal dimension of the pooling layers is $[5, 1, 1]$. As described in Sec. 4.1.1, the input features and the labels have different time resolutions: input features have a time hop length of 0.02 s and 0.1 s for labels. Here, with hop length, we refer to the time step between each time frame. Therefore, as shown in Fig. 5, a filter size of 5 is employed in the first pooling layer to ensure consistency in the time dimension between output features and labels. Instead, the size of the kernels in the frequency domain is $[4, 2, 2]$, respectively. The number of the filters of the CNN is set to 64. Therefore, the output vector of the CNN has size $256 \times 10 \times 2 \times 64$ and after the squeezing operation applied at the end of the CNN, the vector fed to the RNN has size $256 \times 10 \times 128$. The RNN consists of 2 layers with 128 nodes each and tanh activation, obtaining an output vector of $256 \times 10 \times 256$. In order to have more compact data, we reduce the last dimension of the vector from 256 to 128. Then, the feature vector is flattened to feed the FC networks. The source positions are separated in DOA bins with a sampling interval of 5° for both azimuth and elevation. The size of the output of the FC is equal to the number of classes considered. Furthermore, we consider the silence class, which comprehends all the frames where the speech signal is not active. Therefore, the output of the network is classified into 73 azimuth classes (72 for azimuth and 1 for silence) and 16 elevation classes (15 for elevation and 1 for silence). Since we are dealing with multi-class classification, we exploit cross entropy loss

for the training of network parameters, defined as

$$-\sum_{c=1}^{\mathcal{C}} \Omega_c \log(\mathcal{P}(\hat{\Omega}_c)), \tag{52}$$

where $\mathcal{C}$ is the number of classes, Ω_c and $\mathcal{P}(\hat{\Omega}_c)$ represent the label and the predicted probability for class c. In (52), $\log(\cdot)$ is the natural log operator.
We train the network for 300 epochs with early stop at 100 epochs. We adopt Adam optimizer with a momentum of 0.9. The learning rate value $LR = 5 \times 10^{-4}$ is halved if the validation loss does not improve within 25 epochs. For the triplet loss training, we employ the first part of the network that corresponds to the CNN of the aforementioned baseline. As for [17], in triplet loss learning, each batch has $\zeta \times \eta = 8 \times 16$ time sequences. ζ and η denote the number of different DOAs (classes) and the number of sequences per DOA, respectively. The anchor sample and the positive samples are randomly extracted from the same DOA class. Instead, we choose the negative instance from a random class. All the hyperparameters remain the same as in the CRNN training. Finally, we train again the proposed CRNN architecture. The CNN used in the CRNN and in the triplet loss training share the same dimensions and setup. Therefore, we use the CNN network employed in triplet loss training as the pre-trained model for the CNN of the last CRNN in order to start the training from the embeddings learned by the triplet loss.

4.1.3 Metrics

Metrics are measures employed to evaluate the effectiveness of a particular system. Metrics are important in providing insight and understanding the behavior of a method and can be exploited to identify areas of improvement. In deep learning, metrics are used to evaluate the effectiveness of the models in predicting outcomes. In this work, we are dealing with DOA classification. Therefore, we employed metrics that measure how the model is able to localize the source position. Hence, we considered the location recall, the gross error, and the mean absolute estimated error between the estimated and true DOAs.

Localization Recall Localization recall is a metric used to evaluate the performance of the DOA estimation models. The LR is calculated by comparing the ground truth DOA with the estimated DOA by the model. In this work, we consider true positive (TP) when for a frame t, the source signal is active, i.e. the ground truth class is not the silence class, and the estimated DOA is not silence. Furthermore, we refer to false negative (FN) as the estimations when the source signal is active, similar to the TP definition, but the silence class is the estimated class. In this context, we define LR as follows

$$\mathrm{LR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{53}$$

Therefore, the LR measures the ability of the model to recognize when the source is active. A low LR denotes that the model is not capable to detect when the signal source is active. On the contrary, a high LR represents the high proficiency of the model in discerning the instances where the source is active.

Gross Error In DOA estimation, the gross error metric is a measure of the performance of the DOA estimator in detecting the DOA correctly. In practical scenarios, the signal may be affected by noise and reflections, which can result in erroneous DOA estimates. Therefore, the GE measure is employed to evaluate the robustness of the proposed method to outliers. Similarly to [26] it is calculated as:

$$\mathrm{GE}_{\phi,\theta} = \frac{1}{Z_{\phi,\theta}} \sum_{z=1}^{Z_{\phi,\theta}} \Delta((|\phi_z - \hat{\phi}_z|, |\theta_z - \hat{\theta}_z|) - \lambda), \tag{54}$$

where the variables Z_ϕ and Z_θ represent the number of estimated azimuth and elevation values, respectively. Furthermore, ϕ_z, θ_z are the ground truth locations and $\hat{\phi}_z, \hat{\theta}_z$ are the predicted locations, as shown in Fig. 5. $\Delta(z)$ is the indicator function which takes the values of 0 when the argument z is less than 0, and 1 when z is greater equal to 0. $\lambda = 10°$ is the threshold which is considered based on classification of angles, similarly to [26]. Hence, for GE the $(\hat{\phi}, \hat{\theta})$ estimations that fall outside of the threshold λ are discarded. The GE is computed separately for azimuth and elevation. We average GE_ϕ and GE_θ in order to obtain a single metric value, which we can interpret as the percentage of correct DOA estimations that fall inside the range λ from the ground truth DOA. A smaller GE represents a high location effectiveness of the model, while larger GE shows the inability of the model to determine the correct DOA.

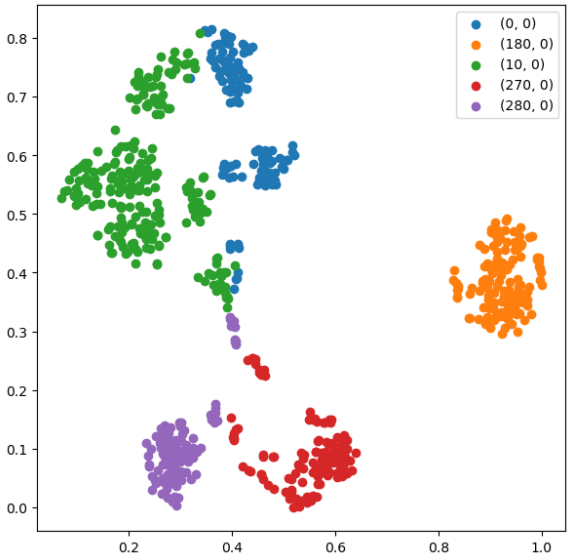
Mean Absolute Estimation Error In the context of DOA estimation, MAEE is employed to evaluate the performance of an estimator by measuring the average absolute difference between the estimated DOAs and the true DOAs. As in [40], the MAEE is defined as the mean of all the absolute differences between estimated values and reference values:

$$\text{MAEE} = \frac{1}{Z}\sum_{z=1}^{Z} |\phi_z - \hat{\phi}_z| + |\theta_z - \hat{\theta}_z|, \tag{55}$$

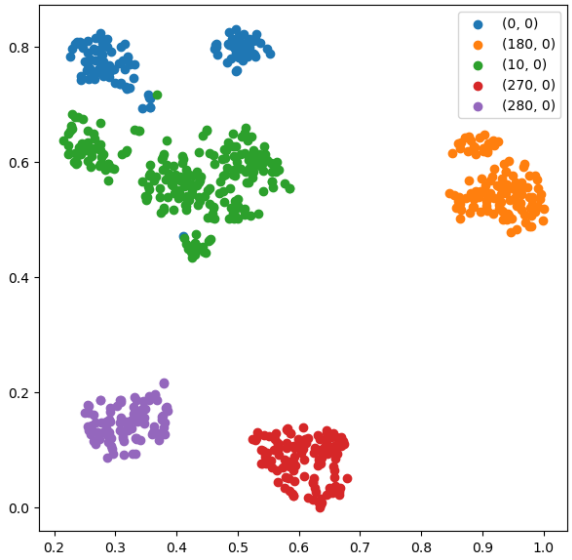
where Z is the total number of joint DOA estimations. Similarly to GE, MAEE measures the performance of the proposed method in estimating the correct DOA. In this case, the MAEE allows us to have a measure of the localization performance expressed in degrees. For this reason, we employed the MAEE in order to have a direct measurement of the localization performance, which can be easily interpreted. Similarly to GE, a smaller MAEE value represents a high location effectiveness of the model because the estimated DOAs are close to the real DOAs.

4.2. Visualization of learned embeddings

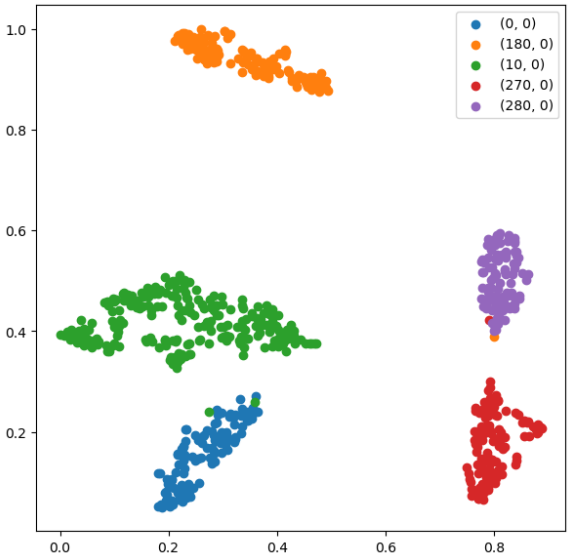
In this section, we discuss and compare the learned embeddings of the models proposed in this work: the free design embedding CRNN, the siamese network trained with triplet loss, and the pre-trained CRNN. For the evaluation, we randomly selected rooms from the simulated dataset without considering the room employed during the training stage. The motivation behind this decision was to demonstrate that the models can abstract the information from data of a single room. Furthermore, we apply T-SNE [56] method to visualize 2D projections of feature embeddings of test data learned from the training of the models. Then, we focus on the effectiveness of the siamese network training with triplet loss, comparing its feature embedding with the free-designed and pre-trained embedding. We expect that the feature embedding learned with triplet loss will result in an effective representation of the input data, where features belonging to the same class will be in the same cluster in the features space, and features of different classes will be more clearly separated.
Based on T-SNE visualization method, we can observe the learned feature embedding in Fig. 8. We represented 5 different DOA classes with different distances in the spatial domain. We can observe that all three considered models can separate the different DOA classes creating clusters. As expected, the free design embedding and the pre-trained embedding are comparable to each other. However, as we can see in Fig. 8c the feature embedding created exploiting triplet loss training results in a more separated feature embedding since the samples of each class are closer to the centroid of the class cluster. Therefore, as we expected, the triplet loss embedding has a better interpretability compared to the other embeddings. Indeed, we can observe that the features embedding of the free design and the pre-trained models are more sparse, creating sub-clusters for the same class. However, the pre-trained model can achieve similar performance as the freely trained network, while obtaining a less sparse features space. In this way, through pre-training with triplet loss, we can have a good DOA classification, while having a more interpretable embedding. It is worth noting that classes that are spatially close to each other in terms of DOA tend to remain close in the feature embedding space. This property of the learned representation can be leveraged to facilitate the classification and spatial interpretation of the features space.
In Fig. 9, we represented the entire test set where the samples have been divided into macro classes based on the azimuth values. As expected, the separation between the samples is more pronounced in the triplet loss feature embedding space. The division between the classes is clear for all three methods. However, the pre-trained appears to have a more homogeneous distribution in the features space thanks to the triplet loss pre-training. Interestingly, we notice that in Fig. 9c-d, the azimuth classes divide the feature embedding space into equal parts. The pre-trained network exhibits a clear division between the classes comparable to the DOA division in the spatial domain, i.e., indicating a stronger correlation between the features space and the spatial dimension.



(a) free design.

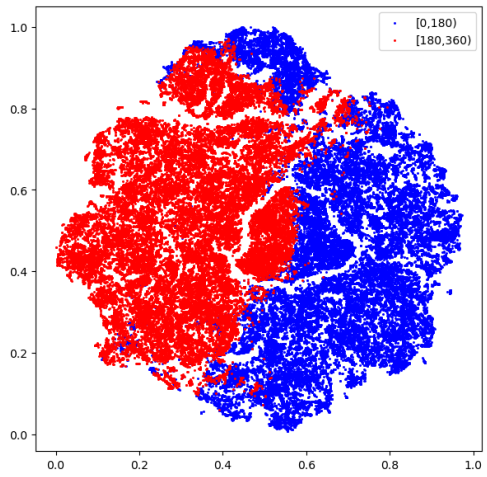


(b) pre-trained.

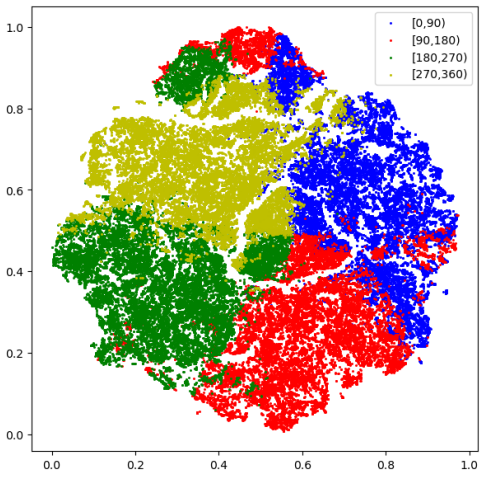


(c) triplet loss.

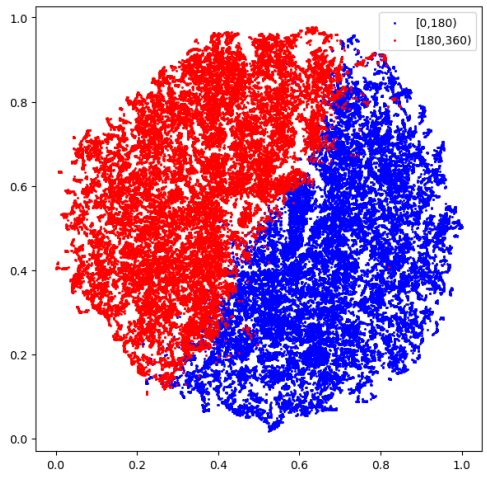
Figure 8: The T-SNE 2D visualization of the learned embeddings for 5 different DOA classes. The free design embedding and the pre-trained embedding are comparable. However, the free design embedding shows separated clusters that are closer to each other with respect to the other methods. Instead, the triplet loss model displays well-separated clusters, while the pre-trained model maintains the cluster separation given by the triplet loss model.



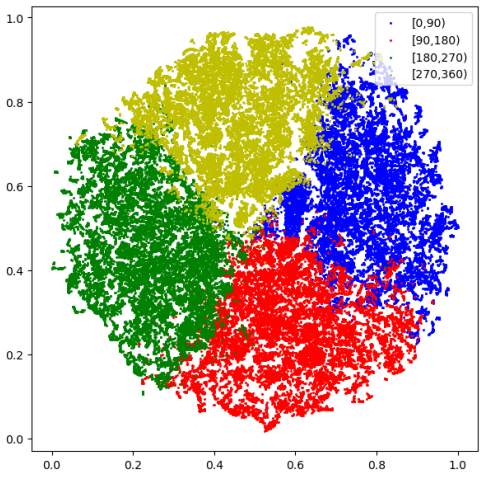
(a) free design.



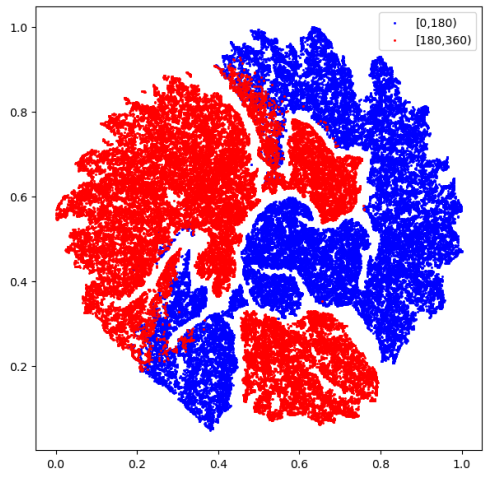
(b) free design.



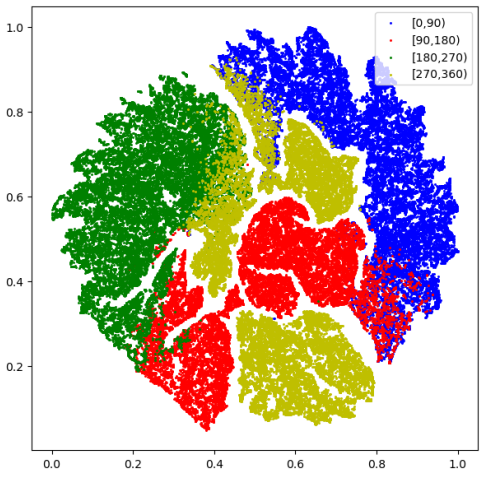
(c) pre-trained.



(d) pre-trained.



(e) triplet loss.



(f) triplet loss.

Figure 9: 2D T-SNE projection visualization of all the samples in the test set. (a,c,e) The samples has been divided in two macro classes based on the azimuth values: those with azimuth between 0° and 179° (blue) and those with azimuth between 180° to 359° (red). (b,d,f) The samples has been divided in four macro classes based on the azimuth values: $[0°, 90°]$ (blue), $[90°, 180°]$ (red), $[180°, 270°]$ (green), $[270°, 360°]$ (yellow).

	GE	LR	MAEE
free design	0.36	0.97	6.83
pre-trained	0.37	0.95	6.86

Table 2: The comparison of the three models presented in this work performance on unseen data

4.3. DOA Estimation Results

As shown in Tab. 2, the free design embeddings method and the triplet loss pre-trained methods have similar performance on the test set. Therefore, we selected the pre-trained model to compare the performance of the proposed approach, referenced in this section as PT-CRNN. The performance is evaluated in terms of GE and MAEE, and compared with the conventional method MUSIC [67, 75], RMUSIC [39], SHD-MUSIC [5, 50], and SHD-RMUSIC [39]. To perform a complete evaluation of the performance, we calculate both the azimuth GE GE_{ϕ} and the elevation GE GE_{θ}. By computing both of these metrics, we can obtain a comprehensive understanding of the performance of the method in both horizontal and vertical planes. The results are obtained for various azimuth and elevation angles. In the case of the GE metrics, the values were calculated by summing all the DOA estimation frames in the test set. For MAEE, instead, the average of all the observations is plotted.

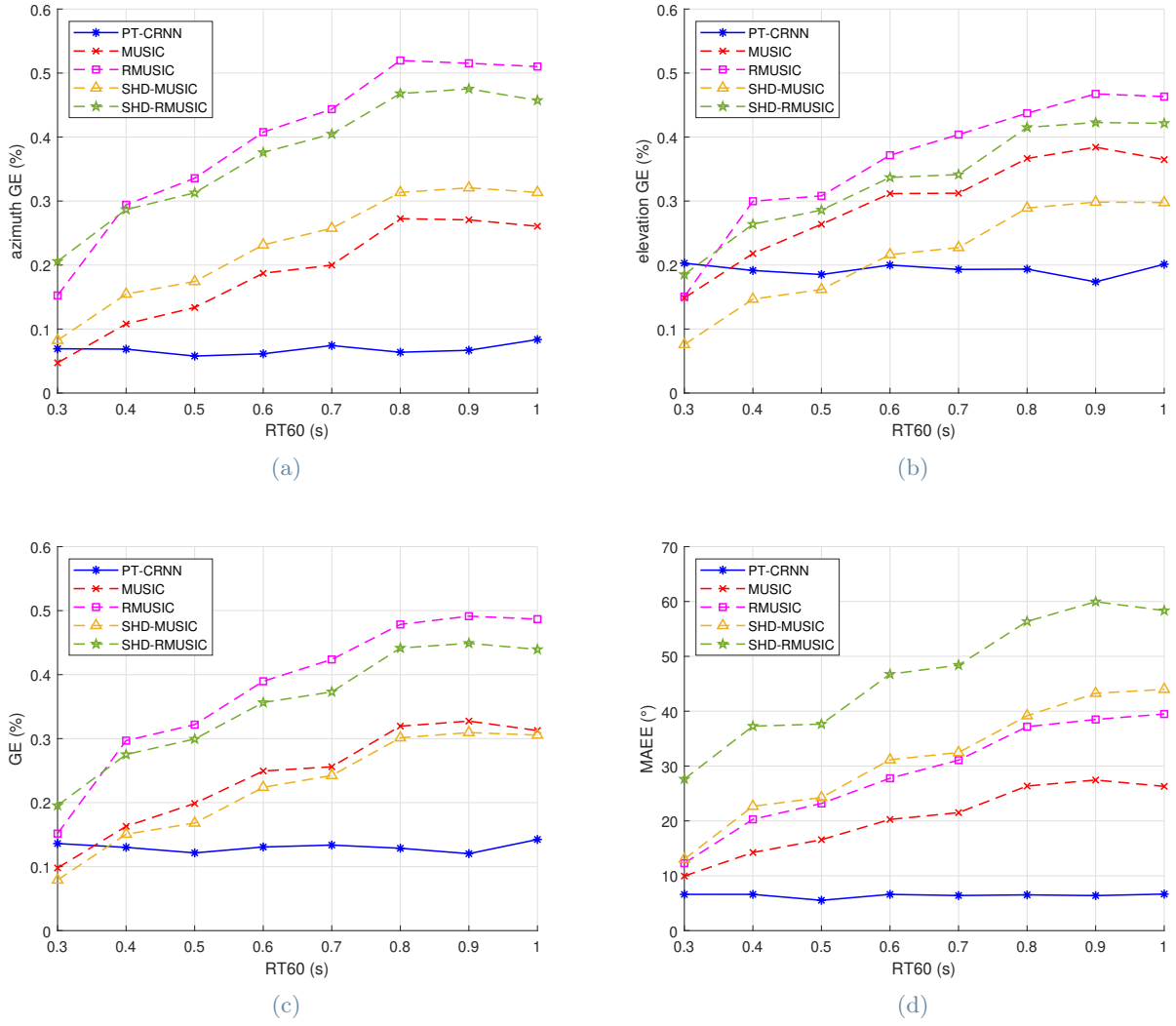


Figure 10: GE values against RT_{60} for (a) azimuth, (b) elevation and (c) combined axis. In (d), it is represented the MAEE performance against RT_{60}. The test samples has been selected for various azimuth and elevation, considering only samples with SNR>40 dB at different distances from the microphone array.

In Fig. 10, we illustrated GE_ϕ, GE_θ, GE, and MAEE by considering different rooms with different RT_{60} values. From the graphs in Fig. 10, it can be observed that the performance of PT-CRNN is superior to the performance of the subspace methods in terms of the GE metrics. Additionally, it can be seen that the proposed method exhibits a lower GE_ϕ than the GE_θ, indicating higher performance in the localization of the sound sources in the horizontal plane than in the vertical plane. In terms of MAEE, it can be observed that the technique proposed in this work has higher performance with respect to the other methods, suggesting that it is able to estimate the DOA more precisely than the considered methods. PT-CRNN, however, is the technique showing a more robust behaviour. We observe that the baseline methods present an important performance drop, while the performance of the proposed method is constant.
Then, we evaluated GE_ϕ, GE_θ, GE, and MAEE with varying SNR, selecting the rooms with RT_{60} from 0.45 s to 0.55 s. The results showed in Fig. 11 demonstrate that the PT-CRNN method performance is better to subspace methods in terms of both GE metrics and MAEE. As observed in the varying RT_{60} case, PT-CRNN presents lower azimuth GE than elevation GE, suggesting better localization in the azimuth plane than the elevation plane. In terms of MAEE, the proposed model outperforms the baselines, demonstrating the effectiveness of the system in DOA estimation under various SNR conditions. In this case, as well, the PT-CRNN method demonstrates a more robust behaviour, indicating a lower sensitivity to noise than the other methods.
Finally, in Fig. 12, we represented the considered localization metrics against the distance between the sound source and the center of the microphone array. The results demonstrate similar behaviour to previous tests, where the PT-CRNN method exhibits a more robust behaviour compared to the conventional methods considered in the evaluation. PT-CRNN outperforms subspace methods at every considered distance. The performance of the baseline methods degrade at higher source distance, where PT-CRNN shows a higher capacity to estimate

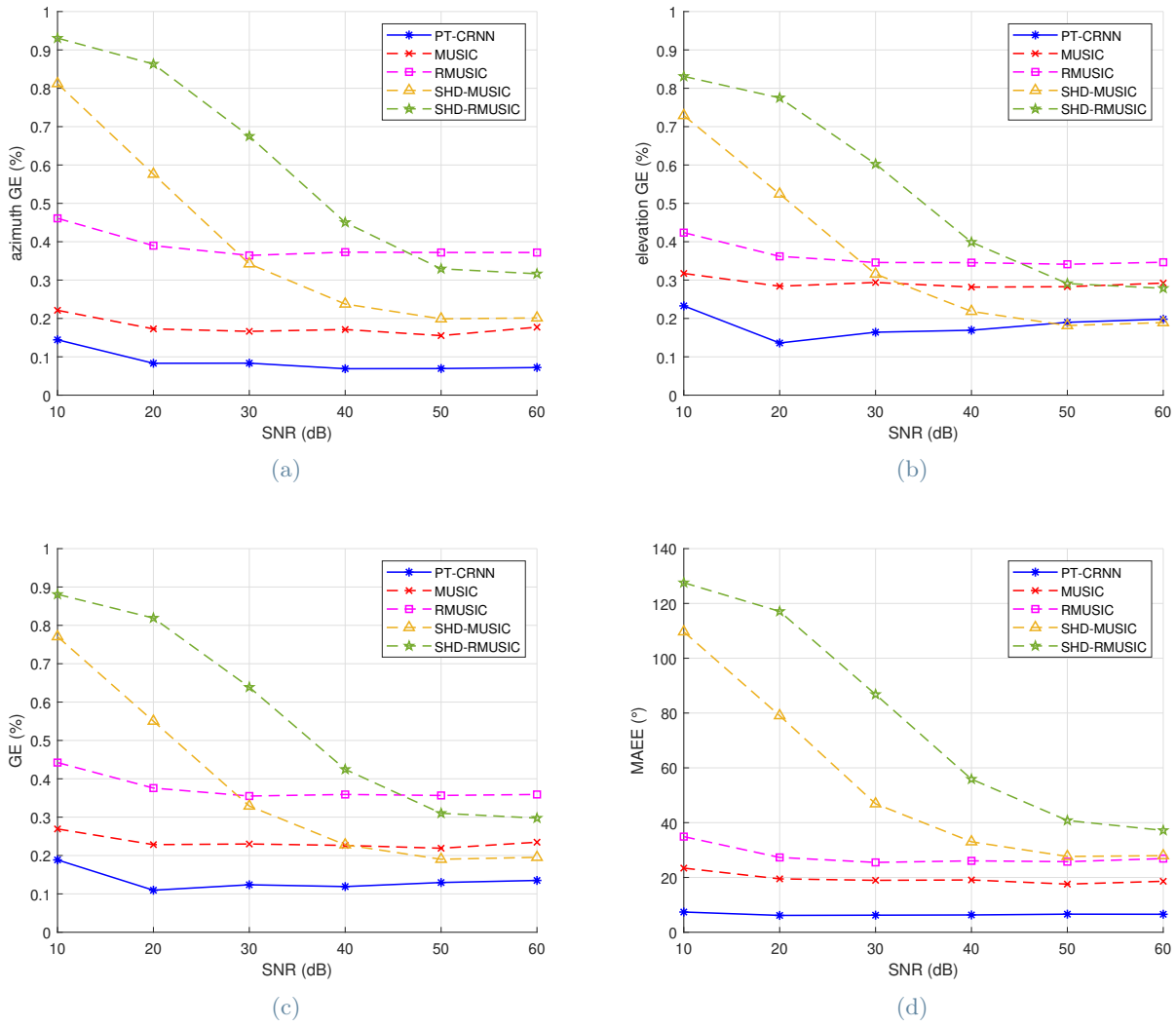


Figure 11: GE and MAEE values against SNR. The test samples has been selected for various azimuth and elevation, considering only rooms with RT_{60} between 0.45 s and 0.55 s.

the source DOA correctly.
In general, PT-CRNN has shown higher performance compared to the other methods for all the considered metrics. In the end, the model proposed in this work outperforms the the baseline approaches in non-ideal conditions of the acoustic scenario, i.e., low SNR and high reverberation, indicating more robust performance than the conventional methods.

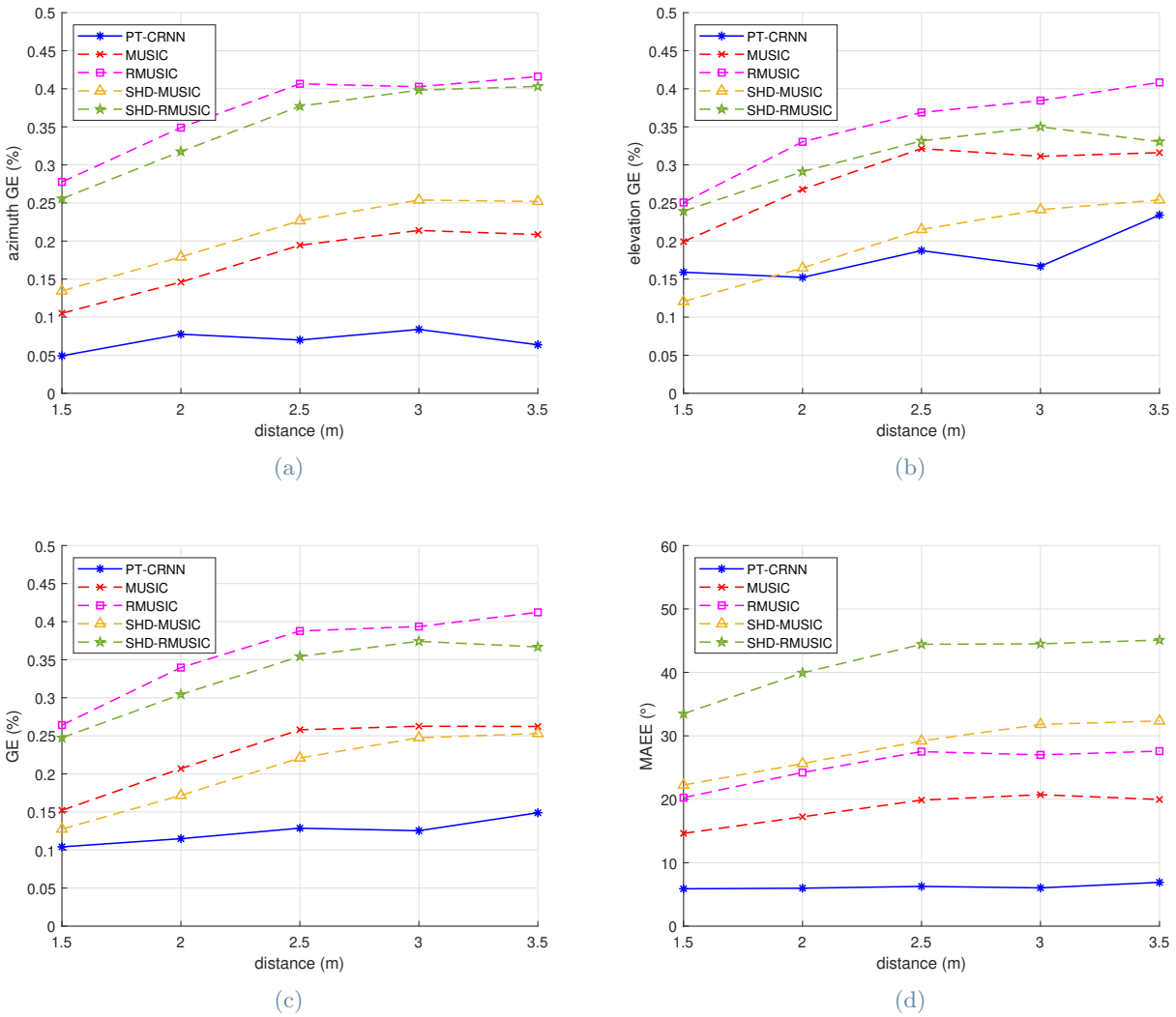


Figure 12: GE and MAEE values against the distance between the sound source and the microphone array. The test samples has been selected for various azimuth and elevation, considering only rooms with RT_{60} between 0.45 s and 0.55 s.

5. Conclusions and Future Works

In this thesis, we have developed a deep-learning based method for the sound source DOA estimation from SHD-based signals. SHD-domain based features can be extracted from microphone arrays with different geometries, making the proposed approach suitable for different microphone setups. We adopted a recently introduced feature representation known as RHC which provides meaningful spatial information about the sound field. In this context, we presented a siamese network architecture for the triplet loss training. Triplet loss promotes a structured and meaningful feature space with improved clustering of similar samples and better discrimination of dissimilar samples.

In order to tackle the DOA estimation problem, we divided the possible DOAs into azimuth and elevation classes. Then, inspired by recent works in the DOA estimation field [8, 26, 27], we proposed a CRNN-based architecture for the joint estimation of azimuth and elevation of the sound source location. The proposed model has been trained in a single room and tested in different rooms, thus, unseen during the training phase, demonstrating that the network is able to generalize the information from input features. As a result, the network showed improved performance in DOA estimation with respect to baseline methods in complex scenarios with low SNR and high RT_{60}.

Then, our solution involved the implementation of a CNN-based siamese neural network. The purpose of this neural network architecture was to employ triplet loss, in order to generate a structured features embedding. With the triplet loss, the network was able to learn a structured feature space where similar samples are close and dissimilar samples are separated. The structured embedding preserved a good correspondence with the spatial domain, i.e., data from two different DOA classes that are close in the spatial domain correspond to two different clusters close to each other in the feature embedding. We demonstrated that this property of the features space results in improved intelligibility.

Next, we trained the proposed CRNN architecture initializing the state of the CNN that composes the CRNN architecture with the weights learned during the triplet loss training. The simulations on the synthetic dataset showed that the pre-trained network is able to estimate the source position with similar performance with respect to the proposed CRNN network with free-designed embedding. The feature space of the pre-trained network demonstrates the beneficial effect of the network initialization with the triplet loss-trained network. The analysis of the features embeddings showed that the feature space of the pre-trained network has a structured correspondence with the spatial domain, indicating higher interpretability.

Finally, we presented the localization results by performing tests on rooms with different dimensions and reverberation times with respect to the room on which the model is trained. Regarding the metrics of both GE and MAEE, the proposed approach exhibited the capacity of the pre-trained network to generalize well when confronted with new, unseen data. In the comparison with subspace methods, the network showed a more robust performance, with minimal degradation of the performance in different acoustic scenarios.

We believe this works represents an exploratory step towards the possibilities of the triplet loss in the DOA estimation problem. The proposed method can be easily extended to the multi-source localization case and we intend to propose it in the near future. We expect that a more structured feature embedding can be exploited for multitask problems, such as sound event detection and localization, and speech enhancement. Furthermore, we suppose that a structured space could be useful for domain adaptation cases, for instance, when transitioning from simulated data to real-world recordings.

References

[1] Dcase challenge. http://dcase.community/.

[2] L3das. http://www.l3das.com.

[3] Spherical harmonics synthesis and analysis tools (shsat). http://github.com/wiseman/py-webrtcvad.

[4] Fahim A. Spherical harmonics synthesis and analysis tools (shsat). http://github.com/abdfahim/audioprocessing.

[5] T.D. Abhayapala and H. Bhatta. Coherent broadband source localization by modal space processing. In *10th International Conference on Telecommunications, 2003. ICT 2003.*, volume 2, pages 1617–1623 vol.2, 2003.

[6] Thushara D Abhayapala and Aastha Gupta. Spherical harmonic analysis of wavefields using multiple circular sensor arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1655–1666, 2009.

[7] Thushara D Abhayapala and Darren B Ward. Theory and design of high order sound field microphones using spherical microphone array. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1949. IEEE, 2002.

[8] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.

[9] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466, 2018.

[10] Sylvain Argentieri, Patrick Danes, and Philippe Souères. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112, 2015.

[11] Jacob Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *the Journal of the Acoustical Society of America*, 107(1):384–391, 2000.

[12] Carlos Busso, Sergi Hernanz, Chi-Wei Chu, Soon-il Kwon, Sung Lee, Panayiotis G Georgiou, Isaac Cohen, and Shrikanth Narayanan. Smart room: Participant and speaker localization and identification. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–1117. IEEE, 2005.

[13] Soumitro Chakrabarty and Emanuël AP Habets. Multi-speaker doa estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):8–21, 2019.

[14] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings 13*, pages 258–266. Springer, 2017.

[15] Xianyu Chang, Chaoqun Yang, Xiufang Shi, Pengfei Li, Zhiguo Shi, and Jiming Chen. Feature extracted doa estimation algorithm using acoustic array for drone surveillance. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2018.

[16] Shlomo E Chazan, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot. Multi-microphone speaker separation based on deep doa estimation. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[17] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

[18] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[19] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

[20] Maximo Cobos, Amparo Marti, and Jose J. Lopez. A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling. *IEEE Signal Processing Letters*, 18(1):71–74, 2011.

[21] David L Colton, Rainer Kress, and Rainer Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93. Springer, 1998.

[22] Emily E Cust, Alice J Sweeting, Kevin Ball, and Sam Robertson. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *Journal of sports sciences*, 37(5):568–600, 2019.

[23] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein. Robust localization in reverberant rooms. *Microphone arrays: signal processing techniques and applications*, pages 157–180, 2001.

[24] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[25] Karthikeyan Duraisamy, Ze J Zhang, and Anand Pratap Singh. New approaches in turbulence and transition modeling using data-driven techniques. In *53rd AIAA Aerospace sciences meeting*, page 1284, 2015.

[26] Priyadarshini Dwivedi, Raj Prakash Gohil, Gyanajyoti Routray, Vishnuvardhan Varanasiy, and Rajesh M Hegde. Joint doa estimation in spherical harmonics domain using low complexity cnn. In *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2022.

[27] Priyadarshini Dwivedi, Siddesh Bharat Hazare, Gyanajyoti Routray, and Rajesh M Hegde. Long-term temporal audio source localization using sh-crnn. In *2023 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2023.

[28] Abdullah Fahim, Prasanga N. Samarasinghe, and Thushara D. Abhayapala. Psd estimation and source separation in a noisy reverberant environment using a spherical microphone array. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1594–1607, 2018.

[29] Abdullah Fahim, Prasanga N Samarasinghe, and Thushara D Abhayapala. Multi-source doa estimation through pattern recognition of the modal coherence of a reverberant soundfield. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:605–618, 2019.

[30] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, 2017.

[31] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, 2015.

[32] Pierre-Amaury Grumiaux, Srdan Kitic, Laurent Girin, and Alexandre Guérin. A survey of sound source localization with deep learning methods. *CoRR*, abs/2109.03465, 2021.

[33] Eric Guizzo, Christian Marinoni, Marco Pennese, Xinlei Ren, Xiguang Zheng, Chen Zhang, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello. L3das22 challenge: Learning 3d audio sources in a real office environment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9186–9190. IEEE, 2022.

[34] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[35] Hedi Harzallah, Frédéric Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *2009 IEEE 12th international conference on computer vision*, pages 237–244. IEEE, 2009.

[36] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32:582–596, 2019.

[37] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[38] Jinbo Hu, Yin Cao, Ming Wu, Qiuqiang Kong, Feiran Yang, Mark D. Plumbley, and Jun Yang. Sound event localization and detection for real spatial sound scenes: Event-independent network and data augmentation chains. Technical report, DCASE2022 Challenge, June 2022.

[39] Yonggang Hu, Thushara D. Abhayapala, and Prasanga N. Samarasinghe. Multiple source direction of arrival estimations using relative sound pressure based music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:253–264, 2021.

[40] Yonggang Hu, Thushara D. Abhayapala, Prasanga N. Samarasinghe, and Sharon Gannot. Decoupled direction-of-arrival estimations using relative harmonic coefficients. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 246–250, 2021.

[41] Yonggang Hu, Prasanga N. Samarasinghe, and Thushara D. Abhayapala. Sound source localization using relative harmonic coefficients in modal domain. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 348–352, 2019.

[42] Yonggang Hu, Prasanga N. Samarasinghe, Thushara D. Abhayapala, and Sharon Gannot. Unsupervised multiple source localization using relative harmonic coefficients. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–575, 2020.

[43] Yonggang Hu, Prasanga N. Samarasinghe, Sharon Gannot, and Thushara D. Abhayapala. Semi-supervised multiple source localization using relative harmonic coefficients under noisy and reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3108–3123, 2020.

[44] Yonggang Hu, Prasanga N. Samarasinghe, Sharon Gannot, and Thushara D. Abhayapala. Decoupled multiple speaker direction-of-arrival estimator under reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3120–3133, 2022.

[45] Y. LeCun E. Säckinger J. Bromley, I. Guyon and R. Shah. Signature verification using a 'siamese' time delay neural network. *Advances in Neural Information Processing Systems*, 6, 1993.

[46] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous. Unsupervised learning of semantic audio representations. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 126–130. IEEE, 2018.

[47] Daniel P Jarrett, Emanuel AP Habets, Mark RP Thomas, and Patrick A Naylor. Rigid sphere room impulse response simulation: Algorithm and applications. *The Journal of the Acoustical Society of America*, 132(3):1462–1472, 2012.

[48] Byeongho Jo and Jung-Woo Choi. Direction of arrival estimation using nonsingular spherical esprit. *The Journal of the Acoustical Society of America*, 143(3):EL181–EL187, 2018.

[49] Sameh Khamis, Cheng-Hao Kuo, Vivek K Singh, Vinay D Shet, and Larry S Davis. Joint learning for attribute-consistent person re-identification. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*, pages 134–146. Springer, 2015.

[50] Dima Khaykin and Boaz Rafaely. Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 221–224, 2009.

[51] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[52] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. 2015.

[53] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[54] Muhammad Usman Liaquat, Hafiz Suliman Munawar, Amna Rahman, Zakria Qadir, Abbas Z Kouzani, and MA Parvez Mahmud. Localization of sound sources: A systematic review. *Energies*, 14(13):3910, 2021.

[55] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, 2017.

[56] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[57] Dejan Markovic, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro. Soundfield imaging in the ray space. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2493–2505, 2013.

[58] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001.

[59] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655, 2017.

[60] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[61] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin. Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):22–33, 2019.

[62] Lam Pham, Ian McLoughlin, Huy Phan, Ramaswamy Palaniappan, and Alfred Mertins. Deep feature embedding and hierarchical classification for audio scene classification. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[63] B. Rafaely. *Fundamentals of Spherical Array Processing.* Academic Press, 1999.

[64] Fuji Ren and Siyuan Xue. Intention detection based on siamese neural network with triplet loss. *IEEE Access*, 8:82242–82254, 2020.

[65] P.N. Samarasinghe, T.D. Abhayapala, M. Poletti, and T. Betlehem. An efficient parameterization of the room transfer function. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2217–2227, 2015.

[66] Prasanga N Samarasinghe, Hanchi Chen, Abdullah Fahim, and Thushara D Abhayapala. Performance analysis of a planar microphone array for three dimensional soundfield analysis. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 249–253. IEEE, 2017.

[67] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[68] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[69] Petre Stoica and Arye Nehorai. Music, maximum likelihood, and cramer-rao bound. *IEEE Transactions on Acoustics, speech, and signal processing*, 37(5):720–741, 1989.

[70] Ryu Takeda and Kazunori Komatani. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 405–409. IEEE, 2016.

[71] Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, volume 2018, pages 3229–3233, 2018.

[72] H. Teutsch. *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, volume 348. Springer, Berlin, Germany, 2007.

[73] Masahito Togami, Shunsuke Suganuma, Yohei Kawaguchi, Takayuki Hashimoto, and Yasunari Obuchi. Transient noise reduction controlled by doa estimation for video conferencing system. In *2009 IEEE 13th International Symposium on Consumer Electronics*, pages 26–29, 2009.

[74] Nicolas Turpault, Romain Serizel, and Emmanuel Vincent. Semi-supervised triplet loss based learning of ambient audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 760–764. IEEE, 2019.

[75] Pascal Vallet, Xavier Mestre, and Philippe Loubaton. Performance analysis of an improved music doa estimator. *IEEE Transactions on Signal Processing*, 63(23):6407–6422, 2015.

[76] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[77] Qing Wang, Li Chai, Huaxin Wu, Zhaoxu Nian, Shutong Niu, Siyuan Zheng, Yuyang Wang, Lei Sun, Yi Fang, Jia Pan, Jun Du, and Chin-Hui Lee. The nerc-slip system for sound event localization and detection of dcase2022 challenge. Technical report, DCASE2022 Challenge, June 2022.

[78] Ruishuang Wang, Zhao Li, Jian Cao, Tong Chen, and Lei Wang. Convolutional recurrent neural networks for text classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2019.

[79] E.G. Williams. *Fourier acoustics: sound radiation and nearfield acoustical holography*. Springer Cham, 2 edition, 2 2019.

[80] Angeliki Xenaki, Jesper Bünsow Boldt, and Mads Græsbøll Christensen. Sound source localization and speech enhancement with sparse bayesian learning beamforming. *The Journal of the Acoustical Society of America*, 143(6):3912–3921, 2018.

[81] S. Gannot T.D. Abhayapala Y. Hu, P.N. Samarasinghe. Evaluation and comparison of three source direction-of-arrival estimators using relative harmonic coefficients. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 815–819, 2021.

[82] Kung Yao, Joe C. Chen, and Ralph E. Hudson. Maximum-likelihood acoustic source localization: Experimental results. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages III–2949–III–2952, 2002.

[83] Ozgur Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 52(7):1830–1847, 2004.

[84] Dong Yu and Li Deng. *Automatic speech recognition*, volume 1. Springer, 2016.

[85] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

Abstract in lingua italiana

La localizzazione di sorgenti sonore in ambienti rumorosi e riverberanti è ancora una questione aperta nel campo dell'elaborazione dei segnali. Tipicamente, l'identificazione della direzione di arrivo di un suono viene eseguita a partire da una registrazione multicanale. L'informazione della posizione di una sorgente sonora può essere fondamentale in diverse applicazioni, come il riconscimento di una voce o di un altoparlante, sorveglianza audio, realtà virtuale e aumentata. Recenti approcci al problema sono basati su modelli che sfruttano una particolare trasformazione dei segnali nel dominio delle armoniche sferiche, chiamati coefficienti armonici relativi. Altri recenti approcci propongono tecniche di deep learning per affrontare la stima della posizione della sorgente sonora, apprendendo le sue caratteristiche da reti neurali. In questo elaborato, proponiamo un nuovo metodo per la classificazione della direzione di arrivo esplorando la rete neurale convoluzionale ricorrente attraverso l'impiego dei coefficienti armonici relativi. In modo da classificare simultaneamente orientamento e ed elevazione della sorgente sonora, la parte finale della rete convoluzionale ricorrente è composta da due reti fully connected indipendenti. Successivamente, presentiamo una rete neurale siamese allenata con la tecnica nota come triplet loss. L'allenamento con la triplet loss permette alla rete di apprendere una rappresentazione strutturata dei dati, organnizzando i campioni della stessa classe vicini tra loro e allo stesso tempo separando i campioni di classi diverse. A tal proposito, abbiamo dimostrato che impiegando la triplet loss nell'allenamento della rete neurale, la rete è capace di localizzare la sorgente acustica in modo efficace anche in simulazioni con un basso rapporto segnale-rumore e un alto tempo di riverberazione. Gli esperimenti effettuati confermano che l'approccio proposto in questo elaborato producono una rappresentazione dei dati meno sparsa, implicandone una superiore interpretabilità. Infine, le prestazioni del metodo proposto sono confrontati con i risultati di metodi convenzionali, esibendo una maggiore robustezza in presenza di riverbero e rumore, e una maggiore accuratezza nella localizzazione della sorgente sonora.

Parole chiave: localizzazione, deep learning, coefficienti armonici relativi, triplet loss