POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Music and acoustic Engineering

Author: Luca Cattaneo

Advisor: Prof. Fabio Antonacci

Co-advisor: Mirco Pezzoli

Academic year: 2021-2022

1. Introduction

Nowadays, the number of applications based on sophisticated audio systems is increasingly spreading and the capacity to provide high audio quality is more and more significant. Smart speakers, voice assistants, teleconferencing systems, and virtual and augmented reality are all examples of such applications. In this context, the ability to extract the so-called direction of arrival (DOA) from multichannel recordings is crucial for various applications such as speech enhancement, speech separation, and efficient noise reduction techniques.

The DOA estimation is still an open problem in the field of signal processing that typically concerns the localization of acoustic sources from microphone array acquisitions. Conventional solutions of source localization rely on signal processing making assumptions about the statistics of both the target signal and noise. A popular class of source localization methods is represented by subspace methods, whose most popular approach is multiple signal classification (MUSIC). In recent years, the availability of microphone arrays with a high number of sensors raised the adoption of different sound field transformations, such as the spherical harmonic decomposition. As a result, several source localization techniques have been modified to work in the spherical harmonic domain (SHD), such as SHD-MUSIC. However, these techniques are susceptible to degraded performance with low signal-to-noise (SNR) ratios and reverberation. To address this problem, inspired by the relative transfer function, in [3] were introduced the relative harmonic coefficients (RHC) as valuable features for localizing sources in the SHD.

Researchers have increasingly employed machine learning, including deep-learning approaches, to solve problems of acoustic signal processing such as DOA estimation. In [1], the authors proposed a convolutional recurrent neural network (CRNN), exploiting both magnitude and phase information of the STFT coefficients to perform joint sound event detection and localization. It is shown that the CRNN model provides good performance for solving the localization problem. Fahim *et al.* [2] proposed a deep-learning method based on measured spherical harmonics coefficients. The paper shows that the proposed framework outperforms conventional methods.

In this work, we propose a CRNN-based framework for the joint classification of azimuth and elevation of the DOA. Differently from previous solutions, we present the siamese neural network for triplet loss training. The main idea behind triplet loss is to create a feature embedding where samples from the same class are clustered towards the same point, while instances of different classes are separated. Afterwards, we employed the network trained with triplet loss as the pre-trained model for the CRNN-based network, in order to obtain a more refined and structurally organized feature embedding. For the training and the evaluation of the models, we create a synthetic dataset composed of simulated acoustic environments with different dimensions and reverberation times. We present the analysis of the feature embeddings created by the three networks. Finally, we compared the localization performance of the proposed method with respect to conventional techniques.

2. Proposed Methods

2.1. RHC estimation

We consider a spherical microphone array multichannel signal decomposed into the SHD. The RHC are ideally defined as the ratio between the spherical harmonic coefficient α_{nm} and α_{00},

$$\beta_{nm}(t,k) = \frac{\alpha_{nm}(t,k)}{\alpha_{00}(t,k)}, \tag{1}$$

with order n and mode m at frequency bin k. In (1), dividing by the omnidirectional component α_{00}, we remove the source contribution from the spherical coefficients. Therefore, RHC are not affected by the time-varying source signal and are only determined by the DOA of the sound source. Actually, the array signal is corrupted by noise, hence we adopt the estimator of RHC [3] over the STFT domain as follows

$$\tilde{\beta}_{nm}(k) \approx \frac{S_{\alpha_{nm}\alpha_{00}}(k)}{S_{\alpha_{00}\alpha_{00}}(k)}, \tag{2}$$

where $S_{\alpha_{00}\alpha_{00}}(t,k)$ and $S_{\alpha_{nm}\alpha_{00}}(t,k)$ are respectively the power spectral density (PSD) and the cross-PSD of the measured spherical harmonic coefficients. The RHC in (2) are a robust estimation in presence of noise, making it suitable for application in realistic scenarios.

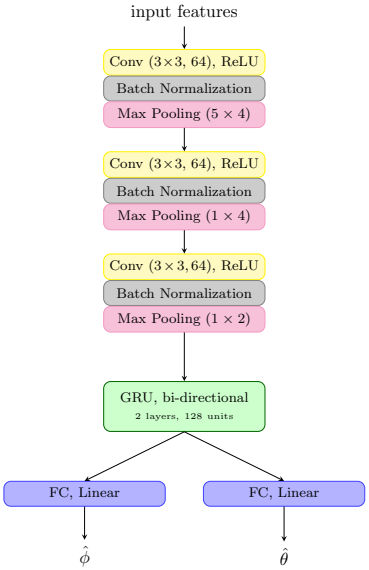


Figure 1: Block diagram of the proposed model architecture.

2.2. Proposed model

Inspired by the model in [1], we propose a CRNN-based model for joint azimuth and elevation estimation. This network receives the RHC features as input at the first convolutional layer. The first part of the model is composed of a convolutional neural network (CNN). The CNN has three sublayers, consisting of a 2D convolutional layer with ReLU activation, a batch normalization layer, and a max pooling layer. The output of the CNN is fed to a bidirectional recurrent neural network (RNN), which is employed to learn temporal information from the received features. For the proposed model, we consider a RNN involving two bi-directional gated recurrent units (GRU) layers with tanh activation. The final section of the CRNN consists in two distinct fully connected (FC) networks, which receive the same feature vector from the RNN. Hence performing the localization of the sound source, each of the FC branches is employed either for the elevation or the azimuth estimation. During the training phase, we exploit the cross entropy loss to implement multi-class classification. The outline of the proposed model is represented in Fig. 1.

2.3. Proposed training strategy

A siamese neural network is a class of neural network architectures that consists of two or

more identical neural networks. The subnetworks share the same structure and configuration. The input received by the network is composed of three samples, an anchor, a positive, and a negative. The anchor and the positive samples are extracted from the same class. Instead, the negative instance is selected from a different class. For this reason, the proposed siamese network comprises three identical CNN. The CNN architecture is the same as the one presented in Sec. 2.2, represented by the first three blocks of the CRNN in Fig. 1. In order to compare the outputs of the subnets, we employed the hard margin triplet loss, defined as

$$\mathcal{L}_{triplet} = [\delta + \mathcal{S}(\boldsymbol{\sigma}_a, \boldsymbol{\sigma}^-) - \mathcal{S}(\boldsymbol{\sigma}_a, \boldsymbol{\sigma}^+)]_+, \quad (3)$$

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 the cosine similarity function and the $[\cdot]_+$ operator is the hinge function $\max(\cdot, 0)$. The aim of triplet loss is to maximize the similarity between samples of the same class while minimizing the similarity between the samples of different classes.

3. Performance Evaluation

3.1. Implementation details

3.1.1 Simulated Dataset

For the training and evaluation of the model, we considered a dataset composed of approximately 165 hours of simulated audio recordings. We generated the data by convolving spherical microphone array impulse response obtained from SMIR generator[1] with speech signals extracted from the Librispeech dataset. We simulated rooms with sizes randomly selected in the range $[4, 8] \times [5, 10] \times [3, 5]$m with uniform distribution. We positioned the spherical array in the center of each room at a height of 1.3 m, which is the average ear height of a seated person. A 4-th order spherical array with 32 microphones and 4.2 cm radius is used. To simulate the thermal noise of the microphones, we employed additive white Gaussian noise with variance set to have a SNR in the range from 5 to 60 dB. For each room, we simulated around 500 random source positions. The azimuth and the elevation of each source location are in the range $\phi \in [0°, 360°]$ and $\theta \in [60°, 130°]$ with distance from the center of the array 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 distinct values in the interval $[0.25, 1.0]$s with a step of 0.05 s. For the training process, we selected samples from one of the 120 simulated rooms. The selected room has size $5.1 \times 6.8 \times 3.3$ m and $RT_{60} = 0.5$. Instead, the evaluation of the model was calculated on a test set composed of 3 randomly chosen rooms, with different sizes and RT_{60} with respect to the one employed for the training.

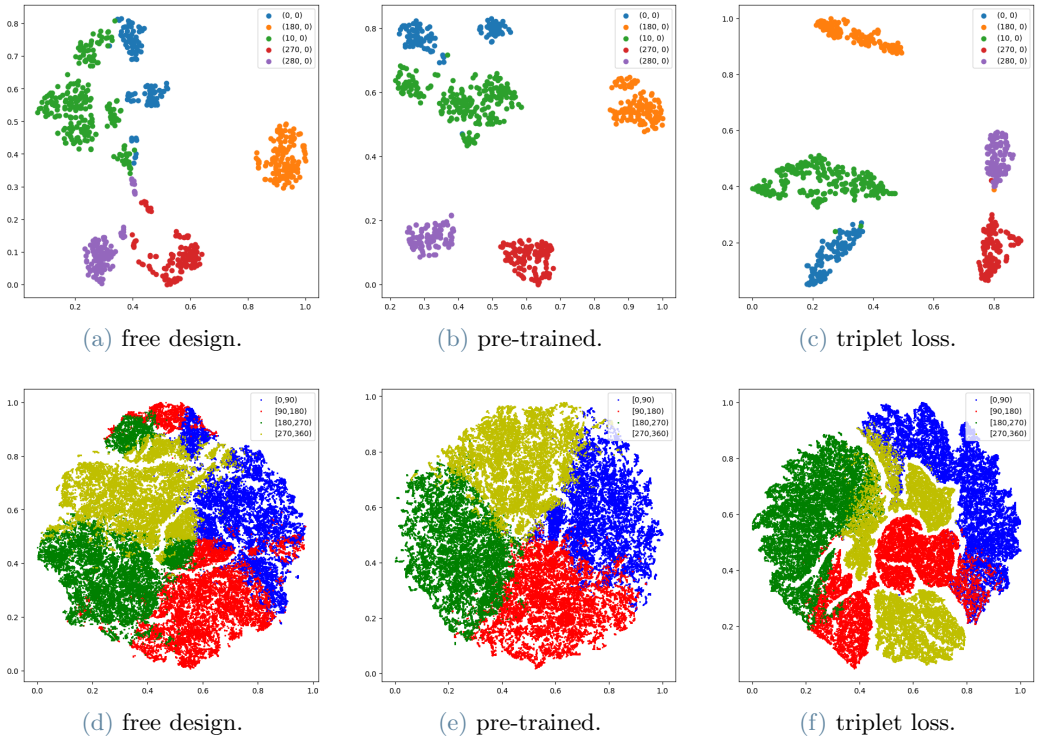
[1] `http://github.com/ehabets/SMIR-Generator`

3.1.2 Training

Pre-processing Before entering the training loop, the simulated data are pre-processed. Since we are considering the spherical harmonics up to the first order, we select the first 4 channels of the simulated recordings. The time domain data are transformed in the STFT domain. For each of the 4 channels, we computed the log mel-spectrogram and the estimate RHC with the biased estimator described in Sec. 2.1. From the RHC definition (2), the 0-th order RHC is always equal to 1. Therefore, we consider only the last three channels of the RHC. Then, in order to have the same frequency dimension of the log mel-spectrograms, we convert the linear frequency axis of the RHC in mel frequency bins. The network is fed with the complex RHC features represented in terms of real and imaginary parts.

For the training stage, we set the time sequence length of the input features to $T_f = 50$ samples and the number of mel-frequency bins to 64. The training can be divided into three main stages: (1) we train the proposed model in Sec. 2.2; (2) we train the siamese network with triplet loss; (3) we train the CRNN model exploiting CNN network trained in stage (2) as pre-trained network, exploiting the pre-trained model as the initial state of the training, allowing the model to optimize the classification embeddings.

The models are trained using Adam optimizer. The starting learning rate is set to 5×10^{-4} and it is halved if the validation loss does not decrease within 25 consecutive epochs. The training is set to 300 epochs, and it is stopped if the validation loss does not improve within 100 epochs. Regarding training of the siamese network



(a) free design. (b) pre-trained. (c) triplet loss.

(d) free design. (e) pre-trained. (f) triplet loss.

Figure 2: (a-c) The T-SNE 2D representation of 5 different DOA classes. (d-f) Representation of the evaluation dataset divided in 4 macro classes based on the azimuth quadrant.

model, the anchor sample and the positive samples are extracted randomly from the same DOA class. Instead, we choose the negative feature vector by selecting a sample of a random class. The triplets are organized in batches of $\zeta \times \eta = 8 \times 16$ time sequences, where ζ and η denote the number of different DOAs (classes) considered in a single batch and the number of different sequences per DOA class, respectively. We considered the triplet loss function in (3), with margin $\delta = 2$.

3.1.3 Metrics

In order to evaluate the effectiveness of the model, we employed metrics that measure the performance of sound source localization systems. Hence, the effectiveness of the proposed models is computed using gross error (GE) and mean absolute estimated error (MAEE).

Gross error The GE metric is a measure of the performance of the DOA estimator in detecting the correct DOA. The GE is defined as

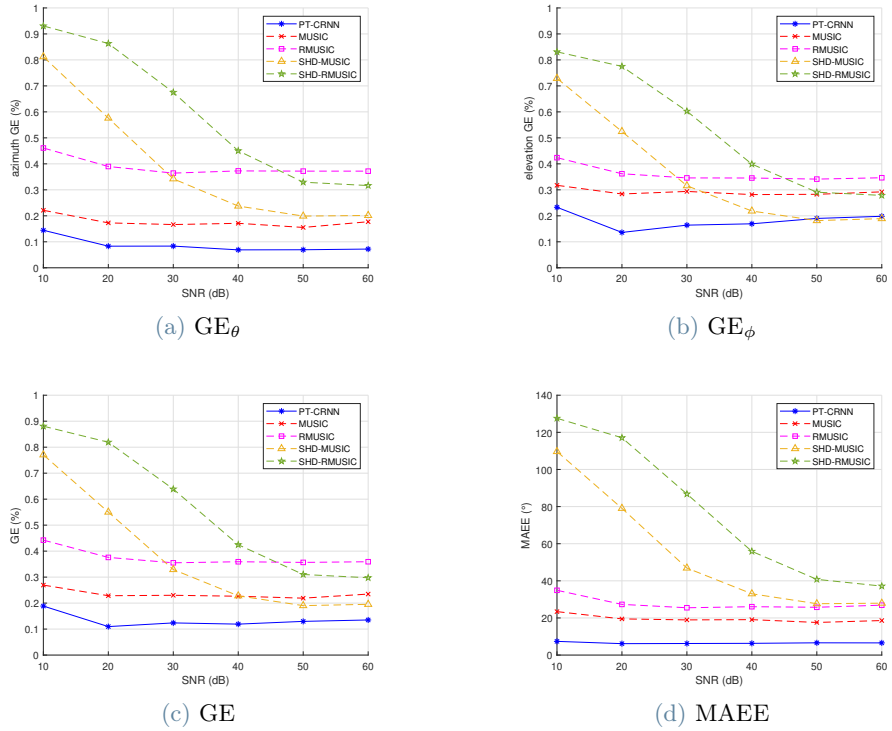
$$\mathrm{GE}_\Omega = \frac{1}{Z_\Omega} \sum_{z=1}^{Z_\Omega} \Delta((|\omega - \hat{\omega}|) - \lambda)\,, \qquad (4)$$

where Z_Ω represents the number of estimated DOAs. Furthermore, ω_z and $\hat{\omega}_z$ are the ground truth and the estimated DOAs, respectively. $\Delta(z)$ is the indicator function which takes the values of 0 when its argument is less than 0, and 1 when is greater or equal to 0. Therefore, the DOA estimation is considered correct when the distance between the ground truth and the estimated angles is lower than $\lambda = 10°$.

Mean absolute estimation error The MAEE is the measure expressed in degrees of the average error between the estimated DOAs and the ground truth. The MAEE is defined as the mean of all the absolute differences between estimated and real values:

$$\mathrm{MAEE} = \frac{1}{Z} \sum_{z=1}^{Z} |\phi_z - \hat{\phi}_z| + |\theta_z - \hat{\theta}_z|, \qquad (5)$$

where Z is the total number of joint DOA estimations, ϕ and θ are the ground truth azimuth and elevation, and $\hat{\phi}$ and $\hat{\theta}$ are the predicted azimuth and elevation.



(a) GE_θ (b) GE_ϕ

(c) GE (d) MAEE

Figure 3: GE_ϕ, GE_θ, GE and MAEE performance comparison against SNR.

3.2. Visualization of the learned embeddings

We analyze the learned embeddings for three proposed methods. For the evaluation, we selected rooms from the simulated dataset discarding the room employed during the training stage. We apply the T-SNE method to visualize the 2D feature embedding of test data. In Fig. 2 we can observe the learned feature embedding. From the T-SNE representation in Fig. 2a-c, we can observe that for all the three considered models, we can identify the different DOA classes. As expected, 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. Therefore, as expected, the triplet loss embedding has better interpretability compared to the other embeddings. The pre-trained model achieves similar performance as the free design embedding network, but with a less sparse feature space. Therefore, pre-training with triplet loss results in an effective DOA classification and a more interpretable embedding. In Fig. 2d-f, we represented the entire test set and divided the samples into four main classes based on the azimuth values. In the feature embedding space the distinction between the classes is evident in all three methods, but in triplet loss, the samples are more clearly separated. Interestingly, we notice that in Fig. 2e, the azimuth classes divide the feature space into four equal parts. Moreover, the pre-trained model displays a structured space that can be related to the direction of arrival (DOA) division in the spatial domain, indicating a more robust correlation between the feature space and the spatial dimension.

3.3. DOA estimation results

We compare the DOA classification performance of the proposed pre-trained model with the baseline. The localization performance is evaluated in terms of GE and MAEE, and compared with the conventional method MUSIC, RMUSIC, SHD-MUSIC, and SHD-RMUSIC. The results are computed for various azimuth and elevation angles. For the GE metrics, the values were obtained by summing all DOA estimation frames in the test set, while for MAEE

the reported values represent the average of all observations. In Fig. 3, we demonstrated azimuth GE (GE_ϕ), elevation GE (GE_θ), GE, and MAEE by evaluating various rooms with different SNR levels. We have evidence that the performance of PT-CRNN is superior with respect to the subspace methods in both GE and MAEE metrics, suggesting that the pre-trained model is able to estimate the DOA more precisely than the baseline methods. Furthermore, PT-CRNN model exhibits constant performance over different SNR levels, while for the baseline methods, we observe a performance drop for low SNR values. Then, we study the metrics varying the RT_{60}, considering SNR $> 40\,\text{dB}$. The results demonstrated higher localization performance, especially in the azimuth plane. The model also outperformed other methods in terms of MAEE. Finally, we computed the considered localization metrics against the distance of the sound source from the center of the microphone array. The results demonstrate similar behaviour to previous tests, where PT-CRNN method exhibits a constant behaviour over different distances. Overall, PT-CRNN demonstrates higher performance than the baseline approaches for both GE and MAEE metrics. Moreover, in non-ideal acoustic scenarios such as low SNR and high reverberation, the proposed model outperformed the other methods, and indicating improved robustness than the conventional approaches. All the graphs and a deeper analysis of these results are reported in the thesis.

4. Conclusions

In this thesis, we have developed a deep-learning based method for the sound source DOA estimation. We adopted a recently introduced feature representation known as RHC which provides meaningful spatial information about the sound field. In this context, we introduced the triplet loss training, which promotes a structured and meaningful features space with improved clustering of similar samples and better discrimination of dissimilar samples. We proposed a CRNN-based architecture for the joint estimation of azimuth and elevation. The proposed models have been trained in a single room and tested in different rooms, demonstrating the ability to generalize on unseen data. As a result, the proposed solution showed improved performance in DOA estimation with respect to baseline methods in complex scenarios with low SNR and high RT_{60}. Then, we proposed a CNN-based siamese neural network. During the training with triplet loss, the network was able to create a feature space encouraging similar samples to be close together. Instead, dissimilar samples, i.e, source with different DOAs were pushed further apart, while preserving a structured correspondence with the spatial domain. This promotes the interpretability of the features embedding. Furthermore, the analysis of the features embedding shows that the feature space of the pre-trained network has higher spatial correspondence with respect to the free design embedding while maintaining accurate DOA estimation performance. Finally, we presented the localization results by performing tests on rooms with different dimensions and reverberation times. The proposed method demonstrated to be able to generalize when dealing with unseen data, providing improved performance from baseline methods. Furthermore, the network exhibits higher robustness, with minimal reduction of the performance even in complex environments.

References

[1] 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.

[2] 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.

[3] 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.