Direction of Arrival Estimation with Microphone Arrays Using SRP-PHAT and Neural Networks

David Díaz-Guerra Aparicio, José Ramón Beltrán Blázquez

Grupo de Investigación en Interfaces Avanzadas (AffectiveLab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: *{646988, jrbelbla}@unizar.es*

Abstract

The Steered Response Power with phase transform (SRP-PHAT) is one of the most employed techniques for Direction of Arrival (DOA) estimation with microphone arrays, but its computational complexity grows when the search space increases. To solve this issue, we propose the use of Neural Networks (NN) to obtain the DOA from low-resolution SRP-PHAT power maps.

Introduction

Due its robustness against acoustical conditions as reverberation or noise, the Steered Response Power (SRP) is one of the most employed technique for Direction of Arrival (DOA) estimation and Sound Source Localization (SSL) with microphone arrays.

In [1], [2] the SRP-PHAT was introduced, using the phase transform to make the algorithm more robust against reverberation [3] and presenting a new formulation of the SRP in terms of the Generalized Cross-Correlation (GCC) functions. With this new formulation, most of the computational complexity is in the computation of the GCCs, which is common to all the search directions, so the number of operations grows slower with the search space. Despite this, the complexity may still be an issue for very large search spaces, especially when they have two, or even three, dimensions.

In the field of microphone arrays, despite widely used for blind source separation, Neural Networks (NNs) have been barely employed for DOA estimation yet. [4] propose solving the DOA estimation as a classification problem using the GCCs as the inputs of a NN which have an output for each point in the search space. This approach has the same problem as the SRP-PHAT algorithm: increasing the number of points of the search space (to improve the precision or to add new dimensions) increase its complexity, as a higher number of outputs are needed and the classification becomes harder.

To solve this problem, we propose to formulate the DOA estimation as a regression problem. Because [4] claims that they obtained worse results with regression than with classification, we use low resolution SRP-PHAT power maps as input instead of GCCs. Due to the good amount of existing research on the implementation of NN in FPGAs [5], [6], and even the existence of chips designed for real-time NN inference commercially available [7], we believe that, if the power map resolution is low enough, it could reduce the computational complexity of the entire system.

Obtaining a complete dataset with array recordings of a room would be too complicated and time consuming and may be unsuitable for some plug-and-play applications, so the training stage is a critical point of these techniques. In [4], the network is trained with simulated signals obtained for different room acoustical properties and get good results when they test it with real recordings. Alternatively, we propose training the NN with real recordings done in the room where the array will be used. As getting a perfect labelled dataset is unfeasible, we use a high-resolution SRP-PHAT power map to label it.

The SRP-PHAT algorithm

The Steered Response Power (SRP) of a sensor array is defined as the power of the output of an array steered to the desired direction using a delay-and-sum beamformer. It can be written in terms of the Cross-Correlation functions between sensors as:

$$P(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} R_{nm}(\Delta\tau_{nm}(\boldsymbol{\theta}))$$

Where N is the number of sensors, $\boldsymbol{\theta}$ is the desired direction, R_{nm} is the Cross-Correlation Function between the sensors n and m, and $\Delta\tau_{nm}(\boldsymbol{\theta})$ is the time difference of arrival between the same sensors.

The Cross-Correlation Functions can be substituted by the Generalized Cross-Correlation Functions (GCCs) [8] in equivalence to substitute the delay-and-sum beamformer by a filter-and-sum beamformer:

$$R_{nm}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Psi_{nm}(\omega) X_n(\omega) X_m^*(\omega) e^{j\omega t} d\omega$$

Where $X_n(\omega)$ is the Fourier Transform of the signal received at the sensor n, * is the complex conjugate operator, $j = \sqrt{-1}$, and $\Psi_{nm}(\omega)$ is a weighting function. The most common weighting function in microphone arrays is the PHAT transform:

$$\Psi_{nm}(\omega) = \frac{1}{|X_n(\omega) X_m^*(\omega)|}$$

To perform a DOA estimation using the SRP-PHAT algorithm we must compute $P(\boldsymbol{\theta})$ for a fine enough grid of directions $\boldsymbol{\theta}$. Figure 1 shows a high-resolution SRP-PHAT power map whose maximum would be lost by an 8x8 power map.

MultiLayer Perceptron (MLP)

A MultiLayer Perceptron (MLP) is one of the most basic NNs. The output of a dense, i.e. fully connected, layer with D inputs and H perceptrons is:

$$\boldsymbol{o} = f(\mathbf{W}^T\mathbf{a} + \mathbf{b})$$

Where T denotes transposition, $\mathbf{a} = (a_1, \ldots, a_D)^T$ is a vector with the inputs of the layer, $\mathbf{W}$ is a DxH matrix with the weight of each perceptron for each input, $\mathbf{b} = (b_1, \ldots, b_H)^T$ is a vector with the bias of each perceptron, and $f(x)$ is a nonlinear function such as the *relu* function: $f(x) = \max(x, 0)$.

The parameters of the MLP network, i.e. $\mathbf{W}$ and $\mathbf{b}$, are optimized using backward propagation techniques to minimize an error function. Several regularization techniques, such as the dropout [9], has been proposed to improve the optimization, i.e. training stage, avoiding the MLP to overfit its parameters to the training dataset obtaining worse results with the test dataset.

Proposed DOA estimator

Network architecture and training

In order to reduce the computational complexity of the SRP-PHAT algorithm we propose to reduce the number of directions where $P(\boldsymbol{\theta})$ is computed and infer the DOA estimation using a neural network instead of looking for the maxima of the SRP-PHAT function. Specifically, we use a MLP with two hidden layers with 128 perceptrons. As we have focused in 2D DOA estimation with a circular microphone array, our output layer has only 2 perceptrons that represent the DOA estimation in spherical coordinates.

We train the network using the ADELTA gradient optimizer [10] implemented in Keras [11] using as error function the mean square angular distance. To reduce the overfitting, we use the dropout regularization technique in the hidden layers.

Dataset

To create the dataset to train the network, we placed the array, a miniDSP UMA-8 with 6 MEMS microphones equispaced in a circumference of approximately 90 mm in diameter, in the center of a conference table. With this configuration, we perform a recording of 5 minutes at 44.1 kHz emitting white noise through the speaker of a smartphone while walking around the table at different heights.

We divide the recording in frames of 1024 samples with an overlap of 512 samples and apply the SRP-PHAT algorithm to get 2 power maps, the first with high resolution (90x360) and the second with a lower one. We use the former to obtain the position of the sound source and the last as input of the network. Finally, we randomly permuted the frames and took 18,000 for training and 6,000 for test.

Results

Table I shows the angular Root Mean Square Error (RMSE) in the test dataset for different power map resolutions. It can be seen that the DOA inferred by the MLP from a power map of resolution 8x8 has an error very similar to taking the maximum of a 32x32 power map, but using 16 times fewer evaluations of the SRP-PHAT functional. Figure 2 shows the 8x8 power map corresponding to Figure 1. Despite the real DOA is not captured, the MLP is able to estimate the DOA estimation with an error of only 1.7º.

Conclusions

This work can be seen as a full low-complexity NN-based DOA estimation framework. In a first step, the user records some minutes of white noise.

Secondly, a non real-time NN training process is performed. Finally, the DOA is obtained by real-time inference from low-resolution SRP-PHAT power maps.

REFERENCIAS

[1]. DIBIASE, J.H. A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays [online]. Brown University, 2000. Available from: http://www.glat.info/ma/av16.3/2000-DiBiaseThesis.pdf

[2]. DIBIASE, J.H., SILVERMAN, H.F. and BRANDSTEIN, M. Robust Localization in Reverberant Rooms. In : Microphone Arrays: Signal Processing Techniques and Applications. Berlin, Heidelberg : Springer Berlin Heidelberg, 2001.

[3]. ZHANG, C., FLORENCIO, D. and ZHANG, Z. Why does PHAT work well in lownoise, reverberative environments? In : 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. March 2008. p. 2565–2568

[4]. XIAO, X., ZHAO, S., ZHONG, X., JONES, D.L., CHNG, E.S. and LI, H. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. In : 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). April 2015. p. 2814–2818.

[5]. VOGEL, S., GUNTORO, A. and ASCHEID, G. Efficient hardware acceleration for approximate inference of bitwise deep neural networks. In : 2017 Conference on Design and Architectures for Signal and Image Processing (DASIP). September 2017. p. 1–6.

[6]. GUAN, Y., LIANG, H., XU, N., WANG, W., SHI, S., CHEN, X., SUN, G., ZHANG, W. and CONG, J. FP-DNN: An Automated Framework for Mapping Deep Neural Networks onto FPGAs with RTL-HLS Hybrid Templates. In : 2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). April 2017. p. 152–159.

[7]. BARRY, B., BRICK, C., CONNOR, F., DONOHOE, D., MOLONEY, D., RICHMOND, R., O'RIORDAN, M. and TOMA, V. Always-on Vision Processing Unit for Mobile Applications. IEEE Micro. March 2015. Vol. 35, no. 2, p. 56–66. DOI 10.1109/MM.2015.10.

[8]. KNAPP, C. and CARTER, G. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing. August 1976. Vol. 24, no. 4, p. 320–327. DOI 10.1109/TASSP.1976.1162830.

[9]. HINTON, G.E., SRIVASTAVA, N., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R.R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 [cs] [online]. 3 July 2012.

[10]. ZEILER, M.D. ADADELTA: An Adaptive Learning Rate Method. arXiv:1212.5701 [cs] [online]. 22 December 2012.

[11]. CHOLLET, F. and OTHERS. Keras. [online]. 2015. Available from: https://github.com/keras-team/kerasbibtex[publisher=GitHub]

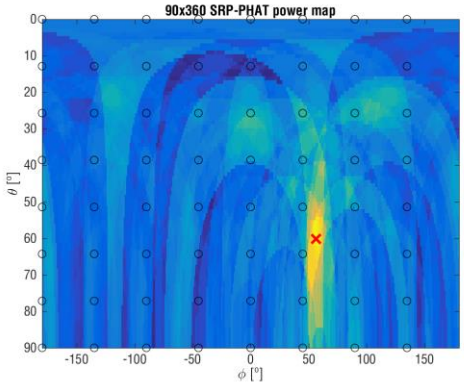

Fig. 1. High-resolution SRP-PHAT power map. The red cross indicate the maximum and the black circles represent an 8x8 equispaced grid.

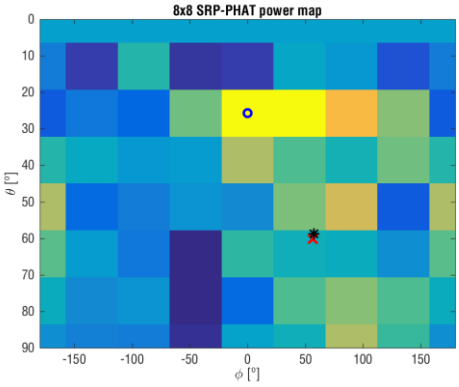

Fig. 2. Low-resolution SRP-PHAT power map. The red cross indicate the maximum of the high-resolution map (Figure 1), the black * is the DOA inferred by the MLP and the blue square the maximum of the low-resolution power map.

Table I. Root Mean Square Error (RMSE)

Θ resolution	Φ resolution	RMSE maximum (deg)	RMSE MLP (deg)
32	32	4.6021	4.0488
16	16	12.6915	4.0303
8	8	29.4648	4.6960
4	4	41.2628	5.7129