SPHERICAL CONVOLUTIONAL RECURRENT NEURAL NETWORK FOR REAL-TIME SOUND SOURCE TRACKING

Tianle Zhong[13⋆], *Israel Mendoza Velázquez*[12⋆], *Yi Ren*[1], *Héctor Manuel Pérez Meana*[2], *Yoichi Haneda*[1]

[1] Graduate School of Informatics. The University of Electro-Communications. Tokyo, Japan.
[2] ESIME Culhuacán. National Polytechnic Institute of Mexico. Mexico City, Mexico.
[3]University of Electronic Science and Technology of China. Chengdu, China

ABSTRACT

Neural networks have been widely applied in direction-of-arrival (DOA) estimation and source tracking systems. In this paper, we introduce a spherical convolutional recurrent neural network that utilizes Deepsphere, a graph-based spherical convolutional neural network, employing the steered response power with phase transform (SRP-PHAT) power maps as input features for real-time robust sound source DOA estimation and tracking applications. The proposed method achieves a performance similar to that of state-of-the-art 3D convolutional neural networks (3D-CNNs) method and reduces the processing time by 88.6%, the parameter count by 85.5%, and the training memory usage by 54.0% respectively. The shallow structure of proposed network demonstrates effectiveness and efficiency.

Index Terms— Microphone arrays, direction-of-arrival, SRP-PHAT, spherical CNNs, recurrent neural networks.

1. INTRODUCTION

Direction-of-arrival (DOA) estimation is one of the main applications of sensor arrays. It can be considered as a stage in a variety of applications such as hearing aids [1], robotics [2], and teleconference systems [3].

The avant-garde proposals have increased the relevance of deep neural networks (DNNs) for DOA estimation. Some of these define the inputs for feeding the network by using parameters from conventional DOA estimation methods [4, 5, 6]. The pattern-recognition capabilities and noise robustness of these networks enable increasing the performance or overcome their intrinsic defects [2, 7, 8, 9].

Recently, in [10], SRP-PHAT maps were proposed with a 3D convolutional neural network (CNN) based only on a DNN architecture (Cross3D) with causal convolution for temporal context and source tracking. Because of the inherent trade-off between SRP-PHAT map resolution and estimation accuracy, high-resolution maps provide more accurate results but involve a high computational cost. Therefore, Cross3D provides robust results even at low-resolution maps and enables real-time processing to a greater extent. However, for Cross3D, to process the SRP-PHAT maps as the input features, the azimuth and elevation axes must be adopted in a Cartesian projection in order to operate with the conventional convolution operation. This may not be appropriate because of the discontinuities present at the edges of the rectangle. In addition, the areas of each pixel do not symbolize equivalent angular values (especially at the poles).

Deepsphere is a graph based spherical CNN tailored for the hierarchical equal area iso-Latitude pixelization (HEALPix) [11] algorithm for spherical manifold sampling, through which the pooling operation on the hierarchical structure can be performed. Because of its graph-based approach, this CNN is a computationally efficient alternative that presents relevant properties, such as rotation equivariance on $SO(3)$, which means that the rotation applied on the input leads to the same rotation at the output.

Given that SRP-PHAT maps for 3D microphone arrays should be processed as spherical data, we propose a spherical convolutional recurrent neural network (CRNN) that employs Deepsphere spherical CNN followed by a long short-term memory (LSTM) layer [12] for stable performance temporal modeling in long-time-sequence circumstances as a source tracking extension of our previous work [13]. Moreover, the HEALPix algorithm is employed for pixelization on SRP-PHAT power maps to adopt them as input features. To make a fair comparison with Cross3D, we performed training and testing of both networks under the same simulated acoustic conditions of noise and reverberation, and assuming an NAO robot head configuration. We conclude the paper with a brief analysis of accuracy and computational efficiency.

2. METHODOLOGY

2.1. Steered-Response Power Phase Transform

Assuming a q element microphone array and a broadband sound source $s(t)$, the captured signals $x_q(t)$ from the q^{th} sensors can be modeled as follows:

$$x_q(t) = s(t) * h_q(\Omega_r, t) + \nu_q(t) \quad (1)$$

⋆Equal Contribution

where $h_q(\Omega_r, t)$ is the impulse response given by the q^{th} sensor, the source direction $\Omega_r = (\theta_r, \phi_r)$ is expressed in terms of the elevation θ_r and azimuth ϕ_r angles, and $\nu_q(t)$ denotes an uncorrelated white Gaussian noise with the source signal and noises of the remaining sensors.

The steered response power algorithm [5] searches for the maximum value given by the output power of a filter-and-sum-beamformer with steered direction Ω, which can be expressed as

$$P(\Omega) = \sum_{p=1}^{Q} \sum_{q=1}^{Q} \int_{-\infty}^{\infty} \Psi_{pq}(\omega) X_p(\omega) X_q^*(\omega) e^{j\omega\tau_{pq}(\Omega)} d\omega, \tag{2}$$

where Q is the number of elements in the array, which p^{th} and q^{th} signals has an equivalence in the frequency domain expressed as $X_p(\omega)$, $X_q(\omega)$ respectively, $\tau_{pq}(\Omega)$ is the time difference between the p^{th} and q^{th} sensors determined by the DoA Ω_r of the sound source, and $\Psi_{pq}(\omega)$ is a weighting function in which the PHAse transform (PHAT) is a popular robust alternative against reverberation defined as $\Psi_{pq}(\omega) = 1 \,/\, |X_p(\omega) X_q(\omega)|$. The search over all directions derives a spatial statistical map $P(\Omega)$ in which the highest power value is related to the approximated DoA $\widehat{\Omega_r}$.

2.2. Laplacian Graph-based Spherical Convolution

Given that SRP-PHAT power maps for 3D arrays can be tailored as spherical data, their operation with a spherical convolutional neural network might be more appropriate. Although there are multiple definitions of the convolution operation applied to the sphere, according to the convolution theorem, convolution is accessible through the spectral domain of the spherical harmonics transform, an adaptation of this procedure can be found in [14]. However, the computational cost of this approach is high, with $\mathcal{O}(n^{3/2})$ on equiangular grids and $\mathcal{O}(n^2)$ in general, which is significant in each forward and backward propagations [15].

Alternatively, an approach to access spherical convolution can be made using the graph Fourier domain. A weighting scheme for the graph is defined by the weights of the adjacency between vertices i:

$$W_{ij} = e^{-\frac{1}{4t}\|x_i - x_j\|^2} \tag{3}$$

where t is the optimal kernel width determined by the sampling resolution of the spherical data and the number of nearest neighbors considered in (4). This weighting scheme was proved by Belkin & Niyogi [16] using a random uniform sampling.

By defining the graph Laplacian L, through its eigenvectors forming an orthogonal basis, the projection of the graph signal f on them can be defined as the graph Fourier transform. Thus we can establish a relation with the convolution theorem. Although the computational cost of the eigendecomposition and matrix multiplication required in the graph Fourier transform is high, the convolution kernel h can be expressed as a polynomial $h_\theta(\lambda) = \sum_{i=0}^{P} \alpha_i \lambda^i$ of degree P (kernel size) and trainable coefficients α_i. This allows us simplifying the convolution operation on graph $h(L)$ as follows:

$$h(L)f = (\sum_{i=0}^{P} \alpha_i L^i) f, \tag{4}$$

For a stable performance, Chebshev polynomials can be employed instead of monomials [17], whose recursive relation is exploitable for computational cost reduction; and this method was adopted in [15].

2.3. Proposed Method

Deepsphere exploits the above principles of convolution in the graph [15], adapting it to represent the discrete points according to the HEALPix algorithm [11], which is a uniform and hierarchical sphere sampling method. In a previous study, we adopted the SRP-PHAT maps with Deepsphere for DOA estimation. [13]

HEALPix is a general class of schemes that hierarchically subdivides a base polyhedron, in which the sphere is constructed as a rhombic dodecahedron that contains 12 congruent rhombic faces. The parameter N_{side} is always a power of two and indicates the pixel number on each rhombic face as N_{side}^2. Thus, such that the resolution (total pixel count) on the sphere is determined by $N_{pix} = 12 \times N_{side}^2$. Each pixel covers the same size of surface area and is beneficial for integrating the white noise generated by the microphone into the pixel space. In addition, given that HEALPix is a hierarchical sampling, down-sampling on the graph enables a pooling operation, which summarizes pixels by a permutation invariant function including the maximum and averaging.

Following the Deepsphere spherical CNN, we introduce an Long Short-Term Memory (LSTM) [12] for source tracking by temporal context analysis. Compared to the conventional recurrent neural networks, LSTM units can overcome the vanishing gradient problem, thereby enabling learning over longer time sequences. This may have benefits when high levels of reverberation are present, in addition to the fact that LSTM does not imply a fundamental problem in learning on noisy sequences.

Concerning output format, the predicted DOA positions on the Cartesian coordinate system are generated in a unitary vector format from the network output as a regression task. The Cartesian coordinate system has a continuous coordinate field, while the spherical periodic coordinate system has an abrupt value break between 0 and 2π that can reduce the efficiency of model training, as reported in [18].

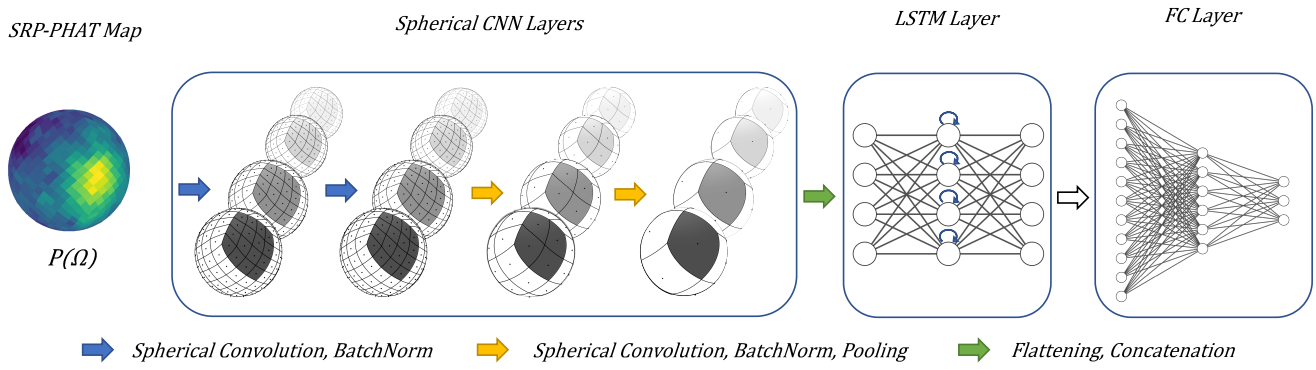
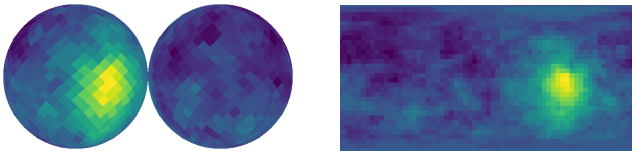


Fig. 1. Spherical Recurrent Neural Network architecture employed in performance estimation



(a) pixelized by HEALPix (b) cartesian projection

Fig. 2. Examples of SRP-PHAT maps sampling

3. EXPERIMENTS

3.1. Model Architecture

As shown in Fig.1, four spherical CNN layers are stacked as the encoding part of the network. The first layer operates on the SRP-PHAT map, which is sampled using the HEALPix algorithm with a resolution of $N_{SIDE} = 8$ yielding 7.32° of angular distance between pixels. The kernel size of the spherical convolution is always three, and the number of output channels for each convolutional layer is 8, 16, 32, and 32. For the last two convolutional layers, a pooling layer follows the corresponding batch normalization layer to down-sample the resolution by half from their input respectively.

For each time frame, the output channels of the spherical convolutional layers are flattened and then concatenated to formulate the time sequence input for an LSTM layer with a hidden size of 120.

Finally, three fully connected layers are added in the last part of the model to summarize the output of the LSTM layer. The last fully connected layer contains three perceptrons to encode the DOA as a vector in Cartesian coordinates. Given that this is formulated as a regression task, the mean squared error loss function is utilized during model training to minimize the Euclidean distance between the output of the network and the ground-truth DOA position coordinates.

Table 1. Room impulse response data generation parameters

Parameter	Value interval	Unit
Room size	3×3×2.5 - 10×8×6	meter
T60	0.2 - 1.3	second
Absorption weights	0.5 - 1.0	not applicable
Array position	(0.1, 0.1, 0.1) - (0.9, 0.9, 0.5)	meter
SNR	5 - 30	dB

3.2. Data Generation

To ensure the consistency of the training data set with Cross3D [10] for subsequent comparison, gpuRIR [19] was employed to generate the room impulse response (RIR) and source moving trajectories data for training and testing data on the fly under random parameters, including room size, microphone array position, reverberation and omnidirectional noise levels, as shown in Table 1. We used the LibriSpeech corpus [20] which contains severe clean speech hours for sound source synthesis. Although a spherical microphone array shape could be preferred owing to its uniform sampling over all directions, we used the microphone array same with that of NAO robot head in order to have a same set up with the experiments of Cross3D in addition to the fact that this geometry may have real world task applications.

Given that the SRP-PHAT maps arise mainly from the conforming directivity pattern, the network structure of our proposal and Cross3D does not depend on the array, but only on the map resolution.

4. RESULTS AND DISCUSSION

4.1. Performance

To estimate the model accuracy under different conditions, we utilized the same set of speech data and trajectories to generate test datasets with SNR values of 5, 15 and 30 dB and reverberation times of 0.0, 0.3, 0.6, 0.9, 1.2, and 1.5 seconds.

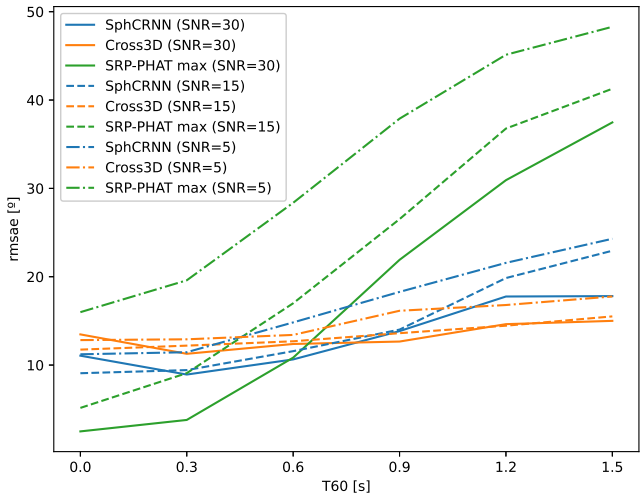


Fig. 3. Root mean square angle error (rmsae) of models under different SNR and reverberation time

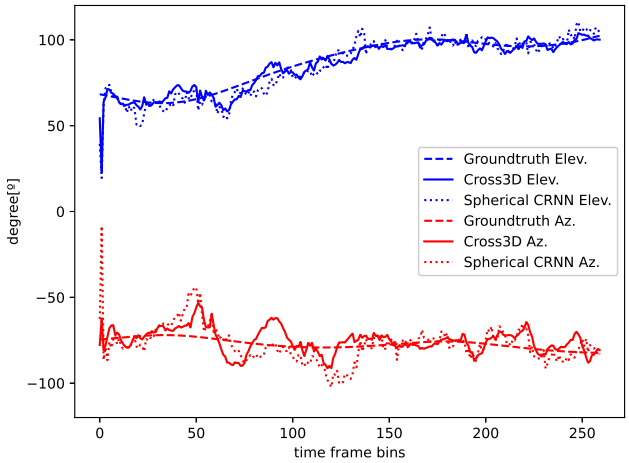


Fig. 4. Azimuth and Elevation of a trajectory and its estimation by both methods under SNR = 15 dB and T60 = 0.7 s

The method selected for comparison is Cross3D, a state-of-the-art 3D neural network [10]. The input of the Cross3D model is the Cartesian projection of the SRP-PHAT power maps with a resolution of 32×64 and a maximum distance between pixels of 5.625°, as shown in Fig.2 (b). Both our model and the Cross3D model were trained for 10 hours on the same machine in this estimation test. The baseline method employed for this performance estimation was the SPR-PHAT max point, which simply selects the position with the highest value among all the pixels on the SRP-PHAT maps. We considered that the baseline model can perform better for very low values of the SNR and short reverberation time.

The estimation results are shown in Fig.3. The spherical CRNN has a relatively robust performance under different conditions. Even though our model does not always outperform the 3D CNN, especially in long reverberation time situations, slight improvement can be observed with the spherical CRNN in situations with SNR equal to 15 dB and reverberation time approximately ranging from 0.3 to 0.9 s. Such situations are more similar to real application environments.

In particular, sound source tracking estimation results for the same trajectory under SNR = 15 dB and T60 = 0.7 second are presented in the Fig. 4, which shows the similar performance of these two methods in a typical acoustic scene.

Given that our model parameters have not been specialized yet for the targeted task and our model architecture employed for performance estimation is a rather shallow model, the effectiveness of our proposed model can still be demonstrated. A deeper Spherical CRNN is promising in this regard.

4.2. Efficiency

For real-time applications, a low processing time delay is needed. This time delay can be regarded as the inference time in neural networks, which provides insight about the computational efficiency of a model. To evaluate and compare the inference time, 8900 (this number arises from the memory limitation of the graphic card) sound source moving trajectories were generated to be processed by spherical CRNN and Cross3D. For a fair comparison, the inference tests of both models were run on the same single graphic card (Nvidia GTX 1080 Ti) based on the same generated testing database. Batch sizes were adjusted to utilize the GPU utilization rate as much as possible during inference.

Table 2. Models employed for the evaluation

Model	Inference		Training
	time	parameter number	GPU memory usage
Cross3D	266.29s/8900	5,626,148	4300MiB
Spherical CRNN	30.32s/8900	816,003	1980MiB

As it is shown in Table 2, the spherical CRNN requires 3.41 ms to process a trajectory, whereas Cross3D needs 29.92 ms in average. Note that the spherical CRNN reduces the inference time by 88.6% and the parameter count by 85.5%. It is also worth mentioning that a 54.0% reduction in GPU memory usage during training was observed for the spherical CRNN.

5. CONCLUSION

In this paper, we present a spherical CRNN based on pixelized SRP-PHAT power maps and Laplacian graph-based spherical convolution. By evaluating its performance on DOA estimation and sound source tracking in comparison with 3D CNNs, its high efficiency was demonstrated. Given that the model architecture employed for performance estimation is fairly shallow and simple, the potential of deeper and more complicated models for this method is promising and deserves further investigation.

6. REFERENCES

[1] Sebastian Braun, Wei Zhou, and Emanuel A.P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing AIDS using relative transfer functions," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015*, 2015.

[2] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, "Deep Neural Networks for Multiple Speaker Detection and Localization," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 74–79, 2017.

[3] Hong Wang and Peter Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1997.

[4] Charles H. Knapp and G. Clifford Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976.

[5] Joseph Hector DiBiase, *A High Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments using Microphone Arrays*, Ph.D. thesis, Brown University, 2000.

[6] Ralph Schmidt and X W Af, "Multiple Emitter Location and Signal Parameter," *IEEE Transactions on Antennas and Propagation*, 1986.

[7] Soumitro Chakrabarty and Emanuel A.P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2017-Octob, pp. 136–140, 2017.

[8] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference*, 2018.

[9] Laureline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guerin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," 2019.

[10] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran, "Robust Sound Source Tracking Using SRP-PHAT and 3D Convolutional Neural Networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2021.

[11] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, "HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere," *The Astrophysical Journal*, 2005.

[12] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.

[13] Israel Mendoza Velázquez, Yi Ren, Yoichi Haneda, and Hector Perez-Meana, "DOA estimation for spherical microphone array using spherical convolutional neural networks," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE) (IEEE GCCE 2021)*, Kyoto, Japan, Oct. 2021.

[14] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling, "Spherical CNNs," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[15] N. Perraudin, M. Defferrard, T. Kacprzak, and R. Sgier, "DeepSphere: Efficient spherical convolutional neural network with HEALPix sampling for cosmological applications," *Astronomy and Computing*, 2019.

[16] Mikhail Belkin and Partha Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *Journal of Computer and System Sciences*, 2008.

[17] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, 2016.

[18] Zhenyu Tang, John D. Kanu, Kevin Hogan, and Dinesh Manocha, "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019.

[19] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, 2020.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015.