A fusion method based on class rotations for DNN-DoA Estimation on Spherical Microphone Array

Israel Mendoza Velázquez*†, Yi Ren*, Yoichi Haneda*, Héctor Manuel Pérez Meana†
*Graduate School of Informatics. The University of Electro-Communications. Tokyo, Japan.
{israel.mv, ren.yi, haneda.yoichi}@uec.ac.jp
†ESIME Culhuacán. National Polytechnic Institute of Mexico. Mexico City, Mexico.
hmperezm@ipn.mx

***Abstract*—Deep Neural Networks (DNN) has been used to estimate the Direction-of-Arrival (DoA) with spherical array under highly reverberant and noisy environments. In this paper, we propose a fusion technique for DNN as a solution to DNN-DOA estimations to obtain a joint decision and outperform the prediction's accuracy of a single network, which may also suggest a data augmentation technique. This proposed fusion consists of averaging the results of multiple networks with their spatially rotated categories to obtain a final approximation. The experiments carried in this work were performed using a 3D Convolutional Recurrent Neural Network (3DCRNN) structure for classification, which classes are determined by the t-design method as a pseudo-uniform spherical sampling. The performed simulations suggests an improvement over a single DNN performance, showing a reduction of the average angular error using 6 Networks can be achieved with 38.96% for highly reverberant and noisy environment.**

***Index Terms*—spherical microphone array, direction of arrival, spherical harmonics, convolutional recurrent neural network, late fusion,**

I. Introduction

The task of Direction-of-Arrival estimation aims to estimate the incident angle of the observed signals relative to the sensor array. Its approach to acoustics with microphone arrays finds diverse applications covering automatic speech recognition, robotics, hearing aids, etc.

Some conventional approaches include the Time Difference of Arrival based methods, such as those derived from Generalized Cross Correlation (GCC) functions [1], [2], [3]. Beamforming based methods as the Steered Response Power with PHAse Transform (SRP-PHAT) [4], [5], [6] and Subspace based methods, such as MUltiple SIgnal Classification (MUSIC) [7], [8]. Additionally, these methods also have special adaptations to spherical arrays, as they have robust results over 3D directions [9], [10], [11].

However, the main reasons that DOA estimation's accuracy degenerates are often due to a high presence of noise, high reverberation in the environment, or even present high computational complexity, and several efforts have emerged as ways to overcome it.

Recently, DNN based methods have gained popularity since the network is not only expected to be capable of learning the right patterns on mapping the observed information to the estimation angle, but also to overcome the shortcomings of the mentioned methods thanks to its high abstraction capacity.

Several proposals with different perspectives have emerged recently, including the study from general input features such as raw signals [12], [13] to robust and closer to angular information based on Spherical Harmonics such as Modal coherence coefficients [14] or Signal Invariant Spherical Harmonic Features [15], regression and classification approach [16], [17], multiple sources estimation [18] or joint task performance, such as Sound Event Localization and Detection [19], source tracking [20] or distance estimation [21].

Whereas DoA estimation for 2D arrays only requires azimuthal estimation, 3D arrays also contemplate elevation. In the case of classification with a joint elevation and azimuth estimation, a uniform sampling scheme is defined on the sphere to define each class, such as seen in [22], [23]. However, in order to reduce the inter-class angular distance to achieve a refined grid, a considerable increase of classes is necessary and may result in a slower training convergence.

In this paper, we propose a multi-network fusion approach based on the score of DNN models as classifiers, each of them representing a rotated version of classes from a primary classifier, this is based on the fact that in certain circumstances, the combination of individual classifier's scores to obtain a joint decision may consistently outperform the results of an individual classifier [24], [25]. To perform classes combination representing a DoA, we consider that individual estimation can be formulated as a vector pointing to the corresponding class, and a final joint estimation is achieved with a weighted average, using each maximum posterior probability as vector's weights.

For our experiments, we use a conventional 3DCRNN with few classes determining a discrete angle given by the t-design sphere sampling method [26]. We use multiple fine-tuned networks with classes rotated angularly of a primary trained model to jointly generate a final estimation to achieve fast convergence. We show that proposed fusion method can

not only progressively compensate for misclassifications of individual networks but also can estimate a numerical value closer to the ground-truth DoA.

II. Feature Extraction

In this section, we briefly describe the choosen input features for our experimental DNN, based on the decomposition of the sound field on a solid spherical microphone array using spherical harmonics.

A. Spherical harmonics expansion

The continuous sound pressure $P(\Omega, k)$ over a sphere S^2 can be reinterpreted by using spherical harmonics (SH) as [27],

$$p_{nm}(k) = \int_{S^2} P(\Omega, k) Y^*_{nm}(\Omega)\, d\Omega, \tag{1}$$

with $\Omega = (\theta, \phi)$ as the elevation and azimuth obr the sphere respectively, $k = 2\pi f/c$ as the wavenumber and c as the speed of sound, $Y_{nm}(\Omega)$ as the spherical harmonic of order n and degree m and $p_{nm}(k)$ as coefficient of expansion. Moreover, using a sampling over sphere made through a spherical microphone array of radius r and Q elements, an approximation is held as

$$p_{nm}(kr) \approx \sum_{q=1}^{Q} \omega_q P(\Omega_q, k) Y^*_{nm}(\Omega_q), \tag{2}$$

where ω_q refers to a constant factor given by sampling method on the sphere, this also determines the maximum order of spherical harmonics N of the observed sound field. As stated in [14], $p_{nm}(kr)$ can be separated as $p_{nm}(kr) = a_{nm}(k) b_n(kr)$ given $a_{nm}(k)$ a sensor independent and a dependent part $b_n(kr)$. For solid spherical arrays and assuming a plane wave model, due to the scattering of sound field, $b_n(kr)$ known also as modal coefficients are defined as [28]

$$b_n(kr) = \frac{j}{kr^2 h'_n(kr)}, \tag{3}$$

with $h'_n(x)$ as the spherical hankel function of second kind. Simplification in (3) is derived by making use of the wronskian relation of spherical hankel function. Thus, with the appropriate transformation, $a_{nm}(k) = p_{nm}(k)/b_n(kr)$ are considered as independent of the array's radius and the position of the microphones.

B. Signal Invariant Spherical Harmonic Features

As proposed in [15], a signal independence could be formulated from a_{00} as it does not have any influence from source's DoA. In this sense, new features are derived in magnitude and phase as follows

Spherical Harmonic Magnitude (SH-M) Features:

$$|q^t_{nm}(k)| = \frac{|a_{nm}(k)|}{|a_{00}(k)|}. \tag{4}$$

Spherical Harmonic Phase (SH-P) Features:

$$\angle q^t_{nm}(k) = \angle a_{nm}(k) - \angle a_{00}(k). \tag{5}$$

Finally, because (5) takes the same values at 0 and 2π, $\cos(\angle q^t_{nm}(k)))$ and $\sin(\angle q^t_{nm}(k))$ can be derived.

It is worth to say that an interesting demonstration of Elevation and Azimuth independence in (4) and (5) is also shown and additionally for real applications. Moreover due to zeroth SH order's spatial averaging, a spatially-white noise reduction is also reported [29].

III. Proposed Method

We consider a primary network whose i-th classes represent a single source direction given by the points Ω_i on the unit sphere determined by a uniform sphere sampling method. Additionally, consider J parallel networks whose i-th classes are derived by a R_j three dimensional rotation operation on the sampling points $\Omega_{i,j}$ of the primary network.

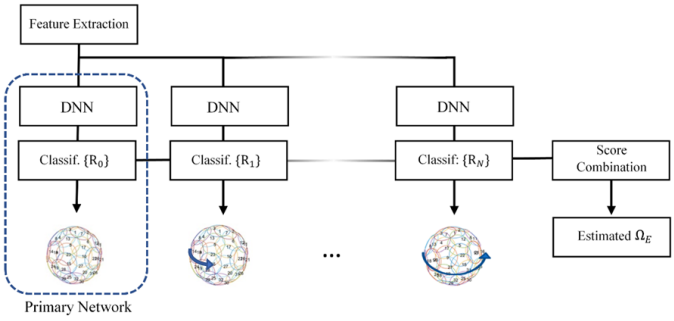

Fig. 1. Joint DoA estimation by the proposed method

The last layer of each network considers the softmax function $C_{i,j}(x)$ to compresses arbitrary real values of an i-th vector into real values within the range [0, 1], this is correlated as the decimal probabilities of the observed feature x to belong to i-th class, by this, an individual decision of the estimated DoA $\Omega_{E,j}$ is related with the class of highest probability.

A score combination could be achieved to increase accuracy and classification certainty, a simple way to achieve it is using weights $\rho_{i,j}$ [30], such that

$$C_i(x) = \frac{1}{J} \sum_{j=1}^{J} \rho_{i,j} C_{i,j}(x). \tag{6}$$

A joint decision can be made again by taking the resulting class with the highest probability. From our approach, due to the angular inequivalence of classes $\Omega_{i,j}$ we can simply propose a fusion scheme formulated as the angle given by the sum unit vectors $\overrightarrow{v}$ pointing in the direction of the individual predicted DoA class with a magnitude defined by its maximum probability, such that

$$\Omega_E \approx \angle\Big(\sum_{j=1}^{J} (\max_i C_{i,j}(x))\, \overrightarrow{v}(\Omega_{E,j})\Big). \tag{7}$$

In summary, by contemplating a reduced number of classes for each network, it is possible to cover information that the individual angular distance cannot contemplate through each

rotation. Based on the above, the main task lies in considering a number of nets and rotations to ensure an estimation closer to the ground-truth DoA.

IV. Experimental Setup

In this section, we describe the details and conditions to corroborate our proposed method.

A. Data Generation

We assume a single source estimation with a solid spherical array following the Eigenmike specifications, consisting in $Q = 32$ microphones and radius $r = 0.043\,\mathrm{m}$, simulating its acoustic scenario in a rectangular room by using Simulated Room Impulse Responses (SRIRs).

Although we are strongly interested by the acoustical simulation environment seen in [20], with a constant generation of SRIRs using GPU resources and used for unlimited training data generation as training occurs [31] to achieve a better generalization, we opted to simulate a finite set of SRIRs considering the sound pressure's scattering effect [32]. Thus, as similar as [22], we generated the acoustic conditions by RIRs considering random room configurations between $2.5 \times 2.5 \times 2.5\,\mathrm{m}$ and $10 \times 10 \times 3\,\mathrm{m}$ with uniformly distributed dimensions. We conformed the training, validation and evaluation datasets of the network convoling calculated RIRS on Librispeech corpus, as a speech database extracted from 960 hours of audiobook and consecutively, uniformly random distribuited Omnidirectional Gaussian Noise from 5 to 30 dB SNR was added.

B. Input Features

First, we obtain Q microphone input time signals after convolution for each utterance of LibriSpeech with simulated RIRS and then a transformation to time-frequency was held, yielding Q complex spectrograms of T time and F frequency bins. Lastly, we transform time-frequency bins to the Spherical Harmonics Domain following (II-A), resulting on $(N+1)^2$ number of time-frequency channels. From there, next stage to Signal Invariant spherical Harmonic Features is held by an appropriate calculation of $a_{nm}(k)$ using (3) and finally, SH-M and SH-P can be derived from (4) and (5). For SH-M features´we take $((N+1)^2 - 1)$ time-frequency channels since $a_{00}(k)$ is discarded. In case of SH-P, $\cos\left(\angle q^t_{nm}(k)\right)$ and $\sin\left(\angle q^t_{nm}(k)\right)$ take $2((N+1)^2 - 1)$ time-frequency channels into account. Stacking all of them over SH axis a give us $3((N+1)^2 - 1)$ channels of a 3D Tensor with T, F time-frequency bins.

For this experiment, we performed STFT with 512 points over all utterances of 16 kHz using a hamming windowing and 50% of overlapping and calculated data sets for DNN input with 0.3 seconds of estimation over a range from 10 Hz to 7000 Hz, resulting in 220 frequency bins and 19 time bins. Finally, the eigenmike structure allows us to sample sound fields with spherical harmonics up to $N = 4$ maximum order, we consider a maximum of $N = 2$. This yields a 3D tensor of dimension $19 \times 220 \times 24$. Normalization of generated data was achieved according to the original proposal [15].

C. Network Architecture

We based our DNN architecture using a CRNN as seen in [22], with a difference over number of classes and for the first 3DCNN layer, this is because a preference on [15] to process derived features jointly, with also a previously reported advantage over inter-feature convolution [23]. Moreover in this case, we believe that this features already have the possibility to form patterns solely dependent on the DoA over SH Orders dimension [14] so using this structure also allows us to take the temporary advantage of the Recurrent Layer.

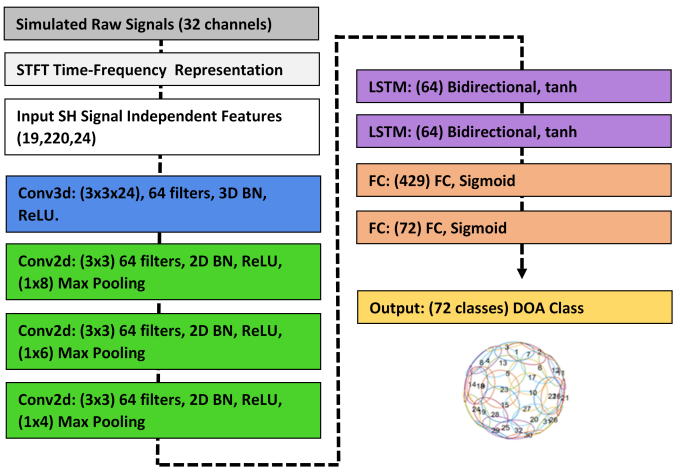

Fig. 2. CRNN Structure with 3DCNN for intra-channel learning. Classes are determined by t-design method with 72 points over sphere. Given the $19 \times 220 \times 24$ input tensor this let us achieve 19 DoA estimations for each 0.3 s.

With respect to the number of classes representing the DoA, the use of 429 classes in [22] yields an angular distance between nearest classes of approximately 10°. For our experiments, the set 72 classes for a primary CRNN model is given by t-design sphere sampling method which yields and angular distance between nearest classes of approximately 25°.

This first model was trained with a step size of 0.001 using a loss entropy criterion and Adam optimizer. Next we made CRNN Fine-tuned replicas of the primary model 5 rotations only on azimuth such as $\{\pi/10, 2\pi/10, 3\pi/10, 4\pi/10, \pi/2\}$ by making use of Euler rotations. The final estimation is made with the same features for all models.

V. Experimental Results

Obtaining the trained networks, we performed the tests on the evaluation set as part of the unseen samples in the training set. In addition, sources are also assumed to be directionally static.

First, we study the temporal behavior for different quantity of fused networks using the angular error between given estimations and the real DoA, this criteria is given by.

$$\theta_d = \arccos\left(\sin\left(\theta_1\right)\sin\left(\theta_2\right) + \cos\left(\theta_1\right)\cos\left(\theta_2\right)\cos\left(\theta_1 - \theta_2\right)\right) \tag{8}$$

In addition, we compare the results obtained with a baseline method based on the formulation of a pseudo intensity vector (PIV) using the first order spherical harmonics [11].

This is depicted from Fig. 3 with a random simulation. In some occasions, a single network is not able to angularly approach to the ground-truth DoA due to its own sampling limitation, even if the individual classification is correct, this also can be visualized from Fig. 4. At the same time we see an overcoming of this phenomenon and a tendency to approach consistently with proposed fused networks.

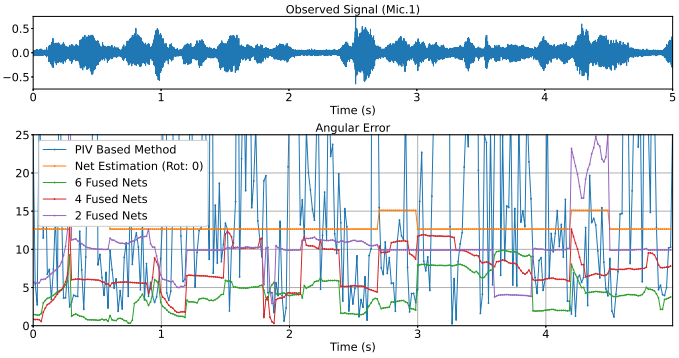

Fig. 3. Top figure illustrate the simulated signal observed on the first channel of the Eigenmike array. Bottom figure shows the temporal behavior of the angular error of different fused networks. A single 72-class primary network with no rotation (Orange line), 2-Net Fusion (Cyan line), 4-Net Fusion (Red line), 6-Net Fusion (Green line) and the Baseline PIV Based method (Blue line). This simulation consists of a source coming from $\theta = 71, \phi = 269$, with 15 dB SNR and TR60 of 0.5 s.

However, this is not always achieved and a destructive fusion can also be observed, this issue can be solved with a more sophisticated weighting criterion.

To suggest a large-scale idea and study how different quantities of fused networks behave in unfavorable environments, we calculated the percentage of correctly predicted values under the tolerance error using about 200,000 DoA estimates in two acoustic conditions. A first experiment consisted in evaluation over simulated scenarios with 15 dB SNR omnidirectional WGN and TR60 going from 0.2s to 0.6s, the results of these conditions can be observed in Table I. This not only suggests a slight advantage of the primary model over Baseline method, but also we corroborate a consecutively performance increase since second Network is fused, we define this as soon as an approximation to the groundtruth DoA is suggested by angular error of less than 5° is approached and the number of outlying estimations tends to reduce.

The results of a second experiment with stronger conditions was held with 5 dB SNR and 0.6 s to 1.0 s of TR60 can be observed in Table II. From this, the fusion scheme continues to improve the results as more networks are added, however the measurement does not seem to be the same as in the first experiment, and the improvement seems to plateau at a certain point.

VI. Conclusion

In this paper, we proposed a multi-network fusion approach with rotated classes as a way to increase the accuracy of a primary network. We evaluated fusion performance given a few quantity of angular distant classes and under different noise and reverberation levels in order to study how far and how this proposal of conformation can operate constructively. For future work we would like to consider different azimuth and elevation rotations, modify the dimensions of the primary network and comparing the training times of networks with a large number of classes, as well as determine a more suitable fusion scheme that makes a better decision on the scores of each classifier.

TABLE I
Angular error for different values of tolerance

15 dB SNR, 0.2 - 0.6 TR60						
	<5°	<10°	<15°	<20°	Error Avg. (°)	Improve (%)
PIV Based	10.50	29.51	45.67	57.53	25.56	-
1 Nets	11.34	34.00	63.99	74.64	19.26	-
2 Nets	20.21	54.93	76.23	86.36	14.85	22.89
3 Nets	28.35	64.59	83.26	89.79	12.38	35.72
4 Nets	32.22	70.69	86.37	91.61	11.12	42.26
5 Nets	36.36	74.62	88.28	92.60	10.33	46.36
6 Nets	40.26	77.84	89.85	93.26	9.69	49.68

TABLE II
Angular error for different values of tolerance

5 dB SNR, 0.6 - 1.0 TR60						
	<5°	<10°	<15°	<20°	Error Avg. (°)	Improve (%)
PIV Based	3.17	11.19	21.25	31.47	39.9	-
1 Nets	6.53	20.83	40.39	53.31	31.82	-
2 Nets	10.50	31.43	51.01	64.61	26.60	16.40
3 Nets	13.47	37.29	57.19	69.39	23.67	25.61
4 Nets	15.47	41.99	61.85	72.57	21.72	31.74
5 Nets	17.60	45.79	64.91	74.90	20.26	36.32
6 Nets	19.57	48.78	66.82	76.12	19.42	38.93

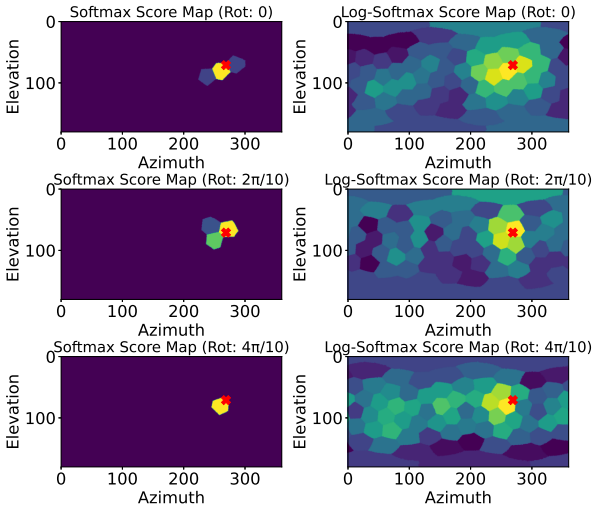

Fig. 4. Classification scores in an scenario of SNR=15dB and 0.5 s TR60 given by softmax and a log-softmax representation to observe hidden details, color scale represents the scores and the center of each area symbolizes the class. Red cross represent the ground-truth DoA. Scores of each classifier have a spatial sense following the expected way of class rotation.

REFERENCES

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

[2] J. VanDecar, "Determination of teleseismic relative phase arrival times using multi-channel cross-correlation and least squares," *Bulletin - Seismological Society of America*, 1990.

[3] A. Lopes, I. S. Bonatti, P. L. D. Peres, R. F. Colares, and C. A. Alves, "A DOA Estimator Based on Linear Prediction and Total Least Squares," *Journal of Communication and Information Systems*, 2002.

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

[5] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," in *Adaptive Antennas for Wireless Communications*, 2009.

[6] D. Salvati, C. Drioli, and G. L. Foresti, "Frequency map selection using a RBFN-based classifier in the MVDR beamformer for speaker localization in reverberant rooms," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

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

[8] R. Roy and T. Kailath, "ESPRIT-Estimation of signal parameters via rotational invariance techniques," in *Adaptive Antennas for Wireless Communications*, 2009.

[9] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2011.

[10] B. Jo and J.-W. Choi, "Sine-based EB-SEPRIT for source localization," *Sensor Array and Multichannel Signal Processing*, vol. 10, 2018.

[11] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3D Source localization in the spherical harmonic domain using a pseudointensity vector," in *European Signal Processing Conference*, 2010.

[12] T. Hirvonen, "Classification of spatial audio location and content using Convolutional neural networks," in *138th Audio Engineering Society Convention 2015*, 2015.

[13] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors (Switzerland)*, 2018.

[14] A. Fahim, P. N. Samarasinghe, and T. 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*, 2020.

[15] V. Varanasi, H. Gupta, and R. M. Hegde, "A Deep Learning Framework for Robust DOA Estimation Using Spherical Harmonic Decomposition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2020.

[16] Z. Tang, J. D. Kanu, K. Hogan, and D. 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.

[17] L. Perotin, A. Defossez, E. Vincent, R. Serizel, and A. Guerin, "Regression versus classification for neural network based audio source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019.

[18] S. Chakrabarty and E. A. P. Habets, "Multi-Speaker Localization Using Convolutional Neural Network Trained with Noise," no. Nips, 2017. [Online]. Available: http://arxiv.org/abs/1712.04276

[19] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," *IEEE Journal on Selected Topics in Signal Processing*, 2019.

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

[21] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound Source Localization in a Multipath Environment Using Convolutional Neural Networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018.

[22] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," 2019.

[23] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference*, 2018.

[24] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, 2005.

[25] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[26] R. H. Hardin and N. J. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions," *Discrete and Computational Geometry*, 1996.

[27] E. G. Williams and J. A. Mann, "Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography," *The Journal of the Acoustical Society of America*, vol. 108, p. 1373, 2000.

[28] K. Bando and Y. Haneda, "Interactive Directivity Control Using Dodecahedron Loudspeaker Array," *Journal of Signal Processing*, vol. 20, no. 4, pp. 209–212, 2016.

[29] S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented Intensity Vectors for Direction of Arrival Estimation in the Spherical Harmonic Domain," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2017.

[30] O. Dehzangi, M. Taherisadr, and R. ChangalVala, "IMU-based gait recognition using convolutional neural networks and multi-sensor fusion," *Sensors (Switzerland)*, 2017.

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

[32] D. P. Jarrett, E. A. Habets, M. R. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2011.