Acoustic Source DOA Tracking Using Deep Learning and MUSIC

Eduardo Alves da Silva and Luiz Wagner Pereira Biscainho

***Abstract*— Sound source tracking is a classical problem in the array processing field that finds applications from human-robot interaction to acoustic navigation. The emergence of deep learning has produced a new class of state-of-the-art solutions, including the Cross3D, based on a convolutional neural network architecture. In this article, the authors present a new system for sound source tracking via microphone array that modifies the pre-processing stage of Cross3D by using MUSIC maps as inputs, and applying appropriate minor modifications to its architecture. The proposed system, called Spectral Cross 3D, improves overall accuracy by 18%, when compared with the original solution.**

***Keywords*— sound source tracking, microphone array, deep learning, convolutional neural networks, MUSIC.**

I. Introduction

The applications of sound source localization (SSL) include security surveillance [1], rescue robotics [23] and indoor navigation [20], among others. With moving acoustic sources, the problem evolves to sound source tracking (SST), and both applicability and complexity increase. In either case, the core of the hardware used are microphone arrays [4].

In SSL, there is a distinction between finding the 3D coordinates of a source or finding the direction of arrival (DOA) of its sound in terms of azimuth and elevation angles. While the former approach is more general, the latter is sufficient for most applications. Also, whenever the source is far away from the microphone array, it is preferable to resort to DOA estimation.

In general, classical solutions for SSL fall into 3 categories:

1) Time delay estimation: Time-difference of arrival (TDOA) is estimated from cross correlation functions [14] and mapped to the source location [12]. This approach has low complexity, but is prone to errors when TDOAs are not precisely estimated.
2) Beamforming: A spatiotemporal filter that estimates the energy from a specific direction is steered to different positions in a grid. Among these methods, the steered response power (SRP) is the most robust, but requires strategies to simplify the grid search and thus reduce its high complexity [15], [5].
3) Subspace methods: Exploring the eigenstructure of the narrowband spatial correlation matrix, they generate a frequency-dependent heatmap of source directions in a grid. An important example of this family is the multiple signal classifier (MUSIC) method [4].

Although there exist methods for specific microphone arrangements (e.g. the spherical array [12]), the aforementioned methods can be applied to a broader range of arrays.

For the tracking problem, the most common approach is based on Bayesian filters [21]: the preferred choices are the Kalman filter and its variants [11] among parametric, and particle filters [6] among non-parametric solutions.

Although each approach has its pros and cons, they tend to behave poorly in scenarios where the signal-to-noise ratio (SNR) is low and/or strong reverberation is present. In order to increase performance under these more stringent conditions, the use of deep neural networks (DNN) has been proposed in the literature, for both SSL and SST tasks. Specifically, different classes of convolutional neural networks (CNN) and recurrent neural networks (RNN) have found applicability.

In the pre-processing stage, a wide range of features are used with the different models. The short-time Fourier transform (STFT) is the input attribute in many systems, using either phase information only [16], [17] or the full spectrogram [2], [3]. The eigenvalues of the narrowband spatial correlation matrix are a low-level feature in [19], and the SRP with phase transform (SRP-PHAT) maps are used in [8], [24].

In this work, we combine the MUSIC method (as a pre-processing stage) with an architecture based on that of the Cross3D [8] (with minor modifications to accommodate the change in dimensionality) to tackle the DOA tracking problem. As the MUSIC power maps are frequency dependent, we generate maps across all the discrete spectrum to serve as input features to the model. A set of experiments demonstrate that, at the cost of an increase in complexity, our proposed solution (called Spectral Cross3D) performs consistently better than the original Cross3D, especially in low-SNR environments.

Following this introduction, the paper is organized as follows. In Section II, we describe the pre-processing methods used in our approach as well as in Cross3D, and detail their architecture. Section III is dedicated to explaining the simulations and training setups. We analyze the results in Section IV, and draw the final remarks on Section V.

II. Methods

Consider an anechoic environment in which a single sound source at position $\mathbf{r}_s$ emits signal $s(t)$. The acoustic scene is captured by an M-microphone array whose $m-$th microphone

Eduardo Alves da Silva, SMT, UFRJ, Rio de Janeiro-RJ, e-mail: eduardo.alves@smt.ufrj.br; Luiz Wagner Pereira Biscainho, SMT, UFRJ, Rio de Janeiro-RJ, e-mail: wagner@smt.ufrj.br. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ)

receives signal

$$y_m = \alpha_m s(t - t_1 - \tau^*_{m1}) + v_m(t), \tag{1}$$

where α_m is the attenuation of the sound wave from the source to the mth microphone (which can be taken as 1 for most applications); v_m is the noise picked up by the mth microphone (usually modeled as white and uncorrelated with the source signal); t_1 is the time of flight (TOF) of the sound wave from the source to the first microphone, and $\tau^*_{m1} = t_m - t_1$ is the time-difference of arrival (TDOA) of the sound wave at the m and first microphones.

A. Steered Response Power – Phase Transform

The steered response power method is based on a delay-and-sum (DS) beamformer [4] whose output is given by

$$z_{DS}(\mathbf{r}, t) = \frac{1}{M}\sum_{m=1}^{M} y_m(t + \tau_{m1}(\mathbf{r})), \tag{2}$$

where $\tau_{m1}(\mathbf{r})$ is the expected TDOA expected from a potential sound source positioned at $\mathbf{r}$ (such that $\tau_{m1}(\mathbf{r}_s) = \tau^*_{m1}$). This expression evidences the main idea of the DS beamformer: aligning the signals from the different microphones in time.

From the Fourier Transform of (2), it is possible to evaluate the average power of $z_{DS}(\mathbf{r}, t)$ for any position as

$$P_{\text{SRP}}(\mathbf{r}) = \int_{-\infty}^{\infty} |Z_{DS}(\mathbf{r}, \omega)|^2 \mathrm{d}\omega. \tag{3}$$

In DOA estimation, $\mathbf{r} = (\phi, \theta)$; for a grid of R_ϕ azimuth and R_θ values, the SRP map will have $R_\phi R_\theta$ elements.

In the classical formulation, one performs a grid search in the map defined by (3) and estimate the source location as

$$\hat{\mathbf{r}}_{s,\text{SRP}} = \arg_{\mathbf{r}} \max P_{\text{SRP}}(\mathbf{r}) \tag{4}$$

If in the more general form of (3) given by [9]

$$P_{\text{SRP}}(\mathbf{r}) = \sum_{m=1}^{M}\sum_{l=1}^{M}\int_{-\infty}^{\infty} \Psi_{lm}(\omega) Y_l(\omega) Y^*_m(\omega) e^{j\omega\tau_{ml}(\mathbf{r})} \mathrm{d}\omega \tag{5}$$

a spectral weighting function

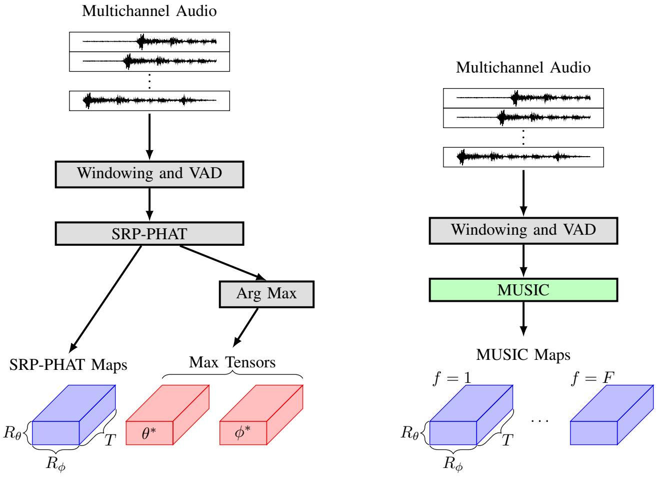
$$\Psi_{lm}(\omega) = \frac{1}{|Y_l(\omega)Y^*_m(\omega)|} \tag{6}$$

is applied, this so-called phase transform (PHAT) eliminates the disturbing effects of magnitude (unnecessary to identify time relations) from the energy map. The resulting SRP-PHAT method generates much sharper maps than the original SRP.

B. Cross3D

The use of the SRP-PHAT map as input feature for a CNN was proposed in the Cross3D [8]. It was shown that the resolution attained by the trained system surpassed that of the original DOA maps, indicating that map patterns could be learned in order to improve the estimation of the DOA.

As the position of the maximum of the SRP-PHAT map is crucial to the determination of the DOA but the $\arg\max$ function is non-linear and difficult to learn conventionally, in [8] additional channels inform to the network the relative position of the maxima as constant tensors, as seen in Figure 1a. Voice activity detection (VAD) is also used in the pre-processing stage, since inference in quiet frames is error-prone.



(a) Cross3D

(b) Spectral Cross3D

Fig. 1: Preprocessing

In order to capture the temporal evolution for DOA tracking, 3D causal temporal convolutions is used. The implemented architecture can be seen in Figure 2a. CNNs commonly use pooling layers to reduce dimensionality through the neural path. In order to avoid a consequent reduction in spatial resolution, the network is split in two paths, each performing pooling in one DOA dimension. The output of the network is a 3×1 vector pointing to the direction of the sound source present in the acoustic scene.

C. MUSIC

Defining $\vec{\mathbf{y}}(\omega) = \begin{bmatrix} Y_1(\omega) & Y_2(\omega) & \ldots & Y_M(\omega) \end{bmatrix}^T$, the MUSIC localization method explores the eigenstructure of the narrowband spatial correlation matrix (SPCM) given by[1] [4]

$$\mathbf{R}_y = \mathbb{E}[\vec{\mathbf{y}}\,\vec{\mathbf{y}}^H] = \sigma_s^2 \boldsymbol{\varsigma}(\mathbf{r}_s)\boldsymbol{\varsigma}(\mathbf{r}_s)^H + \sigma_v^2 \mathbf{I} = \mathbf{R}_s + \mathbf{R}_v, \tag{7}$$

where

$$\begin{aligned} \sigma_s^2 &= \mathbb{E}[|S(\omega)|^2], \\ \boldsymbol{\varsigma}(\mathbf{r}_s) &= \begin{bmatrix} e^{-j\omega\tau_{11}(\mathbf{r}_s)} & \ldots & e^{-j\omega\tau_{M1}(\mathbf{r}_s)} \end{bmatrix}^T, \\ \sigma_v^2 &= \mathbb{E}[|V_1(\omega)|^2] = \ldots = \mathbb{E}[|V_M(\omega)|^2]. \end{aligned} \tag{8}$$

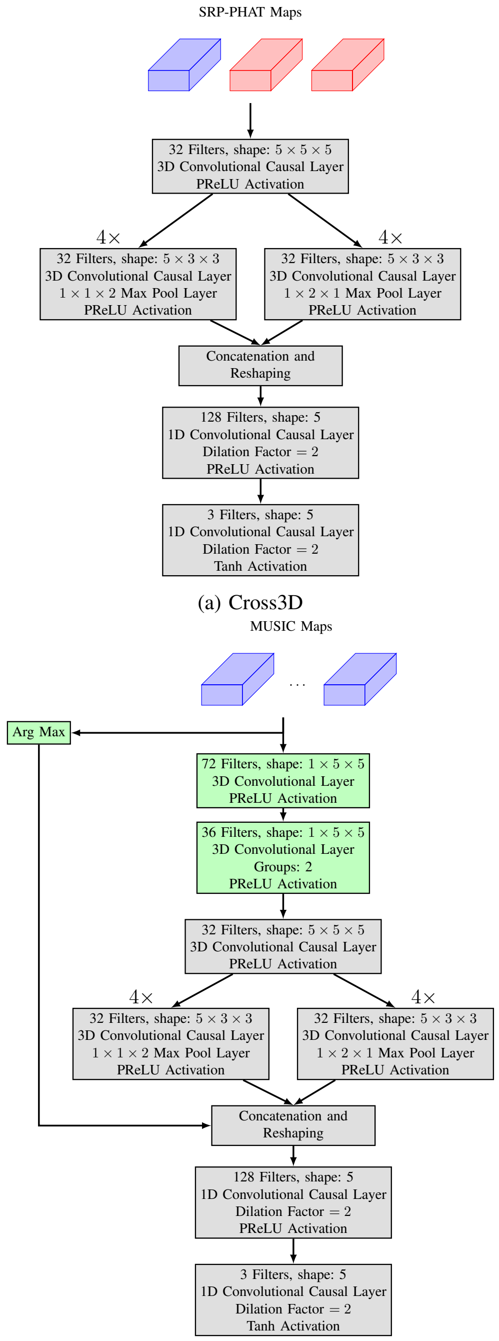
The so-called steering vector $\boldsymbol{\varsigma}(\mathbf{r})$ is essential to the method. It is clear from equation (7) that the $\mathbf{R}_s$ is rank deficient, while $\mathbf{R}_v$ has M eigenvalues equal to σ_v^2. This leads to a specific eigenvalue decomposition of $\mathbf{R}_y$:

$$\mathbf{B} = \begin{bmatrix} \mathbf{b_1} & \mathbf{b_2} & \ldots & \mathbf{b_M} \end{bmatrix}^T \tag{9}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}[\lambda_s + \sigma_v^2, \sigma_v^2, \ldots, \sigma_v^2], \tag{10}$$

where $\mathbf{b}_m$, $m = 1, \ldots, M$, are the eigenvectors ordered according to the eigenvalues in matrix $\boldsymbol{\Lambda}$. Since b_m, $m \geq 2$, are associated with matrix $\mathbf{R}_v$, they are often called the noise

[1] The dependency on ω is omitted in order to keep the notation uncluttered.



(a) Cross3D

(b) Spectral Cross3D

Fig. 2: Architectures

subspace eigenvectors. Applying the eigenvalue equation for them to (7) and using the result in (9), we arrive at

$$\mathbf{b}_m^H \boldsymbol{\varsigma}(\mathbf{r}) \begin{cases} = 0 & \mathbf{r} = \mathbf{r}_s \\ \neq 0 & \text{elsewhere.} \end{cases} \tag{11}$$

Equation (11), which is the core of the MUSIC method, states that the steering vectors for the actual source position are orthogonal to the noise subspace eigenvectors. This allows to define a power-like map

$$P(\mathbf{r}) = \frac{1}{\sum\limits_{m=2}^{M} |\mathbf{b}_m^H \boldsymbol{\varsigma}(\mathbf{r})|^2}, \tag{12}$$

from which source position can be estimated by

$$\hat{\mathbf{r}}_{\text{MUSIC}} = \arg_{\mathbf{r}} \max P_{\text{MUSIC}}(\mathbf{r}). \tag{13}$$

D. SpectralCross3D

Dependence on frequency is inherent to the MUSIC method: it succeeds if $\lambda_s \gg \sigma_v^2$ for a specified value of ω. The broadband nature of signals such as speech and music makes it particularly difficult to determine a suitable frequency for tracking. Furthermore, the typical non-stationarity of such signals would require frame-by-frame frequency selection. To circumvent this issue, we calculate power-maps for all F frequencies of the discrete spectrum, an feed them to the CNN as different channels, as seen in Figure 1b (where the green block highlights the modified step). Although using some criteria to select a subset of frequencies may allow for the reduction of the overall complexity, the authors chose to keep the full spectrum in this first approach.

Given the greater dimensionality of our input tensors, it is reasonable to increase the number of parameters in the subsequent CNN. For this reason, as illustrated in Figure 2b (where green blocks highlight the modified steps), the authors added two convolutional layers to the original Cross3D architecture, with filters selected so as to not increase the temporal receptive field. Also, the maximum positions on the different frequency maps are fed directly into later layers of the network, instead of at the input as originally in Cross3D.

III. Training

We followed the same data generation setup presented in [8], which can be described as follows. The sound source signals used for training were provided by the train-clean-100 partition of the LibriSpeech [18] corpus, a collection of audio-book chapter readings sampled at $f_s = 16\,\text{kHz}$; at each epoch, one random $20\,\text{s}$ excerpt is taken from each chapter reading. Random trajectories were generated by interpolating with random 3-directional oscillations the path between two points $\mathbf{r}_0$ and $\mathbf{r}_L$ randomly chosen within the boundaries of the simulated room; in 25% of cases, a stationary source is simulated instead. To simulate the room impulse responses, gpuRIR [7] was used; the audio signals received at the array microphones are simulated by the convolution of the source signal with the respective impulse responses, degraded by additive omnidirectional random noise. The SNR was set to $40\,\text{dB}$ during the first 39 epochs, and sampled uniformly between $30\,\text{dB}$ and $5\,\text{dB}$ afterwards. The WebRTC VAD [22] was used to detect silent frames; whenever more than 66% of a given frame was considered silent, its associated maps were multiplied by zero.

Although the methods can be adapted to any geometry, the array used in the simulations is a pseudo-spherical 12-microphone array, as described in [10]. The resolution of the input maps is determined by $R_\theta = 32$ and $R_\phi = 64$[2]. Since 256-point discrete Fourier transforms were used, $F = 128$ channels are delivered to the network.

For the network training, the mean squared Cartesian distance loss function and the ADAM optimizer [13] were chosen, similarly to [8]. However, we used them together with a step learning rate (LR) scheduler and early stopping. The root mean squared angular error (RMSAE) in the test set was used as stopping criterion. The values of the hyperparameters selected are shown in Table I.

Parameter	Value
Initial LR	0.0003
Mini batch size	25
Step Size	10 epochs
LR decay factor	0.8
Minimal relative decrease	0.001
Early Stopping Patience	15 epochs

TABLE I: Parameters used in training procedures.

IV. Results

The training of Spectral Cross3D took 176 epochs, which translated to approximately 4 days in a system using an Intel Core i7 processor and a NVIDIA TITAN Xp GPU[3].

Having retrained the original Cross3D model using the same parameters specified in Section III, which took around the same number of epochs and approximately 2 days, we observed a slight reduction in the error metrics, specially under harder environmental conditions (low SNR and higher T60). So, to allow for a fair comparison, we will use the results attained by the retrained version of Cross3D. Whenever necessary, SpC3D denotes the proposed Spectral Cross3D, C3D the original Cross3D, and C3D* the retrained Cross3D.

A. First Experiment: Test Set

Model performances were compared using the test-clean LibriSpeech dataset [18] for several values of SNR and T60, in some cases extending the ranges followed by the training samples. We present the corresponding results in Figure 3, where the orange lines refers to Spectral Cross3D model and the blue ones to the retrained Cross3D. Figure 4 illustrates one example trajectory used in the test set.

It can be seen that for every value of SNR evaluated there is a consistent reduction in the error metric, which decreases with the increase of the T60 value. Even for SNR values not observed in the training phase, we notice a decrease in RMSAE. The average relative decrease of RMSAE in the selected scenarios is presented in Table II, including both retrained and original Cross3D models.

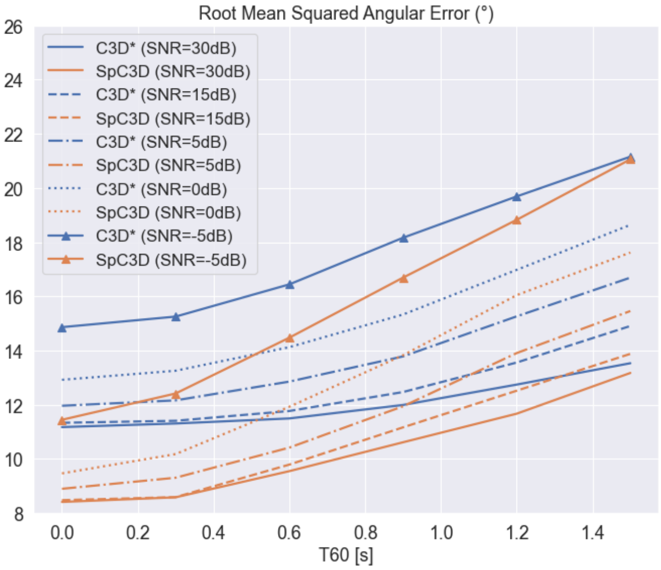


Fig. 3: Comparison of root mean squared angular errors in the test dataset, for multiple values of T60 and SNR

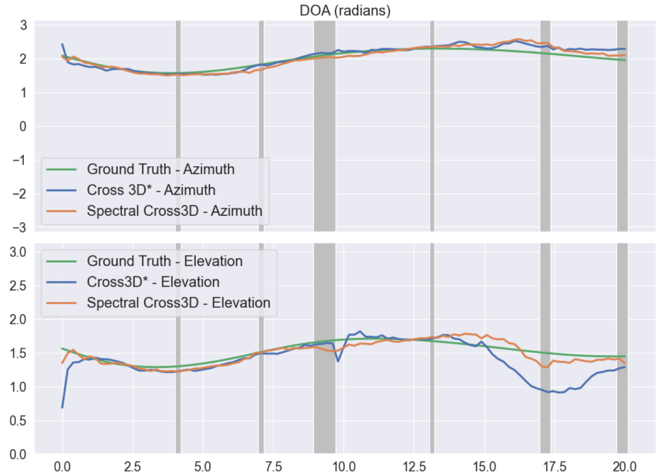


Fig. 4: DOA tracking performed by both models on a sample from the test set, with SNR of 5 dB and T60 of 0.9 s

SNR	Reduction in RMSAE (%) (SpC3D versus C3D*)	Reduction in RMSAE (%) (SpC3D versus Cross3D)
30 dB	14.73	19.35
15 dB	15.30	19.77
5 dB	16.29	20.68
0 dB	14.39	18.52
−5 dB	11.09	15.12
Average	14.36	18.69

TABLE II: Average relative decrease in RMSAE, when comparing both versions of Cross3D with Spectral Cross3D.

B. Second Experiment: LOCATA dataset

The acoustic source localization and tracking (LOCATA) challenge [10] took place in 2018, and provided a small set of audio samples for competitors to evaluate their SSL and SST solutions. Even if the test dataset consists of a small number of recorded signals, comparing the performances of Spectral Cross3D and the baseline (BL) model of the LOCATA Challenge can give an idea of how well our proposed solution operates under completely unseen conditions. We computed the Mean Absolute Azimuthal Error [10] (MAAE) attained on

[2] In [8], the authors demonstrated that finer resolutions did not improve the tracking capabilities significantly.

[3] https://www.nvidia.com/en-us/titan/titan-xp/

the single source tasks (i.e. Tasks 1, 3 and 5), using the same microphone array selected for training. The results, presented in Table III, indicate that our system performs similarly to the baseline model, exhibiting errors in the same order of magnitude.

Task	MAAE (SpC3D)	MAAE (BL)
Task 1	6.6°	4.2°
Task 3	4.5°	9.4°
Task 5	4.2°	5.4°

TABLE III: Mean Absolute Azimuthal Error in the LOCATA Challenge.

V. Conclusions

We presented a new sound source tracking system called Spectral Cross3D, which combines a state-of-the-art neural network architecture [8] with a novel preprocessing stage that delivers MUSIC power-like maps as network input. Our solution has been shown to attain a lower RMSAE metric than the original system in the same test dataset, and results comparable to the baseline provided in the LOCATA challenge dataset.

It should be noticed that in its present form the Spectral Cross3D requires a larger number of parameters (6,591,188 against 5,626,148 of the Cross3D), which increases both training time and memory consumption. The fact that multiple SPCMs have to be estimated and eigendecomposed in order to generate the input maps also increase the system overall complexity. One possible solution to tackle this issue could be the inclusion of a frequency selection mechanism to restrict the input maps to an appropriate subset of the full spectrum. This has the potential to reduce the number of parameters without significantly impacting performance.

Another aspect to be addressed in the continuation of this work is the acoustic simulator used for data generation. It is known that reverberation is a frequency-dependent phenomenon. The traditional image source method used in gpuRIR is fast enough to be used for on-line data generation, at the expense of a simple reverberation model that does not consider the frequency dependence of reflection coefficients. For a better assessment of model capabilities, a more realistic simulator should be used in future works. Moreover, the set of MUSIC narrow-band power-like maps provide the necessary flexibility to describe frequency dependent phenomena. Once this information is taken into account on simulations, a better generalization is expected to be attained in Spectral Cross3D.

References

[1] A.R. Abu-El-Quran, R.A. Goubran, and A.D.C. Chan. Security monitoring using microphone arrays and audio classification. *IEEE Transactions on Instrumentation and Measurement*, 55(4):1025–1032, 2006.

[2] S. Adavanne, A. Politis, J. Nikunen, and T. 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, December 2019.

[3] S. Adavanne, A. Politis, and T. Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466, Rome, Italy, September 2018.

[4] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer, Heidelberg, Germany, 2008.

[5] M.F. Berger and H.F. Silverman. Microphone array optimization by stochastic region contraction. *IEEE Transactions on Signal Processing*, 39(11):2377–2386, November 1991.

[6] C.E. Chen, H. Wang, A. Ali, F. Lorenzelli, R.E. Hudson, and K. Yao. Particle filtering approach to localization and tracking of a moving acoustic source in a reverberant room. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 4, Tolouse, France, May 2006.

[7] 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*, 80(4):5653–5671, February 2021.

[8] 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*, 29:300–311, November 2021.

[9] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. Robust Localization in Reverberant Rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 157–180. Springer, Berlin, Heidelberg, 2001.

[10] C. Evers, H. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. Naylor, and W. Kellermann. The LOCATA challenge: Acoustic source localization and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, April 2020.

[11] J. Fé, S. D. Correia, S. Tomic, and M. Beko. Kalman filtering for tracking a moving acoustic source based on energy measurements. In *International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6, October 2021.

[12] Y. A. Huang, J. Benesty, and J. Chen. Time delay estimation and source localization. In J. Benesty, M. M. Sondhi, and Y. A. Huang, editors, *Springer Handbook of Speech Processing*, pages 1043–1063. Springer, Berlin, Germany, 2008.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, pages 1–15, San Diego, USA, May 2015.

[14] 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, August 1976.

[15] M. V. S. Lima, W. A. Martins, L. O. Nunes, L. W. P. Biscainho, T. N. Ferreira, M. V. M. Costa, and B. Lee. A volumetric SRP with refinement step for sound source Localization. *IEEE Signal Processing Letters*, 22(8):1098–1102, December 2015.

[16] W. Mack, U. Bharadwaj, S. Chakrabarty, and E. A. P. Habets. Signal-aware broadband DOA estimation using attention mechanisms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4930–4934, Barcelona, Spain, May 2020.

[17] S. S. Mane, S. G. Mali, and S. P. Mahajan. Localization of steady sound source and direction detection of moving sound source using CNN. In *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6, Kanpur, India, July 2019.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, Brisbane, Australia, April 2015.

[19] R. Takeda and K. Komatani. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 405–409, Shanghai, China, March 2016.

[20] T.-H. Tan, Y.-T. Lin, Y.-L. Chang, and M. Alkhaleefah. Sound source localization using a convolutional neural network and regression model. *Sensors*, 21(23), December 2021.

[21] S. Thrun, W. Burgard, D. Fox, and R.C. Arkin. *Probabilistic Robotics*. MIT Press, Cambridge, United States, 2005.

[22] J. Wiseman. py-webrtcvad. `https://github.com/wiseman/py-webrtcvad`, 2022. [Online; accessed 17-May-2022].

[23] B. Zhang, K. Masahide, and H. Lim. Sound source localization and interaction based human searching robot under disaster environment. In *SICE International Symposium on Control Systems (SICE ISCS)*, pages 16–20, Kumamoto, Japan, March 2019.

[24] T. Zhong, I. M. Velázquez, Y. Ren, H. M. P. Meana, and Y. Haneda. Spherical convolutional recurrent neural network for real-time sound source tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5063–5067, Marina Bay Sands, Singapore, May 2022.