INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

SR-SRP: Super-Resolution based SRP-PHAT for Sound Source Localization and Tracking

Jae-Heung Cho, and Joon-Hyuk Chang

Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
hhh9431@hanyang.ac.kr, jchang@hanyang.ac.kr

Abstract

Sound source localization and tracking have been extensively studied. Recently, there has been considerable interest in highly reverberant scenarios and steered response power with phase transform (SRP-PHAT) based models have shown a good performance. However, these models still have limitations because the SRP-PHAT algorithm cannot represent the direction of the source in such adverse environments. In this paper, we propose a novel structure combining a super-resolution model and a single sound source localization model that allows to improve direction estimation performance. The proposed method generates a robust power map that accurately represents the direction of the source, even in poor scenarios. Furthermore, the proposed structure has a lower computational cost because it uses a low-resolution map. Experimental results on simulation-based and real-world data show that the proposed method outperforms the state-of-the-art model, Cross3D.

Index Terms: SRP-PHAT, sound source localization and tracking, image super-resolution

1. Introduction

Sound source localization (SSL) involves estimating the direction-of-arrival (DOA) of a sound source by analyzing multi-channel signals captured by a microphone array. SSL is widely used in various real-world applications, such as robot-human interaction and multichannel-based speech separation, speaker diarization, and speech recognition [1–3]. Traditional signal processing-based algorithms, such as generalized cross-correlation (GCC) [4], multiple signal classification (MUSIC) [5, 6], and steered response power with phase transform (SRP-PHAT) [7], are widely used for DOA estimation. Recently, SSL based on various deep learning models, such as deep neural network (DNN) [8], convolutional neural networks (CNNs) [9, 10], and convolutional recurrent neural networks (CRNNs) [11, 12], has been proposed. Signal-processing-based features such as GCC [13] and phase spectrogram [12] are used as input features for deep-learning-based DOA estimation with CNNs and CRNNs [9–12]. These methods demonstrate superior performance compared with traditional methods. However, they have limitations in scenarios with severe reverberation and noise, which are common in real-world environments.

Diaz-Guerra *et al.* [14] proposed 3D CNN-based Cross3D which is a deep learning model based on SRP-PHAT for SSL and tracking, and this model is shown robustness against noise and reverberation. The SSL model based on SRP-PHAT performed well because the input spectrum had high-quality spatial details. However, the SRP-PHAT algorithm has a drawback in that it can not precisely indicate the DOA of the source in a noisy and reverberant environment, which limits the performance of the SRP-PHAT-based models. In particular, as the resolution of SRP-PHAT increases, the impact of noise and reverberation also increases. Furthermore, the experimental results of Cross3D showed that the performance is not improved significantly even if the resolution of the power map increases in such poor situations [14].

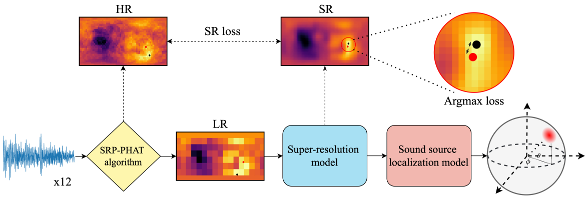


Figure 1: *Overview of the proposed method. LR, HR, and SR are low-resolution, high-resolution, and super-resolution maps, respectively. The red dot indicates the actual DOA of the sound source and the black dot indicates the maximum point of the map.*

In this study, we show that a novel structure SR-SRP that generates robust feature maps can achieve better DOA estimation performance even in poor scenarios. The proposed SR-SRP is an end-to-end structure that combines a super-resolution (SR) model with an SRP-PHAT-based SSL and tracking model. The SR model plays a role for generating a spatial spectrum that can accurately indicate the DOA of the source to overcome the limitation of SRP-PHAT. This model consists of a CNN and transposed CNN and uses a low-resolution (LR) map as an input feature for maintaining the characteristics of SRP-PHAT. The improved spatial feature map, called the SR map, is used as input to the SSL and tracking model Cross3D. We also propose an argmax loss function that improves the spatial characteristics of the SR map. In addition, the proposed method has a low computational cost for the SRP-PHAT algorithm because it uses the LR map as an input. We trained and validated with LibriSpeech [15] and gpuRIR-based [16] simulation data in the same manner as Cross3D and then tested with the localization and tracking (LOCATA) dataset [17]. Consequently, the high peak of SR map indicates the direction of the source well than the SRP-PHAT based high-resolution (HR) map, and SR-SRP outperforms the DOA estimation performance of Cross3D. In particular, our method yields significantly better performance in an environment with severe noise or reverberation in terms of root-mean-squared angular error (RMSAE).

10.21437/Interspeech.2023-2369

2. Related Works

2.1. SRP-PHAT algorithm

Assuming that there are time-domain signal $x(t)$ and M microphones, the signals recorded by the m^{th} microphone, $s_m(t)$ can be expressed as follows:

$$s_m(t) = x(t) * h_m(\mathbf{\Omega}_s, t) + v_m(t), \tag{1}$$

where $\mathbf{\Omega}_s$, $h_m(\mathbf{\Omega}_s, t)$, and $v_m(t)$ denote the DOA of the sound source, the impulse response from $\mathbf{\Omega}_s$ to the m^{th} sensor, and the white Gaussian noise signal, respectively. Additionally, $\mathbf{\Omega}$ is expressed in terms of azimuth $\phi \in [-\pi, \pi]$ and elevation θ $\in [0, \pi]$. The SRP-PHAT algorithm is a beamformer-based approach that searches for the maximum values in a predefined space $\mathbf{\Omega}$.

$$\hat{\mathbf{\Omega}}_s = \underset{\mathbf{\Omega}}{\operatorname{argmax}}\, P(\mathbf{\Omega}), \tag{2}$$

where $\hat{\mathbf{\Omega}}_s$ is the direction of the maximum SRP, and $P(\mathbf{\Omega})$ is the SRP which is expressed in the sum of $R_{m_1,m_2}(\cdot)$, which denotes GCCs of microphones m_1 and m_2.

$$P(\mathbf{\Omega}) = 2\pi \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} R_{m_1,m_2}(\triangle\tau_{m_1 m_2}(\mathbf{\Omega})), \tag{3}$$

$$R_{m_1,m_2}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_{m_1,m_2}(\omega) S_{m_1}(\omega) S^*_{m_2}(\omega) e^{j\omega\tau} d\omega, \tag{4}$$

where $S(\omega)$, *, and τ are the Fourier transform of $s(t)$, complex conjugate, and time difference, respectively. Furthermore, $\Phi_{m_1,m_2}(\omega)$ is the phase transform, defined as $\Phi_{m_1,m_2}(\omega) = 1/|S_{m_1}(\omega) S^*_{m_2}(\omega)|$. Herein, $\hat{\mathbf{\Omega}}_s$ is calculated by all $P(\mathbf{\Omega})$ in a predefined space. Therefore if the resolution of the search space increases, the computational cost also increases.

2.2. Single image super-resolution

Originally, the single image super-resolution (SISR) [18] task played a role in restoring an LR image to an HR image, and was classified into interpolation-based and deep learning-based methods. Interpolation-based SISR [19] can perform upsampling quickly and straightforwardly, but has the disadvantage of poor image recovery performance. A CNN-based SISR [20] has been proposed to overcome the aforementioned disadvantages, and its performance was significantly better than previous methods. In addition, architecture that uses deconvolution [21] and pixel shuffle [22] have been proposed. These architectures reduce the amount of computation and improve performance. The loss function used in SISR depends on the specific purposes of the models and the characteristics of the input features. The loss function is defined in various ways such as pixel [20], content [23], and texture losses [24]. In this study, we intend to upsample and enhance the spatial spectrum and propose suitable objective functions for that goal.

3. Proposed Method

In this section, we introduce the architecture of the proposed SR-SRP. The SR-SRP comprises two models: SR and SSL (Fig. 1). In the following subsection, we describe the SR model in detail, which is the main contribution of this study, and a short description of the Cross3D [14] used as the SSL model.

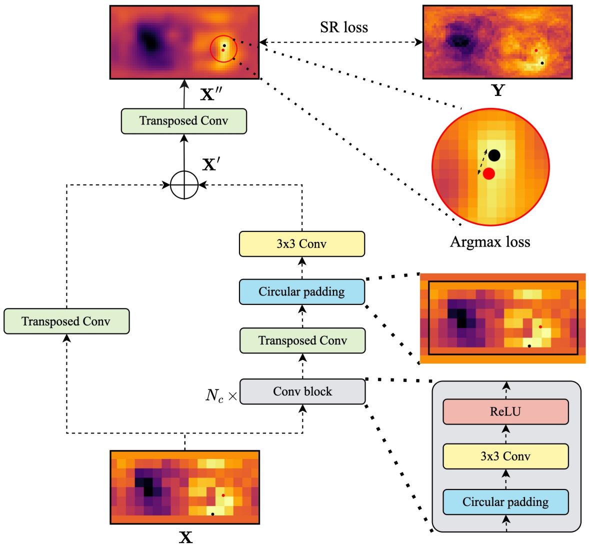


Figure 2: *Model architecture of SR model. Example of upsampling power map from 8 × 16 to 32 × 64. The left branch of the model is the reconstruction layer, and the right branch is the feature extraction layer.*

3.1. Super-resolution model

Let $\mathbf{X} \in \mathbb{R}^{H \times W}$ denotes the SRP-PHAT power map, where H is the resolution of the elevation and W is the resolution of the azimuth. The SR model comprises feature extraction and reconstruction layers, and the maps are passed into a CNN-based feature extraction layer and transposed CNN-based reconstruction layer. The feature extraction layer consists of N_c CNN blocks, a transposed CNN with a scale of 2, and a CNN with a single kernel of size 3 × 3. The CNN blocks contain circular padding [25], a CNN with 64 kernels of size 3 × 3, and a rectified linear unit (ReLU) [26]. The reconstruction layer consists of only a transposed CNN with a scale of 2, and $\mathbf{X}' \in \mathbb{R}^{2H \times 2W}$ is generated by adding it to the output of the feature extraction layer. The SR map $\mathbf{X}'' \in \mathbb{R}^{4H \times 4W}$ is then generated by upsampling with a transposed CNN with a scale of 2. We use circular padding to solve the discontinuity problem for the power map and prevent information loss when using another padding. For circular padding, on the azimuth axis, the left side of the power map is mapped to the right, and vice versa. On the elevation axis, the top left is mapped to the top right, and vice versa.

To earn a power map that has accurate spatial information based on the LR map, we use two types of objective functions in the training process of the SR model: SR loss and argmax loss. The SR loss, which is commonly used in SISR, measures the mean squared error (MSE) loss between the HR and the SR maps such that

$$\mathcal{L}_{SR} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{Y}_n - \boldsymbol{X}''_n)^2, \tag{5}$$

where $\boldsymbol{Y}_n$, $\boldsymbol{X}''_n$, and N are the HR map, the SR map, and the total number of frames, respectively. The SR loss function helps maintain the characteristics of the SRP-PHAT algorithm during the upsampling process. However, the primary objective is to generate a feature map containing clear spatial information. Therefore, we propose an argmax loss function that can further enhance the power map. The argmax loss is the mean absolute error between the maximum position of the SR map and the actual source location $\boldsymbol{y}_n$ in the Cartesian coordinates. The argmax function is usually used to determine the maximum position of the map; however, the argmax function is nonlinear and non-differentiable. Thus, we use the softmax-based soft-argmax function [27].

$$\sigma(\boldsymbol{X}_{i,j}) = \frac{e^{\beta \boldsymbol{X}_{i,j}}}{\sum_{k=1}^{H} \sum_{l=1}^{W} e^{\beta \boldsymbol{X}_{k,l}}}, \qquad (6)$$

where $\sigma(\cdot)$ is the softmax operation, β is the softmax weight and $\boldsymbol{X}_{i,j}$ is the value of power map at location (i, j).

$$\Psi_d(\boldsymbol{X}) = \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{P}_{i,j,d} \sigma(\boldsymbol{X}_{i,j}), \qquad (7)$$

where $\Psi_d(\cdot)$ is the soft-argmax function, d is a given component ϕ or θ and $\boldsymbol{P}_{i,j,d} \in \mathbb{R}^{H \times W}$ is a two-dimensional discrete normalized ramp. In addition, $d = \phi$, then $\boldsymbol{P}_{i,j,\phi} = i/H$ and $d = \theta$, then $\boldsymbol{P}_{i,j,\theta} = j/W$.

$$\mathcal{L}_{argmax} = \frac{1}{N} \sum_{n=1}^{N} |f(\boldsymbol{y}_n) - f(\Psi_\phi(\boldsymbol{X}''_n), \Psi_\theta(\boldsymbol{X}''_n))|, \qquad (8)$$

where $f(\cdot)$ is the transformation function of the spherical to cartesian coordinates. Via argmax loss, the SR model is trained to minimize the DOA error of the SR map and improve the spatial property.

3.2. Sound source localization and tracking model

We use Cross3D as the SSL and tracking model. Cross3D is a 3D CNN-based model, and $\tilde{\mathbf{X}} \in \mathbb{R}^{1 \times T \times 4H \times 4W}$ is input to Cross3D, where T is the number of time frames. Cross3D uses three-channel maps constructed by concatenating feature maps with a tensor of the same size that indicates the maximum position value. This approach is inefficient; however, Cross3D is claimed to be the easiest way to convey the characteristics of SRP-PHAT. However, SR-SRP estimates the DOA using learnable feature maps produced in the SR model; therefore, this process is unnecessary and uses only a single-channel map.

Finally, we use objective functions that perform the weighted sum of the three loss terms in the training process. The $\mathcal{L}_{DOA}$ is the MSE loss between the estimated and actual DOA of the source. The overall loss is configured as follows:

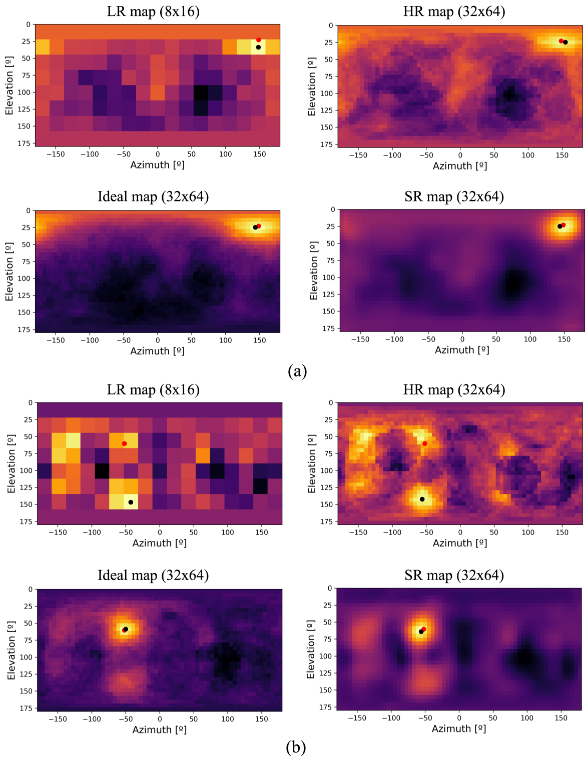
$$\mathcal{L}_{total} = w_1 \mathcal{L}_{argmax} + w_2 \mathcal{L}_{SR} + \mathcal{L}_{DOA}, \qquad (9)$$

where w_1 and w_2 are weighting factors. In this study, we set them to 0.1, which is the optimal value obtained through repeated experiments.

4. Experiments

4.1. Datasets

We used the LibriSpeech [15] and LOCATA [17] datasets, and maintained the same conditions and procedure as the data construction method of Cross3D for a comparison. In the training and validation processes, we used simulated single moving source data based on the LibriSpeech corpus, particularly *train-clean-100* for training and *test-clean* for validation, and generated room impulse responses (RIRs) with gpuRIR. For the details of the RIR data generation parameters, array geometry was set to be the same as the NAO robot head 12-channel array, the room size was 3×3×2.5 m^3 to 10×8×6 m^3, the reverberation time (T_{60}) was 0.2 s to 1.3 s, and the signal-to-noise ratio (SNR) was 5 dB to 30 dB. In addition, we used a LOCATA dataset for the test. Specifically, we used tasks 1, 3, and 5, which were the static microphone and source, static microphone and moving source, and moving microphone and source, respectively. In addition, to prevent tracking errors in situations where speech does not exist for all data, the silent frame was set to zero using the voice activity detector of WebRTC [28].



(a)

(b)

Figure 3: *Result comparison of LR, HR, SR, and ideal maps. (a) the power maps of SNR = 30 dB and* T_{60} *= 0.6 s (b) the power maps of SNR = 5 dB and* T_{60} *= 0.9 s.*

4.2. Evaluation metrics

We used the RMSAE as an evaluation metric for the DOA estimation. The angular error is expressed as the dot-product between the estimated and actual DOA of sound sources, and the RMSAE is given by the formula below:

$$\text{RMSAE} = \sqrt{\frac{\sum_{n=1}^{N} (cos^{-1}(f(\boldsymbol{y}_n) \cdot f(\hat{\boldsymbol{y}}_n)))^2}{N}}, \qquad (10)$$

where $\hat{\boldsymbol{y}}_n$, and $\cdot$ are the spherical coordinates of the estimated source direction and dot-product operator. The Eq. (10) is expressed in radians; however, the RMSAE is expressed in degrees for clarity.

4.3. Experimental setups

To obtain the SRP-PHAT for each frame, we constructed a frame using a Hanning window of 4096 sizes and a 25% overlap of 16 kHz audio. The proposed model was trained for 80 epochs and it took 13 hours. We used the Adam [29] optimizer

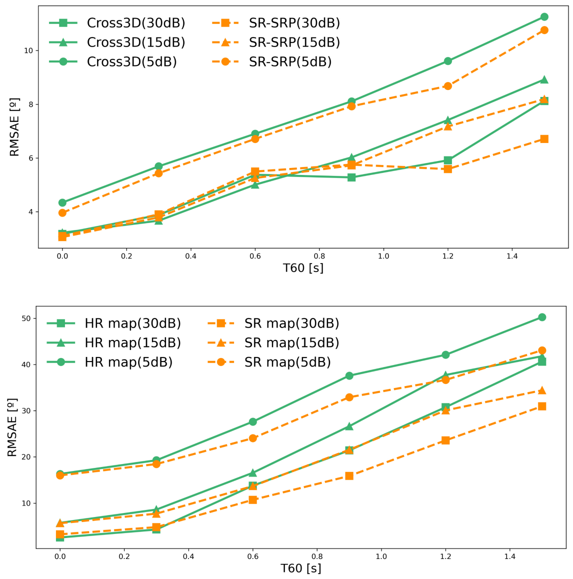


Figure 4: *Result comparison in different levels of noise and reberveration.* ***Top****: RMSAE of SR-SRP and Cross3D.* ***Bottom****: RMSAE of HR map and SR map. The silent frames were not included.*

Table 1: *Results of RMSAE to LOCATA dataset. Map size is the resolution of map. The silent frames were included.*

Model Map size	Cross3D 8×16	 16×32	 32×64	 64×128	SR-SRP 32×64	 64×128
Task1	8.65	5.99	4.26	3.77	**3.61**	3.63
Task3	14.85	11.21	9.49	8.58	8.33	**7.23**
Task5	15.08	12.31	11.19	11.52	**11.04**	11.09
Average	12.86	9.83	8.31	7.96	7.66	**7.32**

Table 2: *Result comparison of the effect of each loss function in the LOCATA dataset on RMSAE.*

Model	SR-SRP (32×64)		
$\mathcal{L}_{SR}$ $\mathcal{L}_{argmax}$	✓ ✓	✓ ✗	✗ ✓
Task1	**3.61**	4.24	4.20
Task3	**8.33**	8.79	9.31
Task5	11.04	**10.90**	11.12
Average	**7.66**	7.98	8.21

on a single NVIDIA RTX 3090 Ti and experimented using PyTorch 1.13.0 based on Ubuntu 20.04. The initial learning rate was set to 1×10^{-3}, and after 20 epochs, it was set to 1×10^{-4}. In addition, similar to Cross3D, we used a curriculum learning technique that fixed the SNR at 30 dB until the 20th epoch, and then set the SNR at 5 dB to 30 dB.

5. Results and Analyses

5.1. Performance comparison

The performance of the DOA estimation is significantly affected by the quality of the spatial spectrum. We compared the LR, HR, ideal, and SR maps under the same scenarios shown in Fig. 3 to demonstrate the superiority of the SR map. The ideal map is extracted in the same scenarios as the other power maps, under the ideal condition (SNR = 30 dB and T_{60} = 0.2 s). Under the condition that SNR = 30 dB and T_{60} = 0.6 s, the SR map shows a high peak only near the actual direction of the source, while the LR and HR maps revealed high SRP values in other regions as well (notably the central area of the map). In particular, in the worst conditions, SNR = 5 dB and T_{60} = 0.9 s, SRP-PHAT-based LR and HR maps failed to estimate the DOA of the source, while the SR map still exhibited distinct and exact peaks around of actual DOA. Compared to the ideal map, these results suggest that differences arise due to the effects of noise and reverberation. Moreover, the results demonstrate that the SR map is more robust in poor scenarios.

We compared the DOA estimation results of SR-SRP and Cross3D to validate the superiority of the proposed method. We used a simulation-based dataset and the LOCATA dataset for the performance comparison. The experimental results for the simulation-based data are shown in Fig. 4. Because random parameters generate the simulation data, we repeated the experiment five times to obtain the general results. The top graph shows that the SR-SRP outperforms Cross3D performance, particularly in noisy and reverberant scenarios. The bottom graph shows the RMSAE between the peaks and the ground truth for the HR map and the SR map, demonstrating that the SR map has more precise spatial information than the HR map. The results for the LOCATA dataset are depicted in Table 1 and show a performance improvement of approximately 8% in terms of the RMSAE compared with that of Cross3D based on the same map size. These results show that the SR-SRP is effective even for real-recording data.

5.2. Ablation studies

The ablation studies are carried out to demonstrate the influence of a proposed argmax loss function, which is suitable for the SRP-PHAT power map. The results in Table 2 show that using only the argmax loss as the unique objective function had lower performance than using only the SR loss, as it ignores the properties of SRP-PHAT and overfits to peak points. However, when using the two losses together, enabling the minimization of DOA estimation error in the feature map while maintaining the characteristics of the SRP-PHAT. As a result, the argmax loss shows superior advantages when used with the SR loss and can achieve competitive performance.

6. Conclusion

In this paper, we introduced a novel SSL and tracking model SR-SRP. The SR-SRP generated high-quality spatial spectrum which can surpass the limitation of SRP-PHAT in worse acoustic conditions. In addition, the advantage of this model was that it used an LR map with low computational cost. Experimental results demonstrate that the SR map clearly represents the actual DOA of the source, and our proposed method outperforms state-of-the-art models in poor situations.

7. Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation)

8. References

[1] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[2] N. Zheng *et al.*, "Multi-channel speaker diarization using spatial features for meetings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2022, pp. 7337–7341.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech Signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[6] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 18–21.

[7] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 20–26, 2016.

[8] X. Xiao *et al.*, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2015, pp. 2814–2818.

[9] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[10] L. Comanducci, M. Cobos, F. Antonacci, and A. Sarti, "Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 4945–4949.

[11] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and Classification for Direction-of-Arrival Estimation with Convolutional Recurrent Neural Networks," in *Proc. Interspeech*, 2019, pp. 654–658.

[12] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. European Signal Processing Conference* (*EUSIPCO*), 2018, pp. 1462–1466.

[13] K. J. Noh, J. H. Choi, D. Y. Jeon, and J. Chang, "Three-stage approach for sound event localization and detection," *Detection Classification Acoust. Scenes Events Challenge, Tech. Rep.*, 2019.

[14] 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*, vol. 29, pp. 300–311, 2021.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2015, pp. 5206–5210.

[16] 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*, pp. 5653–5671.

[17] H. W. Löllmann *et al.*, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop* (*SAM*), 2018.

[18] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[19] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution." in *Proc. European Conference on Computer Vision*, 2014, pp. 184–199.

[21] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE International Conference on Computer Vision* (*ICCV*), 2017.

[22] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2016.

[23] M. S. Rad, B. Bozorgtabar, U. V. Marti, M. Basler, H. K. Ekenel, and J. P. Thiran, "Srobb: Targeted perceptual loss for single image super-resolution," in *Proc. IEEE International Conference on Computer Vision* (*ICCV*), October 2019.

[24] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE International Conference on Computer Vision* (*ICCV*), 2016, pp. 2414–2423.

[25] T. H. Wang *et al.*, "Omnidirectional cnn for visual place recognition and navigation," in *Proc. IEEE International Conference on Robotics and Automation* (*ICRA*), 2018, pp. 2341–2348.

[26] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375.*

[27] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. European conference on computer vision* (*ECCV*), 2018, pp. 529–545.

[28] J. Wiseman, "Wiseman/py-webrtcvad," 2019.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.