INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Time Delay Estimation for Speaker Localization Using CNN-Based Parametrized GCC-PHAT Features

Daniele Salvati, Carlo Drioli, Gian Luca Foresti

Department of Mathematics, Computer Science and Physics
University of Udine, Italy
{daniele.salvati, carlo.drioli, gianluca.foresti}@uniud.it

Abstract

We propose a time delay estimation (TDE) method for speaker localization based on parametrized generalized cross-correlation phase transform (PGCC-PHAT) functions and convolutional neural networks (CNNs). The PGCC-PHAT is used to build a feature matrix, which gives TDE information of two microphone signals with different normalization levels in the cross-correlation functions. The feature matrix is processed by a CNN, composed by several convolutional layers and fully connected layers and by a regression output for the directly estimation of the time difference of arrival (TDOA). Simulations in noisy and reverberant adverse conditions show that the proposed method improves the TDOA estimation performance if compared to the GCC-PHAT.

Index Terms: time delay estimation, parametrized generalized cross-correlation, convolutional neural network, speaker localization, microphone pair, time difference of arrival.

1. Introduction

Time delay estimation (TDE) methods are used to measure the time difference of arrival (TDOA) of a sound source among two spatially separated microphones retaining a central role in the speech technology area. The speaker localization is important in applications such as human-computer interaction, teleconferencing systems, and robotics. The speaker position can be used to electronically steered a beamformer to obtain a selective spatially acquisition of the speech signal, to automatic steer a video camera in multimedia applications, or to determinate binaural cues for stereo imaging.

The generalized cross-correlation phase transform (GCC-PHAT) [1] method is the most popular TDE technique, which is based on a cross correlation function between filtered versions of the received signals. The PHAT is a filter that uses the magnitude information of the cross-correlation to normalize the narrowband components, increasing the resolution of the TDOA function if compared to a simple cross-correlation. The GCC-PHAT is thus computed in the frequency domain using the fast Fourier transform (FFT), calculating the cross-spectrum, applying the PHAT filter, and computing the inverse FFT to obtain the time-domain TDE function. The GCC-PHAT provides good TDOA estimation in moderate noisy and reverberation conditions.

However, TDE performance of the GCC-PHAT deteriorates significantly when reverberation or noise is high. Many methods have been proposed to improve robustness in adverse conditions. A class of TDE methods is based on the blind system identification [2, 3, 4, 5], which focuses on impulse responses between a source and the microphones. These methods require a certain time for the convergence of the filter to estimate the impulse responses, and in particular the direct path dominant peak. Thus, the practical application of this class of methods is very difficult. Other approaches exploit the use of redundant information among several microphones [6, 7, 8]. These methods are thus useful when more than a microphone pair is available.

Recently, the interest around the use of machine learning and multichannel processing methods is growing [9, 10, 11, 12, 13, 14, 15]. Learning-based methods have shown to be able to exploit the multidimensional characteristics of a sensor array and marked the way to new solutions. Few cases have addressed the TDE problem with machine learning [16, 17, 18, 19]. In [16], it is proposed a cross-correlation with time-frequency masking predicted by a deep neural network based on bi-directional long short term memory networks. The goal of the deep learning masking is to emphasize the time-frequency units dominated by the target speech. In [17], the cross-correlation sequences are processed by a deep neural network with an output of 10 dimensional vector for TDOA values. The frequency-sliding GCC with a convolutional neural network (CNN) is proposed in [18]. The frequency-sliding allows the calculation of sub-band GCC for an arbitrary frequency band, and the CNN is used as fully convolutional denoising autoencoder with the output an entire TDE function. In [19], the TDOA is calculated from raw waveforms with a residual CNN scheme including however a joint speaker identification and localization task.

In this paper, we propose a novel TDE technique that is based on parametrized GCC-PHAT (PGCC-PHAT) functions and CNNs. A feature matrix, which consist in the GCC function with different parametrized PHAT filters, is the input of a CNN that has a regression output for the directly estimation of the TDOA. The PGCC-PHAT has the advantage of controlling the PHAT normalization in the GCC, since the PHAT has the problem of emphasizing the noise in the frequency components that have a low signal-to-noise ratio (SNR). The PHAT performs well when the signal is broadband, i.e., when the spectral components span all the frequency range used in the GCC. However, it tends to degrade with narrowband speech components in adverse conditions. We investigate the robustness of the proposed CNN-based PGCC-PHAT with respect to adverse noisy and reverberant condition using simulated experiments.

2. Signal Model and GCC-PHAT

Let us consider a reverberant room, two microphones positioned at coordinates

$$\mathbf{r}_m = [x_m, y_m, z_m]^T, \quad m = 1, 2, \tag{1}$$

where $(\cdot)^T$ denotes the transpose operator, and a single source active at time t and positioned at coordinates

$$\mathbf{r}_s(t) = [x_s(t), y_s(t), z_s(t)]^T. \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2021-988

The signals received at the two microphones $x_1(t)$ and $x_2(t)$ can be modeled as

$$\begin{aligned} x_1(t) &= (h_1 * s)(t) + n_1(t), \\ x_2(t) &= (h_2 * s)(t) + n_2(t), \end{aligned} \tag{3}$$

where $h_1(t)$ and $h_2(t)$ represent the impulse responses of the reverberant channels, $s(t)$ is the speech signal, $n_1(t)$ and $n_2(t)$ correspond to uncorrelated noise, and * denotes linear convolution. In the short-time Fourier transform domain, the data model can be expressed as

$$\begin{aligned} X_1(f,k) &= H_1(f)S(f,k) + N_1(f,k), \\ X_2(f,k) &= H_2(f)S(f,k) + N_2(f,k), \end{aligned} \tag{4}$$

where k is the block index, f is the frequency bin, $X_1(f,k)$ and $X_1(f,k)$ are the discrete time Fourier transforms (DTFTs) of the signals observed at microphones, $S(f,k)$, $N_1(f,k)$ and $N_2(f,k)$ are the DTFTs of $s(t)$, $n_1(t)$, and $n_2(t)$ respectively, and $H_m(f)$ is the time-invariant acoustic transfer function from the source to the microphone m. We assume that the analysis window L is sufficiently long to capture most of the room impulse response such that the multiplicative transfer function approximation holds. The TDOA of the source at the microphones at the time block k is given by

$$\tau_s(k) = \frac{||\mathbf{r}_s(k) - \mathbf{r}_1|| - ||\mathbf{r}_s(k) - \mathbf{r}_2||}{c}, \tag{5}$$

where $\mathbf{r}_s(k)$ is the source position at block time k, $||\cdot||$ denotes Euclidean norm, and c is the speed of sound.

The GCC-PHAT [1] is given by

$$R(\tau,k) = \frac{1}{L}\sum_{f=0}^{L-1} \frac{X_1(f,k)X_2^*(f,k)}{|X_1(f,k)X_2^*(f,k)|} e^{\frac{j2\pi f\tau}{L}}, \tag{6}$$

where τ is the time lag, $(\cdot)^*$ denotes the complex conjugate, j denotes the imaginary unit, L is the size of the frame, and $|\cdot|$ denotes absolute value. The GCC-PHAT is computed in the frequency domain and hence the TDE is calculated on a block-by-block basis. The maximum TDOA in samples τ_{max} for the microphone pair is obtained as

$$\tau_{\text{max}} = \Big\lfloor \frac{||\mathbf{r}_1 - \mathbf{r}_2||f_s}{c} \Big\rfloor, \tag{7}$$

where $\lfloor\cdot\rfloor$ denotes the floor function that maps a real number to the largest previous integer and f_s is the sampling frequency. The admissible range of values for the TDOA is [-τ_{max},τ_{max}], thus the possible discrete TDOA values for the sensor pair are $2\tau_{\text{max}} + 1$. The function $R(\tau,k)$ is hence calculated for τ in the range [-τ_{max},τ_{max}]. The target TDOA at time block k is obtained by searching the maximum as

$$\widehat{\tau}_s(k) = \underset{\tau}{\operatorname{argmax}}[R(\tau,k)]. \tag{8}$$

The PHAT weighting function normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the GCC. It places equal importance on each frequency by dividing the spectrum by its magnitude. The PHAT increases the resolution of the TDOA function if compared to a simple cross-correlation, especially when the speech signal spans all the frequency range. However, when the SNR is low in some frequency bins, the PHAT normalization has the effect of emphasizing the noise in the GCC.

3. CNN-based Parametrized GCC-PHAT

The proposed method is based on the extraction of a feature matrix, which consists on PGCC-PHAT functions, from the two microphone signals and on the use of a CNN for mapping the feature matrix onto the target TDOA estimation.

The PHAT weighting can be generalized to parametrically control the level of influence from the magnitude spectrum [20]. This transform will be referred to as the parametrized PHAT and defined as

$$\overline{R}(\beta,\tau,k) = \frac{1}{L}\sum_{f=0}^{L-1} \frac{X_1(f,k)X_2^*(f,k)}{|X_1(f,k)X_2^*(f,k)|^{\beta}} e^{\frac{j2\pi f\tau}{L}}, \tag{9}$$

where β varies between 0 and 1. When $\beta = 1$, equation (9) becomes the conventional PHAT and the modulus of the Fourier transform becomes 1 for all frequencies, when $\beta = 0$ the PHAT has no effect on the original signal, and we have the cross-correlation function. An intermediate value of β allows the exploitation of a certain amount of the PHAT filter normalization and reduces at the same time the noise in some spectrum components where the SNR is low.

We define a feature matrix based on the PGCC-PHAT with different β values for the input of a CNN. The feature matrix is given by

$$\mathbf{R}(k) = \begin{bmatrix} \overline{R}(\beta_1,\tau_1,k) & \overline{R}(\beta_1,\tau_2,k)) & \dots & \overline{R}(\beta_1,\tau_D,k) \\ \overline{R}(\beta_2,\tau_1,k) & \overline{R}(\beta_2,\tau_2,k) & \dots & \overline{R}(\beta_2,\tau_D,k) \\ \vdots & \vdots & \vdots & \vdots \\ \overline{R}(\beta_B,\tau_1,k) & \overline{R}(\beta_B,\tau_2,k) & \dots & \overline{R}(\beta_B,\tau_D,k) \end{bmatrix}, \tag{10}$$

where $\beta_1, \beta_2, \dots, \beta_B$ are the values for the parametrized PHAT, B is total number of parametrized GCC-PHAT functions, and $\tau_1 = -\tau_{\text{max}}$, $\tau_2 = -\tau_{\text{max}} + 1, \dots, \tau_D = \tau_{\text{max}}$ with $D = 2\tau_{\text{max}} + 1$. The feature matrix has a dimension of $B \times D$.

We aim at designing a nonlinear function $F(\cdot,\mathbf{\Theta})$ ($\mathbf{\Theta}$ are the parameters learned during the training), which maps the feature matrix $\mathbf{R}(k)$ for frame signals of length L, to the output prediction TDOA $\tau_s(k)$ of the speaker

$$\tau_s(k) = F(\mathbf{R}(k),\mathbf{\Theta}). \tag{11}$$

The goal is hence to model the nonlinear function transforming the PGCC-PHAT with different β values into a TDOA estimation value.

The overall network structure is composed of several convolution layers, followed by fully-connected layers. Last, a regression layer provides the prediction of TDOA values. The data undergoes a filtering and activation detection step operated through the convolutional layer, as

$$\mathbf{H}^l = \sigma(\mathbf{W}^l * \mathbf{H}^{l-1} + b^l), \tag{12}$$

where $\mathbf{H}^l$ and $\mathbf{H}^{l-1}$ are feature maps in two consecutive layers, $\mathbf{W}^l$ is a trained kernel, b^l is a bias parameter, $\sigma(\cdot)$ is the activation function, and * denotes convolution. The rectified linear unit (ReLU) [21] is a common operation for generating the output of the convolutional layer. It computes the function $f(x) = \max(0,x)$. The bias guarantees that every node has a trainable constant value.

The output of the convolutional layers is then flattened to create a single feature vector that is used as the input of one or more fully connected layer, in which each neuron is connected to all neurons of the previous layer. A fully connected layer

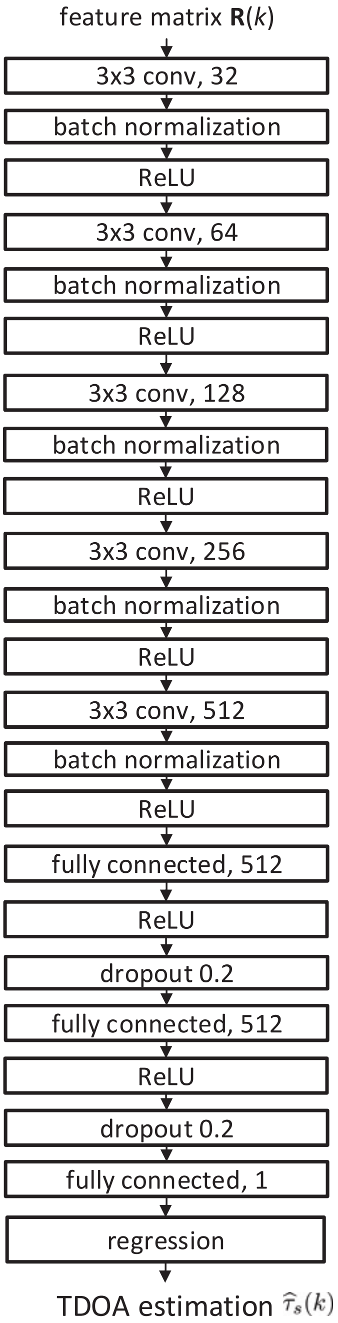

Figure 1: *The architecture of the CNN.*

multiplies the input by a weight matrix and then adds a bias vector

$$\mathbf{h}_{\text{FC}}^{l} = \sigma(\mathbf{W}_{\text{FC}}^{l}\mathbf{h}_{\text{FC}}^{l-1} + \mathbf{b}^{l}). \tag{13}$$

The last component of the nonlinear function $F(\mathbf{R}(k), \boldsymbol{\Theta})$ is the regression output layer, in which the loss function is computed with the mean squared error.

4. CNN Architecture

In this study, we use a length frame L of 1024 samples (64 ms) with a sampling rate of 16 kHz. We consider a distance between microphones of 0.2 m, and assuming $c = 340$ m/s, we have $\tau_{\text{max}} = 9$ samples, and $D = 19$. We consider $B = 11$ PHAT parameters with $\beta_1 = 0$, $\beta_2 = 0.1$, $\beta_3 = 0.2, \ldots,$ and $\beta_B = 1$. The feature matrix has hence a dimension of 11 × 19.

The architecture of the CNN mainly consists of 5 two-dimensional convolutional layers and 3 fully connected layers. After each convolutional layer, the batch normalization and the activation with the ReLU are computed. In the first convolutional layer, the number of filters is 32, and it is doubled for each subsequent convolutional layer. Each kernel of the convolutional layers has dimension 3 × 3. To enhance nonlinearity and to reduce overfitting, 3 fully connected layers are used with two dropout layers between them. The dropout layer is set with a probability of 0.2. The first and the second fully connect layers have 512 neurons. The last fully connect layer has 1 neuron for the regression output. Figure 1 shows the architecture of the CNN.

5. Simulations

The speaker localization performance is illustrated through a set of simulated experiments. The noisy conditions were conducted with different SNR levels, obtained by adding mutually independent white Gaussian noise. The reverberant conditions were simulated with an improved image-source model [22]. The source speech signals used to generate noisy and reverberant speech were taken from the TSP speech database [23]. The TSP speech database consists of 1378 utterances spoken by 23 speakers (12 females, 11 males). Each utterance has a length of about 2 s. The speech was recorded in an acoustic anechoic room. The dataset partitioning is a 70-30 split of the number of segments in training and test subsets. The training and the test subsets consist of 889 and 389 utterances, respectively.

The network parameters of the proposed CNN-based PGCC-PHAT are learned with the training dataset, simulating different source positions with an incident angle on the microphone pair in the range [-90, 90] degrees. The speaker positions were simulated with a distance of 1 m from the center of the two-microphone array. For each utterance a random SNR in the range [0, 30] dB and a random reverberation time (RT$_{60}$) in the range [0, 0.7] s were computed. The reverberation was computed with a simulated room of 5 m × 4 m × 3 m. The positions of the microphones were $(0.5, 1.9, 1.3)$ m and $(0.5, 2.1, 1.3)$ m. The distance between microphones was $d = 0.2$ m. The 889 training source positions were simulated at the same plane of the microphones, i.e., $z = 1.3$ m. The training of the CNN was computed through the Adam method [24]. The learning rate was set to 0.001, the gradient decay factor to 0.9, and the squared gradient decay factor to 0.999. The mini-batch size was set to 128, and the number of epochs to 50.

The test was conducted in a simulated room of 4 m × 7 m × 2.8 m, which is different from the training room. The positions of the microphones were $(0.2, 2.8, 1.7)$ m and $(0.2, 3, 1.7)$ m. The speaker position was randomly selected in the room with $z = 1.7$ m (the same of microphones), with a distance from the center of the array between 0.5 m and 3 m, and a minimum distance from the walls of 0.5 m. The 389 testing source positions were thus with an incident angle on the microphone pair in the range [-80, 85] degrees. The TDOA estimation performance was computed for each frame of length $L = 1024$ samples. The

Table 1: *The RMSE (ms) for the TDOA estimation performance at variation of SNR levels.*

SNR (dB)	Proposed	GCC-PHAT
30	0.0707	0.1551
25	0.0823	0.1524
20	0.0965	0.1613
15	0.1176	0.1781
10	0.1422	0.2154
5	0.1707	0.2613
0	0.2017	0.3141
-5	0.2528	0.3744
-10	0.2993	0.4083
-15	0.3593	0.4621

Table 2: *The RMSE (ms) for the TDOA estimation performance at variation of reverberant conditions with an SNR of 30 dB.*

RT$_{60}$ (s)	Proposed	GCC-PHAT
0.1	0.0768	0.1629
0.2	0.1024	0.1963
0.3	0.1304	0.2357
0.4	0.1671	0.2766
0.5	0.1830	0.2840
0.6	0.2096	0.3180
0.7	0.2209	0.3265
0.8	0.2392	0.3426
0.9	0.2398	0.3492
1.0	0.2502	0.3584

TDE performance is measured with the root mean squared error (RMSE) expressed in ms.

First, a simulation at variation of noisy conditions was conducted. Table 1 reports the results at variation of the SNR level. We can observe the improved accuracy and the robustness to noise of the proposed method if compared to the GCC-PHAT. We have a reduction of RMSE at an SNR of 30 dB with 0.1551-0.0707=0.0844 ms for the proposed method. Note than 1 sample for the GCC-PHAT discretization corresponds to 0.0625 ms. Up to 5 dB of SNR, we have a RMSE reduction minor than 0.1 ms and the RMSE reduction is greater than 0.1 m at low SNR ($\leq$ 0 dB).

Next, an evaluation at variation of reverberant conditions was performed. Table 2 shows the TDE performance with a RT$_{60}$ in the range [0.1,1.0] s. The SNR was 30 dB. The improved accuracy of the proposed method is obtained in all conditions. We can underline the ability of the CNN-based PGCC-PHAT in estimating the TDOA in a room with different reflection characteristics in comparison to the training simulated room.

Last, in Table 3, we can see the TDOA estimation performance in noisy condition (SNR=5 dB) at variation of reverberant times. We can observe also in these simulations the improved performance in comparisons to the GCC-PHAT.

Finally, Table 4 shows the RMSE at variation of the feature matrix size B, i.e., the number of parametrized PHAT functions. We consider $B = 3$ ($\beta_1 = 0$, $\beta_2 = 0.5$, $\beta_3 = 1$), $B = 11$ (the case of previous simulations) and $B = 21$ (from 0 to 1 with step 0.05). We can note the improvement performance with $B = 3$ if compared to the GCC-PHAT. However, we have the lower RMSE when $B = 11$. In case of $B = 21$, a larger number of parametrized PHAT functions does not provide an improvement of the TDOA performance in comparison to the $B = 11$ case.

Table 3: *The RMSE (ms) for the TDOA estimation performance at variation of reverberant conditions in noisy conditions (SNR = 5 dB).*

RT$_{60}$ (s)	Proposed	GCC-PHAT
0.1	0.1749	0.2648
0.2	0.1971	0.3029
0.3	0.2153	0.3245
0.4	0.2461	0.3539
0.5	0.2551	0.3729
0.6	0.2617	0.3780
0.7	0.2804	0.3938
0.8	0.2862	0.3985
0.9	0.2924	0.4055
1.0	0.3037	0.4130

Table 4: *The RMSE (ms) for the TDOA estimation performance at variation of the feature matrix size B. The RT$_{60}$ was 0.8 s and the SNR was 10 dB.*

Proposed			GCC-PHAT
$B = 3$	$B = 11$	$B = 21$	
0.2808	0.2647	0.2676	0.3779

To conclude, we have the following average RMSE performance of all simulations reported in Tables 1, 2, 3:

- 0.2042 ms for the proposed method;
- 0.3047 ms for the GCC-PHAT.

The CNN-based PGCC-PHAT is hence able of decreasing of about 30 % the RMSE of the GCC-PHAT.

6. Conclusions

In this paper, we presented a TDE estimation method for speaker localization using a microphone pair. The proposed method consists in the computation of the feature matrix using different normalized PGCC-PHAT functions. A CNN scheme is proposed for the recognition of the feature matrices to directly estimate the TDOA of the speech source at the microphones. We have demonstrated that the CNN-based PGCC-PHAT increases the accuracy of the TDOA estimation and it is more robust to noise and reverberation if compared to the GCC-PHAT.

Future works include the performance evaluation with real-world data and the analysis of the computational cost for the evaluation in realtime applications.

7. Acknowledgements

This work is partially supported by the ONRG project N62909-20-1-2075 ”Target Re-Association for Autonomous Agents” (TRAAA).

8. References

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

[2] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[3] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[4] J. Cho and H. Park, "Imposition of sparse priors in adaptive time delay estimation for speaker localization in reverberant environments," *IEEE Signal Processing Letters*, vol. 16, no. 3, pp. 180–183, 2009.

[5] D. Salvati and S. Canazza, "Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 507–510, 2013.

[6] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 549–557, 2003.

[7] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1479–1489, 2008.

[8] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "TDOA-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1976–1990, 2018.

[9] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," *Pattern Recognition Letters*, vol. 84, pp. 15–21, 2016.

[10] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, pp. 136–140.

[11] D. Salvati, C. Drioli, and G. L. Foresti, "On the use of machine learning in microphone array beamforming for far-field sound source localization," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2016.

[12] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5210–5214.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[14] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 103–116, 2018.

[15] 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*, vol. 18, no. 10, 2018.

[16] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Proceedings of the Conference of the International Speech Communication Association*, 2018, pp. 322–326.

[17] J. Ding, B. Ren, and N. Zheng, "Microphone array acoustic source localization system based on deep learning," in *Proceedings of the International Symposium on Chinese Spoken Language Processing*, 2018, pp. 409–413.

[18] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 4945–4949.

[19] D. Salvati, C. Drioli, and G. L. Foresti, "Two-microphone end-to-end speaker joint identification and localization via convolutional neural networks," in *Proceedings of the International Joint Conference on Neural Networks*, 2020, pp. 1–6.

[20] K. D. Donohue, J. Hannemann, and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Processing*, vol. 87, no. 7, pp. 1677–1691, 2007.

[21] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the International Conference on Machine Learning*, 2010, pp. 807–814.

[22] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[23] P. Kabal, "TSP speech database," McGill University, Montreal, Quebec, Tech. Rep., 2002.

[24] D. P. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015, pp. 1–13.