Two-Microphone End-to-End Speaker Joint Identification and Localization Via Convolutional Neural Networks

Daniele Salvati, Carlo Drioli, Gian Luca Foresti
Department of Mathematics, Computer Science and Physics
University of Udine
Udine, Italy
{daniele.salvati, carlo.drioli, gianluca.foresti}@uniud.it

***Abstract*—We present an end-to-end scheme based on convolutional neural networks (CNNs) for speaker joint identification and localization. We investigate the possibility to estimate both the direction of arrival (DOA) and the identity of the speaker in far-field noisy and reverberant conditions using a two-channel microphone array. The proposed CNN network is designed to map the raw waveform of the two channels into the speaker identity and into the DOA of its speech signal. We analyze the identification and localization performance with simulated experiments in noisy and reverberation conditions.**

***Index Terms*—Convolutional neural network, end-to-end system, raw waveform, speaker identification, speaker localization, two-microphone array.**

I. Introduction

Microphone array processing techniques retain a central role in applications such as human-computer interaction, teleconferencing systems, audiovisual surveillance, robotics, automation, and in a number of applications in the speech technology area. These include speaker localization [1], speech enhancement [2], speaker/speech recognition [3].

Speaker identification and speaker localization are processing tasks widely investigated in the past years by the signal processing community. Speaker localization, and more in general acoustic localization, is inherently related to multichannel array processing and traditionally relies on measurements of time difference of arrivals (TDOAs) across various combinations of microphones [4] and on geometric considerations to estimate the source position [5], or on steered response power (SRP) beamformers [6]. Recently, however, the interest around the use of machine learning methods for the source localization has increased [7]–[10].

Speaker identification is traditionally faced by single-channel processing techniques and pattern recognition models [11]. However, it is recognized that multichannel processing can be used to enhance the acoustic front-end involved in speaker/speech recognition since it can help reducing background noise, reverberation, source-point interference, especially in distant-talking conditions [12]. Sensor array techniques such as beamforming [13] and multichannel noise reduction [14] can greatly improve the recognition accuracy in adverse acoustic conditions. Some possibilities of exploiting the information gathered from a multichannel system have been discussed for example in [3], [15]–[18].

Since many decades, machine learning and neural network methods have been successfully employed in a wide range of speech and audio processing applications, such as automatic speech recognition [19], audio forensic [20], music information retrieval [21], sound classification [22]. However, their use for the improvement or the new design of multichannel processing localization schemes has been explored only recently [7], [8], [10], [23]–[26]. Moreover, since the new computational and performance advances brought by the recent developments in the field of deep neural networks (DNNs) research, their use is now being investigated in a variety of acoustic and speech oriented applications involving multichannel processing, including in a few cases the specific problem of acoustic source localization. In [9], an approach is developed that uses a discriminative machine learning to compute the location estimator in the frequency domain, in which a DNN encodes the steering vectors by applying the orthogonality principle used in the multiple signal classification (MUSIC) method [27]. The method however resulted ineffective in noisy and reverberant conditions. In [7], the multichannel spectral phase information is used as input of a convolutional neural network (CNN) for the direction of arrival (DOA) estimation. In [10], a CNN-based scheme is proposed to refine the multichannel fusion scheme of the minimum variance distortionless response (MVDR) beamformer and improves the localization of acoustic sources and speakers in far-field noisy and reverberant environments. In [8], a CNN-based end-to-end system is studied, that maps the raw waveforms of a distributed microphone network to the source position in reverberant environments.

While speaker identification and localization have been deeply investigated as independent modules (for an overview see [12]), the study of joint identification and localization is rather limited in the literature. An example is the work in [28], developed in the context of binaural applications using an artificial head. We recently discussed joint speaker classification and localization [29] based on the diagonal unloading beamforming [30], [31] and using a uniform linear

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

array. In both [28] and [29], the approach consists of using different processing blocks performing localization, signal enhancement, and speaker identification, where however each block is connected to the others to improve their respective performance in isolation, without the use of DNN components. An example of multi-output model using DNNs is proposed in [32], in which an end-to-end scheme based on a CNN using spectral information as input provides joint output with localization and speech/non-speech classification in noisy environments. In the field of the sound event localization and detection (SELD), DNN-based methods using multi-output have been recently proposed in [33], [34]. In this regard, to the best of the authors' knowledge, the simultaneous identification and localization of speech signal using only a DNN-based approach has not yet been explored.

In this paper, we discuss the performance of the joint speaker identification and localization scheme based on CNNs. We adopt an end-to-end design, in which the raw audio signal is used as the input of the network, without any further acoustic front-end processing. Moreover, we investigate the possibility of training the CNN to both estimate the DOA and to recognize the identity of the speaker using a two-microphone array. We investigate the robustness of the model with respect to adverse conditions, namely to process speech recorded in noisy and reverberant environments.

II. Raw Waveform CNN Acoustic Model

Let us refer to a two-microphone array with an inter-microphone distance d. Suppose that a single source impinges upon the array and let $s_n(t) \in \mathbb{R}$ denote the signal generated by a nonstationary speech source n ($n = 1, 2, \ldots, N$, where n refers to the nth speaker in a dataset containing recordings from N speakers). The outputs of the two sensors are given by

$$\begin{aligned} x_1(t) &= (h_1 * s_n)(t) + v_1(t), \\ x_2(t) &= (h_2 * s_n)(t) + v_2(t), \end{aligned} \tag{1}$$

where $h_1(t)$ and $h_2(t)$ are the impulse responses from the source to the sensors, and $v_1(t)$ and $v_2(t)$ are additive noise signals that are assumed to be uncorrelated and spatially white Gaussian with zero mean and variance equal to σ^2 for both sensors. Referring to a far-field model for the sound source wave propagation, the source impinges upon the two-microphone array with a DOA θ, which is given by

$$\theta = \arcsin\left(\frac{c\tau}{d}\right), \tag{2}$$

where τ is the TDOA of the wavefront at the two microphones, and c is the speed of sound. Under the hypothesis of ideal reflections, the impulse responses can be expressed as

$$\begin{aligned} h_1(t) &= \sum_{q=0}^{Q_1} \alpha_{q,1}\delta(t - t_{q,1}), \\ h_2(t) &= \sum_{q=0}^{Q_2} \alpha_{q,2}\delta(t - t_{q,2}), \end{aligned} \tag{3}$$

where Q_i is the number of room reflections, $\alpha_{q,i}$ is an attenuation term, δ is the Dirac delta function, and $t_{q,i}$ is the time of arrival of the qth reflection. The direct-path signal is defined as the component corresponding to $q = 0$. The TDOA of the wavefront at the two microphones can be written as

$$\tau = t_{0,1} - t_{0,2}. \tag{4}$$

We aim at designing a nonlinear function $F(\cdot, \mathbf{\Theta})$ ($\mathbf{\Theta}$ being the parameters learned during the training), which maps the input raw waveforms $\mathbf{x}(t)$ of the n-th speaker to the output prediction $\mathbf{o}(t)$

$$\mathbf{o}(t) = F(\mathbf{x}(t), \mathbf{\Theta}). \tag{5}$$

The input corresponding to a frame of length L is a vector composed of the signals from the two channels

$$\begin{aligned} \mathbf{x}(t) = [\mathbf{x}_1(t), \mathbf{x}_2(t)] = [&x_1(t-L+1), x_1(t-L+2), \ldots, \\ &x_1(t), x_2(t-L+1), x_2(t-L+2), \ldots, x_2(t)]. \end{aligned} \tag{6}$$

The input vector has thus dimension $2L$. The multi-output consists of a classification module for the speaker identification and a regression module for the DOA of the source:

$$\mathbf{o}(t) = [\mathbf{p}_n, \theta], \tag{7}$$

where $\mathbf{p}_n$ is the prediction vector of length N for the classification layer.

The overall structure of the one-dimensional convolutional CNN network $F(\cdot, \mathbf{\Theta})$ is made of several convolutional layers, followed by two outputs (speaker classification and DOA regression) provided by a fully-connected layer. The data undergoes a filtering and activation detection step operated through the one-dimensional convolution,

$$\mathbf{h}^l = \sigma(\mathbf{w}^l * \mathbf{h}^{l-1} + b^l), \tag{8}$$

where $\mathbf{h}^l$ and $\mathbf{h}^{l-1}$ are feature maps of 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), computed by the function $f(x) = \max(0, x)$ [35], is a common operation for generating the output of the convolutional layer. The bias guarantees that every node has a trainable constant value. The kernels are computed through an optimization method, which minimizes a loss function measuring the discrepancy between the CNN prediction and the target. We use in this work the Adam optimizer [36]. The loss function for classification is the cross entropy and for regression is the half-mean-squared-error. To speed up the training of CNNs and reduce the sensitivity to network initialization, the batch normalization is used to normalize the data across a mini-batch, back-propagating the gradients through the normalization parameters [37].

The output of the convolutional layers is then used as the input of the fully connected layer, in which each neuron is connected to all neurons of the previous layer. A fully connected layer multiplies the input by a weight matrix and then adds a bias vector

$$\mathbf{h}^l = \sigma(\mathbf{W}^l\mathbf{h}^{l-1} + \mathbf{b}^l). \tag{9}$$

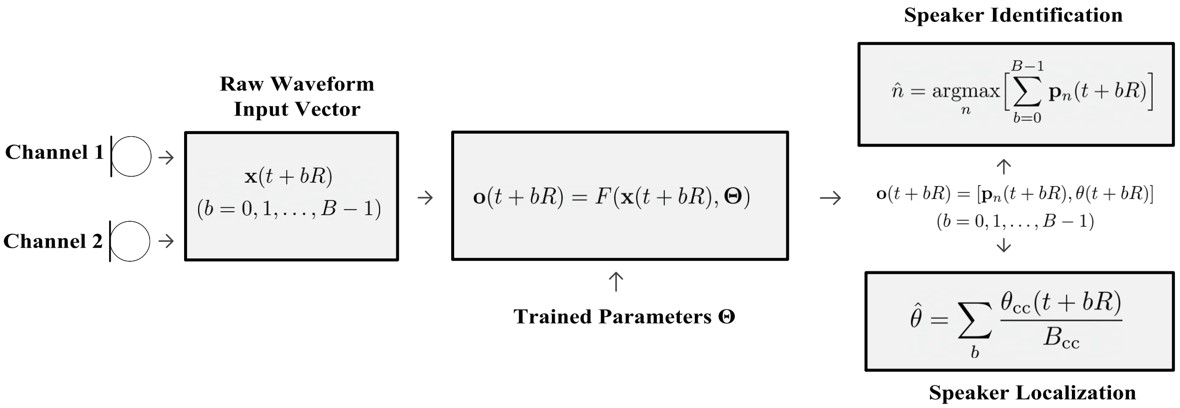
Channel 1
Channel 2
Raw Waveform Input Vector
$\mathbf{x}(t+bR)$
$(b=0,1,\ldots,B-1)$
$\mathbf{o}(t+bR)=F(\mathbf{x}(t+bR),\mathbf{\Theta})$
Trained Parameters $\mathbf{\Theta}$
Speaker Identification
$\hat{n}=\underset{n}{\mathrm{argmax}}\Big[\sum_{b=0}^{B-1}\mathbf{p}_n(t+bR)\Big]$
$\mathbf{o}(t+bR)=[\mathbf{p}_n(t+bR),\theta(t+bR)]$
$(b=0,1,\ldots,B-1)$
$\hat{\theta}=\sum_b\frac{\theta_{\mathrm{cc}}(t+bR)}{B_{\mathrm{cc}}}$
Speaker Localization

Fig. 1. Information flow of the system: a one-dimensional CNN is used to map the raw waveforms from the two channels to the speaker identification and the DOA of the source.

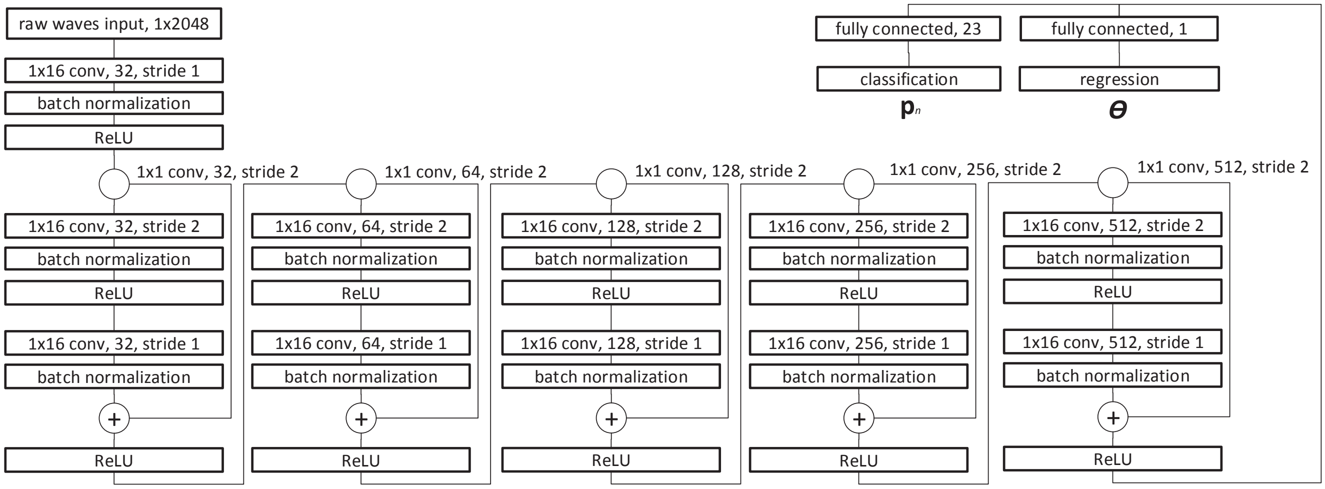

Fig. 2. The architecture of the proposed residual CNN.

III. Speaker Identification and Localization

Since a speech signal can vary in temporal length, we propose a network scheme to handle variable-length signals. Specifically, we design the CNN input using a short signal frame of length L. Using this short-frame analysis setting, the network results suitable to analyze speech segments. A short frame setting is also more advantageous for the network to identify the differences between the two channels of each frame and to estimate the DOA. The speaker identification and localization based on the raw waveform CNN acoustic model is computed using a signal segment composed of B frames of length L. The sequence of input vectors is $\mathbf{x}(t+bR)$, $b=0,1,\ldots,B-1$, where R is the overlap step. Each input vector $\mathbf{x}(t+bR)$ of size $2L$ is processed by the CNN, which estimates B identification prediction values and DOAs. From B outputs, we have that $\mathbf{o}(t+bR)=[\mathbf{p}_n(t+bR),\theta(t+bR)]$, where $\mathbf{p}_n(t+bR)$ and $\theta(t+bR)$ are the prediction outputs at time $t+bR$ for the identification and the DOA respectively.

The speaker identification is calculated as

$$\hat{n}=\underset{n}{\mathrm{argmax}}\Big[\sum_{b=0}^{B-1}\mathbf{p}_n(t+bR)\Big], \tag{10}$$

i.e., the output index providing the maximum sum prediction value over B frames is returned as the identified speaker index.

The speaker localization is calculated by averaging the DOA results over the B frames. Outliers are removed in the average processing using the Chauvenet's criterion [38]. We assume that the source position is stationary in the B frames. The DOA estimation can be written as

$$\hat{\theta}=\sum_b\frac{\theta_{\mathrm{cc}}(t+bR)}{B_{\mathrm{cc}}}, \tag{11}$$

where B_{cc} is the number of elements that are accepted from the Chauvenet's criterion, which is

$$\theta_{\mathrm{cc}}(t+bR)=\Big\{\theta(t+bR):\mathrm{erfc}(\eta)\geq\frac{1}{2B}\Big\}, \tag{12}$$

where erfc$(\cdot)$ is the complementary error function, and

$$\eta = \frac{|\theta(t+bR)-\mu|}{\sigma}, \tag{13}$$

with

$$\mu = \sum_{b=0}^{B-1} \frac{\theta(t+bR)}{B}, \tag{14}$$

$$\sigma = \sqrt{\frac{\sum_{b=0}^{B-1}(\theta(t+bR)-\mu)^2}{B}}. \tag{15}$$

The information flow of the raw waveform CNN acoustic model designed for speaker identification and localization using a two-microphone array is summarized in Figure 1.

IV. Proposed CNN Architecture

In this section, the architecture of the proposed CNN will be described in detail. The network is based on a deep residual learning model [39]. Residual models are implemented with double-layer skips that contain nonlinearities (ReLU) and batch normalization in between. We carefully tune the kernel size and the number of filters of the convolutional layers to obtain an optimal performance. The CNN is composed of an initial convolution layer and of 5 residual modules. Each kernel of the convolutional layers has dimension 1×16. Zero padding is used to obtain equal size of the input and output. The reduction of the feature size is achieved in each residual module by using a stride of 2 in the first convolutional layer. In the first residual module, the number of filters is 32, and it is doubled for each subsequent residual module. The multi-output consists of a classification output, a branch with a fully connected operation of size N (the number of speakers) and a softmax operation, and a regression output, a branch with a fully connected operation of size 1 (the DOA response). Specifically, we consider $N = 23$ speakers (12 females, 11 males) in this study.

We use here a length frame L of 1024 samples (64 ms) with a sampling rate of 16 kHz, resulting in a raw waveform input vector for the CNN of 2048 samples. The size of convolutional kernel is the same for all layers. This setting allows the increasing of the filter resolution at each consequentially residual module due to the downsampling operated by using a stride of 2 in the first convolutional layer of each residual module. The size of the feature maps is hence 64 samples with 512 filters. Figure 2 shows the architecture of the proposed residual CNN.

V. Simulations

The speaker identification and localization performance is illustrated through a set of simulated experiments. The simulations in noisy conditions were conducted with different signal-to-noise ratio (SNR) levels, obtained by adding mutually independent white Gaussian noise. The experiments in reverberant conditions was simulated with an improved image-source model [40]. The source speech signals used to generate noisy and reverberant speech were taken from the TSP speech database [41]. 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 performance was computed for each utterance with an overlap step of $R = 512$ samples. The number of blocks B for the test subset was in the range $[52, 103]$. Each input segment of length $2L = 2048$ samples was normalized (peak normalization) before passing to the CNN.

TABLE I
The identification performance IA (%) in noisy conditions.

SNR (dB)	30	20	10	0
Proposed	99.74	99,14	98.46	85.25
MFCC-CNN	71.72	68.38	60.93	40.36

TABLE II
The localization performance RMSE (degree) in noisy conditions.

SNR (dB)	30	20	10	0
Proposed	3.28	3.32	4.17	5.53
GCC-PHAT	3.13	3.22	3.29	4.27

The model learning was conducted on the training subset simulating different DOAs with a spatial resolution of 1 degree in the range [-90, 90] degrees. The speaker positions were simulated with a distance of 1 m from the center of the 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 $(0.5, 2.1, 1.3)$ m and $(0.5, 1.9, 1.3)$ m. The distance between microphones was $d = 0.2$ m. The training of the CNN was computed through the Adam method. 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 512, and the number of epochs to 100.

We compared the performance of the speaker identification and localization based on the proposed CNN with the speaker identification based on the mel-frequency cepstral coefficient (MFCC) CNN [42] using a single microphone and with the DOA estimation based on the generalized cross-correlation phase transform (GCC-PHAT) [4]. The MFCC-CNN was trained with the same dataset. MFCCs vectors of length 21 were used as input (the zero-th order coefficient is excluded). The MFCCs are calculated with a frame of 1024 samples and they are normalized with zero mean and standard deviation equal to 1. The GCC-PHAT was computed by averaging the functions obtained in each frame and calculating the TDOA and the DOA on the average GCC-PHAT. Performance is reported in terms of the percentage of identification accuracy (IA) and in terms of root mean square error (RMSE) for DOA estimation.

TABLE III
THE IDENTIFICATION PERFORMANCE IA (%) IN REVERBERANT CONDITIONS.

RT$_{60}$ (s)	0.1	0.3	0.5	0.7
Proposed	94.86	89.66	88.60	87.80
MFCC-CNN	90.23	91.43	92.20	89.94

TABLE IV
THE LOCALIZATION PERFORMANCE RMSE (DEGREE) IN REVERBERANT CONDITIONS.

RT$_{60}$ (s)	0.1	0.3	0.5	0.7
Proposed	4.96	6.09	9.82	15.52
GCC-PHAT	3.95	4.67	6.96	12.12

First, a simulation in noisy conditions was conducted. The RT$_{60}$ was 0 (anechoic condition). Each utterance of the test dataset was randomly positioned by considering a minimum distance from the wall of 0.5 m and a minimum and a maximum distance between source and the center of the array of 0.5 and 3 m, respectively. Tables I and II report the results at variation of the SNR level. Looking at the identification results (Table I), we can observe that the robustness to noise of the proposed method is superior if compared to the single-channel MFCC-CNN, and that the performance degrades when the SNR level decreases. On the other hand, the localization accuracy of the proposed method is good and it is comparable to that of the GCC-PHAT, although the GCC-PHAT is slightly more accurate at low SNRs.

Next, an evaluation in reverberant conditions was performed. The SNR was 30 dB. Tables III and IV show the IA performance and RMSE localization, respectively. In low reverberation condition (RT$_{60}$=0.1 s), the proposed method has a slightly better IA performance when compared to the MFCC-CNN (Table III). However, when the reverberation time increases, the proposed method performance degrades resulting more sensitive to the effect of the multi-path propagation and to the relative position between source and microphones. The MFCC-CNN IA instead tends to increase for a RT$_{60}$ of 0.3 s and 0.5 s. We can note that the MFCC-CNN performance in reverberant conditions is higher when compared to the results in Table I, due to the fact that most of the samples of the training dataset are corrupted by reverberation. The localization performance of the proposed method is good, but it tends to degrade at increasing of reverberation time in comparison to the GCC-PHAT (Table IV).

Finally, the results in noisy and reverberation conditions were reported in Table V. The SNR was 5 dB and RT$_{60}$ was 0.4 s. The proposed two-microphone residual CNN model is rather promising since it provides robustness to noise and to reverberation. Future work includes the use of more microphones for the DOA estimation in the entire 3D space.

TABLE V
THE IDENTIFICATION PERFORMANCE IA (%) AND THE LOCALIZATION PERFORMANCE RMSE (DEGREE) IN NOISY AND REVERBERANT CONDITIONS. THE SNR WAS 5 dB AND RT$_{60}$ WAS 0.4 s.

IA (%)		RMSE (degree)	
Proposed	**MFCC-CNN**	**Proposed**	**GCC-PHAT**
67.58	59.04	15.28	12.93

VI. CONCLUSIONS

We presented an end-to-end raw waveform CNN acoustic model for joint identification and localization of a speaker. The residual CNN architecture is designed to operate in a frame-by-frame analysis to handle variable-length signals. We have shown that the two-microphone proposed method provides a good joint identification and localization performance in adverse noisy and reverberant conditions.

REFERENCES

[1] D. Salvati, C. Drioli, and G. L. Foresti, "Sensitivity-based region selection in the steered response power algorithm," *Signal Processing*, vol. 153, pp. 1–100, 2018.

[2] L. Kumar and R. M. Hegde, "Near-field acoustic source localization and beamforming in spherical harmonics domain," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3351–3361, 2016.

[3] I.-J. Ding and J.-Y. Shi, "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots," *Computers & Electrical Engineering*, vol. 62, pp. 719–729, 2017.

[4] 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.

[5] P. Stoica and J. Li, "Source localization from range-difference measurements," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 63–66, 2006.

[6] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, ch. Robust localization in reverberant rooms.

[7] 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 (WASPAA)*, 2017, pp. 136–140.

[8] 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.

[9] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 405–409.

[10] 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.

[11] J. P. Campbell Jr., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[12] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[13] B. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[14] S. Doclo, M. Moonen, T. V. den Bogaert, and J. Wouters, "Reduced bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 38–51, 2009.

[15] Q. Lin, E.-E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 622–629, 1994.

[16] I. A. McCowan and S. Sridharan, “Multi-channel sub-band speech recognition,” *EURASIP Journal on Applied Signal Processing*, pp. 45–52, 2001.

[17] J. W. Stokes, J. C. Platt, and S. Basu, “Speaker identification using a microphone array and a joint HMM with speech spectrum and angle of arrival,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2006, pp. 1381–1384.

[18] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, “Deep beamforming networks for multi-channel speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5745–5749.

[19] Y. Qian, M. Bi, T. Tan, and K. Yu, “Very deep convolutional neural networks for noise robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[20] X. Lin, J. Liu, and X. Kang, “Audio recapture detection with convolutional neural networks,” *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1480–1487, 2016.

[21] S. Durand, J. P. Bello, B. David, and G. Richard, “Robust downbeat tracking using an ensemble of convolutional networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 76–89, 2017.

[22] J. Salamon and J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] T. May, S. van de Par, and A. Kohlrausch, “A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, 2012.

[29] D. Salvati, C. Drioli, and G. L. Foresti, “Joint identification and localization of a speaker in adverse conditions using a microphone array,” in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 21–25.

[30] D. Salvati, C. Drioli, and G. L. Foresti, “A low-complexity robust beamforming using diagonal unloading for acoustic source localization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 609–622, 2018.

[31] D. Salvati, C. Drioli, and G. L. Foresti, “Power method for robust diagonal unloading localization beamforming,” *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 725–729, 2019.

[32] W. He, P. Motlicek, and J.-M. Odobez, “Joint localization and classification of multiple sound sources using a multi-task neural network,” in *Proceedings of the Conference of the International Speech Communication Association*, 2018, pp. 312–316.

[33] T. Hirvonen, “Classification of spatial audio location and content using convolutional neural networks,” in *Audio Engineering Society Convention*, 2015.

[34] 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*, vol. 13, no. 1, pp. 34–48, 2019.

[35] 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.

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

[37] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv:1502.03167*, 2015.

[38] V. Barnett and T. Lewis, *Outliers in statistical data*. John Wiley & Sons, 1994.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[40] 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.

[41] P. Kabal, “TSP speech database,” McGill University, Montreal, Quebec, Tech. Rep., 2002.

[42] D. Salvati, C. Drioli, and G. L. Foresti, “End-to-end speaker identification in noisy and reverberant environments using raw waveform convolutional neural networks,” in *Proceedings of the Conference of the International Speech Communication Association*, 2019, pp. 4335–4339.