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...
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 reverberant 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 use in this work the Adam optimizer [36]. The multi-output consists of a classification module for the speaker identification and a regression module for the DOA of the source:

$$o(t) = F(x(t), \Theta).$$

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:

$$o(t) = [p_n, \theta],$$

where $p_n$ is the prediction vector of length $N$ for the classification layer.

The overall structure of the one-dimensional convolutional CNN network $F(\cdot, \Theta)$ 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,

$$h^l = \sigma(w^l \ast h^{l-1} + b^l),$$

where $h^l$ and $h^{l-1}$ are feature maps of two consecutive layers, $w^l$ is a trained kernel, $b^l$ is a bias parameter, $\sigma(\cdot)$ is the activation function, and $\ast$ 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

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

\[ x(t + bR) \]
\[ \theta(t + bR) = F(x(t + bR), \Theta) \]

**Speaker Identification**
\[ \hat{n} = \arg\max_n \left( \sum_{b=0}^{B-1} p_n(t + bR) \right) \]

**Speaker Localization**
\[ \hat{\theta} = \frac{\sum_b \theta_{cc}(t + bR)}{B_{cc}} \]

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 \( x(t + bR) \), \( b = 0, 1, \ldots, B - 1 \), where \( R \) is the overlap step. Each input vector \( x(t + bR) \) of size \( 2L \) is processed by the CNN, which estimates \( B \) identification prediction values and DOAs. From \( B \) outputs, we have that \( \alpha(t + bR) = [p_n(t + bR), \theta(t + bR)] \), where \( p_n(t + bR) \) 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
\[
\hat{n} = \arg\max_n \left( \sum_{b=0}^{B-1} p_n(t + bR) \right),
\]
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
\[
\hat{\theta} = \frac{\sum_b \theta_{cc}(t + bR)}{B_{cc}} ,
\]
where \( B_{cc} \) is the number of elements that are accepted from the Chauvenet’s criterion, which is
\[
\theta_{cc}(t + bR) = \left\{ \theta(t + bR) : \text{erfc}(\eta) \geq \frac{1}{2B} \right\}.
\]
where \( \text{erfc} (\cdot) \) is the complementary error function, and

\[
\eta = \frac{\theta(t + bR) - \mu}{\sigma},
\]

with

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

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

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 \( 1 \times 16 \). 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 operation 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 were 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.

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 (RT60) in the range [0.2, 0.7] s were computed. The reverberation was simulated with a simulated room of 5 m \( \times \) 4 m \( \times \) 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>
<thead>
<tr>
<th>SNR (dB)</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>99.74</td>
<td>99.14</td>
<td>98.46</td>
<td>85.25</td>
</tr>
<tr>
<td>MFCC-CNN</td>
<td>71.72</td>
<td>68.38</td>
<td>60.93</td>
<td>40.36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>3.28</td>
<td>3.32</td>
<td>4.17</td>
<td>5.53</td>
</tr>
<tr>
<td>GCC-PHAT</td>
<td>3.13</td>
<td>3.22</td>
<td>3.29</td>
<td>4.27</td>
</tr>
</tbody>
</table>

TABLE I

The identification performance IA (%) in noisy conditions.

TABLE II

The localization performance RMSE (degree) in noisy conditions.
First, a simulation in noisy conditions was conducted. The RT₆₀ 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 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 (RT₆₀=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 RT₆₀ 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 RT₆₀ 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.

### 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


