CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings

Lauréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin

CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings

Lauréline Perotin, *Student Member, IEEE,* Romain Serizel, *Member, IEEE,*
Emmanuel Vincent, *Senior Member, IEEE,* and Alexandre Guérin

***Abstract*—Localizing audio sources is challenging in real reverberant environments, especially when several sources are active. We propose to use a neural network built from stacked convolutional and recurrent layers in order to estimate the directions of arrival of multiple sources from a first-order Ambisonics recording. It returns the directions of arrival over a discrete grid of a known number of sources. We propose to use features derived from the acoustic intensity vector as inputs. We analyze the behavior of the neural network by means of a visualization technique called layerwise relevance propagation. This analysis highlights which parts of the input signal are relevant in a given situation. We also conduct experiments to evaluate the performance of our system in various environments, from simulated rooms to real recordings, with one or two speech sources. The results show that the proposed features significantly improve performances with respect to raw Ambisonics inputs.**

***Index Terms*—Audio source localization, direction of arrival, first-order Ambisonics, acoustic intensity, convolutional recurrent neural network, layerwise relevance propagation.**

I. Introduction

More and more applications, such as smart home assistants and spatial audio acquisition, rely on far-field audio recordings. In this context, it is important to know the directions of arrival (DoAs) of the sounds, in order either to enhance the signals of interest or to reproduce the sound scene properly. For instance, DoA estimation is essential for speech enhancement and robust far-field automatic speech recognition in scenarios involving overlapping speakers [1]–[5].

In order to capture the spatial information, the sound scene must be recorded with multiple microphones. Arranging the microphones as a spherical array ensures that no direction in space is favored. The recordings can then be stored in Ambisonics format [6]. This format is more and more employed in the industry, e.g., in the MPEG-H standard [7]. It is also particularly suitable for DoA estimation [8]–[10] as it directly encodes the spatial properties of the sound field.

DoA estimation has been extensively investigated in the past decades [11]–[14]. Time difference of arrival (TDoA) based methods estimate the TDoA for each microphone pair by means of, e.g., generalized cross-correlation with phase transform (GCC-PHAT), and combine it across all microphone pairs to derive the DoA of the dominant source [15]. Steered response power (SRP) based methods explore the space with a beamformer, where the areas of higher energy reveal possible source positions [16]. These methods provide good DoA estimates for a single source, but generally not for multiple sources, especially if they come from close DoAs. Subspace methods such as MUSIC [17] or ESPRIT [18] and their adaptations to Ambisonics [19], [20] are suited to multi-source situations. Another set of methods exploit the sound field characteristics: they mainly rely on the estimation of the acoustic intensity vector, which represents the flow of energy in each frequency band and provides an estimate of the source DoAs [9], [21], [22]. However, in the presence of noise and reverberation, the accuracies of all the aforementioned methods decrease dramatically [10], [23].

Recently, neural networks have improved the robustness of DoA estimation techniques in such adverse conditions. They have been used with various inputs: binaural features [24], GCC features [25], the eigenvectors of the spatial covariance matrix [26], raw short-time Fourier transform (STFT) of signals [27]–[29], including for Ambisonics signals in [29]. Different architectures have been tested: feed-forward neural networks [24], convolutional neural networks (CNNs) [27], [30], deep residual networks [31], convolutional and recurrent networks (CRNNs) [29]. Yet, most of these methods have only been evaluated in simulated environments similar to the training conditions, which is not sufficient to verify their generalization to real-life applications.

In addition to using varied and realistic test data, one can verify the ability of neural networks to generalize by analyzing their behavior, which is rarely done in audio scene analysis despite the increasing use of deep learning. Layerwise relevance propagation (LRP) [32] is a visualization technique which highlights the input features that are relevant for a given output. It can bring new information on the input features, for example in binaural localization where it has been used to identify the relevant elevation cues for a neural network, which can then be compared to human localization [33]. In addition, it is of paramount importance to check that the performance of the network is based on robust reasoning and not, for example, a bias in the dataset, which is made possible by LRP [34].

In this article, we present a neural network based DoA estimation system for multi-source Ambisonics recordings. We consider a normalized expression of the acoustic intensity vector in each time-frequency bin and propose to use its coefficients as input features. We conduct an extensive experimental evaluation for up to two sources in several real and simulated environments, including real-life recordings in reverberant rooms with strong early reflections and background noise. We also analyze the inner working of our neural network with LRP

in order to identify the relevant features on which it relies and compare with those used by classical signal processing methods. This work extends our preliminary study [35], which was limited to a single source in a simulated environment and did not include the analysis by LRP.

Section II provides prerequisites on the Ambisonics format and defines the notations. In Section III, we present our DoA estimation system. The neural network which constitutes the core of the system is analyzed by LRP in Section IV. Section V describes the general experimental settings. Several DoA estimation experiments are then presented and analyzed in Section VI. We conclude in Section VII.

II. Background

A. Ambisonics format

The Ambisonics format relies on the spatial decomposition of the sound field in the orthogonal basis of spherical harmonic functions. In practice, the sound field is recorded by a spherical microphone array and converted into Ambisonics with an encoding matrix.

For the representation of the sound field to be exact, the infinite spherical harmonic basis should be used. However, real-life applications require to use a finite microphone array and to handle a limited amount of channels. As has already been done [9], [29], we consider first-order Ambisonics (FOA) only. It corresponds to the coefficients of the decomposition in the spherical harmonics of order 0 (channel W) and 1 (channels X, Y, and Z). These channels already contain precise spatial information: they can been seen as the recordings obtained from a virtual omnidirectional microphone W and three virtual polarized bidirectional microphones X, Y, and Z, all four being coincident in space [5, Fig. 1]. In an anechoic environment, the azimuth θ and elevation ϕ of a source emitting a plane wave can directly be recovered from the FOA steering vector[1], which appears in the STFT expression of the FOA channels as a function of the sound pressure at the recording point $p(t,f)$ [37]:

$$\begin{bmatrix} W(t,f) \\ X(t,f) \\ Y(t,f) \\ Z(t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{3}\cos\theta\cos\phi \\ \sqrt{3}\sin\theta\cos\phi \\ \sqrt{3}\sin\phi \end{bmatrix} p(t,f). \tag{1}$$

For more complex sound fields involving multiple sources or reverberation, FOA recordings cannot be expressed as a simple function of the impinging sound waves anymore. Advanced techniques hence need to be used to recover the spatial information they contain.

B. Acoustic intensity

Sound fields can be described by various physical quantities. In particular, the active intensity vector

$$\mathbf{I}_{\mathrm{a}}(t,f) = \mathcal{R}\{p(t,f)\mathbf{v}^*(t,f)\} \tag{2}$$

represents the flow of sound energy in a point of space [38], with $\mathbf{v}(t,f)$ the particle velocity. This vector is intrinsic to the sound field, but can be expressed in a simple manner in the Ambisonics formalism. In this framework, the particle velocity of a plane wave is [37]:

$$\mathbf{v}(t,f) = -\frac{1}{\rho_0 c\sqrt{3}} \begin{bmatrix} X(t,f) \\ Y(t,f) \\ Z(t,f) \end{bmatrix} \tag{3}$$

with ρ_0 the density of air. Noting that $p(t,f) = W(t,f)$ and disregarding the constant, we express the active intensity vector as:

$$\mathbf{I}_{\mathrm{a}}(t,f) = -\begin{bmatrix} \mathcal{R}\{W(t,f)X^*(t,f)\} \\ \mathcal{R}\{W(t,f)Y^*(t,f)\} \\ \mathcal{R}\{W(t,f)Z^*(t,f)\} \end{bmatrix}. \tag{4}$$

The reactive intensity is defined as the imaginary counterpart of the active intensity: $\mathbf{I}_{\mathrm{r}}(t,f) = \mathcal{I}\{p(t,f)\mathbf{v}^*(t,f)\}$. It represents dissipative local energy transfers. For FOA contents, it is formulated as:

$$\mathbf{I}_{\mathrm{r}}(t,f) = -\begin{bmatrix} \mathcal{I}\{W(t,f)X^*(t,f)\} \\ \mathcal{I}\{W(t,f)Y^*(t,f)\} \\ \mathcal{I}\{W(t,f)Z^*(t,f)\} \end{bmatrix}. \tag{5}$$

In theory, the sound DoA can be estimated as the opposite direction of the active intensity vector [8]. In practice, however, the estimates obtained across all time-frequency bins are inconsistent in reverberant environments [37].

III. DoA estimation system

In order to deal with noise and reverberation, we propose a neural network based method using appropriate input features. Below we describe the input features, the training targets, and the network architecture.

A. Input features

We propose to exploit both the active and reactive intensity vectors across all frequency bins in the STFT domain as inputs to the neural network in a given time frame. This choice differs from the use of the raw FOA channels in [29]. It is motivated by the fact that the active intensity relates more directly to the DoA and the reactive intensity indicates whether a given time-frequency bin is dominated by direct sound from a single source, as opposed to overlapping sources or reverberation.

To ensure that the inputs remain in a fixed range regardless of the sound power, we normalize them in each time-frequency bin [39] by

$$C(t,f) = |W(t,f)|^2 + \frac{1}{3}(|X(t,f)|^2 + |Y(t,f)|^2 + |Z(t,f)|^2). \tag{6}$$

This results in the following 6-channel input features:

$$\frac{-1}{C(t,f)} \begin{bmatrix} \mathbf{I}_{\mathrm{a}}(t,f) \\ \mathbf{I}_{\mathrm{r}}(t,f) \end{bmatrix}. \tag{7}$$

[1] We use the N3D normalization [36].

B. Target outputs and training cost

We define multiple DoA estimation as the task of estimating whether each DoA on a predefined grid corresponds to the direction of an active source or not. We use a quasi-uniform grid on the 2D (azimuth and elevation) sphere, leading to the following equations for the elevations $\phi_i \in [-90, 90]$ and the azimuths $\theta_j^i \in [-180, 180)$ in degrees:

$$\begin{cases} \phi_i &= -90 + \frac{i}{I} \times 180 \qquad \text{with } i \in \{0, \dots, I\} \\ \theta_j^i &= -180 + \frac{j}{J^i+1} \times 360 \text{ with } j \in \{0, \dots, J^i\}, \end{cases} \tag{8}$$

where $I = \lfloor \frac{180}{\alpha} \rfloor$ and $J^i = \lfloor \frac{360}{\alpha} \cos \phi_i \rfloor$ with α the desired grid resolution in degrees.

The resulting grid contains $n_{DoA} = \sum_{i=0}^{I} \sum_{j=0}^{J^i} j$ points. The target of the CRNN is a binary vector of size $n_{DoA} \times 1$, where each index corresponds to one discrete DoA. For each source in the scene, no matter its power, the element of the target vector that is the closest to the true DoA is set to 1. When several sources are present, more than one element can be set to 1. All other target outputs are set to 0.

We assume that the number of sources is known and we train a specific neural network for each number. We define the training cost as the sum of the binary cross-entropies over all outputs. Note that this does not enforce the sum of the network outputs to be equal to the assumed number of sources. Indeed, we did not find this constraint to bring any benefit.

C. Network architecture

The neural network follows the convolutional recurrent neural network (CRNN) architecture in Fig. 1, which is simpler than the one in [29] and was found to perform better [35]. We also tried to use purely convolutional or recurrent networks, without success: our best CRNN classified correctly 1.5 times more examples than our best CNN on a test set made with simulated spatial room impulse responses (SRIRs).

The first part of the CRNN aims to extract spatial information from the inputs. It consists of three convolutional modules made of a two-dimensional convolutional layer followed by batch normalization [40] and max-pooling (down-sampling by taking the maximum of small regions defined by a sliding window) along frequency. The second part uses this information to estimate the DoAs. It comprises two bidirectional long short-term memory (BiLSTM) layers and two time-distributed fully-connected feedforward (FF) layers.

D. From framewise to global DoA estimation

In the following, the sources are assumed to be static over the duration of the test signal. Therefore, the target DoAs are identical for all time frames. Yet, the network is designed to return DoA scores $\sigma_t(\theta_j^i, \phi_i)$ in each frame. We derive a single DoA estimate for the whole sequence as follows. We first average the network outputs over all frames of the test signal to obtain a global score $\sigma(\theta_j^i, \phi_i)$ for each point on the grid [13]. This global score is then smoothed by averaging with neighboring points within a certain angular distance Δ:

$$\bar{\sigma}(\theta_j^i, \phi_i) = \frac{\sum_{i'j'} w_{iji'j'} \sigma(\theta_{j'}^{i'}, \phi_{i'})}{\sum_{i'j'} w_{iji'j'}}. \tag{9}$$

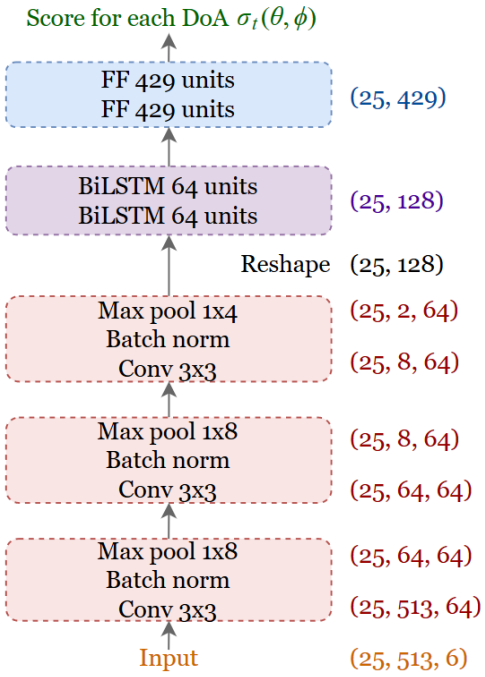


Fig. 1. Architecture of the DoA estimation network.

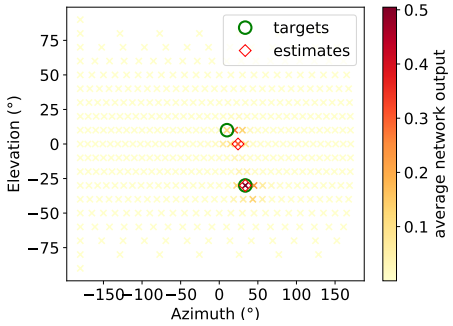


Fig. 2. Example results in a two-source scenario. The crosses represent points on the grid, and the color of each cross encodes the corresponding network output averaged over time. The estimated DoAs are marked by red diamonds, and the true DoAs by green circles.

The weights

$$w_{iji'j'} = \max\left\{0, 1 - \frac{\delta[(\theta_j^i, \phi_i), (\theta_{j'}^{i'}, \phi_{i'})]}{\Delta}\right\}. \tag{10}$$

decay linearly with the angular distance δ, which can be computed via the following formula:

$$\begin{aligned} \delta[(\hat{\theta}, \hat{\phi}), (\theta, \phi)] = \arccos\{ &\sin(\hat{\phi}) \sin(\phi) \\ &+ \cos(\hat{\phi}) \cos(\phi) \cos(\hat{\theta} - \theta)\}. \end{aligned} \tag{11}$$

The estimated DoAs are obtained by picking the largest peaks of the smoothed score. The smoothing step ensures that the peaks are not too close to each other. Figure 2 illustrates the global scores obtained in a two-source scenario, as well as the corresponding estimated and true DoAs.

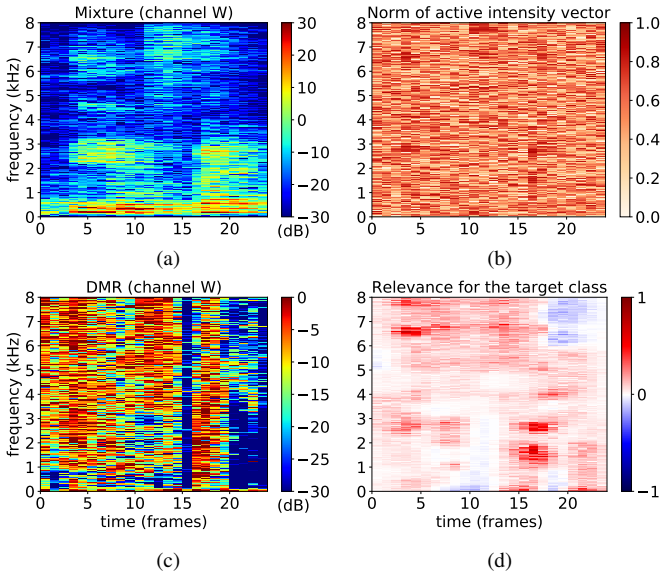


Fig. 3. LRP in the case of an accurate single-source DoA estimation with high SNR. The mixture signal consists of one speech source impinging from $(\theta, \phi) = (139°, -61°)$ and diffuse noise, with $\text{RT}_{60} = 772$ ms and SNR $\approx$ 18 dB. (a) Spectrogram of the mixture at the omnidirectional channel W, (b) norm of the active intensity vector, (c) direct-to-mixture ratio (DMR) at channel W, and (d) relevance map for the accurately estimated class.

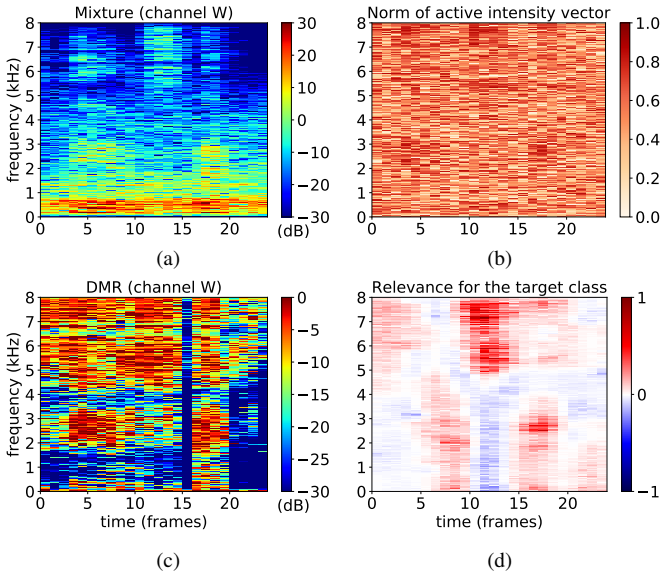


Fig. 4. LRP in the case of an accurate single-source DoA estimation with low SNR. The mixture signal consists of one speech source impinging from $(\theta, \phi) = (41°, -43°)$ and diffuse noise, with $\text{RT}_{60} = 295$ ms and SNR $\approx$ 1 dB. (a) Spectrogram of the mixture at the omnidirectional channel W, (b) norm of the active intensity vector, (c) DMR at channel W, and (d) relevance map for the accurately estimated class.

IV. Analysis by LRP

A. Presentation of the technique

We analyze the inner working of our CRNN using LRP. LRP [41] is a visualization technique that allows for the explanation of a given neural network output via a relevance map indicating which inputs are relevant for that output. LRP has been popularized in the context of image classification, where it has enabled researchers to acquire insight, uncover flaws, and bring specific improvements in data or network design (see [41] for examples).

LRP is based on propagation rules which reportedly provide a better explanation than gradient-based techniques such as sensitivity analysis [42]. The relevance in the last layer is set as the neural network output for the class of interest and to 0 for the other classes. It is then backpropagated down to the input layer. The propagation rules are designed so as to satisfy a layerwise conservation property: the sum of the relevances for all neurons is constant in all layers of the network.

Let us consider the toy case of two successive FF linear layers. The activations in the upper layer are given by $z_j = \sum_i w_{ij} z_i + b_j$, with z_i the activations in the lower layer, w_{ij} the neuron weights, and b_j the biases. The relevance R_j at z_j is distributed on all the lower layer neurons z_i with different shares $R_{i \leftarrow j}$ (different formulas for $R_{i \leftarrow j}$ are discussed below). A lower layer neuron z_i receives relevance shares from all upper layer neurons it is connected to:

$$R_i = \sum_j R_{i \leftarrow j}. \tag{12}$$

The conservation property imposes that the shares coming from an upper layer neuron sum to the relevance at this neuron:

$$\sum_j R_{i \leftarrow j} = R_j. \tag{13}$$

LRP aims to highlight the paths where information flows through the network in order to backpropagate relevance until the significant inputs. This can be achieved in different manners, involving both the weights and the activations of the network after a forward pass. The most used back propagation rules are the ϵ-rule, that ensures stability at the expense of a bending of the conservation property, and the $\alpha\beta$-rule, which is conservative, stable, and treats separately the negative and positive activations [32]. Activation functions, even nonlinear, can be disregarded as long as they are monotonically increasing.

These rules were first designed for FF layers but remain applicable to convolutional and pooling layers [32]. An adaptation to LSTM layers has also been proposed to deal with the gating mechanism [43]. The relevance is fully backpropagated through the signal channels, and set to 0 for the gates. It may seem that the gating factors are then disregarded, but they are in fact already taken into account by their impact on the activation of the LSTM cell, and hence the relevance at the output of the LSTM layer.

B. Settings

In the following, we use the $\alpha\beta$-rule with $\alpha = 1$ and $\beta = 0$ for the BiLSTM and convolutive layers. When used on fully-connected layers, this rule tends to alter the backpropagation of the relevance. We hence use the ϵ-rule for fully-connected FF layers. The parameter ϵ was set to 0.1, as this value was found to stabilize the backpropagation with almost no relevance leak. Furthermore, the $\alpha\beta$-rule was shown to be more stable when the biases of the neurons were forced to be negative [41], which we did with little impact on the network's performance.

We adapt the use of LRP to the context of DoA estimation as follows. For a given estimated DoA, we set the output relevance in each time frame to the corresponding network output $\sigma_t(\theta, \phi)$ for that class and to 0 for the other classes. The relevance is backpropagated separately in each time frame. We then sum the relevances over time and across all 6 channels to obtain a single time-frequency map. Finally, we normalize this map between -1 and 1 for visualization purposes. A positive (resp. negative) relevance in a given time-frequency bin indicates that the features in this time-frequency bin argued in favor of (resp. against) the estimated DoA.

C. Application to DoA estimation

So far, no metric exists for the quantitative analysis of LRP results. In the following, we visualize and seek to interpret the relevance maps obtained for the networks trained to return either one or two DoAs. We consider various signal-to-noise ratios (SNRs) and reverberation times RT_{60}, the time needed for the reverberation to decrease by 60 dB. We also investigate cases when the network returns a wrong estimate. All observations are made on signals generated with simulated SRIRs; see Sections V and VI for details.

1) Single-source DoA estimation: To facilitate comparison, we use the same raw speech signal convolved with different SRIRs in all cases.

Figure 3 presents the case of a single speech source with low noise but strong reverberation. In this case, the network estimates the correct DoA. We compare the relevance map with two other quantities: the norm of the active intensity vector normalized as in (7) and the direct-to-mixture ratio (DMR), defined as the ratio between the power of direct sound (obtained by convolving the raw speech signal with the SRIR truncated after the first main peak) and that of the whole mixture. We notice that the time-frequency bins corresponding to large values of both the intensity vector and the DMR (for instance around frame 17) are particularly used by the network for DoA estimation. The importance of the time-frequency bins corresponding to direct sound has already been used in machine localization by means of various cues, such as estimations of the SNR [13], [15], [44], sound-to-echo ratio [45], or interaural coherence [46]. Interestingly, these bins also correspond to the sound onsets, which is in accordance with psychoacoustic studies showing that onsets are particularly important for DoA estimation by humans. This is known as the precedence effect [47].

Figure 4 presents the same sentence convolved with another SRIR, with lower reverberation but stronger noise. The network still estimates the correct DoA. The observation of channel W, the norm of the intensity vector, and the DMR shows that low frequencies are strongly corrupted by noise. It hence seems natural that the network mostly uses high-frequency features to estimate the DoA, as shown by the relevance map. Nevertheless, the relevance map does not perfectly correlate with either of these simple quantities. This suggests that the network may learn more subtle cues.

In Figure 5, we examine one of the few cases when the network does not estimate the correct DoA and the score

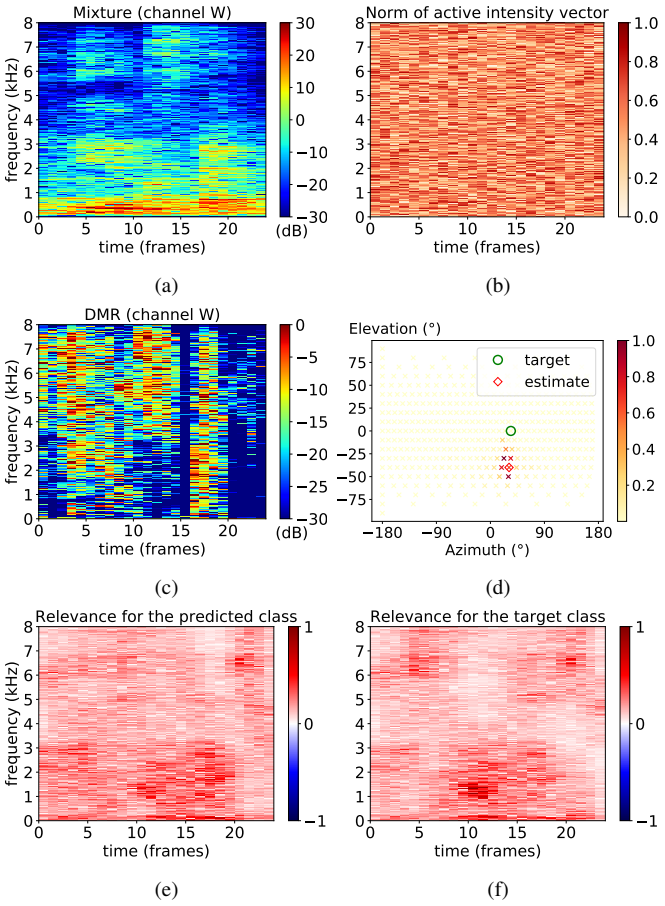


Fig. 5. LRP in the case of an inaccurate single-source DoA estimation. The mixture signal consists of one speech source impinging from $(\theta, \phi) = (29°, -2°)$ and diffuse noise, with $\mathrm{RT}_{60} = 661$ ms and SNR ≈ 5 dB. (a) Spectrogram of the mixture at the omnidirectional channel W, (b) active intensity for channel X, (c) DMR at channel W, (d) network outputs averaged over time, (e) relevance map for the wrongly estimated DoA, (f) relevance map for the true DoA.

returned for the true DoA is very low. Here, the SNR is 5 dB and the RT_{60} is 661 ms. The inputs show that the acoustic intensity is particularly noisy. The relevance backpropagated from the wrongly estimated DoA barely highlights any specific area. The same is observed on the relevance backpropagated from the true DoA. In this case, it appears that the acoustic intensity vector is so noisy that the network is unable to identify the time-frequency bins relevant for DoA estimation. This might be exploited in the future to quantify the confidence in the estimated DoA, which is not well predicted by the network output values.

2) Two-source DoA estimation: Figure 6 illustrates the LRP analysis for the two-source network. The relevance maps for the two sources are similar in many areas, for example between 1 and 2 kHz around frame 5, although this is a high DMR area for the second source but not for the first. This seems to indicate that the network exploits information from all sources at the same time, rather than focusing on the time-frequency bins dominated by a given source. A major difference between the two relevance maps is still obvious in frame 14 at medium frequencies, where source 1 dominates

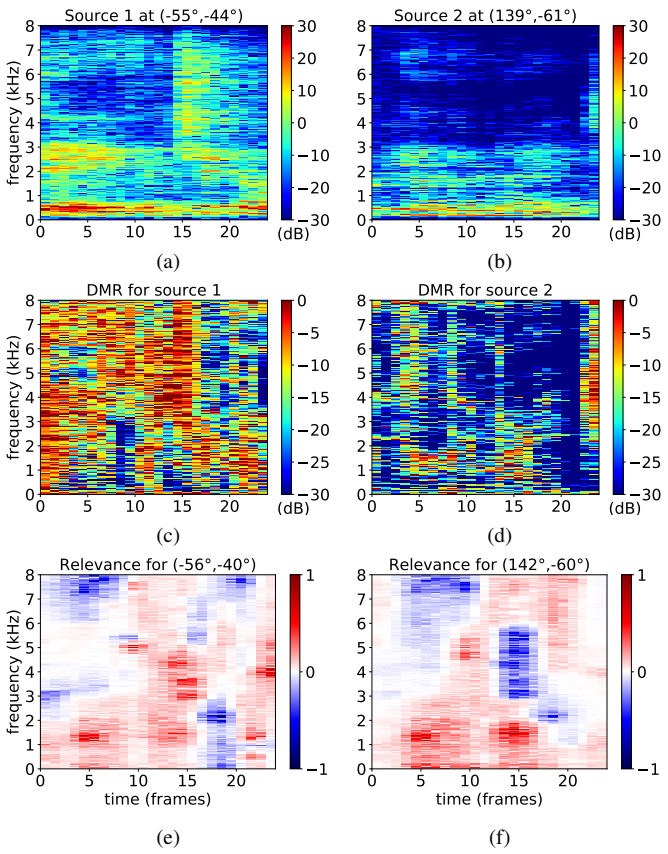


Fig. 6. LRP in the case of an accurate two-source DoA estimation. The first source impinges from $(\theta_1, \phi_1) = (-56°, -40°)$ and the second from $(\theta_2, \phi_2) = (142°, -60°)$. It is approximately 8 dB less loud than the first source. Diffuse noise is present at 20 dB SNR and the room has a $\text{RT}_{60} =$ 772 ms. (a) Spectrogram of source 1 at channel W, (b) spectrogram of source 2 at channel W, (c) DMR for source 1 at channel W, (d) DMR for source 2 at channel W, (e) relevance map for source 1, (f) relevance map for source 2.

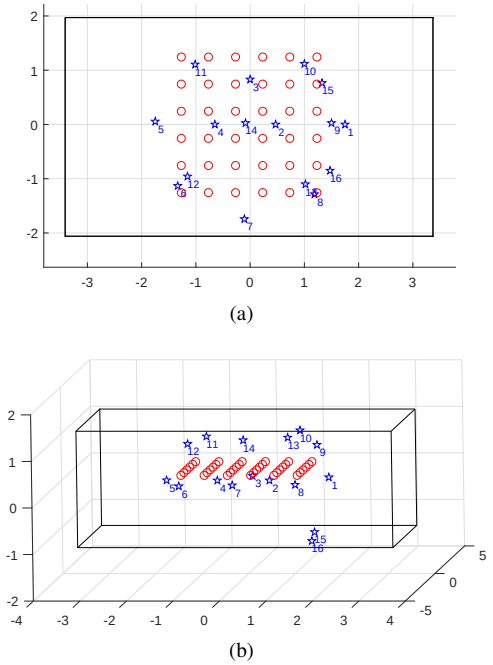


Fig. 7. Recording configuration for the real SRIRs seen (a) from the top and (b) from the side. Loudspeaker positions are denoted by blue stars, and Eigenmike positions by red circles. The loudspeakers all point toward the center of the median plan, except loudspeaker number 4 which points toward the bottom wall of (a). Dimensions are shown in meters.

and source 2 is absent. Accordingly, the corresponding time-frequency bins are positive in the relevance map for the first DoA, and negative in the relevance map for the second DoA.

V. Shared Experimental Settings for DoA Estimation

We evaluated our DoA estimation system in various conditions. In this section, we present the SRIRs and the recordings that were used to generate training or test data and the neural network training procedure. The experiments themselves are described in Section VI.

A. SRIRs and audio recordings

1) Simulated SRIRs: To train the networks, as well as for some test datasets, we synthesized a large number of SRIRs with the image method [48] by adapting the code from Habets [49] to ideal FOA recordings (1). We generated a large number of rooms with random dimensions (length and width between 2.5 and 10 m, height between 2 and 3 m), random RT_{60} between 0.2 and 0.8 s, random microphone array positions and random source-to-microphone distances between 1 and 3 m. To be able to generate several mixtures for each configuration, we synthesized three SRIRs in each room, corresponding to different DoAs. The first of these DoAs could be enforced to follow a certain distribution, which will be specified in the description of the training and test sets. The two other DoAs were picked randomly as long as the corresponding sources were located inside the room, and with at least 10° between all pairs of DoAs.

2) Real SRIRs: In order to perform more realistic experiments, we also generated test signals with real SRIRs recorded with an Eigenmike microphone array [50]. The room had a medium $\text{RT}_{60} \approx 500$ ms. The array was positioned in 36 different locations while 16 loudspeakers emitted sweep signals one after another resulting in 576 SRIRs. The configuration is displayed in Fig. 7. Note that this protocol resulted in situations when the microphone was actually behind the loudspeaker, which cannot happen with synthesized SRIRs where each source has been set as omnidirectional.

3) Real recordings: To validate our system on real data, we recorded speech with an Eigenmike in a living room (see Fig. 8). The microphone was placed just above a coffee table, resulting in very strong early reflections. Three different French speakers read excerpts of "Le Petit Prince" in fixed positions, although some head movements were hardly avoidable. The speakers were located in 14 positions in total around the table, either sitting on a couch or standing. For each position, the speaker read for approximately 5 minutes, resulting in a combined total of 71 minutes of recording. A TV was also separately recorded in the same room while playing various contents (TV shows and advertising containing speech and sound effects, as well as music contents). In all recordings, real ambient noises were present at specific moments, for example

Fig. 8. Layout of the living room for the real recordings. The Eigenmike used for the recordings is circled in red.

footsteps, page turning, or a lawnmower coming from the outside. When those noises were present, the SNR was around 5 to 10 dB.

4) Audio settings: All audio signals and SRIRs were sampled at 16 kHz. The STFT was performed on 1024 sample frames with 50% overlap, resulting in a shift of 32 ms between two frames. We used a sine analysis and synthesis window.

B. Training procedure

1) Network settings: A single-source network and a two-source network were constructed in the same way (see Fig. 1). Both could be used to predict any number of sources, but training each network for a specific number of sources yielded better results. The input shape was (25, 513, 6), where 25 was the number of frames, 513 the number of frequency bins and 6 the number of feature channels. For the convolution layers we used 64 filters of size 3×3. Max-pooling was performed along frequency only, over 8 frequency bins in the first two layers and 4 bins in the last layer. The three convolutional layers were followed by two BiLSTM layers with 64 hidden units and two time-distributed fully-connected FF layers with 429 neurons. Rectified linear unit (ReLU) activations were used after each convolutional layer and after the first FF layer. Hard-sigmoid and tanh were used as recurrent and kernel activations in the BiLSTM layers. A sigmoid was applied after the second FF layer so that the outputs were between 0 and 1. We sampled the sphere with an angular resolution of $\alpha = 10^\circ$ in (8), resulting in $n_{DoA} = 429$ output classes. For peak detection, the threshold defining the neighborhood of a point was set to $\Delta = 2\alpha$.

2) Training and validation sets: The networks were trained on signals generated from a SRIR dataset synthesized as explained in Section V-A1 resulting in a total of 128,700 SRIRs in 42,900 room configurations. Each SRIR was convolved with a different 1 s speech signal randomly extracted from a subset of the Bref corpus [51]. The subset contained over 5 h of French speech from 44 speakers. The total duration of the training set is 36 h. Each 1 s signal was split in two sequences of 25 frames with 12 overlapping frames between sequences (the last 6 frames of the second sequence were padded with zeros).

We enforced the first DoAs in all rooms to lie in the neighborhood of every DoA on the grid, so that each target DoA has been seen a significant number of times. The single-source network was trained on single speech signals generated with all the SRIRs of the dataset, along with diffuse babble noise at a random SNR between 0 and 20 dB. The two-source network was also trained on speech signals generated with all SRIRs of the dataset, to which was added another speech signal generated from a SRIR synthesized in the same room configuration. Eventually, this lead to 127,800 signals of 1 s for both the one-source and the two-source networks. In the two-source case, there was at least 10° angular distance between the sources and a random signal-to-interference ratio (SIR) between 0 and 10 dB. A diffuse babble noise was added at a 20 dB SNR. The babble noises were randomly picked from Freesound[2] and the diffuse field was simulated by averaging the diffuse parts of two SRIRs for a unique room configuration. The validation sets were generated similarly, with 1,287 signals synthesized in different rooms and with different speakers from Bref (among 2 h of speech by 17 speakers).

3) Training settings: We used the Nadam optimizer [52] with an initial training rate of 10^{-3}. We applied dropout after each convolutional block, each FF layer and on the recurrent weights of the BiLSTM layers. The dropout rate was set to 0.2 for the single-source network and 0.3 for the two-source network. Overfitting was also prevented by early stopping with a patience of 20 epochs measured on the validation set, resulting in a maximum number of 80 and 150 epochs for the single-source network and the two-source network, respectively.

VI. Experimental evaluation

A. Baselines

As a baseline for comparison, we used an algorithm based on a histogram analysis of the active acoustic intensity vectors in each time-frequency bins, made more robust by taking into account the estimated SNR and diffuseness in each bin [44]. It uses the same discretization of the unit sphere (8) as our CRNN. This algorithm participated in the 2018 LOCATA challenge [53], where it largely outperformed the MUSIC baseline on real contents. In the following, this baseline will be referred to as the histogram baseline.

Another neural network based DoA estimation system for FOA signals was recently proposed in [29] that directly takes the magnitudes and phases of the FOA signals (1) as inputs, resulting in 8 feature channels. In order to evaluate the added value of our input features (7), we also trained and tested our network (Fig. 1) on our training data with these 8-channel FOA features instead. In the remainder of the paper the latter system is referred to as FOA-CRNN while the proposed system is referred to as Intensity-CRNN.

B. Performance measurement

We evaluated the DoA estimation performance in terms of sequence-wise accuracy, that is the percentage of 25-frame

[2]http://freesound.org/

Room (Section)	Simulated SRIR (V-A1)				Real SRIR (V-A2)				Real recordings (V-A3)			
Angular error (11)	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.
Histogram baseline [44]	15.9	45.9	68.7	19.5	20.8	49.2	67.2	25.8	11.0	36.1	58.4	20.3
FOA-CRNN	48.8	88.8	96.7	54.7	23.9	66.0	87.0	29.7	9.1	39.4	69.5	22.9
Intensity-CRNN (proposed)	**54.3**	**94.4**	**98.9**	**60.9**	**28.6**	**70.2**	**89.6**	**34.1**	**23.4**	**73.7**	**89.5**	**41.2**

TABLE I
SEQUENCE-WISE ACCURACIES AND CLASSIFICATION ACCURACIES (%) FOR SINGLE-SOURCE DOA ESTIMATION WITH 5°, 10° OR 15° ANGULAR ERROR TOLERANCE. THE 95% CONFIDENCE INTERVALS VARY FROM ± 0.4% TO ± 2.9%. THE BEST RESULTS ARE SHOWN IN BOLD.

sequences (and sources in the two-source case) whose DoA is correctly estimated within a certain angular error tolerance (11). We considered 5°, 10°, and 15° tolerances. Note that, due to the chosen grid resolution, the angular distance (11) between a point on the sphere and the closest point on the grid can be up to 7°. This implies that certain signals may be correctly classified, but considered as incorrect according to the 5° tolerance. To account for this issue, we also report the classification accuracy, i.e., the percentage of cases where the estimated DoA is the point on the grid that is closest to the true DoA. In the two-source case, each estimated DoA needs to be associated with the corresponding target in order to compute the accuracy. We chose the permutation that minimizes the sum of angular errors, according to the Hungarian algorithm [54].

C. Single-source DoA estimation

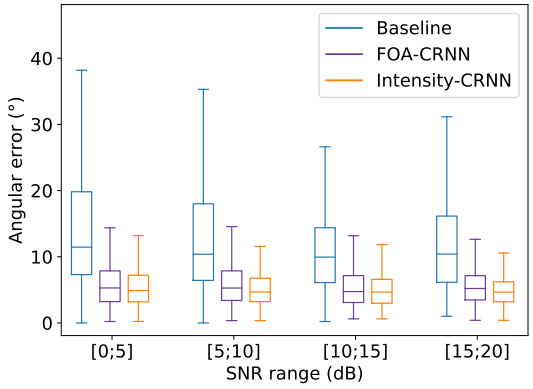
We first evaluated the DoA estimation performance of the network trained to return a single DoA.

1) Test sets: The evaluation was conducted on three test sets. The first test set was generated from 1,287 simulated SRIRs (see Section V-A1) corresponding to 429 unseen room and microphone configurations, with acoustic parameters drawn randomly within the same ranges as for the training SRIRs. The DoAs of the sources were randomly and uniformly drawn on the sphere irrespective of the grid. Each SRIR was convolved with an 1 s English speech signal randomly selected from the SiSEC campaign [55] and corrupted with diffuse babble noise at a random SNR between 0 and 20 dB, resulting in 1,287 mixtures. The babble noise was different from those used for training or validation. Each mixture was split into two overlapping 25-frame sequences, resulting in 2,574 test sequences. The second dataset was created similarly using the 576 real SRIRs from Section V-A2, resulting in 1,152 test sequences. The third test set was obtained by splitting the real recordings in Section V-A3 into 1 s excerpts. A voice activity detection (VAD) was applied to keep only excerpts where speech was present in the two sequences constituting the excerpts. A TV-only recording was added to each 1 s speech-only recording at an SNR of 20 dB. The interfering TV sound was then much more directive than the diffuse babble noise in the first two sets. In total, 5,210 real test sequences were created.

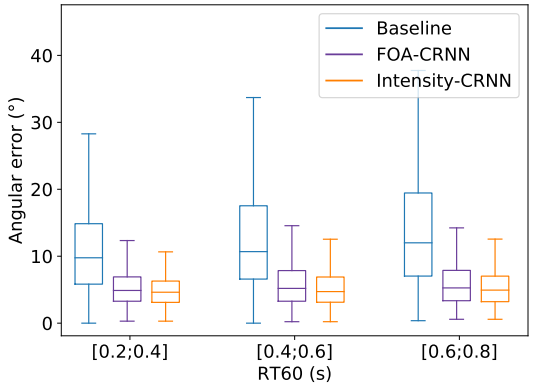
2) Results: Table I summarizes the results for the single-source DoA estimation task. On the signals generated with simulated SRIRs, the two CRNN-based algorithms largely outperform the histogram baseline [44], as they were trained

Azimuth and elevation errors	<5°		<10°		<15°	
	θ	ϕ	θ	ϕ	θ	ϕ
Histogram baseline [44]	67.6	36.3	89.2	51.6	91.9	62.4
FOA-CRNN	65.8	36.4	**92.1**	55.3	**95.0**	75.5
Intensity-CRNN (proposed)	**72.9**	**63.8**	**92.5**	**83.6**	94.2	**92.6**

TABLE II
SEQUENCE-WISE AZIMUTH AND ELEVATION ACCURACIES (%) FOR SINGLE-SOURCE DOA ESTIMATION ON REAL RECORDINGS. THE 95% CONFIDENCE INTERVALS VARY FROM 0.6 TO 1.4%.



(a)



(b)

Fig. 9. Boxplots of the angular errors (11) in degrees for all sequences of the single-source DoA estimation as a function of (a) the SNR and (b) the RT_{60}, on signals generated from simulated SRIRs. The boxes show the first and third quartiles as well as the median. The lower ends of the whiskers correspond to the lowest data still within 1.5 interquartile range (IQR) of the lower quartile, and similarly the higher ends of the whiskers correspond to the highest data within 1.5 IQR of the upper quartile.

Room (Section)	Simulated SRIR (V-A1)				Real SRIR (V-A2)				Real recordings (V-A3)			
Angular error (11)	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.
Histogram baseline [44]	12.6	34.8	52.1	15.1	15.8	40.7	55.0	19.5	6.8	29.3	47.5	14.4
FOA-CRNN	37.2	75.1	87.4	42.4	22.3	58.6	75.7	26.4	6.1	33.4	52.9	10.1
Intensity-CRNN (proposed)	**41.8**	**80.9**	**89.1**	**47.2**	**26.2**	**62.6**	**78.1**	**31.0**	**15.2**	**56.0**	**74.9**	**23.9**

TABLE III
SEQUENCE-WISE ACCURACIES AND CLASSIFICATION ACCURACIES (%) FOR TWO-SOURCE DOA ESTIMATION WITH 5°, 10° OR 15° ANGULAR ERROR TOLERANCE. THE 95% CONFIDENCE INTERVALS VARY FROM ± 0.5% TO ± 2%. THE BEST RESULTS ARE SHOWN IN BOLD.

and tested in matched conditions. The proposed Intensity-CRNN outperform the FOA-CRNN and allows a 11% relative improvement in terms of classification accuracy.

With real SRIRs, CRNNs still perform better than the histogram baseline [44]. In particular, they are much less prone to outliers, with more than 87% of the sequences whose DoA is estimated within less than 15° error, compared to 67.2% only for the histogram baseline. Once again, the Intensity-CRNN is more precise than the FOA-CRNN baseline, with a 15% relative improvement on classification accuracy.

Finally, on real recordings, the Intensity-CRNN largely outperforms both baselines. It proves to be much more robust to the reflections on the coffee table impinging on the microphones right after the direct sound. The accuracy with a 5° tolerance is not very relevant here due to the aforementioned grid resolution issue, which affects 6 positions out of 14. When considering a 15° tolerance, the Intensity-CRNN achieves 89.5% accuracy compared to 69.5% for the FOA-CRNN and 58.4% for the histogram baseline [44].

Table II reports the accuracy in terms of azimuth and elevation separately. This confirms that the collapse of the performance of the baselines can indeed be attributed to the early reflections on the coffee table: the FOA-CRNN performs almost as good as the Intensity-CRNN in terms of azimuth, but significantly worse in terms of elevation. For the 5° tolerance for instance, the FOA-CRNN is 10% relatively worse than the Intensity-CRNN in azimuth but 43% worse in elevation. This was further confirmed by examining the results depending on the speaker position (not shown in the Table). The microphone is located on one side of the table (see Fig. 8). When the speaker is on the same side of the table as the microphone, there are few reflections on the table and the FOA-CRNN performs well. When the speaker is on the opposite side, the network systematically returns an estimated elevation between -5° and 0°, which corresponds to the reflections.

3) Robustness with respect to the SNR and the RT_{60}: Figure 9 illustrates the performance of the three systems as a function of the SNR and the RT_{60}. It can be observed that the CRNNs are quite robust: contrarily to the histogram baseline [44], the distribution of angular errors appears to be mostly independent of the SNR and the RT_{60}.

D. Two-source DoA estimation

1) Test sets: The two-source networks were tested on three datasets, similarly to the single-source networks. In the case of simulated SRIRs, the two SRIRs of a given mixture came from the same room and microphone configuration (unseen during training). In the case of real SRIRs, they were recorded with the same microphone. The two DoAs were chosen such that they are at least 25° apart. Both SRIRs were convolved with two distinct 1 s signals randomly extracted from the SiSEC corpus. The level difference between the two sources varied between 0 and 10 dB, and a diffuse babble noise was added at an SNR of 20 dB. Similarly to the single-source case, this resulted in 2,574 sequences generated from 1,287 simulated SRIRs and 1,152 sequences generated from real SRIRs. For each sequence, the DoAs of two sources must be estimated, which means that the accuracies are computed on a total of 5,148 and 2,304 estimations, respectively. The third test set was obtained by splitting the real speech-only recordings in Section V-A3 into 1 s excerpts with VAD, summing each excerpt with a random excerpt from another speaker at a random SIR between 0 and 10 dB, and further adding TV-only recordings together at an SNR of 20 dB. The sources were never closer than 25° angular distance. 5,210 sequences were created, for a total of 10,420 DoAs to be estimated.

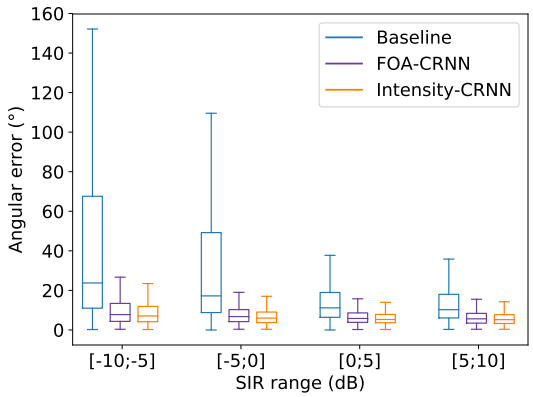


Fig. 10. Boxplots of the angular errors (11) in degrees for all sequences of the two-source DoA estimation as a function of the SIR on signals generated from simulated SRIRs.

2) Results: Table III presents the results for the two-source DoA estimation task. On the simulated SRIR dataset, the same observations as in the single-source task can be made, although the accuracies are globally lower. This is expected since half of the sources have a positive SIR, and the remaining half have a negative SIR, which is a difficult scenario. The CRNNs remain much better than the histogram baseline, and the Intensity-CRNN remains the most effective system.

On the real SRIR dataset, the histogram baseline [44] now performs significantly worse than the CRNNs. It seems unable

to locate multiple sources, especially in scenarios with weak direct sound due to the loudspeaker orientation. The Intensity-CRNN input still outperforms the FOA-CRNN.

Finally, on mixtures of real recordings, we observe once more that the baselines are confused by the first reflections on the table. The Intensity-CRNN is the only model which is able to deal both with multiple sources and early reflections, with 74.9% of the sources located within less than 15° error, compared to 47.5% for the histogram baseline [44] and 52.9% for the FOA-CRNN.

3) Robustness with respect to the SIR: Figure 10 displays the boxplots corresponding to the angular errors of the estimates depending on the SIR. Results are shown for the synthesized SRIRs but were similar for the other test sets. For the CRNN estimates, the quartile and median values only slightly increase for negative SIRs and stay concentrated within 15° of angular error, which shows the robustness of the estimate even when the source is less loud than its competitor. On the contrary, the performance of the histogram baseline [44] drastically drops for negative SIRs.

VII. Conclusion

We introduced a CRNN-based multiple DoA estimation system for FOA signals. Experiments on simulated data as well as in real living-room conditions showed that using the raw FOA channels as inputs fails in the case of real recordings with strong early reflections. We proposed input features based on the acoustic intensity vector that enable the network to properly estimate the source DoAs even in this challenging scenario. Although this was done in the context of Ambisonics recordings, the method could be adapted to any other format allowing the computation of acoustic intensity. We also analyzed and validated the behavior of the network in typical cases by LRP visualization. In the single-source case, the time-frequency bins used by the CRNN tend to correspond to sound onsets, where the norm of the acoustic intensity vector is particularly large. In the two-source case, the CRNN tends to focus on the areas where each source is present with little interference. Nevertheless, the relevance map does not perfectly correlate with simple cues such as the DMR, which suggests that the network may learn more subtle cues. Also, in the case when a wrong estimate is returned, the relevance map does not exhibit any noticeable pattern. This might be exploited in the future to quantify the confidence in the estimated DoA, which is not well predicted by the CRNN output value. In future work, we plan to train a single network to locate any number of sources, which will enable us to perform tracking and deal with appearing and disappearing sources, and to exploit LRP to bring improvements in data or network design.

Acknowledgment

The authors would like to thank S. Kitić for discussions.

References

[1] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[3] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[4] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[5] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multichannel speech separation with recurrent neural networks from high-order Ambisonics recordings," in *Proc. of ICASSP*, 2018, pp. 36–40.

[6] M. A. Gerzon, "Periphony: with-height sound reproduction," *JAES*, vol. 21, no. 1, pp. 2–10, 1973.

[7] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio - The new standard for coding of immersive spatial audio," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 770–779, 2015.

[8] V. Pulkki, "Spatial sound reproduction with directional audio coding," *JAES*, vol. 55, no. 6, pp. 503–516, 2007.

[9] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. of EUSIPCO*, 2010, pp. 442–446.

[10] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. of IWAENC*, 2014, pp. 258–262.

[11] T. E. Tuncer and B. Friedlander, *Classical and modern direction-of-arrival estimation*. Academic Press, 2009.

[12] J. Dibiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 8.

[13] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[14] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo, "Multichannel source activity detection, localization, and tracking," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Wiley, 2018, ch. 4.

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

[16] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. of ICASSP*, vol. 1, 1997, pp. 375–378.

[17] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[18] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[19] O. Nadiri and B. Rafaeli, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 10, pp. 1494–1505, 2014.

[20] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *Proc. of ICASSP*, 2011, pp. 117–120.

[21] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *JAES*, vol. 53, no. 12, pp. 1115–1127, 2005.

[22] S. Tervo, "Direction estimation based on sound intensity vectors," in *Proc. EUSIPCO*, 2009, pp. 700–704.

[23] S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 10, pp. 1956–1968, 2017. [Online]. Available: https://doi.org/10.1109/TASLP.2017.2736067

[24] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. of Interspeech*, 2015, pp. 3302–3306.

[25] X. Xiao *et al.*, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. of ICASSP*, 2015, pp. 2814–2818.

[26] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. of ICASSP*, 2016, pp. 405–409.

[27] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. of WASPAA*, 2017, pp. 136–140.

[28] ——, "Multi-speaker localization using convolutional neural network trained with noise," in *ML4Audio Worskhop at NIPS*, 2017.

[29] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. EUSIPCO*, 2018.
[30] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Trans. Em. Topics Comput. Intell.*, vol. 2, no. 2, pp. 103–116, 2018.
[31] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Proc. Interspeech*, 2018, pp. 312–316.
[32] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015.
[33] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial audio feature discovery with convolutional neural networks," in *Proc. of ICASSP*, 2018, pp. 6797–6801.
[34] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proc. CVPR*, 2016, pp. 2912–2920.
[35] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector," in *Proc. of IWAENC*, 2018, pp. 241–245.
[36] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Thèse de doctorat, Univ. Paris VI, 2000.
[37] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric time-frequency domain spatial audio*. John Wiley & Sons, 2017.
[38] F. Jacobsen, "A note on instantaneous and time-averaged active and reactive sound intensity," *J. of Sound and Vibration*, vol. 147, no. 3, pp. 489–496, 1991.
[39] M. Baqué, "Analyse de scène sonore multi-capteurs," Ph.D. dissertation, Univ. du Maine, 2017.
[40] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. of ICML*, 2015, pp. 448–456.
[41] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.
[42] J. M. Zurada, A. Malinowski, and I. Cloete, "Sensitivity analysis for minimization of input data dimension for feedforward neural network," in *Proc. of ISCAS*, vol. 6, 1994, pp. 447–450.
[43] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *Proc. of WASSA*, 2017, pp. 159–168.
[44] S. Kitić and A. Guérin, "TRAMP: Tracking by a Real-time AMbisonic-based Particle filter," in *LOCATA workshop at IWAENC*, 2018.
[45] J. Huang, N. Ohnishi, and N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, 1997.
[46] C. Faller and J. Merimaa, "Source localization in complex listening situations: selection of binaural cues based on interaural coherence," *JASA*, vol. 116, no. 5, pp. 3075–3089, 2004.
[47] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *JASA*, vol. 106, no. 4, pp. 1633–1654, 1999.
[48] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, 1979.
[49] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., 2006.
[50] M. Acoustics, "EM32 Eigenmike microphone array release notes (v17. 0)," Tech. Rep., 2013. [Online]. Available: www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf
[51] L. F. Lamel, J.-L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," in *Proc. of Eurospeech*, 1991, pp. 505–508.
[52] T. Dozat, "Incorporating Nesterov momentum into Adam," Univ. of Stanford, Tech. Rep., 2015.
[53] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. of SAM*, 2018, pp. 410–414.
[54] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[55] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: a community-based approach to large-scale evaluation," in *Proc. of ICA*, 2009, pp. 734–741.