CRNN-based multiple DoA estimation using Ambisonics acoustic intensity features

Lauréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin

CRNN-based multiple DoA estimation using Ambisonics acoustic intensity features

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 a neural network based system to estimate the direction of arrival of multiple sources from a first-order Ambisonics recording. It is built from stacked convolutional and recurrent layers and 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 or 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]–[13]. 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 [14]. Steered response power (SRP) based methods explore the space with a beamformer, where the areas of higher energy reveal possible source positions [15]. 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], [16], [17]. All these methods provide good DoA estimates for a single source. Some of them are also efficient when facing multiple sources. Nevertheless, it is widely admitted that, in the presence of noise and reverberation, their accuracy decreases dramatically [10], [18].

Recently, neural networks have improved the robustness of DoA estimation techniques in such adverse conditions. They have been used with binaural features [19], GCC features [20], the eigenvectors of the spatial covariance matrix [21], or the cosines and sines of interchannel phase differences [22] as inputs. Convolutional neural networks (CNNs) have also been applied to raw short-time Fourier transform (STFT) phases [23]–[25], including for Ambisonics signals in [25]. 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 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 layerwise relevance propagation (LRP) [26] in order to identify the relevant features on which it relies and compare with those used by classical signal processing methods. To the best of our knowledge, this is the first analysis of this kind for source localization and one of the first times LRP is being used in the field of audio. This work extends our preliminary study [27], 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

Ambisonics rely on the decomposition of the sound field on the orthogonal basis of spherical harmonics. For a point referenced by its spherical coordinates (r, θ, ϕ), and under the

condition that no source is present inside the sphere of radius r, the sound field may be expressed by the equation [28]:

$$p(t,f,r,\theta,\phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm}(t,f) j_n\left(\frac{2\pi f r}{c}\right) Y_{nm}(\theta,\phi), \tag{1}$$

with $p(t,f,r,\theta,\phi)$ the sound pressure at time t, frequency f, distance r, azimuth θ, and elevation ϕ, $j_n(.)$ the spherical Bessel function, c the speed of sound, $Y_{nm}(.,.)$ the spherical harmonic functions, and $p_{nm}(t,f)$ the Ambisonics coefficients of order n and mode m. These coefficients can be computed by integrating over a fixed sphere of radius r:

$$p_{nm}(t,f) = \frac{1}{j_n\left(\frac{2\pi f r}{c}\right)} \iint p(t,f,r,\theta,\phi) Y_{nm}^*(\theta,\phi) \sin\theta d\theta d\phi, \tag{2}$$

where * denotes complex conjugation.

In practice, similarly to [9], [25], we consider first-order Ambisonics (FOA) only. After replacing $Y_{nm}(.,.)$ and $j_n(.)$ by their closed-form expressions, we obtain the Ambisonics coefficients of order $n=0$ (named W) and $n=1$ (named X, Y, and Z).

These coefficients can been seen as the recordings obtained from an omnidirectional microphone W and three polarized bidirectional microphones X, Y, and Z, all four being co-incident in space. Indeed, for a plane wave impinging from azimuth θ and elevation ϕ, they can be expressed as[1] (see also [5, Fig. 1]):

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

For more complex sound fields involving multiple sources or reverberation, this expression does not hold any more.

B. Acoustic intensity

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

$$\mathbf{I}_{\mathrm{a}}(t,f) = \mathcal{R}\{p(t,f)\mathbf{v}^*(t,f)\} \tag{4}$$

represents the flow of sound energy in a point of space [30], with $\mathbf{v}(t,f)$ the particle velocity. The Ambisonics particle velocity of a plane wave is [28]:

$$\mathbf{v}(t,f) = -\frac{1}{\rho_0 c\sqrt{3}} \begin{bmatrix} X(t,f) \\ Y(t,f) \\ Z(t,f) \end{bmatrix} \tag{5}$$

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

$$\mathbf{I}_{\mathrm{a}}(t,f) = \begin{bmatrix} \mathcal{R}\{W(t,f)X^*(t,f)\} \\ \mathcal{R}\{W(t,f)Y^*(t,f)\} \\ \mathcal{R}\{W(t,f)Z^*(t,f)\} \end{bmatrix}. \tag{6}$$

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:

$$\mathbf{I}_{\mathrm{r}}(t,f) = \begin{bmatrix} \mathcal{I}\{W(t,f)X^*(t,f)\} \\ \mathcal{I}\{W(t,f)Y^*(t,f)\} \\ \mathcal{I}\{W(t,f)Z^*(t,f)\} \end{bmatrix}. \tag{7}$$

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 [28].

III. DoA estimation system

In order to deal with noise and reverberation, we propose a neural network based method using appropriate input features. We describe below 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 [25]. 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 intensity, we normalize them by the sound power in each time-frequency bin [31]. This results in the following 6-channel input features:

$$\frac{1}{|W(t,f)|^2 + \frac{1}{3}(|X(t,f)|^2 + |Y(t,f)|^2 + |Z(t,f)|^2)} \begin{bmatrix} \mathbf{I}_{\mathrm{a}}(t,f) \\ \mathbf{I}_{\mathrm{r}}(t,f) \end{bmatrix}. \tag{8}$$

Examples of normalized active and reactive intensities are plotted in Fig. 1. They appear to be very noisy, which makes them difficult to interpret by humans. The DoA information actually mainly derives from the combination of all channels [32]. However, we can already notice that the active intensity seems to be stronger on sound attacks.

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:

$$\begin{cases} \phi_i &= -90 + \frac{i}{I} \times 180 \qquad \text{with } i \in \{0,\dots,I\} \\ \theta_j^i &= -180 + \frac{j}{J^i+1} \times 360 \text{ with } j \in \{0,\dots,J^i\}, \end{cases} \tag{9}$$

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 target DoAs are one-hot encoded: if a source is present, no matter its intensity, the output corresponding to the point

[1] We use the N3D normalization [29].

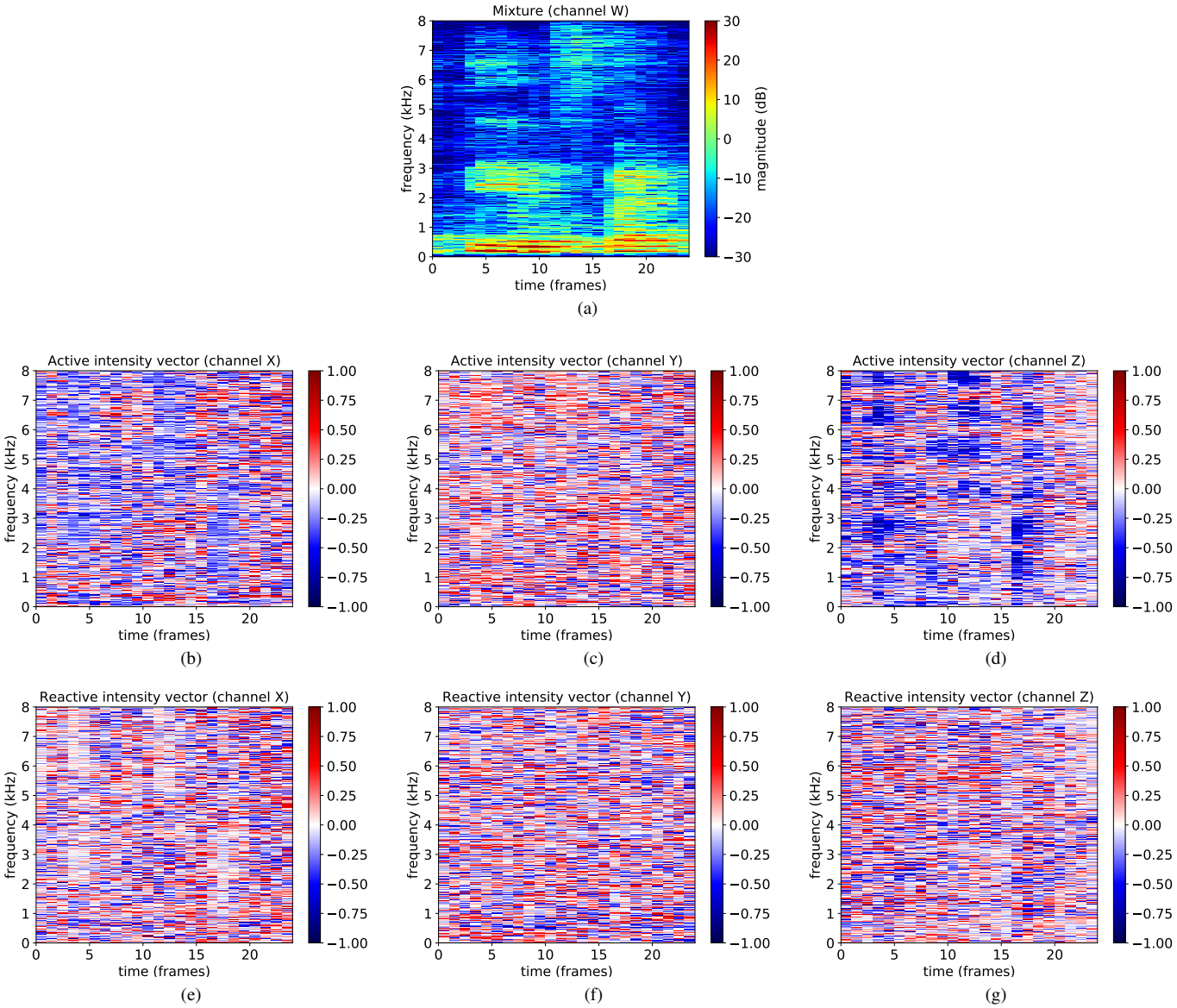


Fig. 1. Normalized active and reactive intensity features for a mixture signal consisting of one speech source impinging from $(\theta, \phi) = (139°, -61°)$ and diffuse noise, with $\text{RT}_{60} = 772$ ms and SNR ≈ 18 dB.

on the grid that is the closest to the true DoA is 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. 2, which is simpler than the one in [25] and was found to perform better [27]. The first part 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 [33] and max-pooling 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 estimates 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 [12]. This global score is then smoothed by averaging with neighboring points within a certain angular distance Δ:

$$\bar{\sigma}(\theta_j^i, \phi_i) = \frac{\sum_{i'j'} w_{iji'j'} \sigma(\theta_{j'}^{i'}, \phi_{i'})}{\sum_{i'j'} w_{iji'j'}}. \tag{10}$$

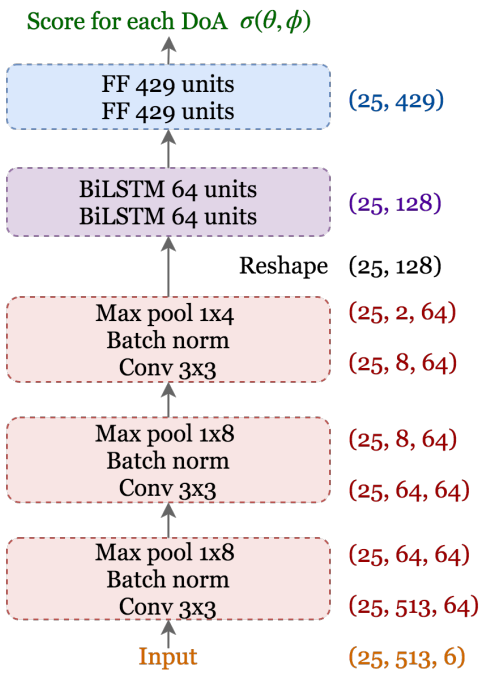


Fig. 2. Architecture of the DoA estimation network.

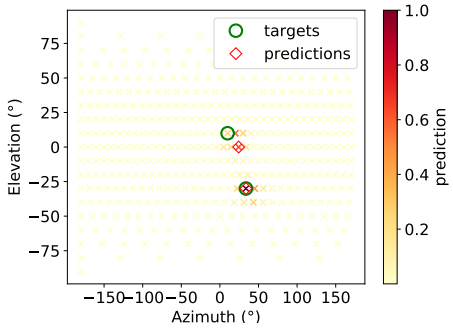


Fig. 3. Example results in a two-source scenario. The crosses represent points on the grid, and the color of each cross encodes the corresponding network output averaged over time. The estimated DoAs are marked by red diamonds, and the true DoAs by green circles.

The weights

$$w_{iji'j'} = \max\left\{0, 1 - \frac{\delta[(\theta^i_j, \phi_i), (\theta^{i'}_{j'}, \phi_{i'})]}{\Delta}\right\}. \tag{11}$$

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

$$\begin{aligned}\delta[(\hat{\theta}, \hat{\phi}), (\theta, \phi)] = \arccos\{&\sin(\hat{\phi})\sin(\phi) \\ &+ \cos(\hat{\phi})\cos(\phi)\cos(\hat{\theta} - \theta)\}.\end{aligned} \tag{12}$$

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 3 illustrates the global scores obtained in a two-source scenario, as well as the corresponding estimated and true DoAs.

IV. Analysis by LRP

A. Presentation of the technique

We analyze the inner working of our CRNN using LRP. LRP [34] 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 [34] for examples). This and other similar techniques have not yet been employed in the context of audio source localization.

LRP is based on propagation rules which reportedly provide a better explanation than gradient-based techniques such as sensitivity analysis [35]. 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 and it is 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:

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

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

$$\sum_j R_{i \leftarrow j} = R_j. \tag{14}$$

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 with the following simple rule for relevance shares:

$$R_{i \leftarrow j} = \frac{w_{ij} z_i}{z_j} R_j. \tag{15}$$

However, this rule does not take into account the biases b_j in which some relevance can get stuck. It can also be unstable when the denominator is close to zero. It was hence modified to the so-called ϵ-rule [36]:

$$R_{i \leftarrow j} = \frac{w_{ij} z_i + \frac{\epsilon \,\mathrm{sign}(z_j) + b_j}{N}}{z_j + \epsilon \,\mathrm{sign}(z_j)} R_j, \tag{16}$$

where ϵ is a small positive constant that acts as a regularizer, N is the number of lower layer neurons connected to z_j and $\mathrm{sign}(.)$ returns the sign of a number. Because of the ϵ factor, the relevance is conserved only approximately. If ϵ is too big, the backpropagation of the relevance to the inputs is altered and can become insignificant.

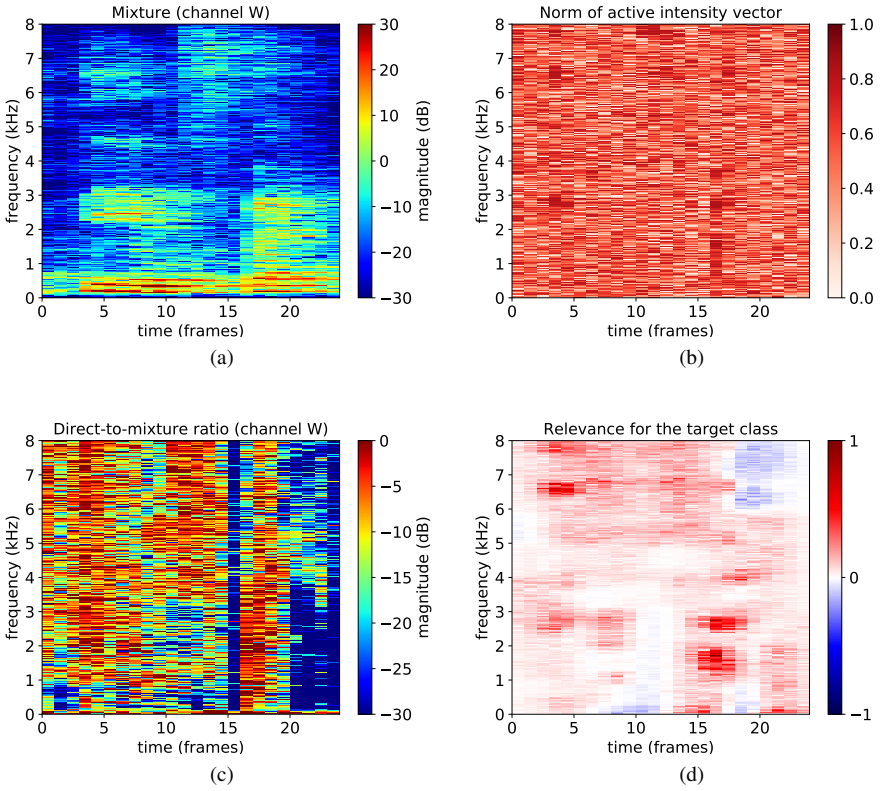


Fig. 4. LRP in the case of an accurate single-source DoA estimation with high SNR. The mixture signal consists of one speech source impinging from $(\theta, \phi) = (139°, -61°)$ and diffuse noise, with $\text{RT}_{60} = 772$ ms and SNR ≈ 18 dB, as in Fig. 1. (a) Spectrogram of the mixture at the omnidirectional channel W, (b) norm of the active intensity vector, (c) DMR at channel W, and (d) relevance map for the accurately estimated class.

The $\alpha\beta$-rule was proposed as an alternative that is conservative and stable [26]. It treats separately the negative and positive activations:

$$R_{i \leftarrow j} = \left[\alpha \frac{(w_{ij} z_i)^+}{z_j^+} - \beta \frac{(w_{ij} z_i)^-}{z_j^-} \right] R_j \text{ with}$$
$$z_j^+ = \sum_i (w_{ij} z_i)^+ + b_j^+ \text{ and } z_j^- = \sum_i (w_{ij} z_i)^- + b_j^-, \quad (17)$$

where $(.)^+$ and $(.)^-$ are the positive and negative parts of a real scalar, respectively.

The above formulas apply to linear layers, yet a nonlinear activation function $f(.)$ commonly follows the linear neurons. The above formulas remain valid as long as this function is monotonically increasing. The z_i in the formulas are then replaced by the activated neurons $y_i = f(z_i)$.

These rules were first designed for FF layers but remain applicable to convolutional and pooling layers [26]. An adaptation to LSTM layers has also been proposed to deal with the gating mechanism [36]. The activations in the upper layer are computed in the forward pass as $z_j = z_g z_s$, with z_g the gate whose activation is comprised between 0 and 1 and z_s the source which carries the information from lower or previous layers. The backpropagation rule is then simply $R_g = 0$ and $R_s = R_j$. It may seem that the values of z_g and z_s are disregarded, but they are in fact already taken into account in R_j which depends on z_j (for example according to (17) if the LSTM layer is followed by a FF 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 [34], 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 $\sigma(\theta, \phi)$ 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

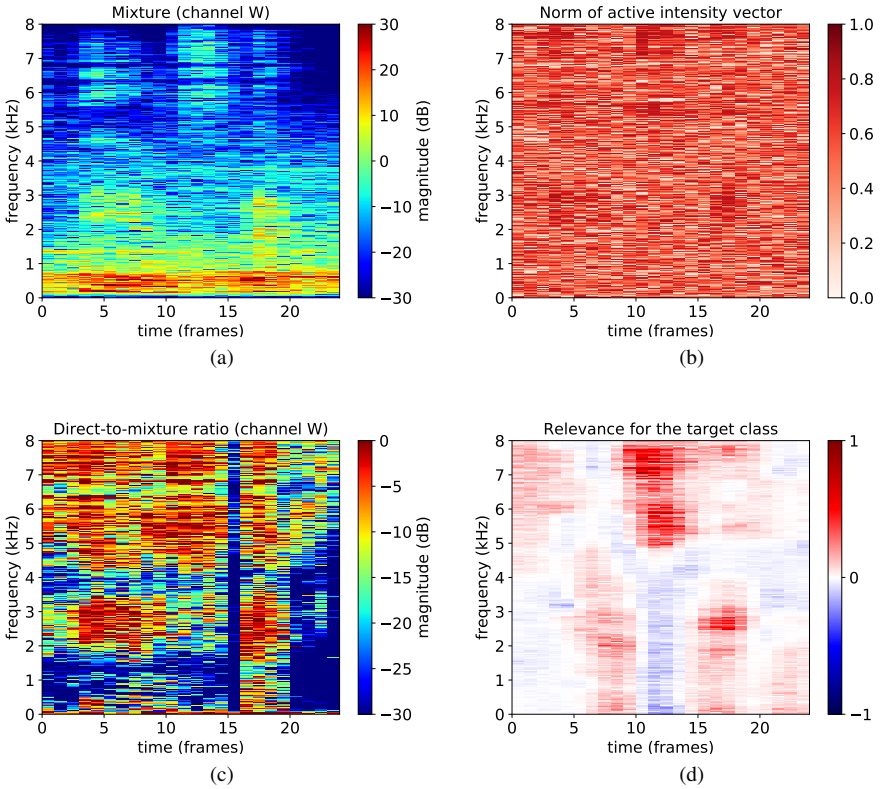


Fig. 5. LRP in the case of an accurate single-source DoA estimation with low SNR. The mixture signal consists of one speech source impinging from $(\theta, \phi) = (41°, -43°)$ and diffuse noise, with $\text{RT}_{60} = 295$ ms and SNR ≈ 1 dB. (a) Spectrogram of the mixture at the omnidirectional channel W, (b) norm of the active intensity vector, (c) DMR at channel W, and (d) relevance map for the accurately estimated class.

return either one or two DoAs. We consider various signal-to-noise ratios (SNRs) and reverberation times (RT_{60}). We also investigate cases when the network returns a wrong estimate. All observations are made on signals generated with simulated spatial room impulse responses (SRIRs); see Sections V and VI for details. To facilitate comparison, we use the same raw speech signal convolved with different SRIRs in all cases.

1) Single-source DoA estimation: Figure 4 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 (8) 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. This suggests that the network tends to focus on sound attacks. This is in accordance with psycho-acoustic studies which have shown that attacks are particularly important for DoA estimation by humans. This is known as the precedence effect [37] and has already been used to weight the importance of the time-frequency bins in some DoA estimation methods [38].

Figure 5 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 6, we examine one of the few cases when the network does not estimate the correct DoA and the score returned for the true DoA is very low. Here, the SNR is 5 dB and the RT_{60} is 661 ms. By observing the inputs, it can be noticed 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 seems 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 7 illustrates the LRP analysis for the two-source network. The relevance maps and the corresponding DMRs 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

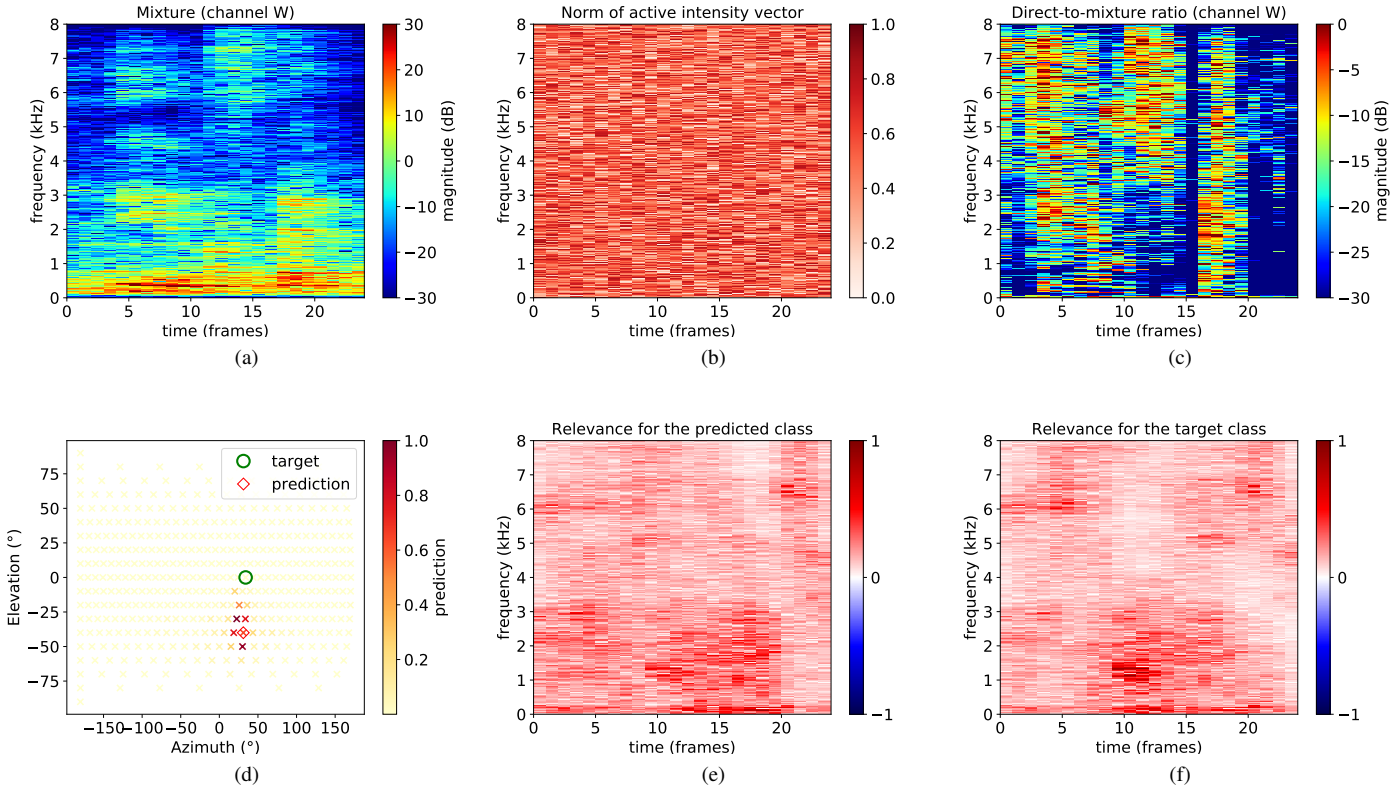


Fig. 6. LRP in the case of an inaccurate single-source DoA estimation. The mixture signal consists of one speech source impinging from $(\theta, \phi) = (29°, -2°)$ and diffuse noise, with $\text{RT}_{60} = 661$ ms and SNR ≈ 5 dB. (a) Spectrogram of the mixture at the omnidirectional channel W, (b) active 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.

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 can still be noticed in frame 14 at medium frequencies, where source 1 dominates 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 generated a large number of SRIRs with the image method [39] by adapting the code from Habets [40] to ideal FOA recordings (3). We generated SRIRs for a first source for a large number of DoAs (DoA_0), in rooms with random dimensions, a random RT_{60} and a random source-to-microphone distance. The ranges for these quantities are summarized in Algorithm 1. In the same room, we also synthesized two other SRIRs for sources coming from DoA_1 and DoA_2 so as to be able to generate several mixtures in each rooms.

Algorithm 1 Protocol to generate the SRIRs.

for each DoA_0 **do**
  repeat
    procedure ROOM
      $l = rand(2.5, 10)$
      $L = rand(2.5, 10)$ ▷ in meters
      $h = rand(2, 3)$
      $RT_{60} = rand(0.2, 0.8)$ ▷ in seconds
    end procedure
    procedure MICPOS
      $x_{mic}, y_{mic}, z_{mic} \in room$
        ▷ at least 0.5 m from walls
      $d_{mic-src} = rand(1, 3)$ ▷ in meters
    end procedure
    procedure SRCPOS
      Pick $DoA_{1,2}$
    end procedure
  until a compatible configuration is found
end for

2) Real SRIRs: In order to perform more realistic experiments, we also generated test signals with real SRIRs recorded with an Eigenmike microphone array. 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

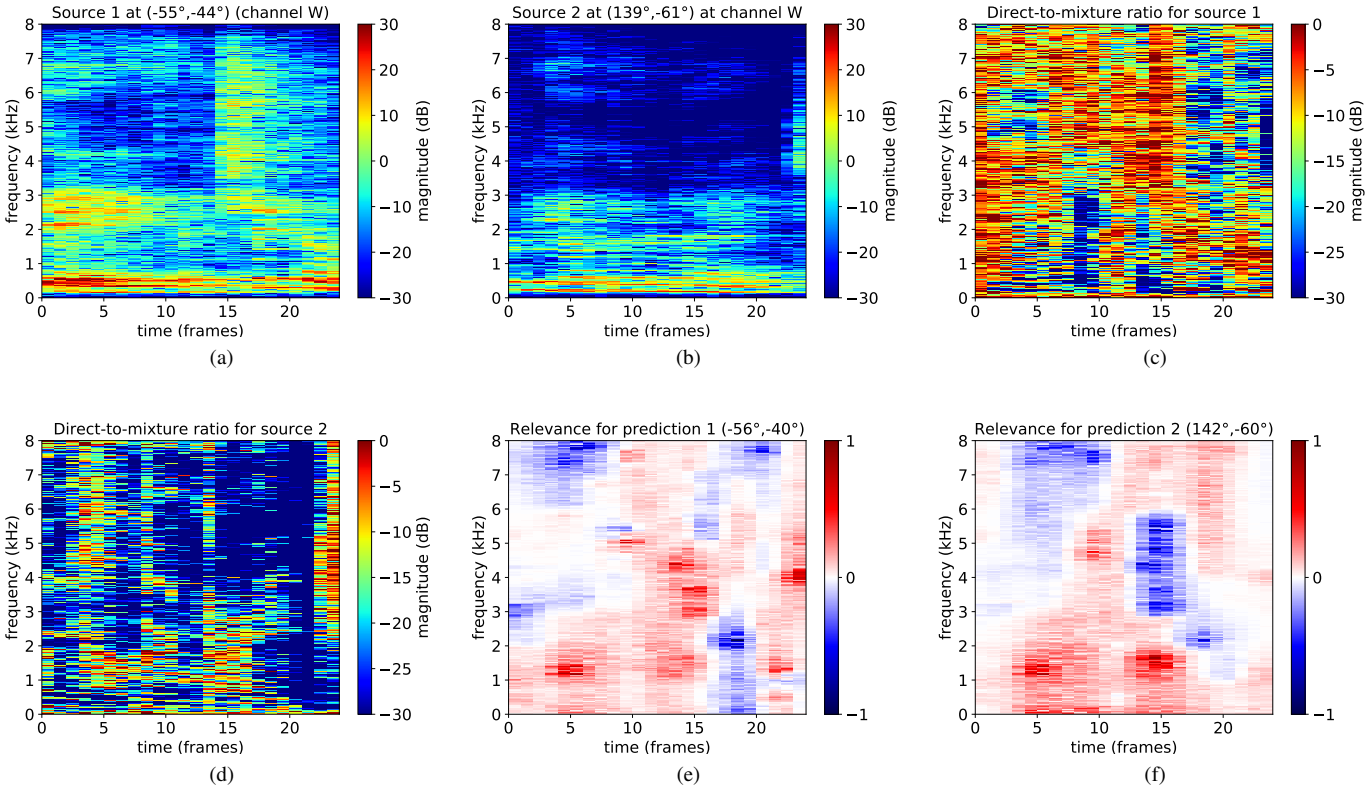


Fig. 7. LRP in the case of an accurate two-source DoA estimation. The first source impinges from $(\theta_1, \phi_1) = (-56°, -40°)$ and the second from $(\theta_2, \phi_2) = (142°, -60°)$. It is approximately 8 dB less loud than the first source. Diffuse noise is present at 20 dB SNR and the room has 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.

another resulting in 576 SRIRs. The configuration is displayed in Fig. 8. Note that this protocol resulted in situations when the microphone was actually behind the loudspeaker, which cannot happen with synthesized SRIRs when the source is supposed to be omnidirectional.

3) Real recordings: To validate our system on real data, we recorded speech with an Eigenmike in a living room (see Fig. 9). 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 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.

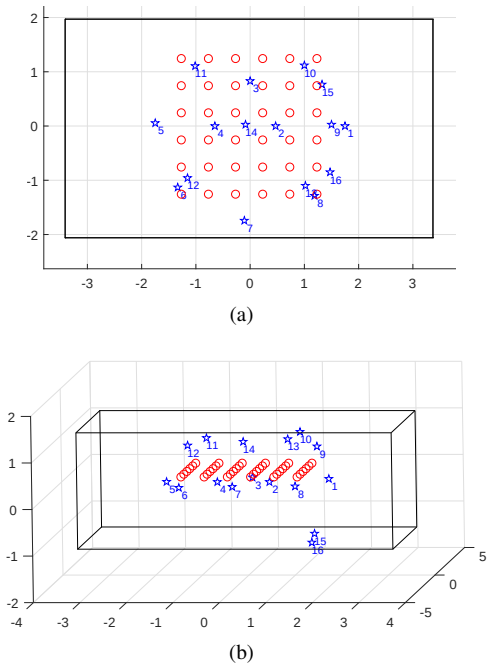


Fig. 8. Recording configuration for the real SRIRs seen (a) from the top and (b) from the side. Loudspeaker positions are denoted by blue stars, and Eigenmike positions by red circles. The loudspeakers all point toward the center of the median plan, except loudspeaker number 4 which points toward the bottom wall of (a). Dimensions are shown in meters.

Fig. 9. Layout of the living room for the real recordings. The Eigenmike used for the recordings is circled in red.

B. Training procedure

1) Network settings: The single-source and the two-source networks were constructed in the same way (see Fig. 2) and fed with sequences of 25 frames. The input shape is therefore (25, 513, 6), where 513 is the number of frequency bins and 6 is the number of feature channels. For the convolution layers we use 64 filters of size 3×3. Max pooling is performed along frequency only (over 8 frequency bins in the first two layers and 4 bins in the last layer). The three convolutional layers are 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 are used after each convolutional layer and after the first FF layer. Hard-sigmoid and tanh are used as recurrent and kernel activations in the BiLSTM layers. A sigmoid is applied after the second FF layer so that the outputs are between 0 and 1. We sample the sphere with an angular resolution of $\alpha = 10^\circ$ in (9), resulting in the 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. Each SRIR was convolved with a different 1 s speech signal randomly extracted from a subset of the Bref corpus [41]. 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). For training, the DoA_0 were forced to be equally distributed in the neighborhood of all grid points, that is to say each point of the grid has the same number of DoA_0 of which it is the target DoA. This ensures the network has seen every prediction class during training. The single-source network was trained on single speech signals generated with all SRIRs of the dataset (whether it was constructed as DoA_0,DoA_1 or DoA_2) 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. 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. A validation set with 1,287 signals was generated similarly, with different room configurations and different speakers from Bref (among 2 h of speech by 17 speakers).

3) Training settings: We used the Nadam optimizer [42] 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 the algorithm in [31], that is the best DoA estimation algorithm for FOA signals that does not employ deep learning. It relies on the estimation of 4 directions by means of wideband independent component analysis (ICA) based on entropy rate bound minimization (ERBM) [43], followed by Bayesian classification of the 4 estimated directions so as to distinguish between direct paths and false alarms due to early reflections.

Another neural network based DoA estimation system for FOA signals was recently proposed in [25] that directly takes the magnitudes and phases of the FOA signals (3) as inputs, resulting in 8 feature channels. In order to evaluate the added value of our input features (8), we also trained and tested our system 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 sequences (and sources in the two-source case) whose DoA is correctly estimated within a certain angular error tolerance. We considered 5°, 10°, and 15° tolerances. Note that, due to the chosen grid resolution, the angular distance (12) 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 correctly estimated DoA classes. 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.

[2]http://freesound.org/

Room (Section)	Simulated SRIR (V-A1)				Real SRIR (V-A2)				Real recordings (V-A3)			
Angular error	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.
Baseline [31]	28.0	57.7	72.8	n/a	24.6	55.0	70.7	n/a	**34.2**	68.9	82.3	n/a
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.

C. Single-source DoA estimation

We first evaluate 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 simulated SRIRs (see Section V-A1) different from those used for training or validation. The DoA_0 were uniformly drawn on the sphere irrespective of the grid. Each SRIR was convolved with a 1 s English speech signal randomly selected from the SiSEC campaign [44] 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 is 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 baseline, as they were trained and tested in matched conditions. The proposed Intensity-CRNN outperform the FOA-CRNN and allows a 6% absolute improvement in terms of classification accuracy.

The difference with respect to the baseline [31] is smaller with real SRIRs, especially for the 5° tolerance, which is unfavorable to CRNNs due to the grid resolution as mentioned earlier. The FOA-CRNN performs worse than the baseline, but the Intensity-CRNN still improves the 5° accuracy compared to the baseline. Furthermore, CRNNs are much less prone to outliers, with more than 87% of the sequences whose DoA is estimated within less than 15° error, compared to 70.7% only for the baseline.

Finally, on real recordings, the performance of the baseline improves compared to the other test sets while the performance of the FOA-CRNN drops dramatically. Here the reflections on the coffee table are very strong and impinge on the microphones right after the direct sound. The observed improvement for the baseline indicates that it was disturbed by diffuse noise in the previous cases, while it seems quite robust to early reflections. This is due to the combination of the ERBM-ICA algorithm, which is particularly efficient at separating direct sound and very early reflections [45], and the identification of these reflections by Bayesian classification. The Intensity-CRNN is much more robust to early reflections than the FOA-CRNN. Its accuracy with a 5° tolerance is smaller than the baseline due to the aforementioned grid resolution issue, which affects 6 positions out of 14. The results with 10 or 15° tolerance are more relevant here. When considering a 15° tolerance for instance, the Intensity-CRNN achieves 89.5% accuracy compared to 82.3% only for the baseline.

Angular error	<5°		<10°		<15°	
	θ	ϕ	θ	ϕ	θ	ϕ
Baseline [31]	**73.0**	60.9	**93.7**	77.8	**97.8**	85.2
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 0.4 TO 1.4%.

Table II reports the accuracy in terms of azimuth and elevation separately. This confirms that the collapse of the performance of the FOA-CRNN can indeed by attributed to the early reflections on the coffee table. For the 5° tolerance for instance, the FOA-CRNN is 10% relatively worse than the baseline in azimuth but 40% 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. 9). 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. Note also that the Intensity-CRNN performs comparably to the baseline in terms of azimuth, but significantly better in terms of elevation.

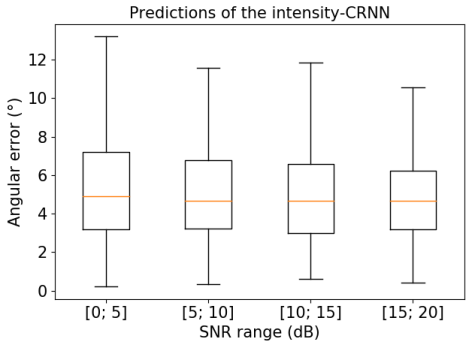
3) Robustness with respect to the SNR and the RT_{60}: Figure 10 illustrates the performance of the Intensity-CRNN as a function of the SNR and the RT_{60}. It can be observed that it is quite robust: contrary to the baseline [31] (not shown here), the distribution of angular errors appears to be mostly independent of the SNR and the RT_{60}.

D. Two-source DoA estimation

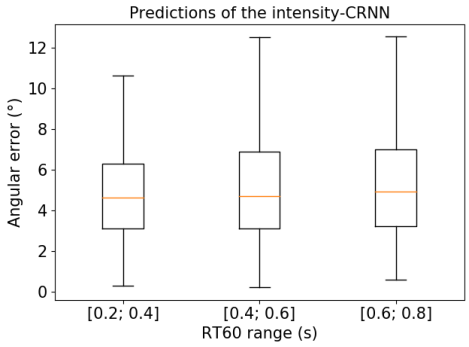
1) Test sets: The two-source network was tested on three datasets, similarly to the single-source network. In the case of

Room (Section)	Simulated SRIR (V-A1)				Real SRIR (V-A2)				Real recordings (V-A3)			
Angular error	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.	<5°	<10°	<15°	classif.
Baseline [31]	12.8	30.4	43.4	n/a	12.6	30.5	43.7	n/a	**16.3**	41.1	55.0	n/a
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.



(a)



(b)

Fig. 10. Boxplots of the angular errors (°) 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 in black and the median in orange. 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.

simulated SRIRs, the two SRIRs came from the same room and microphone configuration and, 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 are convolved with two distinct 1 s signals randomly extracted from the SiSEC corpus. The level difference between the two sources varies between 0 and 10 dB, and a diffuse babble noise is added at an SNR of 20 dB. Similarly to the single-source case, this resulted in 2,574 sequences generated from 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.

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 baseline, and the Intensity-CRNN remains the most effective system.

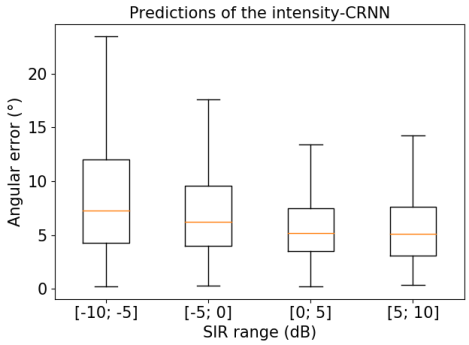
On the real SRIR dataset, the baseline now performs significantly worse than the CRNNs. It seems unable to locate multiple sources, especially in scenarios with weak direct sound due to the loudspeaker orientation. The Intensity-CRNN input still outperforms the FOA-CRNN in terms of classification accuracy.

Finally, on mixtures of real recordings, we observe once more that the FOA-CRNN is confused by the first reflections on the table. The Intensity-CRNN is actually 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 55.0% for the baseline and 52.9% for the FOA-CRNN.

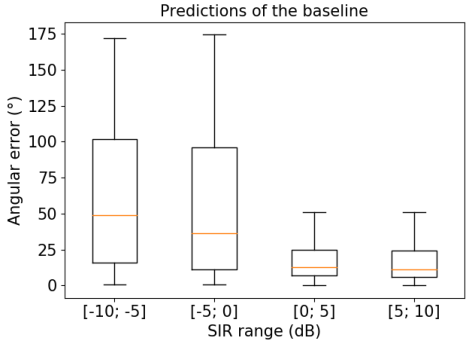
3) Robustness with respect to the SIR: Figure 11 displays the boxplots corresponding to the angular errors of the predictions returned by the Intensity-CRNN and the baseline depending on the SIR. Results are shown for the synthesised SRIRs but were similar for the other test sets. For the Intensity-CRNN predictions, 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 predictions even when the source is less loud than its competitor. On the contrary, the baseline's performance 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 Ambisonics acoustic intensity vector that enable the



(a)



(b)

Fig. 11. Boxplots of the angular errors (°) for all sequences of the two-source DoA estimation with (a) the Intensity-CRNN and (b) the baseline [31] as a function of the SIR on signals generated from simulated SRIRs.

network to properly estimate the source DoAs even in this challenging scenario. We also analyzed 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 attacks, 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 DRNN may learn more subtle cues. Also, in the case when the CRNN returns a wrong estimate, 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 DRNN 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.

[6] M. A. Gerzon, "Periphony: with-height sound reproduction," *J. Audio Eng. Soc.*, 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," *J. Audio Eng. Soc*, 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] 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.

[12] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, pp. 1950–1960, 2012.

[13] 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.

[14] 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.

[15] 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.

[16] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, 2005.

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

[18] 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, Language Process.*, vol. 25, no. 10, pp. 1956–1968, 2017.

[19] 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.

[20] X. Xiao, others, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. of ICASSP*, 2015, pp. 2814–2818.

[21] 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.

[22] V. Varanasi, R. Serizel, and E. Vincent, "DNN based robust DOA estimation in reverberant, noisy and multi-source environment," in preparation.

[23] 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.

[24] ——, "Multi-speaker localization using convolutional neural network trained with noise," in *Proc. of NIPS*, 2017.

[25] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *arXiv preprint arXiv:1710.10059*, 2017.

[26] 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, 2015.

[27] 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.

[28] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric Time-Frequency Domain Spatial Audio*. John Wiley & Sons, 2017.

[29] 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, France, 2000.
[30] 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.
[31] M. Baqué, "Analyse de scène sonore multi-capteurs," Ph.D. dissertation, Univ. du Maine, 2017.
[32] C. Dimoulas, G. Kalliris, K. Avdelidis, and G. Papanikolaou, "Improved localization of sound sources using multi-band processing of ambisonic components," in *Proc. of AES Conv. 126*, 2009, pp. 1–11.
[33] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[34] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.
[35] 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.
[36] 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.
[37] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, 1999.
[38] K. W. Wilson and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Trans. Audio, Speech, Lang.*, vol. 14, no. 6, pp. 2156–2164, 2006.
[39] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
[40] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., 2006.
[41] 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.
[42] T. Dozat, "Incorporating Nesterov momentum into Adam," Univ. of Stanford, Tech. Rep., 2015.
[43] X. L. Li and T. Adali, "A novel entropy estimator and its application to ICA," in *Proc. of MLSP*, 2009, pp. 1–6.
[44] 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.
[45] M. Baqué, A. Guérin, and M. Melon, "Separation of direct sounds from early reflections using the entropy rate bound minimization algorithm," in *Proc. of AES Conf. DREAMS*, 2016.