CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector

Lauréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin

CRNN-BASED JOINT AZIMUTH AND ELEVATION LOCALIZATION WITH THE AMBISONICS INTENSITY VECTOR

Laurélineé Perotin[⋆†] *Romain Serizel*[†] *Emmanuel Vincent*[†] *Alexandre Guérin*[⋆]

[⋆] Orange Labs, 4 rue du Clos Courtel, BP 91226, 35512 Cesson-Sévigné, France
[†] Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

ABSTRACT

We present a source localization system for first-order Ambisonics (FOA) contents based on a stacked convolutional and recurrent neural network (CRNN). We propose to use as input to the CRNN the FOA acoustic intensity vector, which is easy to compute and closely linked to the sound direction of arrival (DoA). The system estimates the DoA of a point source in both azimuth and elevation. We conduct an experimental evaluation in configurations including reverberation, noise, and various speaker w.r.t. microphone orientations. The results show that the proposed architecture and input allow the network to return accurate location estimates in realistic conditions compared to another recent CRNN-based system.

***Index Terms*—** Direction of arrival, first-order Ambisonics (FOA), acoustic intensity, CRNN

1. INTRODUCTION

Estimating the direction of arrival (DoA) of audio sources is a key prerequisite for many applications, in particular for source separation and speech recognition [1, 2]. Among traditional methods, two trends have achieved notable success : methods based on correlation, either to estimate the time difference of arrival [3] or directly the DoA [4], and subspace methods such as MUSIC [5].

When dealing with the spatial properties of a soundfield, the Ambisonics format [6] is particularly well-suited. This format is based on the decomposition of the soundfield on the basis of spherical harmonic functions. It was initially developped to unify the different spatial recording and broadcasting techniques. It is isotropic and enables easy spatial manipulation of the signal, hence its increasing use in the industry, as shown by its inclusion in the MPEG-H standard [7].

The spatial information in Ambisonics signals can be analyzed by means of the so-called acoustic intensity vector. Its computation from a first-order Ambisonics (FOA) signal is simple and it has hence been used for localization [8–10]. Alternative localization methods based on the raw FOA signals [11], on their covariance matrix [12], or on independent component analysis (ICA) [13] have also been proposed. All these methods degrade in reverberant or noisy conditions.

Recently, neural network-based localization systems have shown to be more robust to challenging conditions. Multilayer perceptrons have been applied to binaural cues [14], generalized cross correlation (GCC) features [15], eigenvectors of the spatial covariance matrix [16], or cosines and sines of interchannel phase differences [17]. Convolutional neural networks (CNNs) have also been applied to raw short time Fourier transform (STFT) phases [18]. All these works aim at recovering only the azimuth of the source in the median plan (except in [16] where the source can also have a fixed elevation). Adavanne et al. [19] were the first to use Ambisonics signals as inputs to a neural network for localization. Their network is a stacked convolutional and recurrent neural network (CRNN) that estimates the DoA on the whole sphere for up to two speakers. It achieved encouraging results, showing the relevance of the FOA format. Nevertheless, it was only tested with simulated spatial room impulse responses (SRIRs) and the source DoAs were located on the same discrete grid as the one used for training the network.

In this paper, we propose a CRNN-based system to estimate the DoA of a single static source from an FOA recording. We introduce a new feature vector based on the intensity vector which makes the network more robust to realistic conditions. We train it on a large variety of simulated SRIRs and evaluate it on unseen rooms, including a real room with difficult source/microphone configurations. Furthermore, we evaluate our system on DoAs that lie anywhere on the sphere and not only on a discrete grid.

We introduce the FOA format in Section 2. We propose our solution in Section 3. Section 4 establishes the experimental protocol used to validate the solution, and Section 5 presents the results. We conclude in Section 6.

2. AMBISONICS FORMAT

2.1. FOA format

The Ambisonics representation decomposes the soundfield in a point of space on the basis of spherical harmonic functions [6]. When using the whole infinite basis, this decomposition is exact. In practice, however, we will only use the first order which consists of four channels. The first chan-

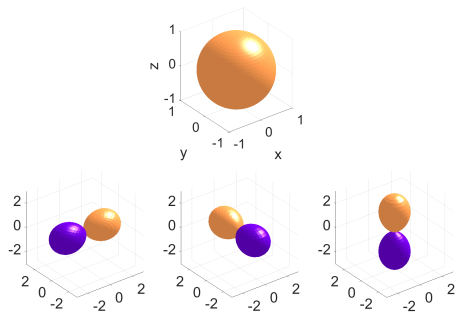

Fig. 1. Power of the first spherical harmonics for order 0 (top) and 1 (bottom). The spherical harmonics are positive in the light areas and negative in the darker areas.

nel, named W, corresponds to the order-0 spherical harmonic function, namely what an omnidirectional microphone placed at the observation point would record. The three other channels, X, Y and Z, correspond to the order-1 functions of the basis, and to what three polarized bidirectional microphones aligned on the axes $\mathbf{e}_X$, $\mathbf{e}_Y$, $\mathbf{e}_Z$ would record (see Fig. 1).

For a plane wave with azimuth θ and elevation ϕ creating a sound pressure p, the STFT FOA components are :

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

with t and f the time frame and frequency bin indexes. In anechoic conditions, the DoA of the sound source can be directly obtained from the FOA signals [11].

2.2. Intensity vector

In reverberant or noisy conditions, spatial information is often analyzed by means of the complex-valued intensity vector $\mathbf{I}(t,f) = p(t,f)^* \mathbf{v}(t,f)$ instead, where $\mathbf{v}(t,f)$ is the particle velocity vector and * denotes complex conjugation [20]. The active intensity vector $\mathbf{I}_\mathrm{a}(t,f) = \mathcal{R}\{p(t,f)^* \mathbf{v}(t,f)\}$ represents the transfer of energy in a given point in space. The reactive intensity vector $\mathbf{I}_\mathrm{r}(t,f) = \mathcal{I}\{p(t,f)^* \mathbf{v}(t,f)\}$ represents dissipative local energy transfers.

In the FOA format, the components X, Y and Z are proportional to the coordinates of $\mathbf{v}$ [9]. Disregarding this constant, the active and reactive intensity are then

$$\mathbf{I}_\mathrm{a}(t,f) = \begin{bmatrix} \mathcal{R}\{W(t,f)^* X(t,f)\} \\ \mathcal{R}\{W(t,f)^* Y(t,f)\} \\ \mathcal{R}\{W(t,f)^* Z(t,f)\} \end{bmatrix} \tag{2}$$

$$\mathbf{I}_\mathrm{r}(t,f) = \begin{bmatrix} \mathcal{I}\{W(t,f)^* X(t,f)\} \\ \mathcal{I}\{W(t,f)^* Y(t,f)\} \\ \mathcal{I}\{W(t,f)^* Z(t,f)\} \end{bmatrix}. \tag{3}$$

The DoA of the sound source can be estimated in each (t,f) bin by the opposite direction of the active intensity vector $\mathbf{I}_\mathrm{a}(t,f)$. The main direction is then recovered by majority vote [10] or averaging [8, 9] over time and frequency. However, these methods exhibit limited robustness to noise and reverberation.

3. PROPOSED METHOD

Instead of merely looking at the average direction pointed by the active intensity vector, we feed its value in all time-frequency bins to a neural network that will predict the DoA as a classification problem.

3.1. Input features

The input to the network is composed of the 6 channels of active (2) and reactive (3) parts of the intensity vector. The network is fed with several time frames at a time. We additionally normalize each time-frequency bin by its energy [13], resulting in the inputs

$$\frac{1}{|W(t,f)|^2 + \frac{1}{3}(|X(t,f)|^2 + |Y(t,f)|^2 + |Z(t,f)|^2)} \begin{bmatrix} \mathbf{I}_\mathrm{a}(t,f) \\ \mathbf{I}_\mathrm{r}(t,f) \end{bmatrix}. \tag{4}$$

Compared to the raw magnitudes and phases of the FOA signals (1) used in [19], the normalized intensity vector (4) encodes spatial information in a more invariant way. This is expected to be beneficial for neural network training.

3.2. Target

Eventually, our aim is to recover the DoA of a source, which is a continuous quantity. It would seem natural to state this as a regression problem. However, it appears harder to perform than classification [15]. We thus formulate the problem as follows : find on a pre-defined grid the DoA that is the closest to the actual one. The grid should be approximately uniform on the 2D (joint 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 :

$$\begin{cases} \phi_i & = -90 + \frac{i}{I} \times 180 \qquad \text{with } i \in \{0, \dots, I\} \\ \theta_j^{(i)} & = -180 + \frac{j}{J^{(i)}+1} \times 360 \text{ with } j \in \{0, \dots, J^{(i)}\} \end{cases} \tag{5}$$

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 DoA among those classes was one-hot encoded. The source being static, the target is the same for all frames of an utterance.

3.3. Network architecture

The network, depicted in Fig. 2, is a CRNN similar to the first half of the network used in [19]. First, features are extrac-

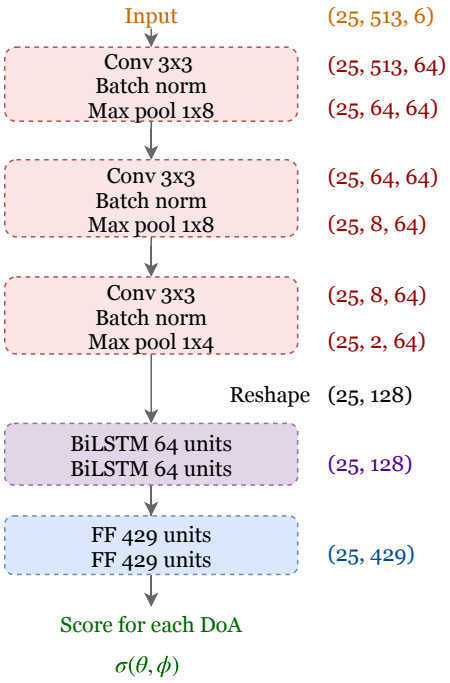

Fig. 2. Architecture of the network.

ted with convolutional layers across time and frequency with 3×3 filters, followed by batch normalization and max-pooling across frequency. A second part of the network made of two bidirectional long short-time memory (BiLSTM) layers and two time-distributed feed-forward (FF) layers returns a DoA for each time frame.

Rectified linear unit (ReLU) activations are used after each convolutional layer. Hard-sigmoid and tanh are used as recurrent and kernel activations in the BiLSTM layers. A sigmoid is finally applied after the second feed-forward layer so that the outputs are between 0 and 1. These outputs can be seen as probabilities, except that they do not sum to 1. We also tried the softmax function, which is more commonly used for classification and ensures that the outputs sum to 1, with no better results. We used the binary cross-entropy loss function for training with the Nadam [21] optimizer.

The network has to be regularized to avoid overfitting. Adavanne et. al [19] stack two networks similar to the one above, and force the intermediate output to match a MUSIC powermap. The full network is then jointly trained on the sum of the intermediate and final losses. However, we did not find this intermediate output to improve learning in the following experiments using thousands of training rooms. It even hindered it, possibly because MUSIC powermaps themselves are not accurate in highly reverberant and noisy environments, or when sources or microphones are close to a wall.

Along the same lines, we investigated a softer target than one-hot encoding, targetting 1 for the closest DoA on the grid and 0.5 for neighbouring classes. This did bring improvement when training and testing were made only on the median-plan subsets, which was already observed with similar soft targets [22]. However, we did not observe such improvement when considering the whole sphere sets. This tends to show that dataset size and variety suffice to achieve good regularization.

Eventually, we only used classical dropout after each convolutional block, on the recurrent weights of the BiLSTM layers, and after the first feed-forward layer.

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

4. EXPERIMENTAL SETUP

4.1. Data

For training, we generated many room configurations and source/microphone positions so that the network can generalize to unseen conditions. Specifically, we generated 42,000 room configurations with random dimensions and reverberation times (RT_{60}) between 200 ms and 800 ms. In each room, we picked 3 DoAs randomly on the sphere as summarized in Algorithm 1, with at least 10° angular distance as defined in (6) between each pair. In order to ensure the network sees all prediction class during training, the DoA_0 were forced to be equally reparted in the neighborhood of all grid points. We simulated SRIRs via the image method [23] by adapting the software in [24] to the FOA format, resulting in a total of 128,700 SRIRs. We convolved each SRIR with a 1 s French speech signal randomly selected from the Bref corpus [25] and added diffuse babble noise at a random SNR between 0 and 20 dB. The diffuse field was simulated by averaging the diffuse parts of two SRIRs in the dataset for the same room configuration. 1,287 signals were generated similarly for validation, with different room configurations and different speakers from Bref.

We tested the algorithm on two datasets. On the one hand, 1,287 signals simulated similarly to the training and validation sets in 429 different simulated rooms with random DoAs picked uniformly on the sphere regardless of the grid, but this time with at least 25° between each pair of sources. On the other hand, 576 signals generated using real SRIRs measured in a room with $RT60 = 500$ ms. 16 loudspeakers with different heights and orientations emitted towards 36 microphone positions uniformly placed in the room (at least 50 cm from any wall). The microphones sometimes ended up behind the loudspeaker, which had not be seen in training where the sources were supposed to be omnidirectionnal. For both test sets, the SRIRs were convolved with 1 s of English speech from the SiSEC campaign [26] and diffuse babble noise was added at an SNR between 0 and 20 dB.

4.2. Algorithm parameters

All signals were sampled at 16 kHz. The STFT was performed on 1024 points with half-overlapping sine windows. Each 1 s utterance was split in two sequences of 25 frames with 12 overlapping frames between sequences. The dimensions of the input to the network are therefore $25 \times 513 \times 6$. The angular step for the prediction grid in (5) was chosen as $\alpha = 10°$, resulting in 429 classes. Dropout was applied with a rate of 0.2. Early stopping with a patience of 20 epochs was used to further regularize the system.

We used the localization system in [13], which provided the best results among pre-deep learning methods, as a baseline. It first applies ICA to the mixture in order to estimate the mixing matrix [27]. DoAs are estimated for each column of this matrix (4 in the case of FOA). Naive Bayesian classification then discriminates between the direct path of the source and false alarms due to reflections.

In order to assert the impact of input features, we also used a system similar to ours that is trained and tested directly on the magnitude and phase of the FOA signals (1) as in [19].

4.3. Evaluation measure

For each test sequence, we average the network outputs over all frames in order to obtain a global score for each DoA. The estimated DoA for the sequence then corresponds to the averaged output with the highest score. We assess performance by computing the angular distance on the sphere between the predicted DoA $(\hat{\theta}, \hat{\phi})$ and the actual DoA (θ, ϕ) :

$$\begin{aligned}\delta[(\hat{\theta}, \hat{\phi}), (\theta, \phi)] = \arccos\{ &\sin(\hat{\phi}) \sin(\phi) \\ &+ \cos(\hat{\phi}) \cos(\phi) \cos(\hat{\theta} - \theta)\}.\end{aligned} \tag{6}$$

We define the accuracy as the proportion of test sequences for which the angular error is below a certain tolerance threshold (5°, 10°and 15°). The angular distance between a point on the sphere and the closest point on the grid being up to 7°, we also compute the classification accuracy for the DRNN models.

Room	Simulated SRIR			Real SRIR		
Angular error	<5°	<10°	<15°	<5°	<10°	<15°
Baseline [13]	27.5	56.6	70.2	24.6	55.0	70.7
CRNN + (1)	45.9	85.1	92.7	23.9	66.0	87.0
CRNN + (4) (proposed)	**51.6**	**91.1**	**95.2**	**28.6**	**70.2**	**89.6**

Table 1. Accuracy (%) of the tested systems for 5°, 10° or 15° angular error tolerance. The 95% confidence intervals vary from $\pm 1.2\%$ to $\pm 1.9\%$. The best results are shown in bold.

5. RESULTS

The accuracies of the three systems are shown in Table 1. When testing on data generated from simulated SRIRs, the two CRNN-based systems perform much better than the baseline, as they were trained in similar conditions. By using the normalized intensity vector (4) as input to the CRNN, our system improves the accuracy by 12% relative for a 5° tolerance compared to using the raw FOA signals (1). The classification accuracies are respectively 52.1% with the raw FOA inputs (1) and 58.0% with the intensity vector input (4).

Data generated from real SRIRs provide the most significant testing conditions. The CRNN with raw FOA inputs doesn't surpass the baseline for the accuracy with 5° tolerance, but using the intensity vector (4) brings 16% relative improvement. CRNNs are less prone to outliers : the 15° tolerance accuracy is improved by a relative 27% between the baseline and our proposed model. The classification accuracies are respectively 29.7% with (1) and 34.1% with (4).

It is worth mentioning that the computation time is more than 10 times faster with the CRNN based methods than with the baseline which includes a time-consuming ICA step.

6. CONCLUSION

We proposed a new CRNN-based method for estimating the DoA of a sound source from FOA contents. The proposed method relies on using the normalized acoustic intensity vector as input. Experiments were carried out with real SRIRs measured in adverse conditions : reverberation, background noise, random microphone and loudspeaker positions leading to unsual orientations, wall interferences, and sources coming from any direction in space. They showed the superiority of CRNN over previous Ambisonics localization systems and of the proposed input over raw Ambisonics signals. In the future, we plan to extend this work to the case of overlapping speech.

7. ACKNOWLEDGMENT

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

8. REFERENCES

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

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

[3] 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.

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

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

[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] Ville 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*, Aug. 2010, pp. 442–446.

[10] H. Khaddour, J. Schimmel, and M. Trzos, "Three-dimensional sound source localization using B-format signals," *Int. J. of Advances in Telecommunications, Electrotechnics, Signals and Systems*, vol. 2, no. 2, 2013.

[11] 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.

[12] 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.

[13] M. Baqué, *Analyse de scène sonore multi-capteurs*, Ph.D. thesis, Univ. du Maine, 2017.

[14] 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.

[15] 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.

[16] 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.

[17] V. Varanasi, R. Serizel, and E. Vincent, "DNN based robust DOA estimation in reverberant, noisy and multi-source environment," *to be submitted*, 2018.

[18] 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.

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

[20] 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.

[21] T. Dozat, "Incorporating Nesterov momentum into Adam," Tech. Rep., Univ. of Stanford, 2015.

[22] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," *Proc. of ICRA*, 2018.

[23] 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.

[24] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2006.

[25] 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.

[26] 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.

[27] X. L. Li and T. Adali, "A novel entropy estimator and its application to ICA," in *Proc. of MLSP*, 2009, pp. 1–6.