Quaternion Neural Networks for 3D Sound Source Localization in Reverberant Environments

Michela Ricciardi Celsi, Simone Scardapane and Danilo Comminiello
Department of Information Engineering, Electronics and Telecommunications
Sapienza University of Rome, Italy

***Abstract*—Localization of sound sources in 3D sound fields is an extremely challenging task. This task turns out to be even more difficult when the environments are reverberant and involve multiple sources. In this work, we propose a deep neural network to analyze audio signals recorded by 3D microphones and localize sound sources in a spatial sound field. In particular, we consider first-order Ambisonics microphones to capture 3D acoustic signals and represent them by spherical harmonics decomposition in the quaternion domain. Moreover, in order to improve the localization performance, we use quaternion input features derived from the acoustic intensity vector, which is strictly related to the direction of arrival (DOA) of a sound source. The proposed network architecture involves both quaternion-valued convolutional and recurrent layers. Results show that the proposed method is able to exploits both the quaternion-valued representation of ambisonic signals and to improve the localization performance with respect to existing methods.**

***Index Terms*—Quaternion neural networks, Hypercomplex-valued neural networks, 3D audio, Source localization, Convolutional recurrent neural networks**

I. Introduction

Sound source localization (SSL) is a fundamental task for many audio applications, including acoustic scene analysis, blind source separation, noise reduction, speech recognition and enhancement, among others [1]–[3].

Traditional SSL methods based on the estimation of the direction of arrival (DOA) are still adopted for many applications. In particular, different advances have been recently proposed based on the gridless DOA estimation [4], super-resolution DOA estimation [5], Bayesian learning approaches [6], [7] and subspace-based approaches [8], [9], among others. Such methods are very effective, but often for specific applications only, most of which are also limited to bidimensional cases or to short audio streams. However, when there is a need to process large audio streams and to obtain a robust DOA estimation to noisy and challenging environmental conditions, deep neural networks are often adopted [10]–[12].

Recently, there has been a remarkable growth of 3D audio industry, involving the development of applications requiring the analysis of huge audio streams acquired through 3D microphone techniques, such as the Ambisonics [13]. Such microphones are basically arrays of coincident capsules, which are based on the decomposition of the sound field into a linear combination of spherical harmonics [14]. In order to analyze ambisonic signals and estimate the DOA in the spatial sound field, some deep neural network architectures have been adopted, involving convolutional and recurrent layers [15], [16].

Among such advances 3D audio analysis, neural architectures in the hypercomplex domain have been proposed. In particular, deep quaternion neural networks (DQNNs) have proven to be very effective for the analysis of 3D audio signals captured by first-order Ambisonics (FOA), involving 4 coincident microphone capsules [17]. DQNNs are able to exploit the spatial harmonic decomposition and take advantage of the intrinsic correlations among the ambisonic signals [17].

Based on these results, we propose here a DQNN for 3D SSL in reverberant and multisource environments. When dealing with neural networks for SSL, one of the key choices to reduce the DOA estimation error is represented by the input features to be passed to the network. To this end, several solutions have been proposed in the literature, based on raw waveform [18], or on short time Fourier transform (STFT) frames [12], [17], rather than on generalized cross correlation [19], [20], binaural cues [21] or spatial covariance matrices [11]. In this work, we adopt the acoustic intensity vectors as input features [22], [23], which have already shown promising results for 3D SSL [16]. In particular, we propose a representation of such features in the quaternion domain in order to be passed as input to a quaternion-valued convolutional recurrent neural network. Results show that the proposed method, involving the quaternion-valued acoustic intensity vectors, will be able to improve the SSL performance with respect to existing methods in reverberant and noisy environments with the presence of overlapping sound sources.

The paper is organized as follows. In Section II, the representation of the 3D sound field in the quaternion domain is described, while the derivation of the quaternion-valued acoustic intensity vectors is presented in Section III. The architecture of the proposed DQNN is described in Section IV, while experimental results are shown in Section V. Finally, conclusion are drawn in Section VI.

II. Quaternion-Valued Ambisonic Signals

The Ambisonics technique performs a sampling of the spatial sound field based on its decomposition into a linear combination of orthogonal basis of spherical harmonics [13], [14]. In this work, we consider the first-order Ambisonics (FOA), which is composed of 4 microphone capsules. The first one

This work has been supported by “Progetti di Ricerca Grandi” of Sapienza University of Rome under grant number RG11916B88E1942F.

Corresponding author’s email: danilo.comminiello@uniroma1.it.

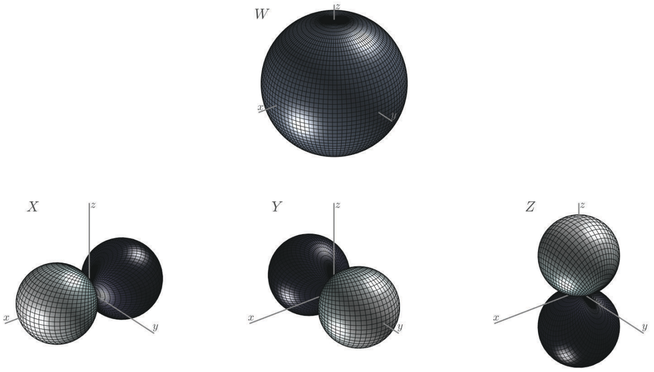

Fig. 1. Spherical harmonics representation of first-order Ambisonics.

is a pressure microphone, e.g., omnidirectional microphone with a unitary gain for all directions. It is usually denoted with W and corresponds to the spherical harmonic function of order 0. The other three microphones are orthogonal figure-of-eight microphones related to the gradient of the pressure, i.e., the acoustic velocity. These microphones are denoted with X, Y, Z, and correspond to the spherical harmonic functions of order 1. The graphical representation of the spherical harmonic decomposition of the FOA is depicted in Fig. 1.

The first-order spherical harmonic decomposition of a plane wave of a sound pressure signal in discrete-time domain $s\left[n\right]$ with angles $\left(\theta, \varphi\right)$ can be represented by the following 4 ambisonic components:

$$\begin{cases} x_W\left[n\right] = s\left[n\right]/\sqrt{3} \\ x_X\left[n\right] = s\left[n\right]\cos\left(\theta\right)\cos\left(\varphi\right) \\ x_Y\left[n\right] = s\left[n\right]\sin\left(\theta\right)\cos\left(\varphi\right) \\ x_Z\left[n\right] = s\left[n\right]\sin\left(\varphi\right) \end{cases} \tag{1}$$

that defines the B-format Ambisonics representation, which is equivalent to apply real gains to the received sound source signal $s\left[n\right]$. The factor $1/\sqrt{3}$ to the omnidirectional microphone allows to attenuate the average energy of that channel by 3 dB about on the full sphere, thus making it equal to the average energy of all the other channels [24].

Since we want to deal with the above 4 FOA components as a single entity, we represent the B-format Ambisonics of (1) in the quaternion-valued domain. To this end, we enclose the individual components of (1) in a quaternion-valued signal:

$$x\left[n\right] = x_W\left[n\right] + x_X\left[n\right]\hat{\imath} + x_Y\left[n\right]\hat{\jmath} + x_Z\left[n\right]\hat{\kappa}. \tag{2}$$

The above expression defines a quaternion-valued ambisonic signal $x\left[n\right]$, in which the omnidirectional microphone signal $x_W\left[n\right]$ represents the real component of the quaternion, while the three figure-of-eight microphone signals, $x_X\left[n\right]$, $x_Y\left[n\right]$ and $x_Z\left[n\right]$, are the imaginary components of the quaternion.

III. Quaternion-Valued Input Features for 3D SSL

In this section, we derive the quaternion-valued input features that will be passed to the network, based on the definition of acoustic intensity.

A. Acoustic Intensity

The acoustic intensity vector, or sound field intensity [25], is defined as the time averaged product of sound pressure $p\left[n\right]$ and particle velocity $\mathbf{v}\left[n\right]$, i.e.:

$$\mathbf{I}\left[n\right] = p\left[n\right]\mathbf{v}\left[n\right]. \tag{3}$$

The acoustic intensity vector $\mathbf{I}\left[n\right]$ provides the magnitude and the direction of the flow of the sound energy.

With reference to the B-format Ambisonics, the sound pressure is defined by the signal captured by the omnidirectional signal, i.e., $p\left[n\right] = x_W\left[n\right]$, while the particle velocity is defined by the three orthogonal figure-of-eight microphone signals [1]:

$$\mathbf{v}\left[n\right] = -\frac{1}{\rho_0 c\sqrt{3}}\begin{bmatrix} x_X\left[n\right] \\ x_Y\left[n\right] \\ x_Z\left[n\right] \end{bmatrix} \tag{4}$$

where ρ_0 is the mean density of the air and c is the speed of sound.

In order to consider harmonic fluctuations associated to a sound field, we express the acoustic intensity in the discrete time-frequency domain in terms of complex-valued pressure $p\left[k,n\right]$ and particle velocity $\mathbf{v}\left[k,n\right]$, where k denotes the frequency bin index. Therefore, we can define the short-time Fourier transform (STFT) of the acoustic intensity for a general harmonic fluctuation in pressure and velocity as [1], [26]:

$$\begin{aligned} \mathbf{I}\left[k,n\right] &= p^*\left[k,n\right]\mathbf{v}\left[k,n\right] \\ &= \mathbf{I}_a\left[k,n\right] + j\mathbf{I}_r\left[k,n\right] \end{aligned} \tag{5}$$

where * denotes the complex conjugation. From (5), we can see that the acoustic intensity may be split into two components: an *active* intensity component $\mathbf{I}_a\left[k,n\right]$, and a *reactive* intensity component $\mathbf{I}_r\left[k,n\right]$.

The active intensity is nothing but the time average of the acoustic intensity, i.e.:

$$\mathbf{I}_a\left[k,n\right] = \Re\left\{p^*\left[k,n\right]\mathbf{v}\left[k,n\right]\right\}, \tag{6}$$

which corresponds to the local net transport of sound energy and whose mean value is non-zero. On the other hand, the reactive intensity is defined as the imaginary counterpart of the active intensity:

$$\mathbf{I}_r\left[k,n\right] = \Im\left\{p^*\left[k,n\right]\mathbf{v}\left[k,n\right]\right\}, \tag{7}$$

which corresponds to the dissipative local energy transport and whose mean value is zero [25]. For a sound wave propagating in free field, the reactive intensity is zero in the far field. These conditions hold in the open air or in an anechoic room where all the sound striking the walls is absorbed.

Since the intensity vectors contain the information of the acoustical energy direction of a sound wave, the acoustic intensity vector can be directly used for DOA estimation. In particular, from the knowledge of $\mathbf{I}_a\left[n\right]$, it is possible to ideally derive the most prominent DOA as the opposite direction of the active intensity vector [1]:

$$\mathrm{DOA} = \angle\,\mathrm{E}\left\{-\mathbf{I}\right\} \tag{8}$$

where $\mathrm{E}\{\cdot\}$ is the expectation operator and $\angle$ gives the 3D angle. However, (8) holds only in anechoic environments, since the estimates are inconsistent in reverberant conditions [1], [16]. To this end, we propose a more robust method for DOA estimation in reverberant and noisy environments.

B. Quaternion-Valued Input Features

The active intensity relates more directly to the DOA, while 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. In this work, both active and reactive intensity vectors are extracted by the spectrogram from each audio channel and are used as separate features. The proposed method takes a sequence of features in consecutive spectrogram frames as input and predicts all the sound event classes active for each of the input frames along with their respective spatial location. Therefore, similarly to [16], we use both the active and reactive intensity vectors across all frequency bins in the STFT domain as input features. However, we represent the input features in the quaternion domain in order to leverage the analysis by the proposed quaternion-valued neural network.

In particular, in order to represent the active and reactive intensity vectors we encapsulate the information in two quaternions. Since we have three components for each intensity vector, we also consider one more channel related to the magnitude of the omnidirectional microphone signal in order to improve the performances of the localization task. Therefore, the input features can be expressed by the two quaternions $q_a[k,n]$ and $q_r[k,n]$:

$$\begin{aligned} q_a[k,n] = {} & \Re\{x_W^*[k,n]\, x_W[k,n]\} \\ & + \Re\{x_W^*[k,n]\, x_X[k,n]\}\,\hat{\imath} \\ & + \Re\{x_W^*[k,n]\, x_Y[k,n]\}\,\hat{\jmath} \\ & + \Re\{x_W^*[k,n]\, x_Z[k,n]\}\,\hat{\kappa} \end{aligned} \tag{9}$$

$$\begin{aligned} q_r[k,n] = {} & \Im\{x_W^*[k,n]\, x_W[k,n]\} \\ & + \Im\{x_W^*[k,n]\, x_X[k,n]\}\,\hat{\imath} \\ & + \Im\{x_W^*[k,n]\, x_Y[k,n]\}\,\hat{\jmath} \\ & + \Im\{x_W^*[k,n]\, x_Z[k,n]\}\,\hat{\kappa} \end{aligned} \tag{10}$$

Then, we normalize the quaternion inputs in each time-frequency bin, as also suggested by [16], in order to keep the inputs in a fixed range regardless of the sound power. In particular, we normalize each time-frequency bin by its total energy $\varepsilon_{\mathrm{T}} = \varepsilon_{\mathrm{P}} + \varepsilon_{\mathrm{K}}$, where $\varepsilon_{\mathrm{P}} = |x_W[k,n]|^2$ is the potential energy density related to the sound pressure and $\varepsilon_{\mathrm{K}} = \frac{1}{3}\left(|x_X[k,n]|^2 + |x_Y[k,n]|^2 + |x_Z[k,n]|^2\right)$ is the kinetic energy density related to the particle velocity. Therefore, the normalized quaternion inputs can be expressed as:

$$\begin{aligned} \overline{q}_a[k,n] &= \frac{q_a[k,n]}{\varepsilon_{\mathrm{T}}} \\ \overline{q}_r[k,n] &= \frac{q_r[k,n]}{\varepsilon_{\mathrm{T}}}. \end{aligned} \tag{11}$$

The proposed model receives the quaternion ambisonic recording, from which it extracts the spectrogram in terms of magnitude and phase components using a Hamming window of length M, an overlap of 50%, and considering only the $M/2$ positive frequencies without the zeroth bin, similarly to [17], [27]. Therefore, from the two quaternion inputs, we obtain a feature sequence of T frames, with an overall dimension of $T \times M/2 \times 8$.

IV. Proposed QSSLnet Architecture

We introduce now the quaternion-valued sound source localization network (QSSLnet), depicted in Fig. 2, that we use to perform the 3D DOA estimation. The QSSLnet involves a series of convolutional and recurrent layers in the quaternion domain and yields the coordinates estimates of sound sources in the continuous spatial space as a multi-output regression task.

A. Quaternion Convolutional Recurrent Neural Network

In quaternion-valued neural networks (QNNs), all the parameters are quaternions, including inputs, outputs and weights. Moreover, operations between these quantities are performed by using quaternion algebra $\mathbb{H}$ [28]. One of the fundamental operations in quaternion-valued convolutional neural networks (QCNNs) is the convolution process performed in the quaternion domain. In particular, we consider a generic quaternion input vector, $\mathbf{x}$, defined similarly to (2), and a generic quaternion filter matrix defined as $\mathbf{W} = \mathbf{W}_W + \mathbf{W}_X\hat{\imath} + \mathbf{W}_Y\hat{\jmath} + \mathbf{W}_Z\hat{\kappa}$. The quaternion convolution is performed by the following Hamilton product:

$$\begin{aligned} \mathbf{W} \otimes \mathbf{x} = {} & (\mathbf{W}_W\mathbf{x}_W - \mathbf{W}_X\mathbf{x}_X - \mathbf{W}_Y\mathbf{x}_Y - \mathbf{W}_Z\mathbf{x}_Z) \\ & + (\mathbf{W}_W\mathbf{x}_X + \mathbf{W}_X\mathbf{x}_W + \mathbf{W}_Y\mathbf{x}_Z - \mathbf{W}_Z\mathbf{x}_Y)\,\hat{\imath} \\ & + (\mathbf{W}_W\mathbf{x}_Y - \mathbf{W}_X\mathbf{x}_Z + \mathbf{W}_Y\mathbf{x}_W + \mathbf{W}_Z\mathbf{x}_X)\,\hat{\jmath} \\ & + (\mathbf{W}_W\mathbf{x}_Z + \mathbf{W}_X\mathbf{x}_Y - \mathbf{W}_Y\mathbf{x}_X + \mathbf{W}_Z\mathbf{x}_W)\,\hat{\kappa} \end{aligned} \tag{12}$$

The Hamilton product allows quaternion neural networks to capture internal latent relations within the features of a quaternion. Indeed, in a real-valued neural network, the multiple weights required to code latent relations within a feature are considered at the same level as for learning global relations between different features, while the quaternion weight codes these internal relations within a unique quaternion output during the Hamilton product [29].

The forward phase for a generic quaternion dense layer can be defined by the following expression:

$$\mathbf{y} = \alpha(\mathbf{W} \otimes \mathbf{x} + \mathbf{b}) \tag{13}$$

where $\mathbf{y}$ is the output of the layer, $\mathbf{b}$ is the quaternion-valued bias offset and α is a quaternion activation function. A simple and suitable choice for α is represented by the quaternion split activation function [17], [29], defined for a generic quaternion q as:

$$\alpha(q) = f(q_W) + f(q_X) + f(q_Y) + f(q_Z) \tag{14}$$

where $f(\cdot)$ is a rectified linear unit (ReLU) activation function, but in general it can be any standard activation function.

The QCNN layers are composed of P filter kernels with size $3 \times 3 \times 8$. Filter kernels have a significant role for localization

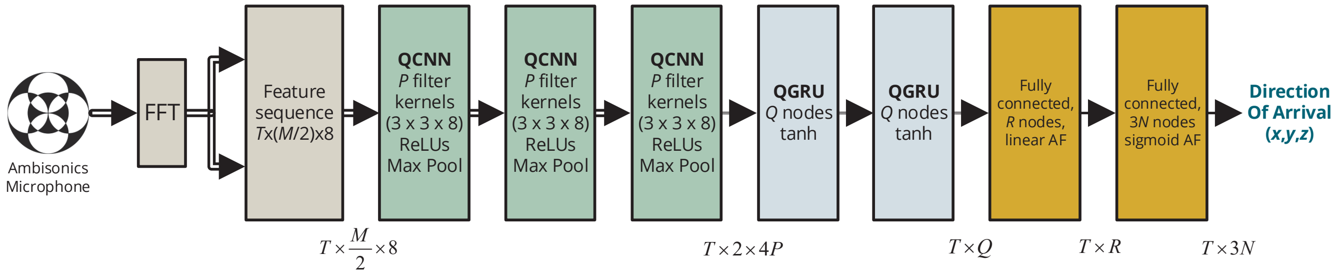
Ambisonics Microphone
FFT
Feature sequence Tx(M/2)x8
QCNN P filter kernels (3 x 3 x 8) ReLUs Max Pool
QCNN P filter kernels (3 x 3 x 8) ReLUs Max Pool
QCNN P filter kernels (3 x 3 x 8) ReLUs Max Pool
QGRU Q nodes tanh
QGRU Q nodes tanh
Fully connected, R nodes, linear AF
Fully connected, 3N nodes sigmoid AF
Direction Of Arrival (x,y,z)
$T\times\frac{M}{2}\times 8$
$T\times 2\times 4P$
$T\times Q$
$T\times R$
$T\times 3N$

Fig. 2. Scheme of the proposed quaternion-valued sound source localization network (QSSLnet).

since they allow learning inter-channel features. At the output of the ReLU activation function a batch normalization is performed and a max-pooling is applied along the frequency axis for dimensionality reduction while preserving the sequence length T. The output of the final QCNN layer has a dimension of $T \times 2P$, where the frequency dimension 2 is reduced by the max-pooling, while the number of output feature maps is 4 times larger, with respect to a standard CNN, due to the quaternion convolution.

The output of the QCNN is reshaped into a $T \times 8P$ frame and fed to bidirectional QRNN layers to better catch the time progress of the input signals. In particular, Q nodes of quaternion gated recurrent units (QGRU) are used in each layer. Let us consider a generic input vector $\mathbf{x}_n$ at time instant n, the hidden state $\mathbf{h}_n$, the output vector $\mathbf{p}_n$, the update gate vector $\mathbf{z}_n$, the reset gate vector $\mathbf{r}_n$, the memory vector $\mathbf{m}_n$, the weight matrices $\mathbf{W}$, $\mathbf{U}$, $\mathbf{V}$ and the bias vector $\mathbf{b}$. The forward phase can be summarized as follows:

$$\mathbf{z}_n = \alpha\left(\mathbf{W}^{(z)} \otimes \mathbf{x}_n + \mathbf{U}^{(z)} \otimes \mathbf{h}_{n-1} + \mathbf{b}^{(z)}\right) \tag{15}$$

$$\mathbf{r}_n = \alpha\left(\mathbf{W}^{(r)} \otimes \mathbf{x}_n + \mathbf{U}^{(r)} \otimes \mathbf{h}_{n-1} + \mathbf{b}^{(r)}\right) \tag{16}$$

$$\mathbf{m}_n = \alpha\left(\mathbf{W}^{(m)} \otimes \mathbf{x}_n + \mathbf{U}^{(m)} \otimes (\mathbf{r}_n \odot \mathbf{h}_{n-1}) + \mathbf{b}^{(r)}\right) \tag{17}$$

$$\mathbf{h}_n = (\mathbf{1} - \mathbf{z}_n) \odot \mathbf{h}_{n-1} + \mathbf{z}_n \odot \mathbf{m}_n \tag{18}$$

$$\mathbf{p}_n = \beta\left(\mathbf{V} \otimes \mathbf{h}_n\right) \tag{19}$$

where β represents any split activation function and α represents any quaternion split activation function. In particular, we use a sigmoid function for (15) and (16), while a tanh for (17). Concerning the back propagation through time, it follows the same logic of the standard QRNN [30].

As depicted also in Fig. 2, after the recurrent layers, we have two further layers that perform the task. The first of the two layer involves R nodes with linear activation functions, while the last layer involves $3N$ nodes, representing the Cardinal coordinates for each sound event class, and a hyperbolic tangent activation function. We use a mean square error (MSE) loss for the localization task, as done in [17], [27].

B. Weight Initialization

A possible solution for the weight initialization in the quaternion domain is based on a normalized purely quaternion $u^{\triangleleft}$ generated for each weight w by following a uniform distribution in $[0, 1]$ [30]:

$$w = |w|\, e^{u^{\triangleleft}\theta} = |w|\left(\cos(\theta) + u^{\triangleleft}\sin(\theta)\right), \tag{20}$$

whose quaternion-valued components are:

$$\begin{cases} w_W = \phi \cos(\theta) \\ w_X = \phi u_X^{\triangleleft} \sin(\theta) \\ w_Y = \phi u_Y^{\triangleleft} \sin(\theta) \\ w_Z = \phi u_W^{\triangleleft} \sin(\theta) \end{cases} \tag{21}$$

where θ is randomly generated in the range $[-\pi, \pi]$ and ϕ is a randomly generated variable related to the variance of the quaternion weight, defined as $\mathrm{var}(\mathbf{W}) = 4\sigma^2$, where σ is the standard deviation [30]. The variable ϕ in (21) can be randomly generated in the range $[-\sigma, \sigma]$.

V. Experimental Results

A. Datasets

In order to assess the proposed QSSLnet we consider two datasets involving 3D audio recordings in FOA format containing sound events in reverberant and noisy environments. Both the datasets consider stationary sources associated with spatial coordinates.

The first dataset is the *Ambisonic, Reverberant and Synthetic Impulse Response* (RESYN) dataset, consisting of spatially located sound events in a reverberant scenario using simulated impulse responses. In particular, a room of size $10 \times 8 \times 4$ m is considered with reverberation times 1.0, 0.8, 0.7, 0.6, 0.5 and 0.4 for each octave band, and 125 to 4000 Hz band center frequencies. The dataset is divided in three subsets, O1, O2, O3, involving respectively a maximum number of 1, 2 and 3 simultaneously active sound events. Each subset is composed of three validation splits with 240 training and 60 testing Ambisonics recordings, each one during 30 seconds at 44100 Hz. The dataset contains 11 isolated sound event classes, each one composed of 20 examples, 16 of which randomly chosen for the training set and the remaining 4 are used for the test set.

The second dataset is the *Ambisonic, Reverberant and Real-life Impulse Response* (REAL) dataset, generated by collecting impulse responses from a real environment using the Eigenmike spherical microphone array. The recorded multichannel audio has been then converted to FOA format. Similarly to the RESYN, the REAL dataset is divided in three subsets

TABLE I
RESULTS ON THE RESYN DATASET IN TERMS OF THE DOA SCORE. BEST SCORES ARE IN BOLD FONT.

	QSELDnet	Proposed QSSLnet
1 Overlapping Sound Event	0.25	**0.10**
2 Overlapping Sound Events	0.44	**0.34**
3 Overlapping Sound Events	0.45	**0.40**

TABLE II
RESULTS ON THE REAL DATASET IN TERMS OF THE DOA SCORE. BEST SCORES ARE IN BOLD FONT.

	QSELDnet	Proposed QSSLnet
1 Overlapping Sound Event	0.27	**0.25**
2 Overlapping Sound Events	0.45	**0.43**
3 Overlapping Sound Events	0.43	**0.40**

composed of three validation splits with 240 training and 60 testing Ambisonics recordings, each one during 30 seconds at 44100 Hz. This dataset consists of 10 sound event classes. More details on the datasets can be found in [15].

B. Metrics

In order to measure the performance of the localization task, a DOA estimation error DOA_{err} can be used as evaluation metric, based on estimated and ground truth DOAs [27]. Moreover, a frame recall metric K (ideally $K = 1$) can be used based on the percentage of true positives. A joint DOA score can be defined as:

$$S_{\text{DOA}} = \frac{(\text{DOA}_{\text{err}}/180 + (1 - K))}{2}. \qquad (22)$$

The lower the S_{DOA} score, the better the results in terms of 3D localization.

C. Simulations

We compare the proposed QSSLnet model with the quaternion-valued neural network architecture (QSELDnet) of [17] for the SSL task. The proposed quaternion network has more or less 400K parameters, while the QSELDnet [17] has 1M parameters. So, the total number of parameters is considerably reduced, being 600K parameters less than the QSELDnet, despite achieving also better performance. To this end, we set a number of $P = 64$ filters, sequence length of $T = 512$ frames, window length $M = 512$, batch size of 10, $Q = 128$ nodes for the recurrent networks and $R = 32$ nodes for the fully connected layers. The models have been trained over 300 epochs.

The reduction of the number of parameters does not result in poor performance in the proposed method. Results are collected in Table I and Table II for the RESYN and the REAL dataset, respectively. From the results, it is possible to note clearly that the new network outperforms the QSELDnet. We can notice from the results that as the presence of reverberations and real-life noise increases, e.g. in splits O2 and O3, the SSL scores are slightly higher.

Overall, results have proved that the use of quaternion-valued acoustic intensity vector as input features leads to an improvement of 3D SSL performance in challenging environments, e.g., including both reverberant and real scenarios.

VI. CONCLUSION

In this paper, we propose a quaternion-valued deep neural network, denoted as QSSLnet, for the localization of 3D sound events recorded by first-order Ambisonics. The proposed QSSLnet involves the convolution process, as well as the rest of the processing, in the quaternion domain, thus resulting in improved performance and reduced number of parameters. The proposed method involves quaternion-valued acoustic intensity vectors as input features, which provide a more accurate DOA estimation. The method is assessed on two datasets, the RESYN involving reverberant environments, and the REAL involving real data. Both the datasets involve up to 3 overlapping sound events to be localized. Results have shown that the proposed method is able to exploit the correlated nature of the ambisonic signals and the quaternion-valued intensity vectors to outperform existing methods for 3D sound localization tasks.

REFERENCES

[1] V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds., *Parametric Time-Frequency Domain Spatial Audio*. John wiley & Sons, Oct. 2018.

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

[3] J.-T. Chien, *Source Separation and Machine Learning*. Elsevier, Oct. 2018.

[4] M. Wagner, P. Gerstoft, and Y. Park, "Gridless DOA estimation via alternating projections," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Brighton, UK, May 2019, pp. 4215–4219.

[5] A. G. Raj and J. H. McClellan, "Single snapshot super-resolution DOA estimation for arbitrary array geometries," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 119–123, Jan. 2019.

[6] S. Chakrabarty and E. A. P. Habets, "A Bayesian approach to informed spatial filtering with robustness against DOA estimation errors," *IEEE/ACM Trans. Audio, Speech and Lang. Process.*, vol. 26, no. 1, pp. 145–160, Jan. 2018.

[7] J. Dai and H. C. So, "Sparse Bayesian learning approach for outlier-resistant direction-of-arrival estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 744–756, Feb. 2018.

[8] M. Esfandiari, S. A. Vorobyov, S. Alibani, and M. Karimi, "Non-iterative subspace-based DOA estimation in the presence of nonuniform noise," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 848–852, Jun. 2019.

[9] E. D. Di Claudio, R. Parisi, and G. Jacovitti, "Space time MUSIC: Consistent signal subspace estimation for wideband sensor arrays," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2685–2699, May 2018.

[10] X. Xiao, S. Zhao, X. Zhong, D. J. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 2814–2818.

[11] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 405–409.

[12] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop Applications of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2017, pp. 136–140.

[13] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," in *50th Convention of the Audio Eng. Soc. (AES)*, London, UK, Mar. 1975, pp. 402–404.

[14] B. Rafaely, *Fundamentals of Spherical Array Processing*, 2nd ed., ser. Springer Topics in Signal Processing. Cham, Switzerland: Springer Nature, 2019.

[15] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural networks," in *26th Europ. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 1476–1480.

[16] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 22–33, Mar. 2019.

[17] D. Comminiello, M. Lella, S. Scardapane, and A. Uncini, "Quaternion convolutional neural networks for detection and localization of 3D sound events," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Brighton, UK, May 2019, pp. 8533–8537.

[18] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Brighton, UK, May 2019, pp. 451–455.

[19] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, Vietri sul Mare, Italy, Sep. 2016, pp. 1–6.

[20] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 2386–2390.

[21] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech and Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[22] F. Jacobsen, "A note on instantaneous and time-averaged active and reative sound intensity," *J. Sound and Vibration*, vol. 147, no. 3, pp. 489–496, Jun. 1991.

[23] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *18th Europ. Signal Process. Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[24] M. A. Gerzon, "Practical periphony: The reproduction of full-sphere sound," in *65th Convention of the Audio Eng. Soc. (AES)*, London, UK, Feb. 1980, pp. 1–12.

[25] F. Fahy, *Sound Intensity*, 2nd ed. CRC Press, Oct. 1995.

[26] F. Fahy and D. Thompson, Eds., *Fundamental of Sound and Vibration*, 2nd ed. CRC Press, Dec. 2019.

[27] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[28] J. P. Ward, *Quaternions and Caley Numbers. Algebra ans Applications*, ser. Mathematics and Its Applications. Kluwer Academic Publishers, 1997, vol. 403.

[29] T. Parcollet, M. Morchid, and G. Linarès, "A survey of quaternion neural networks," *Artif. Intell. Rev.*, Aug. 2019.

[30] T. Parcollet, M. Ravanelli, M. Morchid, G. Linarès, C. Trabelsi, R. De Mori, and Y. Bengio, "Quaternion recurrent neural networks," in *Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, May 2019, pp. 1–19.