QUATERNION CONVOLUTIONAL NEURAL NETWORKS FOR DETECTION AND LOCALIZATION OF 3D SOUND EVENTS

Danilo Comminiello, Marco Lella, Simone Scardapane, and Aurelio Uncini

DIET Dept., Sapienza University of Rome
Via Eudossiana, 18 - 00184 Rome, Italy

ABSTRACT

Learning from data in the quaternion domain enables us to exploit internal dependencies of 4D signals and treating them as a single entity. One of the models that perfectly suits with quaternion-valued data processing is represented by 3D acoustic signals in their spherical harmonics decomposition. In this paper, we address the problem of localizing and detecting sound events in the spatial sound field by using quaternion-valued data processing. In particular, we consider the spherical harmonic components of the signals captured by a first-order ambisonic microphone and process them by using a quaternion convolutional neural network. Experimental results show that the proposed approach exploits the correlated nature of the ambisonic signals, thus improving accuracy results in 3D sound event detection and localization.

Index Terms— Quaternion neural networks, Hypercomplex machine learning, 3D audio, Ambisonics

1. INTRODUCTION

Recently, 3D audio processing has been gaining increasing attention due to significant development of spatial audio technology, which paved the way to this emerging field of application. Immersive audio has changed the way people make use of audio services, placing a greater attention to the satisfaction of the user-required quality [1, 2]. In this context, the last few years have been characterized by a wide spread of commercial intelligent acoustic interfaces, basically composed of acoustic interfaces equipped with intelligent signal processors [1, 3, 4]. This kind of devices can be found in many applications, as well as in human everyday life, such as home automation, voice assistance, safety and security by robots, audio surveillance, virtual reality in gaming and entertainment, up to speech recognition applications.

One of the most suited acoustic interfaces for high-definition capturing of the spatial sound field is represented by the Ambisonics, which is basically an array of coincident microphones. The Ambisonics technique is able to capture 3D sounds while minimizing unwanted artifacts caused by cross-talk. One of the main features of the Ambisonics is the decomposition of the sound field into a linear combination of spherical harmonics. Traditionally, each ambisonic signal is processed as a separate real-valued signal. However, also due to the physical arrangement of the microphone capsules, ambisonic signals show strongly correlated components. Thus, they lend themselves to a more exotic algebraic description in the quaternion domain that allows signals to be treated as a single multidimensional entity [5, 6].

Recently, an increasing interest has been shown on signal processing and machine learning algorithms in quaternion and hypercomplex domains [7–14]. In such a context, significant advances have been proposed on quaternion neural networks (QNNs) [15–18]. In this paper, we want to exploit the capabilities of both QNNs and Ambisonics to analyze 3D sounds, and in particular we focus on the localization and detection of 3D sound events. Both tasks have been widely investigated recently by using convolutional neural networks (CNNs) [19–25]. They are also considered as a joint task in [26] for 3D sounds, but considering each microphone signal as a separate real-valued signal.

Here, we want to exploit the characteristics of ambisonic signals by processing them as a single multidimensional entity. To this end, we propose a quaternion convolutional neural network (QCNN) for the joint 3D sound event localization and detection (SELD) task. We assess the effectiveness of the proposed method in two different 3D acoustic scenarios and we show improved performance for the SELD task with respect to real-valued CNNs proposed in the existing literature.

The paper is organized as follows. In Section 2, the representation of the 3D sound field in the quaternion domain is described, while the QCNN is introduced in Section 3. Experimental results on SELD problems are shown in Section 4. Finally, conclusion are drawn in Section 5.

2. 3D SOUNDS IN THE QUATERNION DOMAIN

The Ambisonics technique is one of the most popular 3D microphone recording techniques, which is based on a local-space sampling of the sound field by using a coincident microphone array. Such approach involves the decomposition of the sound field into a linear combination of spherical harmonics. Here we show how to deal with spherical harmonics and to consider them in the quaternion domain.

2.1. 4D Representation of Spatial Sound Fields

Spherical harmonics are orthonormal functions which can be used to represent the sound field in terms of its basic components. The sound pressure, in absence of impressed sources, can be expressed by the following wave equation depending on sound speed c, radius r, azimuth θ and elevation φ:

$$\nabla^2 p\left(r,\theta,\varphi,t\right) - \frac{1}{c^2}\frac{\partial^2 p\left(r,\theta,\varphi,t\right)}{\partial t^2} = 0. \tag{1}$$

The solution of the wave equation can be achieved by using a Fourier-Bessel series decomposition:

$$p\left(\vec{r}\right) = \sum_{m=0}^{\infty}\left(2m+1\right) j^m J_m\left(kr\right) \sum_{\substack{0\leq n\leq m,\\ \sigma=\pm 1}} X_{mn}^{\sigma} H_{mn}^{\sigma}\left(\theta,\varphi\right) \tag{2}$$

being m the decomposition degree, n the order, σ the spin and $k = 2\pi f/c$ the wave number, $J_m\left(kr\right)$ spherical Bessel functions, and X_{mn}^{σ} a signal component. The previous expression is nothing but a decomposed representation of the sound field.

Each signal component is weighted by an orthonormal function $H_{mn}^{\sigma}\left(\theta,\varphi\right)$, i.e., a *spherical harmonic*, which can be expressed in a normalized form as:

$$H_{mn}^{\sigma}\left(\theta,\varphi\right) = \tilde{P}_{mn}\sin\left(\varphi\right) \times \begin{cases} \cos\left(n\theta\right) & \text{if } \sigma = +1 \\ \sin\left(n\theta\right) & \text{if } \sigma = -1 \end{cases}. \tag{3}$$

The linear combination of spherical harmonics results in functions on the surface of a sphere.

Ambisonics is based on the previous description of the sound field by (2) and, in particular, it is described by the order n, which is also referred to as *ambisonic order*. In this paper, we focus on the so-called *B-Format Ambisonics*, whose order is $n = 1$ (which is the reason why it can be also denoted as first-order ambisonics). The B-Format Ambisonics is composed of an array of 4 coincident microphones (1 omnidirectional and 3 orthogonal figure-of-eight microphones) orthogonal to each other. Each of the 4 microphones is related to a spherical harmonic (1 related to order 0 and 3 to order 1), specifically denoted in this case by W (omnidirectional microphone), X, Y, Z (figure-of-eight microphones).

2.2. Quaternion-Valued Ambisonics Signals

Traditionally, the sound field is defined by spherical harmonics using Euler angles. Here, we aim at dealing with spherical harmonics in the quaternion-valued domain, thus we consider the four ambisonic signals, namely $x_W\left[n\right]$, $x_X\left[n\right]$, $x_Y\left[n\right]$ and $x_Z\left[n\right]$, as a single quaternion signal:

$$x\left[n\right] = x_W\left[n\right] + x_X\left[n\right]\hat{\imath} + x_Y\left[n\right]\hat{\jmath} + x_Z\left[n\right]\hat{\kappa}, \tag{4}$$

which defines a 4-dimensional spatial sound signal, i.e., a quaternion-valued ambisonic signal. In (4), the imaginary units, $\hat{\imath} = (1,0,0)$, $\hat{\jmath} = (0,1,0)$, $\hat{\kappa} = (0,0,1)$, represent an orthonormal basis in $\mathbb{R}^3$ and satisfy the fundamental properties of quaternion algebra [27]. It is worth noting that, in (4), the omnidirectional microphone signal $x_W\left[n\right]$ is considered as the real component of the quaternion signal, 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 considered as the imaginary components.

Once defined the expression of the quaternion-valued ambisonic signal, we can use it to effectively process 3D sounds in the quaternion domain, thus fully exploiting the statistical properties of multidimensional signals.

3. QUATERNION CONVOLUTIONAL NEURAL NETWORKS FOR 3D SELD

We introduce now the QCNN method used to jointly perform the 3D SELD task in the quaternion domain when signals are captured by Ambisonics.

3.1. Quaternion-Valued Convolution

The main peculiarity of a QCNN is the convolution process that is performed in the quaternion domain. Here, a quaternion filter matrix is convolved with a quaternion vector by exploiting real-valued representations of quaternions [27]. Let us consider a quaternion input vector[1], $\mathbf{x}$, defined similarly to (4), 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 obtained from the following Hamilton product:

$$\begin{aligned} \mathbf{W}\otimes\mathbf{x} = &\left(\mathbf{W}_W\mathbf{x}_W - \mathbf{W}_X\mathbf{x}_X - \mathbf{W}_Y\mathbf{x}_Y - \mathbf{W}_Z\mathbf{x}_Z\right) \\ &+ \left(\mathbf{W}_W\mathbf{x}_X + \mathbf{W}_X\mathbf{x}_W + \mathbf{W}_Y\mathbf{x}_Z - \mathbf{W}_Z\mathbf{x}_Y\right)\hat{\imath} \\ &+ \left(\mathbf{W}_W\mathbf{x}_Y - \mathbf{W}_X\mathbf{x}_Z + \mathbf{W}_Y\mathbf{x}_W + \mathbf{W}_Z\mathbf{x}_X\right)\hat{\jmath} \\ &+ \left(\mathbf{W}_W\mathbf{x}_Z + \mathbf{W}_X\mathbf{x}_Y - \mathbf{W}_Y\mathbf{x}_X + \mathbf{W}_Z\mathbf{x}_W\right)\hat{\kappa} \end{aligned} \tag{5}$$

3.2. Learning in the Quaternion Domain

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

$$\mathbf{y} = \alpha\left(\mathbf{W}\otimes\mathbf{x} + \mathbf{b}\right) \tag{6}$$

where $\mathbf{y}$ is the output of the layer, $\mathbf{b}$ is the quaternion-valued bias offset and α is a quaternion activation function. The choice of the activation function for the QCNN, as in the real- and complex-valued domains, needs to meet the property of differentiability. A suboptimal but suitable choice is represented by the *quaternion split activation function*, defined for a generic quaternion q as:

$$\alpha\left(q\right) = f\left(q_W\right) + f\left(q_X\right) + f\left(q_Y\right) + f\left(q_Z\right) \tag{7}$$

[1] We consider monodimensional signals for notational simplicity. As in the real case, everything extends immediately to multidimensional inputs.

being $f(\cdot)$ any standard activation function. In our case, we choose $f(\cdot)$ as a rectified linear unit (ReLU) activation function. The cost function to be optimized is a standard real-valued loss. In particular, in our case, we use a binary class-entropy loss for the SED task and a mean square error (MSE) loss for the localization task, as done in [26].

3.3. Weight initialization

The appropriate and correct initialization of the network parameters in the quaternion domain must take into account the interactions between quaternion-valued components, thus a simple random and component-wise initialization may result in an unsuitable choice [17]. Instead, a possible solution may be derived by considering a normalized purely quaternion $u^{\triangleleft}$ generated for each weight w by following a uniform distribution in $[0, 1]$. Each weight can be written in a polar form as:

$$w = |w| \, e^{u^{\triangleleft}\theta} = |w| \left(\cos(\theta) + u^{\triangleleft} \sin(\theta)\right), \tag{8}$$

from which it is possible to derive the quaternion-valued components of w:

$$\begin{cases} w_W = \phi \cos(\theta) \\ w_X = \phi u_X^{\triangleleft} \sin(\theta) \\ w_Y = \phi u_Y^{\triangleleft} \sin(\theta) \\ w_Z = \phi u_W^{\triangleleft} \sin(\theta) \end{cases} \tag{9}$$

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. The variance of the weight matrix can be defined as $\mathrm{var}(\mathbf{W}) = \mathrm{E}\{|\mathbf{W}|\} - (\mathrm{E}\{|\mathbf{W}|\})^2$, where the second term is null due to the symmetric distribution of the weight around 0 [17]. Since $\mathbf{W}$ follows a Chi distribution with four degrees of freedom, the variance can be expressed as:

$$\mathrm{var}(\mathbf{W}) = \mathrm{E}\left\{|\mathbf{W}|^2\right\} = \int_0^{\infty} w^2 f(w) \, \mathrm{d}\, w = 4\sigma^2 \tag{10}$$

being σ the standard deviation. Denoting with n_i the number of neurons of the input layer and considering the He criterion [28], σ can expressed as $\sigma = 1/\sqrt{2n_i}$ [17]. It follows that the variable ϕ in (9) can be randomly generated in the range $[-\sigma, \sigma]$.

3.4. Network Architecture

The model receives the quaternion ambisonic input, 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 [26]. Therefore, we obtain a feature sequence of T frames, with an overall dimension of $T \times M/2 \times 8$. The network has a similar architecture to the SELDnet [26], in which each input frame is mapped into two parallel outputs, the first one performs the sound event detection (SED), by predicting the active sound event class, and the second one estimates the direction of arrival (DOA) for the detected sound event by a multi-class regression.

In particular, each input frame is processed by the neural network in which the learning of the local shift-invariant features of the spectrogram is performed by using multiple layers of 2D QCNN. The QCNN layers are composed of P filter kernels with size $3 \times 3 \times 8$ and ReLU activation functions. At the output of the 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, which is then processed by a bidirectional recurrent neural network, as in the SELDnet, with the aim of learning the temporal information. Then, two branches of fully connected layers are used in parallel, one for each task. The first layer in both the branches involves R nodes with linear activation functions, while the last layer for the branch related to the SED task has N nodes, each one corresponding to a sound event class to be detected. A sigmoid function is used for multi-class detection, i.e., multiple sounds detected simultaneously. On the other hand, the last layer of the branch related to the localization task involves $3N$ nodes, representing the Cardinal coordinates for each sound event class, and hyperbolic tangent activation functions. As for the SELDnet, we use a cross-validation for the hyperparameter optimization. The network training involves a weighted combination of binary cross-entropy and MSE using Adam optimizer as also done in [26].

4. EXPERIMENTAL RESULTS

4.1. Datasets

We evaluate the proposed method involving the QCNN on two datasets involving 3D sound events in the Ambisonics format recorded in anechoic and reverberant environments. Both the datasets consider stationary sources associated with spatial coordinates.

The first dataset is the *Ambisonic, Anechoic and Synthetic Impulse Response* (ANSYN) dataset [22, 26], consisting of spatially located sound events in an anechoic scenario using simulated impulse responses. 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 Synthetic Impulse Response* (RESYN) dataset, similar to the ANSYN with the only difference that the environment is reverberant. Indeed, 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. More details on the datasets can be found in [22].

4.2. Metrics

The SELD task can use individual SED and localization metrics [26]. For the SED task, we use the polyphonic SED metrics that are the F-score (ideally $F = 1$), based on the number of true and false positives, and the error rate (ER) (ideally $ER = 0$), based on the total number of active sound event classes in the ground truth. A joint SED score can be considered as $S_{\text{SED}} = (ER + (1 - F))/2$.

On the other hand, a DOA estimation error DOA_{err} can be used as evaluation metric for localization task, based on estimated and ground truth DOAs [26]. 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 $S_{\text{DOA}} = (DOA_{\text{err}}/180 + (1 - K))/2$.

Finally, an overall SELD score can be defined based on the previous metrics as $S_{\text{SELD}} = (S_{\text{SED}} + S_{\text{DOA}})/2$.

4.3. Evaluation

We compare the proposed quaternion model with the SELDnet architecture [26] on the ANSYN and RESYN datasets. In order to provide a fair comparison, we use a configuration such to have a comparable number of parameters for both the models. In particular, we have about 760k parameters for the proposed quaternion network and about 530k parameters for the SELDnet, which is the most similar configuration possible considering the higher number of parameters generated by the QCNN, as described in Section 3. To this end, we set a number of $P = 64$ filters, sequence length of $T = 512$ frames, window length $M = 512$, batch size of 16, $Q = 128$ nodes for the recurrent networks and $R = 32$ nodes for the fully connected layers. The models have been trained over 1000 epochs[2].

Results for the ANSYN dataset are shown in Table 1. In terms of the overall SELD score, the proposed quaternion method clearly outperforms the standard SELDnet in each validation split and considering different overlapping sounds. In particular, it is worth noting from Table 1 that, while achieving better performance also in terms of localization score, the most significant part of the improvement is represented by the SED score, which is largely reduced with respect to the standard SELDnet.

[2]Experiments were run thanks to TensorFlow Research Cloud.

Table 1. Results on the ANSYN dataset in terms of the SED, DOA and overall SELD score. Best SELD scores in bold.

		SELDnet			Proposed Method		
Val. split		1	2	3	1	2	3
O1	S_{SED}	0.22	0.21	0.31	0.14	0.12	0.16
	S_{DOA}	0.20	0.21	0.21	0.12	0.10	0.10
	S_{SELD}	0.21	0.21	0.26	**0.13**	**0.11**	**0.13**
O2	S_{SED}	0.47	0.44	0.47	0.33	0.33	0.34
	S_{DOA}	0.35	0.34	0.33	0.29	0.29	0.30
	S_{SELD}	0.41	0.39	0.40	**0.31**	**0.31**	**0.32**
O3	S_{SED}	0.53	0.57	0.55	0.48	0.46	0.45
	S_{DOA}	0.47	0.45	0.45	0.42	0.40	0.41
	S_{SELD}	0.50	0.51	0.50	**0.45**	**0.43**	**0.43**

Table 2. Results on the RESYN dataset in terms of the SED, DOA and overall SELD score. Best SELD scores in bold.

		SELDnet			Proposed Method		
Val. split		1	2	3	1	2	3
O1	S_{SED}	0.22	0.24	0.30	0.23	0.22	0.29
	S_{DOA}	0.38	0.24	0.26	0.27	0.24	0.26
	S_{SELD}	0.30	0.24	0.28	**0.25**	**0.23**	**0.27**
O2	S_{SED}	0.57	0.54	0.61	0.47	0.40	0.46
	S_{DOA}	0.45	0.46	0.41	0.43	0.41	0.42
	S_{SELD}	0.51	0.50	0.51	**0.45**	**0.41**	**0.44**
O3	S_{SED}	0.64	0.59	0.57	0.51	0.53	0.55
	S_{DOA}	0.46	0.49	0.49	0.47	0.49	0.51
	S_{SELD}	0.55	0.54	0.53	**0.49**	**0.51**	**0.53**

Similar conclusions can be drawn also from the results achieved for the RESYN dataset and shown in Table 2. It can be noted that scores are slightly worse with respect to previous results due to reverberations. However, even in this case, the proposed quaternion method is able to improve both individual SED and DOA scores and the overall SELD performance.

5. CONCLUSION

In this paper we propose a SELDnet method involving a QCNN for the detection and the localization of 3D sound events captured by first-order Ambisonics. Ambisonic microphone signals are represented in their spherical harmonics form, which enables the processing in the quaternion domain. In particular, the convolution process of the neural network is performed in the quaternion domain, as well as the learning. Results are evaluated on the ANSYN and RESYN datasets and they have shown that, due to the processing in the quaternion domain, the proposed method is able to exploit the correlated nature of the ambisonic signals, thus providing improvements with respect to the standard SELDnet in terms of the simultaneous detection and localization scores.

6. REFERENCES

[1] J. Edwards, "Signal processing supports a new wave of audio research: Spatial and immersive audio mimics real-world sound environments," *IEEE Signal Process. Mag.*, vol. 35, no. 2, pp. 12–15, Mar. 2018.

[2] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Mag.*, vol. 28, pp. 20–32, Jan. 2011.

[3] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, "Intelligent acoustic interfaces for immersive audio," in *134th Audio Engineering Society Convention*, Rome, Italy, May 2013.

[4] D. Comminiello, S. Cecchi, M. Scarpiniti, M. Gasparini, L. Romoli, F. Piazza, and A. Uncini, "Intelligent acoustic interfaces with multisensor acquisition for immersive reproduction," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1262–1272, Aug. 2015.

[5] F. Ortolani, D. Comminiello, and A. Uncini, "The widely linear block quaternion least mean square algorithm for fast computation in 3D audio systems," in *26th IEEE Workshop on Machine Learning for Signal Process. (MLSP)*, Vietri sul Mare, Italy, Sept. 2016.

[6] F. Ortolani, D. Comminiello, M. Scarpiniti, and A. Uncini, "Advances in hypercomplex adaptive filtering for 3D audio processing," in *2017 IEEE First Ukraine Conf. on Elect. and Comput. Eng. (UKRCON)*, Kiev, Ukraine, May 2017, pp. 1125–1130.

[7] T. Mizoguchi and I. Yamada, "Hypercomplex tensor completion with Cayley-Dickson singular value decomposition," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 3979–3983.

[8] M. Xiang, S. Kanna, and D. P. Mandic, "Performance analysis of quaternion-valued adaptive filters in nonstationary environments," *IEEE Trans. Signal Process.*, vol. 66, no. 6, pp. 1566–1579, Mar. 2018.

[9] T. Ogunfunmi and C. Safarian, "A quaternion kernel minimum error entropy adaptive filter," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 4149–4153.

[10] Y. Xia, M. Xiang, Z. Li, and D. P. Mandic, *Adaptive Learning Methods for Nonlinear System Modeling*, chapter Echo State Networks for Multidimensional Data: Exploiting Noncircularity and Widely Linear, pp. 267–288, Elsevier, June 2018.

[11] F. Ortolani, D. Comminiello, M. Scarpiniti, and A. Uncini, "Frequency domain quaternion adaptive filters: Algorithms and convergence performance," *Signal Process.*, vol. 136, pp. 69–80, July 2017.

[12] L. Xiaodong, L. Aijun, Y. Changjun, and S. Fulin, "Widely linear quaternion unscented Kalman filter for quaternion-valued feedforward neural network," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1418–1422, Sept. 2017.

[13] S. Sanei, C. C. Took, and S. Enshaeifar, "Quaternion adaptive line enhancer based on singular spectrum analysis," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 2876–2880.

[14] X. Xiao and Y. Zhou, "Two-dimensional quaternion sparse principle component analysis," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 1528–1532.

[15] T. Minemoto, T. Isokawa, H. Nishimura, and N. Matsui, "Feed forward neural network with random quaternionic neurons," *Signal Process.*, vol. 136, pp. 59–68, July 2017.

[16] C. Gaudet and A. Maida, "Deep quaternion networks," in *IEEE Int. Joint Conf. on Neural Netw. (IJCNN)*, Rio de Janeiro, Brazil, July 2018.

[17] T. Parcollet, M. Ravanelli, M. Morchid, G. Linarès, C. Trabelsi, R. De Mori, and Y. Bengio, "Quaternion recurrent neural networks," *arXiv preprint arXiv:1806.04418v2*, July 2018.

[18] T. Parcollet, Y. Zhang, M. Morchid, C. Trabelsi, G. Linarès, R. De Mori, and Y. Bengio, "Quaternion convolutional neural networks for end-to-end automatic speech recognition," in *Interspeech 2018*, Hyderabad, India, Sept. 2018.

[19] 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.

[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] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial audio feature discovery with convolutional neural networks," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 6797–6801.

[22] 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, Sept. 2018, pp. 1476–1480.

[23] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *IEEE Int. Conf. on Acoust., Speech and signal Process. (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 771–775.

[24] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1291–1303, June 2017.

[25] I.-Y. Jeong, S. Lee, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network," in *Workshop on Detection and Classification of Acoust. Scenes and Events (DCASE)*, Munich, Germany, Nov. 2017.

[26] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *arXiv preprint arXiv:1807.00129v2*, July 2018.

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

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *IEEE Int. Conf. on Comput. Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1026–1034.