Efficient Sound Event Localization and Detection in the Quaternion Domain

Christian Brignone, Gioia Mancini, Eleonora Grassucci, *Graduate Student Member, IEEE*, Aurelio Uncini, *Member, IEEE*, and Danilo Comminiello, *Senior Member, IEEE*

***Abstract*—In recent years, several approaches have been proposed for the task of Sound Event Localization and Detection (SELD) with multiple overlapping sound events in the 3D sound field. However, accuracy improvements have been often achieved at the expense of more complex networks and a larger number of parameters. In this brief, we propose an efficient and lightweight Quaternion Temporal Convolutional Network for the SELD task (QSELD-TCN), which combines the advantages of the quaternion-valued processing and the effectiveness of the Temporal Convolutional Network (TCN). The proposed approach involves a representation of the Ambisonic signal components as a single quaternion and, accordingly, the use of quaternion-valued layers through the whole structure of the neural network. This results in a considerable saving of parameters with respect to the corresponding real-valued model. In particular, a quaternion implementation of the TCN block is presented, exploiting TCN ability in capturing long-term dependencies and the effectiveness of quaternion convolutional layers in grasping correlations among input dimensions. The proposed approach implies less runtime memory and lower storage memory, and it achieves faster inference time with respect to the state-of-the-art methods, making its implementation possible even in devices with limited resources.**

***Index Terms*—Quaternion neural networks, sound event localization and detection, efficient neural networks, quaternion domain, lightweight neural networks.**

I. Introduction

NOWADAYS, multichannel audio techniques and approaches are spreading in many fields of applications, such as voice assistance, home automation, security and surveillance, virtual reality and autonomous robotic systems [1]–[5], where the acoustic information of the environment could be a crucial aspect. With the recent growth of the 3D audio industry [6], the Ambisonic microphone array [7] became an important tool for the 3D estimation of the direction of arrival problem as well. Here, we focus our attention on the Sound Event Localization and Detection (SELD) task [8]–[11], which consists of identifying jointly the temporal and spatial location of a sound source and the class to which it belongs.

Manuscript received February 9, 2022; accepted March 8, 2022. Date of publication March 17, 2022; date of current version May 3, 2022. This brief was recommended by Associate Editor D. Linaro. *(Christian Brignone and Gioia Mancini are contributed equally to this work.) (Corresponding author: Eleonora Grassucci.)*

The authors are with the Department of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome, 00185 Rome, Italy (e-mail: eleonora.grassucci@uniroma1.it).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2022.3160388.

Digital Object Identifier 10.1109/TCSII.2022.3160388

Although convolutional recurrent neural networks (CRNNs) have been successfully employed for tasks like sound event detection (SED) and direction of arrival (DOA) estimation [12], [13], their implementation on embedded devices can be very arduous, especially for memory and power requirements. Convolutions are powerful to extract features from spectrograms while recurrent units are crucial to encapsulate information through time. However, while convolutional layers are highly parallelizable, their receptive field is limited to the neighbors of the considered data point. Temporal convolutional networks (TCNs) have been successfully proposed as an alternative to CRNNs, due to their ability in capturing long-term dependencies while allowing parallel processing of the input sequence [14], [15]. However, TCNs introduce many more parameters than CRNNs, thus requiring much more computational resources. In this brief, we attempt to reduce the complexity of such models by exploiting the properties of quaternion algebra.

In the recent years, quaternion neural networks (QNNs) have been widely employed for processing multidimensional input due to the quaternion algebra properties that allow to grasp correlations and relations among input dimensions [16]–[21]. Thanks to the Hamilton product, which regulates the vector multiplication between two quaternions by modelling the interactions among the imaginary units, QNNs adopt $1/4$ of the number of parameters of their real-valued counterpart. Firstly, the parameters reduction results in efficient models in terms of computational time and required memory [22]–[25]. Secondly, the parameters sharing due to the Hamilton product allows the model to perceive latent information among dimensions, leading to overall better performances [26], [27]. For these reasons, QNNs are particularly appropriate to deal with Ambisonic signals that show strongly correlated components [25], [28]–[31].

In this brief, we propose a novel neural model for the SELD task which leverages the quaternion algebra properties. In fact, our neural network architecture combines the advantages of the TCN structure and the ones of the quaternion domain. The proposed method encapsulates the Ambisonic signal components in a quaternion and then pass it to a series of quaternion-valued operation blocks. This results in a lightweight and more efficient model with respect to its real-valued counterpart, allowing us to save storage memory for inference and to reduce the required computational time. We show that, besides obtaining a dramatic reduction of the computational complexity, our approach also achieves state-of-the-art results in the SELD tasks, according to different accuracy metrics.

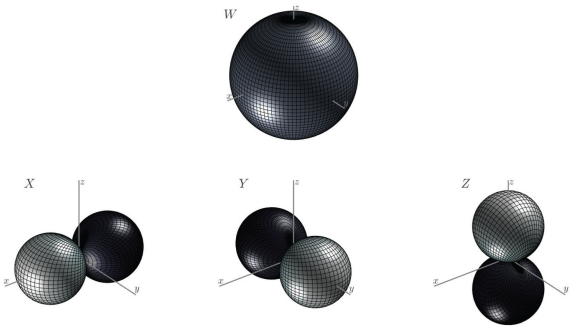

Fig. 1. Spherical harmonics representation of first-order Ambisonics.

The rest of the brief is organized as follows. In Section II, the theoretical background concerning Ambisonic signals and quaternion algebra is introduced. In Section III, the model architecture and the pre-processing are expounded. Section IV reports the experimental setup and the evaluation results. Finally, in Section V, we draw the conclusions.

II. Ambisonic Signals in the Quaternion Domain

The sound field can be described through the method of Spherical Harmonics Decomposition (SHD) [32]. One of the most popular methods for sampling the spatial sound field based on SHD is the Ambisonics technique. The Ambisonics systems are arrays of ideally coincident microphones, with a number of capsules depending on the order of the spherical components considered. Here, we consider the B-format First-Order Ambisonics (FOA), which is composed of four microphone capsules: an omnidirectional microphone (channel denoted by W), which captures the pressure field, and three figure-of-eight microphones (channels denoted by X, Y, Z) that capture the acoustic velocity information. A graphical representation of the spherical harmonic decomposition of the FOA is depicted in Fig. 1.

The FOA SHD of a sound pressure signal $s[n]$, with angles θ, ϕ, is defined as:

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

which describes the discrete-time representation of the four ambisonic components [30], [31]. These components, can be easily encapsulated in a single quaternion, as explained in the following.

Quaternions are a family of number systems that extend the complex-valued domain. A quaternion is composed of four components, a real one and three imaginary ones:

$$q = q_W + q_X\hat{\imath} + q_Y\hat{\jmath} + q_Z\hat{\kappa}, \tag{2}$$

whereby q_W, q_X, q_Y, q_Z are real coefficients and the imaginary units $\hat{\imath} = (1,0,0)$, $\hat{\jmath} = (0,1,0)$ and $\hat{\kappa} = (0,0,1)$ represent orthonormal basis in $\mathbb{R}^3$ with the properties $\hat{\imath}\hat{\jmath} = \hat{\imath} \times \hat{\jmath} = \hat{\kappa}$, $\hat{\jmath}\hat{\kappa} = \hat{\jmath} \times \hat{\kappa} = \hat{\imath}$, $\hat{\kappa}\hat{\imath} = \hat{\kappa} \times \hat{\imath} = \hat{\jmath}$, $\hat{\imath}^2 = \hat{\jmath}^2 = \hat{\kappa}^2 = -1$, and with the non-commutative product:

$$\hat{\imath}\hat{\jmath} = -\hat{\jmath}\hat{\imath}, \quad \hat{\jmath}\hat{\kappa} = -\hat{\kappa}\hat{\jmath}, \quad \hat{\kappa}\hat{\imath} = -\hat{\imath}\hat{\kappa}. \tag{3}$$

Accordingly, the representation of the Ambisonics signals as a single quaternion-valued Ambisonic signal is given by

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

in which the omnidirectional microphone signal $x_W[n]$ is the real part of the quaternion and the three signals $x_X[n]$, $x_Y[n]$, $x_Z[n]$ are the imaginary parts of the quaternion [30].

III. The Proposed QSELD-TCN Model

The structure of the proposed model is similar to the SELD-TCN [15], but completely defined in the quaternion domain, thus involving new layers in the quaternion domain, and quaternion-valued weights and inputs.

A. Pre-Processing

The quaternion ambisonic input is passed to the feature extractor, in which the spectrogram of the signal is extracted by performing the short-time Fourier transform (STFT) on each of the four channels, with an Hann window of length M. Then, channels magnitude and phase are stacked together obtaining a feature sequence of T frames, so the input dimension is $T \times M/2 \times 8$.

B. Quaternion-Valued Convolution

Due to the non-commutativity of quaternion vector multiplications in (3), the Hamilton product has been introduced to wisely model interactions among components [33]. The Hamilton product is at the core of quaternion layers, including convolutional ones. In this way, the internal latent relations within the features of a quaternion can be perceived by the model. Considering a quaternion filter matrix $\mathbf{W} = \mathbf{W}_W + \mathbf{W}_X\hat{\imath} + \mathbf{W}_Y\hat{\jmath} + \mathbf{W}_Z\hat{\kappa}$ and given an input quaternion vector $\mathbf{x}$ with the same form, the quaternion convolution is expressed as:

$$\mathbf{W} * \mathbf{x} = \begin{bmatrix} \mathbf{W}_W - \mathbf{W}_X - \mathbf{W}_Y - \mathbf{W}_Z \\ \mathbf{W}_X + \mathbf{W}_W - \mathbf{W}_Z + \mathbf{W}_Y \\ \mathbf{W}_Y + \mathbf{W}_Z + \mathbf{W}_W - \mathbf{W}_X \\ \mathbf{W}_Z - \mathbf{W}_Y + \mathbf{W}_X + \mathbf{W}_W \end{bmatrix} * \begin{bmatrix} \mathbf{x}_W \\ \mathbf{x}_X \\ \mathbf{x}_Y \\ \mathbf{x}_Z \end{bmatrix}. \tag{5}$$

As a consequence, the weight matrix can be written as a composition of smaller submatrices with a significant reduction of parameters to train [17].

C. Model Architecture

The model architecture can be summarized in three main parts: a series of three quaternion convolutional (Q-Conv2D) blocks, a quaternion temporal convolutional (Q-TCN) block and two parallel branches of quaternion fully-connected blocks (Q-FC), as Fig. 2 shows. In each of the three Q-Conv2D blocks, a 2D quaternion convolution is applied to the input feature with P filters of *kernel size* 3×3, followed by a quaternion split rectified linear unit (*ReLU*) activation function [30]. Then, 2D max-pooling is applied to the frequency axis before the final dropout, with a *pool_size* $= (1, mp_i)$, where mp_i is the i-th component of $mp = [mp_1, mp_2, mp_3]$. After that, the output needs to be manipulated in order to be passed to the Q-TCN block. In particular, this can be achieved by stacking the frequency axis together with the channels axis, resulting in a final dimension equal to $T \times L$,

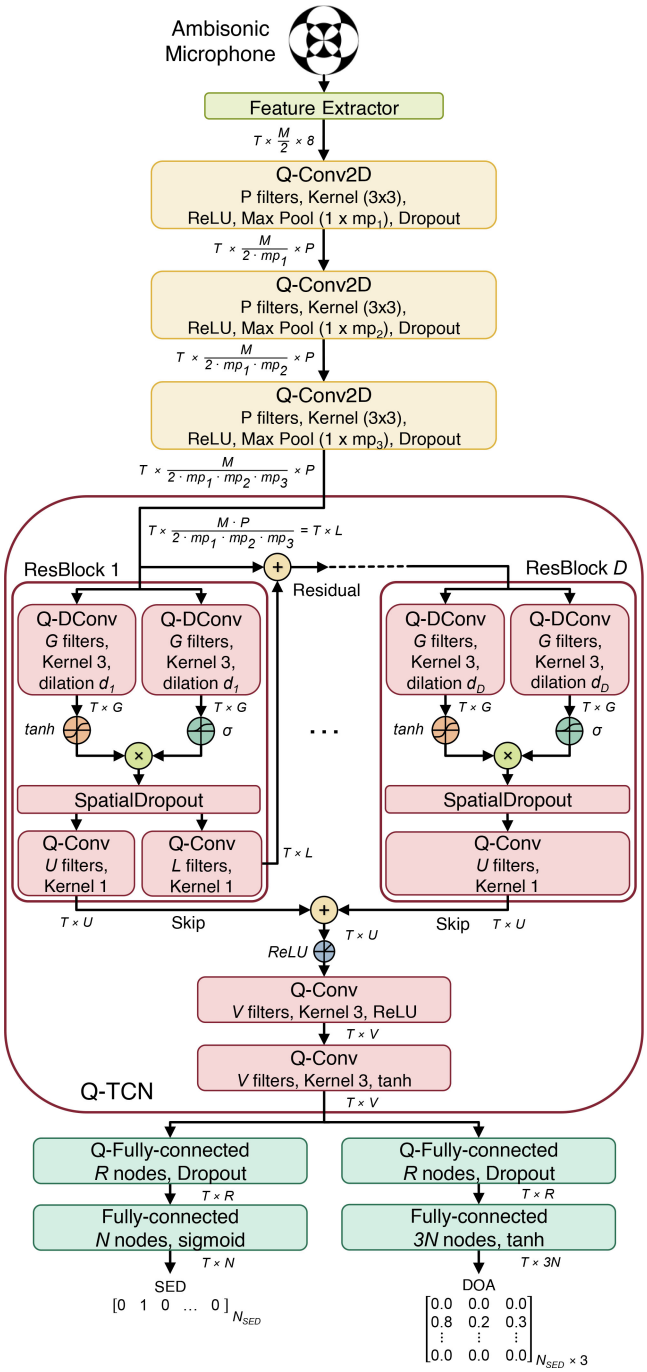

Fig. 2. QSELD-TCN architecture. In the blocks, we set $P = 64$, $G = 128$, $U = 128$, $V = 128$ and $R = 128$. Also, the pooling sizes are $mp_1 = 8$, $mp_2 = 8$, $mp_3 = 2$. The Q-TCN block is composed of $D = 10$ resblocks and the dilation rate is $d = [2^0, \ldots, 2^9]$.

with $L = (M \cdot P)/(2 \cdot mp_1 \cdot mp_2 \cdot mp_3)$. Since, even in the real-time applications, audio is captured in frames, when a specific sample is analyzed, not only previous but also future samples are available. Therefore, non-causal dilated convolutions are used, allowing to enlarge the receptive field considering also future knowledge rather than only present and past samples [15], [34], [35]. The Q-TCN block is characterized by *D resblocks*, similarly to the SELD-TCN [15]. Inspired by WaveNet [35], we do not employ batch normalization and we use spatial dropout. Each resblock contains a gating mechanism called Gated Tanh Unit (GTU) [36], [37]:

$$y = tanh(\mathbf{W}_f * \mathbf{x}) \odot \sigma(\mathbf{W}_g * \mathbf{x}), \tag{6}$$

where W_f and W_g represent, respectively, the convolutional filters of the 1D quaternion dilated convolution (Q-DConv) associated to the *tanh* activation and the one of the Q-DConv associated to the sigmoidal gate. These Q-DConvs have G filters of *kernel size* 3 and an exponentially increasing dilation rate d_i, i-th element of $d = [2^0, \ldots, 2^{D-1}]$. The output is then passed through a spatial dropout, followed by two parallel 1D quaternion convolutional (Q-Conv) layers, one for the skip and one for the residual connection, both with *kernel size* 1 and with a number of filters equal to U and L, respectively. The number of filters of the latter Q-Conv has to be equal to L in order to bring back the dimension to $T \times L$, since each resblock is fed with the sum between the input and the output of the previous resblock through the residual connection.

Once the resblocks computation is completed, the skip connections are summed together, and a *ReLU* activation function is applied. Next, there are two 1D quaternion convolutions with V filters and *kernel size* 3, each followed respectively by a *ReLU* and a *tanh* activation functions.

The dimension of the receptive field RF of a Q-TCN block with D resblocks, each performing dilated convolutions with exponentially increasing dilation rate and *kernel size* k, can be computed as:

$$RF = 1 + (k-1)\sum_{i=0}^{D-1} 2^i. \tag{7}$$

Finally, the output of the Q-TCN block flows into two parallel branches of fully-connected (FC) blocks, that solve the SED task and DOA estimation respectively. The first Q-FC layer in the SED branch has R nodes with *linear* activation function and dropout. The second FC has a number of neurons equal to the number of classes N with a *sigmoid* activation function for multi-class detection, in order to have the output in $[0, 1]$. Instead, in the DOA branch we have a Q-FC layer with R nodes, a *linear* activation function and dropout. The following FC layer has $3N$ nodes, each of which represents the 3D Cartesian coordinates, followed by a *tanh* activation function to produce an output in the range $[-1, 1]$.

IV. Experimental Evaluation

A. Dataset

We consider the *TAU Spatial Sound Events 2019 - Ambisonic dataset* [38], [39] to conduct the experimental evaluation of the proposed model. It is composed of 4-channel FOA one minute long recordings, consisting of stationary point sources from multiple sound classes, each associated with a temporal onset and offset time and DOA spatial coordinates. The development set consists in four cross-validation splits made up of 100 recordings sampled at 48000 Hz. Spatial room impulse responses (IRs) are captured from five indoor locations using an Eigenmike spherical microphone array, at 504 combinations of azimuth and elevation angles and distance. Then, they are convolved with *Isolated Sound Events Dataset* from *DCASE 2016 task 2* [40], which is balanced and consists of 11 classes. At the end, the dataset contains 100 recordings in each of the four splits, with a maximum of two temporally

TABLE I
RESULTS OF SELD-TCN AND QSELD-TCN OVER OV1 AND OV2 TEST SPLITS. THE PROPOSED METHOD OUTPERFORMS THE BASELINE ON THE AVERAGE $\text{SELD}_{\text{SCORE}}$ WHICH MEASURES THE ACCURACY OF THE OVERALL MODEL PREDICTION

	Overlap 1				
Model	F ↑	ER ↓	$\text{SED}_{\text{score}}$ ↓	$\text{DOA}_{\text{score}}$ ↓	$\text{SELD}_{\text{score}}$ ↓
SELD-TCN	0.865	0.237	0.186	**0.047**	0.117
QSELD-TCN	**0.893**	**0.171**	**0.139**	0.052	**0.095**
	Overlap 2				
Model	F ↑	ER ↓	$\text{SED}_{\text{score}}$ ↓	$\text{DOA}_{\text{score}}$ ↓	$\text{SELD}_{\text{score}}$ ↓
SELD-TCN	0.843	0.225	0.191	**0.128**	0.159
QSELD-TCN	**0.850**	**0.179**	**0.164**	0.145	**0.155**
	Average				
Model	F ↑	ER ↓	$\text{SED}_{\text{score}}$ ↓	$\text{DOA}_{\text{score}}$ ↓	$\text{SELD}_{\text{score}}$ ↓
SELD-TCN	0.854	0.231	0.189	**0.087**	0.138
QSELD-TCN	**0.872**	**0.175**	**0.152**	0.098	**0.125**

TABLE II
MODELS COMPARISON CONCERNING THE MEAN INFERENCE TIME, THE STORAGE MEMORY ON DISK, THE MAXIMUM BATCH SIZE ALLOWED FOR EACH MODEL (THAT IS, HOW MANY DATA POINTS CAN BE ALLOCATED IN THE GPU MEMORY) AND THE TOTAL NUMBER OF PARAMETERS. IN THE BOTTOM, WE REPORT THE INFERENCE DELAY AND THE EXCESS OF MEMORY AND PARAMETERS OF THE BASELINE METHODS WITH RESPECT TO OUR METHOD QSELD-TCN

Model	Inference Time (s)	Memory (KB)	Max Batch Size	# Parameters (M)
SELDnet	0.8244	2068	20	0.51
QSELDnet	1.2565	**1166**	21	**0.28**
SELD-TCN	0.0199	6246	21	1.52
QSELD-TCN	**0.0176**	1719	**51**	0.39
Summary of the proposed QSELD-TCN improvements				
Model	Inference Delay	Storage Memory	Runtime Memory	# Parameters
SELDnet	46.84×	+20%	+155%	+32%
QSELDnet	71.39×	-32%	+143%	-28%
SELD-TCN	1.13×	+263%	+143%	+293%

overlapping sound events (Overlap 1 and 2, which will be referred to as OV1 and OV2 in the following).

B. Experimental Setup

In the pre-processing we employ a Hann window of dimension $M = 512$ and a sequence length of dimension $T = 512$. Therefore, the input features have dimension equal to $512 \times 256 \times 8$ and the outputs dimensions are 512×11 for SED and 512×33 for DOA, where $N = 11$ is the number of classes. The number of filters and nodes is chosen to maintain the same data flow of the SELD-TCN architecture and so, we choose the following number of filters for QSELD-TCN: $P = 64$, $G = 128$, $U = 128$, $V = 128$, $R = 128$. The max pooling is $mp = [8, 8, 2]$, the dropout rate and spatial dropout rate are 0.2 and 0.5 respectively, the number of resblocks is $D = 10$ and the receptive field is $RF = 2047$ for both the models. In this setting, we perform the training of the model for a maximum of 6000 epochs with *early stopping* using Fold 1, being the one in [39] over which the best overall performances are obtained with their baseline. The initial learning rate for the Adam optimizer is set equal to $lr = 0.0001$. The model optimization involves a weighted combination of binary cross-entropy for SED output and mean squared error for DOA output, weighting the latter 50 times more than the former. In order to compare the results with the baseline SELD-TCN architecture [15], the same settings are used for this model too.

C. Results

To evaluate the performance of the proposed approach, we consider the metrics described in [8], so the F score (F), error rate (ER), DOA error (DE) and frame recall (FR). The SED score is defined as $\text{SED}_{\text{score}} = (\text{ER} + (1 - \text{F}))/2$, while the DOA score is $\text{DOA}_{\text{score}} = (\text{DE}/180 + (1 - \text{FR}))/2$. At the end, the overall SELD score based on the previous metrics is expressed as $\text{SELD}_{\text{score}} = (\text{SED}_{\text{score}} + \text{DOA}_{\text{score}})/2$. The computed metrics for the baseline SELD-TCN and our method QSELD-TCN on test splits are summarized in detail in Table I. The first two row blocks report the metrics computed over the OV1 and OV2 test sets, while the last one shows their average values. It is evident that the proposed method outperforms SELD-TCN in most of the metrics. Most important, QSELD-TCN leads to better results in terms of mean SELD score which is the most relevant metric for SELD task evaluation. Our approach far exceeds the real-valued baseline despite the lower number of parameters, thus proving the effectiveness of the quaternion layers in capturing correlations among channels.

Furthermore, in Table II, we provide some interesting measurements that put in light the efficiency of the considered neural networks. For the baselines SELDnet [8], QSELDnet [30], SELD-TCN [15] and for our approach, we report the mean inference time (averaged over 1000 iterations) and the storage memory required by the model. Moreover, we also include the maximum batch size allowed in GPU memory for each model, since it is an indicator of the amount of memory utilized by the neural network at runtime (the larger is the maximum batch size, the lower is the memory required). This computation is performed on a NVIDIA RTX 2060 with 6GB. Finally, the last column reports the total number of parameters. From the efficiency point of view, the proposed QSELD-TCN is faster in inference with respect to all the other models considered. As well, our method can fit in GPU memory the maximum batch size. In fact, the QSELD-TCN is 1.13× faster than SELD-TCN, which utilizes 143% more runtime memory and occupies 263% more disk space than QSELD-TCN. Furthermore, comparing our approach with SELDnet, it results to be 46.84× faster. Moreover, SELDnet uses 155% more runtime memory and it occupies 20% more memory on disk. While the QSELDnet has the lower impact on storage memory, our method is 71.39× faster than it, which also requires 143% more runtime memory than our approach. It worth noting that the proposed approach has a higher parameters reduction (−75%) on its real-valued baseline SELD-TCN with respect to the one of the QSELDnet (−45%) with respect to its corresponding baseline SELDnet, due to the lightweight Q-TCN modules.

The quaternion implementation of the TCN block gives a significant saving of parameters and, despite this, QSELD-TCN proves to be still able to achieve better overall performances. These efficiency performances underline that our approach may be more suitable for embedded applications with respect to the heavier real-valued models.

V. Conclusion

In this brief, we propose a Quaternion Temporal Convolutional network to perform efficient Sound Event Localization and Detection (SELD) of 3D sound events acquired through first-order Ambisonics. The quaternion domain representation of Ambisonic signals processed with the proposed quaternion-valued model allows us to capture the internal latent relations between the signal components while reducing the number of parameters by 75%. Indeed, our approach outperforms the real-valued baseline in terms of SELD scores while being more efficient for inference in terms of time and memory requirements. Future works will involve a generalization of the proposed quaternion-valued approach in the hypercomplex domain, being able to involve even higher-order Ambisonics.

References

[1] J.-H. Kim, J.-H. Choi, J. Son, G.-S. Kim, J. Park, and J.-H. Chang, "MIMO noise suppression preserving spatial cues for sound source localization in mobile robot," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2021, pp. 1–5.

[2] H. A. Sánchez-Hevia, R. Gil-Pita, and M. Rosa-Zurera, "Efficient multichannel detection of impulsive audio events for wireless networks," *Appl. Acoust.*, vol. 179, Aug. 2021, Art. no. 108005.

[3] H. Liu, Y. Sun, Y. Li, and B. Yang, "3D audio-visual speaker tracking with a novel particle filter," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 7343–7348.

[4] Y. Gong, J. Yang, and C. Poellabauer, "Detecting replay attacks using multi-channel audio: A neural network-based method," *IEEE Signal Process. Lett.*, vol. 27, pp. 920–924, 2020.

[5] M. Narbutt, S. O'Leary, A. Allen, J. Skoglund, and A. Hines, "Streaming VR for immersion: Quality aspects of compressed spatial audio," in *Proc. Int. Conf. Virtual Syst. Multimedia (VSMM)*, 2017, pp. 1–6.

[6] 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, Aug. 2015.

[7] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," *J. Audio Eng. Soc.*, to be published.

[8] 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.

[9] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by DNN-based denoising and source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 651–655.

[10] T. Komatsu, M. Togami, and T. Takahashi, "Sound event localization and detection using convolutional recurrent neural networks and gated linear units," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 41–45.

[11] E. Guizzo *et al.*, "L3DAS22 challenge: Learning 3D audio sources in a real office environment," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Singapore, Jun. 2022.

[12] L. Perotin, R. Serizel, E. Vincent, and A. Guarin, "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.

[13] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Proc. Detect. Classification Acoust. Scenes Events Workshop (DCASE)*, 2019, pp. 119–123.

[14] C. Lea, M. Flynn, R. Vidal, A. Reiter, and G. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.

[15] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, "SELD-TCN: Sound event localization & detection via temporal convolutional networks," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 16–20.

[16] C. J. Gaudet and A. S. Maida, "Deep quaternion networks," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Rio de Janeiro, Brazil, Jul. 2018, pp. 1–8.

[17] T. Parcollet, M. Morchid, and G. Linarès, "A survey of quaternion neural networks," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2957–2982, Aug. 2019.

[18] T. Parcollet *et al.*, "Quaternion recurrent neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019, pp. 1–19.

[19] R. Vecchi, S. Scardapane, D. Comminiello, and A. Uncini, "Compressing deep-quaternion neural networks with targeted regularisation," *CAAI Trans. Intell. Technol.*, vol. 5, no. 3, pp. 172–176, 2020.

[20] E. Grassucci, D. Comminiello, and A. Uncini, "An information-theoretic perspective on proper quaternion variational autoencoders," *Entropy*, vol. 23, no. 7, p. 856, 2021.

[21] E. Grassucci, D. Comminiello, and A. Uncini, "A quaternion-valued variational autoencoder," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 3310–3314.

[22] 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, Jul. 2017.

[23] Y. Tay *et al.*, "Lightweight and efficient neural natural language processing with quaternion networks," in *Proc. Assoc. Comput. Linguist. (ACL)*, 2019, pp. 1494–1503.

[24] A. Zhang *et al.*, "Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters," in *Proc. Int. Conf. Mach. Learn (ICML)*, 2021.

[25] E. Grassucci, A. Zhang, and D. Comminiello, "PHNNs: Lightweight neural networks via parameterized hypercomplex convolutions," 2021, *arXiv:2110.04176*.

[26] T. Parcollet, M. Morchid, and G. Linarès, "Quaternion convolutional neural networks for heterogeneous image processing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 8514–8518.

[27] E. Grassucci, E. Cicero, and D. Comminiello, "Quaternion generative adversarial networks," 2021, *arXiv:2104.09630v1*.

[28] F. Ortolani, D. Comminiello, and A. Uncini, "The widely linear block quaternion least mean square algorithm for fast computation in 3D audio systems," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2016, pp. 1–6.

[29] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, "Frequency-domain adaptive filtering: From real to hypercomplex signal processing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brighton, U.K., 2019, pp. 7745–7749.

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

[31] M. R. Celsi, S. Scardapane, and D. Comminiello, "Quaternion neural networks for 3D sound source localization in reverberant environments," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2020, pp. 1–6.

[32] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer Topics in Signal Processing). vol. 2. Cham, Switzerland: Springer, 2015.

[33] J. P. Ward, *Quaternions and Caley Numbers. Algebra and Applications* (Mathematics and Its Applications). Boston, MA, USA: Kluwer, 1997, vol. 403.

[34] A. Van Den Oord *et al.*, "Wavenet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[35] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5069–5073.

[36] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, Dec. 2016, pp. 933–941.

[37] A. Van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. 30th Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1–9.

[38] S. Adavanne, A. Politis, and T. Virtanen. (2019). *TAU Spatial Sound Events 2019—Ambisonic and Microphone Array, Development Datasets*. [Online]. Available: https://doi.org/10.5281/zenodo.2580091

[39] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc. Detect. Classification Acoust. Scenes Events (DCASE)*, Jan. 2019, pp. 10–14.

[40] A. Mesaros *et al.* (2017). *DCASE2016 Challenge Submissions Package*. [Online]. Available: https://doi.org/10.5281/zenodo.926660