IMPROVEMENT OF DOA ESTIMATION BY USING QUATERNION OUTPUT IN SOUND EVENT LOCALIZATION AND DETECTION

Yui Sudo, Katsutoshi Itoyama, Kenji Nishida

Department of Systems and Control Engineering,
School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo
1528552, Japan
{sudo, itoyama, nishida}@ra.sc.e.titech.ac.jp

Kazuhiro Nakadai

Department of Systems and Control Engineering,
School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo
1528552, Japan
Honda Research Institute Japan Co., Ltd., 8-1
Honcho, Wako, Saitama 351-0188, Japan
nakadai@jp.honda-ri.com

ABSTRACT

This paper describes improvement of Direction of Arrival (DOA) estimation performance using quaternion output in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Task 3. DCASE 2019 Task3 focuses on the sound event localization and detection (SELD) which is a task that simultaneously estimates the sound source direction in addition to conventional sound event detection (SED). In the baseline method, the sound source direction angle is directly regressed. However, the angle is a periodic function and it has discontinuities which may make learning unstable. Specifically, even though -180 deg and 180 deg are in the same direction, a large loss is calculated. Estimating DOA angles with a classification approach instead of regression can solve such instability of discontinuities but this causes limitation of resolution. In this paper, we propose to introduce the quaternion which is a continuous function into the output layer of the neural network instead of directly estimating the sound source direction angle. This method can be easily implemented only by changing the output of the existing neural network, and thus does not significantly increase the number of parameters in the middle layers. Experimental results show that proposed method improves the DOA estimation without significantly increasing the number of parameters.

Index Terms— Sound event localization and detection, direction of arrival, inter-channel phase difference, quaternion, convolutional recurrent neural networks

1. INTRODUCTION

Sound event detection (SED) is a rapidly developing research area that aims to analyze and recognize a variety of sounds in urban and natural environments. Compared to audio tagging, event detection also involves estimating the time of occurrence of sounds. Automatic recognition of sound events would have a major impact in several applications. For example, SED has been drawing a surging amount of interest in recent years with applications including audio surveillance [1], healthcare monitoring [2], urban sound analysis [3], multimedia event detection [4] and bird call detection [5]. In an actual application, more convenient application can be realized by simultaneously performing sound source localization as well as detection of a sound event occurrence interval. For example, in the case of an audio surveillance, it is useful to detect the direction of the anomalous sound. Alternatively, individual sound events can be identified even when events of the same class are overlapped. Also, regarding the detection of overlapping sound events, more rational detection is possible by using spatial information.

Task 3 of the DCASE 2019 Challenge focuses on locating and detecting sound events (SELD) for overlapping sound sources [6]. A recently developed system called SELDnet was used as a baseline system. SELDnet uses magnitude spectrograms and phase spectrograms as input features to jointly train SED and DOA estimation purposes [7]. Regarding input features, it has been reported that simply using sinIPD and cosIPD (inter-channel phase difference) as input features for the neural network improves the performance in speech separation [8]. Meanwhile, DOA angle has been directly predicted in many research. During training, the difference between the correct angle and the estimated angle is calculated as a loss. Since the angle is a periodic function, it has discontinuities. Specifically, even though -180 deg and 180 deg are in the same direction, a large loss is calculated, which may make learning unstable. Regarding the discontinuity problem in rotation angle estimation, camera pose regression has been proposed that estimates camera position and orientation by using quaternion in computer vision [9-12].

This paper proposes a model that replaces input features of baseline system and DOA output with sinIPD, cosIPD and quaternion respectively. Since this method can be implemented without changing the middle layer of the network, it is easy to implement with almost no increase in the number of parameters of the existing model. Details of the proposed method are explained in the following sections.

2. METHOD

The entire network of Sound event localization and detection is shown in Fig. 1. The time-series sound source is input to the convolutional recurrent neural network (CRNN) [13] after feature extraction block. The CRNN is consists of three blocks, including three layers of convolutional neural network (CNN), two layers of bi-directional recurrent neural network (RNN) and two fully connected layers. There are two branches throughout the joint layer block. One is for SED, and the other is for DOA estimation.

https://doi.org/10.33682/jj50-hm12

The difference from the baseline is that the source direction angle is not directly estimated but through regressing quaternions. The estimated quaternions are converted to source direction angles by post-processing. Details of feature extraction, network, and post-processing is described in the following sections.

2.1. Feature extraction

The input to this method is multi-channel audio with a sampling rate of 48 kHz. At first, short time Fourier transformation (STFT) is applied using a 40 ms long Hanning window and M points (M=2048) from 20 ms hop length. Then, for each STFT obtained, select a reference microphone, p, and other non-reference microphones, q. As a spectral feature, an amplitude spectrogram of only the reference microphone is used. Meanwhile, we use the following equations to extract spatial features,

$$\cos IPD(t,f,p,q) = \cos\left(\theta_{t,f,p,q}\right), \quad (1)$$

$$\sin IPD(t,f,p,q) = \sin\left(\theta_{t,f,p,q}\right), \quad (2)$$

where $\theta_{t,f,p,q} = \angle x_{t,f,p} - \angle x_{t,f,q}$ is the phase difference between the STFT coefficients $x_{t,f,p}$ and $x_{t,f,q}$ at time t and frequency f of the signals at microphones p and q. In this paper, 6-channel sinIPD and cosIPD are used as spatial features for IPD of three combinations (1ch-2ch, 1ch-3ch, 1ch-4ch). That is, a total of 7-channel features consisting of one amplitude spectrogram of the reference microphone and the 6-channel spatial features are input to the neural network.

2.2. Network architecture

In order to verify the effect of quaternion estimation, basically the same network as the baseline system shown in Fig. 1 is used. A sequence of T spectrogram frames (T = 128), extracted in the feature extraction block, is fed to the three convolutional layers that extract shift-invariant features using P filters each (P=64). Batch normalization is used after each convolutional layers. Dimensionality reduction of the input spectrogram feature is performed using max pooling operation only along the frequency axis, which is called frequency pooling in [13]. The temporal axis is untouched to keep the resolution of the output unchanged from the input dimension. The temporal structure of the sound events is modeled using two bi-directional recurrent layers with Q gated recurrent units (GRU) each (Q=128). Finally, the output of the recurrent layer is shared between two fully connected layer (FC) branches each producing the SED as multiclass multilabel classification and DOA as multi-output regression; together producing the SELD output. The first FC layer contains R nodes each with linear activation. The SED output obtained is the class-wise probabilities for the C classes in the dataset at each of the T frames of input spectrogram sequence, resulting in a dimension of $T \times C$. The localization output estimates, for each time frame T, quaternions representing rotation in the azimuth direction and elevation direction for each of the C classes i.e., if multiple instances of the same sound class occur in a time frame the SELDnet localizes either one or oscillates between multiple instances. The overall dimension of localization output is $T \times 4C$, where $4C$ represents the class-wise $\sin(\theta_{azimth})$, $\cos(\theta_{azimuth})$ and $\sin(\theta_{elevation})$, $\cos(\theta_{elevation})$, which describes quaternion around z-axis and y-axis. Note that $\theta_{azimuth}$ and $\theta_{elevation}$ do not represent the phase of STFT but represent the sound source direction angles of azimuth and elevation. A sound event class is said to be active if its probability in SED output is greater than the threshold of 0.5, otherwise, the sound class is considered to be inactive. The presence of sound class in consecutive time frames gives the onset and offset times, and the corresponding DOA estimates from the localization output gives the spatial location with respect to time. A crossentropy loss is employed for detection output, while a mean square error loss on the quaternion distance between reference and estimated locations is employed for the localization output. The combined convolutional recurrent neural network architecture is trained using Adam optimizer and a weighted combination of the two output losses. Specifically, the localization output is weighted ×50 more than the detection output same as the baseline. The number of parameters is 615,799 while baseline has 613,537. This method does not

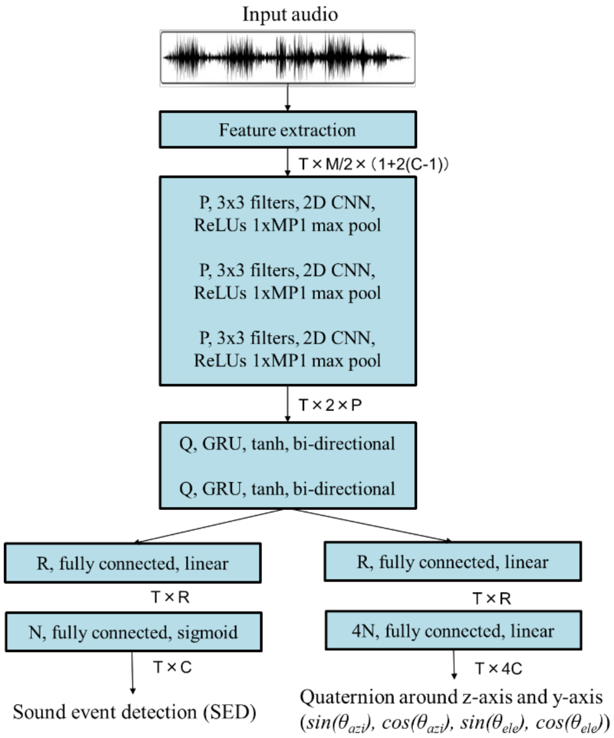
Input audio
Feature extraction
T × M/2 × (1+2(C-1))
P, 3x3 filters, 2D CNN, ReLUs 1xMP1 max pool
P, 3x3 filters, 2D CNN, ReLUs 1xMP1 max pool
P, 3x3 filters, 2D CNN, ReLUs 1xMP1 max pool
T × 2 × P
Q, GRU, tanh, bi-directional
Q, GRU, tanh, bi-directional
R, fully connected, linear
R, fully connected, linear
T × R
T × R
N, fully connected, sigmoid
4N, fully connected, linear
T × C
T × 4C
Sound event detection (SED)
Quaternion around z-axis and y-axis
(sin(θazi), cos(θazi), sin(θele), cos(θele))

Figure 1: Convolutional recurrent neural network for SELD.

Table 1: An example of the outputs and post-processing results.

Output of SED branch			Output of DOA branch				DOA angle (post processing)	
Sound event class	SED prediction	Sound activity	Ground truth of quaternion around z-axis		Ground truth of quaternion around y-axis		Azimuth	Elevation
			sin(θazi)	cos(θazi)	sin(θele)	cos(θele)		
SPEECH	0.8	active	1.0	0.0	0.5	0.9	90	30
CAR	0.1	inactive	0.1	0	-0.1	-0.1	inactive	inactive
…	0.2	inactive	0	0.1	-0.1	0	inactive	inactive
DOG	0.7	active	0.0	-1.0	0.0	1.0	-180	0
…	0.1	inactive	0.1	0	-0.1	0	inactive	inactive
TRAIN	0.1	inactive	0	-0.1	0.1	-0.1	inactive	inactive

significantly increase the number of parameters compared to baseline.

2.3. Output and post processing

In the baseline system, the sound source direction angle is directly estimated by regression, but since the angle is a periodic function, it has discontinuities. Specifically, -180 deg and 180 deg are in the same direction but are trained as a large loss as shown in Fig. 2. Therefore, in this paper, the source direction angle is estimated using quaternions defined by (4). Specifically, the output of the regression branch is changed to quaternion, and the post-processing converts it into the sound source direction angle.

$$\boldsymbol{a} = \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix}, \tag{3}$$

$$\boldsymbol{q} = \begin{pmatrix} \cos(\theta/2) \\ a_x \sin(\theta/2) \\ a_y \sin(\theta/2) \\ a_z \sin(\theta/2) \end{pmatrix}, \tag{4}$$

where, $\boldsymbol{a}=[a_x,a_y,a_z]$ is a unit vector representing the rotation axis and $\boldsymbol{q}$ is the definition of quaternion. For the angle estimation in the azimuth direction, $\boldsymbol{a}=[0, 0, 1]$ is substituted and in the elevation direction angle, $\boldsymbol{a}=[0, 1, 0]$ is substituted. That is, the output of the network can be trained with $\sin(\theta/2)$ and $\cos(\theta/2)$ as ground truth. In order to simplify the calculation, θ is substituted instead of $\theta/2$. Since $\sin(\theta)$ and $\cos(\theta)$ are continuous function, it is possible to train efficiently. Additionally, $\sin(\theta)$ and $\cos(\theta)$ always have different values. If the sound source is inactive i.e., if the ground truth of the SED is 0, then $\sin(\theta) = 0$, $\cos(\theta) = 0$ are used as ground truth. During inference, post-processing is performed as in the following equation to calculate the sound source direction angle as described in (5).

$$\theta = \begin{cases} arctan\left(\frac{\sin(\theta)}{\cos(\theta)}\right) + \pi & (if \sin(\theta) \geq 0, \cos(\theta) < 0) \\ arctan\left(\frac{\sin(\theta)}{\cos(\theta)}\right) - \pi & (if \sin(\theta) < 0, \cos(\theta) \geq 0). \\ arctan\left(\frac{\sin(\theta)}{\cos(\theta)}\right) & (\text{otherwise}) \end{cases} \tag{5}$$

As shown in the Fig. 2, if both $\sin(\theta)$ and $\cos(\theta)$ values are known, the sound source direction angle can be uniquely calculated in the range of -180 to 180 degrees. However, if both $\sin(\theta)$ and $\cos(\theta)$ are within the range of -0.2 to 0.2, it is regarded as inactive, calculation of the sound source direction angle is not performed.

3. DEVELOPMENT RESULTS

Polyphonic sound event detection and localization are evaluated with individual metrics for SED and DOA estimation. For SED, segment-based error rate (ER) and F-score [14] are calculated in one-second lengths. A lower ER or a higher F-score indicates better performance. For DOA, DOA error and frame recall are used. A lower DOA error and a higher frame recall are better. Using the cross-validation split provided for this task, Tab. 1 shows the development set performance for the proposed method. As shown in Tab. 1, the DOA error decreased but the index related to SED did not improve. The reason is that the quaternion output model proposed in this paper backpropagates an error to the DOA branch but does not directly affect the SED branch. Moreover, since the SED results are not improved, it is not considered that the spatial information using sinIPD and cosIPD did not show a significant improvement. However, in terms of the number of parameters, the baseline uses features of a total of 8-channel features consisting of amplitude spectrum and phase spectrum, while the proposed method uses 7-channel features consisting of a reference spectrogram and sinIPD, cosIPD. Therefore, the proposed method is considered to have some dimensional reduction effect.

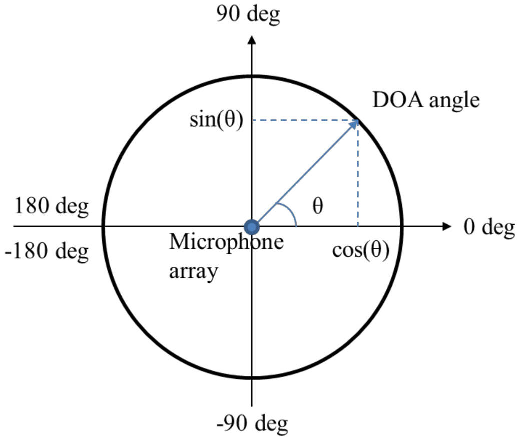

Figure 2: Unit circle with origin of microphone array position, and DOA.

Table 2: Cross validation results for the development set.

	Error rate	F score	DOA error	Frame recall
baseline-ambisonic	0.34	0.799	28.5	0.854
baseline-microphone array	0.35	0.80	30.8	0.840
proposed method	0.35	0.81	**11.5**	0.835

4. CONCLUSION

In this paper, as an approach applicable to the existing neural network model, we propose a method to replace the output and input with quaternion and sinIPD, cosIPD respectively. In the DCASE 2019 Task 3, verification was performed by replacing only the baseline input and output with the proposed method. From the experimental results, it was found that changing the output of the DOA branch to quaternion can improve the DOA estimation without changing the existing neural network model. However, because there were no changes in the SED branch, the performance associated with SED remained comparable. Regarding spatial features using sinIPD and cosIPD, although the performance was not improved, the dimension of feature was reduced. Since this method can be implemented with almost no change to existing network models, further improvement of DOA estimation is expected by using it in combination with other high-performance models.

5. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant No.16H02884, 17K00365, and 19K12017.

6. REFERENCES

[1] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, “Reliable detection of audio events in highly noisy environments,” Pattern Recognition Letters, vol. 65, pp. 22–28, 2015.

[2] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, “Acoustic monitoring and localization for social care,” Journal of Computing Science and Engineering, vol. 6, no. 1, pp. 40–50, 2012.

[3] J. Salamon and J. P. Bello, “Feature learning with deep scattering for urban sound analysis,” in 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 724–728.

[4] Y. Wang, L. Neves, and F. Metze, “Audio-based multimedia event detection using deep recurrent neural networks,” in 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 2742–2746.

[5] D. Stowell and D. Clayton, “Acoustic event detection for multiple overlapping similar sources,” in 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1–5

[6] http://dcase.community/challenge2019/ task-sound-event-localization-and-detection

[7] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 1, pp. 34–48, 2019.

[8] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2018

[9] E. B. Dam, M. Koch, and M. Lillholm. Quaternions, interpolation and animation, volume 2. 1998.

[10] S. Altmann. Rotations, Quaternions, and Double Groups. Dover Publications, 2005.

[11] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015.

[12] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[13] Cakir, E., Parascandolo, G., Herittola, T., Huttunen, H. and Virtanen, T., “Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, Issue 6, June, 2017, pp. 1291 - 1303

[14] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” in Applied Sciences, vol. 6, no. 6, 2016.