Journal of Signal Processing, Vol.25, No.6, pp.239-243, November 2021

RESEARCH NOTE

Monitoring of Domestic Activities Using Multiple Beamformers and Attention Mechanism

Yuki Kaneko[1], Takeshi Yamada[1] and Shoji Makino[1,2]

[1]Graduate School of Science and Technology, University of Tsukuba, Ibaraki 305-8573, Japan
[2]Graduate School of Information, Production and Systems, Waseda University, Fukuoka 808-0135, Japan
E-mail: takeshi@cs.tsukuba.ac.jp

Abstract **Acoustic scene classification is one of the important technologies for classifying domestic activities. When considering domestic activities as acoustic scenes, unlike the general task of acoustic scene classification, there is the problem that the sounds of the target scene and interference scene can become mixed. To deal with this problem, we propose a classification method using multiple beamformers and an attention mechanism. In the proposed method, multiple beamformers for different target directions are prepared and their outputs are input to a classifier. The proposed method then estimates the importance of each beamformer output by using an attention mechanism. To verify the effectiveness of the proposed method, we generated acoustic data by mixing the sounds of the target scene and the interference scene, and conducted a classification experiment. The experimental results confirmed that the F-score could be greatly improved by the proposed method.**

Keywords: acoustic scene classification, multiple beamformers, attention mechanism

1. Introduction

In recent years, interest in smart homes to realize safe, secure, and comfortable living has been increasing. Typical functions of smart homes include security, monitoring, and home automation [1]. To put these functions into practical use, technology for classifying domestic activities is indispensable, and there are high expectations for acoustic scene classification.

When considering domestic activities as acoustic scenes, unlike the general task of acoustic scene classification, there is the problem that the sounds of the target scene and interference scene can become mixed. One way to deal with this problem is to emphasize the desired sound by using a beamformer [2]. However, which direction should be emphasized is unclear in many situations, so it is difficult to apply the beamformer simply.

In this paper, we propose a classification method using multiple beamformers and an attention mechanism [3]. In the proposed method, multiple beamformers for different target directions are prepared and their outputs are input to a classifier after converting to log Mel-filterbank energy features. The proposed method then estimates the importance of each beamformer output using an attention mechanism. This corresponds to automatically finding the activity to be classified. To verify the effectiveness of the proposed method, we generate acoustic data by mixing the sounds of the target scene and the interference scene using the dataset of DCASE (Detection and Classification of Acoustic Scene and Event) 2018 Task 5 [4] and conduct a classification experiment.

2. Proposed Method

2.1 Overview of the proposed method

Fig. 1 shows the process flow of the proposed method. First, the emphasized sounds are obtained by using M MVDR (minimum variance distortionless response) beamformers [5].

The emphasized sounds are then input to the CNNs (convolutional neural networks) after converting to 128-dimensional log Mel-filterbank energy features. As shown in Fig. 1, the network weights are tied among the CNNs. In the proposed method, we use CNNs consisting of eight convolution layers as in [6]. Fig. 2 shows the details of the CNNs. The number of filters, pooling size, padding size, and filter size are

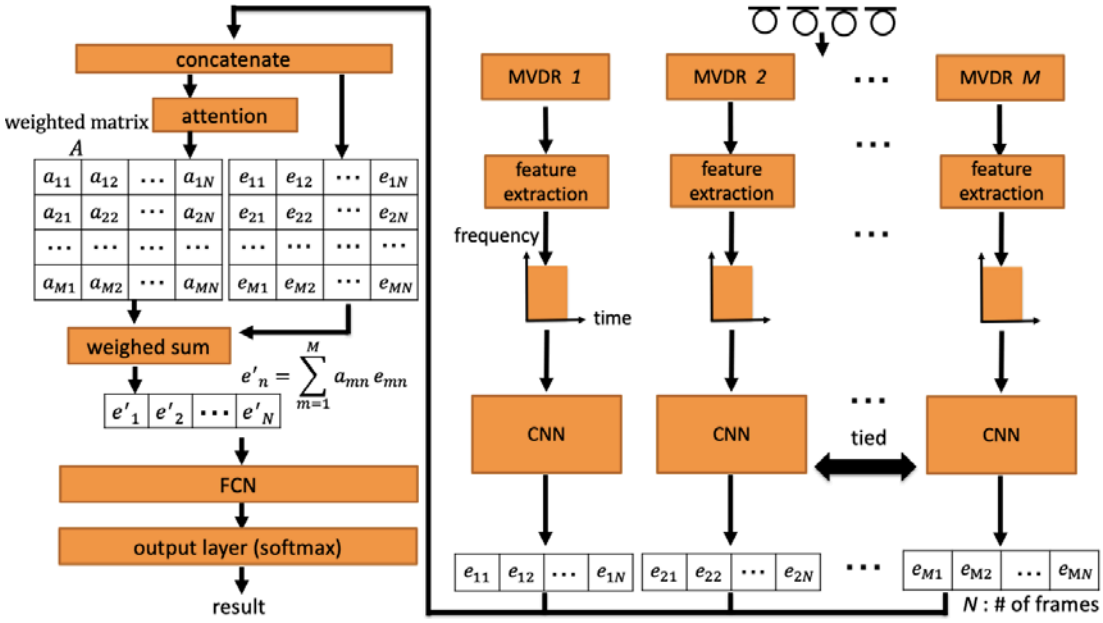

Fig. 1 Process flow of the proposed method

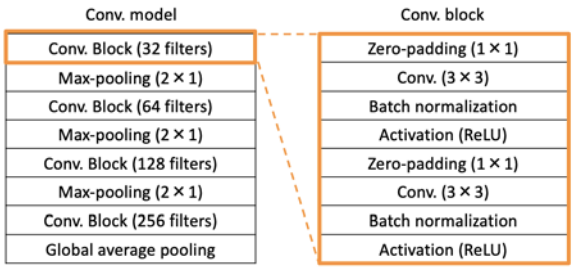

Fig. 2 Configuration of the convolution model

shown in parentheses, respectively. In the conv. block, zero padding, convolution, batch normalization, and application of ReLU function are repeated.

After that, the importance (weight) of each output of CNNs is estimated using the attention mechanism [3], and the weighted sum of each output is obtained. Finally the classification result is obtained by inputing the weighted sum to the FCN (fully connected neural networks). Here, A in Fig. 1 is the weight matrix of $M \times N$, and N is the number of time frames. We compare two types of weight matrices: time-variant (different weights can be set for each time frame) and time-invariant (same weights are set for each time frame) types. The details of the beamformer and attention mechanism are described below.

2.2 MVDR beamformer

In many speech enhancement techniques, the microphone-observed signal is represented in the time–frequency domain by a short-time Fourier transform. Here, $x_i(\omega, t)$ represents the ith observed signal at frequency ω and time frame t. Considering the case of two microphones for the sake of simplicity, a linear beamformer is generally given by

$$y(\omega, t) = \mathbf{w}^H(\omega)\mathbf{x}(\omega, t) \tag{1}$$

$$\mathbf{x}(\omega, t) = [x_1(\omega, t), x_2(\omega, t)]^T \tag{2}$$

$y(\omega, t)$, $\mathbf{w}(\omega)$, $()^T$, and $()^H$ represent the output of the beamformer, the spatial filter, transpose, and complex conjugate transpose, respectively. In the proposed method, the MVDR beamformer [5], which is one of the typical linear beamformers, is used. The formula for calculating the spatial filter of the MVDR beamformer is

$$\mathbf{w}(\omega) = \frac{R^{-1}\mathbf{a}(\omega, \theta(\omega))}{\mathbf{a}(\omega, \theta(\omega))^H R^{-1}\mathbf{a}(\omega, \theta(\omega))} \tag{3}$$

R and $\mathbf{a}(\omega, \theta(\omega))$ represent the spatial correlation matrix and the steering vector for frequency ω and the direction $\theta(\omega)$ of the target, respectively.

Since it is not obvious from which direction the sound of the target scene comes, the proposed method uses M MVDR beamformers with different target directions to obtain the emphasized sound from each target direction. In calculating the spatial filter of each MVDR beamformer, the steering vector for the corresponding target direction is given by considering only the time delay. The deviation between the target direction and the actual direction can be reduced by increasing the number of MVDR beamformers.

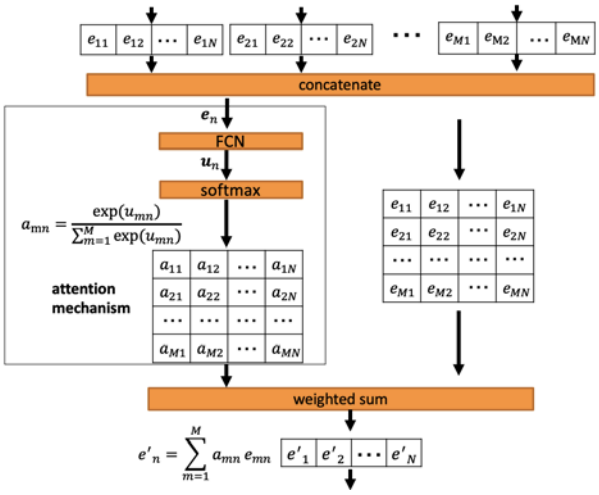

Fig. 3 Details of the attention mechanism

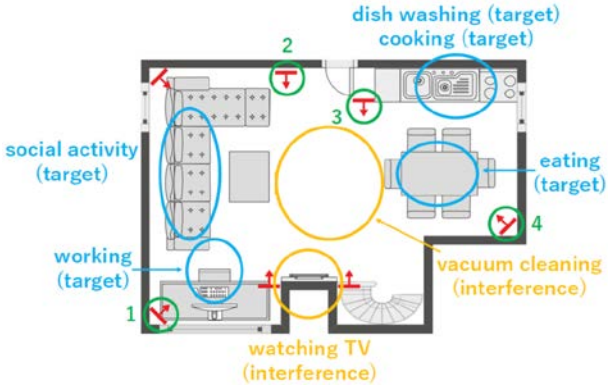

Fig. 4 Layout of the microphone arrays and the sound scenes (quoted from [7])

Table 1 Number of acoustic data of each class

activity	# sessions	# 10 s segments
cooking	13	5124
dishwashing	10	1424
eating	13	2308
working	33	18644
social activity	21	4944

2.3 Classifier with attention mechanism

Fig. 3 shows the details of the attention mechanism [3] in the proposed method. First, the nth column vector $\mathbf{e}_n = (e_{1n}, \cdots, e_{Mn})^T$ of the $M \times N$ matrix, which is obtained by concatenating the outputs of the CNNs, is input to the FCN to obtain the output $\mathbf{u}_n = (u_{1n}, \cdots, u_{Mn})^T$.

$$\mathbf{u}_n = f(\mathbf{e}_n), \quad n = 1, 2, \cdots, N \tag{4}$$

f is the nonlinear function corresponding to the FCN. N and M represent the number of time frames and the number of MVDR beamformers, respectively. In this paper, the FCN has one layer with 32 units. Next, as shown by (5), the weight a_{mn} for each direction and each time frame is obtained using the softmax layer.

$$a_{mn} = \frac{\exp(u_{mn})}{\sum_{m=1}^{M} \exp(u_{mn})}, \quad n = 1, 2, \cdots, N \tag{5}$$

We compare two types of weight matrices: time-variant and time-invariant types. Finally, the weighted sum of the CNN outputs is obtained for each time frame using (6) and then used to obtain the classification result.

$$e'n = \sum_{m=1}^{M} a_{mn} e_{mn}, \quad n = 1, 2, \cdots, N \tag{6}$$

3. Experiment

3.1 Experimental conditions

To verify the effectiveness of the proposed method, we generated acoustic data by mixing the sounds of the target scene and the interference scene using the dataset of DCASE 2018 Task 5 [4] and conducted a classification experiment. DCASE Challenge is an annual event and offers a competitive platform to compare different methods using common datasets. DCASE 2018 Task 5 is an acoustic scene classification task and the dataset comprises the sounds of a person living in a villa for one week, which were recorded using seven 4-channel microphone arrays.

Fig. 4 shows the layout of the microphone arrays and the acoustic scenes. The red arrows, the green circles, the blue circles, and the orange circles indicate the direction of the microphone array, the four microphone arrays used in this experiment, the positions where the target scene occurs and the positions where the interference scene occurs, respectively. In this experiment, the five types of target scenes were cooking, dishwashing, eating, working, and conversation, and the two types of interference scenes were watching TV and vacuum cleaning. The data of the target scene and the interference scene recorded by the same microphone array were randomly selected and mixed.

The total number of acoustic data after mixing is 32,444. Table 1 shows the number of acoustic data of each acoustic scene before mixing. "# sessions" indicates the number of recordings and each recording is divided into segments of 10 s, the number of which is given under "# 10 s segments." The microphone interval in the microphone array is 5 cm. The sampling frequency is 16 kHz and the quantization bits is 12. The features are 128th-order log Mel-filterbank energy. The frame length and frame shift

Table 2 Experimental results (F-score [%])

	time-invariant	time-variant
single channel	64.61	—
proposed (correct weight)	83.46	—
proposed	76.18	75.68

length in the frame analysis are 40 and 20 ms, respectively. The Adam algorithm [8] is used as the optimization method for training the classifier; the number of epochs during training is set up to be 50 and the epoch that gives the best F-score is adopted. The Chainer framework [9] is used for the implementation of the proposed method. In addition, M in Fig. 1 was 3 and the target direction was set to 30°, 90° (direct front), and 150°. Only mixed acoustic data were used for learning the classifier.

In this paper, the macro-averaged F1-score is used as an evaluation metric, which is the average of the class-wise F1-scores. The class-wise F1-score is defined by

$$\text{class-wise F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad (7)$$

where the precision is the number of true positive results divided by the number of all positive results, and the recall is the number of true positive results divided by the number of all positive samples.

3.2 Results and discussion

To verify the effectiveness of the proposed method, we compare the performance in the following three cases.

- single channel
- proposed (correct weight)
- proposed

'Single channel' is the case where only a single channel of the 4-channel microphone array is used for classification, and corresponds to classifying mixed data as is. In addition, 'proposed (correct weight)' is the case in which the correct weight matrix was given in the proposed method and refers to the upper limit performance of the proposed method. In the proposed method, the cases of time-variant and time-invariant weight matrices are also compared.

Table 2 summarizes the experimental results. First, the F-score of 'single channel' was 64.61%. Since the F-score was 89.27% when we conducted a similar experiment without the interference scenes, we can see that the F-score was significantly decreased by the presence of interference scenes. Next, the F-score of 'proposed (correct weight)' was 83.46%. This result means that the F-score was greatly improved by emphasizing the sound of the target scene. Similarly, the F-score of 'proposed' was 76.18%. While it was less than that of 'proposed (correct weight)', about a 12% improvement was obtained compared with that of 'single channel'. It is necessary to optimize the network structure of the attention mechanism in order to approach the classification accuracy of 'proposed (correct weight)'. Finally, the F-score of the proposed method with the time-variant weight matrix was almost the same as that of the time-invariant case. This is because there is no major movement of the sounds in this experiment.

Table 3 F-score [%] of each microphone array

	array 1	array 2	array 3	array 4
proposed	68.20	76.63	79.37	73.46

Next, Table 3 shows the F-score of the proposed method for each microphone array. The microphone array indexes correspond to Fig. 4. The F-score of the microphone array 1 was 68.20%, which is lower than those of other microphone arrays. It is considered that this is because the directions of the target scene and the interference scene are often close to each other, such as for the pair of vacuum cleaning and eating. In addition, the F-score of microphone array 3 is the highest among the four microphone arrays and was 79.37%. It is considered that this is because the directions of the target scene and the interference scene are sufficiently separated. The fact that the three directions set as the target directions matched the actual direction of the target scene would also be a factor in increasing the F-score of microphone array 3. From this result, it is considered that the F-score can be further improved by increasing the number of beamformers.

In addition, Fig. 5 shows an example of the weight matrix estimated by the attention mechanism, when eating and watching TV were recorded by microphone array 3. The horizontal axis represents the time frame, and the vertical axis represents the MVDR beamformer index and target direction. We can see that the weight in the 30° direction, where the target scene is located, is large. However, the weight of 90° or 150° is locally large, and this is considered to be the difference in the performance between 'proposed (correct weight)' and 'proposed'.

4. Conclusions

In this paper, we proposed a classification method of using multiple beamformers and the attention mechanism for the situation in which the sounds of the target scene and interference scene are mixed. To

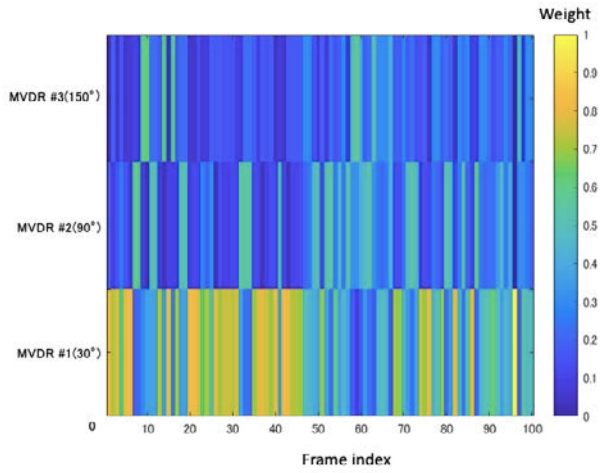

Fig. 5 Weights in each direction estimated by the attention mechanism

evaluate the effectiveness of the proposed method, wee generated acoustic data by mixing the sounds of the target scene and the interference scene taken from the dataset of DCASE2018 Task 5 and conducted a classification experiment. As a result, it was confirmed that the proposed method can automatically find and emphasize the sound to be classified, and the classification accuracy was improved by 12% compared with the method using a single channel.

Acknowledgment

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant No. 20K11880 and 19H04131.

References

[1] M. Alaa, A. A. Zaidan, B. B. Zaidan, M. Talal and M. L. M. Kiah: A review of smart home applications based on Internet of Things, Journal of Network and Computer Applications, Vol. 97, pp. 48–65, 2017.

[2] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi and K. Hamada: Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling, DCASE2018 Challenge Technical Report, 2018.

[3] D. Bahdanau, K. Cho and Y. Bengio: Neural machine translation by jointly learning to align and translate, arXiv:1409.0473, 2014.

[4] G. Dekkers, L. Vuegen, T. Waterschoot, B. Vanrumste and P. Karsmakers: DCASE 2018 Challenge Task 5: Monitoring of domestic activities based on multi-channel acoustics, arXiv:1807.11246v2, 2018.

[5] H. L. Van Trees: Optimum Array Processing, John Wiley & Sons, 2002.

[6] Y. Han and J. Psrk: Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification, DCASE2017 Challenge Technical Report, 2017.

[7] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. V. D. Bergh, T. V. Waterschoot, B. Vanrumste, M. Verhelst and P. Karsmakers: The sins database for detection of daily activities in a home environment using an acoustic sensor network, Proc. DCASE 2017 Workshop, pp. 32-56, Nov. 2017.

[8] D. P. Kingma and J. L. Ba: Adam: A method for stochastic optimization, ArXiv:1412.6980v1, Dec. 2014.

[9] Chainer website, https://chainer.org/.

Yuki Kaneko received his B.E. and M.E. degrees from University of Tsukuba, Japan, in 2019 and 2021, respectively. He was engaged in research on acoustic scene classification.

Takeshi Yamada received his B.E. degree from Osaka City University, Japan, in 1994, and his M.E. and Dr. Eng. degrees from Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. He is presently an associate professor with Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. His research interests include speech recognition, sound scene understanding, multichannel signal processing, media quality assessment, and elearning. He is a member of the IEEE, the IEICE, the IPSJ, and the ASJ.

Shoji Makino received his B.E., M.E., and Ph.D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981 and University of Tsukuba in 2009. He is now a professor at Waseda University, Japan. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech, and acoustic signal processing for speech and audio applications. He is an IEEE Fellow, an IEICE Fellow, a council member of the ASJ, and a member of the EURASIP.

(Received May 11, 2021; revised July 2, 2021)