Neural Beamformer with Automatic Detection of Notable Sounds for Acoustic Scene Classification

Sota Ichikawa*, Takeshi Yamada* and Shoji Makino†*
* University of Tsukuba, Japan
† Waseda University, Japan
Email: s.ichikawa@mmlab.cs.tsukuba.ac.jp

***Abstract*—Recently, acoustic scene classification using an acoustic beamformer that is applied to a multichannel input signal has been proposed. Generally, prior information such as the direction of arrival of a target sound is necessary to generate a spatial filter for beamforming. However, it is not clear which sound is notable (i.e., useful for classification) in each individual sound scene and thus in which direction the target sound is located. It is therefore difficult to simply apply a beamformer for preprocessing. To solve this problem, we propose a method using a neural beamformer composed of the neural networks of a spatial filter generator and a classifier, which are optimized in an end-to-end manner. The aim of the proposed method is to automatically find a notable sound in each individual sound scene and generate a spatial filter to emphasize that notable sound, without requiring any prior information such as the direction of arrival and the reference signal of the target sound in both training and testing. The loss functions used in the proposed method are of four types: one is for classification and the remaining loss functions are for beamforming that help in obtaining a clear directivity pattern. To evaluate the performance of the proposed method, we conducted an experiment on classifying two scenes: one is a scene where a male is speaking under noise and another is a scene where a female is speaking under noise. The experimental results showed that the segmental SNR averaged over all the test data was improved by 10.7 dB. This indicates that the proposed method could successfully find speech as a notable sound in this classification task and generate the spatial filter to emphasize it.**

I. Introduction

The research field of environmental sound classification and detection, which aims to utilize sounds around us, consists of various fundamental technologies. Acoustic scene classification (ASC) is one of them, and it takes an acoustic signal of several seconds in duration as input and classifies it among predefined scenes. For example, if we want to classify the location where an acoustic signal was recorded, a set of scenes may include a restaurant, a train station, and a park. This technology is expected to improve multimedia retrieval and situational awareness for autonomous robots.

Detection and classification of acoustic scenes and events (DCASE) is an international research community and has been holding regular competitions since 2013 [1]. As shown by the methods submitted to the competitions, acoustic features have changed from handcrafted features such as mel frequency cepstral coefficients to more primitive features such as log spectrograms, and classifiers have also changed from support vector machines [2] and Gaussian mixture models [3] to deep-learning-based classifiers such as convolutional neural network (CNN) [4].

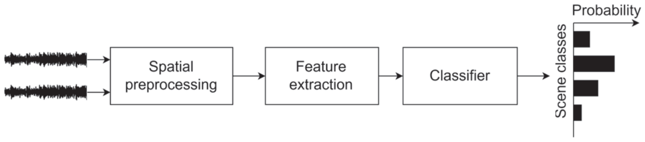


Fig. 1. Overview of ASC incorporating spatial preprocessing

In recent years, ASC methods incorporating spatial preprocessing that is applied to a multichannel input signal as shown in Fig. 1 have been proposed. For example, Han *et al.* employed mid-side processing as spatial preprocessing [5] and Tanabe *et al.* also applied blind source separation [6]. This is because the classification accuracy can be improved if sounds useful for classification can be extracted in advance. However, it is not clear which sound is notable (i.e., useful for classification) in each individual sound scene and thus in which direction the target sound is located and furthermore how many sound sources exist. It is therefore difficult to simply apply spatial preprocessing to ASC.

To solve this problem, we propose a method using a neural beamformer composed of the neural networks of a spatial filter generator and a classifier, which are optimized in an end-to-end manner. The aim of the proposed method is to automatically find a notable sound in each individual sound scene and generate a spatial filter to emphasize that notable sound. The loss functions used in the proposed method are of four types: one is for classification and the remaining loss functions are for beamforming, which are inspired by the minimum variance distortionless response (MVDR) beamformer [7] and help to obtain a clear directivity pattern. Unlike previous neural beamformers [8], [9], [10], the proposed method does not require any prior information such as the direction of arrival and the reference signal of the target sound in both training and testing. To evaluate the performance of the proposed method, we conduct an experiment on classifying two scenes: one is a scene where a male is speaking under noise and another is a scene where a female is speaking under noise.

II. Beamformer

A. Acoustic beamformer

An acoustic beamformer is a technique to enhance a target signal from a specific direction. First, we suppose that the target signal s_{tk} is observed by a microphone array with M microphones. The observed signal vector $\boldsymbol{x}_{tk}$ is given by

$$\boldsymbol{x}_{tk} = s_{tk}\boldsymbol{a}_k + \boldsymbol{n}_{tk}, \tag{1}$$

where $\boldsymbol{a}_k$ is the steering vector of the target signal and $\boldsymbol{n}_{tk}$ the noise signal vector. t and k denote the time frame index and frequency bin index, respectively. Then the enhanced target signal $\hat{s}_{tk}$ can be obtained as

$$\hat{s}_{tk} = \boldsymbol{w}_k^{\boldsymbol{H}}\boldsymbol{x}_{tk}, \tag{2}$$

where $\boldsymbol{w}_k^{\boldsymbol{H}}$ represents the spatial filter coefficient vector and H the Hermitian transpose. The MVDR beamformer [7] is one of the widely used methods and its spatial filter is given by

$$\boldsymbol{w}_k = \frac{\boldsymbol{R}_k^{-1}\boldsymbol{a}_k}{\boldsymbol{a}_k^H\boldsymbol{R}_k^{-1}\boldsymbol{a}_k}, \tag{3}$$

where $\boldsymbol{R}_k$ denotes the spatial covariance matrix. This is obtained by solving the following constrained minimization problem.

$$\text{minimize } E[|\boldsymbol{w}_k^{\boldsymbol{H}}\boldsymbol{x}_{tk}|^2], \quad \text{subject to } \boldsymbol{w}_k^{\boldsymbol{H}}\boldsymbol{a}_k = 1 \tag{4}$$

This means that the MVDR beamformer minimizes the output signal power under the constraint that the sensitivity in the direction of the target signal is set to 1 (undistorted), resulting in forming the directivity pattern with directional nulls in the directions of the undesired noise signals. Therefore, it is appropriate to suppress directional noise signals using a small number of microphones. One drawback is that the MVDR beamformer needs to be given (or estimate) the steering vector in the direction of the target signal. It is therefore difficult to simply apply the MVDR beamformer in preprocessing for ASC as described above.

B. Neural beamformer

Neural beamformers can be divided into two approaches in terms of how the spatial filter is obtained: by acoustic beamformer formulation and by using DNNs.

1) Obtaining spatial filter by acoustic beamformer formulation: In this approach, DNNs are utilized as an auxiliary to obtain the spatial filter according to an acoustic beamformer formulation such as Eq. (3) in the MVDR beamformer. For example, a method that uses a neural beamformer as the frontend of automatic speech recognition (ASR) has been proposed [8]. In this method, DNNs are used for mask estimation and reference microphone estimation, which are simultaneously optimized with ASR to improve recognition performance in noisy environments. A method of replacing matrix inversion and eigenvalue decomposition with DNNs has also been proposed [9]. In both methods, the spatial filter is obtained according to the mask-based MVDR beamformer formulation [11].

This approach has the advantage that the spatial filter with the intended directivity pattern can be obtained reliably. On the other hand, prior information such as the direction of arrival of the target sound must be given or estimated as well as the original acoustic beamformer.

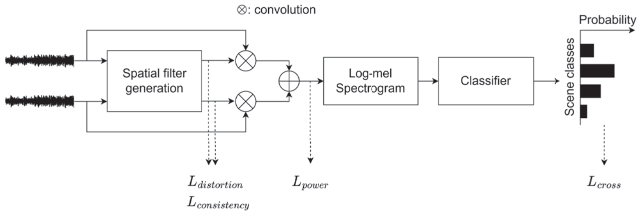


Fig. 2. Overview of the proposed method

2) Obtaining spatial filter using DNNs: In this appoach, the spatial filter is directly obtained using DNNs. For example, Li *et al.* proposed a method that employs a DNN-based spatial filter generator for improving the performance of ASR in noisy environments [10]. The spatial filter generator consists of some long short-term memory (LSTM) layers and generates the spatial filter in each time frame. Although this method requires parallel data of noisy speech and its original clean speech for training, there is no guarantee that the spatial filter with the intended directivity pattern can be obtained because no constraint is imposed on the generated filter.

Our proposed method is under this category but differs from the methods mentioned above in the following respects.

- It does not require any prior information such as the direction of arrival and the reference signal of the target sound in both training and testing.
- The loss functions that help to obtain the spatial filter with the intended directivity pattern are introduced.

III. Proposed method

A. Overview

Fig. 2 shows the overview of the proposed method, where the number of microphones M is set to two for simplicity. In the proposed method, the spatial filter is first generated from only the multichannel input signal. The enhanced signal is then obtained by applying the spatial filter to the multichannel input signal. After the enhanced signal is converted to the log-mel spectrogram and fed to the classifier, the softmax probability over scenes is outputted by the classifier.

Fig. 3 shows the network structure of the spatial filter generator, which was designed on the basis of the work of Li *et al.* [10]. The input signal in each channel is divided into small time frames that overlap one-quarter length of the frame size. Then the sequence of the input signal vectors, which are obtained by concatenating the two channel signals in each time frame, is inputted to the first LSTM layer. The second LSTM layer in each channel outputs the time-domain filter coefficient vector in each time frame. After that, the spatial filter is applied to the two channel signals in each time frame. Finally, the enhanced signal is obtained by the overlap-add method.

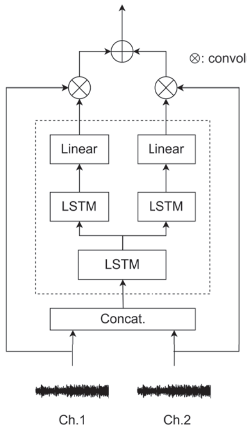


Fig. 3. Network structure of the spatial filter generator

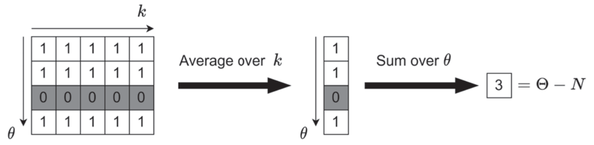


Fig. 4. Meaning of $L_{directivity}$

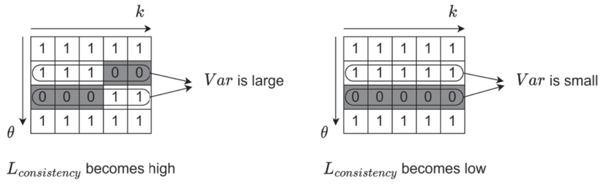


Fig. 5. Meaning of $L_{consistency}$

B. Loss function

In the proposed method, the entire network is trained in an end-to-end manner by minimizing the following loss function.

$$L = \alpha_1 L_{cross} + \alpha_2 L_{power} + \alpha_3 L_{directivity} + \alpha_4 L_{consistency}, \tag{5}$$

where α_1, α_2, α_3 and α_4 represent the weight of each loss function. The loss functions are of four types. L_{cross} is the cross entropy loss used for general softmax-based classification and given by

$$L_{cross} = -\sum_{x} p(x) \log q(x), \tag{6}$$

where p and q represent the true probability distribution and the estimated probability distribution, respectively. By minimizing L_{cross}, we find that the estimated probability distribution approaches the true probability distribution.

The remaining three types of loss functions are for helping to obtain a clear directivity pattern. These loss functions are defined as

$$L_{power} = \frac{1}{T} \sum_{t} \sum_{k} |\boldsymbol{w}_{tk}^{\boldsymbol{H}} \boldsymbol{x}_{tk}|^2, \tag{7}$$

$$L_{directivity} = \frac{1}{T} \sum_{t} \left((\Theta - N) - \frac{1}{K} \sum_{\theta} \sum_{k} |\boldsymbol{w}_{tk}^{\boldsymbol{H}} \boldsymbol{a}_{\theta,k}| \right)^2, \tag{8}$$

$$L_{consistency} = \frac{1}{T} \sum_{t} \left(\frac{1}{\Theta} \sum_{\theta} var_{\theta} \left(|\boldsymbol{w}_{tk}^{\boldsymbol{H}} \boldsymbol{a}_{\theta,k}| \right) \right), \tag{9}$$

where T is the number of time frames, K the number of frequency bins, Θ the number of directions, and N the number of directional nulls ($N = M - 1$). The proposed method aims to form the directivity pattern with directional nulls similarly to the MVDR beamformer. L_{power} in Eq. (7) represents the frame-averaged power of the beamformer output signal. Minimizing L_{power} is identical to the power minimization in Eq. (4). On the other hand, since the target direction is now unknown, it is impossible to impose the same constraint in Eq. (4).

Therefore, we introduce the constraint $L_{directivity}$ in Eq. (8). A directivity matrix is obtained by finding the sensitivity (i.e., the inner product of the generated spatial filter and the steering vector) at each frequency bin and each direction. A directivity evaluation score is then defined by averaging the directivity matrix over frequency bins and then summing over directions. Note that this score can be calculated without the target and the null directions. Minimizing $L_{directivity}$ means bringing the score closer to $\Theta - N$. Fig. 4 shows an example of the directivity matrix when $K = 5$, $\Theta = 4$, and $N = 1$. The directivity matrix in Fig. 4 has a directional null at the third θ and represents a sensitivity of 1 for the other θ. The directivity evaluation score for such a clear directivity pattern with a directional null becomes 3, which is equal to $\Theta - N$. Therefore, it can be expected that bringing the score closer to $\Theta - N$ will lead to obtaining a clear directivity pattern with directional nulls. However there are of course many other spatial filters with a score of $\Theta - N$. Therefore, $L_{directivity}$ is the loose constraint to ensure that the directivity pattern of the generated spatial filter does not deviate significantly from the intended one.

Finally, $L_{consistency}$ in Eq. (9) is the constraint on the consistency of the direction of the directional null. As mentioned above, since $L_{directivity}$ is the loose constraint, the directional nulls can appear in different directions in each frequency bin as shown in Fig. 5. However, this is often undesirable for wideband signals such as speech. Thus, $L_{consistency}$ leads so that the directional null in each frequency bin appears in the same direction by minimizing the variance of the sensitivity in each direction.

Simultaneously minimizing the four loss functions enables us to automatically find a notable sound in each individual sound scene and generate a spatial filter to emphasize that notable sound. The most attractive feature is that our loss function can be calculated without any prior information such as the direction of arrival and the reference signal of the target

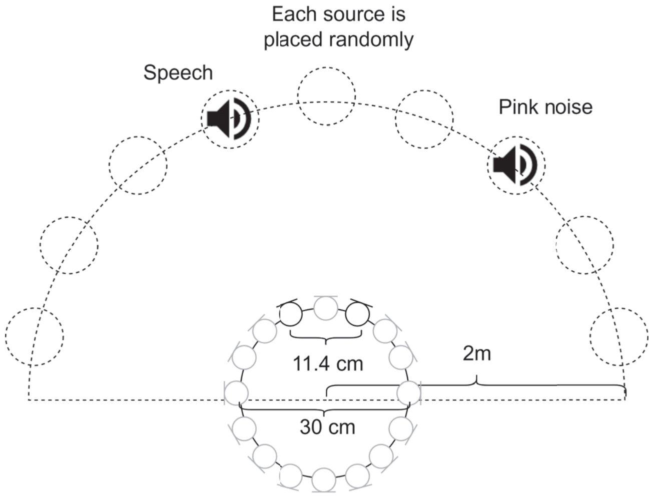


Fig. 6. Arrangement of microphones and sounds

sound.

IV. Experiment and results

A. Overview

The effectiveness of the proposed method is evaluated from the following two viewpoints.

- Can the proposed method improve the classification accuracy?
- Can the proposed method find the notable sound and generate the spatial filter that emphasizes it?

To perform this evaluation reliably, we designed the task of classifying two scenes: one is a scene where a male is speaking under noise and another is a scene where a female is speaking under noise. It is obvious that the notable sound in this task is speech. However, clean speech, which can be used as a cue of a notable sound, is not included in the training data, and the direction of arrival of the notable sound is not given. Thus, this task requires learning that the notable sound is speech and generating the spatial filter that enhances it.

B. Dataset

As shown in Fig. 6, mixed sounds of speech and pink noise arriving from distinct directions at the same time were generated with a signal-to-noise ratio (SNR) of -20 dB. Two microphones of the circular microphone array were used for the experiment. The directions of arrival of speech and pink noise were randomly selected from $10°, 30°, \cdots, 170°$ (9 directions in $20°$ steps) so as not to overlap. Sound propagation was simulated by convolving the measured impulse responses included in the RWCP sound scene database [12] shown in Table I. Male and female speech data were randomly selected from the ASJ continuous speech corpus for research (ASJ-JIPDEC) [13] shown in Table II.

TABLE I
Details of measured impulse responses data

Recording device	circular microphone array (16 ch, 30 cm in diameter)
Microphones used	two adjacent microphones (distance 11.4 cm)
Recording environment	anechoic chamber
Sound source position	2 m from microphone array center
Sample rate	48 kHz
Format	32 bit float

TABLE II
Details of speech data

# of speech data	9600 sentences
# of speakers	64 (30 male and 34 female)
Channels	1
Sample rate	16 kHz
Format	16 bit integer

TABLE III
Details of generated data

# of data	4500 per scene (70% is used for training)
Signal length	2 seconds
Channels	2
Sample rate	16 kHz
Format	16 bit integer

Finally, the mixed-sound dataset shown in Table III was generated. The number of mixed sounds is 4500 for each of the male and female scenes, 70% of which is used for training and the remaining 30% for testing. Note that clean speech, which can be used as a cue of a notable sound, is not included in this dataset.

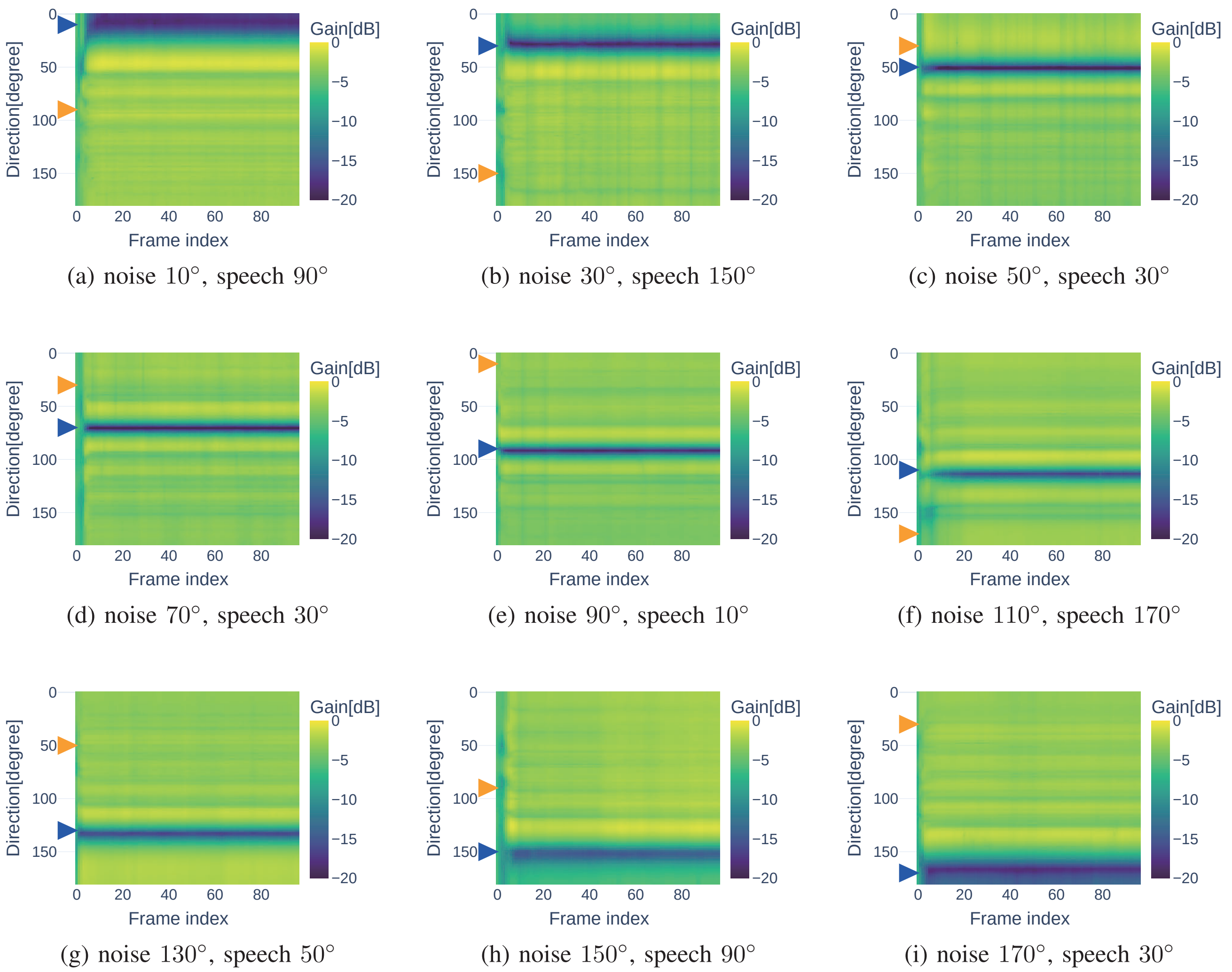
C. Configuration

In the spatial filter generator, the frame size was set to 80 ms, the hop size to 20 ms, and the length of the spatial filter to 5 ms. The sizes of the hidden states of the LSTM layers were 1536 for the first layer and 768 for the 2nd layer. In the experiment, the CNN-based classifier was used, which is the same as that of the baseline method in the DCASE 2019 Task 1A [14] and its network configuration is shown in Table IV. When training, the AdamW optimizer [15] was adopted with the learning rate of 0.001. The batch size was 100 and the entire network was trained for 100 epochs.

For the loss functions, the number of frequency bins K was set to 1024, the number of directions Θ to 9 as described above, and the number of directional nulls to 1 since the number of microphones M was 2. All the weights of the four loss functions, a_1, a_2, a_3, and a_4 were set to 1.

D. Result

Fig. 7 shows the examples of the directivity pattern of the spatial filter generated by the proposed method. In this figure, the pink noise directions are selected in order from $10°, 30°, \cdots, 170°$, whereas the speech directions are randomly selected. In each figure, the vertical axis represents the direction and the horizontal axis the time frame index. The color map shows that the bluer the color, the lower the sensitivity (0 dB means the sensitivity is 1). The blue and orange arrows indicate the pink noise direction and the speech



(a) noise 10°, speech 90°

(b) noise 30°, speech 150°

(c) noise 50°, speech 30°

(d) noise 70°, speech 30°

(e) noise 90°, speech 10°

(f) noise 110°, speech 170°

(g) noise 130°, speech 50°

(h) noise 150°, speech 90°

(i) noise 170°, speech 30°

Fig. 7. Directivity pattern of the spatial filters generated by the proposed method

TABLE IV
NETWORK STRUCTURE OF THE CLASSIFIER

Layer	Kernel	Stride	Feature maps	Dropout rate
input				
conv	(7, 7)	(1, 1)	32	0.3
BN and ReLU				
maxpool	(5, 5)	(2, 2)		
conv	(7, 7)	(1, 1)	64	0.3
BN and ReLU				
maxpool	(4, 20)	(2, 2)		
ReLU				
fc			100	0.3
fc			2	
softmax				
output				

direction, respectively. From these figures, we can see that the directivity pattern with one-directional null in the noise direction was successfully generated for various combinations of the speech and noise directions. This means that the loss function in the proposed method works as intended.

The segmental SNR, which is the average of SNRs over segmented frames, is one of the evaluation metrics of speech enhancement. The segmental SNR averaged over all the test data was improved by 10.7 dB. This indicates that the spatial filter emphasizing speech was successfully generated. As a result, the proposed method achieved a classification accuracy of 95.8%, which was 20.3% higher than when using only the classifier with monaural input data (i.e., the method that excludes the spatial filter generator from the proposed method).

The experimental results confirmed that the proposed method could successfully find speech as a notable sound in this classification task and generate the spatial filter to emphasize that notable sound.

V. CONCLUSIONS

In this paper, we proposed the method using a neural beamformer composed of the neural networks of the spatial filter generator and the classifier, which are optimized in an end-to-end manner. The feature of the proposed method is to automatically find a notable sound in each individual sound scene and generate a spatial filter to emphasize that notable sound, without requiring any prior information such as the direction of arrival and the reference signal of the target sound in both training and testing. To realize it, we devised the four types of loss functions: one is for classification and the remainings are for beamforming that help to obtain a clear directivity pattern. To evaluate the performance of the proposed method, we conducted an experiment on classifying two scenes: one is a scene where a male is speaking under noise and another is a scene where a female is speaking under noise. The experimental results showed that the proposed method can successfully find speech as a notable sound in this classification task and generate the spatial filter to emphasize that notable sound.

Acknowledgment

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

References

[1] DCASE Community, http://dcase.community/.

[2] J. T. Geiger, B. Schuller, G. Rigoll, “Large-scale audio feature extraction and SVM for acoustic scene classification,” Proc. WASPAA2013, pp. 1–4, 2013.

[3] M. Chum, A. Habshush, A. Rahman, C. Sang, “IEEE AASP scene classification challenge using hidden markov models and frame based classification,” DCASE2013 Challenge Technical Report, 2013.

[4] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, T. Virtanen, “A convolutional neural network approach for acoustic scene classification,” Proc. IJCNN, pp. 1547–1554, 2017.

[5] Y. Han, J. Park, K. Lee, “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification,” DCASE2017 Challenge Technical Report, 2017.

[6] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi, K. Hamada, “Multi-channel acoustic scene classification by blind dereverberation, data augmentation, and model ensembling,” DCASE2018 Challenge Technical report, 2018.

[7] H. L. Van Trees, “Optimum Array Processing,” John Wiley & Sons, 2002.

[8] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, “Multichannel end-to-end speech recognition,” Proc. ICML, vol. 70, pp. 2632–2641, 2017.

[9] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, D. Yu, “ADL-MVDR: All deep learning MVDR beamformer for target speech separation,” Proc. ICASSP2021, pp. 6074–6078, 2021.

[10] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, M. Bacchiani, “Neural network adaptive beamforming for robust multichannel speech recognition,” Proc. INTERSPEECH2016, pp. 1976–1980, 2016.

[11] T. Higuchi, N. Ito, T. Yoshioka, T. Nakatani, “Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise,” Proc. ICASSP2016, pp. 5210–5214, 2016.

[12] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,”’ Proc. LREC2000, pp. 965–968, 2000.

[13] T. Kobayashi, S. Itabashi, S. Hayamizu, T. Takezawa, “ASJ continuous speech corpus for research,” J. Acoust. Soc. Jpn. (J), Vol. 48, pp. 888–893, 1992.

[14] A. Mesaros, T. Heittola, T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” Proc. DCASE2018 Workshop, pp. 9–13, 2018.

[15] I. Loshchilov, F. Hutter, “Decoupled weight decay regularization,” Proc. ICLR2019, 2019.