Autofocus Neural Beamformer Based on Steering Vector Estimation

Reiya Marukawa* and Takeshi Yamada*
* Degree Programs in Systems and Information Engineering, University of Tsukuba, Japan
Email: s2420658@u.tsukuba.ac.jp

***Abstract*—In recent years, acoustic scene classification methods using an acoustic beamformer that takes a multichannel signal as input have been developed. Generally, prior information such as the direction of arrival of a target sound is necessary to generate a spatial filter (directivity pattern) for beamforming. However, it is not clear which sound is notable (useful for classification) in each individual sound scene and thus in which direction the target sound is located. This makes it difficult to simply apply a beamformer for preprocessing. To address this problem, we previously proposed an autofocus neural beamformer that automatically finds a notable sound in each individual sound scene and generates a spatial filter to emphasise 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. In this paper, to further improve the performance of our conventional method, we modify it so that a steering vector corresponding to the direction of arrival of a notable sound is estimated instead of the spatial filter itself and the spatial filter is calculated using the minimum variance distortionless response (MVDR) criterion. The effectiveness of this method is demonstrated by an acoustic scene classification experiment in a reverberant environment.**

I. Introduction

Acoustic environment recognition is a research field focused on identifying the sounds present in our surroundings. Within this field, acoustic scene classification is a core technology that takes an audio segments of several seconds as input and determines which predefined scene—such as a restaurant, train station, or park—it belongs to. This technology is expected to support applications such as situational awareness for autonomous vehicles and mobile robots, as well as security systems.

With the rapid progress of deep neural networks (DNNs), a mainstream approach to acoustic scene classification employs a log-Mel spectrogram as the input feature and a convolutional neural network (CNN) as the classifier. More recently, Transformer-based models such as the audio spectrogram Transformer (AST) [1] and its faster, regularized variant patchout audio spectrogram Transformer (PaSST) [2] have been introduced, further improving recognition accuracy. In addition, with a view toward edge deployment, device robustness, model compression, and computational efficiency have become critical issues, and active research is underway to address them [3].

When the input audio is multichannel, applying multichannel signal processing such as beamforming as a preprocessing step has been explored [4][5]. By enhancing sounds useful for sound scene recognition and suppressing others, we can expect that beamforming boosts classification performance. Designing a spatial filter (directivity pattern) for a beamformer normally requires prior information such as the target sound direction. However, in practice, the sounds that should be emphasised—and the directions from which they arrive—are not obvious for each acoustic scene, making beamforming as a preprocessing stage inherently challenging.

To address this issue, we previously proposed an acoustic scene classification method (hereafter, our conventional method) that employs an autofocus neural beamformer [6]. As shown in Fig. 1, our conventional method concatenates two neural networks—one that estimates a spatial filter and another that performs scene classification—and trains them end-to-end using the standard cross-entropy loss for classification together with bespoke losses that evaluate the suitability of the estimated spatial filter. This framework automatically finds the sounds most useful for the downstream classifier and produces a spatial filter that emphasises them, without requiring any prior information such as the target direction. Nevertheless, it has been reported that the spatial filter estimation becomes unstable, especially in reverberant environments.

In this paper, we improve the method by estimating not the spatial filter itself but the steering vector corresponding to the arrival direction of sounds useful for classification, and then computing the spatial filter via a minimum variance distortionless response (MVDR) beamformer. When the spatial filter itself is estimated directly, we must evaluate whether the resulting directivity pattern is appropriate. Although the conventional method uses three loss functions for this purpose, they appear insufficient. In contrast, if we estimate the steering vector, the spatial filter is derived under the MVDR criterion, guaranteeing its suitability by design. We still need to evaluate the validity of the steering vector, but this is expected to be easier than validating an spatial filter. The remainder of the paper details the proposed method and demonstrates its effectiveness through acoustic scene classification experiments in reverberant environments.

II. Beamforming

A. MVDR Beamformer

Beamforming is a technique that extracts a target sound by enhancing or suppressing signals arriving from specified directions. In particular, the MVDR beamformer, which achieves

979-8-3315-7206-8/25/$31.00

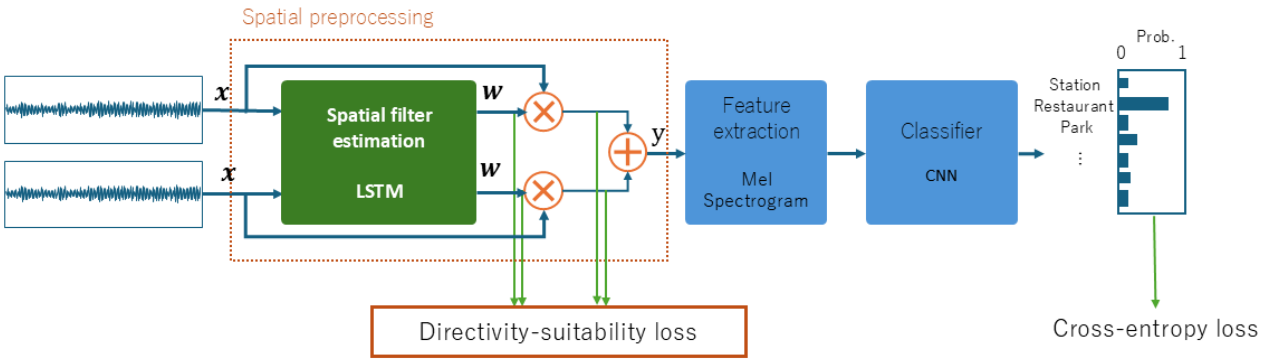


Fig. 1. Network structure of our conventional method

high enhancement even with a small number of microphones, passes sound from the designated direction without distortion while minimising the power of interference. In the MVDR beamformer, the observed multichannel signal $\boldsymbol{x}$ is modelled as

$$\boldsymbol{x}_{tf} = s_{tf}\boldsymbol{a}_f + \boldsymbol{n}_{tf}, \tag{1}$$

where

- $\boldsymbol{x} \in \mathbb{C}^M$: observed signal vector (M microphones),
- $\boldsymbol{a} \in \mathbb{C}^M$: steering vector of the target direction,
- $s \in \mathbb{C}$: target signal,
- $\boldsymbol{n} \in \mathbb{C}^M$: interference-plus-noise vector,
- t: time-frame index,
- f: frequency-bin index.

Here, $\boldsymbol{x}$ is expressed in the time–frequency domain; for simplicity, the indices t and f are omitted below. The steering vector represents, for each microphone, the spatial transfer characteristics (amplitude and phase) between the source and the microphones.

The beamformer output y is obtained with a spatial filter (weight vector) $\boldsymbol{w}$:

$$y = \boldsymbol{w}^H\boldsymbol{x}, \tag{2}$$

where $\boldsymbol{w}^H$ denotes the Hermitian (conjugate transpose) of $\boldsymbol{w}$.

The MVDR beamformer determines $\boldsymbol{w}$ by minimising the output power while forcing the target signal to pass undistorted:

$$\begin{aligned} &\min_{\boldsymbol{w}} \ \mathbb{E}[|y|^2] = \min_{\boldsymbol{w}} \ \boldsymbol{w}^H\boldsymbol{R}\boldsymbol{w}, \\ &\text{subject to} \ \ \boldsymbol{w}^H\boldsymbol{a} = 1, \end{aligned} \tag{3}$$

where $\boldsymbol{R}$ is the spatial covariance matrix of the observations. Using the method of Lagrange multipliers, we obtain the optimal MVDR filter as

$$\boldsymbol{w}_{\mathrm{MVDR}} = \frac{\boldsymbol{R}^{-1}\boldsymbol{a}}{\boldsymbol{a}^H\boldsymbol{R}^{-1}\boldsymbol{a}}. \tag{4}$$

Applying $\boldsymbol{w}_{\mathrm{MVDR}}$ to the observations yields an estimate $\hat{s}$ of the desired source in which interference and noise are suppressed.

B. Neural Beamformers

A neural beamformer estimates the spatial filter $\boldsymbol{w}$ with a DNN. Existing neural beamformers can be broadly grouped into two categories.

1) Direct estimation of the spatial filter: In this paradigm, the DNN outputs the filter coefficients $\boldsymbol{w}$ directly. Li *et al.*, for example, proposed a long short-term memory (LSTM) based system that predicts per-channel, per-frame weights and generates the beamformer output by a filter-and-sum operation as a front-end for noise-robust ASR [7]. Because no prior information such as the target direction is required, the architecture is simple and highly flexible, potentially allowing the network to learn more versatile directivity patterns. The drawback is that the estimated filter is unconstrained by beamforming theory, so its directivity may be sub-optimal and the optimisation can become unstable.

Our previous autofocus neural beamformer addressed this issue by designing loss terms that evaluate the suitability of the filter and training the filter estimator and the classifier end-to-end [6]. Although it successfully focused on the target direction without any prior information, the filter estimation was reported to be unstable in reverberant environments.

2) Leveraging an existing beamformer framework: A representative approach embeds a DNN into the conventional MVDR framework so that the spatial filter $\boldsymbol{w}$ is still computed in a data-driven manner. Typical examples include estimating a time–frequency mask and the reference microphone with a DNN and then deriving the MVDR filter from them, or replacing the explicit matrix inversion in a mask-based MVDR beamformer with a network to stabilise training [8]. The gradients become well behaved, when the analytical inversion is eliminated, making large-scale end-to-end learning feasible.

Because these methods retain the theoretical foundation of the MVDR beamformer, they carry little risk of producing unsuitable directivity patterns and training is generally stable. However, they require prior information such as the target direction or clean reference signals, which must be provided or estimated separately.

The proposed method follows this second paradigm. Since the spatial filter is obtained under the MVDR criterion, its suitability is guaranteed, yet—crucially—the method dispenses with any prior knowledge of the target direction, which is the key advantage of our approach.

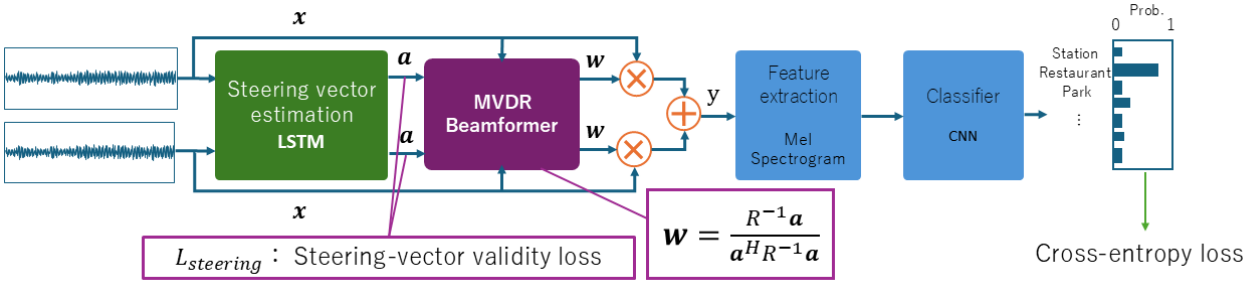


Fig. 2. Network structure of the proposed method

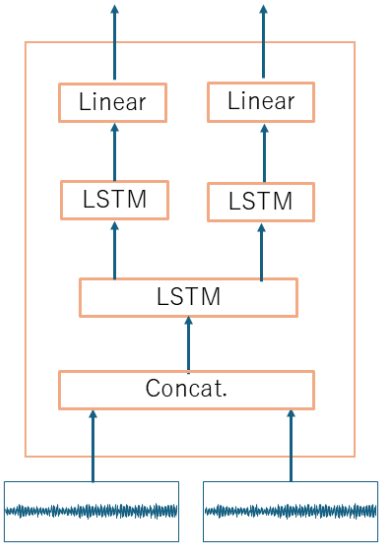


Fig. 3. Structure of the steering vector estimator

III. Proposed Method

A. Network Architecture

Figure 2 illustrates the overall architecture of the proposed method. All components are trained end-to-end with two losses: the standard cross-entropy loss for the classifier and a custom loss that evaluates the validity of the estimated steering vector.

Compared with our previous work, the filter estimator is replaced by a steering-vector estimator. The estimator first predicts the steering vector corresponding to the direction of the sound that is useful for classification. An MVDR spatial filter is then derived from this vector, guaranteeing the suitability of the filter itself. Consequently, only the validity of the steering vector must be assessed—a point addressed in the next subsection.

As shown in Fig. 3, the steering-vector estimator is realised by LSTM. A stereo time–frequency signal, segmented into short-time frames, is fed into a shared LSTM layer; the network then branches into channel-wise LSTM layers followed by fully connected layers, producing complex coefficients for every frame and frequency bin.

The downstream classifier is a CNN. It receives the log-Mel spectrogram of the monaural beamformer output and returns soft-max posteriors over the acoustic-scene classes.

B. Loss Function

To judge the validity of the estimated steering vector, we introduce the loss term L_{steering}. A steering vector can be written as

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix} = \begin{bmatrix} e^{-j2\pi f \tau_1} \\ e^{-j2\pi f \tau_2} \\ \vdots \\ e^{-j2\pi f \tau_M} \end{bmatrix}, \tag{5}$$

where each element represents the transfer function from the source to a microphone. Because $\boldsymbol{a}$ is tightly linked to the direction of arrival (DoA), a valid steering vector should point to a single, well-defined direction. Estimating the DoA independently at every frequency, however, is unreliable at high frequencies owing to spatial aliasing. We therefore adopt cross-power-spectrum phase (CSP) analysis [9], which integrates information over all frequencies, and formulate a loss that encourages the CSP of the estimated steering vector to peak sharply at a physically plausible time difference of arrival (TDOA).

CSP estimates the TDOA from the phase correlation of two signals:

$$\mathrm{CSP}_{1,2}(\tau) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[x_1(n)]\,\mathcal{F}[x_2(n)]^*}{|\mathcal{F}[x_1(n)]\,\mathcal{F}[x_2(n)]^*|}\right], \tag{6}$$

where

- $x_1(n), x_2(n)$: signals at microphones 1 and 2,
- $\mathcal{F}, \mathcal{F}^{-1}$: DFT and IDFT,
- $*$: complex conjugate.

The lag that maximises $\mathrm{CSP}_{1,2}(\tau)$ is $\hat{\tau} = \arg\max_\tau \mathrm{CSP}_{1,2}(\tau)$, and the TDOA is

$$\mathrm{TDOA} = \frac{\hat{\tau}}{F_s}, \tag{7}$$

where F_s is the sampling rate.

Assuming the target direction is unknown, we define

$$L_{\text{steering}} = \alpha\, L_{\text{ROI}} + (1-\alpha)\, L_{\text{Non-ROI}}, \tag{8}$$

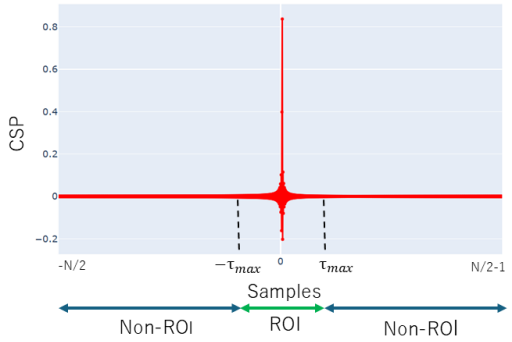


Fig. 4. Example of CSP coefficients

where α is a weighting factor and

$$L_{\text{ROI}} = \frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}\Big(-\sum_{\tau\in\mathcal{T}_{\text{ROI}}} p(\tau)\log p(\tau)\Big), \tag{9}$$

$$L_{\text{Non-ROI}} = \frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}\sum_{\tau\in\mathcal{T}_{\text{Non-ROI}}} \text{CSP}_{1,2}(\tau). \tag{10}$$

Here,

- B: batch size, T: number of frames, τ: sample index,
- $\text{CSP}_{1,2}(\tau)$: CSP coefficient at lag τ,
- $p(\tau)$: soft-max normalised CSP coefficients.

In practice, $\text{CSP}_{1,2}(\tau)$ is computed from the *estimated* steering vector instead of raw observations. $\mathcal{T}_{\text{ROI}}$ denotes the range of physically feasible TDOAs; $\mathcal{T}_{\text{Non-ROI}}$ is its complement (Fig. 4). L_{ROI} becomes small when a sharp CSP peak appears within the ROI, whereas $L_{\text{Non-ROI}}$ penalises large CSP values outside it. Minimising L_{steering}, therefore, drives the steering vector to correspond to exactly one direction.

IV. Experiments

A. Experimental Conditions

Tables I–IV show the experimental settings. To make it straightforward to assess whether the beamformer emphasises sounds useful for classification, we define 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 and the system is trained to decide which scene a given signal belongs to. Each input consists of a mixture of speech signals arriving simultaneously from different directions plus pink noise. Impulse responses were measured in an anechoic chamber and in a variable reverberation chamber; by switching between them, we generate two datasets, *anechoic* and *reverberant*. All parameters were chosen so that the best possible performance could be achieved under the experimental conditions. We ran all experiments on Ubuntu 18.04.4 LTS using an Intel Core i9-10940X processor with 32 GB of memory and NVIDIA Quadro RTX 6000 GPU (24 GB). Our implementation was developed in Python 3.11.0 with PyTorch 2.0.1 and relied on CUDA 11.8.

The direction of arrival is randomly selected from nine angles (10°–170° in 20° steps) as shown in Fig. 5, ensuring that the two sources are not located in the same direction. During training, only the mixtures are provided and the true DoA remains unknown, forcing the model to learn by itself which sound is informative, to enhance that sound (and suppress noise) via the beamformer, and thereby to improve classification accuracy.

TABLE I
SPEECH DATA USED IN THE EXPERIMENTS

Database	ASJ Continuous Speech Corpus (Japanese) [10]
Number of utterances	~9600 sentences
Speakers	64 (30 male, 34 female)
Channels	1
Sampling rate	16 kHz
Quantisation	16 bit

TABLE II
IMPULSE-RESPONSE DATA

Database	RWCP Sound Scene Database [11]
Recording array	Circular array, 16 ch, 30 cm diameter
Environments	Anechoic, reverberation room (RT_{60} = 0.31 s)
Source position	2 m away from the center of the microphone array
Sampling rate	48 kHz
Quantisation	32-bit float

TABLE III
GENERATED DATASET

Number of mixtures samples	9000 (70% used for training)
Duration per sample	2 s
Channels	2
Sampling rate	16 kHz
Quantisation	16 bit

TABLE IV
HYPERPARAMETERS

Optimiser	Adam
Epochs	100
Batch size	100
Learning rate	0.001
α in (8)	0.5
Frame length	75 ms
Frame shift	20 ms
LSTM hidden size (layer 1)	2048
LSTM hidden size (layer 2)	1664
Log-Mel bins	40

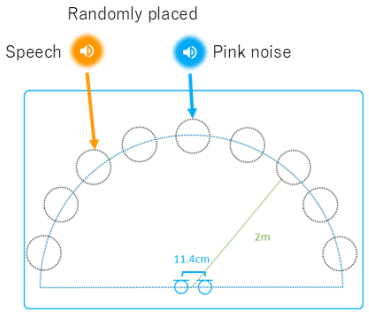


Fig. 5. Geometry of microphones and sound sources

TABLE V
CLASSIFICATION ACCURACY OF EACH METHOD

Accuracy (%)	Classifier only	Ideal	Conventional	Proposed
Anechoic	76.2	98.0	92.8	98.2
Reverberant	65.4	86.6	57.4	82.1

B. Results and Discussion

In Figs. 6–9, we compare the spatial directivity patterns estimated by our proposed and conventional method in anechoic and reverberant rooms. The horizontal axis denotes the time frames, the vertical axis denotes the azimuth, and the colour indicates gain (dB); warmer colours correspond to higher sensitivity. Orange arrows indicate the target direction, whereas blue arrows indicate the interference direction. With the proposed method, a deep null is formed toward the noise source in both rooms, and—especially in the reverberant room—the directivity is noticeably cleaner than that obtained with the conventional method. This improvement stems from predicting the steering vector first and then deriving the spatial filter under the MVDR constraint, rather than estimating the filter coefficients directly.

Table V shows the scene-classification accuracy. The leftmost column lists the evaluation environment (anechoic/reverberant); the four right-hand columns correspond to the following methods:

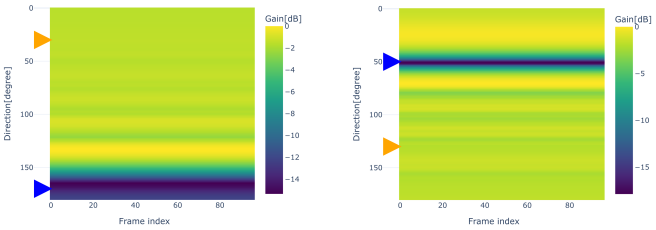
- **Classifier only**: baseline without any beamforming,
- **Ideal**: upper bound obtained by applying an MVDR filter using the *true* steering vector,
- **Conventional**: our previous method that directly estimates the filter,
- **Proposed**: the present method that derives the filter from the estimated steering vector.

In both environments, the proposed method surpasses the conventional method and approaches the ideal accuracy, confirming its effectiveness.

Finally, we observed that the LSTM-based steering-vector estimator occasionally converged to a single, scene-dependent direction. This implies that, once the MVDR beamformer achieves sufficiently good performance for the current classification task, the network may settle on one steerable point rather than exploring a wider range of directions. Investigating how to prevent such premature convergence—e.g., by encouraging greater steering-vector diversity or by regularising the MVDR front-end—remains an open issue and will be the focus of our future work.

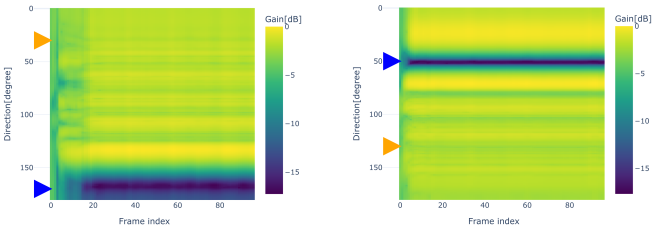
V. Conclusion

We previously introduced an autofocus neural beamformer that finds and enhances sounds useful for downstream classification without any prior directional information. To further improve its performance, in this paper, we estimate the steering vector corresponding to the arrival direction of useful sounds and compute the spatial filter with the MVDR criterion instead of predicting the filter itself. A new loss based on CSP analysis is used to evaluate the steering vector. Experiments conducted



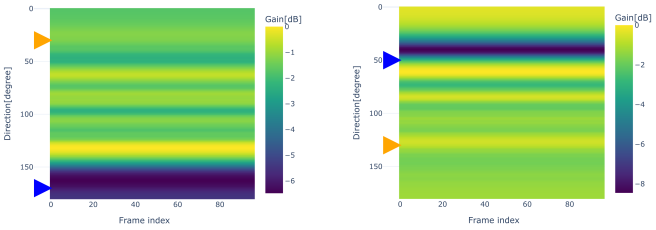
(a) target 30°, noise 170° (b) target 130°, noise 50°

Fig. 6. Directivity in an anechoic chamber estimated by the proposed method



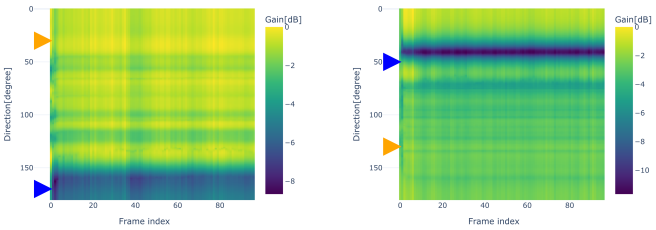
(a) target 30°, noise 170° (b) target 130°, noise 50°

Fig. 7. Directivity in an anechoic chamber estimated by our conventional method



(a) target 30°, noise 170° (b) target 130°, noise 50°

Fig. 8. Directivity in a reverberation room estimated by the proposed method



(a) target 30°, noise 170° (b) target 130°, noise 50°

Fig. 9. Directivity in a reverberation room estimated by our conventional method

in anechoic and reverberant rooms showed that the proposed method outperforms the conventional method and approaches the ideal upper bound, demonstrating its effectiveness.

Acknowledgment

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant No. 23K28113.

References

[1] Y. Gong, Y.-A. Chung, J. R. Glass, "AST: audio spectrogram transformer," Proc. INTERSPEECH 2021, pp. 571–575, 2021.
[2] K. D. Koutini, H. Eghbal-zadeh, G. Widmer, "Efficient training of audio transformers with patchout," Proc. INTERSPEECH 2022, pp. 3763–3767, 2022.
[3] X. Chen, W. Xie, "Data-efficient low-complexity acoustic scene classification in the DCASE2024 challenge," DCASE2024 Challenge Technical Report, 2024.
[4] Y. Han, J. Park, K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," DCASE2017 Challenge Technical Report, 2017.
[5] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi, K. Hamada, "Multichannel acoustic scene classification by blind dereverberation, data augmentation, and model ensembling," DCASE2018 Challenge Technical Report, 2018.
[6] S. Ichikawa, T. Yamada, S. Makino, "Neural beamformer with automatic detection of notable sounds for acoustic scene classification," Proc. APSIPA ASC 2022, pp. 866–871, 2022.
[7] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," Proc. INTERSPEECH 2016, pp. 1976–1980, 2016.
[8] 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. ICASSP 2021, pp. 6074–6078, 2021.
[9] M. Omologo, P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," Proc. ICASSP 1994, pp. II/273–II/276, 1994.
[10] 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.
[11] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," Proc. LREC2000, pp. 965–968, 2000.