APSIPA Transactions on Signal and Information Processing, 2023, 12, e16

Original Paper

Missing Data Completion of Multi-channel Signals Using Autoencoder for Acoustic Scene Classification

Yuki Shiroma[1*], Yuma Kinoshita[2], Keisuke Imoto[3], Sayaka Shiota[1], Nobutaka Ono[1] and Hitoshi Kiya[1]

[1] *Tokyo Metropolitan University, Tokyo, Japan*
[2] *Tokai University, Kanagawa, Japan*
[3] *Doshisha University, Kyoto, Japan*

ABSTRACT

In this paper, we propose an autoencoder-based missing data completion method for multi-channel acoustic scene classification (ASC). It has been reported that many deep-learning-based ASC methods using multi-channel signals have robust performance. The advantage of using multi-channel data is the capture of spatial and frequency information. However, when there is missing data in multi-channel signals, the classification performance declines significantly. We focus on completing the missing data by using an autoencoder as the preprocessor of ASC models. Since positional relationships between multi-channel microphones are modeled in the latent space of the proposed autoencoder, missing information is reconstructed via the latent space from the multi-channel input, including missing data. In an experiment, the missing data is completed by using the proposed autoencoder, and the accuracy of ASC systems is improved by using the completed multi-channel signals.

Keywords: Missing data completion, Autoencoder, Multi-channel signals, Acoustic scene classification

*Corresponding author: Yuki Shiroma, shiroma-yuki@ed.tmu.ac.jp.

Received 01 November 2022; Revised 24 January 2023
ISSN 2048-7703; DOI 10.1561/116.00000074

1 Introduction

Acoustic scene classification (ASC) has been researched actively [1]. Owing to the evolution of machine learning, many deep neural network (DNN)-based methods for ASC have recently been proposed [4, 16, 29]. In ASC, it is well known that multi-channel signals contribute to robust systems because both spatial and frequency information are extracted from multi-channel signals because sound source positions can be estimated from the spatial information [11, 15, 31]. Additionally, multimodal approaches with multi-channel signals that effectively use the spatial information have been researched [18, 19].

However, when there is missing data in multi-channel signals due to microphone malfunctions, packet loss in network errors, faulty connections of the microphone cable, the classification performance of ASC systems is significantly degraded because the multi-channel signals with missing data differ from those without missing data [28]. In conventional methods of addressing the missing data problem, the simulated missing data was regarded as augmentation data and mixed into training data [20, 25]. However, it is difficult to simulate all types of missing data in training data.

As another approach to addressing the missing data problem, we focus on completing the missing data by using an autoencoder as the preprocessor of ASC models. Since an autoencoder models its latent space to reconstruct input data, important information is compressed into the latent space. In the case of inputting multi-channel signals to an autoencoder, it is assumed that the relationships between multi-channel signals are modeled in the latent space. Therefore, by using the reconstruction function of an autoencoder, in this paper, we propose an autoencoder-based missing data completion method for ASC systems with multi-channel signals. Since the proposed method is assumed to be used as the preprocessor of an ASC system, it is expected to mitigate the missing data problem in ASC systems regardless of whether the classification system has a countermeasure against the problem.

To construct the proposed autoencoder for completing the missing data, we use a convolutional autoencoder (CAE) to reconstruct the time-frequency domain. The reason is that since the short-time Fourier transform (STFT) magnitude is unchanged under translation in the time domain and robust to changes in waveform unrelated to the meaning of the data [35], audio signals are often used as spectrograms in DNN-based ASC techniques [16, 22, 29]. To investigate the characteristics of the proposed CAE, we prepared five different architectures. To evaluate the proposed method, we performed acoustic classification tasks with three different ASC models. Our experimental results show that the proposed method completed the missing data satisfactorily, improving the accuracies of all ASC systems.

The rest of the paper is organized as follows. In Section 2, the problems of multi-channel ASC are explained. Then, the proposed missing data completion method using an autoencoder is introduced in Section 3. In Section 4, we evaluate the proposed method and show the experimental results. Finally, our conclusion is shown in Section 5.

2 Problems of Multi-channel Acoustic Scene Classification

The purpose of ASC is to classify scene sounds into predefined categories. Techniques related to ASC are expected to have many applications such as auto-tagging to multimedia [3], localization [26], and anomaly detection [21]. Since 2016, an international competition named the detection and classification of acoustic scenes and events (DCASE) challenge [23] has been held every year. As a result of this competition, many DNN-based ASC techniques have been proposed. Recently, some tasks using multi-channel signals have also been focused on because a multi-channel database has been released by the organizers of the DCASE challenge. Since multi-channel signals include spatial information, they are expected to contribute to improving classification performance, unlike single-channel signals. It has also been reported that multi-channel-based ASC systems obtain a higher accuracy than single-channel-based ones. However, as shown in Figure 1, when multi-channel signals include missing data, the classification performance declines significantly [28] owing to the missing data problem. Multi-channel data may include missing data owing to a recording error, for example, in the databases of the DCASE 2018 challenge task5 [8] and CHiME5 [2].

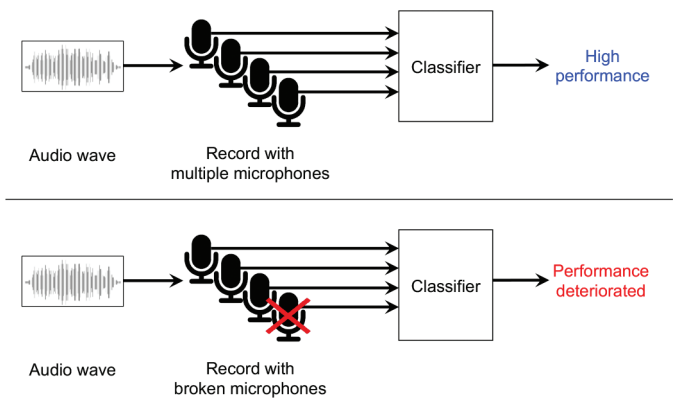

Figure 1: Performance differences between acoustic tasks with clean and missing data.

To address the mismatch problem caused by multi-channel data with missing data between the training and testing data, data-augmentation-based

methods have been proposed [20, 25]. These methods include simulating missing data, which is used as training data to train a robust model. Although the data augmentation-based methods work well, it is difficult to expect them to be effective for all types of missing data and to be able to generate many types of simulated data. Moreover, when the simulation data is changed or added, the ASC models must be retrained with the data. On the other hand, semi-supervised domain adaptation [30] and knowledge transfer [14] were proposed for the device adaptation task to solve the recording device mismatch problem between training and testing data. In this paper, the proposed method is regarded as a preprocessing of the classifiers and aims to avoid retraining the classifiers. Therefore, these methods focusing on improving the robustness of the classifier itself is not regarded as comparison method. The device adaptation task is a task to relieve differences of multi-device and is not focused on the multi-channel signals. However, the device adaptation task is helpful to our proposed method to expand to the application of using multi-device.

3 Missing Data Completion Using Autoencoder

3.1 Motivation

We propose a missing data completion method to address the missing data problem described in Section 2. When multi-channel signals include missing data in some channels, it is difficult to extract spatial information satisfactorily, resulting in the performance deterioration of the classification system using multi-channel signals. Therefore, the proposed method focuses on completing a missing channel, and the reconstructed multi-channel signals are used to extract sufficient spatial information to improve the ASC performance. Since the proposed method is assumed to be used as the preprocessor of an ASC system, it is expected to mitigate the missing data problem, regardless of whether the classification system has a countermeasure against the problem.

Although the virtual microphone technique performs a similar task [9, 24, 34], it requires details of each microphone position. In contrast, the proposed method simply requires the channel with missing data.

3.2 Completion Procedure

In this section, we describe the completion procedure of the proposed method. an autoencoder is used as a missing data completion model. Originally, an autoencoder was used for dimensionality reduction or ceature extraction. As shown in Figure 2, an autoencoder consists of two components, an encoder and a decoder. As shown in Equation (1), the autoencoder is trained to reconstruct

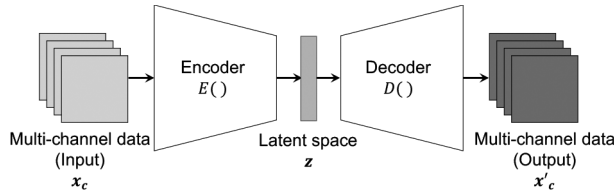

Figure 2: Network architecture of the autoencoder used in the proposed method.

input data by minimizing the mean squared error (MSE) of the input and output as follows.

$$\min\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}-D\left(E\left(\boldsymbol{x}\right)\right)\right)^2\right), \tag{1}$$

where n is the number of data samples and $\boldsymbol{x}$ represents the input spectrograms. $E()$ and $D()$ are the encoder and decoder functions, respectively. Since the encoder layers compress input data into a low-dimensional latent space, the decoder layers reconstruct input from the latent space with interpolation. In our proposed method, we focus on the fact that multi-channel signals with missing data are completed by passing them through the autoencoder. Since passing the signals through the autoencoder only once is insufficient, multi-channel signals with missing data are input repeatedly, similarly to that in the Griffin–Lim algorithm [12]. Since multi-channel signals are represented as spectrograms in many ASC tasks, the multi-channel autoencoder completes the missing data in the time-frequency domain. In DNN-based ASC techniques, audio signals are often represented as spectrograms [16, 22, 29] because the STFT magnitude is unchanged under translation in the time domain and robust to changes in waveform unrelated to the meaning of the data [35]. When a multi-channel signal input into the autoencoder has a missing channel, the output of the autoencoder is expected to complete the missing channel information by using the decoder layers. The proposed completion procedure is illustrated in Figure 3. First, multi-channel spectrograms with missing data are input into an autoencoder, that has already been trained with a clean multi-channel dataset. Second, only the missing channel data of the autoencoder input is replaced with the output data. Then, multi-channel spectrograms with the replaced data are input into the autoencoder again. By repeating this process, the missing data is gradually completed.

3.3 Multi-channel Autoencoder Architechture

We prepared five completion models based on various concepts as listed below. Figure 4 shows the architecture of each CAE, where A, B, and C represent the

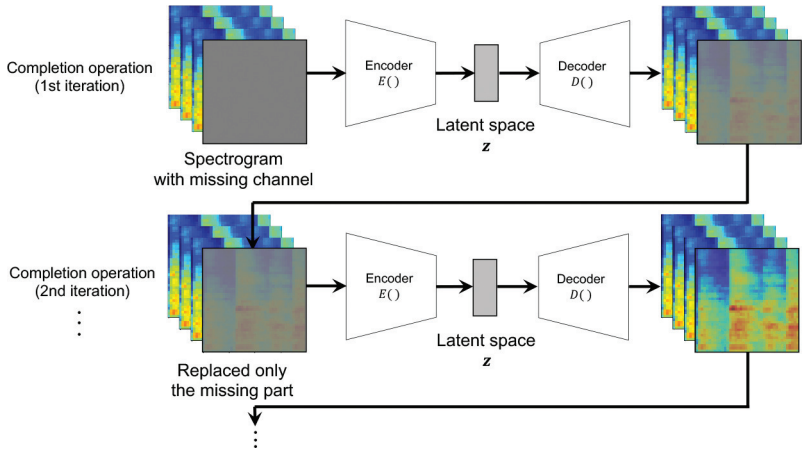

Figure 3: Block diagram of the proposed missing data completion method using an autoencoder. Only the missing channel is replaced. Normal channels are not changed.

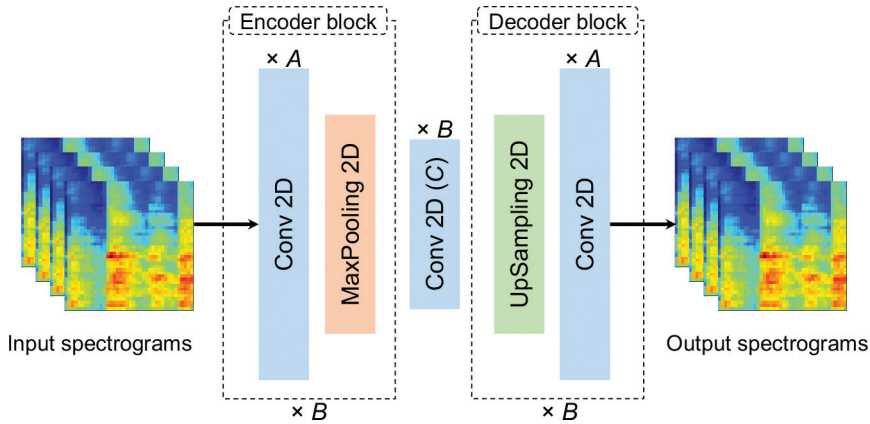

Figure 4: Network structure of the CAE used in the proposed method.

numbers of convolutional layers and encoder–decoder blocks and the maximum number of channels in the latent space, respectively.

1. CAE-S
   CAE-S has five layers; $A = 1$, $B = 1$, and $C = 8$. The maximum number of channel dimensions in the latent space is set to 256. Since the number of parameters is small, the expressive power of the autoencoder is limited. We assumed that the limited expressive power reduces the amount of abnormal output.

2. CAE-M
   CAE-M has 17 layers; $A = 1$, $B = 3$, and $C = 256$. The maximum number of channel dimensions in the latent space is set to 256. In the anomalous sound detection tasks, a key task is to encode audio features into latent space to detect unusual sounds. In the proposed method, we expect that the CAE-M will encode channel relationships correctly. CAE-M is based on a model that was introduced for anomalous sound detection tasks [10].

3. CAE-L
   CAE-L has 28 layers; $A = 2$, $B = 3$, and $C = 256$. The maximum number of channel dimensions in the latent space is set to 256. Since the decoder of CAE-L has many convolutional layers, we expect that the decoder layers will generate a more detailed output than the other models.

4. CAE-latent-S
   CAE-latent-S has 17 layers; $A = 1$, $B = 3$, and $C = 8$. The maximum number of channel dimensions in the latent space is limited to 8. The autoencoder strongly compresses extracted information into the latent space. Thus, it is expected that channel relationships will be extracted explicitly.

5. CAE-latent-L
   CAE-latent-L has 27 layers; $A = 1$, $B = 3$, and $C = 1024$. The maximum number of channel dimensions in the latent space is set to 1024. We expect that the autoencoder will reconstruct input more accurately than the other models because the autoencoder holds a large amount of information in the latent space.

Each CAE has different numbers of convolutional layers and encoder–decoder blocks and a different maximum number of channel dimensions in the latent space. Therefore, the completion ability of each CAE is different.

4 Experiment

To evaluate the proposed method, we performed three experiments to investigate the completion performance of each CAE structure, the performance of some classification systems by using the completed signals, and the robustness of the completion models.

4.1 Database

The SINS database [7] was used as a multi-channel signal scene database. The SINS database contains continuous multi-channel audio recordings of one person living in a vacation home over a period of one week, and it was provided as a database for the ASC task. In our experiment, microphone arrays Nos. 2, 4, 6, and 8 were selected, where the numbers correspond to those in Figure 5, and one channel of each microphone array was combined to prepare four-channel audio signals. For the classification task, nine classes of acoustic scenes, "Absence," "Cooking," "Dishwashing," "Eating," "Other," "Social activity," "Vacuum cleaner," "Watching TV," and "Working," were used. Since the sample numbers of "Calling" and "Visit" were insufficient, the class "Social activity" was made by concatenating "Calling" with "Visit" [6]. The database was divided into two parts, one to construct the completion models and one to construct the ASC models. Tables 1 and 2 show overviews of the database for constructing the completion and ASC model, respectively. The completion model part was split into training and validation data, and the ASC model part was split into training, validation, and testing data. The audio was divided into 10 seconds. The direct current components of the audio waves were also removed.

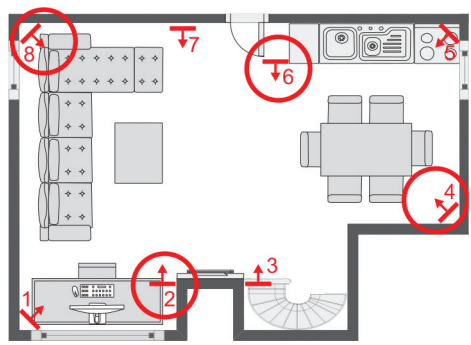

Figure 5: Two-dimensional floorplan of combined kitchen and living room [8] with selected microphone array numbers in SINS database.

4.2 Experimental Setup

As an input feature, four-channel log-Mel spectrograms were used. The input feature was extracted by calculating the STFT, where the window length was 1024, the window shift was 320, and the dimension of the Mel filter bank was 40. To investigate the characteristics of the best architectures for completing the missing data, the five model architectures described in Section 3 were compared. As the conventional method of image reconstruction, U-Net [27]

Table 1: Overview of the dataset for constructing completion models [hours].

Classes	Training	Validation	Total
Absence	8.4	2.1	10.5
Cooking	2.3	0.5	2.8
Dishwashing	0.6	0.2	0.8
Eating	1.0	0.3	1.3
Other	0.9	0.2	1.1
Social activity (Calling + Visit)	2.2	0.5	2.7
Vacuum cleaner	0.4	0.1	0.5
Watching TV	8.3	2.1	10.4
Working	8.3	2.1	10.4
Total	32.4	8.1	40.5

Table 2: Overview of the dataset for constructing ASC models [hours].

Class	Training	Validation	Testing	Total
Absence	8.4	2.1	2.6	13.1
Cooking	1.0	0.2	0.3	1.5
Dishwashing	0.3	0.1	0.1	0.5
Eating	0.5	0.1	0.2	0.8
Other	0.6	0.2	0.2	1.0
Social activity	1.0	0.3	0.3	1.6
Vacuum cleaner	0.2	0.1	0.1	0.3
Watching TV	8.3	2.1	2.6	13.0
Working	7.9	2.0	2.5	12.3
Total	28.2	7.1	8.8	44.1

was prepared. Two metrics were used to assess completion models. The first one was the MSE between the input and output spectrograms of the completio n models in the missing data channel, hereafter referred to as In-Out MSE. The other metric was the MSE between the completion model output and the true values of the spectrograms, hereafter referred to as Out-True MSE. These metrics were measured in each completing iteration. The training conditions for the completion models were 1000 epochs using the Adam optimizer [17], whose parameters were set as a learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. In-Out MSE was used as a loss function. The number of completing iterations was 30. The differences among the four channels were basically small because the spatial information was captured from the small differences in each channel. Therefore, the completion performance of the proposed method strongly depended on the initial value of the missing data spectrogram. From our preliminary experiments, we recognized that a constant

value was not suitable for the initial value of the missing data spectrogram. Thus, the initial value of the missing data spectrogram was assigned as the average of the other channels. In the experiment, we assumed that the channel causing the problem was known, and the missing channel was set in microphone No. 2.

In the experiment to evaluate the performance of some classification systems by using the completed signals, we used three ASC models: a convolutional neural network (CNN), a convolutional recurrent neural network (CRNN), and a patch-based convolutional network. ResNet50 [13] was used as a representative CNN model. CRNN has been used for document classification [32]. It combines the advantages of a CNN for capturing local features and an RNN for the temporal summarization of the input. In this experiment, a modified CRNN [5] inspired by a model proposed for musical classification was used. As a patch-based convolutional network, ConvMixer [33], which is a state-of-the-art classification model for image classification tasks was used. It has been reported that the patch embeddings installed in the first layer of ConvMixer help it to catch input features effectively. The training settings for the ASC models were as follows. The number of training iterations was 100 epochs, SpecAugment [25] was adopted for data augmentation, and the Adam optimizer [17] was used, where the parameters were set as a learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. To evaluate the performance of the proposed method in the ASC task, we prepared three conditions for testing.

1. Clean
   Clean refers to the condition that four-channel log-Mel amplitude spectrograms have no missing data. This condition was used to one of the upper bounds of the proposed method.

2. Missing
   Missing refers to the condition of missing data in microphone No. 2. The missing data was represented as the average of the other channels.

3. Completed
   Completed refers to the output of the proposed method. Missing data was completed by passing it through the completion model repeatedly.

As the metric of the ASC evaluation, the macro F-score [8], which considers the difference in the amount of data in each class was used.

In the evaluation of the robustness of the completion models, we investigated the performance when the completion and ASC models were trained using the microphones in different positions. We evaluated the behavior of the completion models for two situations.

> Situation 1: the positions of the microphones used for training the completion models were shifted compared with those of the missing

and completing conditions as shown in Figure 6. Even though the microphone positions were different for the completion and ASC models, both relationships of the microphone positions were similar.

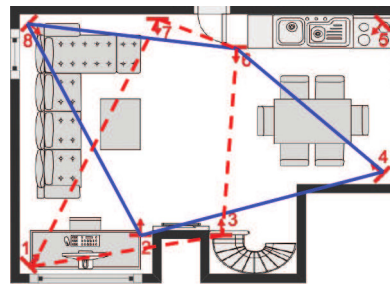

Figure 6: Microphone positions for completion models and ASC models for Situation 1.

Situation 2: The positions of the microphones for the ASC models were in the crossed position of those for the completion models, as shown in Figure 7.

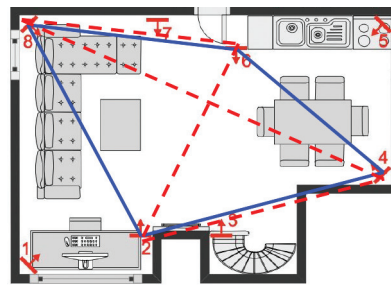

Figure 7: Microphone positions for completion models and ASC models for Situation 2.

4.3 Comparison of Each Completion Model

Figure 8 shows the example signals of Clean, Missing, and Completed conditions. By comparing Clean with Missing, Missing, which was the average of the other channels, has large values at the red circle. In multi-channel acoustic scene classification, since the spatial information is calculated from slight differences between microphones, the average of the other channels degrades classification performance. In the Completed condition, the values at the red circle are getting closer to that of Clean. This indicated that the proposed method could perform to complete the relative value of Clean from the average values in Missing. Same as the previous example, some signals at the red box in Missing were getting weak in Clean and Completed conditions.

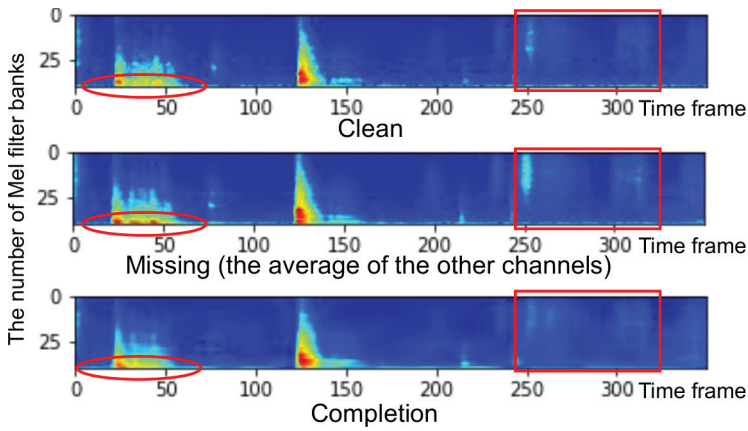

Figure 8: Example signals of Clean, Missing (the average of the other channels), and Completed conditions.

This denoted that the proposed method completed some proper signals for representing some missing data.

Figure 9 shows the trajectories of In-Out MSE and Out-True MSE for each completion iteration in each completion model. For CAE-M, it can be seen that In-Out MSE converged and Out-True MSE gradually increased, but the value did not increase. It is considered that the missing data was satisfactorily completed since CAE-M had sufficient number of parameters to maintain acoustic features and channel relationships. For CAE-L and CAE-latent-S, In-Out MSE of CAE-L and CAE-latent-S converged to zero, in contrast to CAE-M. This result indicates that the completion model hardly changed the input data. It is considered that CAE-L failed to learn acoustic features and channel relationships because it has too many parameters for the dataset. For CAE-latent-S, since the latent space is too small to maintain acoustic features and channel relationships, CAE-latent-S was unable to generate details of the spectrograms. In-Out MSE had similar values for CAE-M with CAE-latent-L. Unlike In-Out MSE, Out-True MSE of CAE-latent-L had an almost constant value. It is considered that CAE-latent-L performed an identical transformation because the input data was not compressed owing to the large latent space.

For CAE-S and U-Net, both In-Out MSE and Out-True MSE were increased with the number of iterations. It is considered that the modeled latent space of CAE-S was insufficient to generate a sufficiently detailed spectrogram owing to the lack of parameters. Since the deficient spectrograms generated by CAE-S were input into CAE-S repeatedly, In-Out MSE increased. For U-Net, In-Out MSE was low at the beginning of the completing iterations. It is considered that the output data was similar to the input data because of the

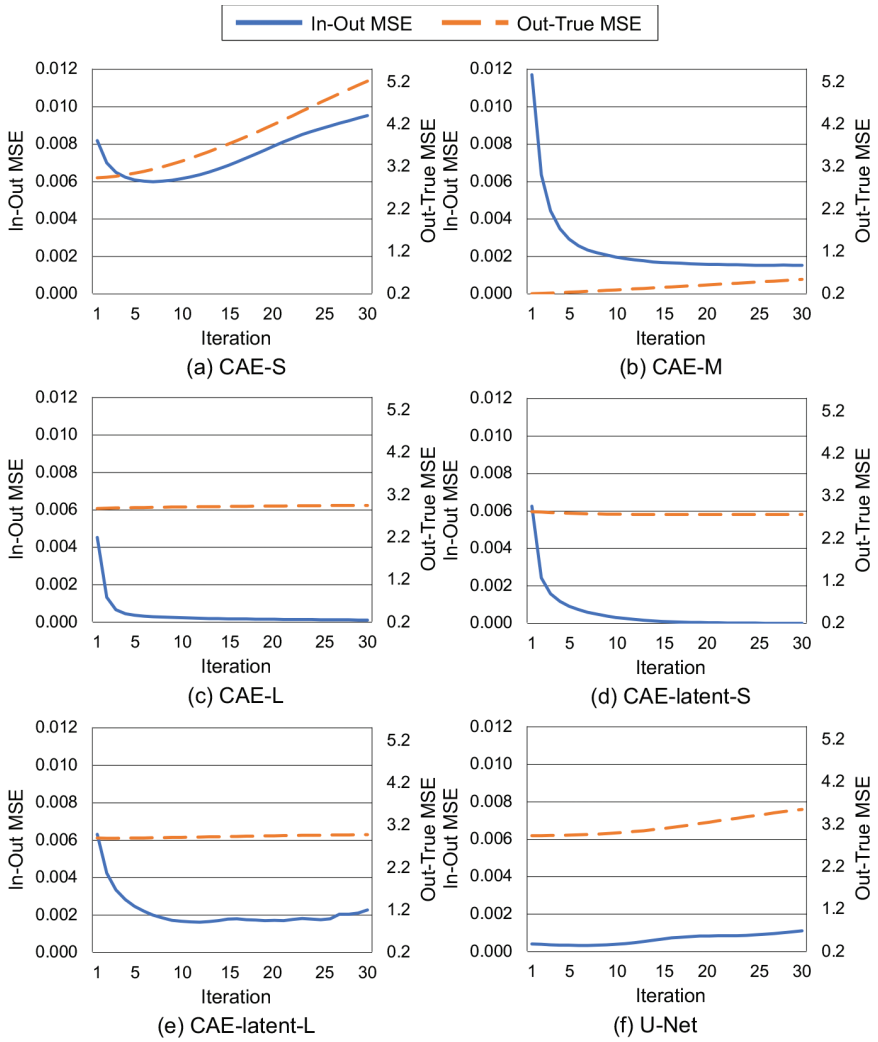

Figure 9: Trajectories of In-Out MSE and Out-True MSE for each completion iteration in each completion model.

skip connection. The increases in In-Out MSE and Out-True MSE for U-Net indicate that U-Net did not complete data.

4.4 Classification Performance with Completed Signals

Table 3 shows the F-scores of the clean, missing, and completed conditions for each completion model. Focusing on the CRNN, the F-score of missing was

Table 3: Classification results of clean, missing, and completed conditions with different completion models described in F-score [%].

Completion model	Conditions	CNN	CRNN	ConvMixer
-	Clean	95.94	95.48	95.13
	Missing	71.21	30.25	79.71
CAE-S	Completed	72.93	63.30	76.35
CAE-M	Completed	83.36	81.94	**91.39**
CAE-L	Completed	91.24	**87.75**	86.42
CAE-latent-S	Completed	**91.35**	87.63	88.55
CAE-latent-L	Completed	87.36	84.85	84.67
U-Net	Completed	69.89	37.24	80.59

Table 4: Precision and recall scores of the experimental results described in Table 3.

Completion model	Condition	CNN	CRNN	ConvMixer
-	Clean	96.07 / 96.13	95.91 / 95.22	95.45 / 95.41
	missing	84.69 / 75.24	83.00 / 39.89	86.76 / 80.91
CAE-S	Completed	85.03 / 76.34	84.11 / 68.17	85.56 / 78.33
CAE-M		87.73 / 81.75	89.82 / 79.50	**92.26 / 91.35**
CAE-L		91.38 / 91.41	89.57 / 86.73	86.60 / 86.47
CAE-latent-S		**91.82 / 91.70**	**89.12 / 87.32**	**89.29 / 88.83**
CAE-latent-L		88.96 / 87.67	87.28 / 84.55	85.21 / 84.49
U-Net		83.75 / 73.42	58.75 / 30.36	86.92 / 81.57

extremely low. This indicates that the CRNN was vulnerable to a mismatch between the training and testing data. Comparing clean and missing, the F-scores of missing of CNN and ConvMixer were also lower than that of clean even though ConvMixer is the state-of-the-art classification model in image classification. The results indicated that missing data degraded the ASC systems markedly. On the other hand, comparing missing and completed, the F-score of completed of almost all CAE-based completion models was improved for all ASC models. In particular, completion using CAM-M, CAE-L, and CAE-latent-S, whose In-Out MSE converged as shown in Figure 9, improved the F-score greatly. The results indicate that the completion method for multi-channel missing data is an effective preprocess regardless of the ASC model.

Table 4 represents the precision and recall scores, which corresponded to the F-scores of Table 3 in the submitted paper. CAE-M with ConvMixer, which achieved the highest F-score, has a small difference between the precision and recall scores but CAE-M with CNN and CRNN the larger differences such as 5.98 and 10.32 than 0.91 of ConvMixer. On the other hand, CAE-latent-S, which achieved almost the same F-score as CAE-M has a small difference

between the precision and recall score with any classification models. This indicated that CAE-latent-S is more stable than CAE-M in the proposed method.

Figure 10 shows the trajectories of the F-score using three types of ASC models for each completion iteration in each completion model. The result for iteration 0 represents the initial value input in the completion models used for ASC. By the end of the completing iterations, the F-score had decreased in

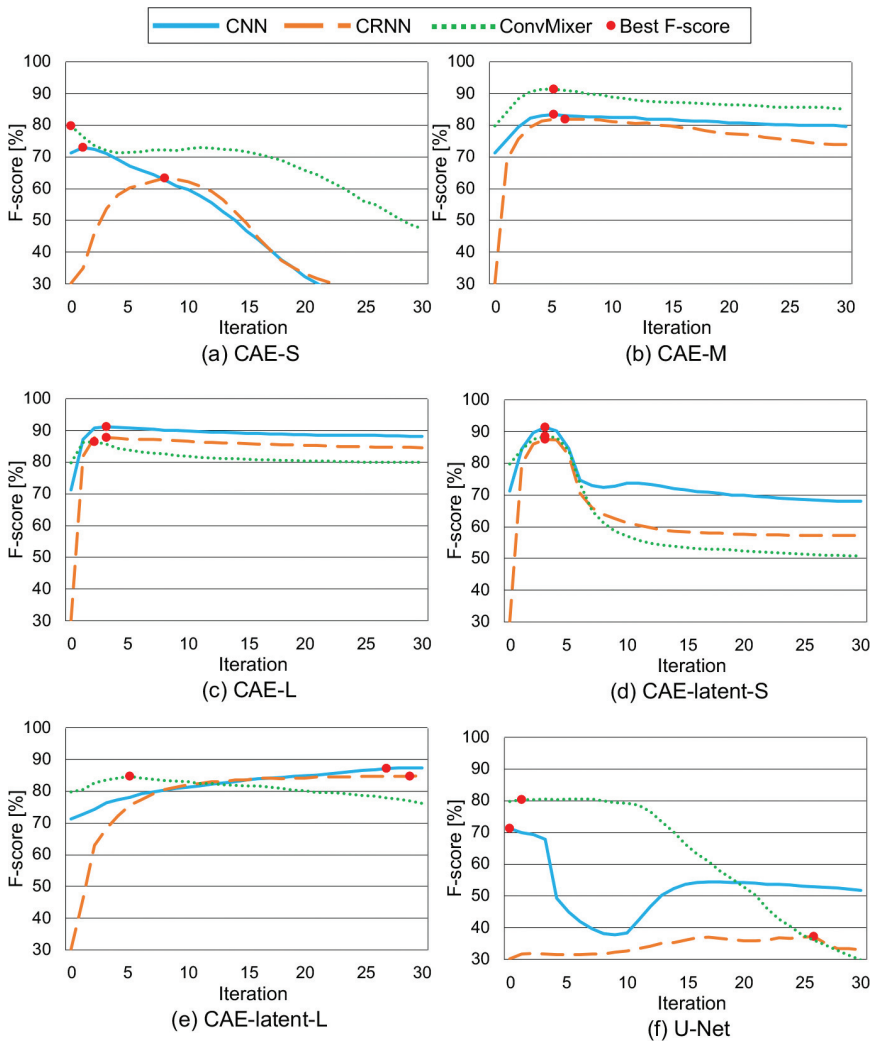

Figure 10: Trajectories of F-score using three types of ASC models for each completion iteration in each completion model.

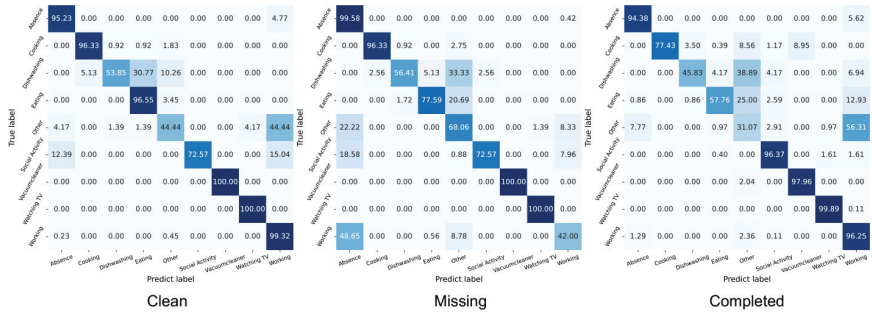

Figure 11: Confusion matrices of the experimental results using ConvMixer and CAE-M described in Table 3.

Table 5: F-scores of clean, missing, and completed conditions with CAE-M and CAE-latent-S in Situations 1 and 2[%].

Completion model	Conditions	CNN	CRNN	ConvMixer
-	Clean	95.94	95.48	95.13
	Missing	71.21	30.25	79.71
CAE-M	Completed (Situation 1)	**80.14**	**74.79**	**81.97**
	Completed (Situation 2)	74.29	60.38	78.26
CAE-latent-S	Completed (Situation 1)	**82.38**	**76.22**	**85.41**
	Completed (Situation 2)	70.11	58.63	75.91

almost all the completion models. This indicates that over-completion occurred in the latter half of the completing iterations. From this investigation, an assessment to stop the iteration for obtaining an adequate completion signal.

Comparing CAE-S with U-Net in Figure 10, the F-score of the completion models deteriorated. In contrast, In-Out MSE of these models gradually increased (Figure 9). According to these results, a completion model whose In-Out MSE gradually increases is inappropriate for ASC.

Figure 11 shows the experimental results using ConvMixer and CAE-M described in Table 3. Comparing Clean with Missing, the F-score of "Working" and "Eazting." was degraded. This denoted that the lack of spatial information affected the classification performance of "Working" and "Eating." Comparing Missing with Completed, the F-score of "Working" was increased considerably. This indicated that the proposed method contributed that classifier trained with clean data identified "Working" correctly by completing spatial information.

4.5 Robustness of Completion Models

Table 5 shows the F-scores in Situations 1 and 2 with CAE-M and CAE-latent-S. In both situations, the proposed method was evaluated using the

ASC task with three ASC models, namely, the CNN, CRNN, and ConvMixer. Comparing Situations 1 and Situation 2, all F-scores of the completed data in Situation 1 were higher than those in Situation 2. The results indicate that the proposed CAE modeled the relationship of the microphone position in the latent space. Therefore, even though the ASC systems were trained with microphones in different positions, when the relationship of the microphone position is similar, the proposed method satisfactorily perform.

5 Conclusion

In this research, we proposed an autoencoder-based missing data completion method for multi-channel signals. As a preprocess of ASC models, the proposed method iteratively completed the missing data by using the relationship between multi-channel signals. From the experimental results, we confirmed that the proposed method, which had a sufficient number of model parameters, reconstructed the missing data adequately for improving the ASC models. In future work, we will use the proposed system in more complicated situations and perform some assessments by selecting sufficient number of iterations to reconstruct the missing data. We will use other completion models such as transformer and variational autoencoder. Moreover, we will consider semi-supervised domain adaptation and knowledge transfer learning as one of the applications for the proposed method.

Acknowledgement

This work was supported in part by JSPS KAKENHI Grant numbers JP20H00613, and ROIS DS-JOINT (021RP2022) to S. Shiota.

Biographies

Yuki Shiroma received his B.E. degree in Tokyo Metropolitan University, Tokyo, Japan in 2021. His research interests include acoustic scene classification and musical instrument classification.

Yuma Kinoshita received his B.Eng., M.Eng., and the Ph.D. degrees from Tokyo Metropolitan University, Japan, in 2016, 2018, and 2020 respectively. In April 2020, he started to work with Tokyo Metropolitan University, as a project assistant professor. He moved to Tokai University, Japan, as an associate professor/lecturer in April 2022. He became a project associate professor at Tokyo Metropolitan University in April 2023. His research interests are in the

area of signal processing, image processing, and machine learning. He is a Member of IEEE, APSIPA, IEICE, and ASJ. He received the IEEE ISPACS Best Paper Award, in 2016, the IEEE Signal Processing Society Japan Student Conference Paper Award, in 2018, the IEEE Signal Processing Society Tokyo Joint Chapter Student Award, in 2018, the IEEE GCCE Excellent Paper Award (Gold Prize), in 2019, the IWAIT Best Paper Award, in 2020, and the APSIPA ASC 2021 Best Paper Award. He was a Registration Chair of DCASE2020 Workshop.

Keisuke Imoto received his B.E. and M.E. degrees from Kyoto University in 2008 and 2010, respectively. He received his Ph.D. degree from SOKENDAI (The Graduate University for Advanced Studies) in 2017. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 2010 and the Ritsumeikan University as an Assistant Professor in 2017. He moved to Doshisha University as an Associate Professor in 2020. He has been engaged in research on sound event detection, acoustic scene analysis, anomalous sound detection, and microphone array signal processing. He is a member of IEEE, EURASIP, APSIPA, ASJ, and IEICE.

Sayaka Shiota received her B.E., M.E., and Ph.D. degrees in intelligence and computer science, Engineering, and engineering simulation from Nagoya Institute of Technology in 2007, 2009, and 2012, respectively. From February 2013 to March 2014, she worked as a project assistant professor at the Institute of statistical mathematics. In 2014, she joined Tokyo Metropolitan University as an assistant professor and became an associate professor in 2023. Her research interests include statistical speech recognition and speaker verification. She is a member of ASJ, IPSJ, IEICE, APSIPA, ISCA, and IEEE.

Nobutaka Ono received his B.E., M.S., and Ph.D. degrees from the University of Tokyo, Japan, in 1996, 1998, and 2001, respectively. He became a research associate in 2001 and a lecturer in 2005 at the University of Tokyo. He moved to the National Institute of Informatics in 2011 as an associate professor and then to Tokyo Metropolitan University in 2017 as a full professor. His research interests include acoustic signal processing, especially microphone array processing, source localization and separation, machine learning, and optimization algorithms. He is a member of IEEE, EURASIP, APSIPA, IPSJ, IEICE, and ASJ. He was a member of IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee from 2014 to 2019. He served as Associate Editor of IEEE Transactions on Audio, Speech, and Language Processing from 2012 to 2015. He received the best paper award at APSIPA ASC in 2018 and 2021 and Sadaoki Furui Prize Paper Award from APSIPA in 2021.

Hitoshi Kiya received his B.E and M.E. degrees from Nagaoka University of Technology, in 1980 and 1982 respectively, and his Dr. Eng. degree from

Tokyo Metropolitan University in 1987. In 1982, he joined Tokyo Metropolitan University, where he became a Full Professor in 2000. From 1995 to 1996, he attended the University of Sydney, Australia as a Visiting Fellow. He is a Fellow of IEEE, IEICE and ITE. He served as President of APSIPA from 2019 to 2020, and as Regional Director-at-Large for Region 10 of the IEEE Signal Processing Society from 2016 to 2017. He was also President of the IEICE Engineering Sciences Society from 2011 to 2012. He was Editorial Board Member of eight journals, including IEEE Trans. on Signal Processing, Image Processing, and Information Forensics and Security. He has organized a lot of international conferences, in such roles as TPC Chair of IEEE ICASSP 2012 and as General Co-Chair of IEEE ISCAS 2019. He has received numerous awards, including 12 best paper awards.

References

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying Environments from the Sounds They Produce," *IEEE Signal Processing Magazine*, 32(3), 2015, 16–34.

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association*, 2018, 1561–5.

[3] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A System for Monitoring, Analyzing, and Mitigating Urban Noise Pollution," *Communications of the ACM*, 62(2), 2019, 68–77.

[4] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the Data Augmentation Scheme With Various Classifiers for Acoustic Scene Modeling," *tech. rep.*, Detection, Classification of Acoustic Scenes, and Events (DCASE) 2019 Challenge, 2019.

[5] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional Recurrent Neural Networks for Music Classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, 2392–6.

[6] "DCASE 2018 Challenge Task5," https://dcase.community/challenge2018/task-monitoring-domestic-activities.

[7] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van 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," in *The Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE)*, 2017, 32–6.

[8] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of Domestic Activities Based on Multi-Channel Acoustics," *tech. rep.*, Detection, Classification of Acoustic Scenes, and Events (DCASE) 2018 Challenge, 2018.

[9] G. Del Galdo, O. Thiergart, T. Weller, and E. A. Habets, "Generating Virtual Microphone Signals Using Geometrical Information Gathered by Distributed Arrays," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2011, 185–90.

[10] T. B. Duman, B. Bayram, and G. İnce, "Acoustic Anomaly Detection Using Convolutional Autoencoders in Industrial Processes," in *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)*, 2020, 432–42.

[11] M. C. Green and D. T. Murphy, "Acoustic Scene Classification Using Spatial Features.," *tech. rep.*, Detection, Classification of Acoustic Scenes, and Events (DCASE) 2017 Challenge, 2017, 42–5.

[12] D. Griffin and J. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2), 1984, 236–43.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 770–8.

[14] H. Hu, S. M. Siniscalchi, C.-H. H. Yang, and C.-H. Lee, "A Variational Bayesian Approach to Learning Latent Variables for Acoustic Knowledge Transfer," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, 4041–5.

[15] K. Imoto and N. Ono, "Spatial Cepstrum as a Spatial Feature Using a Distributed Microphone Array for Acoustic Scene Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6), 2017, 1335–43.

[16] B. Kim, S. Yang, J. Kim, and S. Chang, "QTI Submission to DCASE 2021: Residual Normalization for Device-Imbalanced Acoustic Scene Classification with Efficient Design," *tech. rep.*, Detection, Classification of Acoustic Scenes, and Events (DCASE) 2021 Challenge, 2021.

[17] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[18] Y. Kinoshita and N. Ono, "Analysis on Roles of DNNs in End-to-End Acoustic Scene Analysis Framework with Distributed Sound-to-Light Conversion Devices," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, 1167–72.

[19] Y. Kinoshita and N. Ono, "End-to-End Training for Acoustic Scene Analysis with Distributed Sound-to-Light Conversion Devices," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, 1010–4.

[20] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, 5220–4.

[21] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and Discussion on DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," in *the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, 81–5.

[22] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Acoustic Scene Classification and Audio Tagging with Receptive-Field-Regularized CNNs," *tech. rep.*, Detection, Classification of Acoustic Scenes, and Events (DCASE) 2019 Challenge, 2019.

[23] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2), 2018, 379–93.

[24] T. Ochiai, M. Delcroix, T. Nakatani, R. Ikeshita, K. Kinoshita, and S. Araki, "Neural Network-Based Virtual Microphone Estimator," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 6114–8.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019 - 20th Annual Conference of the International Speech Communication Association*, 2019, 2613–7.

[26] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2020, 684–98.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, 234–41.

[28] Y. Shiroma, K. Imoto, S. Shiota, N. Ono, and H. Kiya, "Investigation on Spatial and Frequency-Based Features for Asynchronous Acoustic Scene Analysis," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, 1161–6.

[29] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing Acoustic Scene Classification Models with CNN Variants," *tech. rep.*, Detection, Classification of Acoustic Scenes, and Events (DCASE) 2020 Challenge, 2020.

[30] Y. Takahashi, S. Takamuku, K. Imoto, and N. Natori, "Semi-Supervised Domain Adaptation for Acoustic Scene Classification by Minimax Entropy and Self-Supervision Approaches," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, 1–5.

[31] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi, and K. Hamada, "Multichannel Acoustic Scene Classification by Blind Dereverberation, Dlind Source Separation, Data Augmentation, and Model Ensembling," *tech. rep.*, Detection, Classification of Acoustic Scenes, and Events (DCASE) 2018 Challenge, 2018.

[32] D. Tang, B. Qin, and T. Liu, "Document Modeling With Gated Recurrent Neural Network for Sentiment Classification," in *The 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, 1422–32.

[33] A. Trockman and J. Z. Kolter, "Patches Are All You Need?" *arXiv preprint*, 2022.

[34] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Abnormal Sound Detection by Two Microphones using Virtual Microphone Technique," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, 478–82.

[35] K. Yatabe, Y. Masuyama, T. Kusano, and Y. Oikawa, "Representation of Complex Spectrogram via Phase Conversion," *Acoustical Science and Technology*, 40(3), 2019, 170–7.