Training Acoustic Scene Classification Models Robust to Asynchrony in Distributed Microphone Arrays

Takao Kawamura* and Nobutaka Ono*
* Tokyo Metropolitan University, Japan
E-mail: kawamura-takao@ed.tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this study, we explore whether phase-based spatial features can still provide informative cues for acoustic scene classification under asynchronous conditions. Distributed microphone arrays can capture spatial information over a large area, such as time differences between microphone recordings, which are useful for acoustic scene classification. However, in real-world applications, this information becomes unreliable due to sampling time offsets (STOs) caused by unsynchronized recording start times or clock drift. In our previous work, we showed that models trained on phase-based spatial features using synchronized data suffer significant performance degradation when tested on asynchronous recordings. Nevertheless, we hypothesize that even under STOs, phase-based features may still contain useful spatial cues, such as whether a sound source is moving or stationary. To examine this hypothesis, we apply a data augmentation strategy that simulates asynchronous conditions by introducing circular shifts to synchronized data and then compute phase-based spatial features. This approach allows the model to learn representations that are robust to STOs. Experimental results demonstrate that our method mitigates performance degradation across various asynchronous conditions.**

I. Introduction

Acoustic scene recognition is a technology that identifies the surrounding environment from sound. It has a wide range of applications, such as monitoring infants and older people [1], [2] and surveillance systems [3], [4]. Acoustic scene classification, one of the key technologies in acoustic scene recognition, aims to classify a pre-defined scene (e.g., “Cooking” or “Watching TV”) from a few seconds of recorded signal (hereinafter referred to as an audio clip) [5]. In acoustic scene classification, many studies have used spectral features [6]–[10]. Meanwhile, when multiple microphones are available, we can also utilize spatial information such as time differences and power ratios between microphones. Combining spatial information with spectral features is expected to improve performance. In particular, phase-based spatial features are considered to have information complementary to spectral features [11].

In multiple microphone frameworks, distributed microphone arrays have been studied. Distributed microphone arrays consist of spatially distributed microphones. Distributed microphone arrays enable obtaining spatial information over a wide area and have been applied to tasks such as sound source localization [12] and acoustic scene classification [11], [13]–[16].

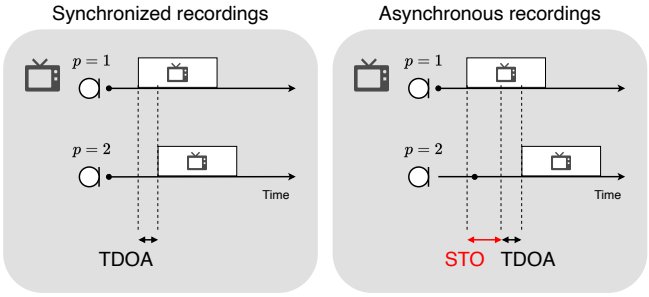

Fig. 1. Time differences between microphones calculated from synchronized and asynchronous recordings. These time differences vary due to the presence of STO.

However, practical applications need to handle asynchronous recordings caused by factors such as sampling time offsets (STOs) and sampling rate offsets (SROs) [17]–[20]. For example, STOs can arise from slight timing mismatches when recordings are manually started. Even if the initial recording start times are synchronized, SROs can cause time drift that accumulates over time, leading to time offsets similar to STOs in practical scenarios. In this study, we focus on acoustic scene classification in the presence of STOs.

In microphone array signal processing, spatial information such as the time difference of arrivals (TDOAs) between microphone recordings plays an important role. However, when STOs exist, the TDOAs between microphone recordings cannot be correctly computed. Figure 1 illustrates an example of the observed time difference between microphone recordings using synchronized data and asynchronous data affected by STO. As shown in the figure, when STO exists, the time difference between microphone recordings may correspond to the sum of the actual TDOA and STO. This indicates that the observed time differences no longer correspond to the actual TDOA, making the spatial information unreliable in the presence of STOs.

This motivates a key research direction: investigating whether spatial information can still be leveraged under asynchronous conditions. One approach to address this issue has involved the use of amplitude-based spatial features related to the distribution of microphone power values [13], [14]. These features do not retain time difference information but may still capture useful spatial cues, provided that frame-level alignment

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

is preserved. In contrast, phase-based features contain time difference information and are therefore inherently sensitive to STOs. While synchronization methods can be applied as a preprocessing step to compensate for STOs, they pose challenges such as the difficulty of distinguishing between TDOAs and STOs and the additional computational cost.

In this study, we investigate whether phase-based features can still provide informative cues for classification under asynchronous conditions, such as indicating whether the sound source is moving or stationary. To this end, we adopt a data augmentation method that simulates asynchronous conditions by applying circular shifts to synchronized recordings. The model is trained using phase-based features computed from these augmented signals. This approach is expected to encourage the model to learn representations that are invariant to STOs. In evaluation experiments under various STOs, we confirmed that the proposed method mitigated the degradation in classification performance. These findings suggest that phase-based features contain spatial information that can be exploited for classification even in the presence of STOs.

II. Acoustic Scene Classification Using Asynchronous Data

A. Problem Setting

We consider a scenario where audio clips are synchronized during training but asynchronous during testing. Let the audio clip recorded by microphone p ($p = 1, \ldots, \mathrm{P}$) be denoted as $x_p[n]$ during training and as $x_p[n - \tau_p]$ during testing, where $n = 1, \ldots, \mathrm{N}$. Here, P and N represent the number of microphones and the length of each audio clip, and τ_p denotes the STO of microphone p (with $\tau_1 = 0$), respectively. When STOs exist, spatial information such as TDOA cannot be accurately calculated. Figure 1 illustrates the example of time differences between microphone recordings under both synchronized and STO-affected conditions. As shown in the figure, the presence of STOs prevents the acquisition of spatial information.

In acoustic scene classification, a model is trained using acoustic features computed from audio clips as input. As illustrated in the above example, spatial features vary depending on whether the data are synchronized or asynchronous due to STOs. Such differences are expected to degrade classification performance [11]. To address this issue, we investigate a training strategy for achieving robustness to STOs in acoustic scene classification, assuming that only synchronized data are available. We propose a data augmentation method to mitigate the impact of STOs on spatial features.

B. Spatial Feature

In this study, we use GCC-PHAT [21] as the spatial feature. GCC-PHAT is commonly used to estimate the TDOA. It has been employed as a spatial feature in various tasks, including sound event detection [22], sound event localization and detection [23], [24], and acoustic scene classification [11].

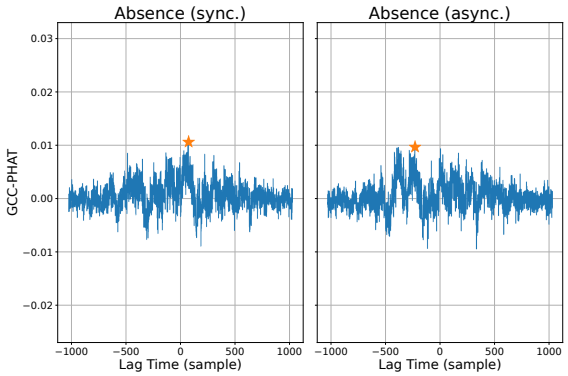

Fig. 2. Example of GCC-PHAT for the label "Absence." Here, "sync." and "async." represent GCC-PHAT calculated from synchronized and asynchronous recordings (with an STO of 300 samples), respectively. "*" indicates the peak position.

GCC-PHAT, which is used in this study, is calculated by the following equation:

$$\phi_{p,q}(\tau) = \mathcal{F}^{-1}_{f\to\tau}\left(\frac{1}{T}\sum_{t}\frac{X_p(f,t)X_q^*(f,t)}{|X_p(f,t)||X_q(f,t)|}\right), \quad (1)$$

where $X_p(f,t)$ denotes the spectrogram computed from the audio clip $x_p[n]$, and $\mathcal{F}^{-1}_{f\to\tau}(\cdot)$ indicates the inverse Fourier transform. In this study, we simplify the representation by temporal averaging; however, in the future, it would be desirable to consider more complex representations, such as learnable aggregation operations instead of temporal averaging.

To observe the effect of STO, we examined examples of GCC-PHAT feature with and without the STO. GCC-PHAT feature was computed using representative microphones from microphone arrays 1 and 2 in Fig. 4. We set the STO of subarray 2 to 300 samples. Figures 2 and 3 show the GCC-PHATs features for the scenes "Absence" and "Watching TV," respectively. In the figures, "sync." and "async." indicate the synchronized (STO was 0 sample) and asynchronous (STO was 300 sample) conditions. The parameters used to compute GCC-PHAT are described in Section III-A.

In Fig. 2, we observed a shift in the peak position. On the other hand, clear peaks did not appear in either the synchronized or asynchronous case, and the shapes of GCC-PHATs were similar. In Fig. 3, we also observed a shift in the peak position due to the presence of STO. Compared to "Absence," the difference in the shapes of GCC-PHATs caused by STO was more pronounced in the "Watching TV." It suggests that such shape difference could negatively impact acoustic scene classification performance.

C. Data Augmentation for Asynchronous Recording

Data augmentation is a widely used technique for increasing training data to improve noise robustness and generalization, and various methods have been proposed [25]–[28]. When using a single microphone, methods such as Mixup [25] and SpecAugment [26] have been applied to spectral features, and time shifting is also an effective augmentation technique for spectral features [29], [30]. On the other hand, when

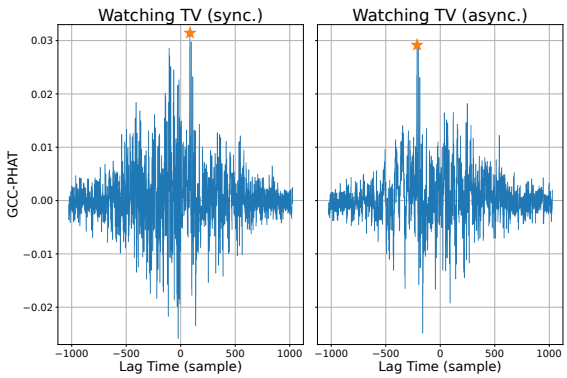

Fig. 3. Example of GCC-PHAT for the label "Watching TV." Here, "sync." and "async." represent GCC-PHAT calculated from synchronized and asynchronous recordings (with an STO of 300 samples), respectively. "*" indicates the peak position.

multiple microphones are available, techniques such as ChannelSwap [27] and change in hop length of short-time Fourier transform (STFT) [28] have been proposed to augment spatial information while preserving spectral information.

In this study, we propose a data augmentation technique that simulates asynchronous signals caused by STOs. By applying the proposed method to synchronized audio clips during training, we aim to mitigate the performance degradation that occurs when using asynchronous audio clips during testing.

The asynchronous signal caused by STO is simulated as a circular shift by an integer number of samples. Unlike previous augmentation methods for single-channel recordings, we adapt this circular shift to augment asynchronous multi-channel signals. The circularly shifted audio clip $\tilde{x}_p[n]$ is defined by

$$\tilde{x}_p[n] = x_p[n - \tilde{\tau}_p], \tag{2}$$

where $\tilde{\tau}_p$ represents the simulated STO for subarray p, with $\tilde{\tau}_1 = 0$. Each $\tilde{\tau}_p$ is sampled from a normal distribution with a mean of 0 and a standard deviation of σ, where σ is treated as a hyperparameter. During training, the GCC-PHATs defined in (1) using the augmented audio clips $\tilde{x}_p[n]$ are calculated, and then the model is trained using the calculated GCC-PHATs. Note that since $\tilde{\tau}_p$ is sufficiently small compared to the clip length N, we assume that the wrap-around effect due to circular shifting can be ignored.

III. Experimental Evaluation

In this experiment, we evaluated the effectiveness of the proposed method on simulated asynchronous conditions.

A. Experimental Setup

We evaluated the effectiveness of the proposed method using the SINS database [9], a real-world dataset that contains continuous recordings of a single person's activities spanning one week. The dataset consists of the recordings of 13 microphone arrays, whose spatial configuration is shown in Fig. 4. Each microphone array is a linear array consisting of four microphones. In this study, we used the first channel of each of the seven arrays (arrays 1–4 and 6–8) in the living room as a distributed microphone array. The dataset contains 10 scenes: "Absence," "Calling," "Cooking," "Dishwashing," "Eating," "Other," "Vacuum cleaner," "Visit," "Watching TV," and "Working." The audio signal was recorded at a sampling frequency of 16 kHz.

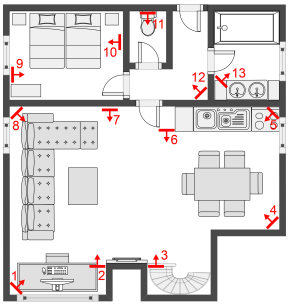

Fig. 4. Microphone array arrangement in the SINS database [9]. Red arrows indicate the linear 4-channel microphone arrays.

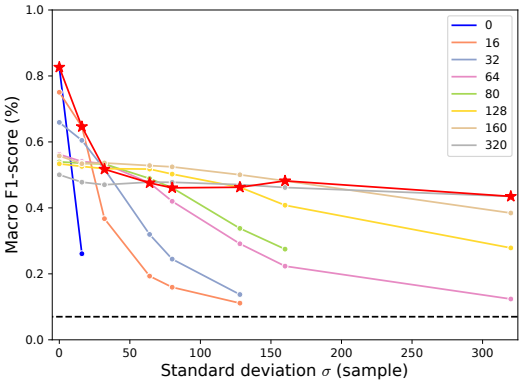

Fig. 5. Performance of acoustic scene classification. The vertical axis indicates the macro F1-score, and the horizontal axis indicates the standard deviation of the STO at testing. The legend represents the standard deviation of the STOs used in training.

In the dataset, synchronization was performed based on time stamps, which were stored every second. For acoustic scene classification, 10-second audio clips were segmented from the continuous recordings. In our experiments, the training, validation, and test sets consisted of 40,794, 6,666, and 4,753 clips, respectively. We used the synchronized dataset for the training and validation sets, while the test set was simulated with various STOs. The same microphones were used in both the training and test sets.

The network architecture and training conditions used in this experiment followed those described in [11], [16], except for the data augmentation settings. For the input features, GCC-PHAT was computed for all 21 pairs from the seven microphones, corresponding to the combination $\binom{7}{2} = 21$. Each GCC-PHAT feature was calculated with a frame length of 128 ms and 50% overlap. The resulting features from all pairs were concatenated and used as input to the model. A three-layer fully connected neural network was employed as the model. The model was trained using an AdamW optimizer [31] for 50 epochs with a learning rate of 1.0×10^{-4} and a weight decay of 1.0×10^{-5}. The loss function used was cross-entropy. To mitigate the data imbalance problem, we sampled audio clips of each scene label equally during training. For

TABLE I
THE F1-SCORE UNDER THE MATCHED CONDITION FOR EACH LABEL.

Std.	F1-score									
	Absence	Calling	Cooking	Dish washing	Eating	Other	Vacuum cleaner	Visit	Watching TV	Working
0	92.5%	64.8%	92.2%	80.0%	92.0%	42.2%	100.0%	72.3%	99.8%	90.7%
16	62.6%	52.3%	90.2%	75.0%	50.5%	20.8%	78.5%	48.8%	93.3%	74.4%
32	53.3%	29.7%	86.5%	60.1%	17.9%	14.6%	58.0%	22.6%	97.9%	76.6%
80	34.4%	29.9%	83.1%	31.1%	8.4%	17.1%	70.4%	11.5%	98.0%	76.8%
160	21.4%	37.0%	84.0%	15.1%	21.8%	27.4%	86.0%	14.7%	97.5%	77.0%
320	37.4%	35.7%	87.7%	9.2%	9.4%	7.6%	81.7%	17.4%	94.5%	54.1%

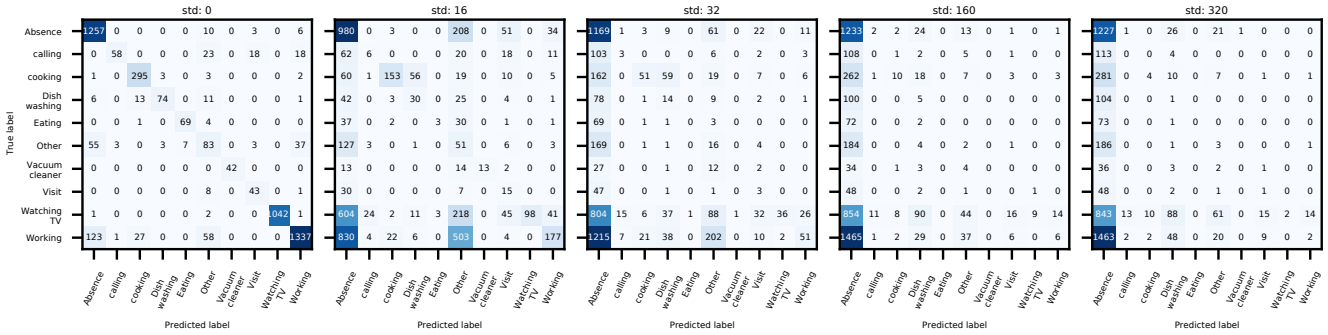

Fig. 6. The confusion matrix obtained when trained with synchronized recordings.

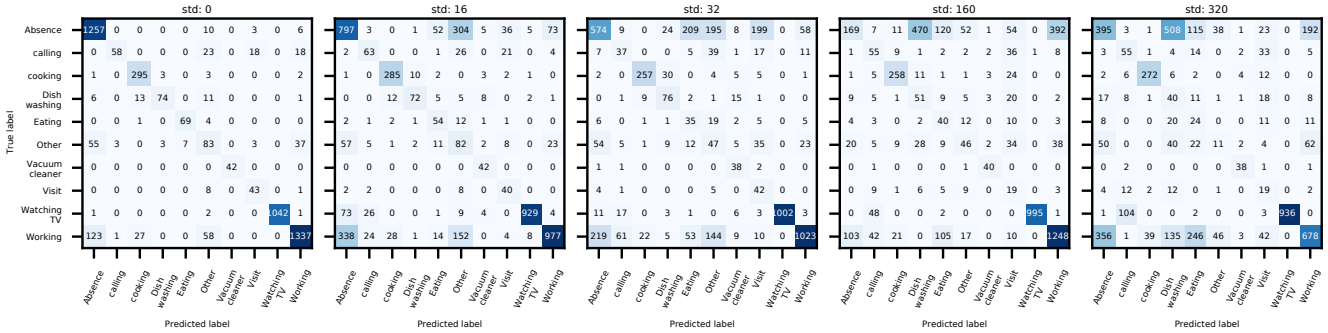

Fig. 7. The confusion matrix under the matched condition.

data augmentation, the standard deviation σ were set to 0 (synchronized data), 16, 32, 64, 80, 128, 160, and 320 samples. A separate model was trained and saved for each standard deviation.

We created a test set for each standard deviation value: 0, 16, 32, 64, 80, 128, 160, and 320 samples. For each test set, integer-sample STOs were independently simulated for each microphone and each audio clip. To evaluate the robustness against STO, all trained models were evaluated on every test set. The evaluation metric was the macro F1-score, calculated as the average of the F1-scores across all labels.

B. Experimental Results

Figure 5 shows the relationship between the macro F1-score and the standard deviation of STO in the test data. The legend indicates the standard deviation σ used in data augmentation during training. When the STO was 0, no data augmentation was applied, and the model was trained only with synchronized data. The red line represents the results under the matched condition, where the standard deviation of STO during training and testing matched. The results show that the macro F1-score decreased as the standard deviation of STO increased. When the model was trained on synchronized data ($\sigma = 0$), the classification performance degraded significantly with increasing STO. In contrast, applying the proposed method mitigated the performance degradation. The macro F1-scores of the matched condition exceeded the chance level of 7% across all STO conditions, with scores reaching approximately 50%. The highest score was when we trained and evaluated on synchronized data, which was an ideal condition. These findings suggest that synchronization of training and test data was one of the effective ways to classify robust acoustic scenes.

We also examined the F1-scores for each label under the matched condition. Table I shows the F1-scores for each label. The results indicate that although the overall F1-scores tend to decrease as the standard deviation of the STO increases, the labels “Watching TV” and “Cooking” exhibited relatively less performance degradation than other labels, maintaining high F1-scores even when the STO standard deviation was 320 samples. In contrast, the F1-scores for the labels “Absence,”

"Calling," and "Visit" degraded significantly.

We show the confusion matrices for detailed classification results to investigate the difference in classification accuracy across labels. Figure 6 shows confusion matrices at testing under different standard deviations of STO, using a model trained on synchronized data. Each confusion matrix corresponds to a different STO standard deviation at testing: 0, 16, 32, 160, and 320 samples. As shown in Fig. 6, misclassification as "Absence" became more frequent as the STO standard deviation increased. In particular, the labels "Watching TV" and "Working" were often misclassified as "Absence" or "Other." On the other hand, the "Absence" label was less misclassified into other labels.

Figure 7 shows confusion matrices under the matched condition. Each matrix corresponds to a case where the standard deviation of the STO at both training and testing is 0, 16, 32, 160, or 320 samples. We observed that labels such as "Watching TV," "Working," and "Cooking," which were more frequently misclassified when trained on synchronized data, are correctly classified under this condition. On the other hand, the label "Absence" was more often misclassified as other labels.

From Fig. 6, we observed that when the model was trained on synchronized data, the classification performance for labels such as "Watching TV" and "Cooking," which likely involve fixed sound source locations, degraded in the presence of STO. This degradation was considered to be caused by changes not only in the peak positions but also in the shape of the GCC-PHAT feature, as observed in Fig. 3. In contrast, Fig. 7 showed that these labels held high classification performance. This suggests that the proposed data augmentation method could simulate asynchronous conditions during training, enabling the model to classify asynchronous test data more robustly.

IV. Conclusions

In this study, we investigated whether the performance of acoustic scene classification under asynchronous conditions can be improved through a training strategy, assuming that only synchronized data is available. To this end, we proposed a data augmentation method that simulates asynchronous conditions by applying circular time shifts based on STOs to synchronized audio clips. Experimental results demonstrated that the proposed approach effectively mitigates performance degradation caused by STOs in test data. As future work, we plan to extend our approach to also address differences in microphone placement between training and testing [32], [33].

Acknowledgment

This work was supported by JST SICORP Grant Number JPMJSC2306 and JSPS KAKENHI Grant Number JP24KJ1866.

References

[1] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 1218–1221. DOI: 10.1109/ICME.2009.5202720.

[2] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen, and L. P. D. Witte, "Literature review on monitoring technologies and their outcomes in independently living elderly people," *Disability and Rehabilitation: Assistive Technology*, vol. 10, pp. 271–294, 4 2015. DOI: 10.3109/17483107.2014.961179.

[3] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 165–168. DOI: 10.1109/ICASSP.2009.4959546.

[4] S. Chandrakala and S. L. Jayalakshmi, "Environmental audio scene and sound event recognition for autonomous surveillance," *ACM Computing Surveys (CSUR)*, vol. 52, pp. 1–34, 3 2020. DOI: 10.1145/3322240.

[5] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, 2024. DOI: 10.1016/j.eswa.2023.121902.

[6] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, N. Greco, and R. Tachibana, "Domestic activities classification based on CNN using shuffling and mixing data augmentation," Detection, Classification of Acoustic Scenes, and Events Challenge (DCASE), Tech. Rep., 2018.

[7] S. Amiriparian, M. Gerczuk, S. Ottl, *et al.*, "Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, p. 19, 2020. DOI: 10.1186/s13636-020-00186-0.

[8] Y. Kaneko, T. Yamada, and S. Makino, "Monitoring of domestic activities using multiple beamformers and attention mechanism," *Journal of Signal Processing*, vol. 25, pp. 239–243, 6 2021. DOI: 10.2299/jsp.25.239.

[9] G. Dekkers, S. Lauwereins, B. Thoen, *et al.*, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2017, pp. 32–36.

[10] Z. Lin, Y. Li, Z. Huang, *et al.*, "Domestic activities clustering from audio recordings using convolutional capsule autoencoder network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 835–839. DOI: 10.1109/ICASSP39728.2021.9414643.

[11] T. Kawamura, Y. Kinoshita, N. Ono, and R. Scheibler, "Acoustic scene classification using inter- and intra-subarray spatial features in distributed microphone array," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, p. 65, 1 2024. DOI: 10.1186/s13636-024-00386-y.

[12] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and*

Mobile Computing, vol. 2017, pp. 1–24, 2017. DOI: 10.1155/2017/3956282.

[13] 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*, vol. 25, pp. 1335–1343, 6 2017. DOI: 10.1109/TASLP.2017.2690559.

[14] K. Imoto, "Graph cepstrum: Spatial feature extracted from partially connected microphones," *IEICE Transactions on Information and Systems*, vol. E103.D, pp. 631–638, 3 2020. DOI: 10.1587/transinf.2019EDP7162.

[15] Y. Shiroma, K. Imoto, S. Shiota, N. Ono, and H. Kiya, "Investigation on spatial and frequency-based features for asynchronous acoustic scene analysis," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1161–1166.

[16] T. Kawamura, Y. Kinoshita, N. Ono, and R. Scheibler, "Effectiveness of inter- and intra-subarray spatial features for acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096935.

[17] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015. DOI: 10.1016/j.sigpro.2014.09.015.

[18] S. Wozniak and K. Kowalczyk, "Passive joint localization and synchronization of distributed microphone arrays," *IEEE Signal Processing Letters*, vol. 26, pp. 292–296, 2 2019. DOI: 10.1109/LSP.2018.2889438.

[19] A. Chinaev, P. Thuene, and G. Enzner, "Double-cross-correlation processing for blind sampling-rate and time-offset estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1881–1896, 2021. DOI: 10.1109/TASLP.2021.3071967.

[20] Y. Masuyama, K. Yamaoka, T. Kawamura, and N. Ono, "Efficient joint optimization of sampling rate offsets using entire multichannel signal," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1816–1828, 2024. DOI: 10.1109/TASLP.2024.3369532.

[21] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 4 1976. DOI: 10.1109/TASSP.1976.1162830.

[22] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "On the effectiveness of spatial and multi-channel features for multi-channel polyphonic sound event detection," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020, pp. 115–119.

[23] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019, pp. 30–34.

[24] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 71–75. DOI: 10.1109/ICASSP40776.2020.9053045.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[26] D. S. Park, W. Chan, Y. Zhang, *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680.

[27] X. Jiang, C. Han, Y. A. Li, and N. Mesgarani, "Exploring self-supervised contrastive learning of spatial sound event representation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1281–1285. DOI: 10.1109/ICASSP48485.2024.10447391.

[28] T. Takahashi, Y. Kinoshita, N. Ueno, *et al.*, "Augmentation of various speed data by controlling frame overlap for acoustic traffic monitoring," in *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 2087–2091. DOI: 10.1109/APSIPAASC58517.2023.10317558.

[29] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 task 4," Detection, Classification of Acoustic Scenes, and Events Challenge (DCASE), Tech. Rep., 2019.

[30] S. Xia, D. Fourer, L. Audin-Garcia, J.-L. Rouas, and T. Shochi, "Speech emotion recognition using time-frequency random circular shift and deep neural networks," in *Proc. Speech Prosody*, 2022, pp. 585–589. DOI: 10.21437/SpeechProsody.2022-119.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[32] M. Tammen, T. Ochiai, M. Delcroix, T. Nakatani, S. Araki, and S. Doclo, "Array geometry-robust attention-based neural beamformer for moving speakers," in *Proc. INTERSPEECH*, 2024, pp. 3345–3349. DOI: 10.21437/Interspeech.2024-2427.

[33] T. Kawamura, Y. Masuyama, and N. Ono, "Domain adaptation for multi-channel acoustic scene classification to different array positions," in *Proc. European Signal Processing Conference (EUSIPCO)*, in press, 2025.