Domain Adaptation for Multi-Channel Acoustic Scene Classification to Different Array Positions

Takao Kawamura, Yoshiki Masuyama, and Nobutaka Ono
Graduate School of Systems Design, Tokyo Metropolitan University, Tokyo, Japan
kawamura-takao@ed.tmu.ac.jp, yoshiki.masuyama@ieee.org, onono@tmu.ac.jp

***Abstract*—In this study, we propose an unsupervised domain adaptation method for multi-channel acoustic scene classification (ASC). Multi-channel ASC leverages spatial features and provides advantages to distinguish scenes with similar spectral features. Meanwhile, it is sensitive to the mismatch of the spatial features due to the difference in the microphone array position. One promising way to mitigate the mismatch is domain adaptation. In ASC, existing studies have mainly focused on unsupervised domain adaptation for single-channel scenarios and compensated for the domain mismatch in the spectral features. In this paper, we explore unsupervised domain adaptation methods for multi-channel ASC. A straightforward approach is to perform unsupervised domain adaptation only with the spatial features. This approach, however, results in limited performance because it is not easy to align the spatial features of the source and target domains per acoustic scene after changing the array position dynamically. To mitigate this issue, we adapt the spatial features along with frozen spectral features. The spectral features are relatively invariant to the microphone positions and are expected to help associate the spatial features across different domains. In the evaluation experiments, we confirmed that our method mitigates performance degradation.**

***Index Terms*—Acoustic scene classification, unsupervised domain adaptation, domain adversarial training, spatial feature**

I. Introduction

Monitoring domestic activities is essential to enhance inhabitants' safety and quality of life. For example, there are many applications, such as monitoring older people and infants [1], [2], surveillance systems [3], [4], and life-logging systems [5]. In monitoring domestic activities, acoustic scene classification (ASC) is an important technology and has been studied [6]. ASC is the technology that classifies an audio recording to a predefined class label (e.g., "Cooking" and "Watching TV"). Many studies have utilized spectral features, such as log-Mel spectrograms and Mel-frequency cepstral coefficients (MFCCs) [7]–[11].

In multi-channel cases, spatial information is available and helpful in ASC [12], [13]. For instance, considering the classification using only spectral features between a TV talk show and everyday conversation, it should be challenging because both scenes contain speech. On the other hand, utilizing spatial features may facilitate their distinction based on the positions of sound sources. Spatial features, such as time differences and power ratios between microphone recordings, have been employed in acoustic scene analysis [14].

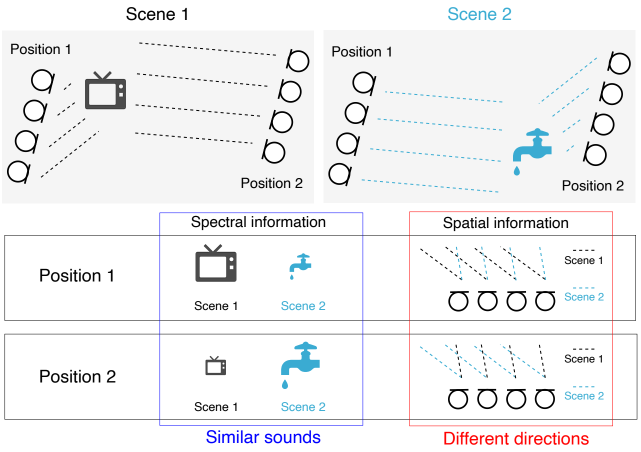

Fig. 1. Behavior of spectral and spatial features in recordings from different microphone array positions.

One of the spatial features is a generalized cross-correlation phase transform (GCC-PHAT) [15]. GCC-PHAT is widely used to estimate the time difference of arrival (TDOA). It can be computed regardless of the number of sound sources, source movement, microphone positions, or microphone array geometry. GCC-PHAT is utilized in ASC and sound event detection (SED) [12], [13], [16]–[19].

When monitoring domestic activities, the microphone array's position should be flexible to account for cases where users move the microphone array. However, if an ASC model is trained using a microphone array at a particular position, a microphone array located in a different position will represent an unknown environment for the trained ASC model. Figure 1 illustrates the behavior of features in different positions. Compared to spectral features, spatial information is expected to change significantly based on the orientation and position of the microphone array. This difference could lead to ASC performance degradation. From a practical perspective, collecting labeled data and retraining the model every time the microphone array's position changes is costly and unrealistic.

In this study, we explore domain adversarial training (DAT) as unsupervised domain adaptation for multi-channel ASC when microphone positions differ between training and evaluation. A straightforward approach is to perform DAT only with the spatial features. This approach, however, results in limited performance because it is not easy to align the spatial features for each acoustic scene when the array positions of the source and target domains differ. To mitigate this issue, we

This work was supported by JST SICORP Grant Number JPMJSC2306 and JSPS KAKENHI Grant Number 24KJ1866.

adapt the spatial features along with frozen spectral features. The spectral features are relatively invariant to the microphone positions and are expected to help associate the spatial features across different domains. In the evaluation experiment, we confirmed that the proposed method mitigates performance degradation.

II. Related work

Domain adaptation has been widely researched in the field of ASC [20]–[24]. For example, there are studies on ASC that adapt to different devices [20]–[22], adapt to different cities [20]–[22], and simultaneously adapt to both factors [24]. These studies apply domain adaptation techniques to spectral features. A widely used approach in domain adaptation is DAT, where a domain classifier is trained to distinguish source and target domains, and feature extractors are adversarially trained to extract domain-invariant features. This method improves robustness to differences between domains.

In multi-channel settings, spatial features can be effective. For instance, sound event localization and detection (SELD) and direction of arrival (DOA) estimation utilize spatial features. He *et al.* have applied domain adaptation for the real and imaginary parts of multi-channel spectrograms in DOA estimation [25]. Yasuda *et al.* have proposed an adaptation method for spectral and spatial features, using echo information as additional information for adaptation to unknown environments in SELD [26]. In these studies, domain adaptation techniques have been applied to spatial information.

III. Method

A. Task Specification

This study aims to conduct unsupervised domain adaptation for multi-channel ASC where a microphone array's position differs from the one used during training. We assume that the microphone array position changes within the same room. Here, we refer to the microphone array used during training as the source domain and the microphone array at different positions from the source domain as the target domain. We consider an ASC model that integrates spectral and spatial features (see Fig. 2). The spectral feature m_i and spatial feature g_i are processed by feature extractors $\mathcal{M}$ and $\mathcal{G}$, respectively. These extracted embeddings are then concatenated and input to classifier $\mathcal{C}$.

Spatial features g_i have been shown to be effective in ASC and SED [12], [16]–[19]. However, spatial characteristics might vary depending on the microphone array's position (see Fig. 1). For example, when the orientation or position of the microphone array changes, the spatial information, such as the direction of the sound source, also changes. Due to these feature differences, the feature extractor trained in the source domain may not work adequately in the target domain, resulting in performance degradation.

To address this issue, we apply DAT as unsupervised domain adaptation. In the source domain, we train the ASC model to enhance classification performance (see Sec. III-B). Then, we apply DAT to the feature extractors for the target domain (see Sec. III-C). Hereafter, the source domain is represented as $S = \{(x_i, y_i) | x_i \in X_S, y_i \in Y_S, i = 1, 2, ..., n_S\}$, and the target domain as $T = \{(x_j) | x_j \in X_T, j = 1, 2, ..., n_T\}$. Here, X_d and Y_d $(d \in \{S, T\})$ denote the sets of audio clips and scene labels, respectively.

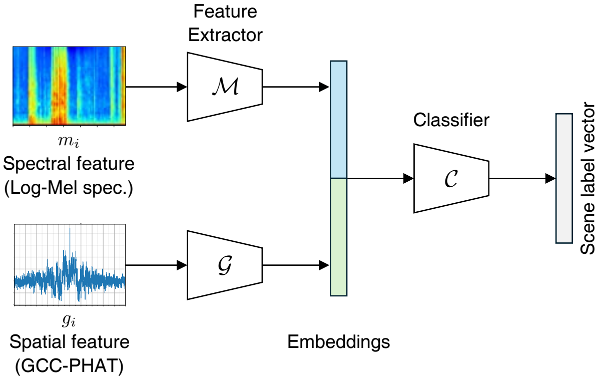

Fig. 2. Overview of the classification network. Feature extractors consist of either parameters trained in the source domain or updated through DAT.

B. Pre-training on Source Domain

We train the ASC model using a two-step approach [13]. In the first step, we train the feature extractors $\mathcal{M}$ and $\mathcal{G}$ for spectral and spatial features, respectively. Because these two types of features have different characteristics, we independently train the corresponding classification networks. This independent training is expected to enable each feature extractor to learn embeddings that are effective for ASC. We train the feature extractors and classifiers by minimizing the following cross-entropy losses:

$$\mathcal{L}_{\text{mel}} = -\frac{1}{n_S} \sum_{i=1}^{n_S} \sum_{k=1}^{K} \mathbb{1}_{[k=y_i]} \log(\mathcal{C}_{\text{mel}}(\mathcal{M}(m_i))), \quad (1)$$

$$\mathcal{L}_{\text{gcc}} = -\frac{1}{n_S} \sum_{i=1}^{n_S} \sum_{k=1}^{K} \mathbb{1}_{[k=y_i]} \log(\mathcal{C}_{\text{gcc}}(\mathcal{G}(g_i))), \quad (2)$$

where K is the number of classes, and $\mathbb{1}_{[k=y_i]}$ is the indicator function that equals 1 if $k = y_i$ and 0 otherwise.

In the second step, we integrate spectral and spatial features by concatenating the embeddings processed by the feature extractors (see Fig. 2). Then, the classifier $\mathcal{C}$ is trained to improve classification performance while freezing the trained feature extractors.

$$\mathcal{L} = -\frac{1}{n_S} \sum_{i=1}^{n_S} \sum_{k=1}^{K} \mathbb{1}_{[k=y_i]} \log(\mathcal{C}(\text{concat}(\mathcal{M}(m_i), \mathcal{G}(g_i)))), \quad (3)$$

where $\text{concat}(\cdot, \cdot)$ indicates the concatenation function. We train only $\mathcal{C}$ to minimize the above equation.

C. Unsupervised Domain Adaptation to Target Domain

We apply DAT for an array position that differs from the array position of the source domain. We train a domain classifier $\mathcal{H}$ that distinguishes the embeddings obtained from the source (S) and target domains (T). In contrast, the feature extractors

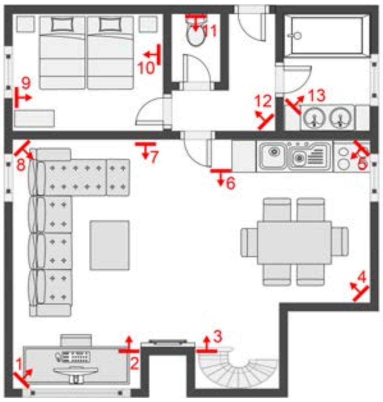

Fig. 3. SINS database [10]. Red arrows indicate 4-channel microphone arrays.

TABLE I
SPLIT SETTINGS FOR TRAINING AND EVALUATION DATA.

	type	domain	audio	label	hours
(a)	train	source	✓	✓	54.1h
(b)	train	target	✓	-	23.1h
(c)	valid	source	✓	✓	19.3h
(d)	test	target	✓	✓	27.1h

are trained adversarially to trick the domain classifier, resulting in the extraction of domain-invariant features. In this study, we input the embedding $f_i^{(d)}$ ($d \in \{S, T\}$) to the domain classifier $\mathcal{H}$. Here, $f_i^{(d)}$ is represented as follows:

$$f_i^{(d)} = \begin{cases} \mathcal{M}^{(d)}(m_i) & \text{(only spectral features)} \\ \mathcal{G}^{(d)}(g_i) & \text{(only spatial features)} \\ \text{concat}\left(\mathcal{M}^{(d)}(m_i), \mathcal{G}^{(d)}(g_i)\right) & \text{(both features)} \end{cases}. \tag{4}$$

In this study, the domain classifier $\mathcal{H}$ is trained using the loss function defined as follows:

$$\mathcal{L}_{\text{disc}} = -\frac{1}{n_S}\sum_{i=1}^{n_S} \log(\mathcal{H}(f_i^{(S)})) - \frac{1}{n_T}\sum_{i=1}^{n_T} \log(1 - \mathcal{H}(f_i^{(T)})). \tag{5}$$

Here, $f_i^{(S)}$ and $f_i^{(T)}$ indicate the embeddings of the source and target domains, respectively. We update the feature extractors $\mathcal{M}^{(T)}$ and/or $\mathcal{G}^{(T)}$ in DAT. The feature extractors are trained using the loss function defined as follows:

$$\mathcal{L}_{\text{feat}} = -\frac{1}{n_T}\sum_{i=1}^{n_T} \log(\mathcal{H}(f_i^{(T)})). \tag{6}$$

This loss function decreases when an embedding from the target domain is misclassified as belonging to the source domain. The domain classifier and feature extractors are updated alternatively.

In the target domain, the ASC model consists of feature extractors adapted using DAT and a classifier trained in the source domain.

IV. EVALUATION EXPERIMENTS

A. Setup

We evaluated the proposed method using the SINS database [10]. The SINS database is a continuous recording of one week of a person's activities collected by a network of 13 microphone arrays (see Fig. 3). Each microphone array consists of four omnidirectional microphones arranged in a linear shape. Here, microphone array 5 is not available due to random crashes/missing data. In this experiment, we trained the ASC model using recordings from subarray 1 located in a corner of the living room. Then, we evaluated domain adaptation on microphone arrays 4 and 8, located in the other corners of the room.

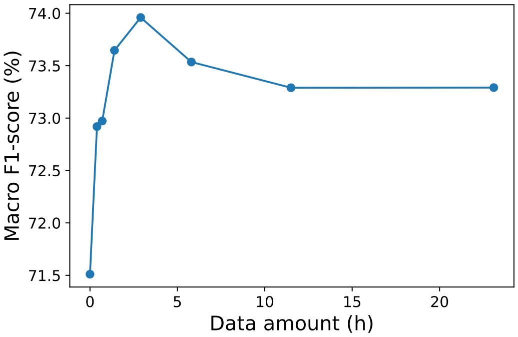

Fig. 4. Average F1-score with an increasing amount of target domain data.

TABLE II
MACRO F1-SCORE OF THE TARGET OF DOMAIN ADAPTATION. THE F1-SCORE IS EVALUATED ON A NETWORK (SEE FIG. 2). "ADAP." INDICATES THE TARGET FEATURE EXTRACTORS OF DOMAIN ADAPTATION. f_i IS THE INPUT OF THE DOMAIN CLASSIFIER. ⑤ IS THE RESULT WHERE DOMAIN ADAPTATION IS FIRST APPLIED TO ONLY SPECTRAL FEATURE, THEN USED TO ADAPT SPATIAL FEATURES.

	Train	Adap.	f_i	Average
①	source	-	-	71.5%
②	source	$\mathcal{G}$	$\mathcal{G}$	72.6%
③	source	$\mathcal{G}$	$\mathcal{M}$, $\mathcal{G}$	73.3%
④	source	$\mathcal{M}$, $\mathcal{G}$	$\mathcal{M}$, $\mathcal{G}$	NaN
⑤	source	$\mathcal{M}$	$\mathcal{M}$	76.0%
		$\mathcal{G}$	$\mathcal{M}$, $\mathcal{G}$	
⑥	target	-	-	82.5%

We divided the continuous audio recording into 10 s audio clips sampled at 16 kHz. From this dataset, we used audio clips including ten daily domestic activities: "Absence," "Calling," "Cooking," "Dishwashing," "Eating," "Other," "Vacuum cleaner," "Visit," "Watching TV," and "Working." We randomly split the dataset, as shown in Table I. In Table I, (a) and (c) consist of audio and labeled data, while (b) consists only of audio data.

We processed the audio clips to obtain spectral and spatial features at the same setting as [12]. The 40-dimension log-Mel spectrogram was calculated for each 64 ms time frame with a 20 ms overlap. The GCC-PHAT was calculated for each 128 ms time frame with a 50% overlap for each microphone pair. We used a log-Mel spectrogram calculated from the first channel of the subarray and three GCC-PHATs calculated with the reference channel fixed (microphone 1-2, 1-3, 1-4). The log-Mel spectrogram was a two-dimensional tensor of shape 40×501 (Freq $\times$ Frame). GCC-PHATs from three microphone pairs, each represented as a 2048-dimensional

TABLE III

AVERAGE F1-SCORE FOR EACH LABEL WITH ARRAY 4 AND 8. THE LEFT FOUR COLUMNS, MIDDLE TWO COLUMNS, AND RIGHT FOUR COLUMNS INDICATE CASES WHERE MEL+GCC WITHOUT DAT IS HIGHER BY MORE THAN 1PT, WITHIN 1PT, AND LOWER BY MORE THAN 1PT COMPARED TO MEL WITHOUT DAT, RESPECTIVELY. MEL+GCC WITH DAT CORRESPONDS TO ⑤ IN TABLE II.

Feature	Train.	Adap.	F1-score										
			Absence	Other	Working	Eating	Vacuum cleaner	Visit	Calling	Cooking	Dish washing	Watch. TV	Ave.
Mel	source	-	77.8%	21.7%	37.3%	33.2%	99.2%	77.7%	88.8%	93.7%	63.6%	99.4%	69.2%
Mel+GCC	source	-	81.1%	25.7%	42.3%	56.5%	99.1%	78.5%	83.5%	88.2%	61.7%	98.3%	71.5%
Mel	source	✓	88.0%	21.0%	54.9%	84.5%	82.6%	77.3%	86.1%	95.3%	62.7%	99.4%	75.2%
Mel+GCC	source	✓	87.4%	15.4%	44.3%	86.9%	96.9%	79.7%	85.0%	95.7%	69.1%	99.4%	76.0%
Mel	target	-	80.5%	32.5%	64.2%	86.3%	98.7%	80.9%	89.8%	97.8%	84.9%	99.7%	81.5%
Mel+GCC	target	-	80.5%	30.2%	66.6%	87.8%	98.7%	84.0%	90.7%	98.3%	88.9%	99.7%	82.5%

vector, were concatenated into a single 6144-dimensional vector. The log-Mel spectrogram and GCC-PHATs were fed into a convolutional neural network $\mathcal{M}$ based on [7] and a fully connected network $\mathcal{G}$, respectively, each extracting a 256-dimensional embedding. The classifiers $\mathcal{C}_{\mathrm{mel}}$ and $\mathcal{C}_{\mathrm{gcc}}$ each received a 256-dimensional embedding as input, while $\mathcal{C}$ received a 512-dimensional embedding.

For pre-training in the source domain, we used datasets (a) and (c). The networks were trained for 50 epochs using the AdamW optimizer [27], with the learning rate of 1.0×10^{-4} and weight decay of 1.0×10^{-5}. To mitigate the data imbalance problem, we sampled audio clips for each scene label equally.

For DAT, we used datasets (a) and (b). The networks were trained for 200 epochs using the AdamW optimizer. For adaptation using only spectral features, the learning rates of the feature extractor and domain classifier were 1.0×10^{-7} and 1.0×10^{-6}, respectively. For other conditions, the learning rates of the feature extractor and domain classifier were both set to 1.0×10^{-5}. The weight decay was set to 1.0×10^{-5}. As the evaluation metric, we averaged the F1-score calculated for each scene label separately.

B. Results

1) Comparison of the Amount of Target Domain Data: First, we investigated the relationship between the amount of target domain data and the macro F1-scores. Figure 4 shows the macro F1-score for changes in the amount of the target domain data. The F1-score without DAT (0 h) was 71.5%. In contrast, we confirmed an improvement in the F1-score through DAT. Especially with 2.9 hours of the target domain data, the F1-score improvement was 2.5%. However, no further improvement in the F1-score was observed with additional target domain data. One possible reason why no additional improvement was observed is that loss functions (5) and (6) were designed for domain classification rather than scene classification. In addition, when the target domain data was increased beyond 2.9 hours, the proportion of several classes that accounted for less than 5 percent of the data decreased incidentally. As a result, the F1-scores for these classes declined, and the overall performance did not improve.

2) Comparison of Domain Adaptation Targets: We compared the feature extractors to which domain adaptation was applied ("Adap.") and the input to the domain classifier (f_i). We evaluated the ASC model that utilizes spectral and spatial features (see Fig. 2). Table II shows the average F1-scores for the feature extractors used in domain adaptation and the input f_i of the domain classifier. Here, ① indicates the result without domain adaptation, where the lowest performance is expected. In contrast, ⑥ indicates the result of training on the target domain data, where the highest performance is expected. First, we compared the DAT results for $\mathcal{G}$ using only spatial features (②) and incorporating spectral features (③). The experimental results show that ③ achieves a 0.7% higher F1-score than ②, suggesting that including spectral features in f_i is beneficial for domain adaptation of $\mathcal{G}$.

Next, we investigated domain adaptation, which includes the spectral feature extractor. We compared the results by adapting both the spectral and spatial feature extractors simultaneously (④), and first adapting the spectral feature, followed by using it to adapt the spatial features (⑤). The experimental results indicated that ⑤ achieved the highest F1-score. In contrast, for ④, the F1-score could not be calculated because the ASC model could not output some labels. This suggests that freezing the spectral feature extractor while adapting the spatial feature extractor was effective. It was considered that spectral information was relatively stable because microphone recordings contained similar sounds. In the following section, we used the results of ⑤ as the Mel+GCC results obtained after DAT.

3) Overall Performance of Domain Adaptation: We compared the classification performance of models trained i) without DAT, ii) with DAT, and iii) on the target domain data. We also compared the results of Mel and Mel+GCC.

Table III shows the overall F1-scores for each label. The experimental results confirmed that the average F1-score of Mel+GCC was higher than that of Mel when domain adaptation was not applied. Examining the F1-scores for each label reveals that Mel+GCC outperformed Mel for labels such as "Absence," "Other," "Working," and "Eating" (left four columns). The labels "Other" and "Working" have been reported to contain many silent intervals, as described in [10]. On the other hand, the F1-scores of Mel+GCC for "Calling," "Cooking," "Dishwashing" and "Watching TV" were lower than those of Mel (right four columns). These labels were considered to be associated with spatial information that changes due to different array positions, leading to performance degradation. To apply DAT, we confirmed that the average F1-score with DAT was higher than that without it. These results indicate that the proposed method mitigates domain mismatch.

We also confirmed that in methods ②, ③, and ④ in Table II, the F1-score of Mel+GCC with DAT was lower than that of Mel with DAT alone, which indicates the difficulty of adapting spatial features. However, our proposed method (⑤), which first adapts the spectral feature and then uses it to guide the adaptation of spatial features, successfully addressed this challenge and achieved F1-score 0.8pt higher than that of Mel with DAT.

V. Conclusion

In this study, we explore unsupervised domain adaptation for multi-channel ASC. We address the problem where the microphone array position differs between the training and evaluation. We apply DAT to the feature extractors for spectral and spatial features. In the evaluation experiments, we demonstrated the effectiveness of DAT when using a different microphone array for training and evaluation in the same room. We also compared the target feature extractors for DAT. We confirmed the highest F1-score when first adapting the spectral feature and then using it to adapt the spatial features. For future work, we plan to incorporate powerful pre-trained models [28] instead of using the log-Mel spectrogram.

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.

[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, 2015.

[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.

[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, 2020.

[5] M. A. M. Shaikh, M. K. I. Molla, and K. Hirose, "Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense," in *Proc. International Conference on Computer and Information Technology (ICCIT)*, 2008, pp. 294–299.

[6] B. Ding, T. Zhang, C. Wang, G. Liu, J. Liang, R. Hu, Y. Wu, and D. Guo, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121902, 2024.

[7] 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 and Classification of Acoustic Scenes and Events Challenge (DCASE), Tech. Rep., 2018.

[8] S. Amiriparian, M. Gerczuk, S. Ottl, L. Stappen, A. Baird, L. Koebe, and B. Schuller, "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.

[9] Y. Kaneko, T. Yamada, and S. Makino, "Monitoring of domestic activities using multiple beamformers and attention mechanism," *Journal of Signal Processing*, vol. 25, no. 6, pp. 239–243, 2021.

[10] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. V. den Bergh, 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 *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2017, pp. 32–36.

[11] Z. Lin, Y. Li, Z. Huang, W. Zhang, Y. Tan, Y. Chen, and Q. He, "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.

[12] 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.

[13] ——, "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, 2024.

[14] P. Giannoulis, G. Potamianos, and P. Maragos, "Room-localized speech activity detection in multi-microphone smart homes," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, p. 15, 2019.

[15] 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, 1976.

[16] 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.

[17] 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.

[18] S. Adavanne, P. Pertila, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 771–775.

[19] 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.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2018, pp. 9–13.

[21] K. Drossos, P. Magron, and T. Virtanen, "Unsupervised adversarial domain adaptation based on the Wasserstein distance for acoustic scene classification," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 259–263.

[22] W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-CRNN: A domain adaptation model for sound event detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 276–280.

[23] S. Singh, H. L. Bear, and E. Benetos, "Prototypical networks for domain adaptation in acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 346–350.

[24] Y. Tan, H. Ai, S. Li, and M. D. Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1286–1297, 2024.

[25] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 770–774.

[26] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 226–230.

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

[28] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. INTERSPEECH*, 2022, pp. 2753–2757.