International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.3, May 2025

Subspace-based Representations for Acoustic Scene Classification

Akansha Tyagi
School of Computing and Electrical Engineering (SCEE)
Indian Institute of Technology Mandi, Himachal Pradesh, India

Padmanabhan Rajan
School of Computing and Electrical Engineering (SCEE)
Indian Institute of Technology Mandi, Himachal Pradesh, India

ABSTRACT

Real-world acoustic scene data has a complex structure that leads to high levels of overlap within an acoustic scene class. This overlap stems from various similar factors, such as different recording devices and recording locations or cities, which act as confounding factors. On the other hand, the same set of confounding factors would be present across different acoustic scene classes and can be considered as a common link across them. Utilizing this common structure, it is possible to perform multi-block analysis to learn the representation of these common links. Two formulations are proposed for the multi-block analysis of acoustic scene data, employing a common orthogonal basis extraction algorithm. The proposed formulations enhance the performance of the acoustic scene classification system by reducing the information pertaining to the recording devices and cities from the learnt acoustic scene representations. Experiments were conducted on five standard Detection and Classification of Acoustic Scenes and Events (DCASE) datasets. Across all datasets, the classification performance achieved using features derived from the multi-block formulations surpassed that of features not incorporating these formulations.

Keywords

Acoustic Scene Classification, Multi-block Analysis, Subspace-based representations, Intra-scene variation, Recording device , Recording city, Detection and Classification of Acoustic Scenes and Events

1. INTRODUCTION

Acoustic scene classification (ASC), which is a component of acoustic scene analysis, has been a well-researched problem for a number of years now [28]. The aim of an ASC system is to identify the environment from where an audio recording has been obtained, based on the sounds present in it [2]. Some of the challenges faced by ASC systems include high intra-class variation and the presence of similar sounds across different acoustic scenes. There can be several factors that cause variation in audio data that represents acoustic scenes. These can include the inherent variability of complex environments. For example, the interior of a moving bus might sound different from the interior of the bus at a bus stop. Also, differences could arise due to different recording devices and recording locations. Each device imparts its own specific characteristics to the recorded audio signal and thus can be considered as a direct source of variation. Additionally, the recording location referring to the city where the acoustic scene is captured, can be considered as an indirect source of variation. For example, the buses commonly used in London might be different (and sound different) from the buses used in Barcelona. Previous work [26] supports the consideration of city information as a variation causing-factor for the problem of ASC.

This work addresses the variations introduced by differing recording conditions, including various recording devices and cities, by using a multi-block subspace-based approach. Through subspace-based approaches, acoustic scene data is transformed to reduce unwanted variations, resulting in more discriminative acoustic scene representations. These methods' main objective is to compute the basis vectors of the subspace, which offer a concise and insightful representation of the data. Subspace analysis techniques include multi-block methods, which are designed to compute data representations by learning an optimal subspace by exploiting the natural linkage across multiple blocks of the data. Face image datasets are a type of multi-block data, where the facial images that share a common pose or illumination across different subjects naturally creates an association among them [33]. This common information is not useful for the task of face recognition.

For ASC, the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018, 2019 and 2020 datasets include device and city labels for their acoustic scene data recordings, enabling the use of multi-block formulation. These labels act as separate sources of linkage across different data blocks representing different acoustic scene classes. In this work, domain is described as the common linking factor present in all the blocks, and common and individual feature extraction (CIFE) multi-block analysis method [33] is applied to remove this domain information. For acoustic scene classification, device and city information act as domain information. Similarly, for device and city classification tasks, the acoustic scene information acts as domain information. This work aims to learn representations where the domain information is suppressed, and discriminative information is enhanced.

The key contributions of this paper are summarized as follows:

(1) This work presents a multi-block formulation of acoustic scene data where the data corresponding to different acoustic scene classes constitute different blocks and the recording conditions (city and device) of these acoustic scene classes provide a common source of association between these different blocks.

(2) Two methods for constructing a multi-block matrix are presented. In the first formulation, acoustic scene data from all devices or cities are considered together to form a multi-block matrix. In the second formulation, acoustic scene data from different devices or cities are considered separately to form multiple multi-block matrices (one corresponding to each device or city).

(3) The multi-block formulation results in the construction of discriminative acoustic scene representations which in turn improves the accuracy of an ASC system. The multi-block formulation is also applied to the auxiliary tasks of device classification and city classification [3] using acoustic scene data.

The remainder of this manuscript is organized as follows: In section 2 a few subspace-based methods for ASC are briefly reviewed. The multi-block formulation for ASC data along with the comprehensive overview of the common and individual feature extraction framework and Common Orthogonal Basis Extraction algorithm are described in section 3. Section 4 presents the experimental details. Finally, the conclusion is presented in section 5 of the paper.

2. RELATED WORK

Several subspace-based methods have been proposed in the acoustic scene literature. Notably, the authors in [5] demonstrate the effectiveness of unsupervised features derived from matrix factorization techniques like Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) [30] for acoustic scene classification. The same work also investigates the performance of the sparse [8], kernel [31] and convolutive [19] variants of these methods. Authors in [5] also, focus on learning dictionary elements from spectrograms of training data and use the features obtained by projecting data onto the learned dictionary for classifying different acoustic scene classes. Another work of the same authors [6] is built upon a supervised matrix factorization method called task-driven dictionary learning (TDL) [16], where a non-negative formulation of TDL is proposed for ASC.

Another form of subspace-based methods are the multi-block analysis methods which include canonical correlation analysis (CCA) [12]. CCA maximizes the correlation between random variables. For classification tasks, these methods have been widely used in computer vision [24], natural language processing [22] and speech processing [1]. In this paper, an attempt is made to formulate the classification of various acoustic scene classes as a multi-block problem by using the CIFE framework [33]. A related method is nuisance attribute projection or NAP which attempts to remove unwanted (or nuisance) components through projections [23]. NAP has been used earlier [27] for ASC, though in a different context.

Recently, some works like [25] in ASC have addressed the presence of differing recording conditions such as recording cities and devices in the acoustic scene data. Typically, data augmentation methods such as random temporal cropping [17] [13], mixup [32], spectrum correction [14], and specaugment [20] are effective in countering variations due to recording devices. To address variations due to difference in recording cities authors in [3] applied a multi-task learning framework, while [4] used a convolutional neural network (CNN)-based classifier. The task of city classification based on audio has evolved from acoustic scene classification [4, 15], but has remain relatively underexplored.

3. MULTI-BLOCK FORMULATION

Real-world data often occurs in the form of multiple block matrices either implicitly or explicitly. For instance, a single matrix $\mathbb{B} = [\mathbf{B}_1\mathbf{B}_2\cdots\mathbf{B}_C]$ can be constructed by concatenating multiple linked block matrices $\mathbf{B}_c \in \mathbb{R}^{d\times N_{B_c}}$ where $c \in [1\cdots C]$ denotes the block index. Here each column of $\mathbf{B}_c$ represents a d-dimensional example and there are N_{B_c} such examples. The existence of a common link between the block matrices allows the application of multi-block analysis methods. The common and individual feature extraction (CIFE) [33] framework exploits such structure. As a part of this framework, Common Orthogonal Basis Extraction (COBEC) algorithm is used to compute the representation of this common linkage, termed as *common basis*. The common basis represented by $\mathbf{D} \in \mathbb{R}^{d\times k}$ (k denotes number of common components) is obtained by applying COBEC on $\mathbb{B}$ as described in Algorithm 1. Basically, the optimal $\mathbf{D}$ is estimated by performing

Algorithm 1 COBEC Algorithm

Input : $\mathbb{B} = [\mathbf{B}_1\mathbf{B}_2\cdots\mathbf{B}_C]$ (multi-block matrix), k (number of common components)
Output: $\mathbf{D}$ (common basis)
(1) QR decomposition on $\mathbf{B}_c = \mathbf{Q}_c\mathbf{U}_c$ where $\mathbf{Q}_c \in \mathbb{R}^{d\times d}$, $\mathbf{U}_c \in \mathbb{R}^{d\times N_{B_c}}$ and $c \in [1\cdots C]$.
(2) Randomly initialize $\mathbf{Z}_c \in \mathbb{R}^{d\times k}$
(3) **while** until convergence **do**
(4) $\quad\mathbf{P} = \sum_c \mathbf{Q}_c\mathbf{Z}_c$
(5) $\quad[\mathbf{E}, \mathbf{\Lambda}, \mathbf{V}] = \text{tSVD}(\mathbf{P}, k)$ `// truncated SVD`
(6) $\quad\mathbf{D} = \mathbf{E}\mathbf{V}^T$
(7) $\quad\mathbf{Z}_c \leftarrow \mathbf{Q}_c^T\mathbf{D}$
(8) **end while**
(9) **return** $\mathbf{D}$

QR decomposition on each block matrix $\mathbf{B}_c$, followed by truncated singular value decomposition (tSVD) and iterative updates with respect to a randomly initialized matrix $\mathbf{Z}_c$. The readers are referred to [33] for more details about the algorithm.

The obtained common basis is used to segregate each block matrix $\mathbf{B}_c$ into its common and specific matrix components as:

$$\mathbf{B}_c = \mathbf{C}_c + \mathbf{S}_c \tag{1}$$

where $\mathbf{C}_c \in \mathbb{R}^{d\times N_{B_c}}$ contains the common features and $\mathbf{S}_c \in \mathbb{R}^{d\times N_{B_c}}$ contains the specific features. The splitting of each block matrix $\mathbf{B}_c$ is performed as per the following steps:

(1) The common matrix $\mathbf{C}_c$ which contains the common features corresponding to block matrix $\mathbf{B}_c$ is determined by using $\mathbf{D}$ as an orthogonal projection matrix:

$$\mathbf{C}_c = \mathbf{D}(\mathbf{B}_c^\intercal\mathbf{D})^\intercal \tag{2}$$

(2) The specific matrix $\mathbf{S}_c$ containing the specific features is obtained by subtracting the common matrix $\mathbf{C}_c$ from the block matrix $\mathbf{B}_c$ which contains the original features:

$$\mathbf{S}_c = \mathbf{B}_c - \mathbf{C}_c \tag{3}$$

Compared to the common matrix $\mathbf{C}_c$, the specific matrix $\mathbf{S}_c$ contains the discriminating information and is therefore better suited for the task of classification.

Two variants of the multi-block formulation *all-domains-multi-block formulation* and *domain-wise-multi-block formulation* are now presented, in which the multiple block matrices are connected through a common source of linkage or belong to the same domain.

Dataset	Cities	Devices
DCASE 2018 Subtask A	Barcelona, Helsinki, London, Paris, Stockholm, Vienna	Device A (Soundman OKM II Klassik/studio A3)
DCASE 2018 Subtask B	Barcelona, Helsinki, London, Paris, Stockholm, Vienna	Device A, Device B (Samsung Galaxy S7), and Device C (iPhone SE)
DCASE 2019 Subtask A	Barcelona, Helsinki, London, Paris, Stockholm, Vienna, Lisbon, Lyon, Milan, Prague	Device A
DCASE 2019 Subtask B	Barcelona, Helsinki, London, Paris, Stockholm, Vienna, Lisbon, Lyon, Milan, Prague	Devices A, B and C
DCASE 2020	Barcelona, Helsinki, London, Paris, Stockholm, Vienna, Lisbon, Lyon, Milan, Prague	Devices A, B, C and 6 simulated devices namely : S1, S2, S3, S4, S5, S6

Table 1. : Overview of Detection and Classification of Acoustic Scenes and Events (DCASE) datasets

3.1 All-domains-multi-block formulation

Different examples from all domains can be rearranged in the form of multiple blocks as presented in figure 1. A multi-block matrix $\mathbb{B}$ is constructed by concatenating C block matrices as follows:

$$\mathbb{B} = [\mathbf{B}_1 \ \mathbf{B}_2 \ \cdots \ \mathbf{B}_C] \tag{4}$$

where each block matrix $\mathbf{B}_c$ corresponds to the c^{th} class and C is the total number of classes. In this formulation, the columns of each block matrix contain the data domain-wise as:

$$\mathbf{B}_c = [\mathbf{B}_c^1 \ \mathbf{B}_c^2 \ \cdots \ \mathbf{B}_c^L] \tag{5}$$

where $\mathbf{B}_c^l \in \mathbb{R}^{d \times N_{B_c^l}}$ corresponds to the data of the c^{th} class for the l^{th} domain, $l \in [1 \cdots L]$ where L is the total number of domains.
It is assumed that the common basis $\mathbf{D}$, learnt from the multi-block matrix $\mathbb{B}$ contains the domain information (pertaining to all domains) present across the data of multiple classes. Thus, $\mathbf{D}$ represents the domain information that is not helpful for the classification task.
Using $\mathbf{D}$, the common feature matrix $\mathbf{C}_c$ is computed as described in equation 2. These common features are then subtracted from each block matrix $\mathbf{B}_c$ to obtain a specific matrix $\mathbf{S}_c$ as described in equation 3; thus the specific features are computed by subtracting the domain information from each block matrix. After this, the columns of $\mathbf{S}_c$ represent the examples with reduced domain information; which presumably reduces the difficulty in the classification task.

3.2 Domain-wise-multi-block formulation

A natural extension of the above formulation would be to construct multi-block matrices for each domain separately by making use of the domain labels; it can be considered as a more fine-grained version of the above formulation. Figure 2 shows the multi-block construction for this formulation in which a set of common basis is determined for each domain separately. Thus, the domain-wise-multi-block formulation is the all-domains-multi-block formulation repeated L times, with each multi-block matrix containing the data of one domain only.
A multi-block matrix $\mathbb{L}^l$ is constructed for each domain as follows :

$$\mathbb{L}^l = [\mathbf{B}_1^l \ \mathbf{B}_2^l \ \cdots \ \mathbf{B}_C^l] \tag{6}$$

where $\mathbb{L}^l \in \mathbb{R}^{d \times N_{\mathbb{L}^l}}$ and $N_{\mathbb{L}^l}$ is the number of examples across all the classes from the l^{th} domain.
Corresponding to every $\mathbb{L}^l$ multi-block matrix, a set of common basis $\mathbf{D}^l \in \mathbb{R}^{d \times k}$ is obtained which represents the l^{th} domain. The CIFE framework splits each block matrix $\mathbf{B}_c^l$ into its common and specific counterparts as:

$$\mathbf{B}_c^l = \mathbf{C}_c^l + \mathbf{S}_c^l \tag{7}$$

where $\mathbf{C}_c^l$ corresponds to the common features containing the domain information and $\mathbf{S}_c^l$ corresponds to the specific features which contain the class information. The specific features $\mathbf{S}_c^l$ are obtained as:

$$\mathbf{S}_c^l = \mathbf{B}_c^l - \mathbf{C}_c^l \tag{8}$$

where $\mathbf{C}_c^l$ is obtained via orthogonal projection similar to equation 2. Thus, $\mathbf{C}_c^l$ does not provide any useful information for classification, but $\mathbf{S}_c^l$ contains the discriminative information that is vital for the task.

4. EXPERIMENTAL EVALUATION

This section presents the experimental details and results for classification using the multi-block formulation. It begins with an overview of the datasets used and a description of the feature extraction process. Subsequently, results are outlined for three related tasks: acoustic scene classification, recording device classification, and recording city classification. The baseline and proposed systems are then described, followed by an analysis of the results and additional insights.

4.1 Dataset Description

The experiments were conducted using five datasets from the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges, spanning the years 2018 to 2020 and encompassing Subtasks A and B. These datasets contain audio recordings from ten distinct acoustic scenes: ‘airport (air)’, ‘metro’, ‘metro station (m-st)’, ‘bus’, ‘park’, ‘tram’, ‘street traffic (s-tr)’, ‘street pedestrian (s-ped)’, ‘public square (p-sq)’ and ‘shopping mall (s-mall)’. Over the years, the datasets have increased in complexity, incorporating a broader range of cities and recording devices.
Additionally, these acoustic scenes can be grouped into three broader categories: indoor, outdoor, and vehicle. The vehicle cat-

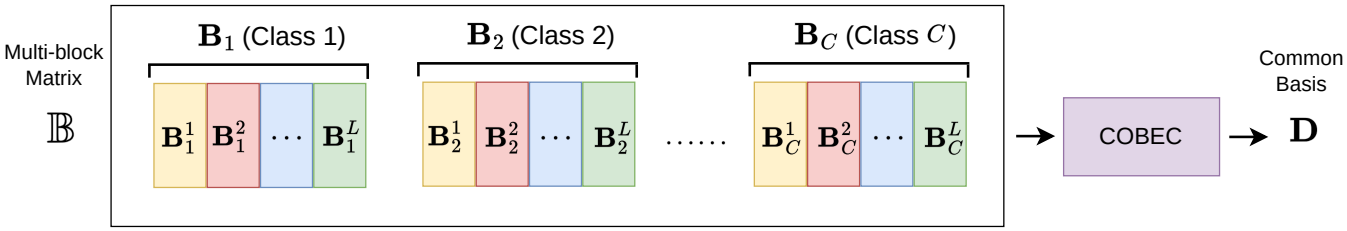

Fig. 1: All-domains-multi-block formulation; each block $\mathbf{B}_c$ contains data from one class and different colors in each block represents data from different domains. For example, in context of acoustic scene classification, different blocks correspond to different acoustic scenes and different colors in each block represent the common information pertaining to all cities, which we aim to suppress.

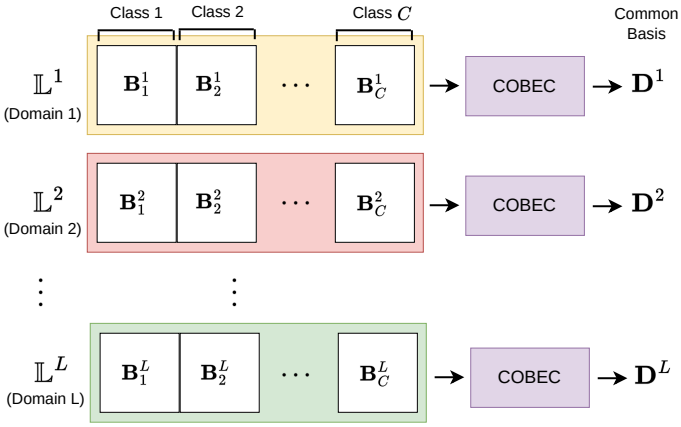

Fig. 2: Domain-wise-multi-block formulation; each multi-block matrix $\mathbb{L}^l$ distinguished by different colors represents data from a single domain. Within each multi-block matrix, separate block matrices $\mathbf{B}_c^l$ represent data for separate classes. For example, in context of acoustic scene classification, different multi-block matrices represent data from different cities and the common information is learned separately for each city.

egory includes bus, tram, and metro scenes; the indoor category consists of airport, metro station, and shopping mall; and the outdoor category includes park, public square, street pedestrian, and street traffic.

A detailed overview of the datasets, including the cities covered and devices used, is provided in table 1. DCASE 2018[1], and 2019[2] datasets primarily focus on recordings from a smaller set of devices, while DCASE 2020[3] introduces additional simulated devices, making it more complex dataset as compared to 2018 and 2019.

4.2 Feature Extraction

The experiments are performed using Patchout faSt Spectrogram Transformer (PaSST)[4] as a frozen feature extractor. PaSST has the transformer-based architecture and is an optimized version of the Audio Spectrogram Transformer (AST) [10], pre-trained on ImageNet [7] and fine-tuned on the Audioset [9] dataset. The network takes a raw audio signal as input and returns a feature vector of dimension 768. These feature representations are referred as *original features* in the subsequent sections.

As mentioned previously, that DCASE dataset has labels for acoustic scene, city, and device, allowing for three separate classification tasks: acoustic scene classification, device classification, and city classification. This structure also enables both all-domains and domain-wise multi-block approaches, where the domain can be chosen based on the task at hand. For example, in acoustic scene classification, the recording device or city can be considered as the domain.

Subsections *4.3.1* and *4.3.2* cover acoustic scene classification with device and city as domains, while subsections 4.4 and 4.5 describe device and city classification tasks by considering acoustic scene as the domain. The multi-block formulations capture and subsequently remove domain information to refine the discriminative features for the respective classification tasks. The features obtained by using multi-block formulations are referred to as *specific features* as per equations 3 and 8.

4.3 Acoustic Scene classification (ASC)

For ASC, the recording device and city information act as confounding factors and negatively impact the classification performance. To address this, the confounding factors are considered as domains and multi-block formulations are applied to learn the bases representing the domain information present across different acoustic scenes. In subsection *4.3.1*, the problem of ASC is framed with device as domain, while subsection *4.3.2* frames it using city as domain.

4.3.1 Device as domain. For this scenario, the different recording devices are considered as different domains and all-devices and device-wise multi-block formulations are used to learn information pertaining to recording devices, which is not useful for classification. In the former formulation, the multi-block matrix $\mathbb{B}$ is constructed by concatenating data from all devices, facilitating the learning of a common basis $\mathbf{D}$ that captures the information pertaining to all devices. In contrast, the latter formulation, constructs separate multi-block matrices and learns basis $\mathbf{D}^l$ for each device separately.

Table 2 presents the ASC results for DCASE 2018 Subtask B, 2019 Subtask B, and DCASE 2020 datasets using both original and specific features. Original features don't use multi-block formulations, while specific-all-devices and specific device-wise features are obtained using all-domains and domain-wise multi-block formulations respectively. The number of domains (L) is set to 3 for 2018, 2019 and 6 for the 2020 dataset. The values for number of

[1]https://dcase.community/challenge2018/index
[2]https://dcase.community/challenge2019/index
[3]https://dcase.community/challenge2020/task-acoustic-scene-classification
[4]https://github.com/kkoutini/PaSST

common components (k) are chosen through validation data and are mentioned with the classification results in table 2.

4.3.2 City as domain. Similar to devices, recording cities also introduce variations in the acoustic scene data. City information is extracted using both multi-block formulations. This information is then removed from the original features to construct specific-all-cities features using the all-domains method and specific-city-wise features using the domain-wise formulation. Table 2 presents the results obtained using original, specific-all-cities and specific-city-wise features for DCASE 2018 Subtask A, 2019 Subtask A and 2020 datasets. When city is considered as the domain, the optimal value of k is determined using validation data, with L set to 6 for the 2018 dataset and 10 for the 2019 and 2020 datasets.

4.4 Device Classification

Acoustic scenes captured with different recording devices contain device-specific characteristics. For device classification, the acoustic scene information can be considered as a source of variation. Multi-block formulations can be applied to minimize scene-specific similarities by viewing acoustic scenes as domains and considering data from different devices as different blocks.
The results for device classification using original, specific-all-scenes, and specific-scene-wise features are shown in Table 2, for the DCASE 2018 Subtask B, 2019 Subtask B, and 2020 datasets. The values $k = 10$ and $L = 10$ were chosen for all datasets, with L corresponding to the number of domains, i.e., acoustic scenes.

4.5 City Classification

City classification is relatively new and a related task to ASC. Compared to ASC this task is a harder problem as there are greater variations within the city data due to scene-specific similarities [11]. To reduce these similarities, multi-block formulations can be used by considering different acoustic scenes as different domains and data corresponding to different cities as different blocks.
Removing acoustic scene information will result in features better suited for city classification. The classification results using original, specific-all-scenes, and specific-scene-wise features are shown in Table 2, for the DCASE 2018 Subtask A, 2019 Subtask A, and 2020 datasets. L is fixed to 10, and k is arbitrarily set to 10, as city classification is considered an auxiliary task.

4.6 Systems Description

For all three tasks, two systems are proposed: the first uses features derived from the all-domains multi-block framework, and the second uses features from domain-wise framework. The baseline systems for the respective tasks use original features and a deep neural network (DNN). The DNN architecture consists of two hidden layers with 128 and 64 neurons each and ReLU activation function, and a classification layer with neurons equal to the number of classes for each task. To ensure a fair comparison, the architecture, training hyperparameters, and validation data remain consistent across both the baseline and proposed systems. The DNN is trained using standard procedures, including categorical cross-entropy loss and the ADAM optimizer.

4.7 Results and Discussion

The all-domains-multi-block formulation processes the training data of all classes from all domains together. On the other hand, the domain-wise-multi-block formulation processes the training data of all classes for each domain separately. Each column in $\mathbf{S}_c$ as per

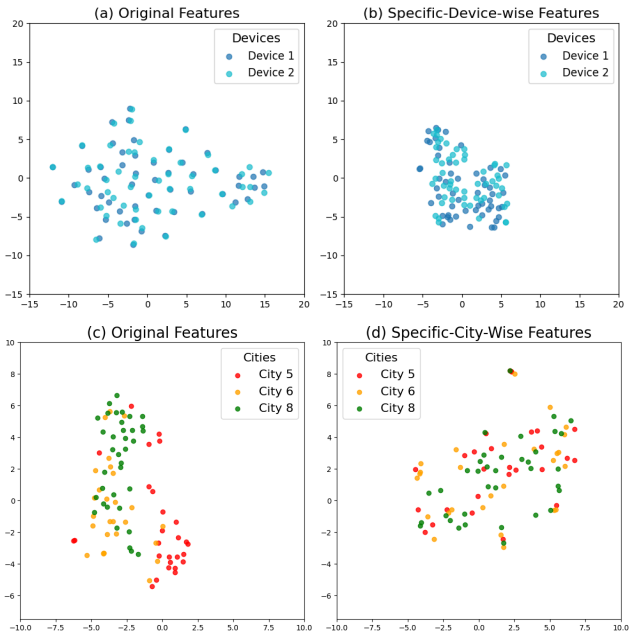

Fig. 3: t-SNE visualization of original features vs. specific-domain-wise features. Figures (a) and (b) represent embeddings from the DCASE 2019 Subtask B test dataset, showcasing data from the scene 'airport', and two devices distinguished by different colors. Figures (c) and (d) represent embeddings from the DCASE 2020 test dataset, presenting data from the scene 'airport' and three cities distinguished by different colors.

equation 3 represents a class example after reducing the information pertaining to all domains whereas each column in $\mathbf{S}_c^l$ as per equation 8 represents a class example with the information of a single domain reduced. The multi-block formulation in this paper utilized the domain information as the common link between the various blocks. The following main inferences can be drawn from the results obtained as per table 2:

(1) The use of specific features led to improved classification performance for the primary task of acoustic scene classification, as well as for the two secondary tasks of device and city classification. In the DCASE 2020 dataset, however, city classification accuracy decreases slightly may be due to the increased variability introduced by a wider range of devices, simulated as well as real. Device classification could not be performed on this dataset because the test data includes unseen devices.

(2) For ASC, using all-domains features resulted in a relative improvement of 3% on average as compared to the original features. This observation is also supported by the t-SNE plot for four acoustic scene classes of DCASE 2020 test data presented in figure 4 which shows the formation of tight clusters for all-cities and all-devices embeddings as compared to the original features.

(3) Additionally, domain-wise features provide a further improvement of 2% over all-domains features, which indicates the effectiveness of domain-wise embeddings. This improvement is supported by the t-SNE plot in figure 3 which shows the removal of city and device information from the original features. This is evident from the intermingling of colors, indicating the merging of devices and cities in the specific-device-

Datasets	Acoustic Scene Classification Accuracy (%) Device as Domain			Device Classification Accuracy (%) Acoustic Scene as Domain		
	Original	Specific-All-Devices	Specific-Device-Wise	Original	Specific-All-Scenes	Specific-Scene-Wise
DCASE 2018_SB	69.49	72.24 (k=5)	72.86 (k=10)	98.54	98.89	99.17
DCASE 2019_SB	68.96	71.72 (k=10)	72.14 (k=5)	97.42	97.87	98.37
DCASE 2020	57.95	61.35 (k=10)	NA	NA	NA	NA
Datasets	Acoustic Scene Classification Accuracy (%) City as Domain			City Classification Accuracy (%) Acoustic Scene as Domain		
	Original	Specific-All-Cities	Specific-City-Wise	Original	Specific-All-Scenes	Specific-Scene-Wise
DCASE 2018_SA	71.13	72.95 (k=10)	74.70 (k=5)	47.02	49.44	51.35
DCASE 2019_SA	72.35	74.55 (k=10)	76.11 (k=5)	35.65	37.78	40.29
DCASE 2020	57.95	60.75 (k=5)	64.86 (k=5)	30.42	29.55	30.09

Table 2. : Comparison of classification results for different tasks across five DCASE datasets, with and without multi-block formulation. 'Original' denotes results without multi-block, while 'Specific' corresponds to the multi-block formulation. NA stands for not applicable as DCASE 2020 dataset consists of unseen test devices.

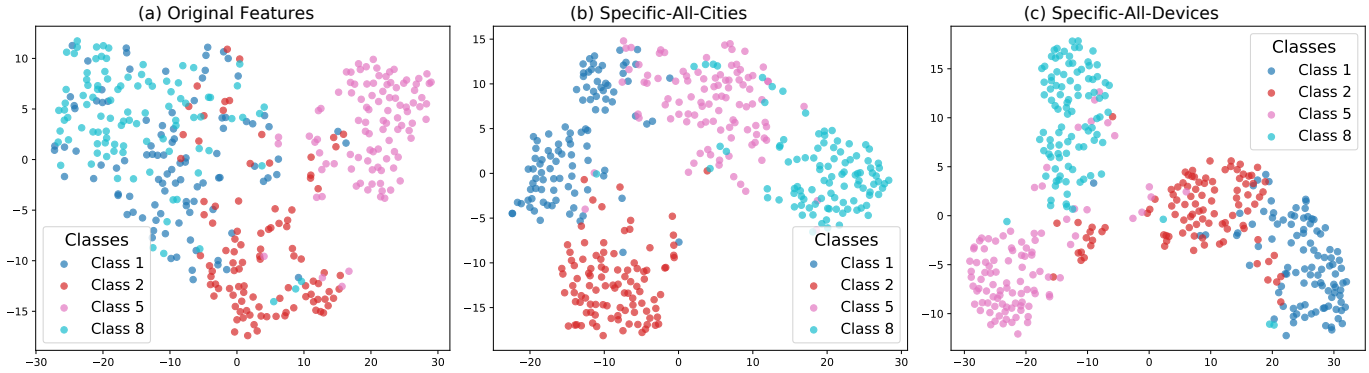

Fig. 4: t-SNE visualization of data for four acoustic scene classes namely: 'metro', 'metro station', 'public square' and 'tram' represented as class 1 , 2 , 5 and 8 respectively for DCASE 2020 test data: (a) original features, (b) specific-all-cities, and (c) specific-all-devices.

Datasets	Classification Accuracy (%)		
	Proposed (Best)	Other Methods	
DCASE 2018_SA	74.70	77.5 [21]	71.2 [26]
DCASE 2018_SB	72.86	70.6 [21]	58.2 [18]
DCASE 2019_SA	76.11	76.8 [21]	70.5 [26]
DCASE 2019_SB	72.14	76.6 [25]	72.8 [21]
DCASE 2020	64.86	70.9 [25]	63.9 [29]

Table 3. : Comparison of classification performance of proposed systems with other non-ensemble-based systems.

wise embeddings (Figure 3 (b)) and the specific-city-wise embeddings (Figure 3 (d)) respectively.

(4) In comparison, the device variation is better reduced than the city variation, as the improvement using device as the domain is slightly more than with city as the domain for the respective datasets.

Table 3 compares the performance of multi-block formulations with systems of similar complexity and ASC systems based on non-ensemble methods such as those proposed in [21], [26], [18], [25] and [29]. For some cases, the proposed systems provided better performance while for some it provided comparable performance. Furthermore, figure 5 (a) represents the confusion matrix for domain-wise-multi-block formulation using city as the domain for DCASE 2020 dataset. An observation is made that the model exhibits difficulty in distinguishing certain classes. For instance, the 'street traffic (s-tr)' category is often misclassified as 'public square (p-sq)' or 'park'. Similarly, the 'bus' class is frequently confused with 'tram' or 'airport (air)'. Among all categories, 'park' achieves the highest classification accuracy, while 'bus' records the lowest. Figure 5 (b) displays the classification performance at a higher categorical level (see section 4.1 for grouping criteria), revealing that

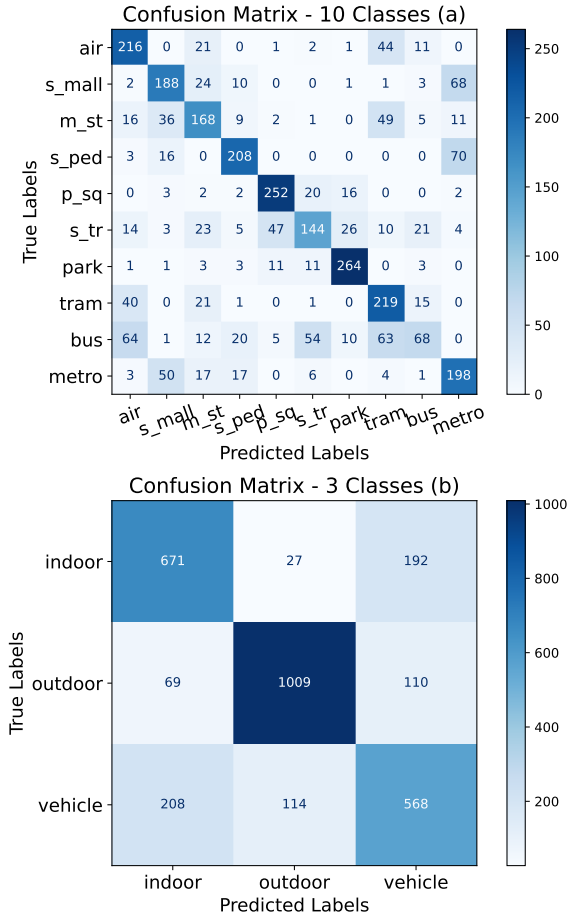

Fig. 5: 10-class confusion matrix (top panel) and 3-class confusion matrix (bottom panel) for the proposed domain-wise-multi-block formulation using city as the domain for DCASE 2020 dataset.

the system has greater difficulty in differentiating vehicle acoustic scenes as compared to indoor and outdoor ones.

5. CONCLUSION

In this paper, it is demonstrated that multi-block formulations can be used to remove unwanted domain information, and hence improve the classification performance. The domain information present in DCASE data formed the common link between the multiple blocks. Using domain-wise information provided improvements over information about all-domains. Both these formulations outperformed the baseline systems, and are comparable to systems of similar complexity. For the domain-wise-multi-block formulation, it is assumed that the domain of the test example is known at the test time. Out of all the DCASE datasets, DCASE 2020 proved to be most challenging as it contained data from unseen recording devices as well.

The scope of this work is limited to mitigate domain information from either city or device at a time. As part of future work, it is planned to use the proposed methods for simultaneously suppressing the domain information from both types of domains.

6. REFERENCES

[1] Raman Arora and Karen Livescu. Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[2] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D. Plumbley. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 2015.

[3] Helen L. Bear, Toni Heittola, Annamaria Mesaros, Emmanouil Benetos, and Tuomas Virtanen. City classification from multiple real-world sound scenes. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 11–15, 2019.

[4] Helen L Bear, Veronica Morfi, and Emmanouil Benetos. An evaluation of data augmentation methods for sound scene geotagging. *arXiv preprint arXiv:2110.04585*, 2021.

[5] Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard. Acoustic scene classification with matrix factorization for unsupervised feature learning. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[6] Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard. Feature learning with matrix factorization applied to acoustic scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1216–1229, 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[8] J. Eggert and E. Korner. Sparse coding and NMF. *2004 IEEE International Joint Conference on Neural Network*, 4:2529–2533 vol.4, 2004.

[9] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[10] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *Interspeech*, 2021.

[11] David Heise and Helen L Bear. Visually exploring multipurpose audio data. *IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2021.

[12] Harold Hotelling. Relations between two sets of variates. *Biometrika, Oxford University Press, Biometrika Trust*, 28(3/4):321–377, 1936.

[13] Liu Jie. Acoustic scene classification with residual networks and attention mechanism. *Detection and classification of acoustic scenes and events (DCASE) challenge*, 2020.

[14] Michał Kosmider. Spectrum correction: Acoustic scene classification with mismatched recording devices. *Interspeech*, 2020.

[15] Anurag Kumar, Benjamin Elizalde, and Bhiksha Raj. Audio content based geotagging in multimedia. *arXiv preprint arXiv:1606.02816*, 2016.

[16] Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[17] Mark D McDonnell and Wei Gao. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 141–145, 2020.

[18] Seongkyu Mun and Suwon Shon. Domain mismatch robust acoustic scene classification using channel information conversion. pages 845–849, 2019.

[19] Paul D. O'Grady and Barak A. Pearlmutter. Convolutive nonnegative matrix factorisation with a sparseness constraint. *16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2006.

[20] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech*, 2019.

[21] Lam Pham, Huy Phan, Truc Nguyen, Ramaswamy Palaniappan, Alfred Mertins, and Ian McLoughlin. Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework. *Digital Signal Processing*, 110:102943, 2021.

[22] William Phillips and Ellen Riloff. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, page 125–132, 2002.

[23] A. Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for svm speaker recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[24] Krishna Somandepalli, Rajat Hebbar, and Shrikanth Narayanan. Multi-face: Self-supervised multiview adaptation for robust face clustering in videos. *arXiv preprint arXiv:2008.11289*, 2020.

[25] Yizhou Tan, Haojun Ai, Shengchen Li, and Mark D Plumbley. Acoustic scene classification across cities and devices via feature disentanglement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[26] Akansha Tyagi and Padmanabhan Rajan. Location-invariant representations for acoustic scene classification. *30th European Signal Processing Conference (EUSIPCO)*, 2022.

[27] Devalraju Dhanunjaya Varma, Padmanabhan Rajan, and Aroor Dinesh Dileep. Learning to separate: Soundscape classification using foreground and background. *28th European Signal Processing Conference (EUSIPCO)*, 2021.

[28] Tuomas Virtanen, Mark D. Plumbley, and Dan Ellis. Computational analysis of sound scenes and events. *Springer Publishing Company, Incorporated*, 1st, 2017.

[29] Peiyao Wang, Zhiyuan Cheng, and Xinkang Xu. Acoustic scene classification with device mismatch using data augmentation by spectrum correction. *Detection and classification of acoustic scenes and events (DCASE) challenge*, 2020.

[30] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 2013.

[31] Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. Nonnegative matrix factorization on kernels. *Pacific Rim International Conference on Artificial Intelligence*, pages 404–412, 2006.

[32] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

[33] Guoxu Zhou, Andrzej Cichocki, Yu Zhang, and Danilo P. Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2426–2439, 2016.