Location-invariant representations for acoustic scene classification

Akansha Tyagi, Padmanabhan Rajan
School of Computing and Electrical Engineering
Indian Institute of Technology, Mandi
d19030@students.iitmandi.ac.in, padman@iitmandi.ac.in

***Abstract*—High intra-class variance is one of the significant challenges in solving the problem of acoustic scene classification. This work identifies the recording location (or city) of an audio sample as a source of intra-class variation. We overcome this variation by utilising multi-view learning, where each recording location is considered as a view. Canonical correlation analysis (CCA) based multi-view algorithms learn a subspace where samples from the same class are brought together, and samples from different classes are moved apart, irrespective of the views. By considering cities as views, and by using several variants of CCA algorithms, we show that intra-class variation can be reduced, and location-invariant representations can be learnt. The proposed method demonstrates an improvement of more than 8% on the DCASE 2018 and 2019 datasets, when compared to not using the view information.**

***Index Terms*—Acoustic scene classification, intra-class variation, multi-view learning, canonical correlation analysis.**

I. Introduction

The task of acoustic scene classification (ASC) involves assigning a label such as ‘park’, ‘metro station’ etc. to an audio recording, indicating the surroundings from where the audio was captured[1]. Overcoming high intra-class variation continues to be a challenge faced by practical ASC systems. In many situations, the recording location adds to the intra-class variability. For example, in the DCASE 2018 dataset, data for various acoustic scenes are captured in different cities (Helsinki, London etc.) Nevertheless, there are certain sounds in the acoustic scenes which are common irrespective of their recording location. For example, the ‘airport’ scene of London shares several acoustic events with the ‘airport’ scene of Helsinki.

In this paper, we try to overcome the intra-class variation due to different recording locations by considering them as different *views*. Typically, the multiple views in multi-view learning correspond to different modalities of data from an observation: for example, in a person identification task, one view could be the video of the speaker, and another view could be the audio. In computer vision, different poses or viewing angles of a scene can be considered as different views [1].

The main contribution of this work is to consider recording locations (cities) as different views, and use multi-view learning to bring together the views from the same class and move apart the views from different classes. Specifically, the contributions of this paper are:

- We show that the city corresponding to an audio scene can be regarded as a source of intra-class variation.
- We formulate the problem of ASC in a multiview paradigm to deal with this variation.
- We compare the performance of single-view and multi-view frameworks, and also compare the performance of different multi-view algorithms for this task.

Before applying multi-view analysis, we show that the ASC data we use satisfy the requirements of multi-view learning. After this, unsupervised canonical correlation analysis-based multi-view learning algorithms [2] are used on the views to derive location-invariant representations. The presence of classwise common acoustic events across the multiple recording locations motivates us to focus on the consensus principle of multi-view learning, which in turn is used by the CCA-based multi-view learning methods.

The remainder of this manuscript is organised as follows: In section II, multi-view learning is briefly reviewed. Section III describes the multi-view CCA method and its application on ASC task. Section IV describes the experimental evaluations. Finally, the conclusion and future work is presented in section V of the paper.

II. Related Work

Various methods have been proposed to reduce intra-class variation. In [3], the authors proposed adversarial multi-task learning to learn domain-invariant representations for speech recognition, where different noise conditions correspond to different domains. In [4], a new loss function was suggested to increase the intra-class compactness to learn better representations for the task of speaker verification. For the same task, the authors in [5] introduced a training method that reduces the intra-class variation by decreasing the mutual information between speaker-related and speaker-unrelated embeddings.

Multi-view learning methods have been used to reduce intra-class variation in domains like computer vision [6] [7], and natural language processing [8]. Xu et al. [1] provide an overview of multi-view learning, its fundamental principles, and its different types (co-training, multiple-kernel learning, and subspace learning). CCA-based multi-view learning falls under the umbrella of subspace based multi-view methods. A good review on CCA-based methods can be found in [9]. CCA [2] is a two-view subspace learning method constructed on the

[1] https://dcase.community/challenge2018/index

idea of maximizing correlation between the views. Livescu et al. [10] used it for speaker recognition wherein acoustics and videos of the speaker's face form the two views. However, CCA learns only linear transformations. Kernel Canonical Correlation Analysis (KCCA) deals with this limitation. It is a kernel variant of CCA and computes a non-linear relationship between the views. Raman et al. [11] used it to learn non-linear transformations of features for phonetic frame classification. The same authors, in another work [12] tested the same features for domain or speaker independence. Andrew et al. [13] proposed a deep variant of CCA called Deep Canonical Correlation Analysis which learns better representations than both CCA and KCCA.

The above mentioned methods use only two views in their respective multi-view formulations. CCA variants like Multiset Canonical Correlation Analysis (MCCA) works on two and more than two views. Somandepalli et al. [14] introduced a deep version of MCCA called Deep Multiset Canonical Correlation Analysis (dMCCA) which uses deep learning to learn maximally correlated non-linear representations and used it for Noisy-MNIST dataset classification. The authors in [15] used dMCCA for speaker and speech command classification. Speech command classification uses multiple speakers as multiple views, whereas speaker classification considers various speech commands from the same speaker as multiple views. In other wok, phan et al. [16] used multiple feature representations of audio data (Mel-scaled spectrogram, Gammatone spectrogram, Constant-Q transform spectrogram and raw-audio) as multiple views to perform audio and music classification tasks.

III. Multi-view CCA

The benchmark datasets for ASC task, like DCASE 2018 and 2019, contain data for different scenes (classes) collected across multiple cities. We attempt to develop location-independent representations for an acoustic scene by considering data across multiple cities as its multiple views. The foundation of multi-view learning algorithms is based upon either consensus or complementary principle [1]. In this work, CCA-based multi-view learning methods are used that are built upon the consensus principle, which aim to maximize the agreement amongst the multiple views.

We begin by considering the data of all classes from two cities, as two views represented by the matrices $C_1, C_2 \in \mathbb{R}^{d\times N}$ containing N examples, each of d dimensions. Exploiting the consensus principle, CCA-based methods aim to find maximally correlated projection matrices W_1, $W_2 \in \mathbb{R}^{d\times K}$ corresponding to the two cities, where K represents the projection dimension. The problem of determining these projection matrices can be formulated as a constrained optimization problem, with the objective function and constraints given by [9]:

$$\begin{aligned} &\max_{\mathbf{w}_1,\mathbf{w}_2} \ \mathbf{w}_1^T C_1 C_2^T \mathbf{w}_2 \\ &\text{s.t} \ \ \mathbf{w}_1^T C_1 C_1^T \mathbf{w}_1 = 1, \\ &\qquad \mathbf{w}_2^T C_2 C_2^T \mathbf{w}_2 = 1 \end{aligned} \tag{1}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are projection vectors forming the columns of W_1 and W_2 respectively. The same procedure is repeated K times for each projection matrix, with an additional constraint that all K sets of projection vectors must be different from each other [17].

In this work, we have used three CCA-based unsupervised multi-view learning algorithms, namely Multiset Canonical Correlation Analysis (MCCA), Kernel Multiset Canonical Correlation Analysis (KMCCA), and Deep Multiset Canonical Correlation Analysis (dMCCA). Each of these algorithms is a formulation of the multi-view problem as a constrained maximization problem, as discussed earlier. In MCCA [17], linear projections are computed corresponding to m views by solving the following constrained optimization problem

$$\begin{aligned} &\max_{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_m} \sum_{i<j} \mathbf{w}_i^T C_i C_j^T \mathbf{w}_j \\ &\text{s.t} \ \ \mathbf{w}_i^T C_i C_i^T \mathbf{w}_i = 1 \quad \forall i,j \in \{1,\ldots,m\} \end{aligned} \tag{2}$$

where $C_i \in \mathbb{R}^{d\times N}$ is the i^{th} view matrix and $\mathbf{w}_i \in \mathbb{R}^d$ is the corresponding projection vector.

In KMCCA [17], non-linear projections are computed for the views. Initially, a kernel method is used to project data to a higher dimension. Following this, the projection vectors can be determined by solving the following constrained optimization problem

$$\begin{aligned} &\max_{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_m} \sum_{i<j} \mathbf{w}_i^T \Phi_i \Phi_j^T \mathbf{w}_j \\ &\text{s.t} \ \ \mathbf{w}_i^T \Phi_i \Phi_i^T \mathbf{w}_i = 1 \quad \forall i,j \in \{1,\ldots,m\} \end{aligned} \tag{3}$$

where Φ_i is the projection of C_i in the higher dimensional space. The power of deep neural networks (DNNs) is used by dMCCA [14] to learn complex transformations present in the data and provide better projections. The optimization problem for dMCCA is re-formulated by replacing the view matrix C_i with $H_i = f_i(C_i)$ in (2), where f_i is the non-linear functional mapping the DNN's input to its output, corresponding to i^{th} view. For more details on MCCA and KMCCA, we refer the readers to the reference [17], and [14] for dMCCA.

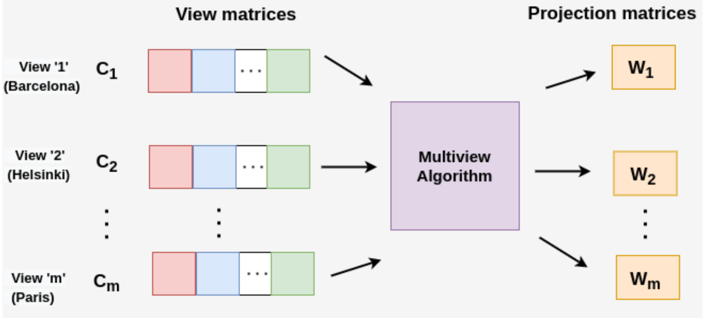

Fig. 1: Projection matrix generation for the multi-view framework. Different colors in a view matrix correspond to different classes.

In this work, we first construct the data in the form of multiple views using 'view generation' (refer subsection IV-B) then obtain the projection matrices corresponding to each view as described in figure 1. Each view matrix C_i is given as an input to the multi-view algorithm to obtain the respective projection matrix W_i. An audio sample $\mathbf{x}$ (a column of C_i) is projected onto the multi-view space (subspace learnt by the multi-view algorithm) as:

$$\mathbf{p} = \left[(W_1^T\mathbf{x})^T \quad (W_2^T\mathbf{x})^T \quad \ldots \quad (W_m^T\mathbf{x})^T\right]^T \qquad (4)$$

where $\mathbf{p} \in \mathbb{R}^{Km}$ is the multi-view projection of $\mathbf{x}$ and Km is the dimension of the multi-view space. The multi-view embeddings computed for all the audio samples are then provided as an input to a classifier.

IV. Experimental Evaluation

In this section, the experiments performed for the multi-view framework for acoustic scene classification are described. The objective is twofold:

- To study the performance gains provided by multi-view processing of recording locations.
- To evaluate different feature representations in the above framework, as well as comparison with other methods.

A. Dataset description

We use DCASE 2018 and 2019 ASC (subtask ‘A’) development datasets containing audio samples for 10 classes [18]. DCASE 2018 contains audio recordings from six European cities namely Barcelona (B), Helsinki (H), London (L), Paris (P), Stockholm (S), and Vienna (V). DCASE 2019 contains data from Lisbon (Li), Lyon (Ly), Milan (M), and Prague (Pr) in addition to the cities present in DCASE 2018. Thus, the number of views m is 6 and 10, respectively for these two datasets.

B. View generation

In the data distributed with DCASE, it is not guaranteed that the number of recordings per class are same across all cities. The underlying assumption in multi-view learning requires the same number of observations per view. The DCASE protocols permit data augmentation, and hence we use a variant of mixup [19] data augmentation scheme to satisfy the above mentioned constraint. We perform class-wise mixup [20] for each view, which takes two features $\mathbf{x}_1$ and $\mathbf{x}_2$ derived from their corresponding raw audio samples, and generates a new feature $\mathbf{x}_3$ of the same view and class by taking a convex combination:

$$\mathbf{x}_3 = \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \qquad (5)$$

where $\alpha \in U(0,1)$ i.e α is drawn from a uniform distribution. Table I show the number of training examples in each city before and after class-wise mixup for DCASE 2018 and 2019. After this step view generation is complete. View generation is followed by view validation. The multiple views must satisfy the properties of ‘view dependency’ (correlation between the views) and ‘view sufficiency’ (each view is sufficient for classification on its own) as apart of view validation [1].

Figure 2 represents the histogram of pairwise inter-view correlation for DCASE 2019 dataset computed by considering correlation coefficients of examples from matched classes, across multiple views. The positive correlation between views indicates view dependency.

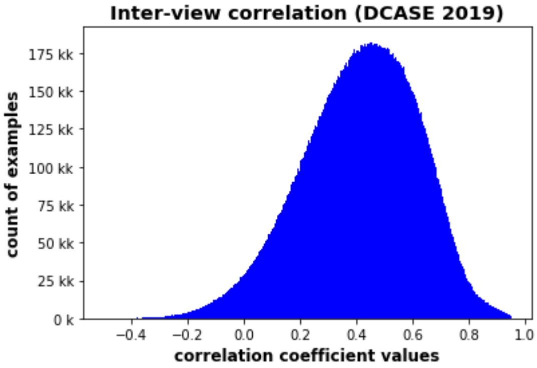

Fig. 2: Inter-view correlation histogram of original features from matched classes, across multiple views for DCASE 2019 dataset.

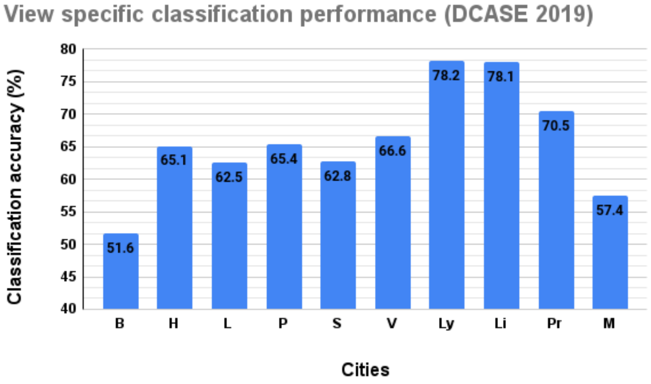

Fig. 3: View specific classification performance for DCASE 2019; refer to section IV-A for city name expansions.

For verification of the ‘view sufficiency’ property, we construct view-specific classifiers and evaluate their respective classification performance by testing on the same view data. Figure 3 shows the classification performance with respect to different cities for DCASE 2019 dataset. All views provide $> 50\%$ accuracy, demonstrating the satisfaction of view sufficiency property. Similar conclusions were obtained for the DCASE 2018 dataset as well.

C. Feature extraction

We have used two different pre-trained networks for feature extraction, namely L^3-Net [21] and Soundnet [22]. The former network takes raw audio signal as input and outputs an embedding of size 512 × 97. We compute the average across the time axis to get a column vector of dimension 512. The latter network also takes raw audio waveform as input, which passes through its 8 layer architecture. Features are extracted from intermediate C5 layer by averaging the layer’s output to obtain a column vector of dimension 256. These features are called ‘original features’, denoted by $\mathbf{x}$ in equation 4.

City name	B	H	L	P	S	V	Li	Ly	M	Pr
Before mixup	1051	1015	964	1014	1013	1065	1061	976	1030	1026
After mixup	1320	1320	1320	1320	1320	1320	1320	1320	1320	1320

TABLE I: City wise training data before and after class-wise mixup for DCASE 2019. Mixup is used only to make the number of examples same for each view, refer to section IV-A for city name expansions.

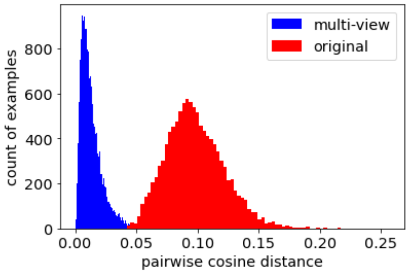

Fig. 4: Histogram of pairwise cosine distance of multi-view and original features from the same class across two different views.

For all the experiments, the number of examples $N = 1320$ and the projection dimension is $K = 9$, which is determined experimentally. The dimension d of the original features $\mathbf{x}$ is 512 for L^3-net and 256 for Soundnet. Since, the dimension of the multi-view features $\mathbf{p}$ depends on the number of views thus, it is 54 and 90 for DCASE 2018 and 2019 respectively.

D. Effect of multi-view representation

The multi-view framework effectively reduces the pairwise distance between the examples of same class across different views. Figure 4 shows a histogram of the pairwise cosine distances of the original data vectors $\mathbf{x}$ and their multi-view features $\mathbf{p}$ for the acoustic scene 'shopping-mall', across 'Barcelona' and 'Helsinki'. Similar observations are obatined for other scenes as well. The transformation of the original features to multi-view features as expressed in equation 4 results in the reduction of pairwise distances between the examples.

Figure 5 represents t-SNE plots for DCASE 2018 train and test data obtained before and after multi-view learning. Inter mixing of different colors in figures 5(a) and 5(c) shows high intra-class variation in train and test data embeddings before multi-view learning. The cluster formation of different colors in figures 5(b) and 5(d) indicate the reduction of variation in both train and test data after multi-view learning.

E. Classification

We use k-NN (k-nearest neighbours) classifier for the demonstration of view sufficiency property and to compare the performance of the system consisting of original features (termed as OF) with that of the multi-view features (MCCA, KMCCA, dMCCA). The value of k is chosen to be 5 experimentally for all the frameworks.

F. Results and Discussion

Figure 6 demonstrates the location invariance of multi-view features $\mathbf{p}$ for the DCASE 2019 dataset. Here, the k-NN classifier is trained with the data of a single city and tested with a subset of data from the remaining cities. This is performed for both the original features $\mathbf{x}$ and the multi-view features $\mathbf{p}$. It can be seen that the multi-view features generalize better than the original features for all locations.

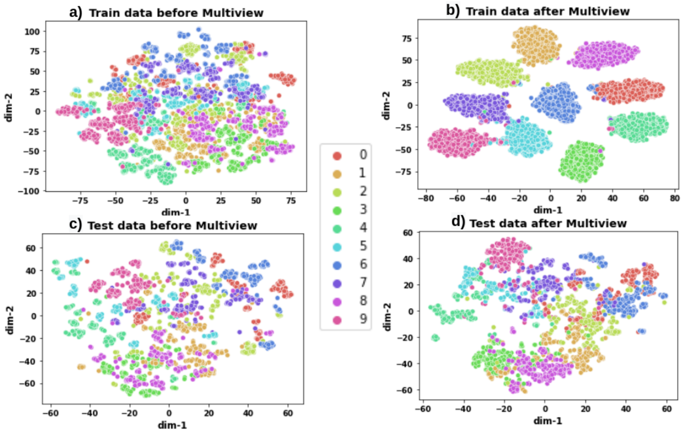

Fig. 5: t-SNE plots a) train data before multi-view b) train data after multi-view c) test data before multi-view d) test data after multi-view. Different colors represent different classes.

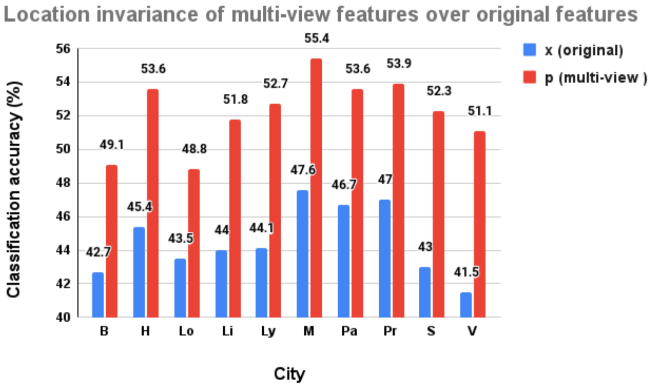

Fig. 6: Plot demonstrating the location invariance of multi-view features $\mathbf{p}$ over original features $\mathbf{x}$.

Figure 7 shows the performance of different systems: OF, MCCA, KMCCA, and dMCCA for L^3-Net and Soundnet features on DCASE 2018 and 2019 datasets. All the systems are trained using training data from all the views. It is evident from figures 7(a) and 7(b) that L_3-net features are more discriminative than those of Soundnet, as the former network is trained on a larger and more diverse dataset.

For L_3-net features, all the multi-view systems perform

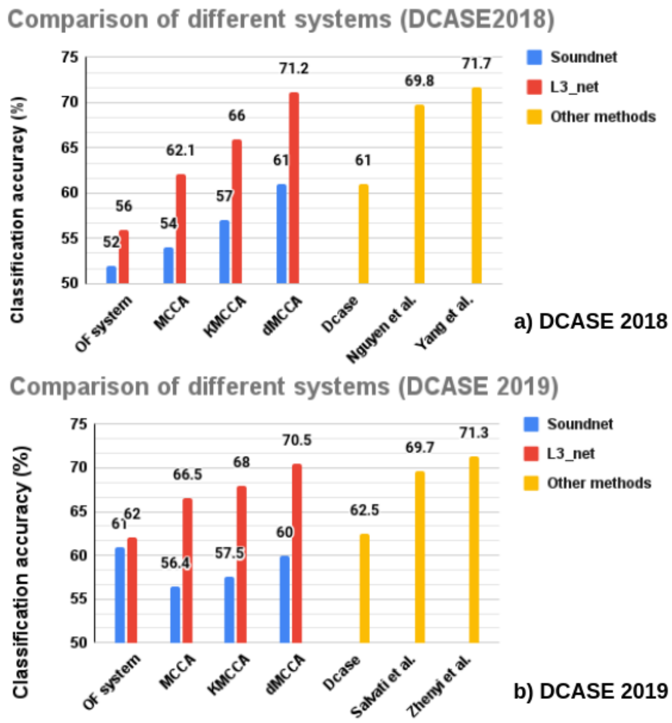

Fig. 7: Comparison of performance of different systems and features for DCASE 2018 and 2019 dataset.

better than the OF system indicating the effectiveness of multi-view learning. Amongst the three multi-view systems, dMCCA performs better than the other two owing to the advantage of using deep learning which effectively learns the complex representations present in the real-world data. Overall, the combination of L^3-Net features and dMCCA gives the best performance with 71.2% classification accuracy for DCASE 2018 and 70.5% for DCASE 2019 datasets.

Figure 7(a) (yellow bars) also shows the comparison with other methods : DCASE baseline [18], Nguyen et al. [23] and Yang et al. [24] for DCASE 2018 dataset. Figure 7(b) (yellow bars) provides the comparison with : DCASE baseline [18], Salvati et al. [25] and Zhenyi et al. [26] for DCASE 2019 dataset.

It is to be noted that, in these experiments, we make no attempt to obtain state-of-the-art results for the respective datasets; rather the objective is to verify the reduction of intra-class variability using multi-view learning. The use of large-scale data augmentation, and classifier ensembles should further improve the performance.

V. Conclusion

This paper explores the effectiveness of using multi-view learning for reducing intra-class variation. By considering recording locations as multiple views and utilising CCA-based algorithms, considerable increase in classification accuracy was obtained for acoustic scene classification. The assumption here is the access to view information while training. Future work includes the use of supervised multi-view learning techniques, and the study of generalizability when one or more views are not available during training. The code is available on Github https://github.com/AkanshaTyagi-06/eusipco_2022

References

[1] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[2] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[3] Y. Shinohara, "Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition," in *Interspeech 2016*.

[4] N. Le and J.-M. Odobez, "Robust and Discriminative Speaker Embedding via Intra-Class Distance Variance Regularization," in *Interspeech 2018*.

[5] Y. Kwon, S.-W. Chung, and H.-G. Kang, "Intra-class variation reduction of speaker representation in disentanglement framework," *arXiv preprint arXiv:2008.01348*, 2020.

[6] K. Somandepalli, R. Hebbar, and S. Narayanan, "Multi-face: Self-supervised multiview adaptation for robust face clustering in videos," *arXiv preprint arXiv:2008.11289*, 2020.

[7] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Transactions on Image Processing*, 2015.

[8] W. Phillips and E. Riloff, "Exploiting strong syntactic heuristics and co-training to learn semantic lexicons," *EMNLP '02*, 2002.

[9] C. Guo and D. Wu, "Canonical Correlation Analysis (CCA) Based Multi-View Learning: An Overview," *ArXiv*, vol. abs/1907.01693, 2019.

[10] K. Livescu and M. Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition Understanding*, 2009.

[11] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," in *MLSLP*, 2012.

[12] ——, "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *ICASSP*, 2013.

[13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," *ICML'13*, 2013.

[14] K. Somandepalli, N. Kumar, R. Travadi, and S. Narayanan, "Multimodal representation learning using deep multiset canonical correlation," *arXiv preprint arXiv:1904.01775*, 2019.

[15] K. Somandepalli, N. Kumar, A. Jati, P. G. Georgiou, and S. Narayanan, "Multiview shared subspace learning across speakers and speech commands." in *INTERSPEECH*, 2019.

[16] H. Phan, H. Le Nguyen, O. Y. Chén, L. Pham, P. Koch, I. McLoughlin, and A. Mertins, "Multi-view audio and music classification," *ICASSP*, 2021.

[17] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *SiKDD*, 2010.

[18] T. Heittola, A. Mesaros, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *DCASE 2018 Challenge*, 2018.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2018.

[20] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, A. Munawar, B. J. Ko, N. Greco, and R. Tachibana, "Shuffling and mixing data augmentation for environmental sound classification," *DCASE 2019 Challenge*, 2019.

[21] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," *ICASSP*, 2019.

[22] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016.

[23] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," *DCASE 2018 Challenge*, 2018.

[24] J. H. Yang, N. K. Kim, and H. K. Kim, "Se-resnet with gan-based data augmentation applied to acoustic scene classification," *DCASE 2018 Challenge*, 2018.

[25] D. Salvati, C. Drioli, and G. L. Foresti, "Urban acoustic scene classification using raw waveform convolutional neural networks," *DCASE 2019 Challenge*, 2019.

[26] Z. Huang and D. Jiang, "Acoustic scene classification based on deep convolutional neuralnetwork with spatial-temporal attention pooling," *DCASE 2019 Challenge*, 2019.