Unsupervised Domain Adaptation for Acoustic Scene Classification Using Band-Wise Statistics Matching

Alessandro Ilic Mezza*, Emanuël A. P. Habets†, Meinard Müller† and Augusto Sarti*
*Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy
Email: alessandroilic.mezza@polimi.it, augusto.sarti@polimi.it
†International Audio Laboratories Erlangen⋆, Am Wolfsmantel 33, 91058 Erlangen, Germany
Email: emanuel.habets@audiolabs-erlangen.de, meinard.mueller@audiolabs-erlangen.de

***Abstract*—The performance of machine learning algorithms is known to be negatively affected by possible mismatches between training (source) and test (target) data distributions. In fact, this problem emerges whenever an acoustic scene classification system which has been trained on data recorded by a given device is applied to samples acquired under different acoustic conditions or captured by mismatched recording devices. To address this issue, we propose an unsupervised domain adaptation method that consists of aligning the first- and second-order sample statistics of each frequency band of target-domain acoustic scenes to the ones of the source-domain training dataset. This approach is devised to adapt audio samples from unseen devices before they are fed to a pre-trained classifier, thus avoiding any further learning phase. Using the DCASE 2018 Task 1-B development dataset, we show that the proposed method outperforms the state-of-the-art unsupervised methods found in the literature in terms of both source- and target-domain classification accuracy.**

***Index Terms*—Unsupervised domain adaptation, mismatched recording devices, acoustic scene classification.**

I. Introduction

Acoustic Scene Classification (ASC) is the task of assigning a categorical label to a test audio recording to characterize the environment in which it was captured — for instance "Metro station", "Park", "Airport". In recent years, deep learning (DL) has proven to be an essential and powerful tool to effectively tackle this problem [1]–[4]. However, as a downside, DL-based ASC systems tend to be susceptible to the effects of domain shift, i.e., the well-known performance degradation that affects machine learning algorithms when trained and tested on data drawn from different distributions [5]. Domain adaptation (DA), despite having been extensively investigated in fields such as natural language processing [6], [7], sentiment analysis [8], [9] and computer vision [10], [11], is still a relatively new topic in the context of ASC. Since 2018, the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) have included a subtask specifically designed to encourage DA, namely Task 1-B on "Acoustic Scene Classification with mismatched recording devices." In this task, each recording device is regarded as a separate domain. Nonetheless, the training datasets adopted by the DCASE challenges in 2018 [3] and 2019 [4], although being highly unbalanced in favour of a single recording device, also contain several acoustic scenes that were simultaneously captured using both source (Device A) and target devices (Devices B and C); we refer to such data as *parallel* data. As a result, most ASC models found in the literature are trained using both source- and target-domain samples and thus are not kept blind to the target domain prior to the adaptation.

To date, only a few studies have applied unsupervised DA techniques to ASC models that were trained solely on source-domain data. In [12] and [13], the authors propose to adapt a DL-based ASC model by means of adversarial learning. In particular, [12] follows the Adversarial Discriminative Domain Adaptation framework presented in [14] to adapt the convolutional layers of a pre-trained CNN so to force the feature extractor into yielding domain-invariant data representations. Furthermore, [13] improves over [12] by replacing the adversarial adaptation process with a module based on Wasserstein Generative Adversarial Networks (WGAN) [15]. Meanwhile, the authors of [16] recently proposed a different paradigm: instead of adapting a pre-trained model, DA is enforced directly on the acoustic scenes using a Factorized Hierarchical Variational AutoEncoder (FHVAE). This method aims to disentangle scene-dependent and channel-related characteristics in terms of a pair of latent variables z_1 and z_2. Afterwards, a "channel conversion" step is performed in the latent space by shifting z_2 by a domain-specific factor $\Delta\mu_2$.

The strategy of applying DA on audio data before model training and evaluation is adopted also by the winning submission to DCASE 2019 Task 1-B [17], where the main idea is to equalize the different frequency responses of mismatched recording devices. To this end, a set of spectral coefficients is computed as the ratio of the spectra of simultaneous recordings from different devices. Coefficients are then averaged over multiple pairs. Finally, spectral correction is applied by multiplying each frequency bin of the short-time Fourier transform of every acoustic scene by the coefficient associated to the corresponding frequency band.

Despite having proven to be quite effective, not only [17] requires several target-domain samples prior to the training phase, but it also makes a further assumption on the availability of parallel audio files. In turn, the adversarial DA methods found in the literature [12], [13] suffer from two critical limitations. First, they entail a whole new adaptation phase every time a novel target domain is encountered. Second,

⋆ A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

they require a suitably sized target-domain dataset to train the adaptation module. The technique presented in [16], while being designed to address the latter shortcomings, requires an auxiliary dataset of acoustic scenes (dubbed "Universal domain") in order to pre-train the FHVAE and thus compute the channel conversion parameter $\Delta\mu_2$. In [16], two additional variants are described: the first one uses source-domain data for the pre-training, while the second employs target-domain data. The latter variant, however, despite being the best performing of the three, violates the requirement of not relying on information from the target devices at training time. Moreover, the classifier is learnt using the reconstructed features decoded by the FHVAE: the adaptation procedure cannot be readily applied to any previously optimized ASC model, as it inevitably entails a training.

In this paper, we present an effective unsupervised DA procedure for ASC that is capable to overcome the limitations of adversarial strategies by performing DA at data level, but without requiring auxiliary audio recordings as in [16]. The main idea, akin to that of the Cepstral Mean and Variance Normalization (CMVN) method proposed for automatic speech recognition [18], is to apply a preprocessing technique prior to the test phase in which the first- and second-order sample statistics of each frequency band of test data are matched to the ones of the source-domain training dataset.

The proposed approach consists of three main steps. First, just before the training phase, we compute the sample mean and standard deviation of each frequency band across every sample in the source-domain training dataset. Second, at inference time, we apply a band-wise standardization to the target-domain test data so to obtain zero-mean and unit-variance frequency bands throughout the dataset. Third, we finally adapt the standardized dataset using the means and standard deviations computed at Step 1. We show that this procedure can significantly increase the target-domain performance of an ASC system, while having a low computational cost and not substantially affecting the results of source-domain classification.

II. Proposed Method

Let $\mathbf{x} \in \mathbb{R}^{M \times K}$ be the log-amplitude mel-frequency spectrogram of an acoustic scene, where M and K represent the number of time frames and mel bands, respectively. Let $\mathbf{X}^{\mathcal{S}} \in \mathbb{R}^{N_{\mathcal{S}} \times M \times K}$ and $\mathbf{X}^{\mathcal{T}} \in \mathbb{R}^{N_{\mathcal{T}} \times M \times K}$ indicate the source- and the target-domain datasets, respectively, where $N_{\mathcal{S}}$ is the number of source spectrograms and $N_{\mathcal{T}}$ the number of target spectrograms. In the following, we assume $\mathbf{X}^{\mathcal{S}}$ and $\mathbf{X}^{\mathcal{T}}$ to be disjoint. Furthermore, we use n, m, k as subscripts to index the tensors $\mathbf{X}^{\mathcal{S}}$ and $\mathbf{X}^{\mathcal{T}}$ in the 1st, 2nd and 3rd dimension, respectively. Finally, let each domain be characterized by a different distribution, i.e., let $\mathbf{x}_l^{\mathcal{S}} \sim \mathcal{X}^{\mathcal{S}}$ for $l = 1, ..., N_{\mathcal{S}}$ and $\mathbf{x}_\ell^{\mathcal{T}} \sim \mathcal{X}^{\mathcal{T}}$ for $\ell = 1, ..., N_{\mathcal{T}}$, where $\mathcal{X}^{\mathcal{S}}$ and $\mathcal{X}^{\mathcal{T}}$ are the source and target data distributions.

The proposed adaptation procedure comprises three steps. First, we compute $\mu_k^{\mathcal{S}}$ and $\sigma_k^{\mathcal{S}}$ from $\mathbf{X}^{\mathcal{S}}$ as in:

$$\mu_k^{\mathcal{S}} = \frac{1}{N_{\mathcal{S}} M} \sum_{n=1}^{N_{\mathcal{S}}} \sum_{m=1}^{M} \mathbf{X}_{nmk}^{\mathcal{S}} \tag{1}$$

$$\sigma_k^{\mathcal{S}} = \sqrt{\frac{1}{N_{\mathcal{S}} M - 1} \sum_{n=1}^{N_{\mathcal{S}}} \sum_{m=1}^{M} \left(\mathbf{X}_{nmk}^{\mathcal{S}} - \mu_k^{\mathcal{S}}\right)^2} \tag{2}$$

for $k = 1, ..., K$. Intuitively, the values of $\mu_k^{\mathcal{S}}$ and $\sigma_k^{\mathcal{S}}$ are computed as the sample mean and standard deviation of the vector obtained by concatenating the k-th row of every spectrogram in $\mathbf{X}^{\mathcal{S}}$. Similarly, we compute $\mu_k^{\mathcal{T}}$ and $\sigma_k^{\mathcal{T}}$ for the target-domain dataset, i.e.,

$$\mu_k^{\mathcal{T}} = \frac{1}{N_{\mathcal{T}} M} \sum_{n=1}^{N_{\mathcal{T}}} \sum_{m=1}^{M} \mathbf{X}_{nmk}^{\mathcal{T}} \tag{3}$$

$$\sigma_k^{\mathcal{T}} = \sqrt{\frac{1}{N_{\mathcal{T}} M - 1} \sum_{n=1}^{N_{\mathcal{T}}} \sum_{m=1}^{M} \left(\mathbf{X}_{nmk}^{\mathcal{T}} - \mu_k^{\mathcal{T}}\right)^2} \tag{4}$$

for $k = 1, ..., K$.

From this, we then standardize $\mathbf{X}^{\mathcal{T}}$ by setting

$$\mathbf{Z}_{nmk}^{\mathcal{T}} = \frac{\mathbf{X}_{nmk}^{\mathcal{T}} - \mu_k^{\mathcal{T}}}{\sigma_k^{\mathcal{T}}} \tag{5}$$

for $n = 1, ..., N_{\mathcal{T}}$, $m = 1, ..., M$ and $k = 1, ..., K$.

In the third step, the first- and second-order statistics of the source domain are finally used to transform the standardized target-domain data as follows:

$$\overline{\mathbf{X}}_{nmk}^{\mathcal{T}} = \sigma_k^{\mathcal{S}} \mathbf{Z}_{nmk}^{\mathcal{T}} + \mu_k^{\mathcal{S}} \tag{6}$$

for $n = 1, ..., N_{\mathcal{T}}$, $m = 1, ..., M$ and $k = 1, ..., K$.

At this point, $\overline{\mathbf{X}}^{\mathcal{T}}$ has been aligned to the source domain and shares the same band means and variances with $\mathbf{X}^{\mathcal{S}}$. Our hypothesis is that $\overline{\mathbf{x}}^{\mathcal{T}}$ in $\overline{\mathbf{X}}^{\mathcal{T}}$ would now be drawn from a distribution $\bar{\mathcal{X}}^{\mathcal{T}}$ which should be closer to $\mathcal{X}^{\mathcal{S}}$ than $\mathcal{X}^{\mathcal{T}}$, and thus that an ASC model trained on $\mathbf{X}^{\mathcal{S}}$ would achieve higher classification rates when evaluated on the aligned dataset $\overline{\mathbf{X}}^{\mathcal{T}}$ rather than on the non-adapted $\mathbf{X}^{\mathcal{T}}$.

The proposed transformation amounts to an identity when applied to the source-domain training dataset: only test data are adapted as described in this section. The proposed method is not tied to a specific target domain and does not involve any learning phase. This makes it effectively independent of the choice of the following classification algorithm which can be trained at an earlier stage. Notably, the entire adaptation procedure is completely unsupervised, i.e., it does not require target-domain labels at any given time.

III. Evaluation

A. Training and Evaluation

We evaluate the proposed adaptation procedure using two different ASC models. The first one, used also in [16], is the baseline system of the DCASE 2018 Challenge [3] (denoted as "DCASE model" from now on). The second one, used in [12] and [13], is the so-called "Kaggle model". The neural network

architectures are implemented in PyTorch as described in Tables I and II.

To be consistent with the evaluation setup of the state of the art, the dataset used for the development and evaluation of the proposed approach is the one provided as the development dataset of Task 1-B of the DCASE 2018 Challenge [3]. The dataset contains 10-seconds long WAV files captured in six different large European cities using three different recording devices — namely, devices A, B and C. Each audio item is categorized by one of ten scene labels.

Before being fed to the learning algorithms, the audio data are transformed into time-frequency features. Specifically, we extract log mel-energies with different parameters depending on the model. For the DCASE model, we use 40 mel bands and a 40 ms Hamming window with 50% overlap. For the Kaggle model, we use 64 mel bands and a 2048 samples ($\sim$ 46 ms) Hamming window with 50% overlap. We adopt the same training, validation and test folds as in [12]. In particular, the training set consists of 5510 audio clips (only Device A) and the validation set of 612 (only Device A). Furthermore, the test dataset contains 2878 clips recorded by different devices, namely 2518 files from Device A, 180 from Device B, and 180 from Device C. Note that training data from Devices B and C are disregarded.

We optimize the DCASE model for 200 epochs using Adam. The learning rate and the batch size were set to 10^{-4} and 16, respectively. For the Kaggle model, to foster evaluation consistency and to show that our method can be effectively decoupled from the model training, we utilize the pre-trained weights made available online[1] by the authors of [12].

To be able to compare the results with prior works, we adopt an evaluation setup similar to the one of the DCASE 2018 Challenge, where the target-domain performance is assessed by averaging the accuracy obtained on Devices B and C. Hence, we evaluate our models on $\overline{\mathbf{X}}_{\text{test}}^{(\text{B,C})}$ and $\mathbf{X}_{\text{test}}^{(\text{B,C})}$, i.e., the adapted and non-adapted test fold of Devices B and C combined. Moreover, to investigate the effect of the proposed adaptation method when applied to source-domain data, we evaluate our models using both $\overline{\mathbf{X}}_{\text{test}}^{(\text{A})}$ and $\mathbf{X}_{\text{test}}^{(\text{A})}$, i.e., the adapted and non-adapted test fold of Device A, respectively.

B. Device-Dependent Adaptation

If we assume to have access to the knowledge of which device captured each acoustic scene in the target-domain dataset, i.e., if we assume that test samples are annotated with *device labels*, we can consider the target devices separately during the adaptation phase. This means that Devices B and C are regarded as target domains in their own right and aligned independently from one another using the statistics of the source-domain training dataset (Device A), i.e.,

$$\mathbf{X}^{\mathcal{S}} := \mathbf{X}_{\text{train}}^{(\text{A})} \qquad \mathbf{X}_{\text{B}}^{\mathcal{T}} := \mathbf{X}_{\text{test}}^{(\text{B})} \qquad \mathbf{X}_{\text{C}}^{\mathcal{T}} := \mathbf{X}_{\text{test}}^{(\text{C})} \tag{7}$$

where $\mathbf{X}_{\text{train}}^{(\text{A})}$ represents the training fold composed of $N_{\mathcal{S}} = 5510$ samples from Device A, while $\mathbf{X}_{\text{test}}^{(\text{B})}$ and $\mathbf{X}_{\text{test}}^{(\text{C})}$ denote the test folds from Devices B and C, each consisting of $N_{\mathcal{T}} = 180$ mel-spectrograms.

Eventually, the two adapted target-domain datasets are concatenated to form the test dataset

$$\overline{\mathbf{X}}_{\text{test}}^{(\text{B,C})} := \left[\overline{\mathbf{X}}_{\text{B}}^{\mathcal{T}} , \overline{\mathbf{X}}_{\text{C}}^{\mathcal{T}} \right] \in \mathbb{R}^{2N_{\mathcal{T}} \times M \times K} \tag{8}$$

and the model evaluation can proceed as usual. In the following, we will refer to this approach as "Device-Dependent Adaptation" (DDA).

C. Device-Independent Adaptation

To account for those cases in which the target-domain device labels are not available, we investigate the performance of the proposed adaptation method when one would consider a single domain comprising the features from both Device B and C. This means that the target domain consists of the concatenated dataset

$$\mathbf{X}_{\text{B|C}}^{\mathcal{T}} := \left[\mathbf{X}_{\text{test}}^{(\text{B})} , \mathbf{X}_{\text{test}}^{(\text{C})} \right] \in \mathbb{R}^{N'_{\mathcal{T}} \times M \times K} \tag{9}$$

where $N'_{\mathcal{T}} = 360$ is the number of samples in the target domain. Then, having applied the proposed method to $\mathbf{X}_{\text{B|C}}^{\mathcal{T}}$, we can define the adapted target-domain test dataset as

$$\overline{\mathbf{X}}_{\text{test}}^{(\text{B,C})} := \overline{\mathbf{X}}_{\text{B|C}}^{\mathcal{T}} \tag{10}$$

In the following, we will refer to this approach as "Device-Independent Adaptation" (DIA).

TABLE I
DCASE MODEL

Input (40 mel bands)
7×7–Conv2D–32–BatchNormalization–ReLU 5×5–MaxPooling2D Dropout(.3) 7×7–Conv2D–64–BatchNormalization–ReLU 4×100–MaxPooling2D Dropout(.3)
Dense–100–ReLU–Dropout(.3)
Output–10–Softmax

TABLE II
KAGGLE MODEL

Input (64 mel bands)
11×11–Conv2D–48–stride(2,3)–padding(5)–ReLU 3×3–MaxPooling2D–stride(1,2) BatchNormalization 5×5–Conv2D–128–stride(2,3)–padding(2)–ReLU 3×3–MaxPooling2D–stride(2) BatchNormalization 3×3–Conv2D–192–stride(1)–padding(1)–ReLU 3×3–Conv2D–192–stride(1)–padding(1)–ReLU 3×3–Conv2D–128–stride(1)–ReLU 3×3–MaxPooling2D–stride(1,2) BatchNormalization
Dense–256–ReLU–Dropout(.25) Dense–256–ReLU–Dropout(.25)
Output–10–Softmax

[1] https://doi.org/10.5281/zenodo.1401995

TABLE III

CLASSIFICATION ACCURACY ON ADAPTED AND NON-ADAPTED SOURCE-DOMAIN (DEVICE A) AND TARGET-DOMAIN TEST DATA (DEVICES B, C) OBTAINED BY BOTH THE DCASE AND THE KAGGLE MODEL. NOTE THAT TWO OF THE THREE VARIANTS OF THE METHOD PRESENTED IN [16] ARE NOT DIRECTLY COMPARABLE WITH OUR APPROACH AND THEREFORE THEY APPEAR HERE IN BRACKETS (SECOND AND THIRD ROW OF THE TABLE).

	DCASE Model				**Kaggle Model**			
	Non adapted		**Adapted**		**Non adapted**		**Adapted**	
	Device A	Devices B,C	Device A	Devices B,C	Device A	Devices B,C	Device A	Devices B,C
[16] $\Delta\mu_2$ derived from Device A	—	—	0.58	0.47				
[16] $\Delta\mu_2$ derived from Devices B, C	—	—	(0.58)	(0.51)				
[16] Universal domain	—	—	(0.58)	(0.50)				
[12]					0.65	0.20	0.65	0.32
[13]					0.65	0.21	0.64	0.45
Proposed method:								
DIA (Device-Independent Adaptation)				0.48				0.45
DDA (Device-Dependent Adaptation)	**0.66**	0.22	0.64	**0.53**	0.65	0.20	**0.66**	**0.51**

D. Influence of the Number of Target-Domain Test Samples

The proposed DDA and DIA methods have so far implied that $N_{\mathcal{T}}$ and $N'_{\mathcal{T}}$ target-domain mel-spectrograms would be available during the adaptation phase. However, many real-life applications cannot rely on such appropriately sized datasets. In these cases, sample statistics are likely to be unreliable in describing device-specific characteristics, especially if the sample size is small. To assess how much the DA capabilities of the proposed method are affected by the amount of available data, we simulate the scenario in which only a limited number of target-domain samples are available. For the sake of brevity, we limit ourselves to the evaluation of the Kaggle model.

To this end, random permutations of the target-domain datasets $\mathbf{X}_{\mathrm{B}}^{\mathcal{T}}$, $\mathbf{X}_{\mathrm{C}}^{\mathcal{T}}$ and $\mathbf{X}_{\mathrm{B|C}}^{\mathcal{T}}$ are partitioned into segments, each containing L mel-spectrograms. Specifically, L is defined so that it takes values in the set of divisors of the number of samples in the respective target-domain datasets, namely $N_{\mathcal{T}}$ for DDA and $N'_{\mathcal{T}}$ for DIA. Each segment is then standardized and adapted independently of the others using the statistics of $\mathbf{X}^{\mathcal{S}}$ and subsequently concatenated to form $\overline{\mathbf{X}}_{\mathrm{B}}^{\mathcal{T}}$, $\overline{\mathbf{X}}_{\mathrm{C}}^{\mathcal{T}}$ and $\overline{\mathbf{X}}_{\mathrm{B|C}}^{\mathcal{T}}$. Finally, the test sets are obtained according to (8) and (10) and the evaluation of the DDA and DIA methods can proceed as described in the previous sections. To reduce the influence of the random indexing involved in the segmentation process, we perform said procedure on 50 different permutations of each target dataset. Systems performance is then assessed by means of the average classification accuracy as a function of L. Note, however, that each class has an equal probability of being represented within a segment due to the preliminary random permutations. It is thus likely that multiple different acoustic scenes are contributing to the computation of the sample statistics, yielding a more robust estimate.

IV. RESULTS AND DISCUSSION

Table III reports the results of the proposed method against the ones of the unsupervised methods in [12], [13] and [16]. Of the three variants presented in [16], only the one that considers Device A as the reference domain used to train the FHVAE and to compute the channel conversion parameter $\Delta\mu_2$ is directly comparable with our method. Indeed, this is the only one not relying on external or target-domain data during the pre-training phase. For completeness, however, the results of the variants which do include such data are reported in brackets.

As can be seen in Table III, our DDA method provides a classification accuracy of 53% (DCASE model) and 51% (Kaggle model) when evaluated on the test fold of Devices B and C. This corresponds to an increase of approximately 6% in target-domain accuracy compared to [16] (47%) and [13] (45%). For what concerns DIA, instead, we can notice that the performance matches the ones of much more complex systems based on FHVAE [16] and WGAN [13]. In particular, the accuracy obtained by DIA is 48% (DCASE model) and 45% (Kaggle model) against the 47% of [16] and 45% of [13]. Moreover, when evaluated on Device A, the non-adapted procedure yields an accuracy of 66% (DCASE model) and 65% (Kaggle model), while the adapted one yields 64% and 66%, i.e., −2% and +1%, respectively. This slight mismatch is probably due to the sample statistics of training and test data being different and appears to be related to the number of mel bands chosen for the feature representation.

In view of the results, it seems that to apply such a band-wise preprocessing procedure across the datasets is beneficial when dealing with mismatched recording devices. A possible interpretation is that, while classic standardization approaches aim at balancing the weights of the features that describe each sample, our method is focusing more on device-specific characteristics which are constant throughout the dataset, rather than on the acoustic content of individual scenes. The result, in practice, is that of an equalization of the frequency response of the recording devices across domains. This would also explain the lower classification accuracy when Devices B and C are combined into a single target-domain dataset before the adaptation (DIA), rather then preprocessed independently (DDA). In the DIA case, indeed, the statistics removed by the standardization are only an average of the device-specific ones and therefore both channels would maintain a larger part of their characteristic response.

In a sense, the effect described here is somewhat reminiscent of the spectral correction presented in [17], where the author finds as many coefficients as frequency bands from a reference

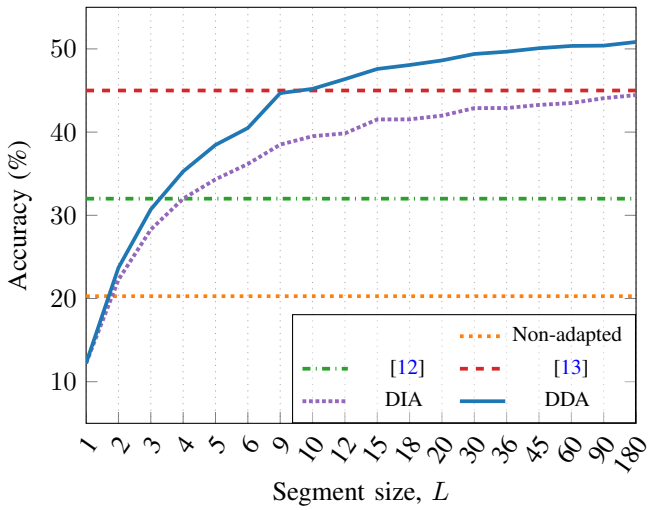

Fig. 1. Average accuracy of the Kaggle model on $\overline{\mathbf{X}}_{\text{test}}^{(\text{B,C})}$ as a function of the number of samples in each test segment. The figure depicts the results of both DDA (blue solid curve) and DIA (purple dotted curve) plotted against the performance of the Kaggle model tested on non-adapted data (orange dotted line), [12] (green dash-dotted line) and [13] (red dashed line).

device (Device A) which are then used to weight the time-frequency representations of audio data from the other devices. Notably, however, our method does not require pair-wise matchings between audio files recorded simultaneously.

As mentioned in [16], a desired quality of unsupervised adaptation strategies is not to rely on a large-sized target-domain dataset. In the following, we show that our method is capable of providing satisfactory results even when just a few target-domain samples are available. By looking at Fig. 1, we can readily observe that, as expected, classification accuracy grows monotonically with L and the maximum is achieved for a segment encompassing all possible mel-spectrograms, i.e., $L = N_{\mathcal{T}}$ (DDA) and $L = N'_{\mathcal{T}}$ (DIA). On the one hand, $L = 1$ (corresponding to trying to align every mel-spectrogram independently of the others) is worse than applying no adaptation at all. On the other hand, the proposed DDA method is already capable of outperforming [13] using segments of $L \geq 10$ samples (i.e., just over one and a half minutes of audio). This is quite remarkable as it suggests that considerable DA can be achieved without the burden of gathering an abundance of target-domain samples, making it feasible for a user to collect the data needed to adapt their own device. However, a more thorough study on the effect of under-representation of certain classes in the test segments is left for future work.

V. Conclusions and Future Work

We proposed an effective approach to unsupervised domain adaptation for acoustic scene classification. Our method, despite its simplicity, is able to outperform the unsupervised methods found in the literature while requiring just over ten test samples to provide state-of-the-art results. Moreover, we showed that our approach is competitive in terms of target-domain classification accuracy even without being given the knowledge of which target device captured which acoustic scene. The adaptation procedure is computationally efficient, is not tied to a particular target domain and does not involve any training. Therefore, our method can be readily applied at inference time to any previously optimized ASC model without the need of further adjustments. For future work, we plan to evaluate the proposed method for other audio classification tasks, such as speech and music recognition.

References

[1] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, 2018.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: An overview of DCASE 2017 challenge entries," in *Proc. of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 411–415.

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

[4] ——, "Acoustic scene classification in DCASE 2019 challenge: Closed and open set classification and data mismatch setups," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE)*, 2019, pp. 164–168.

[5] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, 2010.

[6] H. Daumé III and D. Marcu, "Domain adaptation for statistical classifiers," *J. Artif. Intell. Res.*, vol. 26, pp. 101–126, 2006.

[7] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2006, p. 120–128.

[8] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proc. of the annual meeting of the association of computational linguistics*, 2007, pp. 440–447.

[9] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. of the International Conference on Machine Learning (ICML)*, 2011, pp. 513–520.

[10] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2010, pp. 213–226.

[11] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. of the International Conference on Machine Learning (ICML)*, 2018, pp. 1989–1998.

[12] S. Gharib, K. Drossos, E. Çakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE)*, 2018, pp. 138–142.

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

[14] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7167–7176.

[15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. of the International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.

[16] S. Mun and S. Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 845–849.

[17] M. Kośmider, "Calibrating neural networks for secondary recording devices," Detection and Classification of Acoustic Scenes and Events 2019 Challenge (DCASE), Tech. Rep., 2019.

[18] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.