Domain Adversarial for Acoustic Emotion Recognition

Mohammed Abdelwahab, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE*

***Abstract*—The performance of speech emotion recognition is affected by the differences in data distributions between train (source domain) and test (target domain) sets used to build and evaluate the models. This is a common problem, as multiple studies have shown that the performance of emotional classifiers drop when they are exposed to data that does not match the distribution used to build the emotion classifiers. The difference in data distributions becomes very clear when the training and testing data come from different domains, causing a large performance gap between development and testing performance. Due to the high cost of annotating new data and the abundance of unlabeled data, it is crucial to extract as much useful information as possible from the available unlabeled data. This study looks into the use of adversarial multitask training to extract a common representation between train and test domains. The primary task is to predict emotional attribute-based descriptors for arousal, valence, or dominance. The secondary task is to learn a common representation where the train and test domains cannot be distinguished. By using a gradient reversal layer, the gradients coming from the domain classifier are used to bring the source and target domain representations closer. We show that exploiting unlabeled data consistently leads to better emotion recognition performance across all emotional dimensions. We visualize the effect of adversarial training on the feature representation across the proposed deep learning architecture. The analysis shows that the data representations for the train and test domains converge as the data is passed to deeper layers of the network. We also evaluate the difference in performance when we use a shallow neural network versus a *deep neural network* (DNN) and the effect of the number of shared layers used by the task and domain classifiers.**

***Index Terms*—Speech emotion recognition, adversarial training, unlabeled adaptation of acoustic emotional models.**

I. Introduction

IN many practical applications for speech emotion recognition systems, the testing data (target domain) is different from the labeled data used to train the models (source domain). The mismatch in data distribution leads to a performance degradation of the trained models [1]–[5]. Therefore, it is vital to develop more robust systems that are more resilient to changes in train and test conditions [6]–[9]. One approach to ensure that models perform well on the target domain is to use training data drawn from the same distribution. However, this approach can be expensive, since it requires enough data with emotional labels to build models specific to a new target domain. A more practical approach is to use available labeled data from similar domains along with unlabeled data from the target domain, creating models that generalize well to the new testing conditions without the need to annotate extra data with emotional labels. This study proposes an elegant solution for the problem of mismatch in train-test data distributions based on domain adversary training.

M. Abdelwahab and C. Busso are with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson TX, 75013 USA e-mail: Mohammed.Abdel-Wahab@utdallas.edu; busso@utdallas.edu

Manuscript received March 26, 2018; revised xxx xx, 2018.

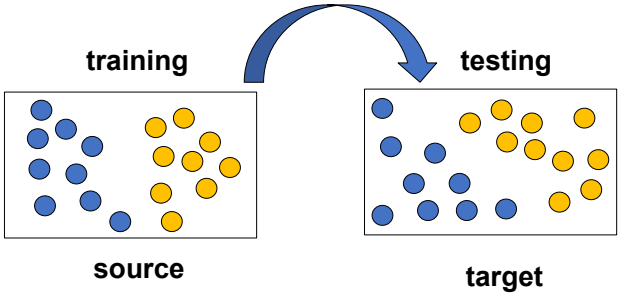


Fig. 1: Formulation of the problem of data mismatch between train (source domain) and test (target domain) datasets. Our solution uses unlabeled data from the target domain to reduce the differences between domains.

We formulate the machine-learning problem as follows. We have a *source* domain(s) with annotated emotional data, which is used to train the models, and a large *target* domain with unlabeled data (see Fig. 1). The testing data come from the target domain. Due to the prohibitive cost of annotating new data every time we change the target domain and the abundance of unlabeled data, we aim to use the unlabeled target data to extract useful information, reducing the differences between the source and target domains. The envisioned system generalizes better and is more robust by maximizing the performance using a shared data representation for the source and target domains. The key principle in our approach is to find a consistent feature representation for the source and target domains. Common approaches to address this problem include feature transformation methods, where the representation of the source data is mapped into a representation that resembles the features in the target domain [10], and finding a common representation between the domains, such that the features are invariant across domains [11]–[13]. The common domain-invariant features do not necessarily contain useful information about the main task. Therefore, it is vital to constrain the learned common feature representation to ensure that it is discriminative with respect to the main task.

This paper explores the idea of finding a common representation between the source and target domains, such that data from the domains become indistinguishable while maintaining the critical information used in emotion recognition. This work is inspired by the work of Ganin et al. [14],

which proposed training an adversarial multitask network. The approach searches the feature space for a representation that accurately describes the data from either domain, containing the relevant information to accurately classify the main task, in our study, the prediction of emotional attributes (arousal, valence, and dominance). The discriminative and domain-invariant features are learned by aligning the data distributions from the domains through back-propagation. This approach allows our framework to use unlabeled target data to learn a flexible representation.

This paper shows that adversarial training using unlabeled training data benefits emotion recognition. By using the abundant unlabeled target data, we gain on average 27.3% relative improvement in *concordance correlation coefficient* (CCC) compared to just training with the source data. We evaluate the effect of adversarial training by visualizing the similarity of the data representation learned by the network from both domains. The visualization shows that adversarial training aligns the data distributions as expected, reducing the gap between the source and target domains. The study also shows the effect of the number of shared layers between the domain classifier and the emotion predictor on the performance of the system. The size of the source domain is an important factor that determines the optimal number of shared layers in the network. This novel framework for emotion recognition provides an appealing approach to effectively leverage unlabeled data from a new domain, generalizing the models and improving the classification performance.

This paper is organized as follows. Section II discusses previous work on speech emotion recognition, with emphasis on frameworks that aim to reduce the mismatch between the train and test datasets. Section III presents our proposed model, providing the motivation and the details of the proposed framework. Section IV presents the experiment evaluation including the databases, network structure and acoustic features. Section V presents the experimental results and the analysis of the main findings. Section VI finalizes the study with conclusions and future research directions.

II. Related Work

The key challenge in speech emotion recognition is to build classifiers that perform well under various conditions. The mismatches in speech emotion recognition include speaker variability in expressing affect, the emotional context of a given interaction, the channel variability, and the acoustic differences in the environment (e.g., noise, reverberation) [1]. Some of these challenges are unique to speech emotion recognition, since they are not observed on other speech-based tasks such as *automatic speech recognition* (ASR). For example, while the ground truth can (almost always) be objectively defined in ASR, this is not the case in speech emotion recognition. The emotional labels to train models come from perceptual evaluations, which are often noisy, reflecting the differences in emotional perception of the evaluators [15]. It is a very subjective problem as different people can perceive different emotional content after listening to the same audio. Speech emotion recognition systems have to deal with larger speaker variability, beyond acoustic differences associated with variation in the feature space. We express emotional differently. One speaker may raise his/her fundamental frequency while being angry, while another may speak faster. Some speakers may partially mask their emotions due to social conventions. Furthermore, contextual information plays a key role in assessing the emotional content of a speech sample [16]. For example, the emotional content in a political debate is quite different from the emotional content in a colloquial conversation between friends. All these challenges impact the performance of speech emotion recognition systems when they are tested on new domains (different speakers, different content, different channels). A clear example of this problem is the performance of cross-corpus evaluation (training in one corpus, testing on another corpus). The cross-corpora evaluation in Shami and Verhelst [17] demonstrated the drop in classification performance observed when training on one emotional corpus and testing on another. Other studies have shown similar results [5], [18]–[22]. It is important to address the drop in performance in the presence of these mismatches.

Several approaches have been proposed to solve this problem. Shami and Verhelst [17] proposed to include more variability in the training data by merging emotional databases. They demonstrated that it is possible to achieve classification performance comparable to within-corpus results. More recently, Chang and Scherer [23] showed that data from other domains can improve the within-corpus performance of a neural network. They employed *deep convolutional generative adversarial network* (DCGAN) to extract and learn useful feature representation from unlabeled data from a different domain. This led to better generalization compared to a model that did not make use of unlabeled data.

The main approach to attenuate the mismatch between train and test conditions is to minimize the differences in the feature space between both domains. Zhang et al. [24] showed that by separately normalizing the features of each corpus, it is possible to minimize cross-corpus variability. Hassan et al. [25] increased the weight of the train data that matches the test data distribution by using *kernel mean matching* (KMM), *Kullback-Leibler importance estimation procedure* (KLEIP), and *unconstrained least-squares importance fitting* (uLSIF). Zong et al. [8] used *least square regression* (LSR) to mitigate the projected mean and covariance differences between the source data and unlabeled target samples while learning the regression coefficient matrix. Because the learned coefficient matrix depends on the samples selected from the target domain, multiple matrices were estimated and used to test new samples. Each matrix was used to predict an emotional label for a test sample, combining the results with the majority vote rule.

Studies have also explored mapping both train and test domains to a common space, where the feature representation is more robust to the variations between the domains. Deng et al. [11] used auto-encoders to find a common feature representation across the domains. They trained an auto-encoder such that it minimized the reconstruction error on both domains. Building upon this work, Mao et al. [26] proposed

to learn a shared feature representation across domains by constraining their model to share the class priors across domains. Sagha et al. [27], also motivated by the work of Deng et al. [11], used *principal component analysis* (PCA) along with *kernel canonical correlation analysis* (KCCA) to find views with the highest correlation between the source and target corpora. First, they used PCA to represent the feature space of the source and target data. Then, the features for the source and target domains were projected using the PCA in both domains. Finally, they used KCCA to select the top N dimensions that maximized the correlation between the views. Inspired by universum learning where unlabeled data is used to regularize the training process for *support vector machine* (SVM), Deng et al. [28] proposed adding an universum loss to the reconstruction loss of an auto-encoder. The added loss function was defined as the addition of the L_2-margin loss and the ϵ-insensitive loss, making use of both labeled and unlabeled data. This approach aimed to learn an auto-encoder classifier that has low reconstruction and classification errors on both domains.

Song et al. [29] proposed a couple of methods based on *non-negative matrix factorization* (NMF) that utilize data from both train and test domains. The proposed methods aimed to represent a matrix formed by data from both domains as two non-negative matrices whose product is an approximation of the original matrix. The factorization was regularized by *maximum mean discrepancy* (MMD) to ensure that the differences in the feature distributions of the two corpora were minimized. The proposed methods aim to learn a robust low dimensional feature representation using either unlabeled data or labels as hard constraints on the problem. Abdelwahab and Busso [6] proposed creating an ensemble of classifiers, where each classifier focuses on a different feature space (each classifier maximized the performance for a given emotion category). The features were selected over the labeled data from the target domain obtained with active learning. This semi-supervised approach learned discriminant features for the target domain, increasing the robustness against shifts in the data distributions between domains.

Instead of finding a common representation between domains, Deng et al. [30] trained a sparse auto-encoder on the target data and used it to reconstruct the source data. This approach used feature transformation in a way that exploits the underlying structure in emotional speech learned from the target data. Deng et al. [31] used two denoising auto-encoders. The first auto-encoder was trained on the target data and the second auto-encoder was trained on the source data, but it was constrained to be close to the first auto-encoder. The second auto-encoder was then used to reconstruct the data from both the source and target domains.

Our proposed approach takes an innovative formulation with respect to previous work relying on *domain adversarial neural network* (DANN) [14]. Liao et al. [32] used domain adversarial training in speech enhancement to learn noise invariant features. Meng et al. [33] used multi-task adversarial training to learn speaker invariant senone for ASR. They showed that adversarial training was able to align the feature representations of different speakers. Similarly, Wand and Schmidhuber [34] used DANN to learn speaker invariant features for lipreading achieving an average relative improvement of 40%. Wang et al. [35] showed that domain adversarial training outperforms state of the art domain adaptation methods for speaker recognition. Shinohara [36] showed that the use of DANN can increase the robustness of an ASR system against certain types of noise. Sun et al. [37] showed that domain adversarial training helps the model to learn features that are stable against accented speech. While these studies have showed that adversarial training can help speech based tasks, this framework has not been used with domains that have multiple forms of mismatches such as speech emotion recognition (e.g., speaker, acoustic conditions and emotional content). This is an important contribution as this framework can reduce the mismatch between train and test sets in a principled way. DANN relies on adversarial training for domain adaptation to learn a flexible representation during the training of the emotion classifier. As the training data changes, both the emotion and domain classifiers readjust their weights to find the new representation that satisfies all conditions. The domain classifier can be considered as a regularizer that prevents the main classifier from over-fitting to the source domain. The final learned representation performs well on the target domain without sacrificing the performance on the source domain.

III. Proposed Approach

A. Motivation

We present an unsupervised approach that reduces the mismatch between source and target domains by creating a discriminative feature representation that leverages unlabeled data from the target domain.

We aim to learn a common representation between the source and target domains, where samples from both domains are indistinguishable to a domain classifier. This approach is useful because all the knowledge learned while training the classifier on the source domain is directly applicable to the target domain data. We learn the representation by using a *gradient reversal layer* (GRL) where the gradient produced by the domain classifier is multiplied by a negative value when it is propagated back to the shared layers. Changing the sign of the gradient causes the feature representation of the samples from both domains to converge, reducing the gap between domains. Ideally, the performance of the domain classifiers should be at random level where both domains "look" the same. When such a representation is learned, the data distributions of both domains are aligned. This approach leads to a large performance improvement in the target domain, as demonstrated by our experimental evaluation (see Section V). A key feature of this framework is that it is unsupervised, since it does not require labeled data from the target domain. However, this framework would continue to be useful when labeled data from the target domain is available, working as a semi-supervised approach.

B. Domain Adversarial Neural Network for Emotion Recognition

Ganin et al. [14], inspired by the recent work on *generative adversarial networks* (GAN) [38], proposed the *domain*

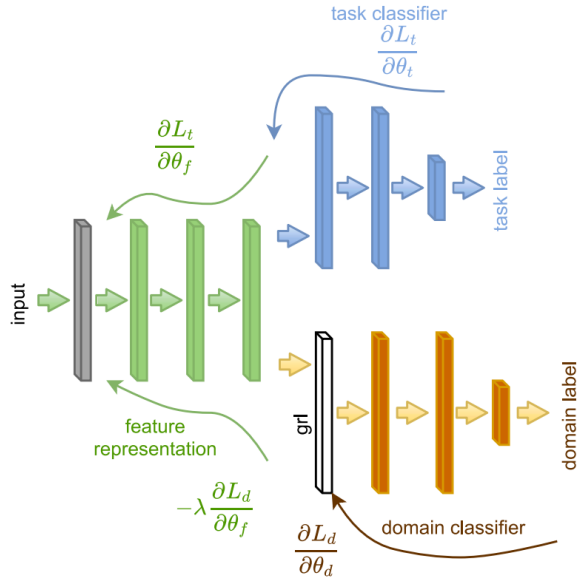


Fig. 2: Architecture of the *domain adversarial neural network* (DANN) proposed for emotion recognition. The network has three parts: a feature representation common to both tasks, a task classifier layer, and a domain classifier layer.

adversarial neural network (DANN). The network is trained using labeled data from the source domain and unlabeled data from the target domain. The network learns two classifiers: the main classification task, and the domain classifier, which determines whether the input sample is from the source or target domains. Both classifiers share the first few layers that determine the representation of the data used for classification. The approach introduced a GRL between the domain classifier and the feature representation layers. The layer passes the data during forward propagation and inverts the sign of the gradient during backward propagation. The network attempts to minimize the task and domain classification errors. By inverting the gradient coming from the domain classifier, the network learns a representation that maximizes the error of the domain classifier. By considering these two goals, the model ensures a discriminative representation for the main recognition task that makes the samples from either domain indistinguishable.

Figure 2 shows an example of the DANN structure. The network is fed labeled source data and unlabeled target data in equal proportions. In our formulation, we propose to predict emotional attribute descriptors as the primary goal. We train the primary recognition task with the source data, for which we have emotional labels. For the domain classifier, we train the classifier with data from the source and target domains. Notice that the domain classifier does not require emotional labels, so we can rely on unlabeled data from the target domain. The classifiers are trained in parallel. The network's objective is defined as:

$$E(\theta_f,\theta_y,\theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f,\theta_y) - \lambda\Big(\frac{1}{n}\sum_{i=1}^{n} L_d^i(\theta_f,\theta_d) + \frac{1}{m}\sum_{i=1}^{m} L_d^i(\theta_f,\theta_d)\Big) \quad (1)$$

where θ_f represents the parameters of the shared layers providing the regularized feature representation, θ_y represents the parameters of the layers associated with the main prediction task, and θ_d represents the parameters of the layers of the domain classifier (n is the number of labeled training samples, m is the number of unlabeled training samples). The optimization process consists of two loss functions. L_y^i is the prediction loss for the main task, and L_d^i is the domain classification loss. The prediction loss and the domain classification loss compete against each other in an adversarial manner. The parameter λ is a regularization multiplier that controls the tradeoff between the losses. This is a minimax problem. It attempts to find a saddle point parametrized by $\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_d$,

$$(\hat{\theta}_f,\hat{\theta}_y) = \arg\min_{\theta_f,\theta_y} E(\theta_f,\theta_y,\hat{\theta}_d) \quad (2)$$

$$\hat{\theta}_d = \arg\max_{\theta_d} E(\hat{\theta}_f,\hat{\theta}_y,\theta_d) \quad (3)$$

At the saddle point, the classification loss on the source domain is minimized and the domain classification loss is maximized. The maximization is achieved by introducing a GRL that changes the sign of the gradient going from the domain classification layers to the feature representation layers (see the white layer in Fig. 2). The updates taken on the feature representation parameters are in opposite direction to the gradient. With this approach, stochastic gradient descent tries to make the features similar across domains, so what is learned from the source domain remains effective for the target domain without loss in performance.

The simple concept of choosing a representation that confuses a competent domain classifier leads to models that perform better in the target domain without impacting the performance in the source domain. This approach is particularly useful for emotion recognition, as most annotated corpora come from studio settings that greatly differs from real-world testing conditions. This unsupervised approach uses available unlabeled data to align the distributions of both domains. The aligned distributions lead to a common representation, causing the domain classifier's performance to drop to random chance levels. The common indistinguishable representation retains discriminative information learned during the training of the models with data from the source domain. We improve the classifier's performance on the target domain without having to collect new annotated data. This is an important contribution in this field, taking us one step closer toward robust speech emotion classifiers that generalize well in most testing conditions.

IV. Experimental Evaluation

We define the main task as a regression problem to estimate the emotional content conveyed in speech described by

the emotional attributes arousal (calm versus activated), valence (negative versus positive), and dominance (weak versus strong). This section describes the databases (Section IV-A), the acoustic features (Section IV-B) and the specific network structures (Section IV-C) used in the experimental evaluation.

A. Emotional Databases

The experimental evaluation considers a multi-corpus setting with three databases. The source domain (train set) corresponds to two databases: the USC-IEMOCAP [39] and MSP-IMPROV [40] corpora. The target domain corresponds to the MSP-Podcast [41] database.

1) The USC-IEMOCAP Corpus: The USC-IEMOCAP database is an audiovisual corpus recorded from ten actors during dyadic interactions [39]. It has approximately 12 hours of recordings with detailed motion capture information carefully synchronized with audio (this study only uses the audio). The goal of the data collection was to elicit natural emotions within a controlled setting. This goal was achieved with two elicitation frameworks: emotional scripts, and improvisation of hypothetical scenarios. These approaches allowed the actors to express spontaneous emotional behaviors driven by the context, as opposed to read speech displaying prototypical emotions [42]. Several dyadic interactions were recorded and manually segmented into turns. Each turn was annotated with emotional labels by at least two evaluators across emotional attributes (valence, arousal, dominance). Dimensional attributes take integer values that range from one to five. The dimensional attribute of an utterance is the average of the values given by the annotators. We linearly map the scores between -3 and 3.

2) The MSP-IMPROV Corpus: The MSP-IMPROV database is a multimodal emotional database recorded from actors interacting in dyadic sessions [40]. The recordings were carefully designed to promote natural emotional behaviors, while maintaining control over lexical and emotional contents. The corpus relies on a novel elicitation scheme, where two actors improvised scenarios that lead one of them to utter target sentences. For each of these target sentences, four emotional scenarios were created to contextualize the sentence to elicit happy, angry, sad and neutral reactions, respectively. The approach allows the actor to express emotions as dictated by the scenarios, avoiding prototypical reactions that are characteristic of other acted emotional corpus. Busso et al. [40] showed that the target sentences occurring within these improvised dyadic interactions were perceived as more natural than the read renditions of the same sentences. The MSP-IMPROV corpus includes not only the target sentences, but also other sentences during the improvisations that led one of the actors to utter the target sentence. It also includes the natural interactions between the actors during the breaks.

The corpus consists of 8,438 turns of emotional sentences recorded from 12 actors (over 9 hours). The sessions were manually segmented into speaking turns, which were annotated with emotional labels using perceptual evaluations conducted with crowdsourcing [43]. Each turn was annotated by at least five evaluators, who annotated the emotional content in terms of the dimensional attributes arousal, valence, and dominance. Dimensional attributes take integer values that range from one to five. The consensus label assigned to each speech turn is the average value of the scores provided by the evaluators, which we linearly map between -3 and 3.

3) The MSP-Podcast Corpus: The *MSP-Podcast* corpus is an extensive collection of natural speech from multiple speakers appearing in Creative Commons licensed recordings downloaded from audio-sharing websites [41]. Some of the key aspects of the corpus are the different conditions in which the recordings are collected, a large number of speakers, and a large variety of natural content from spontaneous conversations conveying emotional behaviors. The audio was preprocessed removing portions that contain noise, music or overlapped speech. The recordings were then segmented into speaking turns creating a big audio repository with sentences that are between 2.75 seconds and 11 seconds. Emotional models trained with existing databases are then used to retrieve speech turns with target emotional content. The candidate segments were annotated with emotional labels using an improved version of the crowdsourcing framework proposed by Burmania et al. [43].

Each speech segment was annotated by at least five raters, who provided scores for the emotional dimensions arousal, valence and dominance using seven-Likert scales. The consensus scores are the average scores assigned by the evaluators, which are shifted such that they are in the range between -3 and 3. The collection of this corpus is an ongoing effort. This study uses 14,227 labeled sentences. From this set, we use 4,283 labeled sentences coming from 50 speakers as our test set, which is consistently used across conditions. For the within corpus evaluation (i.e., training and testing in the same domain), we define a development set with 1,860 sentences from 10 speakers, and a train set with the remainder of the corpus (8,084 sentences). This study also uses 73,209 unlabeled sentences from the audio repository of segmented speech turns. The unlabeled segments are used to train the domain classifier in the DANN approach.

B. Acoustic Features

The acoustic features correspond to the set proposed for the INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) [44]. This feature set includes 6,373 acoustic features extracted at the sentence level (one feature vector per sentence). First, it extracts 65 frame-by-frame *low-level descriptors* (LLDs) which includes various acoustic characteristics such as *Mel-frequency cepstral coefficients* (MFCCs), fundamental frequency, and energy. The externalization of emotion is conveyed through different aspects of speech production so including these LLDs is important to capture emotional cues. After estimating LLDs, a list of functions are estimated for each LLD, which are referred to as *high-level descriptors* (HLD) features. These HLDs include standard deviation, minimum, maximum, and ranges. The acoustic features are extracted using OpenSMILE [45].

We separately normalize the features from each domain (i.e., corpus) to have zero mean and a unit standard deviation. The mean and the variance of the data is calculated

considering only the values of the features within the 5% and 95% quantiles to avoid outliers skewing the values. After normalization, we ignore any value greater than 10 times its standard deviation by setting their values to zero.

C. Network Structure

As discussed in Section III-B, the DANN approach has three main components: the domain classifier layers, task classifier layers, and feature representation layers. The *domain classifier layers* are implemented with two layers across all the experimental evaluation. The *task classifier layers* are also implemented with two layers, except for the shallow network described below. The number of layers in the *feature representation layers* is a parameter of the network that is set on the development set. We consider different number of shared layers, evaluating the performance of the system with one, two, three or four layers. We fix the number of nodes per layers to 256 nodes. This setting provides good performance on the source domain. This parameter is consistently implemented in all the structures evaluated in this paper.

We also study whether a simple shallow network can achieve similar performance compared to the deep network. In the shallow network, the *task classifier layer* and the *feature representation layer* are implemented with one layer each. We implement the *domain classifier layer* with two layers.

D. Baseline Systems

We establish two baselines. The first baseline is a network trained and validated only on the source data. This condition creates a mismatch between the train and test conditions. The second baseline corresponds to within corpus evaluations, where the models are trained and tested with data from the target domain. This model assumes that training data from the target domain is available, so it corresponds to the ideal condition. The parameters of the networks are optimized using the development set. The baselines are implemented with similar architectures, serving as a fair comparison with the proposed method (e.g., number of layers, number of nodes). The key difference with the DANN models is the lack of the *domain classification layers*, where the feature representation layers only consider the primary classification task.

E. Implementation

We train the networks using Keras [46] with Tensorflow as back-end [47]. We use batch normalization and dropout. The dropout rates are p=0.2 for the input layer and p=0.5 for the rest of the layers. We further regularize the models using max-norm on the weights of value four and a clip norm on the gradient of value ten. The loss function used for the main regression task is the *mean square error* (MSE). The loss function for the domain classification task is the categorical cross-entropy. We use Adam as an optimizer with a learning rate of $5e^{-4}$ [48]. We train the models for 100 epochs with a batch size of 256 sentences. A parameter of the DANN model is λ in Equation 1, which controls the trade-off between the task and domain classification losses. We follow a similar approach to the one proposed by Ganin et al. [14], where λ is initialized equal to zero for the first ten training epochs. Then, we slowly increase its value until reaching $\lambda = 1$ by the end of the training. We train each model twenty times to reduce the effect of initialization on the performance of the classifiers. We report the average performance across the trials.

In adversarial training, we need unlabeled data to train the domain classifier. We randomly select samples from the unlabeled portion of the target domain to be fed to the domain classifier. The number of selected samples from the audio repository of unlabeled speech turns is equal to the number of samples in the source domain, keeping the balance in the training set of the domain classification task.

V. Experimental Results

The performance results for the baseline models and the DANN models are reported in terms of the *root mean square error* (RMSE), *Pearson's correlation coefficient* (PR), and *concordance correlation coefficient* (CCC) between the ground-truth labels and the estimated values. While we presented PR and RMSE, the analysis focus on CCC as the performance metric, which combines *mean square error* (MSE) and PR. CCC is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4}$$

where ρ is the Pearson's correlation coefficient, σ_x and σ_y, and μ_x and μ_y are the standard deviations and means of the predicted score x and the ground truth label y, respectively.

A. Number of Layers For the Shared Feature Representation

We first study the effect of the number of shared layers between the domain classifier and the primary regression task on the DANN model's performance (e.g., *feature representation layers* in Fig. 2). The domain and task classifier layers are implemented with two layers each. We vary the number of shared layers between the classifiers and observe how the changes in feature representation affect the regression performance. This evaluation is conducted exclusively on the development set of the target domain (e.g., MSP-Podcast corpus).

Table I shows the average RMSE, PR, and CCC for the models trained with one, two, three and four shared layers. For arousal, we consistently observe better performance (lower RMSE and higher PR and CCC) with one shared layer between the domain and task classifiers. For the CCC values, the differences are statistically significant for all the cases, except when the source domain is the MSP-IMPROV corpus and the DANN model is implemented with two layers (one-tailed t-test over the average across the twenty trials, asserting significance if p-value < 0.05). The performance degrades as more shared layers are added. For valence and dominance, the number of shared layers that provides the best performance varies from one corpus to another. In most cases, two or three shared layers provide the best performance. Based on these results on the development set, we set the number of shared layers for the feature representation networks to one for arousal, two for valence and three for dominance.

TABLE I: Performance of DANN framework implemented with different number of shared layers for the domain and task classifiers [*Iem*: USC-IEMOCAP corpus, *Imp*: MSP-IMPROV corpus, *All*: All corpora combined]

		Arousal			Valence			Dominance		
src	n	RMSE	PR	CCC	RMSE	PR	CCC	RMSE	PR	CCC
Iem	1	**1.26**	**.412**	**.365**	**1.57**	.140	.135	1.30	.151	.124
	2	1.29	.379	.332	1.64	**.152**	**.144**	1.27	.160	.135
	3	1.32	.353	.305	1.67	.150	.140	**1.27**	**.217**	**.181**
	4	1.31	.347	.296	1.67	.147	.135	1.27	.192	.161
Imp	1	1.62	.497	.284	1.58	.171	.137	**0.80**	.485	.448
	2	**1.57**	**.506**	**.303**	1.47	.203	.176	0.82	**.514**	**.476**
	3	1.71	.399	.218	1.48	.202	.173	0.87	.478	.403
	4	1.73	.382	.202	**1.43**	**.210**	**.193**	0.85	.442	.383
All	1	**1.37**	**.432**	**.341**	1.56	.181	.165	**1.01**	**.395**	**.356**
	2	1.41	.396	.313	**1.56**	.173	.160	1.05	.345	.309
	3	1.46	.369	.286	1.58	**.183**	**.170**	1.06	.339	.308
	4	1.45	.371	.282	1.56	.161	.149	1.06	.307	.270

B. Regression Performance of the DANN Model

This section presents the regression results achieved by the DANN model, and the baseline models. Table II lists the performance for the within-corpus evaluation where the models are trained and tested with data from the MSP-Podcast corpus (referred to as *target* in the table), and the cross-corpus evaluations, where the models are trained with other corpora (referred to as *src* on the table). These baseline results are compared with our proposed DANN method (referred to as *dann* on the table). The results are tabulated in terms of emotional dimensions and networks structures. These values are the average of twenty trials to account for different initializations and set of unlabeled data used to train the DANN models. For the rows denoted "All Databases", we combine all the source domain together (IEMOCAP and MSP-IMPROV corpora), treating them as a single source domain. To determine statistical differences between the *src* and *dann* conditions, we use the one-tailed t-test over the twenty trials, asserting significance if p-value < 0.05. We highlight in bold when the difference between these conditions is statistically significant.

To visualize the results in Table II, we create figures showing the average performance under different conditions (Figs. 3-6). We use statistical significance tests to compare different conditions (values obtained from Table II). We test the hypothesis that population means for matched conditions are different using one-tailed z-test. We assert significance if p-value < 0.05. We use an asterisk in the figures to indicate if there is a statistically significant difference relative to the baseline model trained with the source domain.

Figure 3 shows the average concordance correlation coefficient across emotional dimensions, training sources, trials, and networks structures (three emotional dimensions $\times$ twenty trials $\times$ three sources $\times$ five structures = 900 matched conditions). The five structures correspond to four deep networks with a varying number of shared layers (one, two, three and four), and the shallow network. The average performance for the within-corpus evaluation (target) is close to double the performance for the cross-corpus evaluations (source). This result demonstrates the importance of minimizing the mismatch between train and test conditions in emotion recognition. The figure shows that the proposed DANN approach greatly improves the performance, achieving 6.6% gain compared to the systems trained with the source domains. As highlighted by the asterisk, the improvement is statistically significant. The proposed approach reduces the gap between within-corpus and cross-corpus emotion recognition results, effectively using unlabeled data from the target domain.

Figure 4 shows the average concordance correlation coefficient for each emotional dimension (twenty trials $\times$ three sources $\times$ five structures = 300 matched conditions). On

TABLE II: Performance of the proposed DANN approach across different structures and emotional dimensions.

Training Data		Arousal					
Source	Approach	deep			shallow		
		RMSE	PR	CCC	RMSE	PR	CCC
MSP-Podcast	target	0.67	.786	.776	0.66	.787	.767
USC-IEMOCAP	src	**1.01**	**.515**	.452	**1.00**	.510	.434
	dann	1.10	.503	**.489**	1.07	**.520**	**.503**
MSP-IMPROV	src	1.57	.551	.267	1.53	.555	.263
	dann	**1.48**	**.607**	**.381**	**1.46**	**.614**	**.381**
All Databases	src	1.20	.496	.386	1.18	.506	.378
	dann	**1.18**	**.555**	**.499**	1.18	**.551**	**.486**

Training Data		Valence					
Source	Approach	deep			shallow		
		RMSE	PR	CCC	RMSE	PR	CCC
MSP-Podcast	target	1.08	.327	.294	1.03	.344	.283
USC-IEMOCAP	src	**1.30**	.202	.198	**1.19**	.227	.207
	dann	1.44	**.218**	**.215**	1.34	**.267**	**.255**
MSP-IMPROV	src	**1.42**	.142	.122	**1.29**	.154	.127
	dann	1.43	**.163**	**.161**	1.37	**.180**	**.178**
All Databases	src	**1.33**	.214	.201	**1.16**	.245	.228
	dann	1.39	**.299**	**.294**	1.33	**.272**	**.267**

Training Data		Dominance					
Source	Approach	deep			shallow		
		RMSE	PR	CCC	RMSE	PR	CCC
MSP-Podcast	target	0.57	.718	.697	0.57	.723	.704
USC-IEMOCAP	src	**0.95**	.437	.393	**0.87**	.461	.426
	dann	1.10	.437	.401	1.03	**.497**	**.457**
MSP-IMPROV	src	**0.86**	.563	.368	0.87	.520	.353
	dann	0.89	.565	**.456**	0.88	**.578**	**.472**
All Databases	src	**0.85**	.481	.418	**0.81**	.493	.437
	dann	0.92	**.526**	**.499**	0.90	**.550**	**.516**

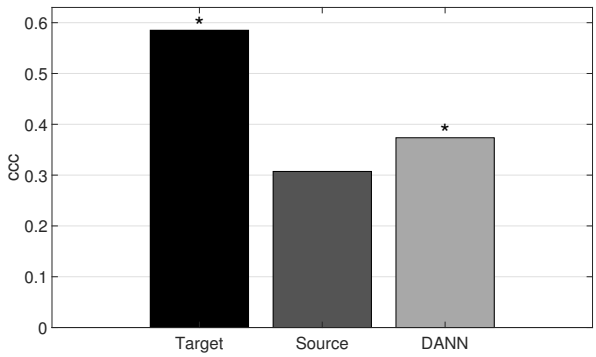


Fig. 3: Average concordance correlation coefficient across conditions. The DANN models provide significant improvements over the baseline model trained with the source domain.

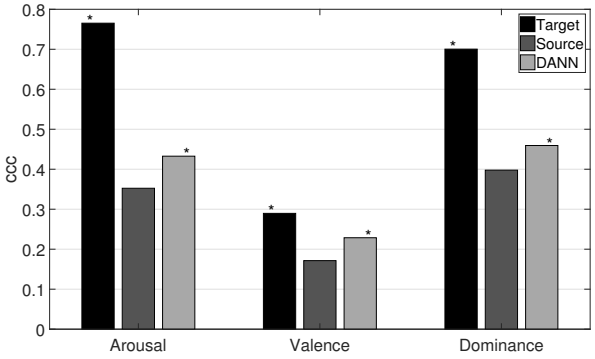


Fig. 4: Average concordance correlation coefficient across conditions for each emotional dimension.

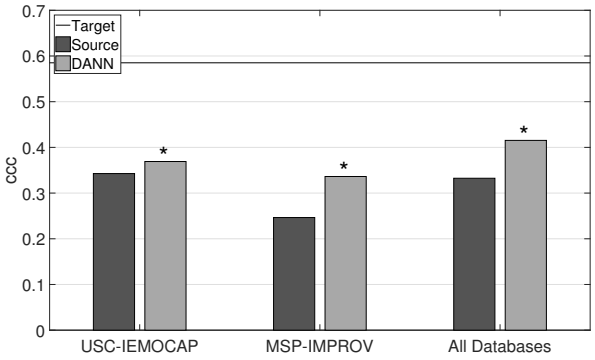


Fig. 5: Average concordance correlation coefficient across condition for each source domain. The solid line represents the within corpus performance.

average, the figure shows that models trained using DANN consistently outperform models trained with the source data. The asterisk denotes that the differences are statistically significant across all emotional dimensions. The relative improvements over the source models are 22.8% for arousal, 33.4% for valence and 15.5% for dominance. In general, the values for CCC are lower for valence, validating findings from previous studies, which indicated that acoustic features are less discriminative for this emotional attribute [49].

Figure 5 shows the average concordance correlation coefficient per source domain (three emotional dimensions $\times$ twenty trials $\times$ five structures = 300 matched conditions). The figure shows the within-corpus performance (target) with a solid horizontal line. The results consistently show significant improvements when using DANN. The relative improvements in performance over training with the source domain are 7.7% for the USC-IEMOCAP corpus, 36.4% for the MSP-IMPROV corpus, and 25% when we combine all the corpora. Figure 5 also shows that, on average, combining all the sources into one domain improves the performance of the systems in recognizing emotions. Adding variability during the training of the models is important, as also demonstrated by Shami and Verhelst [17]. DANN models also benefit from adding variability. By leveraging the added data representations, DANN models are able to find a common representation between the domains without sacrificing the critical features relevant for learning the primary task.

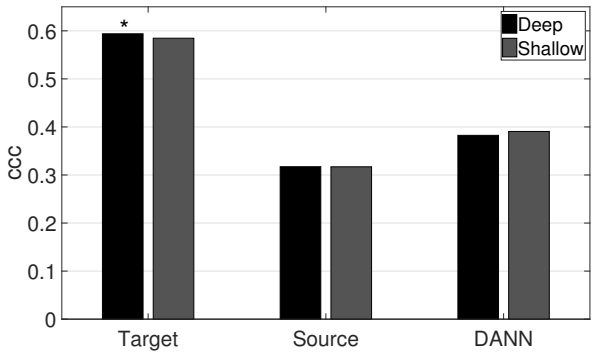
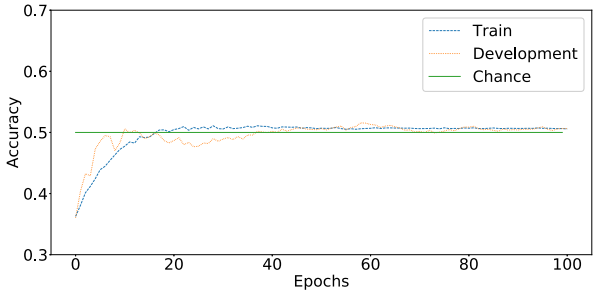
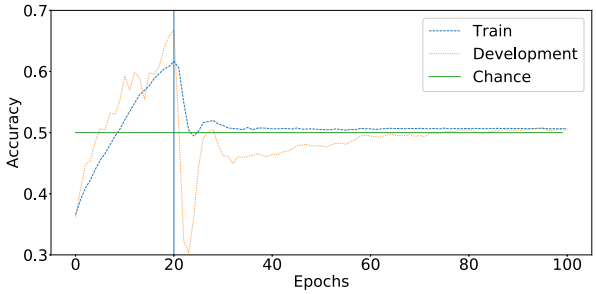


Fig. 6: Average concordance correlation coefficient across conditions for deep and shallow structures.



(a) Domain classifier's accuracy with active GRL.



(b) Domain classifier's accuracy activating the GRL after 20 epochs.

Fig. 7: Domain classifier accuracy during the training process of the DANN models (arousal model trained on the USC-IEMOCAP corpus). The figures include results for the train and development sets.

Figure 6 compares the performance for deep and shallow networks (see Section IV-C). The figure summarizes the results for the within-corpus evaluations (target), cross-corpus evaluations (source) and with our proposed DANN model (three emotional dimensions $\times$ twenty trials $\times$ three sources = 180 matched conditions). For the target models, we observe significant improvements when using deep structures over shallow structures. However, the differences are not statistically significant for the source and DANN models.

C. Performance of Domain Classifier

While the focus of this study is on the prediction of emotional attributes, studying the results on domain classi-

TABLE III: Performance of the domain classifier on the target domain. The results correspond to accuracies in detecting samples belonging to the target domain.

	Source	Target Test
Arousal	USC-IEMOCAP	71.2%
	MSP-IMPROV	70.6%
	ALL	46.8%
Valence	USC-IEMOCAP	60.2%
	MSP-IMPROV	75/0%
	ALL	54.0%
Dominance	USC-IEMOCAP	66.7%
	MSP-IMPROV	72.5%
	ALL	61.2%

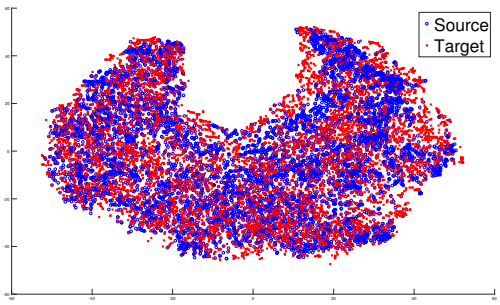
fication can lead to further insights into the use of DANN in speech emotion recognition. We consider the performance of the domain classification during training. Figure 7 shows the accuracy in detecting the source and target domains during training as a function of epochs. The figure includes the results on the train set, consisting of data from the source and target domains, and on the development set, consisting of data from the source domain. Figure 7a shows that the GRL behaves as expected, where the performance of the domain classifier converges to the chance level. We also re-implement the approach without the GRL so the domain classifier is trained to maximize its performance. After 20 epochs, we activate the GRL. Figure 7b shows the results for the train and development sets, where the vertical line indicates the time when the GRL is activated. The figure shows that the domain classifier's performance increases during the first 20 epochs. After activating the GRL, the performance drops to the chance level, as expected.

We also report the performance of the domain classifiers on the test set. Since all the samples belong to the target domain, any sample assigned to the source domain is a mistake. The results are reported in the Table III. The domain classifier recognizes data from the target domain with low confidence reaching performance below 73%. We observe that when the source domain consists of both databases (i.e., higher variability in the source domain), the domain classifier's performance is close to 50%, as expected.

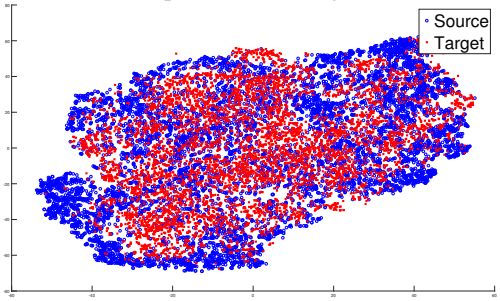
D. Data Representation

The results in Tables I and II demonstrate the benefits of using the proposed DANN framework. This section aims to understand the key aspect of the DANN approach by visualizing the feature representation created during the training process.

We use the t-SNE toolkit [50] to create 2D projections of the feature distributions at different layers of the networks. t-SNE is a popular data visualization technique to project high dimensional data into a smaller subspace. The projection provides a useful tool to visually inspect feature representations learned by the model. t-SNE is built on stochastic neighbor embedding. It attempts to minimize the *Kullback-Leibler divergence* (KLD) between the low-dimensional embedding and the high-dimensional data. This approach results in an embedding that retains the local data structure while also revealing global structures such as clusters of samples on different manifolds.



(a) Feature representation using the DANN model

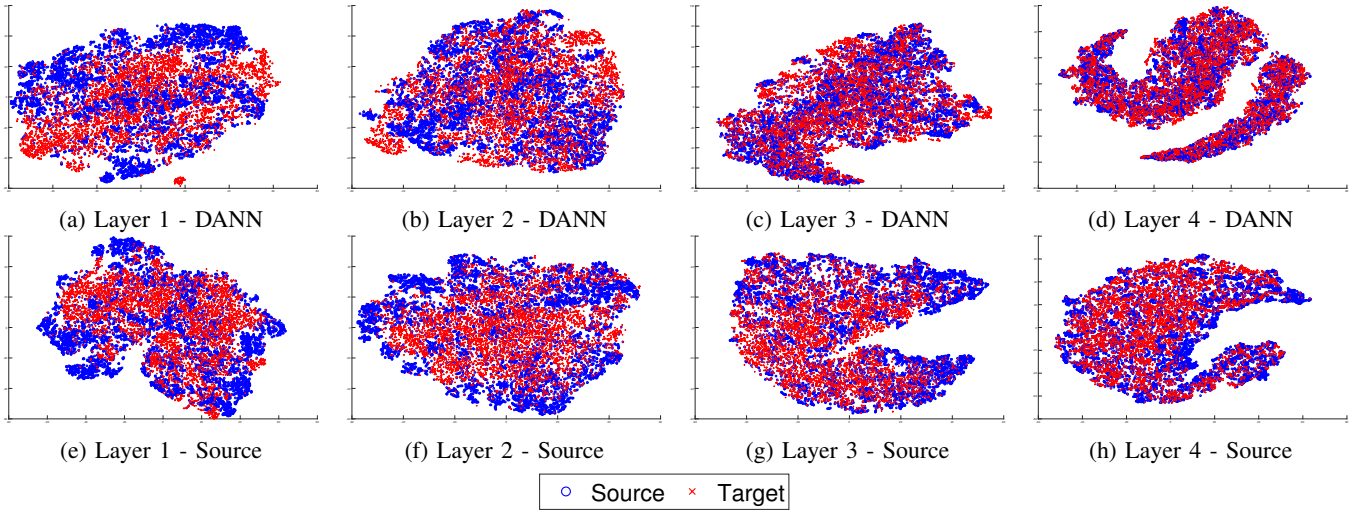


(b) Feature representation using the baseline source model

Fig. 8: Feature representation at the last shared layer for arousal models trained on USC-IEMOCAP corpus. The figures are created with the t-SNE toolkit.

Figure 8 shows the distributions of the data from the source domain (blue circles) and the target domain (red crosses) after projecting them in the feature representation created by two models. This example corresponds to the models trained for arousal using the USC-IEMOCAP corpus as the source domain (as explained in Section V-A, the DANN model for arousal has only one shared layer as a feature representation). Figure 8a shows the data representation at the output of the shared layer of the DANN model. Figure 8b shows the data representation at the equivalent layer of the DNN model trained with the source domain (i.e., the USC-IEMOCAP database). By using adversarial domain training, the feature distributions for samples from both domains are almost indistinguishable, demonstrating that the proposed approach is able to find a common representation. Without adversarial domain training, in contrast, there are large regions in the feature space where it is easy to separate samples from the source and target domains. Figure 8b suggests the presence of a source-target mismatch which affects the performance of the emotion classifier.

We also explore the feature representation when the DANN model is trained with four shared layers using the t-SNE toolkit. The objective of this evaluation is to visualize the distribution of the data in each of the shared layers. This evaluation is implemented using the models for valence, using the MSP-IMPROV corpus as the source domain. Figures 9a-9d show the changes in the data representation for each of the four shared layers in the DANN model. At the first layer (Fig. 9a), the feature representations for the source data (blue circles) and the target data (red crosses) are dissimilar enough for the domain classifier to distinguish them. While the difference



(a) Layer 1 - DANN (b) Layer 2 - DANN (c) Layer 3 - DANN (d) Layer 4 - DANN

(e) Layer 1 - Source (f) Layer 2 - Source (g) Layer 3 - Source (h) Layer 4 - Source

Fig. 9: Feature representation at each layer using the DANN model, and the baseline DNN trained with the source domain. The figures are created with the t-SNE toolkit. The example corresponds to models for valence with four shared layers, using the MSP-IMPROV corpus as the source domain. Figures (a)-(d) report the results for DANN model. Figures (e)-(h) report the results for DNN trained with the source domain.

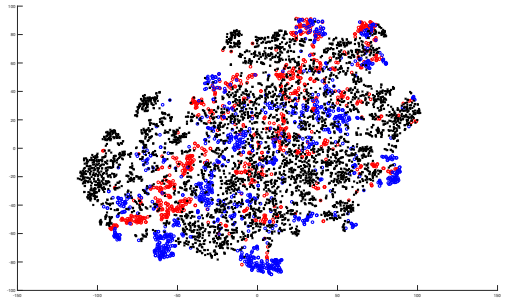
between domains in the feature representation has decreased at the second layer (Fig. 9b), there are still some regions dominated by samples from one of the two domains. At the third layer (Fig. 9c), the feature representations of the domains are similar enough to confuse the domain classifier. The common representation is maintained in the fourth layer (Fig. 9d), where the data from the target domain is indistinguishable from the data from the source domain. This final representation is used by the emotion regression system to predict the emotional state of the target speech. For comparison, we also trained a baseline model with six layers, matching the combined number of shared and task classifier layers in the DANN model. Figures 9e-9h show the feature representation of the corresponding first four layers of this model. In the DANN model, the data representation of the samples from the source and target domains become more similar in deeper layers. This trend is not observed in the model trained with only the source data. The DANN model effectively reduces the mismatch in the feature representation across domains, which leads to significant improvements in the regression models.

E. Analysis of Speaker Mismatch

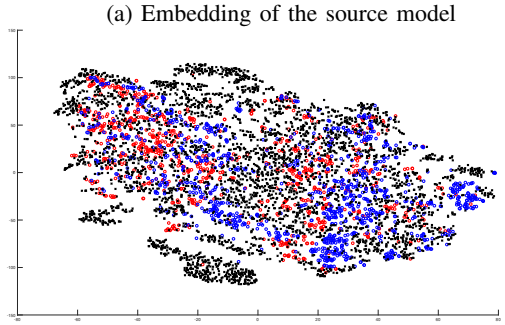
As mentioned before, speaker variability is one of the sources of mismatches in cross-corpus evaluations. The expectation is that after training the DANN model, the representation of the source data will be closer to the representation from the target data. We explore this hypothesis in Figure 10, which shows the feature embeddings of the source and DANN models. After training the models, we project the samples from the test set (target domain), which are highlighted in black. These samples illustrate the data distribution of the target domain on the corresponding embeddings. We consider the data from two speakers in the development set of the source domain (USC-IEMOCAP corpus). We project these samples in both feature representations, highlighting speech segments from the speakers in red and blue, respectively. The source and DANN models have not been trained with data from this set. Figure 10a shows that the data from the source domain (red and blue points) is clustered when projected on the embedding of the source domain. The clusters do not necessarily match the distribution of the target domain (black markers). Instead, Figure 10b shows that the source domain data is more widespread and intertwined with the target data when using the embedding of the DANN models. These figures illustrate that the proposed approach reduces the speaker mismatch in speech emotion recognition tasks.

VI. Conclusions

This study proposed an appealing framework for emotion recognition that exploits available unlabeled data from the target domain. The proposed approach relies on *domain adversarial neural network* (DANN), which creates a flexible and discriminant feature representation that reduces the gap in the feature space between the source and target domains. By using adversarial training, we learned domain invariant representations that can effectively discriminate the primary regression task. The model aims to find a balanced representation that aligns the domain distributions, while retaining crucial information for the primary regression task. The proposed adversarial training leads to significant improvements in the performance of emotion recognition classifier over models exclusively trained with data from the source domain, which was demonstrated by the experimental evaluation. We visualized the data representation of both domains by projecting the features into the shared layers of the proposed DANN model. The results showed that the model converged to a common representation, where the source and target domains became indistinguishable. The experimental evaluation also showed that the amount of labeled data from the source domain plays a small role in determining how many shared layers are needed between the domain and regression tasks. Since the number of shared layers has a strong impact on the system's performance,



(a) Embedding of the source model



(b) Embedding of the DANN model

Fig. 10: Data projection of two speakers (red and blue points) from the source domain in the embedding created by the (a) source and (b) DANN models. The figure considers arousal using the IEMOCAP corpus. The figure shows that samples are better distributed in the embedding of the DANN model.

it is vital to identify the optimal number of shared layers, given a specific source domain.

One challenging aspect in using the proposed approach is the difficulty of training adversarial networks. For example, Shinohara [36] noted that in ASR problems, the improvements of DANN were large for some types of noises, but less effective for others. They suggested that tuning the parameters could lead to better results. We also observed that the framework failed to converge for certain parameters, which is common in minimax problems. It is important to investigate robust strategies to set the hyper-parameters of the model that work in most scenarios. When properly trained, however, this powerful framework can elegantly solve one of the most important problems in speech emotion recognition: reducing the mismatch between train and test domains.

There are many research directions to extend the proposed approach. The current implementation relies on hand-crafted features. The framework can be easily extended so the acoustic features are directly learned during the training of the models using *convolutional neural networks* (CNNs) (e.g., end-to-end system). Likewise, studies have shown the benefit of jointly predicting multiple emotional attributes using *multi-task learning* (MTL) [5], [51]. We hypothesize that combining these approaches with domain adversarial training would lead to further improvements.

In the case of multiple sources, our approach seems to work well when multiple sources are combined, treating them as one. This approach forces the network to learn a representation that is common across all the source domains. We hypothesize that a better approach is to use asymmetric transformations, where the model learns multiple possible representations for the test data, creating one representation for each source. During testing, the network chooses the most useful representation for each data point. Another alternative approach is to transform the available sources to match the target domains. Finally, this unsupervised approach can be easily extended to the cases where limited labeled data from the target domain is available (semi-supervised approach), creating a flexible framework to create emotion recognition systems that can generalize across domain.

ACKNOWLEDGMENT

This study was funded by the National Science Foundation (NSF) CAREER grant IIS-1453781.

REFERENCES

[1] C. Busso, M. Bulut, and S.S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[2] Y. Kim and E. Mower Provost, "Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition," in *ACM International Conference on Multimedia (MM 2014)*, Orlando, FL, USA, November 2014, pp. 27–36.

[3] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *European Signal Processing Conference (EUSIPCO 2004)*, Vienna, Austria, September 2004, pp. 341–34.

[4] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *IEEE International Conference on Multimedia and Expo (ICME 2005)*, Amsterdam, The Netherlands, July 2005, pp. 474–477.

[5] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[6] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5000–5004.

[7] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5160–5164.

[8] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, May 2016.

[9] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 5058–5062.

[10] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[11] J. Deng, R Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014, pp. 4818–4822.

[12] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4490–4494.

[13] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *International conference on machine learning (ICML 2011)*, Bellevue, WA, USA, June-July 2011, pp. 513–520.

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, April 2016.

[15] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.

[16] G.N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 248–255.

[17] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," in *Speaker Classification II*, C. Müller, Ed., vol. 4441 of *Lecture Notes in Computer Science*, pp. 43–56. Springer-Verlag Berlin Heidelberg, Berlin, Germany, August 2007.

[18] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Fuzzy emotion recognition in natural speech dialogue," in *IEEE International Workshop on Robot and Human Interactive Communication*, Nashville, TN, USA, August 2005, pp. 317–322.

[19] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, July-Dec 2010.

[20] L. Vidrascu and L. Devillers, "Anger detection performances based on prosodic and acoustic cues in several corpora," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 13–16.

[21] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: A cross-corpora study," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 2350–2353.

[22] I. Lefter, L.J.M. Rothkrantz, P. Wiggers, and D.A. Van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *Text, Speech and Dialogue (TSD 2010)*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds., vol. 6231 of *Lecture Notes in Computer Science*, pp. 353–36. Springer Berlin Heidelberg, Brno, Czech Republic, September 2010.

[23] J. Chang and S Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2746–2750.

[24] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, Waikoloa, HI, USA, December 2011, pp. 523–528.

[25] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, July 2013.

[26] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 2608–2612.

[27] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5800–5804.

[28] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, April 2017.

[29] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, October 2016.

[30] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 511–516.

[31] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, September 2014.

[32] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *ArXiv e-prints (arXiv:1807.07501)*, July 2018.

[33] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gong, and B.-H. Juang, "Speaker-invariant training via adversarial learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5969–5973.

[34] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 3662–3666.

[35] Q. Wang, W. Rao, S. Sun, L. Xie, E.S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 4889–4893.

[36] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2369–2372.

[37] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 4854–4858.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems (NIPS 2014)*, Montreal, Canada, December 2014, vol. 27, pp. 2672–2680.

[39] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[40] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[41] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.

[42] C. Busso and S.S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[43] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[44] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[45] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[46] F. Chollet, "Keras: Deep learning library for Theano and TensorFlow," https://keras.io/, April 2017.

[47] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Symposium on Operating Systems Design and Implementation (OSDI 2016)*, Savannah, GA, USA, November 2016, pp. 265–283.

[48] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Workshop track at International Conference on Learning Representations (ICLR 2015)*, San Juan, Puerto Rico, May 2015, pp. 1–4.

[49] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.

[50] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *Journal of machine learning research*, vol. 15, no. 1, pp. 3221–3245, October 2014.

[51] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018.

Mohammed AbdelWahab (S'14) received his B.Sc. degree in electrical and electronic engineering at Ain Shams University, Cairo, Egypt in 2010, and his M.S degree in Electrical engineering from Nile university, Cairo, Egypt 2012. He is currently pursuing his Ph.D. degree in electrical engineering at the University of Texas at Dallas. His current research interest includes speech signal processing, emotion recognition, artificial intelligence and machine learning.

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests include digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, in-vehicle active safety system, and machine learning methods for multimodal processing. He was the general chair of ACII 2017. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.