Learning Transferable Features for Speech Emotion Recognition

Alison Marczewski
CS Dept., UFMG & Kunumi
Belo Horizonte, Brazil
amarczew@dcc.ufmg.br

Adriano Veloso
CS Dept., UFMG
Belo Horizonte, Brazil
adrianov@dcc.ufmg.br

Nivio Ziviani
CS Dept., UFMG & Kunumi
Belo Horizonte, Brazil
nivio@dcc.ufmg.br

ABSTRACT

Emotion recognition from speech is one of the key steps towards emotional intelligence in advanced human-machine interaction. Identifying emotions in human speech requires learning features that are robust and discriminative across diverse domains that differ in terms of language, spontaneity of speech, recording conditions, and types of emotions. This corresponds to a learning scenario in which the joint distributions of features and labels may change substantially across domains. In this paper, we propose a deep architecture that jointly exploits a convolutional network for extracting domain-shared features and a long short-term memory network for classifying emotions using domain-specific features. We use transferable features to enable model adaptation from multiple source domains, given the sparseness of speech emotion data and the fact that target domains are short of labeled data. A comprehensive cross-corpora experiment with diverse speech emotion domains reveals that transferable features provide gains ranging from 4.3% to 18.4% in speech emotion recognition. We evaluate several domain adaptation approaches, and we perform an ablation study to understand which source domains add the most to the overall recognition effectiveness for a given target domain.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**;

KEYWORDS

Affective Computing; Emotion Recognition; Deep Learning

1 INTRODUCTION

Humans are increasingly interacting with machines via speech, which is an important impetus for studying the vocal channel of emotional expression. Applications of an interface capable of assessing emotional states from human voice are numerous and diverse, including communication systems for vocally-impaired individuals, call centers, lie detection, airport security, and realistic interaction with empathy. The aim of this work is the development of models capable of recognizing people's emotions from recorded voice, also known as emotion recognition from speech.

Most emotional states involve physiological reactions, which in turn modify different aspects of the voice production process [22]. Emotions produce changes in respiration and an increase in muscle tension, which influence the vibration of the vocal folds and vocal tract shape, thus affecting the acoustic characteristics of the speech. When someone is in a state of anger, fear or joy, the sympathetic nervous system is aroused, the heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. As a result, speech is loud, fast and enunciated with strong high frequency energy. Sadness, by contrast, is associated with a low, hesitant, and lacking in energy speech [31].

While there is considerable evidence that speech features can differentiate emotional states [12, 43, 47], the way in which physiological reactions translate into speech features may vary greatly depending on specific factors such as acoustic signal conditions, speakers, spoken languages, linguistic content, and type of emotion (e.g., acted, elicited, or naturalistic) [16]. Since each possible combination of such factors may define a specific domain, emotion recognition from speech becomes particularly challenging because it is unclear which speech features are the most effective for each domain. Also, it is challenging to train an emotion recognition system exclusively for the target domain due to unavailability of sufficient labeled data which limits the exploration of the feature space. Fortunately, there are potentially shared or local invariant features that shape emotions in different domains, thus transfer learning may alleviate the data demands.

In this paper, we propose a deep architecture for speech emotion recognition composed of a convolutional neural network (CNN) and a long short-term memory network (LSTM). The main hypothesis in this work is that the blend of a CNN with a LSTM exploits both spatial and temporal information of speech features for emotion recognition. That is, while the CNN extracts spatial features of varying abstract levels, the LSTM employs contextual information in order to model how emotions evolve over time. We discuss several feature transference approaches designed to our deep architecture. Such feature transference approaches differ in terms of the choice of which layers to freeze or tune, and whether or not target domain data are used during pre-training.

We conducted rigorous experiments using six standard speech emotion datasets that correspond to different domains. Recognition models are trained using different transference approaches, and we pose the following questions:

- Which feature transference approach is the most appropriate, given factors such as the amount of labels and the discrepancy between domains?
- How effective is the blend of CNN with LSTM networks for domain adaptation?

- How effective is our recognition model compared with the state-of-the-art models for speech emotion recognition based on supervised domain adaptation?

We performed an ablation domain analysis in order to elucidate the benefits of incorporating multi-domain data into the final recognition model. We show that even small amounts of multi-domain data used for adaptation can significantly improve recognition effectiveness, while domain discrepancy poses serious issues to effective model adaptation. Also, the effectiveness of the different feature transference approaches varies greatly depending on the factors that define the target domain. We report gains that vary from 4.3% to 18.4%, depending on the target domain and feature transference approach.

2 RELATED WORK

Research on the recognition of emotional expressions in voices is of great academic interest in psychology [2], neurosciences [21, 41, 42, 44] and affective computing [12, 28, 38, 47]. A number of researchers investigated acoustic correlates of emotions from human speech. In one of the first studies [46], the authors identify parameters in the speech that reflect the emotional state of a speaker. They found that anger, fear, and sorrow situations tend to produce characteristic differences in contour of fundamental frequency, average speech spectrum, temporal characteristics, precision of articulation, and waveform regularity of successive glottal pulses.

Features There are studies on how acoustic correlates of emotions from speech are transformed into features for supervised learning algorithms. In [25, 32], the authors provide reviews on a wide range of features employed for emotion recognition from speech. In [30], the authors present an approach based on hidden semi-continuous Markov models, which are built using specific energy and pitch features. In [24], the authors employ mel frequency cepstral coefficients (MFCCs) as features for a Gaussian mixture model classifier. A similar MFCC model was proposed in [26] and features related to speaking rate are also explored to categorize the emotions. In [33], the authors propose speech prosody and related acoustic features for the recognition of emotion. Methods for emotion recognition from speech relying on long-term global prosodic features were developed. In [3], the authors describe seven acoustic and four linguistic types of features, from which they found the most important ones, and also discuss the mutual influence of acoustics and linguistics. In [35], the authors introduce string kernels as a novel solution in the field.

Data Concerns Background noise, varying recording levels, and acoustic properties of the environment, and how these issues impact speech emotion recognition systems are discussed in [17]. More serious concerns about data used for emotion recognition from speech were presented in [38], where the authors discuss issues related to the overestimation of the accuracy of emotion recognition systems, since experiments are usually performed on acted data (rather than on spontaneous data). Concerns with experiments performed on acted data were also discussed in [39]. Alternatively, more realistic acted data were recently presented in [7].

Transfer Learning and Domain Adaptation Since speech data are usually captured from different scenarios, it is often observed a

significant performance degradation due to the inherent mismatch between training and test set. Thus, domain adaptation is a relevant topic in emotion recognition from speech. In [50], the authors explore a multi-task framework in which speech or song are jointly leveraged in emotion recognition in a cross-corpus setting. In [40], the authors show that training and test data used for system development usually tend to be similar as far as recording conditions, noise overlay, language, and types of emotions are concerned. The authors conclude that a cross-corpus evaluation would provide a more realistic view of the recognition performance. In [20], the authors propose a feature transfer approach using a deep architecture called PCANet, which extracts both the domain-shared and the domain-specific latent features, leading to significant effectiveness improvements. In [27], the authors propose a two-layer network, so that the parameters within the second layer are imposed the common priors between the related classes, so that the classes with few labeled data in target domain can borrow knowledge from the related classes in source domain. In [14], the authors present a feature transfer learning method using denoising autoencoders [45] to build high order sub-spaces of the source and target corpora, where features in the source domain are transferred to the target domain by a specific neural network. Similarly, in [12], the authors employ a denoising autoencoder as a domain adaptation method. In this case, prior knowledge learned from a target set is used to regularize the training on a source set. Finally, in [1], the authors propose a supervised domain adaptation approach which can improve the speech emotion recognition performance in the presence of mismatched training and testing conditions. In [13] the authors propose feature transfer learning based on sparse autoencoders. Their approach consists of learning a representation using a single-layer autoencoder, and then applying a linear SVM using the learned representation.

Feature Learning Deep neural networks were already used for emotion recognition from speech. In [43], the authors propose a generalized discriminant analysis using deep neural networks. They show that low-dimensional features capture hidden information from the acoustic features leading to significant gains compared with typical SVMs. In [11], the authors assume a scenario where speech data are obtained from different devices and varied recording conditions. As a result, data are typically highly dissimilar in terms of acoustic signal conditions. They evaluate the use of denoising autoencoders [45] to minimize this data mismatch problem. In [18], the authors propose the use of deep neural networks to extract high level features from raw recorded voice. The network outperforms SVMs using hand-crafted features. In [23], the authors employ deep belief networks and their results suggest that learning high-order non-linear relationships using these networks is an effective approach for emotion recognition. In [51], the authors employ a feature enhancement method based on an autoencoder with LSTMs, for robust emotion recognition from speech. The enhanced features are then used by SVMs. In [19], the authors propose to learn salient features for speech emotion recognition using CNNs. The network is learned in two stages. In the first stage, unlabeled samples are used to learn local invariant features using sparse autoencoders with reconstruction penalization. In the second step, these features are used as the input to a feature extractor. In [48],

the authors introduce an approach to separate emotion-specific features from general and less discriminative ones. They employ an unsupervised feature learning framework to extract rough features. Then these rough features are further fed into a semi-supervised feature learning framework. In this phase, efforts are made to disentangle the emotion-specific features and some other features by using a novel loss function, which combines reconstruction penalty, orthogonal penalty, discriminative penalty and verification penalty.

Our Work The main differences between this work and aforementioned works are: (i) we consider diverse domain adaptation approaches using CNN and LSTM features, (ii) we perform a domain ablation analysis which reveals the relative value of different domains, (iii) we perform domain blending, that is, we not just transfer features from one domain to another, but we produce generic features using data from multiple domains simultaneously. Further, we investigated the best freezing/tuning cut-off for each target domain.

3 MULTI-DOMAIN NETWORK

The task of learning to recognize emotions from speech is defined as follows. We have as input the *training set* (referred to as $\mathcal{D}$), which consists of a set of records of the form $< a, e >$, where a is an audio sample (i.e., an emotional episode) and e is the corresponding emotion being expressed. Emotions draw their values from a discrete set of possibilities, such as sadness, fear, happiness, surprise, and anger. The training set is used to construct a model which relates features within the audio samples to the corresponding emotions. The *test set* (referred to as $\mathcal{T}$) consists of records $< a, ? >$ for which only the audio sample a is available, while the corresponding emotion e is unknown. The model learned from the training set $\mathcal{D}$ is used to produce estimations of the emotions expressed on audio samples in the test set $\mathcal{T}$.

We consider a learning scenario in which audio samples and their corresponding emotion labels are drawn from different generating distributions. For instance, some audio samples may be obtained from acted speech while other audio samples are obtained from spontaneous speech. The process that produces audio samples may also differ in terms of factors such as recording conditions, spoken language, and linguistic content. A specific combination of these factors defines a *domain*. Speech emotion recognition is a domain-specific problem, that is, a recognition model learned from one domain is likely to fail when tested against data from another domain [4]. As a result, real application systems usually require labeled data from multiple domains, guaranteeing an acceptable performance for different domains. However, each domain has a very limited amount of labels due to the high cost to create large-scale labeled datasets for domain-specific speech emotion recognition. Feature transferability is thus an appealing way to alleviate the demands for domain-specific labels. Thus, for domains that are short of labeled data transferable features enable model adaptation from multiple domains.

3.1 Network Architecture

In this section, we introduce our deep architecture. It first extracts generic features from multi-domain data (or domain-shared features) which are then used to produce domain-specific and highly

discriminative features. The architecture combines a deep hierarchical spatial feature extractor with a model that can learn to recognize and synthesize temporal dynamics of emotions, as illustrated in Figure 1. The network works by passing each audio sample through a feature transformation to produce a fixed-length vector representation.[1] After that, spatial features are computed for the audio input, and then the sequence model captures how emotions evolve over time.

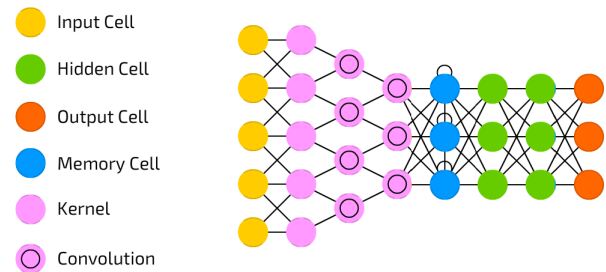**Figure 1: (Color online) Multi-Domain Network architecture for learning transferable features. Convolutional layers are followed by a LSTM layer. Different feature transference approaches are designed using this architecture.**

More specifically, the network receives a 54,000 dimensional input representing audio samples. It has five hidden layers, including two uni-dimensional convolutional layers, one LSTM layer, and two fully connected layers. The convolutional layers apply kernels with 128 dimensions, combined with ReLUs and a dropout level of 0.30. The LSTM layer receives 128 dimensional inputs, and returns two 500 dimensional vector outputs which are then flattened into a single 1,000 dimensional output. The next two fully connected layers are composed of 1,000 units and are combined with the hyperbolic tangent activation. Again, a dropout level of 0.30 is applied. The final classification layer employs a softmax cross-entropy loss and thus the minimization problem is given as:

$$\min \frac{1}{n} \sum_{i=1}^{n} J(\theta(x_i), y_i)$$

where J is the cross-entropy loss function and $\theta(x_i)$ is the conditional probability that the network assigns x_i to emotion label y_i. The network is trained by the AdaDelta method, and six emotions are considered, namely: anger, disgust, fear, happiness, sadness, and surprise. The network architecture is substantially smaller than others commonly used. We also evaluated deeper networks, but the resulting models showed to be less accurate and learning becomes significantly slower.

3.2 Feature Transferability

We assume the presence of few labeled audio samples in the target domain, hence a direct adaption to the target domain via fine-tuning is prone to overfitting. We also assume that the training

---

[1]Padding was applied so that audios with different durations have representations with the same size. Also, features are standardized so that they are centered around 0 with a standard deviation of 1.

Table 1: Summary of the datasets.

Dataset/Domain	Age	Language	Emotion	Gender	Recording	Sampling rate	# samples
AFEW	children/adults	English	natural	balanced	movies	48kHz	568
Emo-DB	adults	German	acted	balanced	studio	16kHz	287
EMOVO	adults	Italian	acted	balanced	studio	48kHz	336
eNTERFACE	adults	English	induced	unbalanced	normal	16kHz	1,047
IEMOCAP	adults	English	acted	balanced	studio	48kHz	1,770
RML	adults	many	induced	balanced	studio	22kHz	650

set is composed of audio samples belonging to different domains, and we can explicitly split $\mathcal{D}$ into n different domains, that is, $\mathcal{D} = d_1, d_2, \ldots, d_n$. Thus, the goal of our deep architecture is to train a multi-domain network to differentiate emotions based on input audios associated with multiple domains. Although audio samples associated with a given domain d_i may be better represented by specific features, there still exist some common features that permeate all other domains. Examples of such low-level features may include pitch, derivative of pitch, energy, derivative of energy, duration of speech segments, among others.

Transference Approaches The main intuition that we exploit for feature transferability is that the features must eventually transition from general to specific along our deep architecture, and feature transferability drops significantly in higher layers with increasing domain discrepancy [49]. In other words, the features computed in higher layers must depend greatly on a specific domain d_i, and recognition effectiveness suffers if d_i is discrepant from the target domain. Since we are dealing with many domains simultaneously, we also considered multiple transference approaches, which are detailed next:

- A1: no fine-tuning is performed, which means that the pre-trained model is used to recognize emotions.
- A2: no layer is kept frozen during fine-tuning, which means that errors are back-propagated through the entire network during fine-tuning.
- A3: only the first convolutional layer is kept frozen during fine-tuning.
- A4: both convolutional layers are kept frozen during fine-tuning.
- A5: convolutional and LSTM layers are kept frozen during fine-tuning. That is, errors are back-propagated only through the fully-connected layers during fine-tuning.
- A6: only the first convolutional layer is kept frozen during fine-tuning. All other layers have their weights randomly initialized for fine-tuning.
- A7: both convolutional layers are kept frozen during fine-tuning. All other layers have their weights randomly initialized for fine-tuning.
- A8: convolutional and LSTM layers are kept frozen during fine-tuning. Weights in fully-connected layers are randomly initialized for fine-tuning.

Further, these transference approaches are applied considering different scenarios:

- S1: target domain data are used during pre-training and fine-tuning.
- S2: target domain data are used exclusively during fine-tuning.

4 EXPERIMENTAL RESULTS

In this section, we present the datasets and baselines used to evaluate our multi-domain network for speech emotion recognition. Then we discuss our evaluation procedure and report the results of our multi-domain network.

In particular, our experiments aim to answer the following research questions:

RQ1 How effective is the blend of CNN with LSTM networks for speech emotion recognition? How do the learned features compare against hand-crafted features?

RQ2 Which feature transference approach is more appropriate to each target domain?

RQ3 Which domain characteristics affect the most the accuracy of the model?

RQ4 How effective is our multi-domain compared with other domain adaptation models?

4.1 Datasets and Domains

Our analysis is carried on six datasets which differ mainly in terms of language, number of speakers, number of emotions and spontaneity of speech. The details about each dataset are given next:

- AFEW [15]: The Acted Facial Expressions In The Wild dataset contains segments from 37 movies in English. The movies have been chosen keeping in mind the need for different realistic scenarios and large age range of subjects to be captured.
- Emo-DB [5]: The Berlin Emotional Speech dataset features actors speaking emotionally defined sentences. The dataset contains emotional sentences from 10 different actors and ten different texts.
- EMOVO [9]: The dataset consists of sentences recorded by six professional actors. Each speaker reads fourteen Italian sentences expressing different emotions.
- eNTERFACE [29]: The dataset consists of recordings of naive subjects from fourteen nations speaking pre-defined spoken content in English. The subjects listened to six successive short stories eliciting a particular emotion.

- IEMOCAP [6]: The Interactive Emotional Dyadic Motion Capture dataset features ten actors performing improvisations in English, specifically selected to elicit emotional expressions. Each sentence is labeled by at least three human annotators.
- RML:[2] The dataset contains audiovisual emotional expression samples that were collected at Ryerson Multimedia Lab. The RML emotion database is language and cultural background independent. The audio samples were collected from eight human subjects, speaking six different languages (English, Mandarin, Urdu, Punjabi, Persian, Italian). Different accents of English and Chinese were also included.

Table 2: UAR numbers for different models. No domain adaptation is performed.

Dataset	SVM–IS	CNN	CNN+LSTM
AFEW	.333	.344	.338
Emo-DB	.645	.622	.659
EMOVO	.411	.440	.459
eNTERFACE	.456	.419	.454
IEMOCAP	.719	.673	.684
RML	.482	.581	.631

Table 1 presents a summary of the datasets. All datasets were normalized to cover the same emotional states. Specifically, we focus on the well-known six emotions [10]: anger, disgust, fear, happiness, sadness, and surprise.

4.2 Baselines

We considered the following methods in order to provide baseline comparison:

- SVM with Interspeech 2010 features (SVM–IS): the 1,582 acoustic features proposed in [36] are fed into an SVM with RBF kernel [37]. The hyper-parameters of the SVM are chosen by cross-validation. The main objective of using this baseline is to answer RQ1.
- Training on Target (TT): a model CNN+LSTM is trained using only the target domain data. No source domain data are used. The main objective of using this baseline is to assess the benefits of the different feature transference approaches.
- Adaptive SVM [1]: this is a supervised domain adaptation algorithm for speech emotion recognition. The approach poses an optimization problem which seeks a decision boundary close to that of an SVM trained from the source domain, while managing to separate the labeled data from the target domain.

4.3 Setup

We implemented our architecture using Keras [8]. The measure used to evaluate the recognition effectiveness of our models is the standard Unweighted Average Recall (UAR),[3] as presented in [36].

---

[2]http://www.rml.ryerson.ca/rml-emotion-database.html
[3]The UAR metric is the sum of the recalls per class divided by the number of classes.

We conducted five-fold cross validation where datasets are arranged into five folds with approximately the same number of audio samples each. At each run, four folds are used as training set and the remaining fold is used as test set. The results reported are the average of the five runs, and are used to assess the overall discrimination performance of the models. To ensure the relevance of the results, we assess the statistical significance of our measurements by means of a pairwise t-test [34] with p−value ≤ 0.05.

4.4 Results and Discussion

The first experiment is concerned with RQ1. We present a comparison between SVM–IS trained with Interspeech 2010 features, and our deep architecture was trained with raw audio. We considered deep architectures with and without the LSTM layer to assess the impact of using both spatial and sequential features. Table 2 shows UAR numbers for the different models. For this experiment, no domain adaptation is performed. Instead, samples from all datasets were used for training and testing the models using five-fold cross-validation. On average, the CNN+LSTM model provides UAR numbers that are statistically superior than the numbers provided by SVM–IS and CNN models (which are statistically equivalent on average), except for the dataset AFEW. Thus, the features learned by CNN+LSTM architecture lead to significantly raised UAR numbers.

The next set of experiments is devoted to answer RQ2. We evaluate diverse feature transference approaches. Table 3 shows UAR numbers when our architecture is trained using solely target domain data (TT). Therefore, if the target domain is short on labeled data, the model will probably suffer from overfitting. The table also shows the gains obtained by each feature transference approach relatively to TT. That is, we investigated the best freezing/tuning cut-off for each target domain. On average, the best performing transference approach is S1−A2, which uses target domain data during pre-training and fine-tuning and no layer is kept frozen during fine-tuning. Further, gains tend to decrease as more layers are kept frozen during fine-tuning. However, the best approach varies greatly depending on the target domain.

Considering AFEW as the target domain, the best transference approaches are S1−A1, S1−A4, and S1−A7. Usually, using target domain data during pre-training is very beneficial, except for EMOVO for which the best performer was S2−A3. Fine-tuning is extremely important in all cases, specially if target domain data are not used during pre-training. Gains for IEMOCAP are significantly lower than the gains obtained for other domains. Notice that IEMOCAP is the largest dataset, and thus TT achieves very high UAR numbers, which are hard to surpass with domain adaptation. For RML, the best transference approaches are those that freeze less layers. This is because RML is composed of highly diverse languages. Thus, freezing layers will only work if target domain data are used during pre-training. Otherwise, freezing layers would be clearly detrimental to domain adaptation. It is also important to mention that for each target domain, many feature transference approaches lead to significant improvements.

The next set of experiments is devoted to answer RQ3. Table 4 shows UAR numbers obtained with a domain ablation analysis. More specifically, the table shows UAR numbers obtained by different feature transference approaches after excluding one of the

Table 3: Different feature transference approaches and scenarios. Numbers in bold indicate the highest gains for each target domain.

	UAR	Gains over TT														
		S1								S2							
Target	TT	A1	A2	A3	A4	A5	A6	A7	A8	A1	A2	A3	A4	A5	A6	A7	A8
AFEW	.288	**.121**	.101	.047	**.120**	.115	.045	**.121**	.116	.042	.015	.024	.052	.019	.007	.103	.029
Emo-DB	.614	.047	**.117**	.088	.051	.052	.102	.057	.086	-.365	.093	.064	.083	.050	.068	.065	.083
EMOVO	.518	-.095	.053	.014	-.061	-.089	-.071	.034	.014	-.372	.093	**.109**	-.060	-.041	-.044	.008	-.017
eNTERF	.441	.032	.133	.114	.061	.032	**.153**	.087	.045	-.353	.027	-.015	-.016	-.034	.002	-.026	-.037
IEMOCAP	.682	.004	.004	-.002	-.009	-.003	.003	**.017**	**.017**	-.363	.003	-.015	-.016	-.034	.003	-.026	-.035
RML	.623	-.014	.032	.041	.005	-.002	**.073**	.035	.028	-.518	**.074**	.062	-.085	-.145	.054	-.087	-.143
Average	–	.016	.073	.050	.028	.017	.051	.058	.053	-.321	.051	.038	-.007	-.031	.015	.006	-.020
Std.	–	.072	.051	.044	.063	.068	.078	.039	.041	.189	.041	.049	.064	.067	.040	.069	.076

source domains from the pre-training. This enables us to grasp the domain characteristics that affect the most the effectiveness of our multi-domain network.

The reference UAR value (All) is given by the model built using data from all domains. We first analyze scenario S1, in which target domain data are used during pre-training and fine-tuning. As can be seen, in almost all cases it is better removing one of the source domains from pre-training. Using AFEW data during pre-training is highly detrimental in all cases. The probable explanation is that the AFEW domain is highly discrepant from all other domains. Similarly, IEMOCAP data are highly detrimental for AFEW, Emo-DB, eNTERFACE and RML target domains. IEMOCAP data are also very discrepant from other domains. Removing out-of-domain data from pre-training is not beneficial only for S1−A1 when RML is the target domain. Thus, we conclude that if target domain data are used during pre-training, it is detrimental to have out-of-domain data during pre-training, specially if out-of-domain data are highly discrepant from the target domain data.

Very different trends are observed when we analyze scenario S2. In this case, target domain data are used exclusively during fine-tuning, and therefore we may expect that out-of-domain data used during pre-training are less discrepant. Using IEMOCAP data during pre-training is highly beneficial. This is probable due to the size of IEMOCAP dataset. This is also a probable explanation for the robustness when removing specific out-of-domain datasets when IEMOCAP is the target domain. The RML domain seems to benefit the most from out-of-domain data. In general, we conclude that if target domain data are not included during pre-training, it is beneficial to have out-of-domain data during pre-training, even if out-of-domain data are highly discrepant from the target domain data.

The last set of experiments is concerned with RQ4, that is, to assess the effectiveness of our multi-domain network when compared with state-of-the-art domain adaptation solutions for speech emotion recognition. Table 5 shows UAR numbers obtained by Adaptive SVM. The table also shows UAR numbers obtained by our multi-domain network. As can be seen, our multi-domain network outperformed Adaptive SVM in all target domains considered in the study. Gains are statistically significant, and range from 4.3% to 18.4%, depending on the target domain.

5 CONCLUSIONS

Automatically recognizing human emotions from speech is currently one of the most challenging tasks in the field of affective computing. In solving this task we are often in the situation that we have a large collection of labeled out-of-domain data but truly desire a model that performs well in a target domain which is short on labeled data. To deal with this situation we proposed a deep architecture which implements a multi-domain network. More specifically, the architecture is a blend of CNN with LSTM networks, and extracts spatial and sequential features from raw audio. In order to evaluate different feature transference approaches, we investigated the best freezing/tuning cut-off for each target domain. We also investigated whether it is beneficial to use target domain data during pre-training. We performed a comprehensive experiment using six domains, which may differ in terms of language, emotions, amount of labels, and recording conditions. Our feature transference approaches provide gains that range from 4.3% to 18.4% when compared with recent domain adaptation approaches for speech emotion recognition.

6 ACKNOWLEDGEMENTS

REFERENCES

[1] Mohammed Abdel-Wahab and Carlos Busso. 2015. Supervised domain adaptation for emotion recognition from speech. In *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*. 5058–5062.

[2] Tanja Banziger, Didier Grandjean, and Klaus Scherer. 2009. Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT). *Emotion* 9, 5 (2009), 691–704.

[3] Anton Batliner, Stefan Steidl, Björn W. Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Vered Aharonson, Loïc Kessous, and Noam Amir. 2011. Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language* 25, 1 (2011), 4–28.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning* 79, 1-2 (2010), 151–175.

Table 4: Domain ablation analysis. The table shows UAR numbers after excluding a domain from the pre-training, so a low UAR number indicates that an important domain was removed from pre-training. Symbol ↑ indicates that UAR has raised significantly. Symbol • indicates that UAR has not changed significantly. Symbol ↓ indicates that UAR has dropped significantly. We omitted UAR numbers for A5 to A8 in order to avoid cutter. Highest UAR numbers for each feature transference approach are highlighted in bold.

		UAR numbers							
		S1				S2			
Target	Source	A1	A2	A3	A4	A1	A2	A3	A4
AFEW	All	.323	.317	.301	.322	.300	.292	**.295**	.303
	− Emo-DB	.356↑	.440↑	.469↑	.517↑	.304↑	**.306**↑	**.299**↑	.283↓
	− EMOVO	.380↑	.442↑	.473↑	.561↑	.307↑	.289•	.262↓	.322↑
	− eNTERFACE	**.390**↑	.464↑	.487↑	.566↑	.284↓	.291•	.269↓	.314↑
	− IEMOCAP	.315•	**.514**↑	**.572**↑	**.625**↑	.237↓	.272↓	.280↓	.275↓
	− RML	.366↑	.424↑	.487↑	.539↑	**.314**↑	**.298**↑	.287↑	**.332**↑
Emo-DB	All	.643	.685	.668	.645	.389	**.671**	.653	.665
	− AFEW	.725↑	.830↑	**.835**↑	**.843**↑	**.397**↑	.652↓	.644↓	.648↓
	− EMOVO	.688↑	.792↑	.789↑	.786↑	.382↓	**.667**•	.658•	.638↓
	− eNTERFACE	.689↑	.775↑	.780↑	.767↑	.393•	.620↓	**.668**↑	.662•
	− IEMOCAP	**.798**↑	**.856**↑	**.840**↑	.824↑	.372↓	.653↓	.639↓	.660↓
	− RML	.692↑	.762↑	.748↑	.778↑	**.401**↑	.655↓	.658↑	**.683**↑
EMOVO	All	.469	.545	.525	.496	.325	.566	**.574**	.487
	− AFEW	**.635**↑	**.691**↑	**.716**↑	**.735**↑	.332↑	.542↓	**.571**•	**.547**↑
	− Emo-DB	.566↑	.664↑	.675↑	.664↑	.320•	.567•	.549↓	.528↑
	− eNTERFACE	.566↑	.655↑	.671↑	.695↑	.334↑	**.595**↑	.561↓	**.543**↑
	− IEMOCAP	.619↑	.634↑	.641↑	.696↑	**.351**↑	.506↓	.495↓	.498↑
	− RML	.544↑	.631↑	.648↑	.611↑	.309↓	.560•	.563↓	**.547**↑
eNTERFACE	All	.455	.500	.491	.468	**.247**	.484	**.499**	**.395**
	− AFEW	.645↑	**.694**↑	**.721**↑	.737↑	.238↓	**.492**↑	.490↓	**.397**•
	− Emo-DB	.538↑	.602↑	.642↑	.644↑	.231↓	**.499**↑	.482↓	**.391**•
	− EMOVO	.632↑	.652↑	.654↑	.681↑	**.248**•	.476↓	.477↓	.375↓
	− IEMOCAP	**.749**↑	**.696**↑	.711↑	**.751**↑	**.244**•	.481•	.483↓	.381↓
	− RML	.624↑	.625↑	.639↑	.674↑	.233↓	.471↓	.472↓	384↓
IEMOCAP	All	.685	.685	.681	.676	.441	**.684**	.672	**.671**
	− AFEW	**.780**↑	**.762**↑	**.771**↑	**.783**↑	**.470**↑	**.688**•	.671•	.656↓
	− Emo-DB	.739↑	.722↑	.737↑	.741↑	.435↓	**.686**•	.669•	.649↓
	− EMOVO	.756↑	.746↑	.751↑	.762↑	.456↑	**.686**•	**.680**↑	**.665**•
	− eNTERFACE	.765↑	.740↑	.755↑	.772↑	.427↓	.680•	**.683**↑	**.667**•
	− RML	.755↑	.735↑	.746↑	.764↑	.459↑	.671↓	**.681**↑	.659↓
RML	All	**.615**	.643	.649	.626	.301	**.669**	**.662**	**.570**
	− AFEW	.470↓	.715↑	**.744**↑	**.722**↑	**.318**↑	.656↓	.644↓	.562↓
	− Emo-DB	.503↓	.709↑	.714↑	.663↑	.297•	.650↓	**.664**•	.543↓
	− EMOVO	.468↓	.690↑	.704↑	.655↑	.287↓	.653↓	.653↓	.555↓
	− eNTERFACE	.471↓	.710↑	.707↑	.687↑	.302•	.648↓	**.660**•	.557↓
	− IEMOCAP	.543↓	**.736**↑	**.754**↑	.693↑	.298•	.656↓	.656↓	.533↓

[5] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *European Conference on Speech Communication and Technology*. 1517–1520.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language*

Table 5: UAR numbers for Adaptive SVM and CNN+LSTM.

Target	Adaptive SVM	CNN+LSTM	Gain
AFEW	.539	.625	.159
Emo-DB	.797	.856	.074
EMOVO	.692	.735	.062
eNTERFACE	.634	.751	.184
IEMOCAP	.751	.783	.043
RML	.721	.760	.054

Resources and Evaluation 42, 4 (2008), 335–359.

[7] Carlos Busso, Srinivas Parthasarathy, Alec Bur mania, Mohammed Abdel-Wahab, Najmeh Sadoughi, and Emily Mower Provost. 2017. MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2017), 67–80.

[8] F. Chollet. 2015. keras. https://github.com/fchollet/keras. (2015).

[9] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO corpus: an Italian emotional speech database. In *Intl. Conference on Language Resources and Evaluation*. 3501–3504.

[10] Roddy Cowie and Randolph R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication* 40, 1-2 (2003), 5–32.

[11] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn W. Schuller. 2017. Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters* 24, 4 (2017), 500–504.

[12] Jun Deng, Zixing Zhang, Florian Eyben, and Björn W. Schuller. 2014. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters* 21, 9 (2014), 1068–1072.

[13] Jun Deng, Zixing Zhang, Erik Marchi, and Björn W. Schuller. 2013. Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition. In *ACII Association Conference on Affective Computing and Intelligent Interaction*. 511–516.

[14] Jun Deng, Zixing Zhang, and Björn W. Schuller. 2014. Linked source and target domain subspace feature transfer learning - exemplified by speech emotion recognition. In *Intl. Conference on Pattern Recognition*. 761–766.

[15] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia* 19, 3 (2012), 34–41.

[16] Matthis Drolet, Ricarda Schubotz, and Julia Fisher. 2012. Authenticity affects the recognition of emotions in speech: Behavioral and fMRI evidence. *Cognitive Affective & Behavioral Neuroscience* 12, 1 (2012), 140–150.

[17] Florian Eyben, Björn W. Schuller, and Gerhard Rigoll. 2012. Improving generalisation and robustness of acoustic affect recognition. In *Intl. Conference on Multimodal Interaction*. 517–522.

[18] Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Annual Conference of the Intl. Speech Communication Association*. 223–227.

[19] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech emotion recognition using CNN. In *ACM Intl. Conference on Multimedia*. 801–804.

[20] Zhengwei Huang, Wentao Xue, Qirong Mao, and Yongzhao Zhan. 2017. Unsupervised domain adaptation for speech emotion recognition using PCANet. *Multimedia Tools and Applications* 76, 5 (2017), 6785–6799.

[21] Tom Johnstone, Carien van Reekum, Terrence Oakes, and Richard Davidson. 2006. The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions. *Social Cognitive and Affective Neuroscience* 1, 3 (2006), 242–249.

[22] P. Juslin and P. Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin* 129, 5 (2003), 770.

[23] Yelin Kim, Honglak Lee, and Emily Mower Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*. 3687–3691.

[24] Shashidhar Koolagudi, Anurag Barthwal, Swati Devliyal, and K. Rao. 2012. Real life emotion classification from speech using gaussian mixture models. In *Intl. Conference Contemporary Computing*. 250–261.

[25] Shashidhar Koolagudi and K. Rao. 2012. Emotion recognition from speech: a review. *Intl. Journal of Speech Technology* 15, 2 (2012), 99–117.

[26] Shashidhar Koolagudi, Sudhin Ray, and K. Rao. 2010. Emotion classification based on speaking rate. In *Intl. Conference Contemporary Computing*. 316–327.

[27] Qirong Mao, Wentao Xue, Qiyu Rao, Feifei Zhang, and Yongzhao Zhan. 2016. Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*. 2608–2612.

[28] Erik Marchi, Florian Eyben, Gerhard Hagerer, and Björn W. Schuller. 2016. Real-time tracking of speakers' emotions, states, and traits on mobile platforms. In *Annual Conference of the Intl. Speech Communication Association*. 1182–1183.

[29] Olivier Martin, Irene Kotsia, Benoit M. Macq, and Ioannis Pitas. 2006. The eNTERFACE'05 Audio-Visual Emotion Database. In *Intl. Conference on Data Engineering Workshops*. 8.

[30] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño. 2001. Speech emotion recognition using hidden Markov models. In *European Conference on Speech Communication and Technology*. 2679–2682.

[31] Pierre-Yves Oudeyer. 2003. The production and recognition of emotions in speech: features and algorithms. *Intl. Journal of Human-Computer Studies* 59, 1-2 (2003), 157–183.

[32] S. Ramakrishnan and Ibrahiem El Emary. 2013. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems* 52, 3 (2013), 1467–1478.

[33] K. Rao, Shashidhar Koolagudi, and Vempada Reddy. 2013. Emotion recognition from speech using global and local prosodic features. *I. J. Speech Technology* 16, 2 (2013), 143–160.

[34] Tetsuya Sakai. 2014. Statistical reform in information retrieval? *SIGIR Forum* 48, 1 (2014), 3–12.

[35] Björn W. Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2009. Emotion recognition from speech: Putting ASR in the loop. In *IEEE Intl. Conference on Acoustics, Speech, and Signal Processing*. 4585–4588.

[36] Björn W. Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *Annual Conference of the Intl. Speech Communication Association*. 312–315.

[37] Björn W Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, Shrikanth S Narayanan, et al. 2010. The INTERSPEECH 2010 paralinguistic challenge.. In *Interspeech*, Vol. 2010. 2795–2798.

[38] Björn W. Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, André Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2015. Cross-corpus acoustic emotion recognition: Variances and strategies. In *Intl. Conference on Affective Computing and Intelligent Interaction*. 470–476.

[39] Dino Seppi, Anton Batliner, Björn W. Schuller, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, and Vered Aharonson. 2008. Patterns, prototypes, performance: classifying emotional user states. In *Annual Conference of the Intl. Speech Communication Association*. 601–604.

[40] Peng Song, Wenming Zheng, Shifeng Ou, Xinran Zhang, Yun Jin, Jinglei Liu, and Yanwei Yu. 2016. Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication* 83 (2016), 34–41.

[41] Katja Spreckelmeyer, Marta Kutas, Thomas Urbach, Eckart Altenmuller, and Thomas Munte. 2009. Neural processing of vocal emotion and identity. *Brain and Cognition* 69, 1 (2009), 121–126.

[42] B. Stienen, A. Tanaka, and B. de Gelder. 2011. Emotional voice and emotional body postures influence each other independently of visual awareness. *Plos One* 10, 6 (2011), e25517.

[43] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Hans-Günter Meier, and Björn W. Schuller. 2011. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *IEEE Intl. Conference on Acoustics, Speech, and Signal Processing*. 5688–5691.

[44] A. Tanaka, A. Koizumi, H. Imai, S. Hiramatsu, E. Hiramoto, and B. de Gelder. 2010. I feel your voice: Cultural differences in the multisensory perception of emotion. *Psychological Science* 21, 9 (2010), 1259–1262.

[45] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Intl. Conference on Machine Learning*. 1096–1103.

[46] Carl Williams and Kenneth Stevens. 1972. Emotions and speech: some acoustical correlates. *The Journal of the Acoustical Society of America* 52, 4 (1972), 1238–1250.

[47] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn W. Schuller, and Gerhard Rigoll. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31, 2 (2013), 153–163.

[48] Went Xue, Zhengwei Huang, Xin Luo, and Qirong Mao. 2015. Learning speech emotion features by joint disentangling-discrimination. In *Intl. Conference on Affective Computing and Intelligent Interaction*. 374–379.

[49] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Annual Conference on Neural Information Processing Systems*. 3320–3328.

[50] Biqiao Zhang, Emily Mower Provost, and Georg Essl. 2016. Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*. 5805–5809.

[51] Zixing Zhang, Fabien Ringeval, Jing Han, Jun Deng, Erik Marchi, and Björn W. Schuller. 2016. Facing realism in spontaneous emotion recognition from speech: feature enhancement by autoencoder with LSTM neural networks. In *Annual Conference of the Intl. Speech Communication Association*. 3593–3597.