LEARNING TRANSFERABLE FEATURES FROM MULTIPLE SOURCE DOMAINS FOR SPEECH EMOTION RECOGNITION

ALISON DE OLIVEIRA MARCZEWSKI

LEARNING TRANSFERABLE FEATURES FROM MULTIPLE SOURCE DOMAINS FOR SPEECH EMOTION RECOGNITION

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Adriano Veloso

Belo Horizonte
Julho de 2017

ALISON DE OLIVEIRA MARCZEWSKI

LEARNING TRANSFERABLE FEATURES FROM MULTIPLE SOURCE DOMAINS FOR SPEECH EMOTION RECOGNITION

Dissertation presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Ciência da Computação.

Advisor: Adriano Veloso

Belo Horizonte
July 2017

Marczewski, Alison de Oliveira

D1234p Learning Transferable Features from Multiple Source Domains for Speech Emotion Recognition / Alison de Oliveira Marczewski. — Belo Horizonte, 2017
xxiv, 44 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Adriano Veloso

1. Computação — Dissertações. 2. Aprendizado de Máquina — Dissertações. I. Orientador. II. Título.

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

Dedico este trabalho a minha familia e amigos, que me proporcionaram tudo e me deram o devido suporte para que eu chegasse ate aqui.

Acknowledgments

Agradeço primeiramente aos meus pais, Elisangela Oliveira e Adilton Marczewski, e ao meu tio Márcio Marczewski, pelo apoio e por proporcionar as condições para que eu chegasse até aqui desde o início de toda essa caminhada. À Juliana Nunes pela paciência e apoio me dado na caminhada em direção da realização desse trabalho. Aos meus amigos atuais e aos que já se foram por me ajudarem nos momentos difíceis e me incentivarem a seguir esse caminho. Aos meus professores por me guiar e incentivar durante tantos anos. Aos meus orientadores, Marco Cristo e Adriano Veloso, por me introduzir e conduzir tão bem no meu crescimento no ramo da pesquisa científica. E a todas as pessoas que me ajudaram de alguma forma.

“Todos nós podemos dirigir um carro. Talvez uns dirijam melhor, mas todos podemos dirigir. Mas quantos de nós têm resistência, a coordenação motora, a concentração e os reflexos para ser um piloto de corrida?”

()

Resumo

Reconhecimento de emoção através da fala é um ponto chave na direção da inteligência emocional em interações homem-máquina avançadas. Identificar emoções na fala humana requer aprender descritores que sejam robustos e discriminativos entre os mais diversos domínios, estes que se distinguem em termos de idioma, espontaneidade da fala, condições de gravação do áudio, além dos tipos de emoções expressas. Isso descreve um cenário de aprendizado que as distribuições disjuntas de descritores e rótulos estão sujeitas a divergências substanciais entre domínios. Neste trabalho, propomos uma arquitetura profunda que explora em conjunto uma rede convolucional para extração de descritores compartilhados entre domínios e uma rede recorrente LSTM (Long Short-Term Memory) para classificar emoções usando descritores específicos de domínio. Utilizamos descritores genéricos para permitir a adaptação de modelos a partir de vários domínios de origem, dada a espacidade de dados de fala e o fato de que os domínios alvo apresentam poucos dados rotulados. Um extenso experimento entre os datasets nos mais variados domínios revela que descritores genéricos proveem ganhos entre 4.3% e 78.6% no reconhecimento de emoção em fala. Nós avaliamos várias abordagens para adaptação de um domínio em outro e realizamos um estudo de ablação para entender quais domínios de origem mais contribuem para a efetividade geral no reconhecimento de emoção para um domínio alvo. Para entender a diferença na efetividade entre domínios e emoções, nós analisamos a divergência entre eles para entender melhor as razões pelas quais o processo de adaptação ao domínio alvo não é efetivo quando alguns outros domínios estão na base de dados fonte.

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 78.6% 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. In addition, to understand the effectiveness difference between domains and emotions, we analyze the divergence among them to understand better the reasons why adaptation process to the target domain is uneffectiveness when some other domains are in the source dataset.

List of Figures

List of Tables

Contents

Chapter 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 [Juslin and Laukka, 2003]. 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 [Oudeyer, 2003].

While there is considerable evidence that speech features can differentiate emotional states [Deng et al., 2014a; Wöllmer et al., 2013; Stuhlsatz et al., 2011], 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) [Drolet et al., 2012]. 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.

It is performed an ablation domain analysis in order to elucidate the benefits of incorporating multi-domain data into the final recognition model. It is showed 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. In addition, to elucidate divergence among datasets and emotions, it is also performed a brief analysis to evaluate the divergence among them. As result, it is reported gains that vary from 4.3% to 78.6%, depending on the target domain and feature transference approach.

1.1 Contributions

In this work, we propose a deep architecture for speech emotion recognition composed of a convolutional neural network (CNN) to extract domain-shared features from multi-domain data, and a long short-term memory network (LSTM) that is fed with the extracted domain-shared features for emotion prediction and uses a limited amount of target-domain data.

The main contributions in this work are:

1. the blending of a CNN with a LSTM to exploit both spatial and temporal information of speech features for improving 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;

2. It is discussed several feature transference approaches in 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;

3. We conducted rigorous experiments using six standard speech emotion datasets that correspond to different domains. Recognition models are trained using different transference approaches;

4. Experimental results show that our network is capable to learn and transfer the learng featuresfrom one domain to other.

1.2 Text Organization

The remainder of this thesis is organized as follows. Chapter 2 presents the background and relevant related works on acoustic features, concerns in audios, transfer learning and domain adaptation, and feature learning. In Chapter 3 we describe the proposed multi-domain networks. In Chapter 4 we describe the datasets used in the experiments, its characterization, as well as the data pre-processing step by step. In Chapter 5 we present, discuss and analyse empirical and subjectively the results of our multi-domain network. In addition we also present an ablation domain analysis. In Chapter 6 we analyse the divergence among datasets and emotions to elucidate and give support for results presented in 5. And finally, Chapter 7 shows conclusions of this work.

Chapter 2

Background and Related Work

In this chapter we describe the importance of emotion recognition and present the main approaches to classify emotions in audio. In Section 2.1, we present the introduction for emotion recognition in audio and its importance. In Section 2.2, the idea of feature engineering for emotion recognition using well known audio features. In Section 2.3, we present concerns related to data. In Section 2.4, we present the main idea of this work that is transfer the learning from some domains to a specific domain to minimize the concerns presented in Section 2.3. In Section 2.5, the concept of feature learning for emotion recognition in audio is presented. As additional analysis, in Section 2.6, we show how the divergence between domains can be estimated. Finally, in Section 2.7, we show the main differences between this work and the works aforementioned.

2.1 Emotion Recognition

Research on the recognition of emotional expressions in voices is of great academic interest in psychology [Velten, 1968; Banziger et al., 2009], neurosciences [Tanaka et al., 2010; Stienen et al., 2011; Spreckelmeyer et al., 2009; Johnstone et al., 2006] and affective computing [Marchi et al., 2016; Schuller et al., 2015; Deng et al., 2014a; Wöllmer et al., 2013]. A number of researchers investigated acoustic correlates of emotions from human speech. In one of the first studies [Williams and Stevens, 1972], 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.

2.2 Features

There are studies on how acoustic correlates of emotions from speech are transformed into features for supervised learning algorithms. In [Koolagudi and Rao, 2012; Ramakrishnan and Emary, 2013], the authors provide reviews on a wide range of features employed for emotion recognition from speech. In [Nogueiras et al., 2001], the authors present an approach based on hidden semi-continuous Markov models, which are built using specific energy and pitch features. In [Koolagudi et al., 2012], the authors employ mel frequency cepstral coefficients (MFCCs) as features for a Gaussian mixture model classifier. A similar MFCC model was proposed in [Koolagudi et al., 2010] and features related to speaking rate are also explored to categorize the emotions. In [Rao et al., 2013], 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 [Batliner et al., 2011], 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 [Schuller et al., 2009a], the authors introduce string kernels as a novel solution in the field.

2.3 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 [Eyben et al., 2012]. More serious concerns about data used for emotion recognition from speech were presented in [Schuller et al., 2015], 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 [Seppi et al., 2008]. Alternatively, more realistic acted data were recently presented in [Busso et al., 2017].

2.4 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 [Yosinski et al., 2014], the authors provide a detailed analysis on how transferable are features in deep neural networks. They found that initializing a network with transferred features from almost any number of layers can produce a boost to generalization that lingers even after fine-tuning to the target dataset. In [Zhang et al., 2016a], the authors explore a multi-task framework in which speech or song are jointly leveraged in emotion recognition in a cross-corpus setting. In [Song et al., 2016], 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 [Huang et al., 2017], 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 [Mao et al., 2016], 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 [Deng et al., 2014b], the authors present a feature transfer learning method using denoising autoencoders [Vincent et al., 2008] 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 [Deng et al., 2014a], 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. In [Abdel-Wahab and Busso, 2015], 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. Finally, in [Deng et al., 2013] 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.

2.5 Feature Learning

Deep neural networks were already used for emotion recognition from speech. In [Stuhlsatz et al., 2011], 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 [Deng et al., 2017], 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 [Vincent et al., 2008] to minimize this data mismatch problem. In [Han et al., 2014], 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 [Kim et al., 2013], 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 [Zhang et al., 2016b], 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 [Huang et al., 2014], 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 [Xue et al., 2015], 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.

2.6 Divergence Analysis

Divergence analysis has been applied in some scenarios to measure how different are two or more parameters, where parameter can be time series, clusters, and data distribution. Kullback-Leibler (KL) divergence has been used in several works for this purpose. In [Helén and Virtanen, 2007, 2009], KL divergence is used to estimate how similar are audios in an audio query problem. In [Huang, 2008], KL divergence is used as similarity metric among documents in a clustering process. An SVM using the KL divergence between single Gaussians was able to classify 84% of songs correctly in [Mandel and Ellis, 2005]. In [Socher et al., 2011], KL divergence is used to evaluate the model's ability to predict sentiment distributions, in other words, to evaluate how similar or divergent are two distributions. In [Nisius et al., 2009], an approach for Fingerprint Reduction is proposed in the basis of KL divergence analysis of bit distributions. In Study on Gesture-Sound Similarity, KL divergence is used to measure how much one signal can be explained by the other [Caramiaux et al., 2010]. KL divergence is also used in Non-negative spectrogram factorization (NSF) in [Parry and Essa, 2007].

KL divergence has been also used in dimension reduction where data distribution in high dimension must have the same distribution in low dimension as presented in [der Maaten and Hinton, 2008]. Gao et al. [2011] present an application of KL divergence more similar with that used in our work that is analyse how divergent are datasets to each other. The authors propose an adjustment in model when the test dataset has a different distribution related to train dataset. To get this, a transfer learning framework for latent variable model is proposed which can utilize the distance (or divergence) of the two datasets to modify the parameters of the obtained latent variable model. So they do not need to rebuild the model and only adjust the parameters according to the divergence, which will adopt different datasets.

2.7 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. Finally, to understand and give a better support for conclusions, (iv) a divergence analysis is performed among datasets and emotions.

Chapter 3

Multi-domain Network

In this chapter we propose a multi-domain network with the objective of learning features for speech emotion recognition in a wild scenario with many domains.

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 [Ben-David et al., 2010]. 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

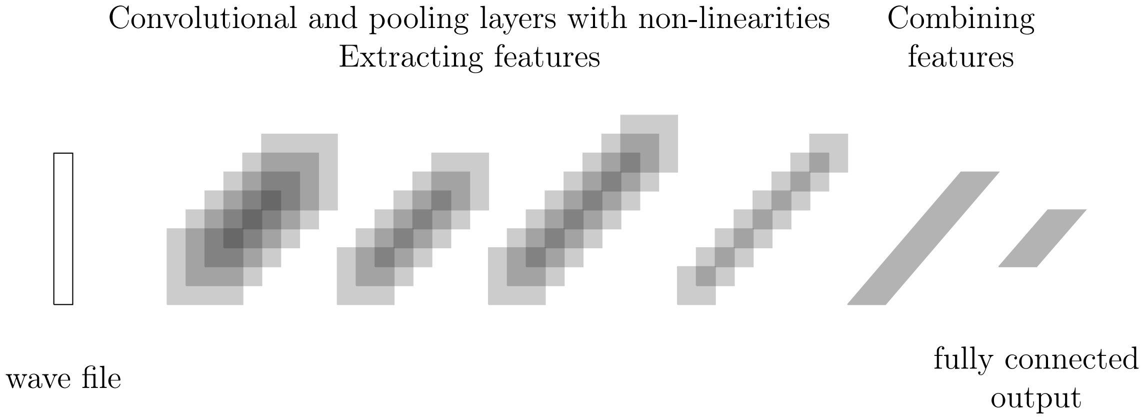


Figure 3.1: Multi-Domain Network architecture for learning transferable features. Convolutional layers are followed by a layer that combines features extracted by convolutional layers. In addition, there is a dropout layer as regularizer after each convolutional and combiner feature layer. Different feature transference approaches are designed using this architecture.

labels. Thus, for domains that are short of labeled data transferable features enable model adaptation from multiple domains.

3.1 Network Architectures

The general idea is 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 3.1. The network works by passing each audio sample through a feature transformation to produce a fixed-length vector representation. After that, spatial features are computed for the audio input, and then a layer captures how emotions evolve over time.

The next two sections introduce our deep architectures: convolutions with fully connected layers and convolutions with LSTM followed by a fully connected layer.

3.1.1 Convolutions combined with a fully connected layer

The network receives a 54,000 dimensional input representing audio samples. It has four hidden layers, including two uni-dimensional convolutional layers followed by two fully connected layers. The convolutional layers apply kernels with 128 dimensions, combined with ReLUs and a dropout level of 0.30. The first fully connected layer

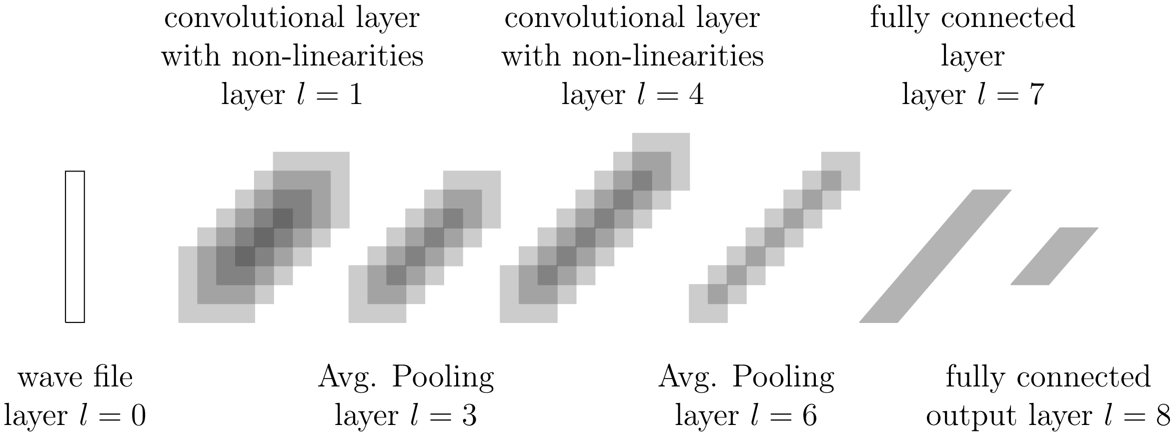


Figure 3.2: In this Multi-Domain Network architecture for learning transferable features, a Fully-Connected layer is used as feature combiner extracted by Convolutional layers. In this illustration, the convolutional layers already include non-linearities and, thus, a convolutional layer actually represents two layers.

receives 128 dimensional inputs, which are then flattened into a single 256 dimensional output, and returns one 1000 dimensional vector output followed by hyperbolic tangent activation. The next fully connected layer is composed of 6 units. Again, a dropout level of 0.30 is applied, as illustrated in Figure 3.2. 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.

3.1.2 Convolutions combined with a LSTM layer

This architecture is similar to that presented in 3.1.1, differences are after convolutions only. The first layer after convolutions is a LSTM that receives 128 dimensional inputs, and returns two 500 dimensional vector outputs which are then flattened into a single 1,000 dimensional output. The next fully connected layer is composed of 6 units and are combined with the hyperbolic tangent activation. Again, a dropout level of 0.30 is applied, as illustrated in Figure 3.3.

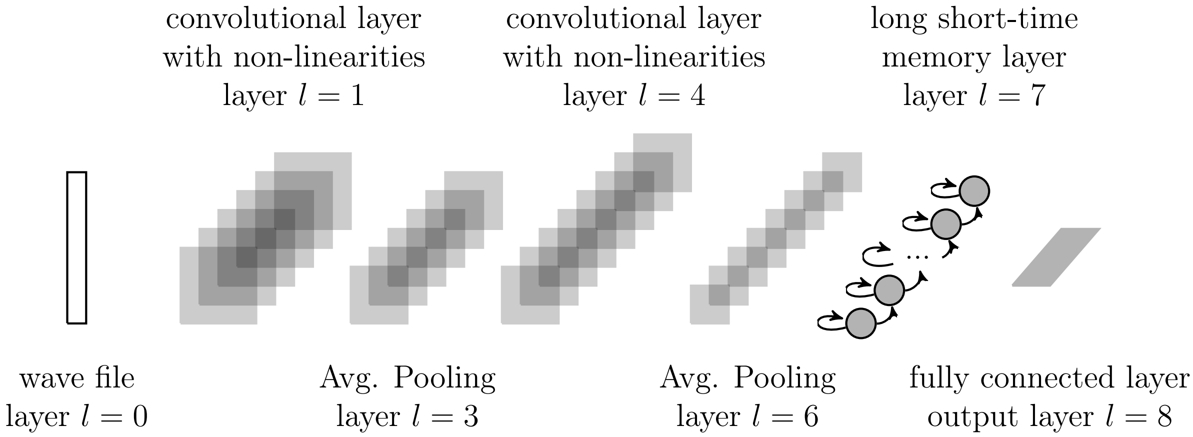


Figure 3.3: In this Multi-Domain Network architecture for learning transferable features, a Long Short-Time Memory (LSTM) layer is used as feature combiner extracted by Convolutional layers. In this illustration, the convolutional layers already include non-linearities and, thus, a convolutional layer actually represents two layers.

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 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.

3.2.1 Transference Approaches

The main intuition that we exploit for feature transferability is that the features must eventually transits from general to specific along our deep architecture, and feature transferability drops significantly in higher layers with increasing domain discrepancy [Yosinski et al., 2014]. 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 thought 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.

Chapter 4

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:

- AFEW [Dhall et al., 2012]: 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 [Burkhardt et al., 2005]: 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 [Costantini et al., 2014]: The dataset consists of sentences recorded by six professional actors. Each speaker reads fourteen Italian sentences expressing different emotions.

- eNTERFACE [Martin et al., 2006]: 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 [Busso et al., 2008]: 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[1]: 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 4.1 presents a summary of the datasets with general information like age, language, and number of samples.

Dataset/Domain	Age	Language	Emotion	Gender	Recording	Sampling rate	# samples
AFEW	children/adults	English	natural	balanced	movies	48kHz	1,156
Emo-DB	adults	German	acted	balanced	studio	16kHz	535
EMOVO	adults	Italian	acted	balanced	studio	48kHz	588
eNTERFACE	adults	English	induced	unbalanced	normal	16kHz	1,292
IEMOCAP	adults	English	acted	balanced	studio	48kHz	10,039
RML	adults	many	induced	balanced	studio	22kHz	720

Table 4.1: Summary by datasets.

4.1 Preprocessings

All datasets were normalized to cover the same emotional states. Specifically, we focus on the well-known six emotions [Cowie and Cornelius, 2003]: anger, disgust, fear, happiness, sadness, and surprise. Beyound that, only audio with at least two and at most six seconds of duration were selected.

4.1.1 Audio files

For this work, all audio files were preprocessed in two steps: normalizing wave files and preparing data for the models.

1. Normalizing wave files:
   - they were transformed to have only one channel of data, in other words, stereo into mono audios and
   - downrated to 9kHz. Thus, are used 9k integers to represent one second of audio.
2. Input for model:

[1] `http://www.rml.ryerson.ca/rml-emotion-database.html`

- for each audio, the values are normalized between 0 and 1.
- as audios do not have the same duration, padding was applied so that audios with different durations have representations with the same size, so all audios are represented by 54k (e.g. 9k by second x 6 seconds = 54k) values between 0 and 1.
- time values are standardized so that they are centered around 0 with a standard deviation of 1, as illustrated in Figure 4.1.

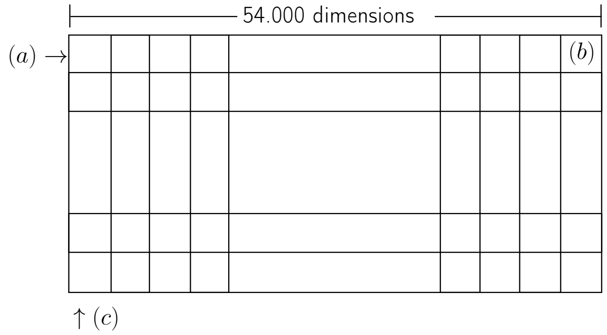


Figure 4.1: Preprocessing before using in models

4.2 Final dataset after pre-processings

In this section, we present a summarization of the final dataset used in this work. Firstly, we have in the Table 4.2 a summary of the datasets after all preprocessings.

It is clear the dataset is unbalanced. For instance IEMOCAP and eNTERFACE together comprise over 60% of the full dataset. Emo-DB and EMOVO together are less than 14% and they do not contain *surprise* and *happiness* emotion, respectively. In addition, RML has an audio average duration 1 second greater than that of other datasets.

Dataset	Emotions	Sampling rate	# samples	Weight	Duration(sec)
AFEW	all	9kHz	568	12.19%	3.05±0.82
Emo-DB	*surprise	9kHz	287	6.16%	3.12±0.90
EMOVO	*happiness	9kHz	336	7.21%	3.33±1.01
eNTERFACE	all	9kHz	1,047	22.48%	3.05±0.77
IEMOCAP	all	9kHz	1,770	38.00%	3.56±1.10
RML	all	9kHz	650	13.95%	4.95±0.64

*does not contain that emotion

Table 4.2: Summary of the datasets

On the other hand when we take a look by emotion, this discrepancy is smaller. It still is unbalanced, *anger* and *sadness* comprise over 50% of the full dataset. *Disgust, fear* and *surprise* are balanced among them with 10% each. In average, the duration is almost the same among emotions, as presented in Table 4.3.

Emotion	# samples	Weight	Duration(sec)
Anger	1,310	28.12%	3.50±1.04
Disgust	467	10.03%	3.62±1.17
Fear	470	10.09%	3.48±1.10
Happiness	802	17.22%	3.46±1.11
Sadness	1,143	24.54%	3.65±1.14
Surprise	466	10.00%	3.44±1.11

Table 4.3: Summary of the emotions

In table 4.4 we analysis the emotion weight by dataset. We can see that half of datasets are balanced by emotion(e.g. *EMOVO, eNTERFACE, RML*). *IEMOCAP* is the worst case with *anger* and *sadness* representing almost 75% of samples, as presented in Table 4.4.

Emotion	AFEW	Emo-DB	EMOVO	eNTERFACE	IEMOCAP	RML
Anger	22.40%	34.49%	19.05%	19.68%	39.55%	17.38%
Disgust	14.29%	14.98%	22.92%	15.95%	0.11%	14.92%
Fear	10.41%	13.94%	18.15%	16.91%	1.24%	17.08%
Happiness	18.87%	18.12%	0.00%	15.38%	20.73%	17.69%
Sadness	22.57%	18.47%	19.64%	16.43%	35.37%	15.08%
Surprise	11.46%	0.00%	20.24%	15.66%	2.99%	17.85%

Table 4.4: % of each emotion by dataset

Beyond concerns presented in section 2, we saw in this section additional challanges for our domain adaptation approach like unbalanced distribution among datasets (Table 4.2), emotions (Table 4.3), and among emotions inside datasets (Table 4.4).

Chapter 5

Experimental Results

In this chapter, we present the 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 in multi-domain scenario for training and test? 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 with a defined target compared with other models?

5.1 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 [Schuller et al., 2010] are fed into an SVM with RBF kernel [Schuller et al., 2009b]. The hyper-parameters of the SVM are chosen by cross-validation. The main objective of using this baseline is to answer RQ1 and RQ4. Although this baseline is not recent, it is a strong baseline. For instance, in RML dataset it get 75.00% of accuracy against 68.57% from [Ooi et al., 2014] and in Emo-DB 84.7% against 85.6% from [Huang et al., 2016].

- 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.

5.2 Setup

We implemented our architecture using Keras [Chollet, 2015] and Theano [Bergstra et al., 2011] as backend. The measure used to evaluate the recognition effectiveness of our models is the standard Unweighted Average Recall (UAR),[1] as presented in [Schuller et al., 2009b]. 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 [Sakai, 2014] with $\mathrm{p}-\mathrm{value} \leq 0.05$.

5.3 Results and Discussion

In this section present experimental results that answer the four research questions presented at the beginning of this chapter.

5.3.1 Blending of CNN and LSTM networks

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 Fully Connected and other with LSTM layer to assess the impact of using both spatial and sequential features. Table 5.1 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+FC 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. Still in the same experiment, Table 5.2 shows UAR numbers related to emotion for different models, we can note some emotions are harder to classify than others. For instance, *Anger* and *Sadness* are the easiest.

[1] The UAR metric is the sum of the recalls per class divided by the number of classes.

5.3.2 Feature Transference

The next set of experiments is devoted to answer RQ2. We evaluate diverse feature transference approaches. Table 5.3 and 5.4 show 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, due the high variance in results for each target dataset, the best approach varies greatly depending on the target domain. There are some reasons for that like domain divergence and weight dataset(i.e., domain). A better analysis on that is provided in Chapter 6.

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.

Dataset	SVM−IS	CNN+FC	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 5.1: UAR numbers related to datasets for different models. No domain adaptation is performed.

Dataset	SVM−IS	CNN+FC	CNN+LSTM
Anger	.817	.733	.758
Disgust	.172	.294	.286
Fear	.261	.304	.357
Happiness	.359	.316	.328
Sadness	.783	.775	.773
Surprise	.277	.325	.324

Table 5.2: UAR numbers related to emotions for different models. No domain adaptation is performed.

	UAR	Gains over TT - S1							
Target	TT	A1	A2	A3	A4	A5	A6	A7	A8
AFEW	.288	**.121**	.101	.047	**.120**	.115	.045	**.121**	.116
Emo-DB	.614	.047	**.117**	.088	.051	.052	.102	.057	.086
EMOVO	.518	-.095	.053	.014	-.061	-.089	-.071	.034	.014
eNTERF	.441	.032	.133	.114	.061	.032	**.153**	.087	.045
IEMOCAP	.682	.004	.004	-.002	-.009	-.003	.003	**.017**	**.017**
RML	.623	-.014	.032	.041	.005	-.002	**.073**	.035	.028
Average	–	.016	.073	.050	.028	.017	.051	.058	.053
Std.	–	.072	.051	.044	.063	.068	.078	.039	.041

Table 5.3: Different feature transference approaches for scenario S1. Numbers in bold indicate the highest gains for each target domain considering scenarios S1 and S2.

	UAR	Gains over TT - S2							
Target	TT	A1	A2	A3	A4	A5	A6	A7	A8
AFEW	.288	.042	.015	.024	.052	.019	.007	.103	.029
Emo-DB	.614	-.365	.093	.064	.083	.050	.068	.065	.083
EMOVO	.518	-.372	.093	**.109**	-.060	-.041	-.044	.008	-.017
eNTERF	.441	-.353	.027	-.015	-.016	-.034	.002	-.026	-.037
IEMOCAP	.682	-.363	.003	-.015	-.016	-.034	.003	-.026	-.035
RML	.623	-.518	**.074**	.062	-.085	-.145	.054	-.087	-.143
Average	–	-.321	.051	.038	-.007	-.031	.015	.006	-.020
Std.	–	.188	.041	.049	.064	.067	.040	.069	.076

Table 5.4: Different feature transference approaches for scenario S2. Numbers in bold indicate the highest gains for each target domain considering scenarios S1 and S2.

5.3.3 Characteristics that affects the accuracy of the model

The next set of experiments is devoted to answer RQ3. Table 5.5 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 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. The probable explanation is that only RML has many languages, so as more diversity in pre-training as better for recognize emotions in a dataset with a diversity of languages. Another support for this explanation is that for Emo-DB (German dataset) and EMOVO (Italian dataset) accuracies are the smallest. 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.

5.3.4 Our multi-domain network with a well defined target

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 solutions for speech emotion recognition focused only one dataset. Table 5.6 shows UAR numbers obtained by SVM-IS. Different from Table 5.1, now the comparison are in results achieved training SVM-IS using only samples from the target dataset. The table also shows UAR numbers obtained by our multi-domain network. As can be seen, our multi-domain network outperformed SVM-IS in all target domains considered in the study. Gains are statistically significant, and range from 4.3% to 78.6%, depending on the target domain.

The second highest gain is 18.4% indicating an outlier behavior for AFEW but this has

		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$^{\uparrow}$	.440$^{\uparrow}$	.469$^{\uparrow}$	.517$^{\uparrow}$	.304$^{\uparrow}$	**.306**$^{\uparrow}$	**.299**$^{\uparrow}$	.283$^{\downarrow}$
	– EMOVO	.380$^{\uparrow}$	.442$^{\uparrow}$	.473$^{\uparrow}$	.561$^{\uparrow}$	.307$^{\uparrow}$	.289$^{\bullet}$	.262$^{\downarrow}$	.322$^{\uparrow}$
	– eNTERFACE	**.390**$^{\uparrow}$	.464$^{\uparrow}$	.487$^{\uparrow}$	.566$^{\uparrow}$	.284$^{\downarrow}$	.291$^{\bullet}$	.269$^{\downarrow}$	.314$^{\uparrow}$
	– IEMOCAP	.315$^{\bullet}$	**.514**$^{\uparrow}$	**.572**$^{\uparrow}$	**.625**$^{\uparrow}$	.237$^{\downarrow}$	.272$^{\downarrow}$	.280$^{\downarrow}$	.275$^{\downarrow}$
	– RML	.366$^{\uparrow}$	.424$^{\uparrow}$	.487$^{\uparrow}$	.539$^{\uparrow}$	**.314**$^{\uparrow}$	**.298**$^{\uparrow}$	.287$^{\downarrow}$	**.332**$^{\uparrow}$
Emo-DB	All	.643	.685	.668	.645	.389	**.671**	.653	.665
	– AFEW	.725$^{\uparrow}$	.830$^{\uparrow}$	**.835**$^{\uparrow}$	**.843**$^{\uparrow}$	**.397**$^{\uparrow}$	.652$^{\downarrow}$	.644$^{\downarrow}$	.648$^{\downarrow}$
	– EMOVO	.688$^{\uparrow}$	.792$^{\uparrow}$	.789$^{\uparrow}$	.786$^{\uparrow}$	.382$^{\downarrow}$	**.667**$^{\bullet}$	.658$^{\bullet}$	.638$^{\downarrow}$
	– eNTERFACE	.689$^{\uparrow}$	.775$^{\uparrow}$	.780$^{\uparrow}$	.767$^{\uparrow}$	.393$^{\bullet}$	.620$^{\downarrow}$	**.668**$^{\uparrow}$	.662$^{\bullet}$
	– IEMOCAP	**.798**$^{\uparrow}$	**.856**$^{\uparrow}$	**.840**$^{\uparrow}$	.824$^{\uparrow}$	.372$^{\downarrow}$	.653$^{\downarrow}$	.639$^{\downarrow}$	.660$^{\downarrow}$
	– RML	.692$^{\uparrow}$	.762$^{\uparrow}$	.748$^{\uparrow}$	.778$^{\uparrow}$	**.401**$^{\uparrow}$	.655$^{\downarrow}$	.658$^{\uparrow}$	**.683**$^{\uparrow}$
EMOVO	All	.469	.545	.525	.496	.325	.566	**.574**	.487
	– AFEW	**.635**$^{\uparrow}$	**.691**$^{\uparrow}$	**.716**$^{\uparrow}$	**.735**$^{\uparrow}$	.332$^{\uparrow}$	.542$^{\downarrow}$	**.571**$^{\bullet}$	**.547**$^{\uparrow}$
	– Emo-DB	.566$^{\uparrow}$	.664$^{\uparrow}$	.675$^{\uparrow}$	.664$^{\uparrow}$	.320$^{\bullet}$	.567$^{\bullet}$	.549$^{\downarrow}$	.528$^{\uparrow}$
	– eNTERFACE	.566$^{\uparrow}$	.655$^{\uparrow}$	.671$^{\uparrow}$	.695$^{\uparrow}$	.334$^{\uparrow}$	**.595**$^{\uparrow}$	.561$^{\downarrow}$	**.543**$^{\uparrow}$
	– IEMOCAP	.619$^{\uparrow}$	.634$^{\uparrow}$	.641$^{\uparrow}$	.696$^{\uparrow}$	**.351**$^{\uparrow}$	.506$^{\downarrow}$	.495$^{\downarrow}$	.498$^{\uparrow}$
	– RML	.544$^{\uparrow}$	.631$^{\uparrow}$	.648$^{\uparrow}$	.611$^{\uparrow}$	.309$^{\downarrow}$	.560$^{\bullet}$	.563$^{\downarrow}$	**.547**$^{\uparrow}$
eNTERFACE	All	.455	.500	.491	.468	**.247**	.484	**.499**	**.395**
	– AFEW	.645$^{\uparrow}$	**.694**$^{\uparrow}$	**.721**$^{\uparrow}$	.737$^{\uparrow}$	.238$^{\downarrow}$	**.492**$^{\uparrow}$	.490$^{\downarrow}$	**.397**$^{\bullet}$
	– Emo-DB	.538$^{\uparrow}$	.602$^{\uparrow}$	.642$^{\uparrow}$	.644$^{\uparrow}$	.231$^{\downarrow}$	**.499**$^{\uparrow}$	.482$^{\downarrow}$	**.391**$^{\bullet}$
	– EMOVO	.632$^{\uparrow}$	.652$^{\uparrow}$	.654$^{\uparrow}$	.681$^{\uparrow}$	**.248**$^{\bullet}$	.476$^{\downarrow}$	.477$^{\downarrow}$	.375$^{\downarrow}$
	– IEMOCAP	**.749**$^{\uparrow}$	**.696**$^{\uparrow}$	**.711**$^{\uparrow}$	**.751**$^{\uparrow}$	**.244**$^{\bullet}$	.481$^{\bullet}$	.483$^{\downarrow}$	.381$^{\downarrow}$
	– RML	.624$^{\uparrow}$	.625$^{\uparrow}$	.639$^{\uparrow}$	.674$^{\uparrow}$	.233$^{\downarrow}$	.471$^{\downarrow}$	.472$^{\downarrow}$	384$^{\downarrow}$
IEMOCAP	All	.685	.685	.681	.676	.441	**.684**	.672	**.671**
	– AFEW	**.780**$^{\uparrow}$	**.762**$^{\uparrow}$	**.771**$^{\uparrow}$	**.783**$^{\uparrow}$	**.470**$^{\uparrow}$	**.688**$^{\bullet}$	.671$^{\bullet}$	.656$^{\downarrow}$
	– Emo-DB	.739$^{\uparrow}$	.722$^{\uparrow}$	.737$^{\uparrow}$	.741$^{\uparrow}$	.435$^{\downarrow}$	**.686**$^{\bullet}$	.669$^{\bullet}$	.649$^{\downarrow}$
	– EMOVO	.756$^{\uparrow}$	.746$^{\uparrow}$	.751$^{\uparrow}$	.762$^{\uparrow}$	.456$^{\uparrow}$	**.686**$^{\bullet}$	**.680**$^{\uparrow}$	**.665**$^{\bullet}$
	– eNTERFACE	.765$^{\uparrow}$	.740$^{\uparrow}$	.755$^{\uparrow}$	.772$^{\uparrow}$	.427$^{\downarrow}$	**.680**$^{\bullet}$	**.683**$^{\uparrow}$	**.667**$^{\bullet}$
	– RML	.755$^{\uparrow}$	.735$^{\uparrow}$	.746$^{\uparrow}$	.764$^{\uparrow}$	.459$^{\uparrow}$	.671$^{\downarrow}$	**.681**$^{\uparrow}$	.659$^{\downarrow}$
RML	All	**.615**	.643	.649	.626	.301	**.669**	**.662**	**.570**
	– AFEW	.461$^{\downarrow}$	**.733**$^{\uparrow}$	**.760**$^{\uparrow}$	**.738**$^{\uparrow}$	**.318**$^{\uparrow}$	.656$^{\downarrow}$	.644$^{\downarrow}$	.562$^{\downarrow}$
	– Emo-DB	.485$^{\downarrow}$	.690$^{\uparrow}$	.687$^{\uparrow}$	.655$^{\uparrow}$	.297$^{\bullet}$	.650$^{\downarrow}$	**.664**$^{\bullet}$	.543$^{\downarrow}$
	– EMOVO	.475$^{\downarrow}$	.696$^{\uparrow}$	.721$^{\uparrow}$	.675$^{\uparrow}$	.287$^{\downarrow}$	.653$^{\downarrow}$	.653$^{\downarrow}$	.555$^{\downarrow}$
	– eNTERFACE	.453$^{\downarrow}$	**.729**$^{\uparrow}$	.717$^{\uparrow}$	.705$^{\uparrow}$	.302$^{\bullet}$	.648$^{\downarrow}$	**.660**$^{\bullet}$	.557$^{\downarrow}$
	– IEMOCAP	.543$^{\downarrow}$	**.743**$^{\uparrow}$	.748$^{\uparrow}$	.695$^{\uparrow}$	.298$^{\bullet}$	.656$^{\downarrow}$	.656$^{\downarrow}$	.533$^{\downarrow}$

Table 5.5: 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.

Target	SVM-IS	CNN+LSTM	Gain
AFEW	.350	.625	.786
Emo-DB	.797	.856	.074
EMOVO	.692	.735	.062
eNTERFACE	.634	.751	.184
IEMOCAP	.751	.783	.043
RML	.721	.760	.054

Table 5.6: UAR numbers for SVM-IS and CNN+LSTM with a specific target.

a quick explanation. The work of [Schuller et al., 2010] assumes an audio without noising and AFEW is a dataset composed of audio segments extracted from movies, thus this is not a good scenario for feature extraction. In addition, this shows that our multi-domain network is not so sensitive for noise in audio.

Chapter 6

Divergence Analysis

The first part of this chapter tries to explain why some emotions are easier than others in a classification process. In addition, to understand why some datasets have a positive and others a negative impact, the second part of this chapter is focused on analysis how divergent are datasets to each other. For both parts, it is used the KL divergence presented in Section 2.6.

6.1 Emotions

In this section, we want to analyse how divergent the signals in raw audio are among emotions. To get this, the following approach was applied: we calcutate the KL divergence for each audio time frame in pairs of emotions in a scheme of one emotion against the others. Thus, we will have six graphs showing how much divergent one emotion is in regard the others. Note that with this approach, the size of each emotion dataset is not a problem because we are analysing the signal distribution over time and the divergence calculated is between these distributions.

6.1.1 Pre-processing

Firstly, it is applied the same pre-processing as described in Section 4.1. Then, to calculate the KL divergence between the emotions A and B, the following steps are applied:

1. For each time frame, considering only samples from the emotions A and B, get the smallest and highest value.
2. Split this interval into 10 blocks equally spaced.
3. For each sample from each emotion, separately, it is checked in which block the value fits into. We count plus one for that block.

After these steps, for each time frame in each emotion we have a vector of size 10 representing an accounting of how many emotions fit into each range of values. Finally, the

KL divergence is calculated for each time step between emotions A and B resulting in a vector of size 54.000 positions.

6.1.2 Analysis

To analyse the divergence among emotions over time, we plot a graph for each emotion showing the divergence of other emotions in regards to it. The Figures 6.1-6.6 are the graphs for emotions *anger, disgust, fear, happiness, sadness,* and *surprise*, respectively.

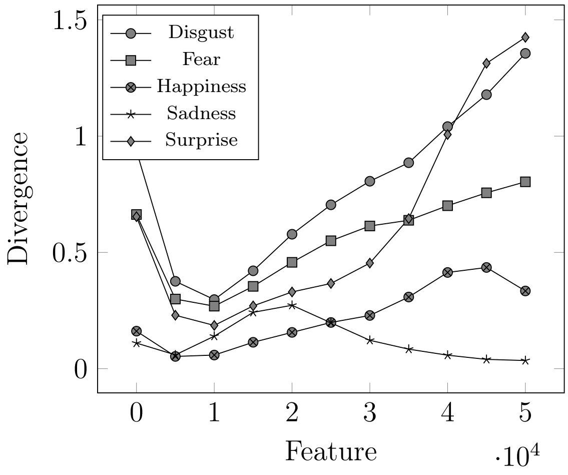


Figure 6.1 Divergence related to *Anger* emotion

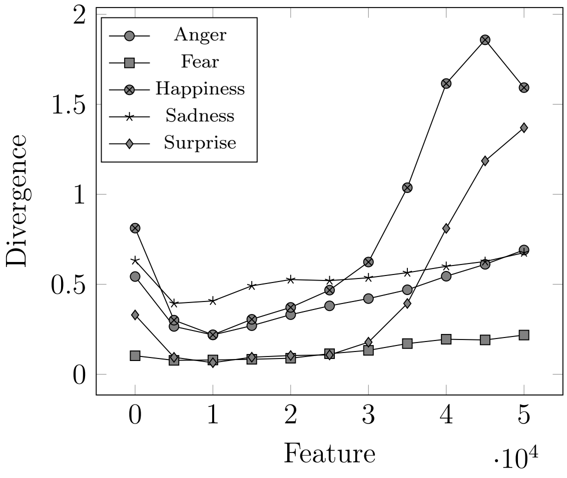


Figure 6.2 Divergence related to *Disgust* emotion

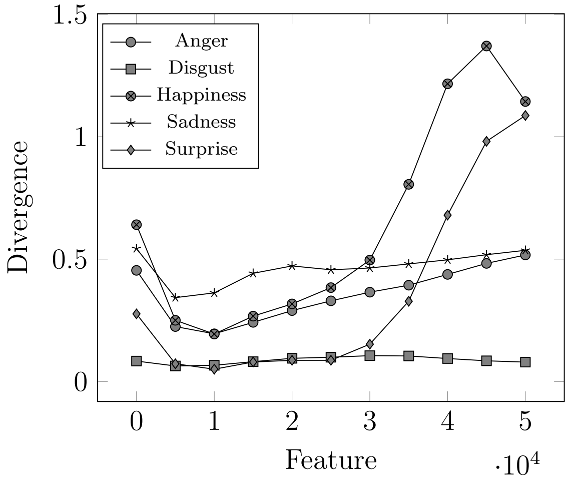


Figure 6.3 Divergence related to *Fear* emotion

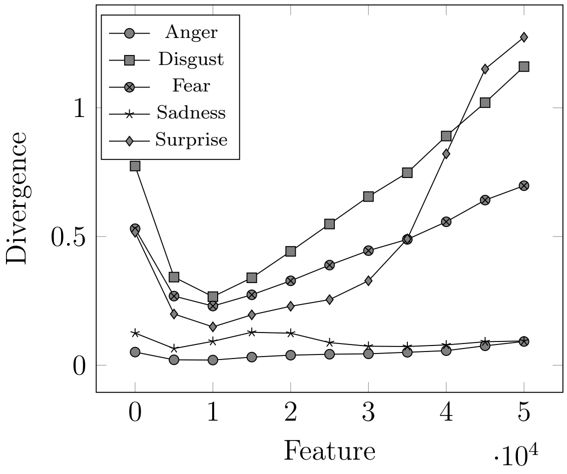


Figure 6.4 Divergence related to *Happiness* emotion

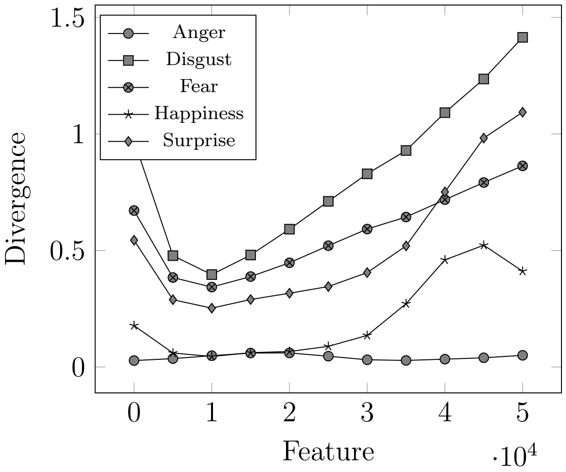


Figure 6.5 Divergence related to *Sadness* emotion

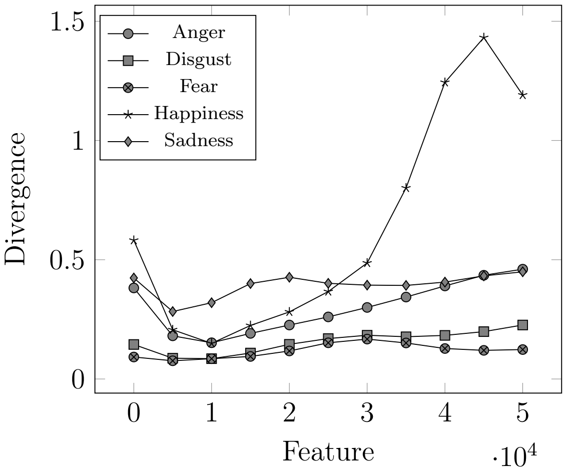


Figure 6.6 Divergence related to *Surprise* emotion

It is easy to realize that there are some emotions that are equally divergent in regards to other emotions. For example, *Anger*, Figure 6.1, and *Sadness*, Figure 6.5, are equally divergent to the other emotions over time, in addition, they have a low divergence to each other. The same analysis is applied between *Disgust*, Figure 6.2, and *Fear*, Figure 6.3. In addition, *Anger* and *Sadness* have similar divergence behavior over time in regards to *Disgust* and *Fear*, the same is applied to *Disgust* and *Fear*.

Analysing emotions *Happiness*, Figure 6.4, and *Surprise*, Figure 6.6, we note *Happiness* is much divergent in regards to *Disgust, Fear,* and *Surprise* in the end of audios. In addition, we note the same behavior for *Surprise* in all other emotions.

The above analysis seem to have connection with results obtained in Table 5.2. In our experiments, for all models, the emotions *Anger* and *Sadness* were the easiest to recognize. This result meet the above analysis that claims *Anger* and *Sadness* have similar divergence and behavior related to the other emotions. It is important to note when two emotions have low divergence to each other this does not mean they have similar signal values, but they have similar signal value distribution over time. For instance, lets suppose emotion A has always high amplitude signal over time and emotion B has always low amplitude. As they have the same behavior over time, they have low divergence between them. KL divergence returns how much information we lost if emotion B distribution is used to estimate the emotion A distribution. With this explanation, now this result makes more sense, since *Anger* and *Sadness* are too different to each other for humans.

6.2 Datasets

In this section, the purpose is to analyse the divergence among datasets. The first approach was to analyse it as we do in Section 6.1, but unfortunately taking a look in graphs, we can only clain that datasets are so much divergent to each other. For this reason, another approach was attempted.

The new approach is to calculate the KL divergence in regards to emotion distribution by dataset or scenario.

6.2.1 Analysis

Based on the Table 5.5, the Table 6.1 was generated. For now, the values representing divergence between source and target. The field *pair* represents divergence between removed and target dataset, *S1* and *S2* represent divergence between source and target dataset. From this table, we are going to analyse results from different views to answer the following questions:

1. What are the most divergent datasets? Are there any explanation for this?
2. Are there some relationship between divergence and weight dataset?
3. Are there some straight relationship between paired divergence and scenario divergence(i.e., S1 and S2)?
4. As we have an unbalanced source dataset, in regards of emotion distribution, can this divergence analysis explain some experimental results?

To answer the first question, for all target scenarios, the two datasets most divergent are *Emo-DB* and *EMOVO*. The main reason is they do not have any sample of *Surprise* and *Happiness* emotions, respectively, as previously presented in Table 4.2. Thus, it is hard to use any of them to approximate the emotion distribution of any other dataset.

For the second question, as can there are interaction between weight dataset and paired divergence, we isolate only scenarios where target dataset were removed from *S1*, so we analyse the impact between *S1* and *S2* because the divergence between target and the removed dataset is zero. The Figure 6.7 shows divergence increases as the weight of removed dataset increases. Based on this figure, we can see a linear relationship between these factors. On the other hand, this seems to have a limit because for IEMOCAP dataset that represents 38% we can note a trending for not increasing linearly. Since ideal scenarios are those with many datasets, so weights larger than this are not common for our purpose.

For the third question, we have in Figure 6.8 for each dataset a graph showing the variance of divergence that it causes for all targets. We decided for this view because it removes the weight factor from each graph. All datasets could be able to have negative variance(i.e., when this dataset is removed, the divergence between source and target datasets is lower) for at least one target dataset, i.e., for all datasets it is not a good idea to have all other datasets in source. We also can note that there is a trend where variance of divergence descrease when paired divergence increases. It is important to note that there are other factors that

			KL divergence			Impact		
Target	Source	Weight	Pair	S1	S2	S2/S1	S1/All-S1	S2/All-S2
AFEW	All			0.0169●	0.0221●	31.21%●		
	– Emo-DB	.062●	3.4872	0.0173	0.0233	34.58%	2.55%●	5.19%●
	– EMOVO	.072	**5.8745**	0.0249	0.0338	35.85%	47.82%	53.04%
	– eNTERFACE	.225	0.0374●	**0.0418**	**0.0608**	45.62%	**147.81%**	**175.03%**
	– IEMOCAP	**.380**	0.8199	0.0241	0.0370	**53.55%**	42.92%	67.25%
	– RML	.140	0.0515	0.0303	0.0419	38.12%	79.79%	89.26%
Emo-DB	All			0.1324	0.1437●	8.53%●		
	– AFEW	.122	0.1523●	0.1329	0.1463	10.06%	0.37%	1.79%
	– EMOVO	.072●	**5.8437**	0.1322	0.1454	9.96%	-0.18%●	1.13%
	– eNTERFACE	.225	0.2087	0.1397	0.1587	13.55%	5.50%	10.37%
	– IEMOCAP	**.380**	0.8775	**0.1829**	**0.2100**	**14.82%**	**38.11%**	**46.11%**
	– RML	.140	0.2504	0.1303●	0.1446	11.00%	-1.63%	0.61%●
EMOVO	All			0.3207	0.3655	13.97%●		
	– AFEW	.122	0.2662	0.3320	0.3879	16.82%	3.54%	6.13%
	– Emo-DB	.062●	**6.4584**	0.3197	0.3679	15.09%	-0.32%	0.66%
	– eNTERFACE	.225	0.1768●	**0.4057**	**0.4952**	**22.06%**	**26.50%**	**35.48%**
	– IEMOCAP	**.380**	1.8363	0.1782●	0.2069●	16.11%	-44.42●%	-43.38%●
	– RML	.140	0.2042	0.3653	0.4321	18.26%	13.93%	18.22%
eNTERFACE	All			0.0779	0.1366	75.37%●		
	– AFEW	.122	0.0393	0.0863	0.1673	93.91%	10.76%	22.47%
	– Emo-DB	.062●	**4.8119**	0.0783	0.1440	83.87%	0.54%	5.41%
	– EMOVO	.072	4.7225	0.0996	0.1889	89.74%	27.86%	38.34%
	– IEMOCAP	**.380**	1.1807	0.0024●	0.0059●	**147.49%**	-96.94%●	-95.68%●
	– RML	.140	0.0054●	**0.1114**	**0.2259**	102.88%	**42.99%**	**65.42%**
IEMOCAP	All			**0.2355**	0.4716	100.26%●		
	– AFEW	.122	0.3311●	0.2241	**0.5117**	128.32%	-4.84%	**8.50%**
	– Emo-DB	.062●	1.1681	0.2349	0.5030	114.13%	**-0.26%**	6.65%
	– EMOVO	.072	**7.0094**	0.2081	0.4475	115.01%	-11.62%	-5.11%
	– eNTERFACE	**.225**	0.5217	0.1721●	0.4453●	**158.77%**	-26.94%●	-5.59%●
	– RML	.140	0.5680	0.1969	0.4494	128.27%	-16.41%	-4.72%
RML	All			0.1002	0.1383	38.04%●		
	– AFEW	.122	0.0538	0.1095	0.1591	45.31%	9.28%	15.04%
	– Emo-DB	.062●	**5.5137**	0.0989	0.1394	40.97%	-1.31%	0.78%
	– EMOVO	.072	5.4586	0.1217	0.1753	43.96%	21.54%	26.75%
	– eNTERFACE	.225	0.0054●	**0.1637**	**0.2610**	59.51%	**63.38%**	**88.80%**
	– IEMOCAP	**.380**	1.1953	0.0133●	0.0222●	**67.70%**	-86.76%●	-83.91%●

Table 6.1: KL divergence ablation analysis. The highest value for each column and target is highlighted in bold. Symbol ● indicates the lowest value.

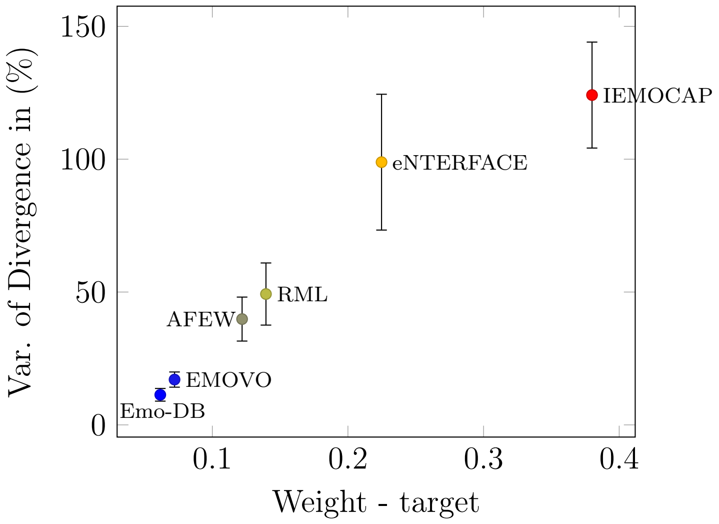


Figure 6.7: % of variance in KL divergence with std. deviation when target dataset is removed from source dataset

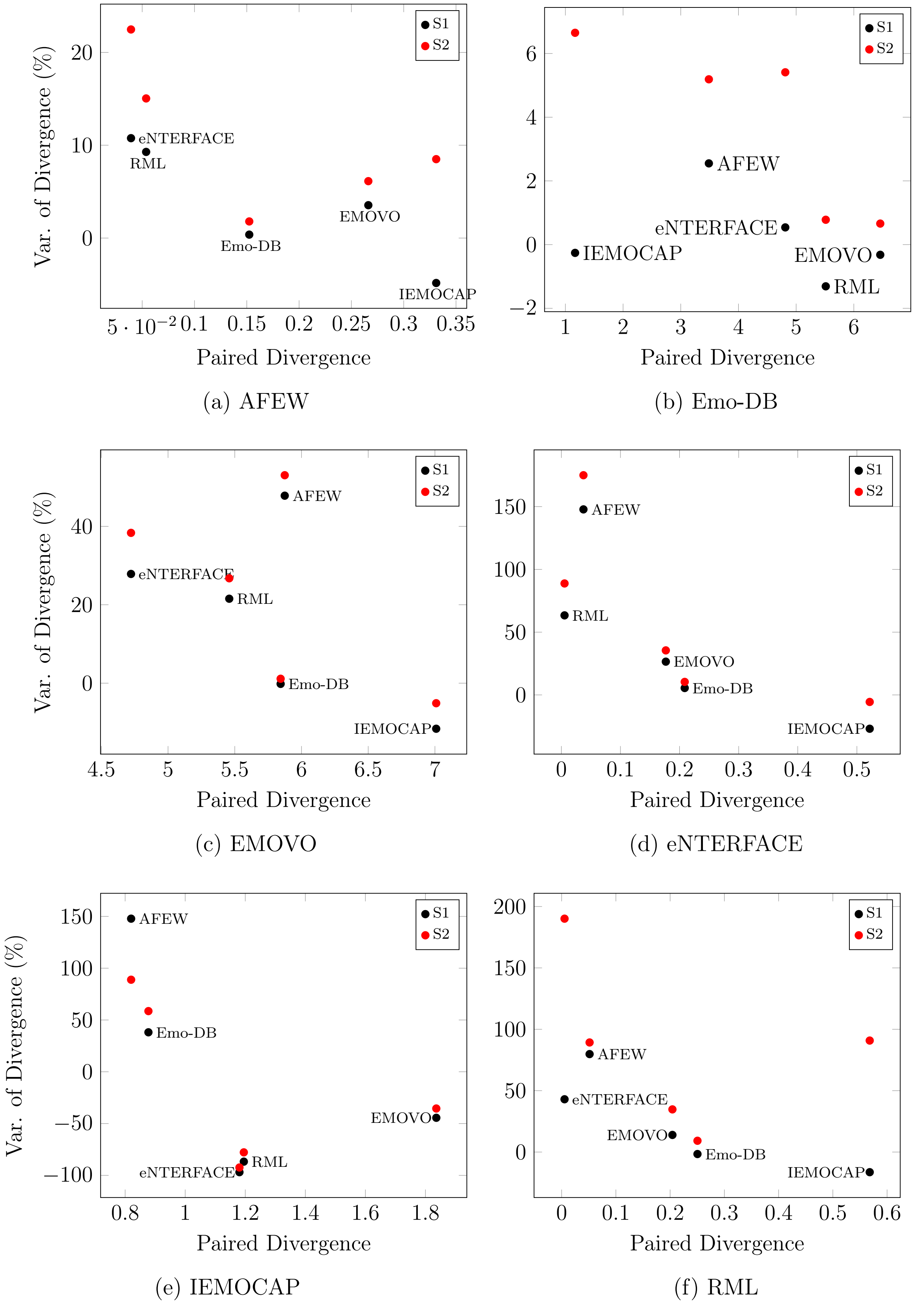
influence the Variance of Divergence beyond Weight of removed dataset like paired divergence in regards to signal audio distribution.

To answer the last question, we are going to try explain some results obtained in Chapter 5, more specifically in Table 5.5, using some insights that we had answering aforementioned questions. In this case, we are going to analysis only scenario S1-A3 and S1-A4 where target dataset is present in source and pre-trained convolutions and/or LSTMs are freezed in fine-tuning, i.e., the features learnt in pre-training are not modified in fine-tuning process. In addition, for paired divergence we are also going to ignore Emo-DB and EMOVO datasets because they present a high paired divergence just because they do not have all six emotions. This will not affect directly our analysis since they are small datasets in pre-training phase.

For AFEW dataset, the best result occurs when IEMOCAP is not present in source dataset for both scenarios. IEMOCAP is the most divergent dataset in regards to AFEW beyond being the largest dataset in source, thus it impacts negativaly the features learnt in Convolutions and LSTM layers for the target dataset. In addition, it has the smallest impact when removed from source.

If we apply the same criteria in EMOVO, the best result would be removing IEMOCAP but in experiments the best was removing AFEW. Although IEMOCAP has the highest divergence in regards to EMOVO and results in a negative impact, it represents the largest dataset in source. Thus smaller source can suffer from overfitting and features learnt in pre-training may not fit well to EMOVO.

When we analyse eNTERFACE and RML, we see a match between divergence analysis and experimental results since the best result is removing IEMOCAP to both. This time, even IEMOCAP representing the largest dataset in source, when it is removed, the source reduces its divergence in almost 87% and 97%, having divergence close to zero, for eNTERFACE and



(a) AFEW

(b) Emo-DB

(c) EMOVO

(d) eNTERFACE

(e) IEMOCAP

(f) RML

Figure 6.8: Variance of divergence for each removed dataset in all target scenarios

RML respectively. Probably this new divergence between source and target compensates the reduce of pre-training dataset.

In IEMOCAP dataset we see that previous insights make sense for it, but the best result in experiments occurs when AFEW is removed. If we do not take in account of AFEW, the previous analysis fit perfectly here. Probably, another factor (e.g., divergence over raw data or over audio recording condition since AFEW is a natural dataset and IEMOCAP is an acted dataset) makes AFEW the best dataset to be removed from source.

Chapter 7

Conclusion

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 that 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 78.6% when compared with feature extraction approaches for speech emotion recognition beyond being more robust to natural audios as noted in AFEW.

Due to the wide range of gain, we also performed a divergence analysis over emotions and later over datasets to get evidences that helped us to explain some experimental results in ablation. As result of this analysis, we found out there is a strong trade-off between paired divergence and the weight of removed dataset, i.e., sometimes even with high paired divergence the result is better with that dataset because it represents a large dataset in source, the opposite is also true.

Bibliography

Abdel-Wahab, M. and Busso, C. (2015). Supervised domain adaptation for emotion recognition from speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5058--5062.

Banziger, T., Grandjean, D., and Scherer, K. (2009). Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert). *Emotion*, 9(5):691--704.

Batliner, A., Steidl, S., Schuller, B. W., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., and Amir, N. (2011). Whodunnit - searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 25(1):4--28.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1-2):151--175.

Bergstra, J., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., Desjardins, G., Goodfellow, I., Bergeron, A., Bengio, Y., and Kaelbling, P. (2011). Theano: Deep learning on gpus with python.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *European Conference on Speech Communication and Technology*, pages 1517--1520.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335--359.

Busso, C., Parthasarathy, S., mania, A. B., Abdel-Wahab, M., Sadoughi, N., and Provost, E. M. (2017). MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67--80.

Caramiaux, B., Bevilacqua, F., and Schnell, N. (2010). Study on gesture-sound similarity. In *Music and Gesture*, pages 1--1.

Chollet, F. (2015). keras. `https://github.com/fchollet/keras`.

Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). EMOVO corpus: an italian emotional speech database. In *International Conference on Language Resources and Evaluation*, pages 3501--3504.

Cowie, R. and Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5--32.

Deng, J., Xu, X., Zhang, Z., Frühholz, S., and Schuller, B. W. (2017). Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 24(4):500--504.

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. W. (2014a). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068--1072.

Deng, J., Zhang, Z., Marchi, E., and Schuller, B. W. (2013). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *ACII Association Conference on Affective Computing and Intelligent Interaction*, pages 511--516.

Deng, J., Zhang, Z., and Schuller, B. W. (2014b). Linked source and target domain subspace feature transfer learning - exemplified by speech emotion recognition. In *International Conference on Pattern Recognition*, pages 761--766.

der Maaten, L. V. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579--2605.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34--41.

Drolet, M., Schubotz, R., and Fisher, J. (2012). Authenticity affects the recognition of emotions in speech: Behavioral and fmri evidence. *Cognitive Affective & Behavioral Neuroscience*, 12(1):140--150.

Eyben, F., Schuller, B. W., and Rigoll, G. (2012). Improving generalisation and robustness of acoustic affect recognition. In *International Conference on Multimodal Interaction*, pages 517--522.

Gao, X., Wang, X., Li, X., and Tao, D. (2011). Transfer latent variable model based on divergence analysis. *Pattern Recognition*, 44(10):2358--2366.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Annual Conference of the International Speech Communication Association*, pages 223--227.

Helén, M. and Virtanen, T. (2007). A similarity measure for audio query by example based on perceptual coding and compression. In *Proc. 10th Int. Conf. Digital Audio Effects (DAFX)*.

Helén, M. and Virtanen, T. (2009). Audio query by example using similarity measures between probability density functions of features. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):179303.

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49--56.

Huang, Y., Wu, A., Zhang, G., and Li, Y. (2016). Speech emotion recognition based on deep belief networks and wavelet packet cepstral coefficients. *International Journal of Simulation: Systems, Science and Technology*, 17(28):28--1.

Huang, Z., Dong, M., Mao, Q., and Zhan, Y. (2014). Speech emotion recognition using CNN. In *ACM International Conference on Multimedia*, pages 801--804.

Huang, Z., Xue, W., Mao, Q., and Zhan, Y. (2017). Unsupervised domain adaptation for speech emotion recognition using pcanet. *Multimedia Tools and Applications*, 76(5):6785--6799.

Johnstone, T., van Reekum, C., Oakes, T., and Davidson, R. (2006). The voice of emotion: an fmri study of neural responses to angry and happy vocal expressions. *Social Cognitive and Affective Neuroscience*, 1(3):242--249.

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

Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3687--3691.

Koolagudi, S., Barthwal, A., Devliyal, S., and Rao, K. (2012). Real life emotion classification from speech using gaussian mixture models. In *International Conference Contemporary Computing*, pages 250--261.

Koolagudi, S. and Rao, K. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99--117.

Koolagudi, S., Ray, S., and Rao, K. (2010). Emotion classification based on speaking rate. In *International Conference Contemporary Computing*, pages 316--327.

Mandel, M. I. and Ellis, D. (2005). Song-level features and support vector machines for music classification. In *ISMIR*, volume 2005, pages 594--599.

Mao, Q., Xue, W., Rao, Q., Zhang, F., and Zhan, Y. (2016). 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*, pages 2608--2612.

Marchi, E., Eyben, F., Hagerer, G., and Schuller, B. W. (2016). Real-time tracking of speakers' emotions, states, and traits on mobile platforms. In *Annual Conference of the International Speech Communication Association*, pages 1182--1183.

Martin, O., Kotsia, I., Macq, B. M., and Pitas, I. (2006). The enterface'05 audio-visual emotion database. In *International Conference on Data Engineering Workshops*, page 8.

Nisius, B., Vogt, M., and Bajorath, J. (2009). Development of a fingerprint reduction approach for bayesian similarity searching based on kullback- leibler divergence analysis. *Journal of chemical information and modeling*, 49(6):1347--1358.

Nogueiras, A., Moreno, A., Bonafonte, A., and Mariño, J. B. (2001). Speech emotion recognition using hidden markov models. In *European Conference on Speech Communication and Technology*, pages 2679--2682.

Ooi, C. S., Seng, K. P., Ang, L.-M., and Chew, L. W. (2014). A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858--5869.

Oudeyer, P. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157--183.

Parry, R. and Essa, I. (2007). Phase-aware non-negative spectrogram factorization. *Independent Component Analysis and Signal Separation*, pages 536--543.

Ramakrishnan, S. and Emary, I. E. (2013). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 52(3):1467--1478.

Rao, K., Koolagudi, S., and Reddy, V. (2013). Emotion recognition from speech using global and local prosodic features. *I. J. Speech Technology*, 16(2):143--160.

Sakai, T. (2014). Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3--12.

Schuller, B. W., Batliner, A., Steidl, S., and Seppi, D. (2009a). Emotion recognition from speech: Putting ASR in the loop. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4585--4588.

Schuller, B. W., Steidl, S., and Batliner, A. (2009b). The INTERSPEECH 2009 emotion challenge. In *Annual Conference of the International Speech Communication Association*, pages 312--315.

Schuller, B. W., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., Narayanan, S. S., et al. (2010). The interspeech 2010 paralinguistic challenge. In *Interspeech*, volume 2010, pages 2795--2798.

Schuller, B. W., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2015). Cross-corpus acoustic emotion recognition: Variances and strategies. In *International Conference on Affective Computing and Intelligent Interaction*, pages 470--476.

Seppi, D., Batliner, A., Schuller, B. W., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., and Aharonson, V. (2008). Patterns, prototypes, performance: classifying emotional user states. In *Annual Conference of the International Speech Communication Association*, pages 601--604.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151--161. Association for Computational Linguistics.

Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., and Yu, Y. (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication*, 83:34--41.

Spreckelmeyer, K., Kutas, M., Urbach, T., Altenmuller, E., and Munte, T. (2009). Neural processing of vocal emotion and identity. *Brain and Cognition*, 69(1):121--126.

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

Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, H., and Schuller, B. W. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5688--5691.

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

Velten, E. (1968). A laboratory task for induction of mood states. *Behavior Research & Therapy*, 6:473--482.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. (2008). Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pages 1096--1103.

Williams, C. and Stevens, K. (1972). Emotions and speech: some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4):1238--1250.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B. W., and Rigoll, G. (2013). Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153--163.

Xue, W., Huang, Z., Luo, X., and Mao, Q. (2015). Learning speech emotion features by joint disentangling-discrimination. In *International Conference on Affective Computing and Intelligent Interaction*, pages 374--379.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Annual Conference on Neural Information Processing Systems*, pages 3320--3328.

Zhang, B., Provost, E. M., and Essl, G. (2016a). Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5805--5809.

Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., and Schuller, B. W. (2016b). Facing realism in spontaneous emotion recognition from speech: feature enhancement by autoencoder with LSTM neural networks. In *Annual Conference of the International Speech Communication Association*, pages 3593--3597.