Multi-task learning from Unlabelled Data to Improve Cross Language Speech Emotion Recognition

Hafiz Shehbaz Ali¹, Siddique Latif²

¹Emulation AI
²University of Southern Queensland, Australia
siddique.latif@usq.edu.au

Abstract
Despite the recent progress in deep learning-based speech emotion recognition (SER), the performance of state-of-the-art systems significantly decreases in cross-language settings. The main reason is the lack of generalisation in SER systems due to the unavailability of larger training emotional labelled data in different languages. In this work, we present a novel multi-task learning (MTL) approach to effectively utilise unlabelled data to improve the generalisation as well as the performance of cross-language SER systems. In particular, we propose to use language and domain identification as auxiliary tasks, which facilitates the proposed framework to learn from abundantly available language identification data. We evaluate the proposed model on publicly available datasets in four languages and achieve state-of-the-art performance.

Index Terms: Cross-language, speech emotion recognition, deep learning, unlabelled data.

1. Introduction
Speech emotion recognition (SER) is a mature area of research, which uses acoustic attributes to perform emotion detection using Deep Neural Networks (DNNs) models [1]. It has many applications in call centres, forensic sciences, media retrieval systems, smart cars, healthcare, to name a few [1, 2]. In future, SER systems will redefine the human-machine interactions (HCI) by enabling effective service delivery in many real-life applications.

Current deep learning (DL)-based SER systems achieve improved performance when training and testing data belong to the same language corpus, however, their performance plummets against cross-language settings. This shows that state-of-the-art SER systems lack generalisation, which makes them susceptible to unknown conditions due to different languages. In SER research, most studies focused on performance improvement using a single corpus for training and testing without considering cross-language emotion recognition. Some studies [3, 4] attempt to perform cross-lingual SER by training the models using multiple languages corpora to learn common or generalised feature representation and achieve improved performance. However, such setups require emotionally labelled data in multiple languages. In practice, we have few emotionally labelled datasets for SER, and most of these datasets are in English. This limits the training of the models to learn generalised representations.

Another way to learn generalised representation is to train the DL models in multi-task learning (MTL) setting. This enables the model to learn shared representation by simultaneously solving multiple related tasks. MTL is widely explored in SER and researchers presented various approaches to improve the performance by learning shared generalised representations [5, 6, 7]. Most of these studies focus to improve SER performance within-corpus and cross-corpus settings without performing cross-language SER. In this paper, we present MTL technique by focusing on utilising the unlabelled data to improve cross-language SER performance.

We summarise the contributions of this work as follow:
• We present a novel MTL framework for cross-language SER, which can effectively utilise abundantly available unlabelled data to improve performance.
• We propose to use language and domain identification as auxiliary tasks for which we can inject unlabelled data into the system.
• We evaluate the proposed model on four publicly available datasets in four different languages to perform cross-language SER. Results show that the proposed framework achieves state-of-the-art results compared to the baseline and recent studies.

2. Related Work
Our proposed framework utilises MTL for cross-lingual SER. It uses unlabelled for performance improvement with a CNN-LSTM-based classifier. We, therefore, cover all these aspects in our literature search section.

2.1. Multitask Learning (MTL) for SER
Multi-task learning (MTL) is successfully becoming popular across different applications areas of machine learning including natural language processing [8] and speech recognition and analysis [9, 10], computer vision [11]. It aims to leverage useful information about multiple related tasks contained in the training data and help improve the generalisation and the performance of the DL systems [12]. In contrast to MTL, conventional machine learning (ML) methods are optimised for single task learning (STL) objective function, which ignores important information that can be optionally utilised for the auxiliary tasks to improve the generalisation and performance of the system.

Speech signal contains information about the intended message, speaker, gender, language, and emotion. This leads to the investigation of MTL to improve the performance of speech-based systems. Eyben et al. [7] are the first to show that MTL helps to improve the performance of the SER system compared to single-task training. Xia et al. [5] utilise deep belief network (DBN) for MTL by utilising activation and valence information as secondary tasks for SER. Based on the results, they show that the use of secondary information in auxiliary
tasks helps improve the performance of the primary SER task. Parthasarathy et al. [6] present a multi-task model to jointly learn three emotional attributes including arousal, valence, and dominance from a given speech utterance. The authors empirically show that joint training of the DNNs for learning emotional attributes significantly helps improve the SER performance in contrast to single-task training. Some other studies (e.g., [13, 14]) also explore the use of speakers and gender identification, and naturalness classification as auxiliary tasks to improve the performance of primary emotion classification task. These studies show that MTL improves the generalisation as well as the performance of SER. Our approach is also motivated by these findings, however, we aim to improve the performance of multi-language emotion recognition.

Some studies also evaluate the MTL approaches for cross-corpus emotion recognition. Zhang et al. [15] use an MTL approach to investigate the influence of the domain, corpus, and gender on cross-corpus emotion recognition systems. The authors show that the performance of the cross-corpus SER system increases with the rising number of emotional corpora used for training. Sagha et al. [16] use MTL to improve the performance of multi-language emotion recognition by utilising language and gender classification as auxiliary tasks and achieved improved results in English and Japanese languages. Similarly, Latif et al. [9] show that utilisation of additional data for speaker and gender identification auxiliary tasks can improve the performance of the primary task. In contrast, we utilise unlabelled data for auxiliary tasks of language identification and domain identification in our proposed MTL framework. In domain identification, we classify labelled and unlabelled data.

2.2. Cross-Language SER

The performance of state-of-the-art SER systems significantly drops in cross-language SER settings as highlighted in [17, 18, 19, 20]. These studies pointed out the need for in-depth research to build robust SER systems for cross-language SER. Various other studies also explore different DL techniques to design generalised and robust SER systems. For instance, Neumann et al. [21] present an attentive CNN classifier for cross-language SER using French and English languages. They evaluate the proposed model for binary arousal/valence classification and showed that fine-tuning the model on the partial target data can improve cross-language SER performance. Li et al. [22] explore a combination of different speech features and speaker normalisation methods to improve multi-language SER. In [3], the authors utilise DBNs for cross-corpus SER. They empirically show that the DBNs can learn the generalised representation of multiple languages of training data to improve cross-language SER. Another recent study [23] explores the generative adversarial network (GANs) for learning language invariant features to improve cross-language SER. They evaluate the proposed model on four different language corpora and show improvement over the support vector machines (SVMs)-based on baseline results. In contrast to the studies mentioned above, we present a novel multi-task learning framework with language and domain identification as auxiliary tasks. We propose to use unlabelled data in multiple languages for proposed auxiliary tasks to improve the generalisation of cross-lingual SER.

2.3. CNN-RNNs for SER

Convolutional neural networks (CNNs) combined with recurrent neural networks (RNNs) are one of the most widely used deep models in speech emotion recognition (SER) [2]. In CNN-RNN-based approaches, features are learned through CNNs and RNN architectures (i.e., LSTM/GRU) are exploited for contextual modelling. CNN-RNN models can effectively learn temporal relationships suitable for SER [2]. Trigeorgis et al. [24] propose an end-to-end CNN-LSTM model to learn context-aware emotional representation. They empirically show that CNN-LSTM performs better than the traditional techniques. Latif et al. [25] utilise CNN-LSTM for learning emotional features learning from raw speech. They found that CNNs act like a data-driven filter and automatically learn emotional contexts. Some other studies [26, 27] also explore CNN-RNN architectures and show that it is a better choice for SER compared to using CNN or RNNs individually. In contrast, we are using CNN-LSTM in MTL scenarios, which helps the proposed framework to learn emotional contexts and generalised discriminative features for emotion classification.

3. Proposed Model

![Figure 1: Illustration of our proposed multitask framework for multi-language SER. We inject unlabelled data for language and domain identification.](image)

3.1. Multi-Task (MTL) Framework

We proposed an MTL framework for SER that use language identification and domain classification as auxiliary tasks. Figure 1 illustrate the proposed model. It consists of a CNN block that acts as a shared network among three classification networks for emotion, language, and domain identification.

**Convolutional neural networks (CNNs) block** is used to extract temporal features from the given spectrograms. Here, we used a deep residual learning network (ResNet [32]) in the CNN block. ResNet has several stacked residual blocks with identical structures and the skip connection for the identity mapping of input ($x$) as shown in Figure. The residual block is defined as:

$$y = F(x, W_i) + x.$$  

(1)

Where $y$ and $x$ represent the output and input layers considered respectively, and $F$ is the stacked non-linear mapping function. Shortcut connections in ResNet perform identity mapping, and their outputs are added to the outputs of the stacked layers. These shortcut connections do not add extra parameters and computational complexity. However, they help improve convergence and do not degrade in performance with an increase in depth. In ResNet, each residual block has a similar structure. Generally, the residual block has two convolutional layers with
Table 1: Selected emotional corpora and non-emotional corpora, and the mapping for emotional corpora to binary valence.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>Utterances</th>
<th>Emotions</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEMOCAP</td>
<td>English</td>
<td>5854</td>
<td>angry, sadness, interested, happiness</td>
<td>neutral, happy</td>
</tr>
<tr>
<td>EMODB [29]</td>
<td>German</td>
<td>404</td>
<td>angry, sad, fear, joy, surprise</td>
<td>neutral, happiness</td>
</tr>
<tr>
<td>EMOCOVO [30]</td>
<td>Italian</td>
<td>588</td>
<td>angry, sad, fear, joy, surprise</td>
<td>neutral, joy, surprise</td>
</tr>
<tr>
<td>URDU [31]</td>
<td>Urdu</td>
<td>23062</td>
<td>angry, sadness</td>
<td>neutral, happy, neutral</td>
</tr>
<tr>
<td>Librispeech</td>
<td>English</td>
<td>28539</td>
<td>neutral, happiness</td>
<td></td>
</tr>
<tr>
<td>Voxforge</td>
<td>German</td>
<td>25665</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>27700</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>31794</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a small $3 \times 3$ filter, as we used in this work. We modified the ResNet architecture and used it for temporal feature extraction. Features extracted by ResNet are given to the LSTM for emotional context learning.

The Long short-term memory (LSTM) [33] network is a special architecture that eliminates the problems of exploding gradient or vanishing in standard RNNs. The architecture of LSTM consists of a recurrent memory block that consists of three multiplicative gates help enable LSTM to model long ranges of temporal relationships. Emotions are context-dependent [25] and contexts are embedded in temporal dimension of speech signal [25]. We use LSTM to learn contextual representation from the output feature of ResNet.

In each classifier, we pass the outputs of LSTM to the dense layer as it transforms them into more discriminative space, which helps the classifier for target prediction.

3.2. Training MTL Framework

Our MTL framework has speech emotion classification as a primary task, language identification and domain classification as auxiliary tasks. We use additional unlabelled data for language identification and domain classifier. The language classifier is trained to identify the language of a given sample, whereas the domain classifier is tasked to classify between labelled and unlabelled data. The multi-task objective function presented in Equation 2 that combines three cross-entropy losses for emotion classification ($\mathcal{L}_e$), language identification ($\mathcal{L}_l$), and domain classification ($\mathcal{L}_d$).

$$ L = \mathcal{L}_e + \alpha (\mathcal{L}_l + \mathcal{L}_d), $$

where $\mathcal{L}_e$, $\mathcal{L}_l$, $\mathcal{L}_d$ represent the standard cross-entropy loss function for emotion, language, and domain classifiers. $\alpha$ is the trade-off parameter between primary and auxiliary tasks. In our proposed model, the shared network, language and domain classifiers are updated for all data, however, the emotion classifier is only updated for data with emotional labels. We tried different values of $\alpha$ (0.1-0.5), however, achieved better performance at $\alpha = 0.3$.

4. Experiment Setting

4.1. Datasets Details

We selected four publicly available and popular emotional corpora which have maximum diversity in languages. These databases are annotated differently, therefore, one way to investigate cross-language SER learning is to consider binary valence classification. We follow [34, 21] valence mapping for categorical emotion. For the auxiliary task of language classification, we use Librispeech and Voxforge as unlabelled data. LibriSpeech dataset is selected for English and four other languages including German, Italian, Spanish, and French data is used from Voxforge data repository. The details of these data and categorical emotional mapping are given in Table 1.

4.2. Pre-processing and Feature Representation

We have represented the speech utterances in the form of spectrograms, which is a popular speech representation used for SER [2]. We compute the spectrograms using a short-time Fourier transform (STFT) with an overlapping Hamming window of size 25 ms with a 10 ms shift. We select the frequency range of 0-8KHz and extract 128 Mel frequency bands. Before spectrogram representation, non-speech intervals at the beginning and end of each utterance are removed as performed in [9]. We select the audio length of 6 seconds for both emotion and language tasks. Therefore, larger utterances are cropped and shorter zero-padded. We empirically tested that increasing the length of speech utterances from 6 seconds does not help significantly improve the performance.

4.3. Model Configuration

In our proposed architecture, we use ResNet a shared network with 8 stacked residual blocks that have shortcut connections from the input layer to the output layer of each block. These shortcut connections shuttle multiple levels of emotional abstractions and help improves the gradient flow through the network [9]. We use batch normalisation before applying the Rectified Linear Unit (ReLU) activation in each convolutional layer. The output from the shared network is fed to the LSTM layer with 128 LSTM units for contextual modelling. We use a dense layer with 256 hidden units for discriminative representation learning before the softmax in each classifier. For the baseline architecture, we implement the single task architecture without language and domain identification.

5. Experiments and Evaluations

In this section, we present the results of different experiments using our proposed model. For baseline results, we perform SER experiments in within-corpus settings. We follow a speaker-independent evaluation scheme for all datasets and results are reported in unweighted average recall rate (UAR). For URDU data, we train the model using 25 speakers and the remaining speakers’ data are for testing with five-fold cross-validation. For other corpora, we use a leave-one-speaker-out evaluation scheme with cross-validation equal to the number of speakers in the respective dataset as per the accepted practice in SER literature [35, 23]. Baseline results are presented in Table 2.

---

http://www.voxforge.org/
Table 2: Baseline results (UAR %) within corpus setting.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>EMO-DB</th>
<th>IEMOCAP</th>
<th>EMOVO</th>
<th>URDU</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAR (%)</td>
<td>81.3</td>
<td>62.1</td>
<td>74.2</td>
<td>83.4</td>
</tr>
</tbody>
</table>

5.1. Cross-Lingual Experiments

In this experiment, we evaluated the proposed model for cross-language SER. We compute our results with the single-task as well as multi-task learning models. We trained these models using IEMOCAP data in the English language and testing is performed on data from other languages including German, Italian, and Urdu. Results are compared with baseline and different studies in Table 3. In [3], authors utilise the deep belief networks (DBNs) for cross-language SER. They evaluated their model for binary valence classification using three different language datasets. In [23] utilise GAN based architecture to learn language-independent emotional representations. We also evaluated different MTL frameworks and the results are reported in Table 3. In [21], authors evaluated MTL CNN on cross-language SER without considering additional unlabelled data. Similarly, DBN [5] (MTL) and LSTM [14] (MTL) also utilised MTL to improve the performance within-corpus settings. In contrast to these studies, we focus to utilise the additional data in MTL to improve the performance by capturing the generalised representations. We consider only language identification as an auxiliary task while computing results without additional unlabelled data. Results are presented in Table 3, which shows that identifying the language in multi-task learning improves the performance compared to the baseline and other studies. In addition, the use of additional data for auxiliary tasks significantly helps the system to achieve better results by learning generalised representations.

Table 3: Comparison of Cross-language SER results in UAR (%).

<table>
<thead>
<tr>
<th>Model</th>
<th>Target datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EMO-DB</td>
</tr>
<tr>
<td>CNN-LSTM (STL) [9]</td>
<td>52.4</td>
</tr>
<tr>
<td>DBN (STL) [3]</td>
<td>54.5</td>
</tr>
<tr>
<td>GAN (STL) [23]</td>
<td>51.7</td>
</tr>
<tr>
<td>CNN (MTL) [21]</td>
<td>53.2</td>
</tr>
<tr>
<td>LSTM (MTL) [14]</td>
<td>54.5</td>
</tr>
<tr>
<td>DBN (MTL) [5]</td>
<td>54.5</td>
</tr>
<tr>
<td>Proposed (without additional data)</td>
<td>57.5</td>
</tr>
<tr>
<td>Proposed (with additional data)</td>
<td>63.7</td>
</tr>
</tbody>
</table>

5.2. Effect of Auxiliary Tasks

In this section, we performed evaluations to highlight the effect of auxiliary tasks on the performance of cross-language SER using our proposed framework. Results are presented in Figure 2, which shows that using both language and domain identification as auxiliary tasks gives better performance compared to either using domain or language identification as an auxiliary task. This shows that using multiple related auxiliary tasks improves cross-language SER by producing generalised features. In addition, we achieve better results using language identification as an auxiliary task compared to using domain identification. This shows that identifying the language in cross-language SER help the MTL system to produce generalised features suitable for cross-language SER.

5.3. Using Partial Target Language Data

In this experiment, we show how a small percentage of target data can be utilised to significantly improve the cross-language SER. We vary the percentage of target samples (25 to 200) in the training data and results are plotted in Figure 3 for three target datasets (EMOVO, EMO-DB, and URDU). These results show that the generalisation of the proposed model for cross-language SER improves significantly by introducing a small percentage of target samples in the training data. Our proposed model is able to achieve performance comparable to the baseline as presented in Table 2 using only a small percentage of target data.

6. Conclusions and Future Works

In this work, we present a multi-task learning (MTL) framework for cross-language speech emotion recognition (SER). In particular, we enable to use of additional unlabelled emotional data to improve the generalisation and performance of the system. We propose to use language and domain identification as an auxiliary for which we utilise the additional data. We empirically showed that the proposed model considerably outperforms the previous studies on cross-language SER and additional data for auxiliary tasks significantly help improve the generalisation of the system. We also showed that the cross-language SER performance considerably improves when a small fraction of target data is used in the training data. In addition, we found that two related auxiliary tasks help produce better generalisation against cross-language SER in contrast to using only one auxiliary task. In contrast to domain identification, language identification as an auxiliary task achieved better SER performance. These findings would help researchers to utilise the abundantly available unlabelled data to build generalised multi-language SER systems with very few target examples. In our future work, we aim to focus on solving multi-modal auxiliary tasks to learn richer generalised representation.
7. References


