# Cross-Lingual Low-Resources Speech Emotion Recognition with Domain Adaptive Transfer Learning

Imen Baklouti[a], Olfa Ben Ahmed[b] and Christine Fernandez-Maloigne[c]

*XLIM Research Institute, UMR CNRS 7252, University of Poitiers, France*

Keywords: Domain Adaptation, Transfer Learning, Cross-Lingual Speech Emotion Recognition.

Abstract: Speech Emotion Recognition (SER) plays an important role in several human-computer interaction-based applications. During the last decade, SER systems in a single language have achieved great progress through Deep Learning (DL) approaches. However, SER is still a challenge in real-world applications, especially with low-resource languages. Indeed, SER suffers from the limited availability of labeled training data in the speech corpora to train an efficient prediction model from scratch. Yet, due to the domain shift between source and target data distributions traditional transfer learning methods often fail to transfer emotional knowledge from one language (source) to (target) to another. In this paper, we propose a simple yet effective approach for Cross-Lingual speech emotion recognition using supervised domain adaptation. The proposed method is based on 2D Mel-Spectrogram images as features for model training from source data. Then, a transfer learning method with domain adaptation is proposed in order to reduce the domain shift between source and target data in the latent space during model fine-tuning. We conduct experiments through different tasks on three different SER datasets. The proposed method has been evaluated on different transfer learning tasks namely for low-resource scenarios using the IEMOCAP, RAVDESS and EmoDB datasets. Obtained results demonstrate that the proposed method achieved competitive classification performance in comparison with the classical transfer learning method and with recent state-of-the-art SER-based domain adaptation works.

## 1 INTRODUCTION

The ability to understand and interpret human emotions is a fundamental aspect of effective communication and social interaction. In recent years, the field of affective computing has gained significant attention as researchers and technologists seek to develop intelligent systems capable of perceiving, analyzing, and responding to human emotions. Among the various modalities explored in affective computing, speech has garnered significant attention due to its rich emotional content, and accessibility. Our voice carries a wealth of information, including vocal pitch, intensity, rhythm, and spectral properties, all of which are influenced by our emotional states.

Speech Emotion Recognition (SER) presents an important field in affective computing that involves the automatic detection and analysis of emotions conveyed through speech signals. SER systems are used in several daily applications such as marketing, gaming, psychology, healthcare, and cognitive sciences. (El Ayadi et al., 2011; Lugović et al., 2016). However, several challenges exist in the field of SER because of the subjective and context-dependent nature of emotion in addition to the individual differences and linguistic variations (Lech et al., 2020). While significant progress has been made in SER, the majority of existing systems focus on single-language settings, limiting their applicability in multilingual and multicultural contexts (Agarla et al., 2022; Cai et al., 2021; Sharma, 2022; Zhang et al., 2021b). Indeed, emotions are very important and can have an impact on what we say and what we do and drive us to take decisions and make choices. However, through different languages and cultures, emotions are expressed differently that's why it's important to take into account these differences (Barth, 2020). Hence, developing an efficient and generalizable SER system remains challenging. Indeed, different languages exhibit unique phonetic and acoustic characteristics posing a challenge to developing language-independent deep learning models for SER. Additionally, the availability of labeled emotion datasets for training across multiple languages

[a] https://orcid.org/0000-0001-5413-8660
[b] https://orcid.org/0000-0002-6942-2493
[c] https://orcid.org/0000-0003-4818-9327

Baklouti, I., Ben Ahmed, O. and Fernandez-Maloigne, C.
Cross-Lingual Low-Resources Speech Emotion Recognition with Domain Adaptive Transfer Learning.
DOI: 10.5220/0012788100003756

In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 118-128
ISBN: 978-989-758-707-8; ISSN: 2184-285X

presents a significant obstacle. Yet, insufficient data for less-resourced languages hinders the development of accurate multilingual SER systems.

In this paper, we propose a new domain adaptation-based transfer learning method for multi-lingual SER. The proposed method is based on 2D Mel-Spectrogram data extracted from voice signals. A source model is learned in the first step and adapted in the second step to reduce the domain shift between the source and the target datasets. The goal is to leverage the knowledge learned from the source task to improve the performance of the model on a related but with a different (language) target SER task. Notably, our proposed method goes beyond conventional approaches by incorporating the extraction of domain-invariant features directly into the loss function in a supervised manner. This supervised adaptation mechanism guides the neural network to acquire features that are less influenced by domain-specific variations while maintaining the critical emotional information necessary for precise SER. The proposed method has been evaluated on different tasks using three different data-sets in different languages namely Interactive Emotional Dyadic Motion Capture (IEMOCAP), Berlin Database of Emotional Speech (EmoDB), and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).

The rest of the paper is organized as follows. First, section 2 exposes some literature review on deep learning and domain adaptation for SER. Second, section 3 presents the proposed SER method. Then in section 4, experiments and results are presented. Finally, section 5 concludes the work and gives some perspectives.

## 2 RELATED WORK

SER has witnessed significant advancements in recent years, fueled by the growing applications of affective computing in human-computer interaction and artificial intelligence. In this section, we will start by reviewing the use of deep learning in SER in the last years, and then we will present and discuss some recent and relevant works on domain adaptation for SER.

### 2.1 Deep Learning for SER

Deep Learning (DL) has been widely used for SER (Khalil et al., 2019) by extracting high-level features from raw speech data. Several works in SER have used Convolutional Neural Networks (CNN) to capture local acoustic patterns and learn discriminative features for spectrograms for emotion recognition (Huang et al., 2014; Badshah et al., 2019). Recurrent Neural Networks (RNNs) have also been investigated to capture both temporal and spectral information (Zhao et al., 2018; Ma et al., 2019). However, DL models trained on one language may not generalize well to other languages due to the differences between languages (Hu et al., 2020). Ensuring cross-lingual generalization and transferability of models are challenges in multilingual SER. In this context, cross-lingual Transfer Learning (TL), where models are pre-trained on large-scale speech datasets, is used for SER in small data. Indeed, pre-trained models can leverage the learned representations from general audio tasks and transfer them to SER, enhancing performance, especially in low-resource settings. For instance the authors in (Padi et al., 2021), proposed the TL of a pre-trained ResNet model with spectrogram data augmentation. the authors in (Liu et al., 2021), proposed to transform interval time parts of speech signals into a spectrogram and a discrete waveform fed separately to the FaceNet model for training. The network was pre-trained on the Chinese Academy of Sciences' Institute of Automation (CASIA) dataset and fine-tuned on the IEMOCAP dataset. The authors in (Gerczuk et al., 2021) presented a combination of residual adapters and ResNet model to be transferred from visual recognition to multi-corpus EmoSet. The authors in (Goel and Beigi, 2020), proposed a Multi-Task Learning framework with arousal, naturalness, and gender tasks for TL across SER datasets for five languages in single and cross corpus data. The authors in (Latif et al., 2018) proposed the Deep Belief Networks (DBN) model for TL evaluated on five corpora in different three languages SER. The authors in (Deng et al., 2013) proposed a sparse AutoEncoder (AE) technique evaluated on six databases for feature TL for SER. More recently, authors in (Scheidwasser-Clow et al., 2022), proposed a benchmark with nine datasets in six languages called SER adaptation benchmark for evaluating the extracted handcrafted features and leverages Deep Neural Networks (DNN).

TL has gained a lot of attention for multi-language deep-learning-based SER systems. However, it still suffers from some limitations. In fact, TL assumes that the source and the target domains are similar which is not the case in real-work speech emotion recognition tasks. This domain shift can degrade the performance of the model. Yet, pre-trained models may not generalize well to new emotional expressions or speakers not represented in the training data. Moreover, the limited availability of labeled data hampers the ability to effectively fine-tune the models, leading to overfitting and poor generalization.

## 2.2 Domain Adaptation for SER

In order to reduce the gap between source and target domains, Domain Adaptation (DA) methods have been recently proposed for cross-lingual SER. Most of the existing works focus on adversarial training for DA. For instance, The authors in (Latif et al., 2019) proposed an unsupervised DA cross-lingual SER using a Generative Adversarial Network (GAN) model. The authors in (Abdelwahab and Busso, 2018) proposed a novel Domain Adversarial Neural Network that exploits target unlabeled data to predict emotional descriptors for dominance, valence, and arousal. Authors proposed in (Gideon et al., 2019) an Adversarial Discriminative Domain Generalization and Multiclass MADDoG methods for cross-corpus SER evaluated on MSP-Improv, IEMOCAP and PRIORI datasets. More recently, authors in (Latif et al., 2022) proposed Adversarial Dual Discriminator (ADDi) and self-supervised (sADDi) methods to minimize the distance between source and target datasets. However, the presence of two discriminators in the ADDi architecture leads to training instability and an increase in both computational complexity and overfitting risk. A DA method based on the constant-Q Transform (CQT) for time-frequency has been proposed for SER in (Singh et al., 2021). The authors mentioned that they had no gain improvement over EmoDB because time-frequency needs monitoring in SER.

Another group of works maximizes the domain confusion to learn a common feature space by minimizing some measures of domain shift namely the Maximum Mean Discrepancy (MMD) distance (Yang et al., 2020; Liu et al., 2020; Zhang et al., 2021a; Kexin and Yunxiang, 2023; Ye et al., 2023). Indeed, the MMD has been used for DA in the computer vision field and proved its effectiveness in feature alignments compared to L1 and L2 distances, especially in high-dimensional spaces, which is often the case with complex data like images or spectrograms (Rezaeianjouybari and Shang, 2020). However, achieving DA for speech emotion presents a heightened level of complexity. This complexity arises from the necessity to simultaneously preserve emotional information while mitigating domain shift. For example, the authors in (Yang et al., 2020) proposed DA based on MMD layers with dynamic transfer coefficients using VGG19 and AlexNet models. The method has been evaluated on an office environment dataset leading to questions regarding its applicability in the real-world SER scenarios. Authors in (Liu et al., 2020) proposed DA using the MMD technique with CNN model based on LeNet and AlexNet models, to validate transfers between the eNTERFACE and CASIA datasets. More recently, the authors in (Zhang et al., 2021a) proposed an unsupervised novel Joint Distribution Adaptation Regression (JDAR) method for DA-based regression. The authors in (ZHAO et al., 2023) extracted 3D-mel-spectrograms using CNN then a Long Short-Term Memory (LSTM) model was trained on these features for SER. The authors in (Kexin and Yunxiang, 2023) combined Linear Discriminant Analysis, MMD and graph embedding to constrain the differences between target and source data while the authors in (Ye et al., 2023) proposed a novel dual-level emotion alignment module based on unsupervised DA. The authors proposed in this paper (Fu et al., 2023) to evaluate gender impact in SER, subdomain adaptation based on LMMD and multi-task learning models. However, integrating multiple tasks or subdomains into a single model often increases the model's complexity. In (Agarla et al., 2024) authors in proposed for Cross lingual SER a Semi-Supervised Learning method based on a transformer to classify unlabeled utterances. Some limitations are mentioned by authors. First pseudo-labeling can deteriorate performance if the model makes incorrect, unlabeled predictions. Second hard-pseudo labels can cause overfitting and the proposed method assumes that the target and source sets have the same number of emotions. The authors in proposed in this paper (Huijuan et al., 2023) a local domain adaptation using the local MMD method, to measure the distance between the source and target data. The authors proposed as future work, a multi-source DA using LMMD to be more suitable across SER.

Within the MMD-based domain adaptation approaches, there has been a notable omission in preserving emotional information while aligning the feature distributions. Moreover, most existing domain adaptation methods are designed for unsupervised or semi-supervised SER tasks due to the lack of annotated data in the target domain (low-resource domain). However, both unsupervised and semi-supervised DA methods lack direct supervision from labeled target domain data. This limits their ability to explicitly guide the adaptation process towards the target domain's specific characteristics. Without explicit supervision, the models may struggle to capture fine-grained details or subtle differences in emotional expressions across languages. In addition, most of the works cited above evaluate DA methods for cross-corpus SER using similar language corpora and few works show the effectiveness of their methods for different languages in cross-corpus SER (Ahn et al., 2021). Yet, few works have tackled the problem of DA from source to low-resource target domain. Hence, in this work, we propose a supervised DA-based transfer learning approach for SER for low-resource target domains. Combining classification loss, which encourages the model to correctly classify data, with MMD, which emphasizes the alignment of data distributions, can result in a more powerful discriminative model. This means the model can both classify well and adapt effectively to different domains.

# 3 PROPOSED APPROACH

Figure 1 illustrates the proposed framework of DA-based transfer learning for SER which is composed of two steps. The first one consists of learning the source model from source data, while the second one implements the domain adaptation-based transfer learning approach for final prediction from the target data.

## 3.1 Step 1: Learning the Source Model: A Mel-Spectrogram-Based Transfer

In order to get a source model for later transfer learning to the target data, we perform low-level feature-based transfer learning using an image-based pre-trained network. First, 2D Mel-Spectrograms are extracted from signals using the Librosa library (McFee et al., 2015). Indeed, a spectrogram is a visual representation of the spectrum of frequencies of the speech signal as it varies with time. Hence, we use spectrogram data to represent speech signals as images. Each point in the spectrogram represents the energy, or magnitude, of a specific frequency component at a particular time. Indeed, the process of fine-tuning the CNN using the spectrogram data involves taking advantage of the pre-trained CNN's ability to detect low-level image features. By initializing the CNN with these learned features and then adapting it to the spectrogram domain through fine-tuning, the network gains the capability to recognize low-level features and spatial details such as edges, and textures in spectrogram images that are informative for speech analysis. This process ultimately helps in transferring the knowledge of low-level image features to the speech domain, enhancing the network's performance on SER the task.

## 3.2 Step 2: Domain Adaption Based Transfer Learning

In this section, we describe our proposed method for transfer learning with domain adaptation in the case of cross-lingual SER. Hence, we propose a hybrid loss function for source model fine-tuning. Based on the Maximum Mean Discrepancy (MMD) method, the DA is proposed to minimize the discrepancies between source and target datasets for cross-lingual SER. The final strategy of model training is to minimize both, the CNN classification loss and the MMD loss. Hence, the model is optimized to classify samples with the classification loss when minimizing the domain distribution difference during transfer learning using the MMD loss.

For a source data $D_s = [(x_s^1, y_s^1), ..., (x_s^N, y_s^N)]$ where input features $X_s = [x_s^1, ..., x_s^N]$ with output labels $Y_s = [y_s^1, ..., y_s^N]$ and target data $D_t = [(x_t^1, y_t^1), ..., (x_t^M, y_t^M)]$ where input features $X_t = [x_t^1, ..., x_t^M]$ with output labels $Y_t = [y_t^1, ..., y_t^M]$, satisfy the marginal distributions $P(D_s)$ and $Q(D_t)$. MMD is often chosen for feature alignment in domain adaptation due to its ability to capture differences in distributions in a high-dimensional space compared to L1 and L2 distances (Attabi and Dumouchel, 2013). The total loss function is presented in Eq 1 :

$$Loss_{Total} = C_e(Y_s, y_s) + \lambda MMD_{FC}(D_s, D_t) \tag{1}$$

Where $C_e$ is a loss function Cross entropy computed from labeled source set, $MMD_{FC}$ presents the MMD loss function (the distance between the target data $D_t$ and the source data $D_s$), $y_s$ is the source labels data and $\lambda$ is the weight of $MMD_{FC}$ to transfer source knowledge to domain target.

The estimation of the MMD method is presented as follow (see Eq 2), where the kernel function k (Gretton et al., 2006) is a gaussian function also P and Q are two marginal different distributions sample sets ($P(D_s) \neq Q(D_t)$). Such explicit alignment in the Reproducing Kernel Hilbert Space (RKHS) could support the structure learning for precise conditional alignment.

$$MMD_{FC}[P,Q] = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(D_s^i, D_s^j\right) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k\left(D_s^i, D_t^j\right) + \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} k\left(D_t^i, D_t^j\right) \tag{2}$$

Two output batches from the source $D_s$ and target $D_t$ datasets are calculated in the first Fully Connected (FC) layer. Eq 2 calculates the $MMD_{FC}$ as a loss function for the output, to the optimize the training process. The objective here is that the distance

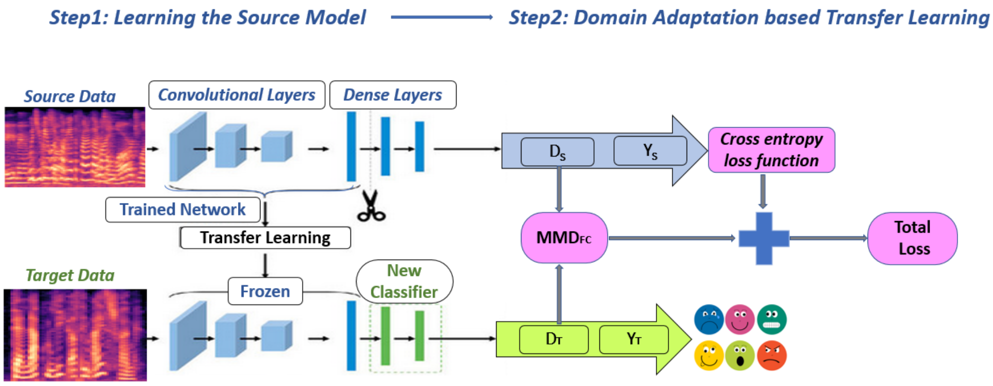


Figure 1: Domain adaptation based transfer learning proposed method for cross lingual SER.

between the source and target data (MMD loss) for cross-lingual SER will be narrowed and the total loss function will be decreased.

By jointly optimizing the cross-entropy loss and the MMD loss, we encourage the model to learn a feature representation that minimizes the distribution discrepancy between the source and the target domains while maintaining good classification performance on the source domain. The weighting factor λ determines the trade-off between the two loss terms.

# 4 EXPERIMENTS AND RESULTS

## 4.1 SER Datasets

In this work, we use three different SER datasets for the evaluation of the proposed method. Table 1 provides the characteristics of these corpus namely IEMOCAP (English, Audio) (Busso et al., 2008), EmoDB (German, Audio) (Burkhardt et al., 2005) and RAVDESS (North American, Audio-Visual) dataset. (Livingstone and Russo, 2018). A total of 535, 9952 and 1248 audio files with seven emotions have been used for EmoDB, IEMOCAP and RAVDESS datasets, respectively.

Table 1: Characteristics of Emotional Speech Databases.

| Database | data | Emotions |
|---|---|---|
| IEMOCAP (English) | 9952 | 1:Angry, 2:Happy, 3:Excited, 4:Sad, 5:Frust, 6:Fear, 7:Neutral |
| EmoDB (German) | 535 | 1:Angry, 2:Bored, 3:Disgust, 4:Fear, 5:Happy, 6:Sad, 7:Neutral |
| RAVDESS (North American) | 1248 | 1:Neutral, 2:Calm, 3:Happy, 4:Sad, 5:Angry, 6:Fear, 7:Disgust |

## 4.2 Training Setting and Model Parameters

The model is fine-tuned for 400 epochs, with a learning rate of $10^{-4}$ and a batch size equal to 16 for all experiments. Categorical cross entropy loss and Adam optimizer are used for model optimization. In this way, the model has been adapted to the new properties of the source data. Source and target datasets are randomly divided into training (80%) and testing (20%) sets. Two evaluation metrics are used:

- Unweighted Accuracy: $Accuracy = (t_p + t_n)/p + n$ where $t_p$ and $t_n$ are predictions of scores true positive and true negative, respectively. $n + p = t_p + t_n + f_n + f_p$ where $f_n$ is a prediction of score false negative and $f_p$ is a prediction of a false positive score.
- Unweighted Average Recall (UAR): $UAR = 0.5 * (t_p/p + t_n/n)$ .

## 4.3 Classification Results

In this section, we present and discuss the obtained classification results namely on the source model and the target model with DA.

### 4.3.1 Classification Results for the Source Model Learning (Only Source Domain)

In order to select the best model for source data prediction, we computed the prediction of the test set for the three source datasets. By employing transfer learning for feature extraction, we have generated Log Mel-Spectrogram from the input audio set. We performed several experiments with different parameters and configurations for fine-tuning the VGG model on the source data. Table 2 presents the best obtained classification accuracy for the source model. We report classification accuracies of 82.24%, 78.5% and

83.17% for the RAVDESS, IEMOCAP and EmoDB datasets, respectively.

Table 2: Classification results for the source model.

| work | Model | Dataset | Accuracy % |
|---|---|---|---|
| (Senthilkumar et al., 2022) | Spec VGG16 | RAVDESS | 77 |
| (Aggarwal et al., 2022) | Spec VGG16 | RAVDESS | 81.94 |
| Ours | TL | RAVDESS | **82.24** |
| (Senthilkumar et al., 2022) | Spec VGG16 | IEMOCAP | 63 |
| Ours | TL | IEMOCAP | **78.5** |
| (Senthilkumar et al., 2022) | SpecVGG16 | EmoDB | 75 |
| Ours | TL | EmoDB | **83.17** |

We also compare the obtained results with recent works on SER in Table 2. In (Aggarwal et al., 2022), the authors proposed a two-way approach for SER. The first one is based on Principal Component Analysis for feature extraction and a DNN for training. In the second method, Mel-spectrogram images are extracted from audio files and given as input to a pre-trained VGG-16 model. Senthilkumar et .al (Senthilkumar et al., 2022) proposed Bi-directional LSTM architecture. They also have some disadvantages such as increased computational complexity, overfitting risk and increased memory requirements. Indeed, Table 2 shows that our proposed transfer learning strategy has better accuracy results compared to (Senthilkumar et al., 2022) and (Aggarwal et al., 2022). For instance, for RAVDESS from Table 2, we can observe that our transfer learning method leads to an average increase of at least 0.3% for the accuracy, 15.5% for the IEMOCAP and 8.17% for the EmoDB dataset compared to the results obtained in recent works.

In order to evaluate the model performance per class, we present confusion matrices on the test data of the source domain (Figure 2). The goal here is to see how well the model is performing for each class, including any imbalances or weakly represented classes. The test data were randomly sampled. Figures 2(a),2(b),2(c) show that the proposed transfer learning gives high accuracy with low misclassifications for all datasets. Table 3 presents further classification results per category.

We can see in the case of RAVDESS, IEMOCAP and EmoDB datasets that precision of fear, recall of fear and precision of disgusted, sad and neutral recall emotions achieve 1, respectively.

#### 4.3.2 Classification Results on Target Data (New Domain)

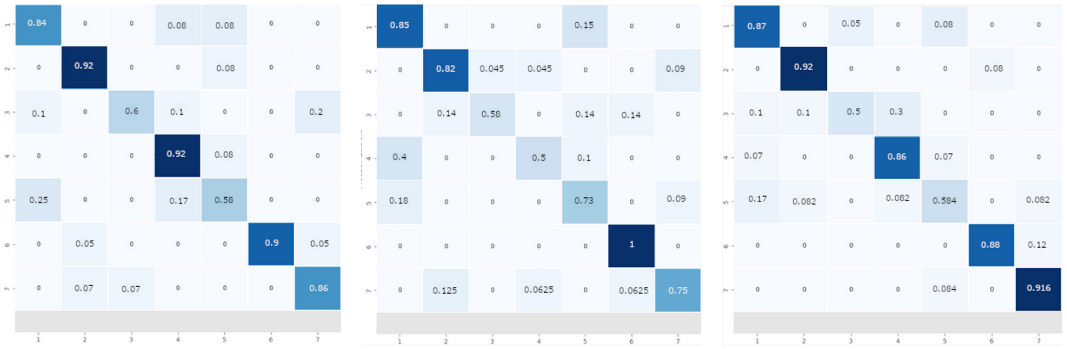
In order to evaluate the proposed method, we perform 4 transfer learning tasks (Source → Target) with DA using different datasets:

Table 3: Classification results on source model.

| | **Neutral** | **Calm** | **Happy** | **Sad** | **Anger** | **Fear** | **Disgust** |
|---|---|---|---|---|---|---|---|
| | | | RAVDESS | | | | |
| Precision | 0.84 | 0.86 | 0.86 | 0.71 | 0.64 | **1** | 0.81 |
| Recall | 0.84 | **0.92** | 0.60 | **0.92** | 0.58 | 0.89 | 0.87 |
| F1-score | 0.84 | 0.89 | 0.71 | 0.80 | 0.61 | **0.94** | 0.84 |
| | **Anger** | **Happy** | **Excited** | **Sad** | **Frust** | **Fear** | **Neutral** |
| | | | IEMOCAP | | | | |
| Precision | 0.79 | 0.86 | 0.71 | 0.57 | 0.64 | 0.87 | 0.80 |
| Recall | 0.85 | 0.83 | 0.50 | 0.73 | 0.58 | **1** | 0.75 |
| F1-score | 0.82 | 0.84 | 0.59 | 0.64 | 0.61 | **0.93** | 0.77 |
| | **Anger** | **Bored** | **Disgust** | **Fear** | **Happy** | **Sad** | **Neutral** |
| | | | EmoDB | | | | |
| Precision | 0.76 | 0.83 | **1** | **0.94** | 0.54 | **1** | 0.89 |
| Recall | **0.92** | **0.91** | 0.46 | **0.94** | 0.64 | 0.79 | **1** |
| F1-score | 0.83 | 0.87 | 0.63 | **0.94** | 0.58 | 0.88 | **0.94** |

- **Task 1:** IEMOCAP (**9952**) → EmoDB (**535**) (English→German)
- **Task 2:** EmoDB (**535**) → IEMOCAP (**9952**) (German→ English)
- **Task 3:** RAVDESS (**1248**) → EmoDB (**535**) (English:North American → German)
- **Task 4:** EmoDB (**535**) → RAVDESS (**1248**) (German → English:North American)

We used the output of the first FC layer in VGG 16 model to calculate the MMD loss between source and target domains (Yang et al., 2020). The model has been trained for 50 epochs. Table 4 presents the classification results using the proposed DA-based transfer learning approach on the 4 transfer tasks. In order to ensure a fair comparison with the works cited in Table 4, we compute the average and the standard deviation values of both accuracy and UAR metrics for task 1 and task 2. All values are calculated with a margin of error or uncertainty associated with the true value is expected to fall within this range centered around the given value. We repeat each experiment ten times and computed the standard deviation and mean. We compare the performance of the DA-based TL method with sADDi (Latif et al., 2022), GAN (Latif et al., 2019), AE (Deng et al., 2013) and CNN-LSTM (Parry et al., 2019) methods. Our proposed method compared to these existing techniques, achieves better results. Moreover, the DA-based TL technique has demonstrated improvements in UAR for tasks 1 and 2 in UAR by 1.1% and 4.1%, respectively, compared to the sADDi method (Latif et al., 2022), 4.9% and 8.5% compared to the GAN method (Latif et al., 2019) and 7% and 9.5% (including 9.5% in term of accuracy for task 1) compared to the CNN-LSTM method (Parry et al., 2019), respectively. For tasks 3 and 4, we have only one state of the art work that have performed cross-corpus in SER (Singh et al., 2021) and there



(a) Confusion Matrix of RAVDESS (b) Confusion Matrix of IEMOCAP (c) Confusion Matrix of EmoDB

Figure 2: Confusion Matrices of the source model on RAVDESS, IEMOCAP and EmoDB datasets.

are not be specific studies directly comparing these two datasets in a cross-lingual context. However, both datasets are widely used for emotion recognition tasks in speech processing. We can see in table 4 when we used our proposed DA-based TL method for tasks 3 and 4, that we achieved better results than CQT (Singh et al., 2021) and MFSC (Singh et al., 2021) methods. Accuracy and UAR metric values are further boosted by using a DA-based TL method. We achieved a 57% accuracy rate against CQT 48% and MFSC 45% (Singh et al., 2021) in task 3 and a 50% accuracy rate against 44% and 41% (Singh et al., 2021) in task 4, respectively. Indeed for UAR values, we obtained 58% against CQT 48% and MFSC 42% (Singh et al., 2021) in task 3 and 50% against CQT 44% and MFSC 41% (Singh et al., 2021) in task 4, respectively. In summary, the proposed method proves that the network is able to produce better-generalized features for cross-language SER by increasing the final UAR and accuracy of the four tasks.

Table 4: Comparative results of DA-based TL cross-lingual results by UAR (%) and Accuracy (%) metrics for the 4 tasks.

| Work | Models | Task 1 | Task 2 |
|---|---|---|---|
| | | UAR/Accuracy | UAR/Accuracy |
| (Parry et al., 2019) | CNN-LSTM | 42.1±1.8/41.99 | 38.9±2.1/NA |
| (Latif et al., 2018) | DBN | 42.5±2.1/NA | 39.5±2.4/NA |
| (Deng et al., 2013) | AE | 43.2±2.3/NA | 40.1±1.8/NA |
| (Latif et al., 2019) | GAN | 44.3±1.7/NA | 40.3±1.7/NA |
| (Latif et al., 2022) | ADDi | 46.1±1.6/NA | 41.2±1.8/NA |
| (Latif et al., 2022) | sADDi | 48.3±1.5/NA | 44.8±1.6/NA |
| Ours | DA based TL | **50.1±0.8/ 50.74±0.3** | **50±0.5/ 50.02±0.4** |
| Work | Models | Task 3 | Task 4 |
| | | UAR/Accuracy | UAR/Accuracy |
| (Singh et al., 2021) | MFSC | 42/45 | 44/41 |
| (Singh et al., 2021) | CQT | 48/48 | 46/44 |
| Ours | DA based TL | **58/ 57** | **50/ 50** |

Table 5 presents the classification results of TL compared to DA-based TL methods for task 1 and task 3 cross-lingual SER. It shows that the classification accuracy of seven emotions in terms of Precision, Recall and F1-score are improved with DA-based TL compared to classification report values of the TL. In the case of the proposed DA-based TL method, RAVDESS→EmoDB (task 3) achieves high F-score, Precision and Recall values (> 70%) over Happy emotion while in the case of TL method F-score, Precision and Recall are the categories with the lowest values (< 30%). As shown in Table 5, IEMOCAP→EmoDB (task 1) with TL method over Boredom emotion achieves a classification performance (< 35%) lower than IEMOCAP→EmoDB with DA-based TL (> 50%) method.

Table 5: Classification results of IEMOCAP → EmoDB (task 1) and RAVDESS→EmoDB (task 3) cross-lingual SER using DA based TL method.

| | Angry | Bored | Disgust | Fear | Happy | Sad | Neutral |
|---|---|---|---|---|---|---|---|
| Task 1 | | | | | | | |
| Classification results of DA based TL method | | | | | | | |
| Precision | 0.32 | **0.55** | 0.32 | 0.33 | 0.58 | 0.60 | 0.45 |
| Recall | 0.47 | **0.55** | 0.29 | 0.39 | 0.61 | 0.39 | 0.49 |
| F1-Score | 0.38 | **0.55** | 0.31 | 0.36 | 0.59 | 0.47 | 0.47 |
| Classification results of transfer learning method | | | | | | | |
| Precision | 0.35 | **0.34** | 0.47 | 0.29 | 0.58 | 0.46 | 0.47 |
| Recall | 0.39 | **0.29** | 0.45 | 0.44 | 0.56 | 0.44 | 0.38 |
| F1-score | 0.37 | **0.31** | 0.46 | 0.35 | 0.57 | 0.45 | 0.42 |
| Task 3 | | | | | | | |
| Classification results of DA based TL method | | | | | | | |
| Precision | 0.45 | 0.72 | 0.61 | 0.50 | **0.70** | 0.53 | 0.61 |
| Recall | 0.69 | 0.82 | 0.41 | 0.32 | **0.73** | 0.65 | 0.74 |
| F1-score | 0.55 | 0.77 | 0.49 | 0.39 | **0.71** | 0.58 | 0.67 |
| Classification results of transfer learning method | | | | | | | |
| Precision | 0.74 | 1 | 0.5 | 0.37 | **0.27** | 0.28 | 0.27 |
| Recall | 0.57 | 0.42 | 0.5 | 0.44 | **0.27** | 0.83 | 0.42 |
| F1-score | 0.64 | 0.59 | 0.5 | 0.40 | **0.27** | 0.42 | 0.52 |

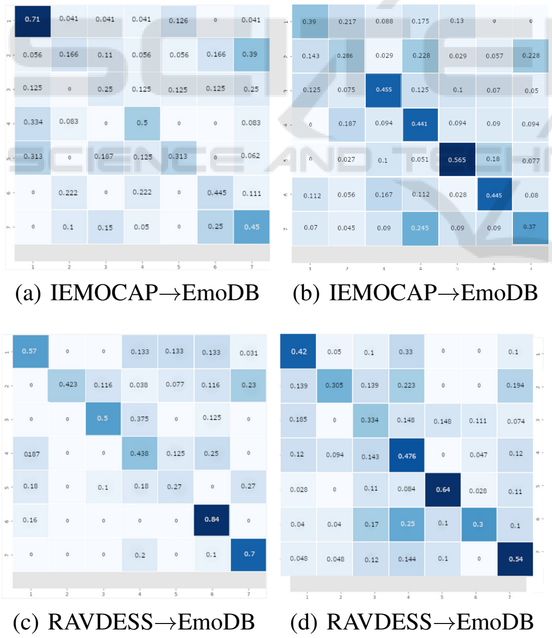
### 4.3.3 Effect of DA on Transfer Learning

**Classification Results:** In this section, we present results for the most representative tasks namely task 1 and task 3 that illustrate an adaptation from English data to a smaller German (low resource) one. Table 6, presents a comparative study of accuracy and UAR results metrics for IEMOCAP→EmoDB (task 1) and RAVDESS→EmoDB (task 3) cross-corpus using DA-based TL and TL methods.

Table 6: Classification results on the target domain.

| Method | Task | Accuracy (%) | UAR (%) |
|---|---|---|---|
| Transfer Learning | Task 1 | 45.8 | 44.73 |
| DA-based TL (ours) | Task 1 | **50.74** | **50.1** |
| Transfer Learning | Task 3 | 50.47 | 50.65 |
| DA-based TL (ours) | Task 3 | **53.47** | **58.01** |

As shown in Table 6 the results of the experimental groups using the DA based TL technique are better than those obtained by the TL method from scratch. Indeed, with task DA based TL for IEMOCAP→EmoDB (task 1), the recognition accuracy and UAR are 5.24% and 6% higher than those obtained by TL method for IEMOCAP→EmoDB (task 1) model, respectively. For RAVDESS→EmoDB (task 3) accuracy and UAR obtained by DA based TL, are 3% and 8.6% higher than those obtained by the TL method for RAVDESS→EmoDB (task 3), respectively.

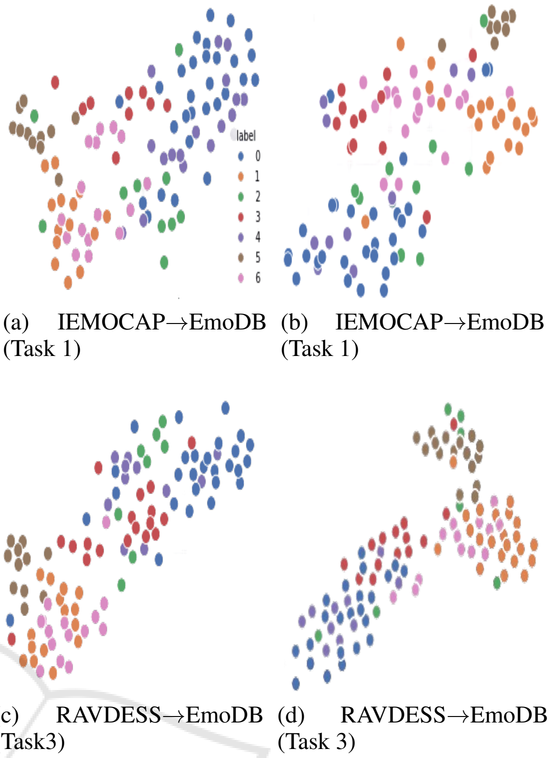


(a) IEMOCAP→EmoDB (b) IEMOCAP→EmoDB

(c) RAVDESS→EmoDB (d) RAVDESS→EmoDB

Figure 3: Confusion Matrices for RAVDESS→EmoDB and IEMOCAP→EmoDB tasks using TL and DA based TL methods with 1:Angry, 2:Boredom, 3:Disgust, 4:Fear, 5:Happy, 6:Sad, 7:Neutral emotions.

We plot confusion matrices on the test data of the target domain (Figure 3) with the same goal as for the evaluation of the source model. Indeed, some emotions might naturally occur less frequently in real-world scenarios, making them weakly represented in the data. Hence, in our experiments, the test data are randomly sampled to contain some underrepresented emotions that may illustrate real-world and low-resource scenarios. We can see from the confusion matrix Figures of IEMOCAP→EmoDB 3(b) and RAVDESS→EmoDB 3(d) cross-lingual, that the distribution of features resulted in a high classification effect using the proposed approach DA based TL compared to the confusion matrix Figures of IEMOCAP→EmoDB 3(a) and RAVDESS→EmoDB 3(c) cross-lingual with TL technique. This is because the good detection emotion abilities of DA-based TL over TL techniques. Comparing Figures 3(b) and 3(d), we observe a general increase in accuracy for different classes of emotions for the DA based TL method, usually at the expense of the ”disgust” class. Further-more, we note that ”Happy, Sad and Neutral” classes, appears to have the largest gain in accuracy.



(a) IEMOCAP→EmoDB (Task 1) (b) IEMOCAP→EmoDB (Task 1)

(c) RAVDESS→EmoDB (Task3) (d) RAVDESS→EmoDB (Task 3)

Figure 4: Visualization of features for the tasks 1 and 3 using transfer learning (Fig.(a,c)) and proposed approach (Fig.(b,d)).

**Features Visualization:** In order to validate the effectiveness of the proposed domain adaptation method for speech emotion recognition, we plot the t-SNE (Van der Maaten and Hinton, 2008) for features extracted from the obtained model by simple transfer learning and features generated by the proposed method. Hence, we used the layer FC-4 to extract the features and to plot labels. The points are colored according to the emotion category.

From Figure 4 we can see that our proposed approach (figures 4(b) and 4(d)) is able to more effectively separate emotional features from different classes and the clusters are better structured than the basic transfer learning method (figures 4(a) and 4(c)) from the tasks IEMOCAP→EmoDB (task 1) and RAVDESS→EmoDB (task 3). As can be observed in figures 4(b) and 4(d), that the inter-class variance is increased and the emotional classes are highly correlated, while the obtained emotional clusters are more compact. The proposed adaptive transfer learning method helps in learning discriminative multi-language features and hence improves the classification performance of emotions from speech.

## 5 CONCLUSION

In this paper, we propose a domain adaptive-based transfer learning approach for cross-lingual Speech Emotion Recognition problem. The proposed method jointly align domain invariant features and improve feature quality for better classification in a supervised manner. The proposed method has been evaluated on different transfer learning tasks namely for low-resource scenarios using IEMOCAP, RAVDESS and EmoDB datasets. Obtained results revealed that our proposed approach outperforms existing SER methods in terms of both UAR and Accuracy metrics. The promising results obtained from this study open avenues for future research in the realm of cross-lingual low-resource speech emotion recognition due to the the modest improvement.

## ACKNOWLEDGEMENTS

Support for this research was provided by a grant from La Région Nouvelle Aquitaine (CPER FEDER P-2019-2022), in partnership with the European Union (FEDER/ ERDF, European Regional Development Fund).

## REFERENCES

Abdelwahab, M. and Busso, C. (2018). Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435.

Agarla, M., Bianco, S., Celona, L., Napoletano, P., Petrovsky, A., Piccoli, F., Schettini, R., and Shanin, I. (2022). Semi-supervised cross-lingual speech emotion recognition. *arXiv preprint arXiv:2207.06767*.

Agarla, M., Bianco, S., Celona, L., Napoletano, P., Petrovsky, A., Piccoli, F., Schettini, R., and Shanin, I. (2024). Semi-supervised cross-lingual speech emotion recognition. *Expert Systems with Applications*, 237:121368.

Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A. A., Alhadlaq, A., and Lee, H.-N. (2022). Two-way feature extraction for speech emotion recognition using deep learning. *Sensors*, 22(6):2378.

Ahn, Y., Lee, S. J., and Shin, J. W. (2021). Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation. *IEEE Signal Processing Letters*, 28:1190–1194.

Attabi, Y. and Dumouchel, P. (2013). Anchor models for emotion recognition from speech. *IEEE Transactions on Affective Computing*, 4(3):280–290.

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., and Baik, S. W. (2019). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, 78:5571–5589.

Barth, I. (2020). When we talk about emotions and plurilingualism.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Cai, X., Wu, Z., Zhong, K., Su, B., Dai, D., and Meng, H. (2021). Unsupervised cross-lingual speech emotion recognition using domain adversarial neural network. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

Deng, J., Zhang, Z., Marchi, E., and Schuller, B. (2013). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587.

Fu, H., Zhuang, Z., Wang, Y., Huang, C., and Duan, W. (2023). Cross-corpus speech emotion recognition based on multi-task learning and subdomain adaptation. *Entropy*, 25(1):124.

Gerczuk, M., Amiriparian, S., Ottl, S., and Schuller, B. W. (2021). Emonet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*.

Gideon, J., McInnis, M. G., and Provost, E. M. (2019). Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog). *IEEE Transactions on Affective Computing*, 12(4):1055–1068.

Goel, S. and Beigi, H. (2020). Cross lingual cross corpus speech emotion recognition. *arXiv preprint arXiv:2003.07996*.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Huang, Z., Dong, M., Mao, Q., and Zhan, Y. (2014). Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804.

Huijuan, Z., Ning, Y., and Ruchuan, W. (2023). Improved cross-corpus speech emotion recognition using deep local domain adaptation. *Chinese Journal of Electronics*, 32(3):1–7.

Kexin, Z. and Yunxiang, L. (2023). Speech emotion recognition based on transfer emotion-discriminative features subspace learning. *IEEE Access*.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345.

Latif, S., Qadir, J., and Bilal, M. (2019). Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 732–737. IEEE.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., and Schuller, B. W. (2022). Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition. *IEEE Transactions on Affective Computing*.

Latif, S., Rana, R., Younis, S., Qadir, J., and Epps, J. (2018). Transfer learning for improving speech emotion classification accuracy. *arXiv preprint arXiv:1801.06353*.

Lech, M., Stolar, M., Best, C., and Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, 2:14.

Liu, J., Zheng, W., Zong, Y., Lu, C., and Tang, C. (2020). Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network. *IEICE TRANSACTIONS on Information and Systems*, 103(2):459–463.

Liu, S., Zhang, M., Fang, M., Zhao, J., Hou, K., and Hung, C.-C. (2021). Speech emotion recognition based on transfer learning from the facenet framework. *The Journal of the Acoustical Society of America*, 149(2):1338–1345.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Lugović, S., Dunder, I., and Horvat, M. (2016). Techniques and applications of emotion recognition in speech. In *2016 39th international convention on information and communication technology, electronics and microelectronics (mipro)*, pages 1278–1283. IEEE.

Ma, A., Filippi, A. M., Wang, Z., and Yin, Z. (2019). Hyperspectral image classification using similarity measurements-based deep recurrent neural networks. *Remote Sensing*, 11(2):194.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

Padi, S., Sadjadi, S. O., Sriram, R. D., and Manocha, D. (2021). Improved speech emotion recognition using transfer learning and spectrogram augmentation. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 645–652.

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., and Hofer, G. (2019). Analysis of deep learning architectures for cross-corpus speech emotion recognition. In *Interspeech*, pages 1656–1660.

Rezaeianjouybari, B. and Shang, Y. (2020). Deep learning for prognostics and health management: State of the art, challenges, and opportunities. *Measurement*, 163:107929.

Scheidwasser-Clow, N., Kegler, M., Beckmann, P., and Cernak, M. (2022). Serab: A multi-lingual benchmark for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7697–7701. IEEE.

Senthilkumar, N., Karpakam, S., Devi, M. G., Balakumaresan, R., and Dhilipkumar, P. (2022). Speech emotion recognition based on bi-directional lstm architecture and deep belief networks. *Materials Today: Proceedings*, 57:2180–2184.

Sharma, M. (2022). Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6907–6911. IEEE.

Singh, P., Saha, G., and Sahidullah, M. (2021). Non-linear frequency warping using constant-q transformation for speech emotion recognition. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Yang, Z., Liu, G., Xie, X., and Cai, Q. (2020). Efficient dynamic domain adaptation on deep cnn. *Multimedia Tools and Applications*, 79(45):33853–33873.

Ye, J., Wei, Y., Wen, X.-C., Ma, C., Huang, Z., Liu, K., and Shan, H. (2023). Emo-dna: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition. *arXiv preprint arXiv:2308.02190*.

Zhang, J., Jiang, L., Zong, Y., Zheng, W., and Zhao, L. (2021a). Cross-corpus speech emotion recognition using joint distribution adaptive regression. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3790–3794. IEEE.

Zhang, S., Liu, R., Tao, X., and Zhao, X. (2021b). Deep cross-corpus speech emotion recognition: Recent advances and perspectives. *Frontiers in Neurorobotics*, 15.

ZHAO, H., YE, N., and WANG, R. (2023). Improved cross-corpus speech emotion recognition using deep local domain adaptation. *Chinese Journal of Electronics*, 32(3):1–7.

Zhao, Z., Zhao, Y., Bao, Z., Wang, H., Zhang, Z., and Li, C. (2018). Deep spectrum feature representations for speech emotion recognition. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pages 27–33.