Signal Processing and Renewable Energy
December 2021, (pp. 83-106)
ISSN: 2588-7327
eISSN: 2588-7335

Parallel Shared Hidden Layers Auto-encoder as a Cross-Corpus Transfer Learning Approach for Unsupervised Persian Speech Emotion Recognition

Yousef Pourebrahim[1], Farbod Razzazi*[1], Hossein Sameti[2]

[1] Department of Electrical and Computer Engineering, Science and Research Branch,
Islamic Azad University, Tehran, Iran.
[2] Speech Processing Laboratory, Department of Computer Engineering,
Sharif University of Technology, Tehran, Iran.

Received: 24-Apr-2021, Revised: 14-May-2021, Accepted: 14-May-2021.

Abstract

Detecting emotions from speech is one of the challenging topics in speech signal processing, especially in low resource languages. Extracting common features between the training and testing set, using unsupervised method, can solve the inconsistency difficulty between training and test data. In this study, a new auto-encoder based structure is proposed as a new unsupervised method for domain adaptation. To this end, the proposed structure is made of shared encoders to learn common feature representations, shared across the source and the target domain datasets to minimize the discrepancy between them. In order to evaluate the performance of the proposed method, five generally available databases in different languages were used as training and testing datasets. Results on various scenarios demonstrated that the proposed method improves the classification performance significantly compared to the baseline and state of the art unsupervised domain adaptation methods for emotional speech recognition. As an example, the proposed method improved the emotion recognition rate in Persian emotional speech dataset (PESD) by 8% compared to cross corpus training when the source training set is EMOVO.

Keywords: Transfer Learning; Emotional Speech Recognition; Deep Neural Networks; Unsupervised Classification.

1. INTRODUCTION

The recognition of emotional speech and its various applications have attracted many researchers in recent years. In general, this approach focuses on teaching a machine to correctly recognize human emotions from

***Corresponding Authors Email:** farbod_razzazi@yahoo.com

speech.

Most research on the recognition of emotions focuses on the supervised classification of a short emotional speech by using a wide variety of classifiers [1], such as Hidden Markov Models (HMMs) [2], Gaussian Mixture Models (GMMs) [3], Support Vector Machines (SVMs) [4], and Deep Neural Networks (DNN) [5, 6].

With the expansion of the use of deep learning, applying deep neural networks in the recognition of categorical emotion is also noted [7, 8]. For example, Deep Convolution network has been used to learn high-level representation and classification of emotional speech [9]. Adversarial auto-encoder has also been used to high dimensional feature reduction. The obtained features have been used to train a classifier [10]. In [11], a recurrent neural network (RNN) with an attention mechanism has been used to learn a new representation in utterance-level by collecting frame-level features over time. Then, by passing these new representations from several dense layers, the classification has been performed. However, all these methods require a large number of emotional utterances to train a classifier and it is assumed that the training and test samples are from the same set, or that the statistical distribution of the training and the test samples are the same [12].

In real applications, these assumptions lead to poor recognition performance. This is because of the fact that training samples are collected under certain conditions and are used to train a classifier, while test samples are obtained from a variety of environmental conditions (e.g. the presence of noise in the environment, the speaker's age, the type of spoken language, and even the genre of the speaker). Hence, in practice, we encounter an inter-data problem in which the classifier is trained under a given data and is tested under another data. This is called Cross-Corpus (CC) problem [4].

In addition, despite the success of the supervised procedures on emotional speech recognition they need to have enough number of labeled data in target domain. Due to the cost of samples labeling, there is a strict limitation to access labeled emotional speech samples. Moreover, there are no specific criteria for confident labeling. In other words, a same sentence may have different labels based on people’s mental understandings. In contrast, providing unlabeled data for this purpose costs much less. Therefore, the use of transfer learning can be useful for the field of emotional speech recognition as well. For example, the labeled corpus may be acted speech and test data to be spontaneous corpus where the training and test data’s features or data’s distributions may be different. As a result, it is not possible to directly apply the emotion models learnt on the acted speech to the new spontaneous data. In such cases, transferring the classification knowledge learnt by the model into the new domain could be helpful. Transfer learning has been proposed as an effective solution to this kind of problem by learning the knowledge gained from a source domain (training set) and adapting it to a target domain (test set) [13-15]. For example, Hassan et al. have introduced three transfer learning approaches to compensate the environmental disparities and differences between speakers in source and target domains [16].

Transfer learning can be categorized to

Inductive and Transductive method. In Inductive transfer learning, it is assumed that there are labeled samples in both source and target domain. In Transductive transfer learning, labeled samples are available only in the source domain and the distributions of the source and target domains are different [17]. In this case, transfer learning can be called domain adaptation. A general method in domain adaptation is to assign a large weight to more similar samples of the source domain to test data and less weight to the other samples. This method is known as importance weighting. One of these methods is unconstrained least-squares importance fitting to estimate the importance weights by a linear model [18].

In general, domain adaptations techniques are divided into supervised, semi-supervised, and unsupervised categories, depending on whether there are labeled samples in the target data or not. In the supervised approach, labeled samples in the target domain are used to train the domain transformation [13]. In semi-supervised approach, it is assumed that there are a small number of labeled samples in target domain. In an unsupervised method, there are not any labeled samples in target domain. In this case, it is assumed that there are a number of discriminative features that are common or invariant in both domains, or that there is a hidden space where the distribution difference between target and source is minimal [19], or there are transformations that can map the source data to the target data [20].

Recently, lots of deep learning approaches based on auto-encoders have achieved a significant performance in domain adaptation. Yang et al. proposed a novel Semi-Supervised Representation Learning framework via Dual auto-encoders for domain adaptation, named SSRLDA [21]. They made full use of label information to optimize feature representations. For this, pseudo-labels are used for target samples.

Deng et al. have proposed an unsupervised domain adaptation method to solve the problem of the difference between source and target domains in emotional speech recognition, in which past knowledge derived from target data has been used to regulate the learning based on source data [4].

Based on the idea of ‘shared learning’, Deng et al. have proposed a new structure of auto-encoder that tries to minimize the reconstruction error on both source and target domains [14]. This structure shares the same weights for the mapping from the input layer to the hidden layer. In contrast, it uses independent weights for the reconstruction process. They have called outlined structures as “Shared hidden layer auto-encoder” (SHLA as short). Using this technique, they have obtained good results for cross-corpus emotional speech recognition.

Similar to the pseudo-label idea and shared hidden layer auto-encoders, in [22] a semi-supervised shared hidden layer auto-encoder is proposed as domain adaptation for emotional speech recognition. In this method it is shown that using residual connection and pseudo-labels for target samples can be useful for increasing the accuracy of recognition.

Therefore, the most challenging part in the emotional speech recognition field is to collect enough training and test data. However, collecting this data is expensive and time consuming. In addition, labeling of collected samples is difficult because of uncertainty (Different people may have different perceptions of a single emotional utterance).

Cross lingual studies have shown that using the datasets in one language in recognition of another language by transfer learning methods can be effective in emotional speech recognition. Research in this area has been well-developed for some languages such as German, French and English languages [14]. These languages have some common properties such as cultural similarity and root of language. These properties can lead to increase the performance of the above-mentioned methods; because of their simple structure. In contrast, in the situation that the roots of the spoken languages are completely different, the complex relation between them cannot be extracted. This is due to the fact that previous methods do not have enough depth in terms of the number of neurons and layers to extract discriminative features between different languages. Our proposed method is the deep unsupervised structure for domain adaptation in the form of auto-encoders which is constructed with parallel encoders that reduces the computational cost as well as its succeed in modeling of a complex function and achieves better performance compared to the state of the art methods. Therefore, we aim to introduce a new auto-encoder with parallel shared encoders as a domain adaptation method to compensate the distribution difference between features obtained from emotional speech corpus of two different languages.

1.2. Contribution

In this study, based on the width expansion of layers in inception neural network and SHLA approach for domain adaptation, we propose a ‘Parallel Shared Encoders as a Hidden Layer for an Auto-encoder’ (PSHLA) that is useful for domain adaptation in emotional speech recognition. It is expected that PSHLA can extract common robust features and therefore it provides a common distribution for different input languages with the corresponding distributions. In addition, to the use of labels information that is hidden in the source domain dataset without using any pseudo-labels for target samples, another version of PSHLA with name ‘Parallel Shared Encoders as a Hidden Layer for an Auto-encoder with Residual connection and Classification task’ (PSHL-RC) is proposed. The residual connection re-injects previous representation into downstream of data by adding the output of lower layer into upper layer. Finally, the effectiveness and efficiency of the proposed method experimentally is evaluated on Persian and some other emotional speech databases. Therefore, the main contributions of this paper can be summarized as follows:

Most of researches in the field of emotional speech recognition are based on the existence of a sufficient number of labeled data. This is the major challenge in recognizing emotional speech in languages where there is not enough labeled data. Therefore, a new structure based on deep neural networks is proposed in this paper to be used as a cross-language domain adaptation problem to adapt the distribution of data which can reach state-of-the-art accuracy with only a few unlabeled data.

In comparison with similar methods of domain adaptation for emotional speech recognition that use one auto-encoder with one shared hidden layer to construct new features in the output of hidden layer, the proposed PSHL uses multiple shared parallel encoders inside of auto-encoders to extract fusion features. Therefore, the final output of SHLA is qualified discriminative representations that will be used for speech emotion recognition.

From our point of view, this is the first time that emotional speech recognition in Persian language was considered using domain adaptation. In this case, Persian emotional speech was considered as the target and labeled emotional speech datasets in other languages were considered as the source.

The remainder of this article is as follows. Section 2 describes the relevant research. In the third section, the proposed method is presented. In the fourth section, the experimental settings and test results for the four datasets are described. Finally, the conclusions are presented in Section five with future works.

2. RELATED WORK

Although cross corpus and cross language emotional speech recognition is an interesting issue, relatively few studies have addressed the subject. Existing studies have mainly studied the feasibility of cross corpus learning and pointed to the need for deeper research. Deep learning approaches are widely used to domain adaptation in speech recognition. In deep neural network, auto-encoders are used to find common features from input samples [23]. This new representation increases the performance of speech emotion recognition systems. In this field, the distribution difference across training and test samples is not considered [24, 25]. To overcome this challenge, auto-encoders are used as an unsupervised learning model and have been very successful in dealing with imbalance in the distribution of training and test samples in emotional speech recognition [13, 26]. In other words, auto-encoders are used to extract a common representation across different domains in unsupervised methods.

Denoising Auto-encoder (DAE) is a type of basic auto-encoder that is used to reconstruct clean samples from noisy samples [27]. In this case, the trained network should be able to extract the distribution of underlying structure of input samples in order to reduce the effect of the destructive process that has caused the samples to be noisy. Therefore, it is necessary to learn robust features compared to the basic network. Based on this ability of the DAE network, transfer learning has been proposed for the source and the target domains adaptation with application in the emotional speech classification [4, 19]. In this case, a DAE is trained using the unlabeled target domain data. Then, using the encoder part of the trained network, a new representation for source and target data is constructed. By this approach, it is expected that the difference between the source and the target data, would be decreased.

SHLA is an alternative structure of auto-encoder that tries to minimize the reconstruction error in both source and target sets simultaneously. It shares the same parameters for the mapping from the input layer to the hidden layer, but it uses independent parameters in the process of reconstruction [14].

Fig. 1. demonstrates the structure of SHLA for two input sets from source and target domains. The outputs $\widetilde{\boldsymbol{X}}^{s}$ and $\widetilde{\boldsymbol{X}}^{t}$ are the reconstructed input sets.

In detail, this method trains an SHLA using source and target data in unsupervised manner. The result of this training is to create a balance between mismatch samples in two different sets. Subsequently, this method yields a new representation on two datasets by using the encoder part of SHLA and trains a supervised classifier on new representation of the source dataset. Finally the classifier is tested using the new representation of the target dataset [14].

Motivated by the success of SHLA, an analytical approach based on Kernel Canonical Correlation Analysis (KCCA) for domain adaptation has been proposed in [28]. This method tries to obtain a new representation of features like the SHLA method using data from both domains in order to cross lingual emotional speech recognition. Although this study is not based on deep neural networks but due to success of KCCA in increasing the accuracy of emotional speech recognition compared to SHLA, we have compared the efficiency of our proposed method with its efficiency.

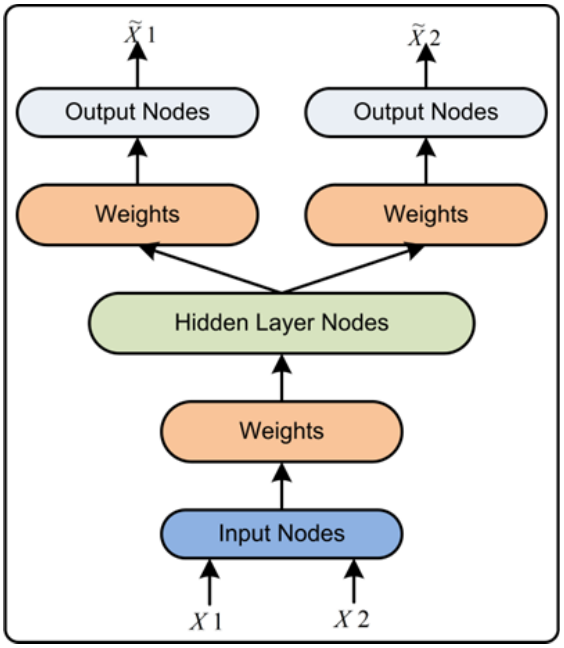

Fig. 1. Structure of SHLA. It shares same parameters from input layer to hidden layer for mappings of input sets, but uses independent parameters for reconstruction.

In KCCA approach, two mappings of the source data $\boldsymbol{X}^{Px}$ and $\boldsymbol{X}^{Py}$ are obtained based on its principle components and the principle components of the target data respectively. Similarly, two mappings of the target data $\boldsymbol{Y}^{Px}$ and $\boldsymbol{Y}^{Py}$ are obtained. Then using the method KCCA, a shared view between the paired mapped data on the source principal components, $[\boldsymbol{X}^{Px};\boldsymbol{Y}^{Px}]$ and the target principal components $[\boldsymbol{X}^{Py};\boldsymbol{Y}^{Py}]$ are extracted. Finally, a classifier is trained using the mapped training data and it is tested on the mapped testing data. KCCA maximizes the correlation between the mappings of kernels.

'Feature transfer learning in subspace' named as DAE-NN is another approach to cope with the inherent difference between the source data and the target data in emotional speech recognition [26]. In this approach, two subspaces are first created by separately training of two DAEs using the source and target data. A high-order feature representation of the target dataset is then obtained on subspaces. These new representations are used to train a regression neural network (NN) to discover the difference between them. It is expected that the regression neural network can compensate the disparity between the source and target domain. Then using a NN, high-order feature representation of the source data is estimated on target subspace. These new features are used to train a classifier on target datasets by applying a supervised learning algorithm (such as SVM).

Deep Belief Networks (DBNs) are another type of deep architectures that have been used for transfer learning in the field of speech emotion recognition. The key reason for using DBNs is their generalization power. Because, their constituent blocks are universal approximators that are very powerful to approximate any distribution [29]. In other words, DBNs consist of the stack of Restricted Boltzmann Machines (RBMs). By training RBMs layer-wise, a probabilistic generative model is created. In [30] a DBN with three RBM layers has been proposed as transfer learning for cross language emotional speech recognition. The first two RBM layers contain 1000 hidden units and the last one has 2000 hidden units. To evaluate the effectiveness of DBN, the authors have used FAU AES dataset as training and EMO-DB, EMOVO and SAVEE datasets as testing. The results have been shown that DBN outperforms sparse auto-encoders.

3. PROPOSED TRANSFER LEARNING METHOD FOR EMOTIONAL SPEECH RECOGNITION

3.1. System Architecture

Due to the success of the auto-encoders in providing a solution to the distribution mismatch in the field of emotional speech recognition, the PSHLA and the PSHL-RC are proposed to extract common hidden attributes across the source and the target domain and they are used as a domain adaptation in the emotional speech recognition structure. Before domain adaptation and classification, the features set pre-processed.

Fig. 2. shows architecture of emotional speech recognition system, which includes pre-processing, domain adaptation and recognition sections. Based on this architecture first, the maximum and minimum of each feature on training set is determined and used to map the source and

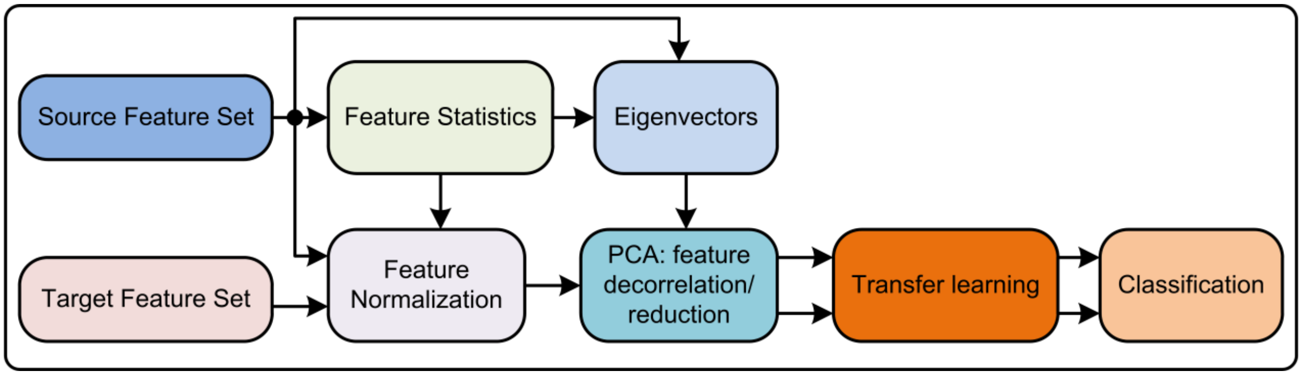

Fig. 2. The architecture of emotional speech recognition system, which includes preprocessing, transfer learning and recognition sections. Inputs are the source features set: $\widetilde{X}^s$ and target features set: $\widetilde{X}^t$. Outputs are the labels belonging to the positive or negative valence classes.

target set onto interval [0, +1]. This normalization is necessary to prevent the saturation of activation function. Then, using the Principal Component Analysis (PCA), the source set and the target set are de correlated. For this, the eigen values and eigenvectors of source set are only calculated. Furthermore, the PCA is used to reduce the dimension of the feature vector. Subsequently, using PSHLA, the domain adaptation takes place between the distribution of the features of the source and the target domain. In other word, PSHLA is trained with the source and target samples, simultaneously.

Because the weights of encoders are shared between the source and target samples features, then the features of two domains are paired to each other by PSHLA to produce discriminative features that are common for two domains with similar distribution. Details on PSHLA are described in the flowing subsection. Finally, by training a conventional classifier (in this paper: SVM) with labeled source samples, the target domain samples are classified.

3.2. PSHLA and PSHL-RC as Transfer Learning for Emotional Speech Classification

In Fig. 3, a schematic of PSHLA and PHLA-RC are presented, which consist of three parallel shared hidden layers and two independent reconstructed outputs. Originally, the structure of PSHLA is made of an auto-encoder which consists of three encoders in parallel form, two decoders to reconstruct the outputs and an extra layer to concatenate the output of encoders (Fig. 3a). This new auto-encoder structure is used as a transfer learning module that uses the same weights in the encoder layers for discovering common new representation from two different input examples; promising to improve speech emotion recognition performance.

In the encoding phase, given an input $\boldsymbol{x}$, the encoder in PSHLA is defined as follows:

$$\boldsymbol{h}_e = concatenation(\boldsymbol{h}_1^{EN1}, \boldsymbol{h}_1^{EN2}, \boldsymbol{h}_2^{EN3}) \qquad (1)$$

where

$$\boldsymbol{h}_1^{EN1} = f(\boldsymbol{W}_1^{EN1}\boldsymbol{x} + \boldsymbol{b}_1^{EN1})\,. \qquad (2)$$

$$\boldsymbol{h}_1^{EN2} = f(\boldsymbol{W}_1^{EN2}\boldsymbol{x} + \boldsymbol{b}_1^{EN2})\,. \qquad (3)$$

$$\boldsymbol{h}_2^{EN3} = f(\boldsymbol{W}_2^{EN3}\boldsymbol{h}_1^{EN3} + \boldsymbol{b}_2^{EN3}), \qquad (4)$$

$$\boldsymbol{h}_1^{EN3} = f(\boldsymbol{W}_1^{EN3}\boldsymbol{x} + \boldsymbol{b}_1^{EN3})\,. \qquad (5)$$

The matrix $\boldsymbol{W}_i^{EN_j}$ and the vector $\boldsymbol{b}_i^{EN_j}$ are referred to as weights and bias, respectively and superscript 'EN_j', indicates the encoder 'j'.

In the decoder phase, the decoder uses the output of hidden layer $\boldsymbol{h}_e$ to reconstruct the original inputs:

$$\widetilde{\boldsymbol{x}}_1 = f(\boldsymbol{W}_3\boldsymbol{h}_e + \boldsymbol{b}_3)\,. \qquad (6)$$

$$\widetilde{\boldsymbol{x}}_2 = f(\boldsymbol{W}_4\boldsymbol{h}_e + \boldsymbol{b}_4)\,. \qquad (7)$$

where $\widetilde{\boldsymbol{x}}_1$ and $\widetilde{\boldsymbol{x}}_2$ are reconstructed inputs belong to the source and target datasets respectively. The cost function in equation (8) is used in this study to optimize a joint distance of two different source and target examples, as follows [14]:

$$J(\boldsymbol{\theta}) = J^{s}(\boldsymbol{\theta}^{s}) + C\, J^{t}(\boldsymbol{\theta}^{t}), \tag{8}$$

where:

$$J^{s}(\boldsymbol{\theta^{s}}) = \textstyle\sum_{\boldsymbol{x}\in \boldsymbol{X}^{s}} \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|^{2}, \tag{9}$$

$$J^{t}(\boldsymbol{\theta^{t}}) = \textstyle\sum_{\boldsymbol{x}\in \boldsymbol{X}^{t}} \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|^{2}, \tag{10}$$

where $\boldsymbol{X}^{s}$ and $\boldsymbol{X}^{t}$ are the source and target dataset respectively and $C \geq 0$ is used to control the trade-off between the labeled and unlabeled objectives. By this term, PSHLA is able to insert information from the target set to the training parameters [14].

To make the maximum use of information contained in the source domain examples in some methods for domain adaptation, classification task is also used with reconstruction task [21, 22]. In this method a pseudo-labels are used for target samples. According to the success of such idea in using source labels and using pseudo-labels for target samples, we also used third task as classification task in our proposed method as variant of PSHLA with the name PSHLA-RC (Fig. 3b), which labels information of source's examples are used to create common new representation between two domains as well as unlabeled data information. In this new variation, we did not use any pseudo- labels for target samples in comparison with the methods explained in [21, 22]. As a result, be any miss labeling in target samples cannot affect the other labels in training time. Equation (11) shows the common objective function that should be optimized in this case. Therefore, it is ensured that the reconstruction error of unsupervised learning task as well as the prediction error of supervised learning task would be minimized on both the labeled and unlabeled data. In addition, a residual connection from low layer into upper layer along classification task branch is used to prevent exploding and vanishing gradients that may occur during training.

$$J(\boldsymbol{\theta}) = J^{s}(\boldsymbol{\theta}^{s}) + BJ^{Classs-labels}(\boldsymbol{\theta}) + C\, J^{t}(\boldsymbol{\theta}^{t}), \tag{11}$$

where B is used to control the impact of the different losses and $J^{Class-labels}$ is defined as follows:

$$J^{Class-labels}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(p(1 - y_i)) \tag{12}$$

where y is the label (1 for positive valence samples and 0 for negative valence samples) and $p(y)$ is the predicted probability of the sample being positive valence for all N input samples.

3.3. Analytical Discussion

DAE method does not attempt to employ useful information from the source domain data and forces them to create a new representation using the characteristics of the target domain. For this reason, some examples of the source domain may not follow these characteristics and therefore the useful information required for the classifier is loosed. This causes the accuracy of classifier to be decreased. Even worse, negative transfer learning may occur [17].

However, DAE based transfer learning approach is used in domain adaptation since it is a simple and efficient method.

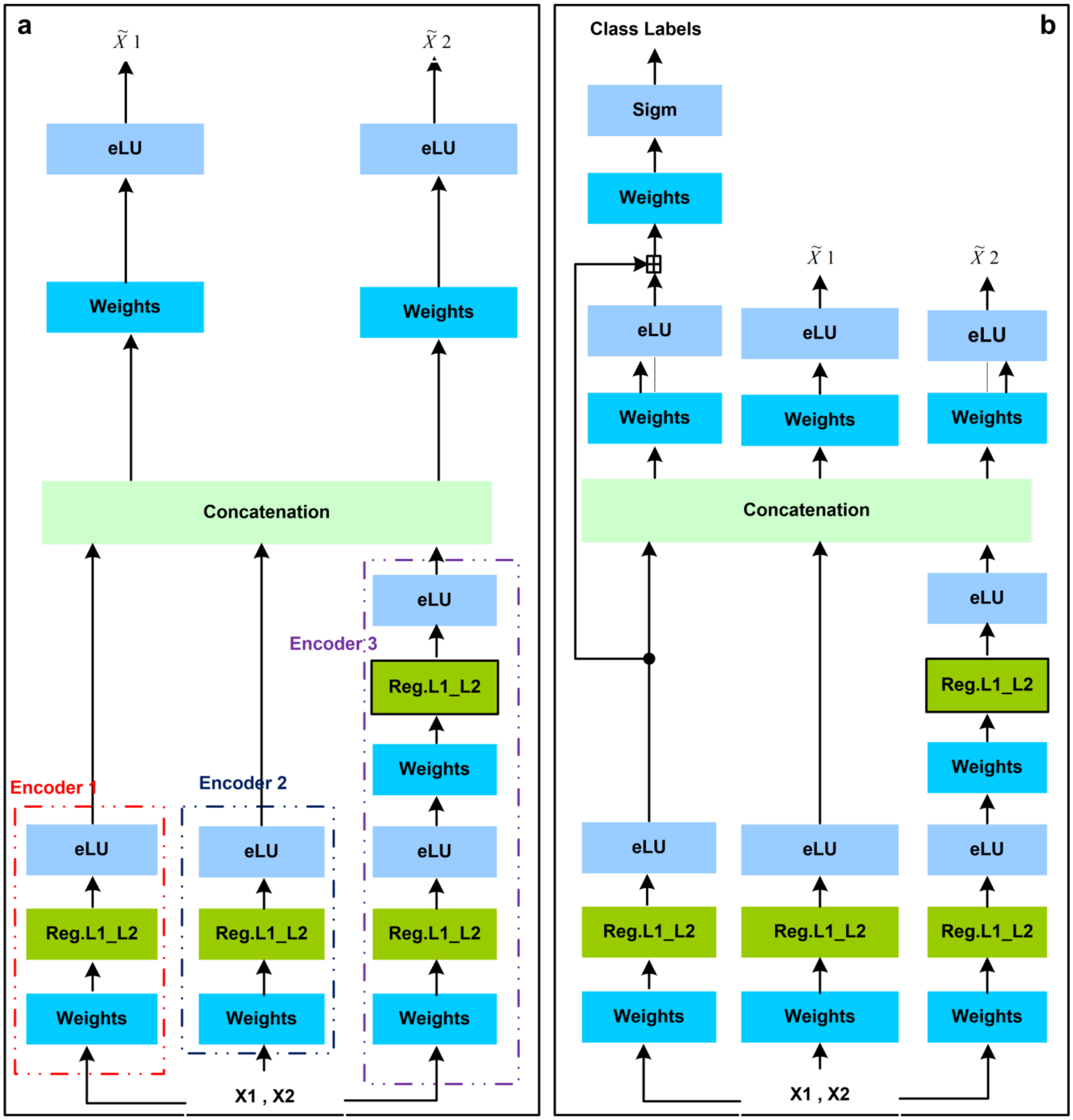

Fig. 3. Architecture of our proposed transfer learning methods. $X1 \in \{X^s\}$ *and* $X2 \in \{X^t\}$*. (a): Parallel Shared Hidden Layer Auto-encoder (PSHLA). (b): Parallel Shared Hidden Layer Auto-encoder with Residual connection and Classification task (PSHLA-RC).*

In contrast to DAE, the transfer learning based on the SHLA utilizes the information available in the source domain samples. Hence, occurrences of negative transfer learning are not possible in comparison with transfer learning based on DAE. Further, sharing information in both domains to create a new representation for the input samples will reduce the difference between the domains and, as a result, the classification accuracy increases. However, if the samples have a significant difference (e.g. Emotional speech expressed in two languages with different linguistic structure, the difference in the duration of emotional utterances in two domains), the SHLA method stated in [14] is not successful in this case. Therefore, SHLA will not be able to extract very complex nonlinear relations between the samples in the source and the target domains. To succeed, it is necessary to increase the number of layers and the number of nodes in the network.

The advantage of KCCA method compared to the SHLA method is its analytical solution. Hence, there is no possibility of falling to local minima and its learning process is faster compared to gradient descent based algorithms. However, if the large number of layers and nodes are used in SHLA, the nonlinear relation in data distribution could be extracted better [28] . However, to represent this nonlinearity, there is a need for a lot of data samples in the training phase.

Transfer learning with the DAE-NN method can reduce the difference between the source domain and the target but it has no significant success compared to other methods such as SHLA. To increase the effectiveness of this method, the number of layers and the nodes is required to increase and this may cause over-fitting in training time; Especially if the number of training data to be sparse. In addition, the complexity of its implementation (execution) is high compared to the SHLA method.

Using DBNs can be effective in reducing source and target domain distribution. However, the main weakness of these networks is the need for a large number of training examples. Labeling such a large number of examples reduces reliability, because different people may have a different understanding of a particular feeling.

On the other hand, the proposed method of this paper has all the advantages of the SHLA and KCCA based transfer learning methods. The encoder part of proposed method consists of several simple parallel AE networks that use both source and target samples information to extract common features between two domains. At the same time, with a new architecture, it does not need a large number of instances when the linguistic structure of the languages and duration of utterances are different. Thus, using small number of sample instances, PSHLA and PSHLA-RC can extract new representations of features that are good attributes for emotions and discriminative for classification. In other words, each auto encoder with a fixed number of nodes in the hidden layer can extract certain attributes. Therefore, two or more parallel auto encoders can extract more efficient features. Note that increasing the number of parallel encoders also increases the need for more training samples. So there is a trade-off between the number of training samples and the choice of the number of parallel encoders. In addition, PSHLA-RC uses the label of source domain samples to produce new representation of features in contrast to other unsupervised domain adaptation methods based on auto-encoders. Thus, the use of multiple simple encoders in parallel and the use of source data labels is the most fundamental difference between the proposed method and other methods to overcome previous method’s weaknesses.

4. EXPERIMENTS AND RESULTS

In this section, the experimental results for recognition of emotional speech in one language using disjoint corpora in other languages as the training set based on the proposed PSHLA and PSHL-RC representation learning are provided.

4.1. Task and Data

To evaluate the performance of the proposed method, we compared the results using six models of transfer learning on five emotional speech databases. Two of the domain adaptation methods are the proposed approaches. The three other approaches are based on Kernel PCA, Kernel CCA, and SHLA [14, 28]. The last one is DBN. The process is rune for ten times and the average of the results is provided. Furthermore, we use PSHLA and PSHLA-RC to classification of Persian emotional speech. These datasets have been selected to cover a wide range of languages. Details about these corpora are tabulated in Table 1.

The German emotional speech (EMODB) dataset consists of six basic emotions (anger, joy, sadness, fear, disgust, and boredom) plus the neutral speech that have been simulated by five females and five males native German actors [31]. Each actor has uttered five short sentences and five long sentences in each of the emotional states. The sentences are interpretable in all applied emotions [32].

The SAVEE database consists of 480 British English utterances which four male actors recorded them in seven different emotions. ten subjects evaluated the recordings under audio, visual, and audio-visual to investigate the quality of performance [33].

The EMOVO is a database contained six actors who uttered 14 sentences simulating six emotional states (disgust, fear, anger, joy, surprise, sadness) as well as the neutral state in the Italian language [31].

The FAU Aibo Emotion Corpus (FAU AEC) is a dataset that includes nine hours of German speech. It contains the emotional speech of 51 children from two different schools at age of 10 to 13 years interacting with the pet robot Aibo [34]. This dataset covers emotional utterances in two classes: IDLe (Motherese, Joyful, Neutral, Rest) and NEGative (Angry, Touchy, Emphatic, Reprimanding).

Table 1. ***Some information of chosen datasets.***

Dataset	Language	Emotion	Negative Valence (#)	Positive Valence (#)	# m/f
PESDB	Persian	acted	Anger, Sadness, Fear, Disgust (234)	Neutral, Happiness (238)	1/1
EMODB	German	acted	Anger, Sadness, Fear, Disgust (352)	Neutral, Happiness (142)	5/5
SAVEE	English	acted	Anger, Sadness, Fear, Disgust (240)	Neutral, Happiness, Surprise (240)	4/0
EMOVO	Italian	acted	Anger, Sadness, Fear, Disgust (336)	Neutral, Joy, Surprise (252)	3/3
FAU-AEC	German	natural	Angry, Touchy, Emphatic, Reprimanding (5823)	Motherese, Joyful, Neutral, Rest (12393)	21/30

Number of utterances per binary valence (# Valence, Positive (+), Negative (-)), Number of female (#f) and male (#m) subjects.

Table 2. ***Overview of the standard feature set provided by the INTERSPEECH 2009 Emotion Challenge.***

LLDs(16×2)	Functional (12)
(Δ)ZCR	Mean
(Δ)RMS Energy	Standard deviation
(Δ)F0	Kurtosis, skewness
(Δ)HNR	Extremes: value, rel, position, range
(Δ) MFCC 1-12	Linear regression: offset, slope, MSE

The Persian Emotional Speech Database (PESD) contains five basic emotions (anger, disgust, fear, happiness, and sadness) plus a neutral. Two native Persian speakers (one man, one woman) have uttered 90 sentences in congruent, incongruent, and basic condition. A group of 34 native speakers has evaluated the validity of the database in a perception test [35].

4.2. Acoustic Features

We used the feature set of the INTERSPEECH 2009 Emotion Challenge in this study for the proposed methods [36]. This collection includes 384 features extracted by applying 12 functions to 32 acoustic Low-Level Descriptors (LLDs), according to. In detail, The LLDs are zero crossing-rate, root mean square of frame energy, pitch frequency, harmonics-to-noise ratio by autocorrelation function and Mel-frequency Cepstral coefficients 1-12. The 12 functional features are minimum, maximum, mean, standard deviation, kurtosis, skewness, relative position, ranges, and two linear regression coefficients with their mean square error. To ensure the reproducibility as well, the open source openSMILE toolkit was used [37]. Additionally, we applied PCA to eliminate those principal components that contribute less than 2% to the total variation in the feature set.

4.3 Experimental Setup and Evaluation Metrics

Since the databases used in this study, are annotated differently; therefore, we used the binary valence mapping per emotion category from [14, 28] as the evaluation metric, because it is the best consistent metric to investigate and compare different feature transfer learning methods.

The proposed architecture has many parameters that must be determined. For this purpose, the grid search was used to search the number of hidden nodes over {20,40,80,160} and the number of parallel encoder over {2,3,4} and the hyper-parameters C over {0.5,1,2} in the form of 5-fold cross validation. In order to reduce the search time as well as to prevent the occurrence of over-fitting, we fixed the number of hidden layers equal to one in enoder1 and encoder2. For the third encoder, the number of hidden layers is fixed to two.

For network training initialization, the Xavier uniform initializer was used [38]. We used Adam optimizer with common parameters indicated in [39]. Early stopping strategy was employed for network training. Furthermore, the maximum epochs were limited to 150. In addition, to balance the training samples between positive and negative classes, Synthetic Minority Oversampling Technique (SMOTE) was used [40].

As a classifier, SVM with sigmoid function as kernel was used. In order to determine the hyper parameters of SVM, a grid search was performed based on independent validation set selected from the labeled source set. The pair of parameters that gives the best result was chosen as final SVM parameters. The toolkit LIBSVM [41] was applied in the experiments.

The performance of PSHLA and PSHLA-RC were evaluated in two scenarios. In the first scenario, the performance was obtained for the proposed PSHLA and PSHLA-RC methods compared with SHLA, KPCA, and KCCA that had been reported in [28]. In addition, we compared the performance of our proposed methods with DBN method. In the second scenario, the proposed PSHLA and PSHLA-RC methods were used as a transfer learning approach to recognize Persian Emotional Speech. To compare the obtained performance, the un-weighted average recall (UAR) metric was used [36]. Because it is the officially-recommended criterion for paralinguistic tasks [36]. UAR is defined as follows:

$$. \; UAR = \frac{\sum_{c}^{C} \frac{\left|\left\{ x \in X_c^{te} \,\middle|\, y = c \right\}\right|}{N_c}}{C}, \qquad (13)$$

where $\mathrm{N_c}$ is the total number of test samples, belonging to the test set $\boldsymbol{X}_c^{te}$, and C is the number of classes. Based on this idea, the accuracy was evaluated for each class separately and then the mean of these values across all classes is given as the final accuracy. Further, to validate the statistical significance of the obtained results, one-sided z-test has been used.

4.4. Results and Discussion

To demonstrate the performance of the proposed transfer learning method, it is necessary to compare its efficiency with the Cross-Training (CT) test as a baseline. In this case, a specific dataset in one language is used for training an SVM classifier. Then, the other dataset in the other language is classified with this trained classifier. For this experiment, we used three datasets: EMODB, SAVEE, and EMOVO. The results of CT method along with other methods are outlined in Table **3**, Table 4, and **Table 5**. For better evaluation of the performances, the error reduction rate [27] is calculated and presented in these tables. ERR is defined as follow:

$$ERR = \frac{error_{CT} - error_{new\ method}}{error_{CT}}, \qquad (14)$$

where $\mathrm{error_{CT}}$ is the error of Cross Training approach and $\mathrm{error_{new\ mthod}}$ is the error of the PSHLA or PSHLA-RC methods.

According to Table **3**, CT method obtained average UAR only around chance level (55.8% and 56.4%) for SAVEE and EMOVO respectively, when EMODB is used for training (source dataset) and two other datasets are used for testing (target datasets). SHLA achieved a 63.7% UAR and 56.8% UAR for SAVEE and EMOVO, respectively. KCCA transfer learning methods increased these accuracies to 65.2% and 57.9%. In addition, these methods (SHLA and KCCA) have not been able to increase the accuracy sufficiently compared to CT method for

Table 3. ***Cross-languages average UAR and ERR. EMODB has been used for the training set EMOVO and SAVEE have been used for test sets. ERR has been calculated based on CT method.***

		Target Datasets			
		SAVEE		**EMOVO**	
		UAR [%]	**ERR [%]**	**UAR [%]**	**ERR [%]**
Methods:	**CT**	55.8	0.00	56.4	0.00
	KPCA[28]	64.4	19.46	57.6	2.75
	KCCA[28]	65.2	21.27	57.9	3.44
	SHLA[28]	63.7	17.87	56.8	0.92
	PSHLA	65.47	21.88	58.26	4.27
	PSHLA-RC	65.4	21.72	60.43	9.24

Table 4. ***Cross-languages average UAR and ERR. SAVEE has been used for the training set and EMODB and EMOVO have been used for test sets. ERR has been calculated based on CT method.***

		Target Datasets			
		SAVEE		**EMOVO**	
		UAR [%]	**ERR [%]**	**UAR [%]**	**ERR [%]**
Methods:	**CT**	62.5	0.00	54.6	0.00
	KPCA[28]	70.1	20.27	58.7	9.03
	KCCA[28]	71.9	25.07	59.5	10.79
	SHLA[28]	67.7	13.87	59	9.69
	PSHLA	73.44	29.17	60.57	13.15
	PSHLA-RC	73.1	28.27	62.03	16.37

3

EMOVO as testing. In the other hand, proposed method PSHLA-RC, boosted the average UAR to 60.43% compared with CT, SHLA, and KCCA, when EMOVO was used for testing dataset. Note that, SHLA is a homogeneous competitive method compared to our PSHLA. KPCA and KCCA are two state-of-the-art methods that are referred to in this article for better comparison. In addition, the proposed method has a new architecture that is different from SHLA and, to maintain the same experiments conditions with the competitive methods, the same datasets were used for SHLA and PSHLA evaluation.

Table 4 presents the results when SAVEE is used for training and two other datasets are used for testing. In this case, CT method obtained a 62.5% UAR for EMODB as testing. SHLA and KCCA boosted the accuracy to 67.7% and 71.9% respectively, while proposed PSHLA achieved a 73.44% UAR, which is slightly higher than a 73.10% UAR obtained by PSHLA-RC.

According to Table 4, when EMOVO was used as testing dataset, CT method achieved a 54.6% UAR, which is near the chance level. SHLA and KCCA obtained an average UAR of 59.0% and 59.5% respectively, while the proposed PSHLA achieved a 60.57% UAR and the proposed PSHLA-RC booted the average UAR to 62.03%.

Table **5** shows the results when EMOVO is used for training. In this case, CT method achieved the average UAR of 58.0% for EMODB and 51.2% for SAVEE. SHLA obtained a 67.3% UAR and 58.2% UAR for EMODB and SAVEE, respectively. While, proposed PSHLA method achieved a 69.03% UAR for EMODB and 60.27% UAR for SAVEE. In addition, using labels of source dataset (EMOVO) in training of PSHLA-RC, the average accuracy improved to 69.40% and 61.39%, for EMODB and SAVEE, respectively.

From the results, it is obvious that the distribution difference between EMOVO and two other datasets (EMODB and SAVEE) is high, that may be due to the difference in the cultures of nations and the linguistic differences. In addition, in the case of using EMODB (or SAVEE) as source set and SAVEE (or EMODB) as target set, the average accuracy of PSHLA-RC has been slightly reduced. This can be due to the similarity of the distribution of domains, where the average accuracy is high and the labels of source set have no more information. In this case over-fitted network may be occurred due to the low number of training samples.

Table 5. Cross-languages average UAR and ERR. EMOVO has been used for the training set EMODB and SAVEE have been used for test sets. ERR has been calculated based on CT method.

		Target Datasets			
		SAVEE		**EMOVO**	
		UAR [%]	**ERR [%]**	**UAR [%]**	**ERR [%]**
Methods:	**CT**	58	0.00	51.2	0.00
	KPCA[28]	66.7	20.71	56	9.84
	KCCA[28]	62.9	11.67	58.5	14.96
	SHLA[28]	67.3	22.14	58.2	14.34
	PSHLA	69.03	26.26	60.27	18.59
	PSHLA-RC	69.4	27.14	61.39	20.88

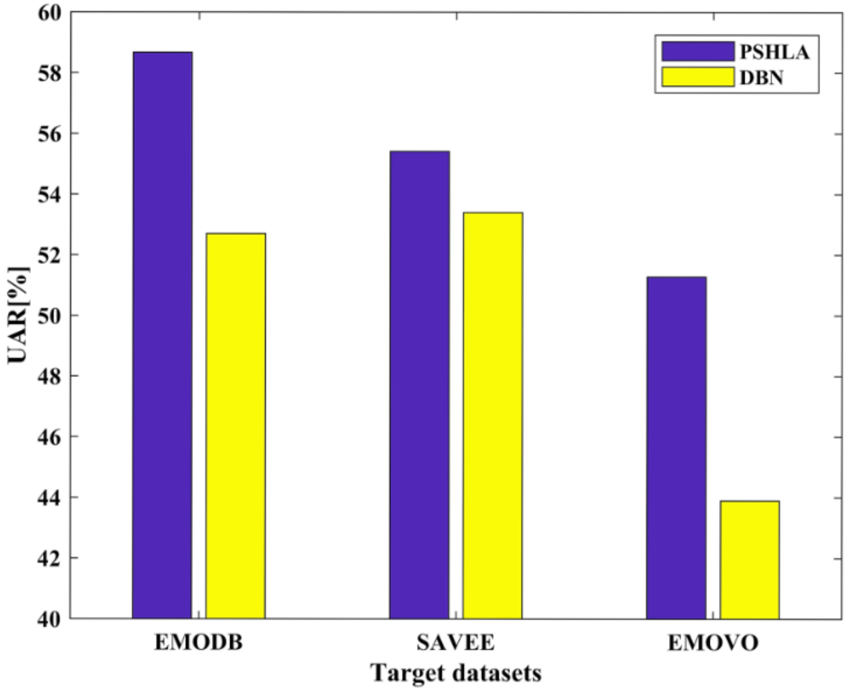

Fig. 4. Efficiency of DBN and PSHLA transfer learning methods. FAU-AEC has been used as source dataset.

As can be seen in all tables, based on ERR, the proposed PSHLA and PSHLA-RC methods are more effective than all the other approaches. However, on average, SHLA yields 13.14% error reduction with respect to CT. Moreover, KPCA and KCCA yield 13.68%, 14.58% error reduction on average over CT, respectively. In addition, the proposed PSHLA and PSHLA-RC methods reduced the error by 18.89% and 20.60%, respectively. The ERR of PSHLA-RC methods indicates that there is some useful information within the source dataset labels that can be effective in reducing the discrepancy of source and target domain distribution.

In another experiment, FAU-AEC dataset was used as source dataset to compare the performance of the proposed method PSHLA with the performance of the DBN method. Hence the three other datasets (EMODB, SAVEE, and EMOVO) were used for target datasets. The results are demonstrated in Fig. 4. According to these results, the proposed PSHLA improved the average accuracy by 4.97% in average. The reason for this improvement in recognition accuracy by PSHLA is that, in the DBN method compared to the proposed PSHLA method, there are many parameters (layer’s weights) that must be determined in training phase while, the number of training samples is small. Therefore, DBN cannot extract the complex relationships between the source and target domains. In contrast, in the encoding method, the number of parameters that need to be trained is very low.

According to Fig. 5, the average accuracy obtained for EMOVO is less than two other datasets (EMODB and SAVEE) and the average accuracy obtained for EMODB is high than two other datasets. These results confirm that the linguistic similarity between the Italian and German languages is low, as mentioned earlier. In general, inter-language similarity plays a key role in determining recognition accuracy using transfer learning methods. Hence, transfer learning methods should, as a matter of fact, be able to minimize inter-language differences as much as possible to increase recognition accuracy. Thus, our proposed method has been able to reduce the inter-linguistic differences more than other methods.

In the second scenario, we examined the proposed PSHLA and PSHLA-RC as a transfer learning approach to adapt PESD dataset with the other datasets. In this scenario, each time one of the datasets was used as the source data, and the PESD dataset was used as the target. PSHLA was trained ten times with different seeds for initialization. This scenario was repeated for SHLA, too. In Fig. 5, the results of the average UAR over ten trials are visualized, including the error bars. As demonstrated in Fig. 5, the proposed methods are more efficient than SHLA. For example, when EMODB was used as source dataset, SHLA obtained a 67.22% UAR and the proposed PSHLA and PSHLA-RC obtained a 72.77% UAR and 73.43% UAR, respectively. These improvements have statistical significance at $p < 0.01$ with a one-sided z-test compared to SHLA.

When SAVEE was used as source dataset, SHLA obtained a 67.18% UAR. The proposed PSHLA and PSHLA-RC methods achieved a 69.51% UAR and 70.44% UAR, respectively. These improvements have statistical significances at $p < 0.02$.

While both databases EMODB and SAVEE are used as source dataset simultaneously, the average UAR boosted to 78.68% and 80.67% by PSHLA and PSHLA-

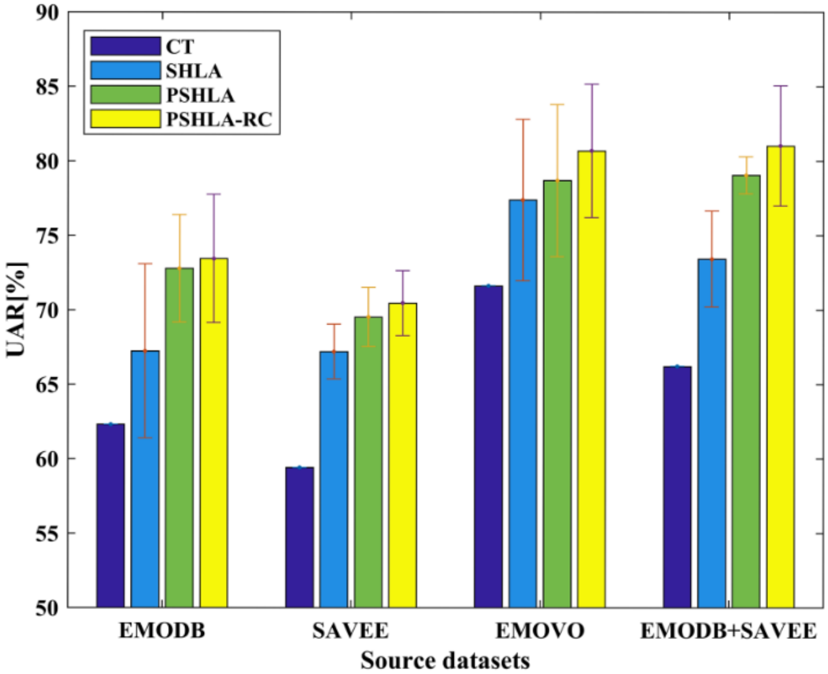

Fig. 5. Efficiency of transfer learning methods over ten trials. PESD has been used as target dataset.

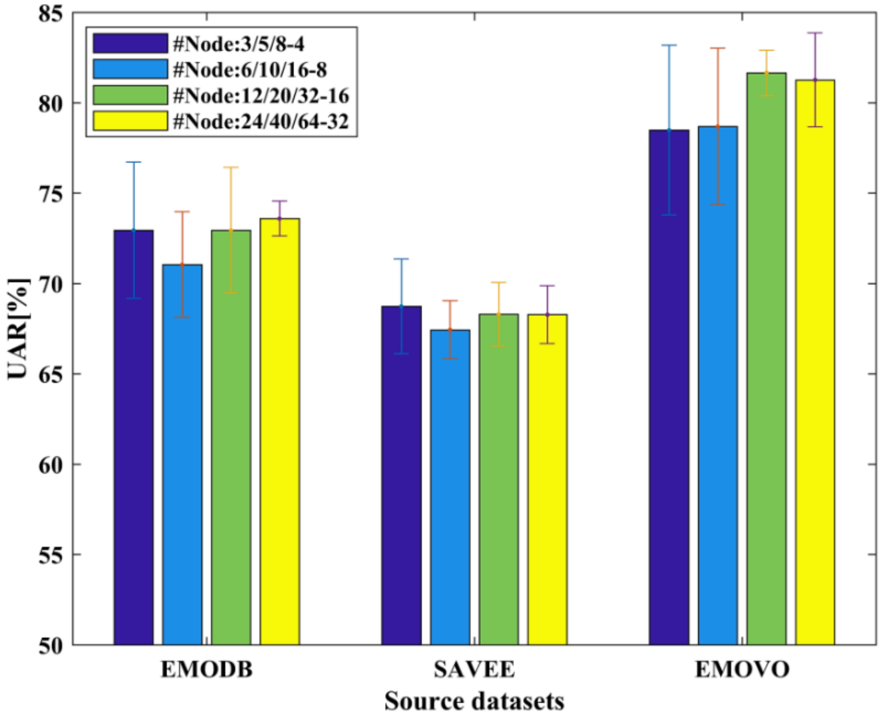

Fig. 6. effect of increasing the number of nodes in hidden layers of PSHLA. #Nodes indicate the number of nodes in hidden layer/layers of encoder1/encoder2/encoder 3 (see Fig. 3).

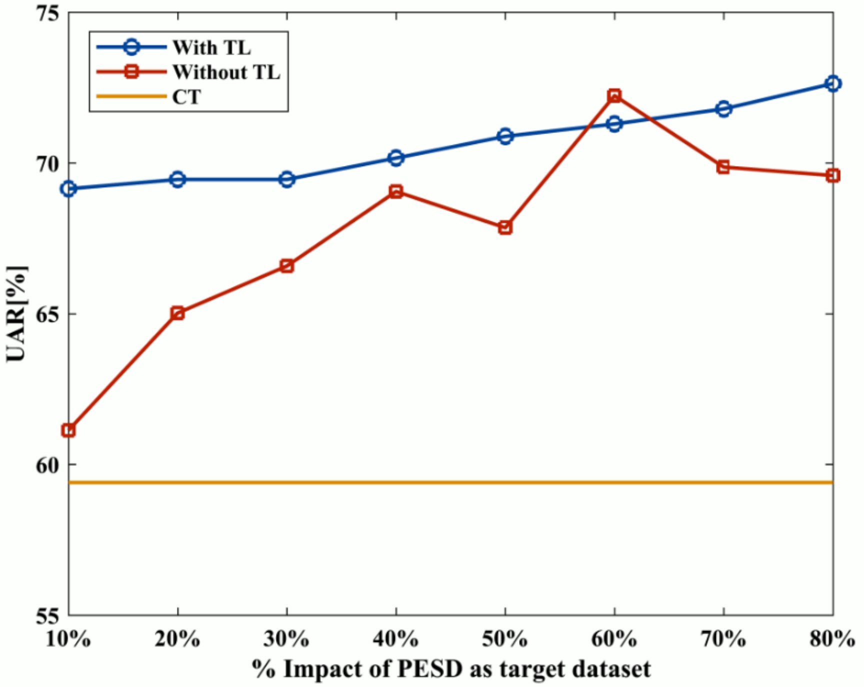

Fig. 7. effects of using a percentage of PESD as target dataset with source dataset SAVEE.

RC, respectively. These improvements are much greater than the time; EMODB and SAVEE were used separately as source dataset. One reason for this improvement is the increasing in the number of training samples. Furthermore, as the linguistic diversity increases, the new representation features at the output of the hidden layer become more specific to minimize the discrepancy between source and target domains.

According to Fig. 5, on the EMOVO dataset, the proposed PSHLA and PSHLA-RC methods give an average UAR of 79.04% and 81.01%, respectively, which is significantly higher than the maximum average UAR obtained by SHLA. In addition, these improvements pass the significant test at $p < 0.03$ and $p < 0.01$, respectively.

By analyzing the results of the second scenario, it can be observed that SAVEE dataset has less similar features with Persian language because of its constituent utterances that are uttered by males. In contrast, in PESD dataset, both male and female speakers are present. In addition, English language may be less similar to Persian language. Therefore, training by SAVEE and testing by PESD has less accuracy than other training datasets. On the other hand, training by EMOVO and testing by PESD has better accuracy than other datasets. This shows that emotional utterances in Italian and Persian languages seem to be very similar. It is also clear that the combination of two databases is effective in increasing the accuracy of recognition.

Fig. 7. shows the effect of increasing the number of nodes. The results show that an increase in the number of nodes can increase the overall accuracy in most cases. But increasing more will not affect the result significantly.

In another experiment, a semi-supervised TL approach was examined using a percentage of target domain samples (10% to 80%) along with the source domain samples. In this case, a percentage of the Persian dataset after domain adaptation was used with the source dataset (e.g. SAVEE) to train a classifier. This experiment was performed before domain adaptation, too. Based on the results (see Fig. 7), it is clear that after transfer learning, the similarity of the distribution of the domains has been increased and therefore, recognition accuracy has been improved.

5. CONCLUSION

In this paper, an innovative auto-encoder based common feature representation approach is presented to compensate the difference between source and target data. This method was used to recognize the emotional speech between different languages. The experimental results on five publicly available datasets revealed that the proposed method effectively and sufficiently improves the accuracy of the recognition of emotional speech compared to the relative transfer learning methods. In the future, we plan to extend this approach to semi-supervised modes using recursive deep neural networks.

REFERENCES

[1] M. Swain, A. Routray, and P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, *International Journal of Speech Technology,* vol. 21, pp. 93-120, 2018.

[2] T. L. Nwe, S. W. Foo, and L. C. De Silva, Speech emotion recognition using hidden Markov models, *Speech communication*, vol. 41, pp. 603-623, 2003.

[3] M. M. El Ayadi, M. S. Kamel, and F. Karray, Speech emotion recognition using Gaussian mixture vector autoregressive models, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 2007, pp. IV-957-IV-960.

[4] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, Autoencoder-based unsupervised domain adaptation for speech emotion recognition, *IEEE Signal Processing Letters*, vol. 21, pp. 1068-1072, 2014.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, *et al.*, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2016, pp. 5200-5204.

[6] G. Keren and B. Schuller, Convolutional RNN: an enhanced model for extracting features from sequential data, in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 3412-3419.

[7] J. Lee and I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition, in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] M. Neumann and N. T. Vu, Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech, *arXiv preprint arXiv:1706.00612*, 2017.

[9] X. Xia, J. Liu, W. Han, X. Zhu, H. Sahli, and D. Jiang, Speech emotion recognition based on global features and dcnn, in *International workshop on Affective Social Multimedia Computing (ASMMC), a satellite workshop of INTERSPEECH*, 2017.

[10] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, Adversarial auto-encoders for speech based emotion recognition, *arXiv preprint arXiv:1806.02146*, 2018.

[11] S. Mirsamadi, E. Barsoum, and C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227-2231.

[12] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, Domain adaptation for speech emotion recognition by sharing priors between related source and target classes, in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2016, pp. 2608-2612.

[13] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, Sparse autoencoder-based feature transfer learning for speech emotion recognition, in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 511-516.

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

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

[16] A. Hassan, R. Damper, and M. Niranjan, On acoustic emotion recognition: compensating for covariate shift, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1458-1468, 2013.

[17] S. J. Pan and Q. Yang, A survey on transfer learning, *IEEE Transactions on knowledge and data engineering*, vol. 22, pp. 1345-1359, 2009.

[18] T. Kanamori, S. Hido, and M. Sugiyama, Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection, in *Advances in neural information processing systems*, 2009, pp. 809-816.

[19] X. Glorot, A. Bordes, and Y. Bengio, Domain adaptation for large-scale sentiment classification: A deep learning approach, in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513-520.

[20] R. Gopalan, R. Li, and R. Chellappa, Unsupervised adaptation across domain shifts by generating intermediate data representations, *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, pp. 2288-2302, 2013.

[21] S. Yang, H. Wang, Y. Zhang, P. Li, Y. Zhu, and X. Hu, Semi-supervised representation learning via dual autoencoders for domain adaptation, *Knowledge-Based Systems*, vol. 190, p. 105161, 2020.

[22] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, Semisupervised autoencoders for speech emotion recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 31-43, 2017.

[23] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, Greedy layer-wise training of deep networks, in *Advances in neural information processing systems*, 2007, pp. 153-160.

[24] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, *IEEE transactions on multimedia*, vol. 16, pp. 2203-2213, 2014.

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

[26] J. Deng, Z. Zhang, and B. Schuller, Linked source and target domain subspace feature transfer learning--exemplified by speech emotion recognition, in *2014 22nd International*

Conference on Pattern Recognition, 2014, pp. 761-766.

[27] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in *Proceedings of the 25th international conference on Machine learning,* 2008, pp. 1096-1103.

[28] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace, in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP),* 2016, pp. 5800-5804.

[29] N. Le Roux and Y. Bengio, Representational power of restricted Boltzmann machines and deep belief networks, *Neural computation,* vol. 20, pp. 1631-1649, 2008.

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

[31] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, Emovo corpus: an italian emotional speech database, in *International Conference on Language Resources and Evaluation (LREC 2014)*, 2014, pp. 3501-3504.

[32] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, A database of German emotional speech, in *Ninth European Conference on Speech Communication and Technology,* 2005.

[33] P. Jackson and S. Haq, Surrey audio-visual expressed emotion (savee) database, *University of Surrey: Guildford, UK,* 2014.

[34] A. Batliner, S. Steidl, and E. Nöth, Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus, in *Proc. of a Satellite Workshop of LREC*, 2008, p. 28.

[35] N. Keshtiari, M. Kuhlmann, M. Eslami, and G. Klann-Delius, Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD), *Behavior research methods,* vol. 47, pp. 275-294, 2015.

[36] B. Schuller, S. Steidl, and A. Batliner, The interspeech 2009 emotion challenge, in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[37] F. Eyben, M. Wöllmer, and B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462.

[38] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249-256.

[39] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980,* 2014.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling

technique, *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[41] C.-C. Chang and C.-J. Lin, LIBSVM: A library for support vector machines, *ACM transactions on intelligent systems and technology (TIST),* vol. 2, p. 27, 2011.