DOMAIN ADAPTATION FOR SPEECH EMOTION RECOGNITION BY SHARING PRIORS BETWEEN RELATED SOURCE AND TARGET CLASSES

Qirong Mao, Wentao Xue, Qiru Rao, Feifei Zhang and Yongzhao Zhan

Department of Computer Science and Communication Engineering, Jiangsu University, China

ABSTRACT

In speech emotion recognition (SER), speech data is usually captured from different scenarios, which often leads to significant performance degradation due to the inherent mismatch between training and test set. To cope with this problem, we propose a domain adaptation method called Sharing Priors between Related Source and Target classes (SPRST) based on a two-layer neural network. The classifier parameters, namely the weights of the second layer, are imposed the common priors between the related classes, so that the classes with few labeled data in target domain can borrow knowledge from the related classes in source domain. The method is evaluated on the INTERSPEECH 2009 Emotion Challenge two-class task. Experimental results show that our approach significantly improves the performance when only a small number of target labeled instances are available.

Index Terms— Domain adaptation, speech emotion recognition, neural network, priors

1. INTRODUCTION

The problem of automatically predicting the emotional states in speech emotion recognition has been the subject of increasing attention among the speech community. Many state-of-the-art speech emotion recognition methods usually assume that the training and test samples are drawn from the same distribution. However, in real world applications, the speech signals obtained from different devices and recording conditions will be typically highly dissimilar in terms of speakers, spoken languages, type of emotion, acoustic signal conditions and type of labeling scheme [1]. A classifier just trained on a specific corpus and then applied directly to another corpus, cannot be expected to have excellent performance.

One outstanding approach to deal with this problem is domain adaptation (DA). DA is one special type of transfer learning problem, in which the source and target data distributions are different, but the source and target tasks remain the same [2, 3]. Based on whether the target domain data is partially labeled or completely unlabeled, DA techniques are commonly classified into two categories: semi-supervised DA and unsupervised DA. It has been theoretically shown that transfer learning can greatly improve the classification performance especially when there exist a small number of labeled samples in the target domain [3, 4]. Here, we mainly address the situation of semi-supervised DA.

For semi-supervised DA for SER, many approaches are proposed [5, 6]. However, priors are not considered. In this paper, we propose a Sharing Priors between Related Source and Target classes (SPRST) approach based on a two-layer neural network model. The major contribution of this paper is: To our best knowledge, this is the first paper introducing the priors to DA in SER. The classifier parameters can be derived from the priors. The same prior is imposed on the classifier parameters of the related source and target classes, so that the target classes with few samples can borrow knowledge from the source classes.

2. RELATED WORK

Research in SER has increasing drawn attention [7, 8, 9]. Most of these works are based on the condition that the training and test set come from the same corpus. This assumption doesn't hold in many real world applications and the performance will degrade due to the inherent dissimilarities between the training and test set.

Transfer learning has been proposed to transfer useful information from one source domain to a related target domain and can solve this problem effectively [4]. Meanwhile, deep neutral network has recently achieved state-of-the-art performance on a number of machine learning tasks [10]. The success of deep learning mainly contributes to the ability of layerwise unsupervised pre-training and extracting abstract hierarchical non-linear features of the input [11, 12, 13]. Deep neural networks have shown to suit well to the transfer learning [14]. Deng et al. [5] have presented a sparse autoencoder-based feature transfer learning method, in which a common emotion-specific mapping rule is learnt from a small set of target labeled data and then the source data reconstructed by this mapping are used to train a classifier. Schuller et al. [15] propose a shared-hidden-layer autoencoder (SHLA) approach to learn common feature representations shared across the training and test set. Also, Deng et al. [16] have introduced an

This work is supported by the National Natural Science Foundation of China (No. 61272211) and the general Financial Grant from the China Postdoctoral Science Foundation (No. 2015M570413).

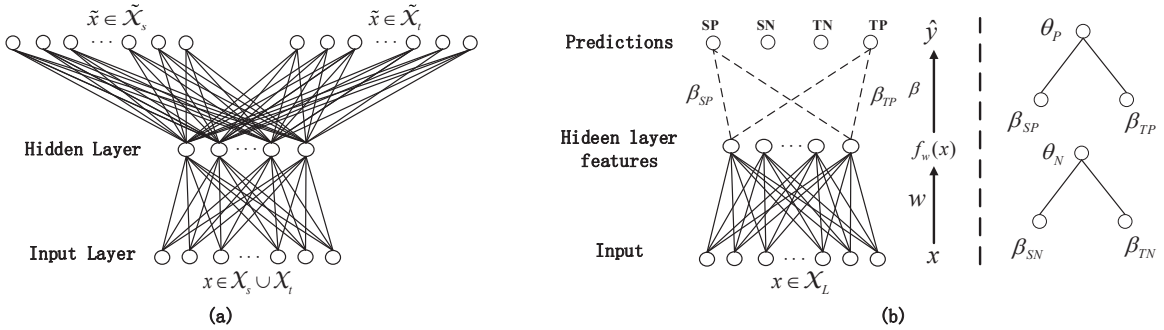


Fig. 1: Overview of SPRST model. (a) Schematic showing the unsupervised pre-training with SHLA method. (b) A two layer neural network (the parameters of first layer initialized with the weights learned by SHLA method) with priors sharing over the classification parameters of related classes.

adaptive denoising autoencoder method where prior knowledge learned from a target set is used to regularize the training on a source set. Srivastava et al. [17] have improved the classification performance for the classes with few training examples by discovering similar classes and transferring knowledge among them.

In [17], tree-based priors are introduced. This interesting work gives us new insight about DA in SER. In this paper, we impose priors on the classifier parameters of the related source and target classes. Different from [17], here the related classes are drawn from different domains. In our two-layer neural network model, we first pre-train the weights of the first layer combining the source and target unlabeled samples as input. This makes the distribution induced by the source samples as similar as possible as the distribution of target samples. Then we impose common priors on the classifier parameters of the related source and target classes, and use the labeled data to train the network.

3. PROPOSED METHODOLOGY

The structure of SPRST model is shown in Fig. 1. It involves two stages: 1) unsupervised pre-training using SHLA; 2) sharing priors between the related classes.

3.1. Unsupervised Pre-training

We employ the SHLA method for unsupervised pre-training. The process is shown in Fig. 1(a). Given the source domain samples $\mathcal{X}_s$, and the target domain samples $\mathcal{X}_t$, the two objective functions, similar to that of autoencoder, are defined as follows:

$$L_s(\theta_s) = \sum_{x \in \mathcal{X}_s} \|x - \tilde{x}\|^2, \tag{1}$$

$$L_t(\theta_t) = \sum_{x \in \mathcal{X}_t} \|x - \tilde{x}\|^2, \tag{2}$$

where $\tilde{x}$ is the reconstruction of x, and the parameters $\theta_s = \{W_1, b_1, {W_2}^s, {b_2}^s\}$, and $\theta_t = \{W_1, b_1, {W_2}^t, {b_2}^t\}$ share the same parameters $\{W_1, b_1\}$. The overall objective function is:

$$L_{SHLA}(\theta_{SHLA}) = L_s(\theta_s) + \gamma L_t(\theta_t), \tag{3}$$

where $\theta_{SHLA} = \{W_1, b_1, {W_2}^s, {b_2}^s, {W_2}^t, {b_2}^t\}$ are the parameters to be optimized during training, and the hyper-parameter γ weighs the contribution of two terms.

3.2. Sharing Priors

Assume $\mathcal{X}_L = \{x_1, x_2, ..., x_N\}$ are the labeled examples drawn from the source and target domains and $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ are the corresponding labels, where each label y_i is a K dimensional vector of targets. Our model is a two-layer neural network (see Fig. 1(b)). Let w denote the parameters (weights and biases) of the first layer, and w is initialized with (W_1, b_1) in section 3.1. Parameter $\beta \in R^{D \times K}$ denotes the second layer weights ($\beta_k \in R^D$ representing the classifier parameters for class k). Here D represents the number of the hidden units, and K is the number of classes.

Generally speaking, the classifier parameters for one class are independent of that of all other classes. It does work well for most applications when large plenty of labeled examples per class are available. However, in semi-supervised DA, there exist only a small number of labeled examples in target domain. For a two-class problem, we consider that the source positive (SP) class and the target positive (TP) class are related, the same as the source negative (SN) and target negative (TN). The related classes share a common prior over their classifier parameters. For example, SP and TP share a common prior, so that TP with few labeled data can borrow knowledge from SP.

Assume that each class k is associated with a weight vector $\beta_k \in R^D$, and β_k can be derived from a prior vector $\theta_s \in R^D$. We define the following generative model for β:

$$\theta_s \sim N(0, \frac{1}{\lambda_1} I_D), \quad \beta_k \sim N(\theta_{parent(k)}, \frac{1}{\lambda_2} I_D), \tag{4}$$

Table 1: Emotion categories mapping onto negative and positive valence for three databases.

Corpus	Negative	Positive
FAU AEC	angry, touchy, emphatic, reprimanding	motherese, neutral, joyful, rest
ABC	aggressive, intoxicated, nervous, tired	cheerful, neutral, rest
Emo-DB	anger, boredom, disgust, fear, sadness	joy, neutral

Table 2: Overview of the standardised feature set provided by the INTERSPEECH 2009 EC.

LLDs (16×2)	Functionals (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, mean square error

where λ_1 and λ_2 are hyper-parameters, I_D is identity matrix of size $D \times D$, and $\theta_{parent(k)}$ denotes the θ that β_k derives from. In this paper, we denote two priors θ_P and θ_N. For related classes, they share the same prior. So β_{SP} and β_{TP} are derived from θ_P, and β_{SN} and β_{TN} are derived from θ_N. This is shown is Fig. 1(b). More formally, we wish to minimize the following loss function:

$$\begin{aligned} L(w,\beta,\theta) &= -\log P(\mathcal{Y}|\mathcal{X},w,\beta) - \log P(w) - \log P(\theta) \\ &\quad - \log P(\beta|\theta) \\ &= -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} 1\{y_{(i)}=k\}\log\frac{e^{\beta_k^T f_w(x_{(i)})}}{\sum_{j=1}^{K} e^{\beta_j^T f_w(x_{(i)})}}\right] \\ &\quad + \frac{\lambda^2}{2}\|w\|^2 + \frac{\lambda_1}{2}\|\theta\|^2 \\ &\quad + \frac{\lambda_2}{2}\sum_{k=1}^{K}\left\|\beta_k - \theta_{parent(k)}\right\|^2, \end{aligned} \tag{5}$$

where $\log P(\mathcal{Y}|\mathcal{X},w,\beta)$ is the log-likelihood function and the other terms are priors over the model's parameters. $1\{\cdot\}$ is the indicator function, $1\{\text{a true statement}\} = 1$, and $1\{\text{a false statement}\} = 0$. $f_w(x_{(i)}) = s(wx_{(i)})$ represents the hidden features, and s is the sigmoid activation function. The choice of normal distributions in Eq.4 leads to a nice property that maximization over θ, given β can be done in closed form. It just amounts to taking a scaled averaged of all β_k's which are derived from θ_s. Let $N_s = \{k|parent(k) = s\}$, then

$$\theta_s{}^* = \frac{1}{|N_s| + \lambda_1/\lambda_2}\sum_{k\in N_s}\beta_k. \tag{6}$$

Therefore, the loss function in Eq.5 can be optimized by iteratively performing the following two steps. First, we maximize over w and β keeping θ fixed by using standard stochastic gradient descent (SGD). Then, we maximize over θ keeping β fixed using Eq.6.

4. DATABASE

To evaluate the effectiveness of our method, we consider the INTERSPEECH 2009 Emotion Challenge (EC) two-class task [18]. It is based on the FAU Aibo Emotion Corpus (FAU AEC), which is a spontaneous corpus that contains 9 hours of German speech of 51 children interacting with Sony's pet robot Aibo at two different schools, Ohm and Mont. We treat FAU AEC as target domain database. The data of school Ohm is used for target training, the data of Mont for target testing.

Additionally, for the source set we choose two publicly available databases, namely the database of German emotional speech (Emo-DB) [19], and the Airplane Behavior Corpus (ABC) [20]. They are highly different from the target set FAU AEC in terms of age, type of emotion and recording situation, and degree of spontaneity. For comparability with FAU AEC, we have to map the diverse emotion classes onto the valence axis of the dimensional emotion model. The mapping strategy is shown in Table 1 according to [1, 15].

4.1. Acoustic Features

We keep in line with the INTERSPEECH 2009 EC [18] and use a baseline feature set which consists of 12 functionals applied to 2×16 acoustic Low-Level Descriptors (LLDs) including their first order delta regression coefficients as shown in Table 2. Therefore, the feature vector per chunk contains $16 \times 2 \times 12 = 384$ attributes. To ensure reproducibility, the open source toolkit openEAR [21] is utilized to extract 384 attributes.

5. EXPERIMENTS

5.1. Experimental Setup

For the first stage of our model, we choose the source set (ABC or Emo-DB), and the target training set Ohm to perform unsupervised pre-training. For the second stage, a small part of labeled examples (the size ranging from 10 to 1000 chunks) are randomly chosen from the target training set Ohm, where the same number of instances are chosen from positive valence and negative valence. These selected target labeled instances, together with the source labeled instances, are used to train the network. Then, the source set and parts of the labeled examples in Ohm set (10 instances) are selected as a validation set to select parameters. Finally, the target test set Mont is fed into the network for classification.

For parameters selection, the number of hidden units is fixed to 200, and attempted hyper-parameters γ, λ_1 and λ_2 are the following: $\gamma \in \{0.1, 0.5, 1, 2, 3\}$, $\lambda_1, \lambda_2 \in \{0.1, 0.5, 1, 3, 5\}$. For performance evaluation, we choose unweighted average recall (UAR), namely the mean accuracy over the accuracy of each class. The reported performance in UAR is the average over 20 runs to avoid 'lucky' or 'unlucky' selection.

5.2. Methods for Comparison

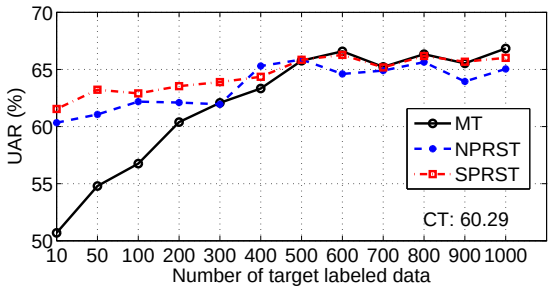
To evaluate the effectiveness of approach, we compare the following methods with the same initialized parameters for the first layer:

- **Matched Training (MT)**: randomly picks a number of instances from the target training set Ohm to train the network , without using the source set.
- **Cross Training (CT)**: only uses the source set ABC or Emo-DB to train the network.
- **NPRST**: uses the source set and a number of instances from target training set Ohm to train the network , without sharing priors.
- **SPRST**: uses the source set and a number of instances from target training set Ohm to train the network , with sharing priors.

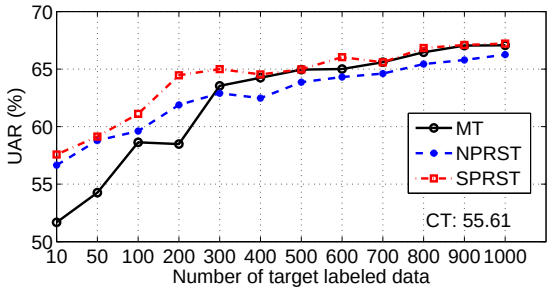
5.3. Results

Fig. 2(a) and (b) report the results of the source set being ABC and Emo-DB respectively. Our approach achieves higher performance when only a small number of target labeled instances are available. Specifically, for the ABC, the SPRST achieves the highest UAR when the number of chosen target labeled instances is in the range of 10 to 300. Afterwards, when the size of target labeled instances continues increasing, the performance of MT gradually overtakes the SPRST since no more extra information in the ABC can be transferred to the target domain. For the source set being Emo-DB, the SPRST performs better within the range of target size ranging from 10 to 300. Afterwards, the performance of SPRST is still comparable with the MT.

Table 3 shows the UAR comparisons for each method when only a small number of labeled instances are available in the target domain, e.g. only 10 instances. As we can see, MT can only achieve a chance level UAR. The SPRST method outperforms the NPRST, which means that the priors shared between the related source and target classes do work for the classification in the speech emotion recognition.



(a) ABC



(b) Emo-DB

Fig. 2: UAR comparison for the increase of number of instances chosen from the FAU training set Ohm for the source set being ABC and Emo-DB. CT:# is the UAR if only using source set.

Table 3: UAR comparison when only 10 labeled instances are chosen from the target training set.

UAR (%)	MT	NPRST	SPRST
ABC	50.69	60.33	**61.54**
Emo_DB	51.68	56.65	**57.58**

6. CONCLUSION

In this paper, we proposed a Sharing Priors between Related Source and Target classes (SPRST) approach based on a two-layer neural network model. We first pre-train the weights of the first layer. Then we impose the common priors on the classifier parameters of the related source and target classes so that the target classes with few labeled data can borrow knowledge from the source classes. Experimental results with two publicly available corpus show that the proposed method can effectively transfer knowledge and enhance the classification performance.

Further work includes extending the single-architecture to a deep architecture in order to further find the useful information in emotional features.

7. REFERENCES

[1] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 119–131, 2010.

[2] H. Daumé III and D. Marcu, "Domain adaptation for statistical classifiers," *Journal of Artificial Intelligence Research (JAIR)*, vol. 26, pp. 101–126, 2006.

[3] S. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[4] L. Torrey and J. Shavlik, "Transfer learning," *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, vol. 1, pp. 242, 2009.

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

[6] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 5058–5063.

[7] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in *Emotion-Oriented Systems*, pp. 71–99. 2011.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[9] D. Yu, M. L. Seltzer, J. Li, J. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[10] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[12] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, pp. 153, 2007.

[13] A. Coate, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," *International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

[14] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *Unsupervised and Transfer Learning Challenges in Machine Learning*, vol. 7, pp. 19, 2012.

[15] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4818–4822.

[16] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *Signal Processing Letters, IEEE*, vol. 21, no. 9, pp. 1068–1072, 2014.

[17] N. Srivastava and R. Salakhutdinov, "Discriminative transfer learning with tree-based priors," *Advances in Neural Information Processing Systems*, 2013.

[18] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *INTERSPEECH*, 2009, vol. 2009, pp. 312–315.

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[20] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Acoustics, Speech and Signal Processing (ICASSP), 2007 IEEE International Conference on*, 2007, vol. 2, pp. II–733.

[21] F. Eyben, M. Wollmer, and B. Schuller, "Openear-introducing the munich open-source emotion and affect recognition toolkit," *Affective Computing and Intelligent Interaction and Workshops (ACII). 3rd International Conference on*, pp. 1–6, 2009.