INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Cross-corpus speech emotion recognition using semi-supervised transfer non-negative matrix factorization with adaptation regularization

Hui Luo, Jiqing Han

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, PRC
luohui0216@163.com, jqhan@hit.edu.cn

Abstract

This paper focuses on a cross-corpus speech emotion recognition (SER) task, in which there are some mismatches between the training corpus and the testing corpus. Meanwhile, the label information of the training corpus is known, while the label information of the testing corpus is entirely unknown. To alleviate the influence of these mismatches on the recognition system under this setting, we present a non-negative matrix factorization (NMF) based cross-corpus speech emotion recognition method, called semi-supervised adaptation regularized transfer NMF (SATNMF). The core idea of SATNMF is to incorporate the label information of training corpus into NMF, and seek a latent low-rank feature space, in which the marginal and conditional distribution differences between the two corpora can be minimized simultaneously. Specifically, in this induced feature space, the maximum mean discrepancy (MMD) criterion is used to measure the discrepancies of not only two corpora, but also each class within the two corpora. Moreover, to further exploit the knowledge of the marginal distributions, their underlying manifold structure is considered by using the manifold regularization. Experiments on four popular emotional corpora show that the proposed method achieves better recognition accuracies than state-of-the-art methods.

Index Terms: cross-corpus, speech emotion recognition, semi-supervised transfer NMF, adaptation regularization

1. Introduction

Since humans can recognize emotion across various vocal sources, the similar emotion recognition systems will have broad application. A typical cross-corpus speech emotion recognition (SER) problem is that the training (or source) speech corpus would be very different from the testing (or target) corpus, e.g., they are from two different languages. The conventional SER methods would not be well applicable for such a cross-corpus SER task because they consider that the source and target feature points come from a same corpus or distribution. Consequently, it is an important and very challenging problem that effectively handles the cross-corpus SER in current SER study. Nevertheless, there are still several methods that had been developed to investigate this problem [1, 2].

In [1], various normalization schemes were addressed for the cross-corpus SER. From then on, a variety of interesting methods were introduced to deal with this challenging problem. In [2], three transfer learning methods, namely kernel mean matching (KMM) [3], Kullback-Leibler importance estimation procedure (KLIEP) [4], and unconstrained least-squares importance fitting [5], were incorporated into support vector machine (SVM) for cross-corpus SER tasks. In the works of [6, 7, 8], a series of auto-encoder based domain adaptation methods were presented by leveraging various auto-encoder based networks to learn a common representation across the source and target samples. Recently, a transfer non-negative matrix factorization (TNMF) method, in which the maximum mean discrepancy (MMD) [9] was introduced to eliminate the feature distribution difference between source and target speech databases, was developed to cope with cross-corpus SER. More recently, a domain-adaptive subspace learning (DoSL) model was applied [10] to cross-corpus SER, which aims at learning a regression coefficient matrix to bridge the source and target speech corpora. In addition, various deep learning architectures were provided for cross-corpus SER [11, 12].

In this paper, we focus on a specific cross-corpus SER task, in which the source speech signals are labeled, while the label information of the target speech signals is entirely unknown. Inspired by [13, 14], a novel semi-supervised adaptation regularized transfer non-negative matrix factorization (SATNMF) method is presented for this task. In our method, the label information of the source speech is incorporated into the NMF via the semi-supervised NMF algorithm [13] to learn the discriminative representations. Meanwhile, the distance of the joint probability distributions between the source and target data are considered on the learned representations. The joint distribution distance, which involves the distance between not only the two different distributions, but also each class within the two distributions, is measured by the maximum mean discrepancy (MMD) approach [9]. Moreover, the manifold structure of the two distributions is also considered for further exploiting their underlying knowledge. In the end, a classifier can be trained on the representations of the labeled source speech signals, and applied to predict the emotional states of the target speech signals. Experimental results show the effectiveness of our proposed method.

2. The proposed framework

The overall diagram of our proposed cross-corpus SER framework is illustrated in Figure 1.

2.1. Speech feature extraction and normalization

The open-source openSMILE tool [15] is popularly used to extract acoustic features in a number of paralinguistic challenges. The idea is to obtain a large pool of potentially relevant features by passing an extensive set of summarizing functionals on the low level descriptor (LLD) contours. In this paper, we use the toolbox with a standard configuration that used in computational paralinguistic challenge of INTERSPEECH 2013 [16].

The feature set contains as large as 6373 supra-segmental features, where 65 low-level descriptors (e.g. pitch, MFCC and loudness) as well as their first order derivatives are summarized by 54 statistical functionals. Note that some functionals are not applied to all LLDs. In addition, five global temporal statistics are considered and contained in this set. The detailed configuration of this feature set can be referred to [16]. After feature ex-

http://dx.doi.org/10.21437/Interspeech.2019-2041

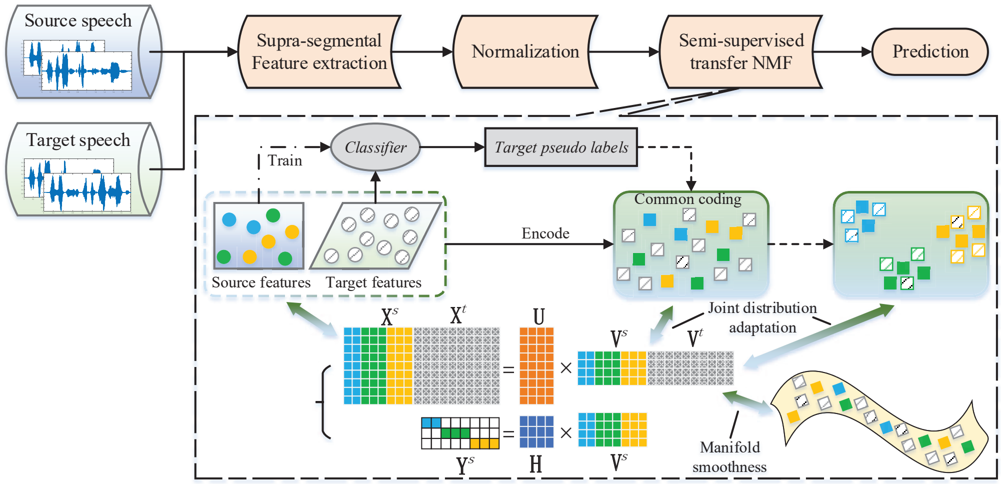

Figure 1: *Illustration of our cross-corpus speech emotion recognition model*

traction, we employ speaker-dependent (SD) strategy [17] and normalize the range of each feature to the interval $[0, 1]$ by linear scaling. In this paper, we normalize the source samples and the target samples independently.

2.2. Semi-supervised transfer NMF

Suppose that we have a feature matrix $\mathbf{X}^s = [\mathbf{x}_1^s, \ldots, \mathbf{x}_n^s] \in \mathbb{R}^{d\times n}$ extracted from a source corpus $\mathcal{D}^s$, and a feature matrix $\mathbf{X}^t = [\mathbf{x}_1^t, \ldots, \mathbf{x}_m^t] \in \mathbb{R}^{d\times m}$ from a target corpus $\mathcal{D}^t$, where d is the dimension of the speech emotion feature vectors, n is the number of source samples, and m is the number of target samples. We denote the label matrix of $\mathcal{D}^s$ by $\mathbf{Y}^s = [\mathbf{y}_1^s, \ldots, \mathbf{y}_n^s] \in \mathbb{R}^{c\times n}$, where c is the number of classes. The entry y_{ij}^s is 1 if $\mathbf{x}_j^s$ belongs to class i, and 0 otherwise.

2.2.1. *Semi-supervised non-negative matrix factorization*

The NMF is an efficient algorithm that can obtain a low dimensional representation of the non-negative data [18], which has been successfully applied to various pattern recognition fields, e.g., face recognition and document clustering. It aims at finding two non-negative matrices to well approximate the original matrix data. In [13], a semi-supervised NMF (SNMF) was presented to incorporate the label information into NMF. Given a non-negative feature matrix $\mathbf{X} = [\mathbf{X}^s, \mathbf{X}^t] \in \mathbb{R}^{d\times(n+m)}$ and the corresponding label matrix $\mathbf{Y} = [\mathbf{Y}^s, \mathbf{Y}^t] \in \mathbb{R}^{c\times(n+m)}$, the following problem is then needed to be solved:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{H},\mathbf{V}} \quad & \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \beta\|\mathbf{E} \circ (\mathbf{Y} - \mathbf{H}\mathbf{V})\|_F^2, \\ s.t. \quad & \mathbf{U} \geq 0,\ \mathbf{H} \geq 0,\ \mathbf{V} \geq 0, \end{aligned} \tag{1}$$

where $\mathbf{U} = [u_{ik}] \in \mathbb{R}^{d\times r}$ and $\mathbf{H} = [h_{lk}] \in \mathbb{R}^{c\times r}$ are basis matrices for $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\mathbf{V} = [v_{kj}] \in \mathbb{R}^{r\times(n+m)}$ is the coding matrix. $\beta > 0$ is a tradeoff parameter determining the importance of the supervised term. $\|\cdot\|_F^2$ is a Frobenius norm, and $\circ$ is the Hadamard product. It does no matter that the target labels $\mathbf{Y}^t$ are unknown, since they are filtered out by the label indicator matrix $\mathbf{E} = [e_{ij}] \in \mathbb{R}^{c\times(n+m)}$, which can be defined as:

$$e_{ij} = \begin{cases} 0.001, & if\ y_{ij} = 1, \\ 1, & if\ y_{ij} = 0, \\ 0, & if\ y_{ij}\ is\ unknown. \end{cases} \tag{2}$$

The SNMF function (1) is a non-convex problem when calculating $\mathbf{U}$, $\mathbf{H}$ and $\mathbf{V}$ together, so an iterative algorithm [13] has been presented by updating $\mathbf{U}$, $\mathbf{H}$ and $\mathbf{V}$ as follows:

$$\begin{cases} u_{ik} &= u_{ik}\dfrac{(\mathbf{X}\mathbf{V}^T)_{ik}}{(\mathbf{U}\mathbf{V}\mathbf{V}^T)_{ik}}, \\ h_{lk} &= h_{lk}\dfrac{((\mathbf{E}\circ\mathbf{Y})\mathbf{V}^T)_{ik}}{((\mathbf{E}\circ\mathbf{H}\mathbf{V})\mathbf{V}^T)_{ik}}, \\ v_{kj} &= v_{kj}\dfrac{(\mathbf{U}^T\mathbf{X} + \beta\mathbf{H}^T(\mathbf{E}\circ\mathbf{Y}))_{kj}}{(\mathbf{U}^T\mathbf{U}\mathbf{V} + \beta\mathbf{H}^T(\mathbf{E}\circ\mathbf{H}\mathbf{V}))_{kj}}. \end{cases} \tag{3}$$

2.2.2. *Joint distribution adaptation*

By using the SNMF algorithm, a common coding vectors $\mathbf{V}$ shared by the data matrix and label matrix can be obtained for both labeled source and unlabeled target corpus. However, for the cross-corpus task, on one hand, the difference of coding vectors between the source and target data are still large, which will have an adverse impact on the recognition performance. On the other hand, the basis matrix $\mathbf{H}$ may only connect the source data with the label space, while not generalize well to the target data since it requires the labeled and unlabeled data to be sampled from identical probability distribution. Thus the major computational issue is how to reduce the distribution difference. Following [14], we employ the maximum mean discrepancy (MMD) criterion [19] as the distance measure, which compares different distributions based on the distance between the sample means of two corpora in the coding space, namely:

$$D(\mathcal{D}^s, \mathcal{D}^t) = \|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i - \frac{1}{m}\sum_{j=n+1}^{n+m}\mathbf{v}_j\|_2^2, \tag{4}$$

where $\mathbf{v}_i$ is the i-th column of $\mathbf{V}$.

However, it is not enough to only reduce the difference between the marginal distributions, we also need to reduce the differences between the conditional distributions. Unfortunately, it is hard to measure the conditional distribution if we have no labeled data from the target corpus. To address this problem, we manage to utilize the pseudo labels of the target data by adopting some basic classifiers (e.g., SVM) trained on the source data [20]. Although there may be many incorrect labels in the pseudo labels, we look forward to uncovering the underlying structure of the source and target data by transferring the local information. Therefore, we can use both true and pseudo labels to compute the MMD w.r.t. each class $k \in \{1, \ldots, c\}$ and make the intra-class centroids of two distributions closer in the coding space:

$$D_k(\mathcal{D}^s, \mathcal{D}^t) = \| \frac{1}{n_k} \sum_{\mathbf{v}_i \in \mathcal{D}_k^s} \mathbf{v}_i - \frac{1}{m_k} \sum_{\mathbf{v}_j \in \mathcal{D}_k^t} \mathbf{v}_j \|_2^2, \quad (5)$$

where $\mathcal{D}_k^s$ is the set of samples belonging to class k in the source data, and $n_k = |\mathcal{D}_k^s|$. Correspondingly, $\mathcal{D}_k^t$ is the set of samples belonging to class k (pseudo) in the target data, and $m_k = |\mathcal{D}_k^t|$.

Integrating (4) and (5) leads to the regularization for joint distribution adaptation, that is

$$D(\mathcal{D}^s, \mathcal{D}^t) + \alpha \sum_{k=1}^{c} D_k(\mathcal{D}^s, \mathcal{D}^t) = Tr(\mathbf{V}\mathbf{M}\mathbf{V}^T), \quad (6)$$

where $Tr(\cdot)$ is the trace operator of matrix, $\alpha > 0$ is a balance parameter, and $\mathbf{M} = \mathbf{M}_0 + \alpha \sum_{k=1}^{c} \mathbf{M}_c$ is the MMD matrix, in which each $\mathbf{M}_k$ $(k = 0, \ldots, c)$ is calculated as

$$(\mathbf{M}_k)_{ij} = \begin{cases} \frac{1}{n_k^2} & , \mathbf{v}_i, \mathbf{v}_j \in \mathcal{D}_k^s \\ \frac{1}{m_k^2} & , \mathbf{v}_i, \mathbf{v}_j \in \mathcal{D}_k^t \\ \frac{-1}{n_k m_k} & , \mathbf{v}_i \in \mathcal{D}_k^s, \mathbf{v}_j \in \mathcal{D}_k^t \ or \ \mathbf{v}_i \in \mathcal{D}_k^t, \mathbf{v}_j \in \mathcal{D}_k^s \\ 0 & , otherwise. \end{cases}$$

Here we denote $n_0 = n$, $m_0 = m$, $\mathcal{D}_0^s = \mathcal{D}^s$ and $\mathcal{D}_0^t = \mathcal{D}^t$ for simplicity.

2.2.3. Manifold regularization

In addition, many previous studies have demonstrated that the naturally occurring data may usually reside on or close to a low dimensional submanifold embedded in a high dimensional space [21], and the intrinsic geometrical information is important to the discrimination of the data [22]. Inspired by this, a manifold regularization can be computed on the coding vectors as following

$$\frac{1}{2} \sum_{i,j=1}^{n+m} \|\mathbf{v}_i - \mathbf{v}_j\|_2^2 w_{ij} = Tr(\mathbf{V}\mathbf{L}\mathbf{V}^T), \quad (7)$$

where $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{(n+m)\times(n+m)}$ is the graph affinity matrix, and $\mathbf{L}$ is the graph Laplacian matrix. $\mathbf{W}$ is defined as:

$$\mathbf{w}_{ij} = \begin{cases} e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}} & , \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \ or \ \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0 & , otherwise, \end{cases}$$

where $\mathcal{N}_k(\mathbf{x}_i)$ is the set of k-nearest neighbors of $\mathbf{x}_i$, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with each item $d_{ii} = \sum_{j=1}^{(n+m)} w_{ij}$.

Finally, by regularizing (1) with (6) and (7), our semi-supervised adaptation regularized transfer NMF based cross-corpus speech emotion recognition is formulated as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{H},\mathbf{V}} \quad & \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \beta\|\mathbf{E} \circ (\mathbf{Y} - \mathbf{H}\mathbf{V})\|_F^2 \\ & + \lambda Tr(\mathbf{V}\mathbf{L}\mathbf{V}^T) + \gamma Tr(\mathbf{V}\mathbf{M}\mathbf{V}^T), \\ s.t. \quad & \mathbf{U} \geq 0, \ \mathbf{H} \geq 0, \ \mathbf{V} \geq 0, \end{aligned} \quad (8)$$

where λ and γ are two positive regularization parameters.

2.3. Cross-corpus speech emotion recognition

We first obtain pseudo labels for the target samples via the classifier trained on the source samples using the original features. Then suppose that $(\mathbf{U}^*, \mathbf{H}^*, \mathbf{V}^*)$ is the optimal solution of (8), a standard supervised classifier (e.g. SVM) is trained on $\mathbf{V}^{s*}$, and the emotion labels of the target speech feature vectors are assigned by feeding $\mathbf{V}^{t*}$ into the trained classifier.

It should be noted that the instability may be introduced by the random initialization in applying SATNMF to the cross-corpus SER. To solve this problem, we propose a simple and efficient voting scheme for better predicting the emotion labels of the target samples. In this method, we conduct several trials of experiments independently and predict the labels of the target samples. Then, for each testing sample from the target corpus, its emotion label is assigned to the label that occurs most among the several trials of predictions.

3. Experiments

Several experiments are conducted to evaluate our proposed SATNMF approach for cross-corpus SER.

Table 1: *Statistics of the datasets. Number of samples (# Sample). Number of emotion categories (# Emotion).*

Datasets	Berlin	CASIA	eNTERFACE	Estonian
Language	German	Chinese	English	Estonian
# Sample	535	1200	1257	1164
# Emotion	7	6	6	4

3.1. Data preparation

We conduct experiments on four typical discrete speech emotion datasets, including Berlin [23], CASIA [24], eNTERFACE [25] and Estonian [26]. The statistics of these data sets are summarized in Table 1.

Table 2: *Schemes of the cross-corpus SER.*

Schemes	Source	Target	Emotions
e2E	eNTERFACE	Estonian	A, H, Sa
E2B	Estonian	Berlin	A, H, N, Sa
B2C	Berlin	CASIA	A, F, H, N, Sa
C2e	CASIA	eNTERFACE	A, F, H, Sa, Su

We select two datasets each time, which are served as source and target corpus, alternatively, and pick out the samples belonging to the common emotion states. Therefore, we design four types of cross-corpus SER schemes, i.e., e2E, E2B, B2C and C2e. The details are listed in Table 2, where the emotional categories Anger (A), Fear (F), Happiness (H), Neutral (N), Sadness (Sa) and Surprise (Su) are used for evaluation.

3.2. Experimental setup

For comparison purpose, we use the linear SVM as the baseline method and choose five state-of-the-art cross-corpus SER methods, including Marginalized Denoising Auto-encoder (mSDA) [27], Transfer NMF (TNMF) [14], Feature Selection based Transfer Subspace Learning (FSTSL) [28], Domain-adaptive Subspace Learning (DoSL) [10] and Deep Belief Networks (DBN) [29], to conduct the same experiments as our SATNMF.

For the baseline, a linear SVM model is trained on the source data, and directly tested on the unlabeled target data. mSDA, TNMF, FSTSL and SATNMF are run on all data to learn the new feature representations of source and target data,

Table 3: *Results (%) of the four cross-corpus speech emotion recognition experiments in terms of UAR and WAR (bold numbers indicate the best UAR and WAR).*

Schemes	SVM		mSDA+SVM		TNMF+SVM		FSTSL+SVM		DBN		DoSL		SATNMF+SVM	
	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
e2E	35.65	37.74	36.90	39.31	40.50	41.49	40.12	42.10	35.52	37.14	39.55	37.98	**43.06**	**43.22**
E2B	38.16	37.35	37.25	33.22	54.51	57.11	47.73	**58.59**	44.48	45.61	46.98	37.64	**56.79**	53.84
B2C	32.90	32.90	34.20	34.20	38.88	38.88	26.80	26.80	32.60	32.60	40.50	40.50	**42.30**	**42.30**
C2e	28.85	28.89	28.85	28.89	**32.98**	33.07	29.86	29.94	31.48	31.57	30.24	30.33	32.79	**34.00**
Average	33.89	34.22	34.30	33.91	41.72	42.64	36.13	39.36	36.02	36.73	39.32	36.61	**43.74**	**43.34**

and then a linear SVM is trained and tested on these new representations. DoSL and DBN are trained on all data in a transductive way to directly induce cross-corpus classifiers.

Under our experimental setup, it is impossible to automatically tune the optimal parameters for the target classifier using cross validation, since we have no labeled data in the target corpora. Therefore, we evaluate the six baseline methods on our datasets by empirically searching the parameter space for the optimal parameter settings, and report the best results of each method. For the linear SVM, we set the trade-off parameter C by searching $C \in \{0.1, 0.5, 1, 5, 10, 50, 100\}$. For TNMF, FSTSL and SATNMF, the number of nearest neighbors p is set by searching $p \in \{1, 3, 5, 10, 15\}$. The size of dictionary r in TNMF and SATNMF is chosen by searching $r \in \{16, 32, 64, 128, 256, 512\}$. We set all regularization parameters (if any) by searching $\{0.01, 0.1, 0.5, 1, 5, 10, 50, 100, 1000\}$. For DBN, a structure with three RBM layers is selected, where the number of hidden units of each layer is set by searching $\{500, 1000, 2000\}$, and the number of epoch is set by searching $\{100, 200, 300, 500\}$. The learning rate is fixed 10^{-3}. The other network parameters are chosen by following the setup in [29].

We adopt the weighted average recall (WAR) and the unweighted average recall (UAR) to report the emotion recognition accuracy. The WAR means the total number of correctly predicted testing samples of all classes averaged by the total number of testing samples, while the UAR is defined as the accuracy per class averaged by total number of classes. It is better to use both WAR and UAR, rather than only single one, to show the overall performance.

3.3. Experimental results

3.3.1. Comparison with SNMF-based methods

The SNMF is a special case of our method with $\alpha = \gamma = \lambda = 0$. Besides, we consider SNMF-γ and STNMF-λ as two other special cases of our SATNMF by setting $\alpha = \lambda = 0$ and $\alpha = 0$, respectively. We plot the UAR and WAR of the four SNMF-based methods in Figure 2. From the results, it is clear that our method obtains the best overall performance among all the four methods, which implies the reduction of the differences between joint distributions and introduction of the manifold information are benefit to improving the cross-corpus SER accuracy.

3.3.2. Comparison with state-of-the-art methods

The experimental results of different approaches are depicted in Table 3. From the results, we have the following observations: Firstly, our SATNMF approach can achieve much better overall performance than the other methods. The corresponding relative improvements are 4.84% (UAR) and 1.64% (WAR) compared to the second best method. Secondly, the five state-of-the-art methods and SATNMF outperform the baseline on the average accuracy. The reason is that they employ various techniques to reduce the difference between source and target corpus. Thirdly, our method performs better than the unsupervised TNMF. This is due to that our method takes the label information and the joint distribution difference into consideration. Lastly, the our SATNMF is superior to the supervised FSTSL, DBN and DoSL. A major limitation of these methods is that they are prone to overfitting, since the lack of ability to simultaneously reduce the differences in both marginal and conditional distributions between corpora.

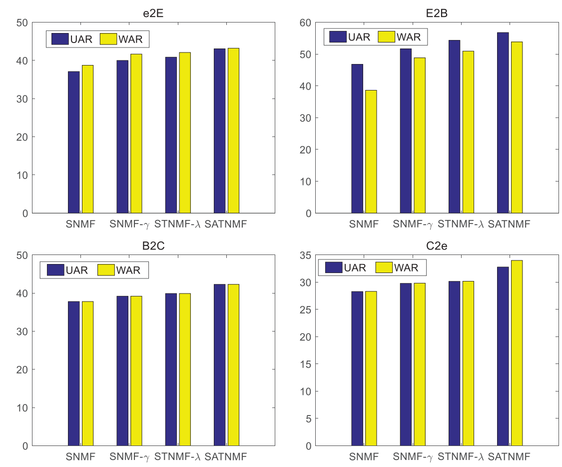

Figure 2: *Recognition performance (%) of SNMF, SNMF-γ, STNMF-λ and SATNMF on the four schemes.*

4. Conclusion

In this paper, a new method called semi-supervised adaptation regularized transfer non-negative matrix factorization (SATNMF) is proposed for cross-corpus speech emotion recognition (SER). SATNMF jointly factorizes the data matrix and label matrix, and simultaneously takes into account the joint distribution adaptation of both the marginal and conditional distributions, and the manifold structure. It can obtain more discriminative feature representations, and significantly improve the cross-corpus SER. Several experiments are carried out on four popular public emotional datasets, and the experimental evidence demonstrates the advantages of our method.

5. Acknowledgements

This work is supported by National Science Foundation of China under grant No. U1736210 and National Key Research and Development Program of China under grant No. 2017YFB1002102.

6. 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," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[2] 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, no. 7, pp. 1458–1468, 2013.

[3] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Advances in neural information processing systems*, 2007, pp. 601–608.

[4] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in neural information processing systems*, 2008, pp. 1433–1440.

[5] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1391–1445, 2009.

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

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

[8] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.

[9] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[10] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5144–5148.

[11] Z. Huang, W. Xue, Q. Mao, and Y. Zhan, "Unsupervised domain adaptation for speech emotion recognition using pcanet," *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 6785–6799, 2017.

[12] H. Kaya, D. Fedotov, A. Yeşilkanat, O. Verkholyak, Y. Zhang, and A. Karpov, "Lstm based cross-corpus and cross-task acoustic emotion recognition," in *Proceedings of INTERSPEECH*, 2018, pp. 521–525.

[13] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4–7, 2010.

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

[15] 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.

[16] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH*, 2013.

[17] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[18] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Advances in neural information processing systems*, 2007, pp. 513–520.

[20] M. Long, J. Wang, G. Ding, S. J. Pan, and S. Y. Philip, "Adaptation regularization: A general framework for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2014.

[21] J. Liu, D. Cai, and X. He, "Gaussian mixture model with local consistency," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, 2010, pp. 512–517.

[22] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

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

[24] ChineseLDC, "CASIA-Chinese emotional speech corpus," http://www.chineseldc.org/, 2005.

[25] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops*, 2006, pp. 8–8.

[26] R. Altrov and H. Pajupuu, "Estonian emotional speech corpus: theoretical base and implementation," in *International Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 2012.

[27] M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 1627–1634.

[28] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2018.

[29] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Proceedings of INTERSPEECH*, 2018, pp. 257–261.