LETTER

Speech Emotion Recognition Based on Sparse Transfer Learning Method

Peng SONG[†a)], **Wenming ZHENG**[††], *Nonmembers*, *and* **Ruiyu LIANG**[†††], *Member*

SUMMARY In traditional speech emotion recognition systems, when the training and testing utterances are obtained from different corpora, the recognition rates will decrease dramatically. To tackle this problem, in this letter, inspired from the recent developments of sparse coding and transfer learning, a novel sparse transfer learning method is presented for speech emotion recognition. Firstly, a sparse coding algorithm is employed to learn a robust sparse representation of emotional features. Then, a novel sparse transfer learning approach is presented, where the distance between the feature distributions of source and target datasets is considered and used to regularize the objective function of sparse coding. The experimental results demonstrate that, compared with the automatic recognition approach, the proposed method achieves promising improvements on recognition rates and significantly outperforms the classic dimension reduction based transfer learning approach.

key words: *speech emotion recognition, sparse coding, transfer learning*

1. Introduction

As an important branch of affective computing, the recognition of human's emotions from speech has been a hot research topic in speech signal processing field [1]. Speech emotion recognition refers to automatically recognizing emotions from speech [2]. It can be applied in various areas, e.g., diagnosing patient's negative emotions for better treatment in the healthcare field, monitoring driver's mental states to avoid accidents in intelligent vehicle systems, managing customer's disputes in emotionally intelligent systems of call centers.

Over the past few decades, various methods have been presented for speech emotion classification [2], e.g., hidden Markov model (HMM), Gaussian mixture model (GMM), artificial neural network (ANN) and support vector machine (SVM). Recently, deep learning techniques, popular in speech recognition and image classification, have also been applied to speech emotion recognition [3]. All these methods can obtain satisfactory results to some extent, however, they are conducted on single corpus. In reality, the training data and testing data are often from different datasets, and the recognition rates will drop significantly.

To tackle the cross-corpus speech emotion recognition problem, several studies have been done in recent years. Schuller et al. investigate different kinds of normalization methods on six standard datasets for cross-corpus evaluation experiments [4]. In [5], we present a transfer learning algorithm to perform cross-corpus speech emotion recognition, which learns a common latent feature space for source and target datasets. Although the transfer learning based speech emotion recognition method can enhance the cross-corpus speech emotion recognition rates to some extent, there still exists too much room for improvement, e.g., current domain adaptation or transfer learning based speech emotion recognition methods perform adaptation and feature selection independently, which will degrade the performance of the speech emotion recognition system. To address these problems, in this letter, a novel speech emotion recognition method using sparse transfer learning is presented. The sparse coding algorithm is first employed to learn the common feature representations for source and target corpora. Then, to ensure the robustness of the learned sparse features, the distance between the feature distributions of the two datasets is considered and used as a regularization for the objective function of sparse coding.

The remainder of this letter is structured as follows. In Sect. 2, the details of our proposed sparse transfer learning method, including the motivations and implementations, are introduced. The results are given and discussed in Sect. 3. Finally, the conclusions are drawn in Sect. 4.

2. Sparse Transfer Learning Based Speech Emotion Recognition

2.1 Sparse Feature Representation

Sparse coding has received a growing interest in recent years. It has been studied in many applications, e.g., face recognition, image classification, image restoration and signal classification [6]–[9]. Inspired by the successful applications in these fields, in this letter, the sparse coding algorithm is also introduced for speech emotion recognition.

Given the feature sequence of the emotional speech $X = [x_1, x_2, \dots, x_n] \in R^{m\times n}$, the dictionary matrix $D = [d_1, d_2, \dots, d_k] \in R^{m\times k}$, and the sparse coefficient matrix $S = [s_1, s_2, \dots, s_n] \in R^{k\times n}$, each feature vector x_i can be represented as a linear combination of the sparse coefficient vectors via the dictionary matrix, and the objective function

Manuscript received January 24, 2015.
Manuscript revised March 12, 2015.
Manuscript publicized April 10, 2015.

[†] The author is with the School of Computer and Control Engineering, Yantai University, Yantai 264005, P.R. China.

[††] The author is with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, P.R. China.

[†††] The author is with the School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, P.R. China.

a) E-mail: pengsongseu@gmail.com

DOI: 10.1587/transinf.2015EDL8028

can be written as

$$\min \sum_{i=1}^{n} \| x_i - Ds_i \|_p \tag{1}$$

where $\| \cdot \|_p$ is a l_p norm, and p can be 1, 2 or ∞. In this letter, we will concentrate on the case of $p = 2$, so the objective function of sparse coding is given as

$$\begin{aligned} &\min \sum_{i=1}^{n} \| x_i - Ds_i \|_F^2 + \lambda f(s_i) \\ &\text{s.t. } \| d_i \|^2 \leq c, \forall i = 1, 2, \ldots, k \end{aligned} \tag{2}$$

where $\| \cdot \|_F$ is a Frobenius norm, $f(\cdot)$ is the regularization term, which refers to the sparsity of the emotional features, $\lambda \geq 0$ is a regularization parameter, which is a trade-off between the empirical loss and feature sparseness, and c is a constant. In general, the l_0 norm will be directly chosen as the regularization function, but it has been proven to be a NP hard problem [10]. For compromise, an l_1 norm is adopted to replace l_0 to make a convex relaxation [11]. So the objective function becomes as follows

$$\begin{aligned} &\min \sum_{i=1}^{n} \| x_i - Ds_i \|_F^2 + \lambda \sum_{i=1}^{n} \| s_i \|_1 \\ &\text{s.t. } \| d_i \|^2 \leq c, \forall i = 1, 2, \ldots, k \end{aligned} \tag{3}$$

The above equation is a non-convex problem to solve D and S together, and can be computed by the iterative alternative optimization method [9]. Firstly, fixing the dictionary D, the sparse coefficients matrix S will be computed, and the Eq. (3) becomes a classic least absolute shrinkage and selection operator (LASSO) problem. Then, fixing S, this problem becomes a quadratic programming (QP) problem to compute D. These two problems can be easily solved by many optimization methods. Repeat these two steps, until the convergence threshold is reached.

2.2 The Sparse Transfer Learning Method

The sparse coding algorithm can perform feature selection and obtain the sparse representations of the emotional features, but it does not take into account the different feature distributions of different emotional corpora, which will lead to the limited scope for the improvements of recognition rates. As is well known, transfer learning is a technique that aims at addressing the problems when the training and testing datasets follow different distributions or even have different features. It has been proven that transfer learning is beneficial in many applications, e.g., web-document classification, indoor WiFi location and sentiment classification [12]. Inspired by the ideas of good feature selection of sparse coding and model adaptation of transfer learning, a novel sparse transfer learning method, which can optimize them together, is presented n this letter.

Given the emotional features of labeled source and unlabeled target datasets, denoted as $X_{src} = [x_1, x_2, \ldots, x_{n_l}]$ and $X_{tar} = [x_{n_l+1}, x_{n_l+2}, \ldots, x_{n_l+n_u}]$, where n_l and n_u are the numbers of source and target features, respectively. The objective of transfer learning is to minimize the distance between the distributions of labeled and unlabeled features. It is noted that after sparse coding, the emotional features are represented by their sparse feature representations, so the aim becomes to minimize the distance between distributions of the corresponding sparse features. Following [13], the maximum mean discrepancy (MMD) is chosen as the criterion for measuring the distance, and the empirical estimation of MMD is written as

$$\begin{aligned} Dist(S_{src}, S_{tar}) &= \| \frac{1}{n_l} \sum_{i=1}^{n_l} s_i - \frac{1}{n_u} \sum_{j=n_l+1}^{n_l+n_u} s_j \|^2 \\ &= \sum_{i,j=1}^{n_l+n_u} s_i^T s_j m_{ij} \\ &= tr(SMS^T) \end{aligned} \tag{4}$$

where $S = [S_{src}, S_{tar}] \in R^{k \times (n_l+n_u)}$ is the sparse representation of emotional features, $S_{src} = [s_1, s_2, \ldots, s_{n_l}] \in R^{k \times n_l}$ and $S_{tar} = [s_{n_l+1}, s_{n_l+2}, \ldots, s_{n_l+n_u}] \in R^{k \times n_u}$ are the sparse representations for source and target emotional features, respectively, $tr(\cdot)$ refers to the trace of the matrix, and $M = [m_{ij}]_{i,j=1}^{n_l+n_u} \in R^{(n_l+n_u) \times (n_l+n_u)}$ is the MMD matrix, which is given as

$$m_{ij} = \begin{cases} \frac{1}{n_l^2} & s_i, s_j \in S_{src}, \\ \frac{1}{n_u^2} & s_i, s_j \in S_{tar}, \\ -\frac{1}{n_l n_u} & \text{otherwise.} \end{cases} \tag{5}$$

By incorporating the distance measurement (4) into the original sparse coding, the objective function (3) will become as

$$\begin{aligned} &\arg\min_{D,S} \sum_{i=1}^{n} \| x_i - Ds_i \|_F^2 + \lambda \sum_{i=1}^{n} \| s_i \|_1 + \alpha tr(SMS^T) \\ &\text{s.t. } \| d_i \|^2 \leq c, \forall i = 1, 2, \ldots, k \end{aligned} \tag{6}$$

where $n = n_l + n_u$, $\alpha \geq 0$ is a regularization parameter, when $\alpha = 0$, this problem becomes the original sparse coding problem.

As mentioned in 2.1, Eq. (6) is not a convex problem when estimating the unknown parameters together. So following [9], the iterative alternative optimization method is also employed to solve this problem. The procedure is summarized as follows:

1) Fix dictionary D and optimize S, the Eq. (6) becomes as

$$\arg\min_{S} \sum_{i=1}^{n} \| x_i - Ds_i \|_F^2 + \lambda \sum_{i=1}^{n} \| s_i \|_1 + \alpha tr(SMS^T) \tag{7}$$

The above problem can be solved by the coordinate descent optimization algorithm. The strategy is that, in each step, update each vector s_i while other $\{s_j\}_{j \neq i}$ are fixed, then the objective function of optimizing s_i is given as

$$\arg\min_{s_i} \parallel x_i - Ds_i \parallel_F^2 + \lambda \sum_{j=1}^{k} |s_i^j| + \alpha tr(m_{ij} s_i^T s_j) \quad (8)$$

where s_i^j is the j−th component of s_i. The feature-sign search algorithm [14] is adopted to solve the above problem.

2) Fix S and update D, the objective function is rewritten as

$$\arg\min_{D} \parallel X - DS \parallel_F^2 \quad (9)$$

It can be solved by the traditional QP algorithm, then the new D will be incorporated into Eq. (8), repeat these two steps until the objective function is converged.

3. Experiments

3.1 Experimental Setup

The proper selection of databases is critically important to the performance evaluations of speech emotion recognition. In this letter, two popular emotional speech corpora are utilized, i.e., Berlin database [15] and eNTERFACE database [16]. To evaluate the performance of the proposed method for speech emotion recognition, two types of tests are considered, i.e., *case*1 and *case*2. In *case*1, the eNTERFACE database is used as the labeled training corpus, while the Berlin database is chosen as the unlabeled testing corpus. In *case*2, the Berlin database is used as the labeled training corpus, while the eNTERFACE database is chosen as the unlabeled testing corpus, respectively. For *case*1 and *case*2, the common five kinds of emotions are used for evaluation, i.e., anger, disgust, fear, happiness and sadness.

The emotional features are extracted by employing the openSMILE toolkit [17], and the emotional feature set of Interspeech 2010 paralinguistic challenge [18] is chosen. It consists of 1582 dimensional features, the 38 low-level descriptors (LLDs) and their first order regression coefficients are extracted, 21 functionals are applied to the above 76 LLDs, while 16 zero-information features are discarded. Additionally, the F0 numbers of onsets and turn durations are added into the feature set.

SVM is one of the most popular classifiers, and a multiple kernel learning (MKL) based SVM [19] is employed in our experiments. Five kinds of methods are compared, including the automatic recognition method (*Automatic*), in which the classifier trained in source corpus is directly adopted to predict the emotion labels of the target corpus, the sparse coding method (*SC*), the transfer learning via dimension reduction method (*TDR*) [5], [13], the proposed sparse transfer learning method (*Ours*), and the baseline method, in which the training and testing procedures are carried out on single corpus (*Baseline*).

Table 1 Average recognition rates using different methods in *case*1 (anger:A, disgust:D, fear:F, happiness:H, sadness:S).

Methods	Recognition rates (%)					
	A	D	F	H	S	Average
Baseline	75.96	85.14	71.98	58.17	84.36	82.13
Automatic	34.05	55.87	18.05	23.64	49.15	36.74
SC	37.25	74.28	19.91	27.26	70.84	41.13
TDR	38.36	75.42	20.84	28.16	77.05	53.95
Ours	39.12	76.93	21.85	27.89	76.91	55.47

Table 2 Average recognition rates using different methods in *case*2 (anger:A, disgust:D, fear:F, happiness:H, sadness:S).

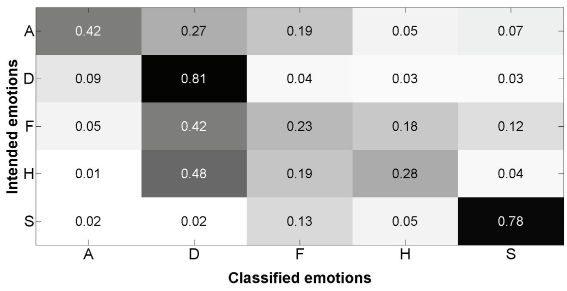
Methods	Recognition rates (%)					
	A	D	F	H	S	Average
Baseline	78.75	58.06	57.15	62.23	63.98	63.15
Automatic	39.83	19.61	18.32	29.58	29.21	23.64
SC	46.76	22.32	31.13	45.05	42.94	31.25
TDR	54.21	26.43	40.89	48.52	49.85	45.27
Ours	55.15	26.84	39.15	50.13	50.42	46.68

3.2 Experimental Results and Discussions

In our experiments, the unlabeled target dataset is divided into five equal parts, among which, 4/5 are used for training, and the other 1/5 are for testing. A 10-fold cross validation is adopted to choose the parameters λ and α, which are optimized as 0.1 and 10^5, respectively. The experiments are conducted 20 times to cover all the cases.

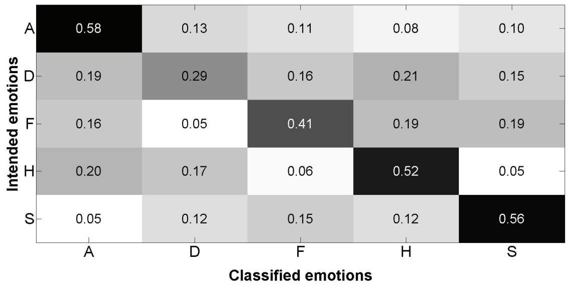
Tables 1 and 2 summarize the recognition rates using different methods in *case*1 and *case*2, respectively. Firstly, it can be found that in both cases, compared to the *Automatic* method, the proposed *SC* method can obtain higher recognition rates, this might be attributed to a good feature selection via sparse coding algorithm. Secondly, it can be easily observed that the *TDR* method can obtain better recognition rates than *SC* and *Automatic* approaches, with over 21% and 14% improvements, respectively, which indicates that the transfer learning techniques are very efficient for cross-corpus speech emotion recognition. Thirdly, our proposed sparse transfer learning method can obtain higher recognition rates than the other methods except the *Baseline* method.

To evaluate the performance of our proposed method, the confusion values between the emotions are also summarized. In Figs. 1 and 2, the confusion matrices of our proposed method are given for *case*1 and *case*2, respectively. It can be observed that, in *case*1, the disgust and sadness achieve the highest recognition rates, while in *case*2, the anger, sadness and happiness obtain better performance with over 50% accuracies. In Fig. 1, the highest confusion value is 0.48, in which, the happiness is more easily mistaken for disgust. Meanwhile, the highest confusion value is 0.21 in Fig. 2, in which, the disgust is more likely to be recognized as happiness. This might be attributed to that the arousal levels are similar when the speakers are in these two emotional states [20].



	A	D	F	H	S
A	0.42	0.27	0.19	0.05	0.07
D	0.09	0.81	0.04	0.03	0.03
F	0.05	0.42	0.23	0.18	0.12
H	0.01	0.48	0.19	0.28	0.04
S	0.02	0.02	0.13	0.05	0.78

Fig. 1 Confusion matrix of emotions in *case*1 (anger:A, disgust:D, fear:F, happiness:H, sadness:S).



	A	D	F	H	S
A	0.58	0.13	0.11	0.08	0.10
D	0.19	0.29	0.16	0.21	0.15
F	0.16	0.05	0.41	0.19	0.19
H	0.20	0.17	0.06	0.52	0.05
S	0.05	0.12	0.15	0.12	0.56

Fig. 2 Confusion matrix of emotions in *case*2 (anger:A, disgust:D, fear:F, happiness:H, sadness:S).

4. Conclusions

In this letter, a novel sparse feature transfer learning approach is presented for speech emotion recognition. The sparse coding algorithm is first presented to achieve a common dictionary and robust feature representations for both labeled source and unlabeled target emotional features. Meanwhile, to cope with the different feature distributions between different corpora, the sparse coding and transfer learning approaches are combined, and the MMD algorithm, which describes the distance between two distributions, is added into the objective function of sparse coding as a regularization term. Experiments are conducted on the public emotional datasets, and the results confirm the efficacy of our proposed method.

Acknowledgements

This study was supported by the Natural Science Foundation of Shandong Province (ZR2014FQ016), the Fundamental Research Funds for the Southeast University (CDLS-2015-04), and the National Natural Science Foundation of China (61231002 and 61301219).

References

[1] D. Ververidis, C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Communication, vol.48, no.9, pp.1162–1181, 2006.

[2] M.E. 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.

[3] M.R. Amer, B. Siddiquie, C. Richey, and A. Divakaran, “Emotion detection in speech using deep networks,” Proc. ICASSP, Florence, Italy, pp.3752–3756, 2014.

[4] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, “Cross-corpus acoustic emotion recognition: variances and strategies,” IEEE Trans. Affective Comput., vol.1, no.2, pp.119–131, 2010.

[5] P. Song, Y. Jin, L. Zhao, and M. Xin, “Speech emotion recognition using transfer learning,” IEICE Trans. Inf. & Syst., vol.E97-D, no.9, pp.2530–2532, 2014.

[6] B.A. Olshausen and D.J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” Nature, vol.381, no.6583, pp.607–609, 1996.

[7] B.A. Olshausen and D.J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” Vision research, vol.37, no.23, pp.3311–3325, 1997.

[8] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.2, pp.210–227, 2009.

[9] K. Huang, and S. Aviyente, “Sparse representation for signal classication,” Proc. NIPS, pp.609–616, Vancouver, Canada, 2006.

[10] B.K. Natarajan, “Sparse approximate solutions to linear systems,” SIAM Journal on Computing, vol.24, no.2, pp.227–234, 1995.

[11] S.S. Chen, and D.L. Donoho, M.A. Saunders, “Atomic decomposition by basis pursuit,” SIAM Journal on Scientific Computing, vol.20, no.1, pp.33–61, 1998.

[12] S.J. Pan, and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol.22, no.10, pp.1345–1359, 2010.

[13] S.J. Pan, J.T. Kwok, and Q. Yang, “Transfer learning via dimensionality reduction,” Proc. AAAI, Chicago, U.S.A., pp.677–682, July 2008.

[14] H. Lee, A. Battle, R. Raina, and A.Y. Ng, “Efficient sparse coding algorithms,” Proc. NIPS, pp.801–808, Vancouver, Canada, 2006.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” Proc. Interspeech, pp.1517–1520, Lisbon, Portugal, 2005.

[16] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The eNTERFACE’05 audio-visual emotion database,” Proc. International Conference on Data Engineering Workshops. Atlanta, USA, pp.8–8, 2006.

[17] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” Proc. ACM Multimedia, Firenze, Italy, pp.1459–1462, Oct. 2010.

[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C.A. Muller, and S.S. Narayanan, “The interspeech 2010 paralinguistic challenge,” Proc. Interspeech, pp.2794–2797, Makuhari, Japan, 2010.

[19] Y. Jin, P. Song, W. Zheng, L. Zhao, and M. Xin, “Speaker-independent speech emotion recognition based on two-Layer multiple kernel learning,” IEICE Trans. Inf. & Syst., vol.E96-D, no.10, pp.2286–2289, 2013.

[20] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, “Emotion recognition in human–computer interaction,” IEEE Signal Process. Mag., vol.18, no.1, pp.32–80, 2001.