LETTER

Cross-Corpus Speech Emotion Recognition Based on Causal Emotion Information Representation*

Hongliang FU[†a)], **Qianqian LI**[†], **Huawei TAO**[†], **Chunhua ZHU**[†], ***Nonmembers***, **Yue XIE**[††], ***Member***, ***and*** **Ruxue GUO**[†††], ***Nonmember***

SUMMARY Speech emotion recognition (SER) is a key research technology to realize the third generation of artificial intelligence, which is widely used in human-computer interaction, emotion diagnosis, interpersonal communication and other fields. However, the aliasing of language and semantic information in speech tends to distort the alignment of emotion features, which affects the performance of cross-corpus SER system. This paper proposes a cross-corpus SER model based on causal emotion information representation (CEIR). The model uses the reconstruction loss of the deep autoencoder network and the source domain label information to realize the preliminary separation of causal features. Then, the causal correlation matrix is constructed, and the local maximum mean difference (LMMD) feature alignment technology is combined to make the causal features of different dimensions jointly distributed independent. Finally, the supervised fine-tuning of labeled data is used to achieve effective extraction of causal emotion information. The experimental results show that the average unweighted average recall (UAR) of the proposed algorithm is increased by 3.4% to 7.01% compared with the latest partial algorithms in the field.

key words: *cross-corpus speech emotion recognition, causal representation learning, domain adaptation*

1. Introduction

Speech and paralanguage recognition technology has always been the research focus in the field of human-computer interaction, it is of great significance in computer intelligence. For a single corpus, the accuracy of machine speech emotion recognition (SER) has exceeded the human level [1]. However, when the training and testing set are from different corpus, the system performance of cross-corpus recognition deteriorates significantly, which means that cross-corpus SER is still the key problem to be solved in the process of practical application of emotion recognition technology.

Manuscript received December 4, 2023.
Manuscript revised March 4, 2024.
Manuscript publicized April 12, 2024.

[†]Key Laboratory of Grain Information Processing and Control, Ministry of Education, Henan University of Technology, Zhengzhou, 450001, China.
[††]School of Communication Engineering, Nanjing Institute of Technology, Nanjing, 211167, China.
[†††]IFLYTEK Research, Hefei, 230088, China.

*This research project was founded in part by Natural Science Project of Henan Education Department (No. 22A520004 and No. 22A510001), Innovative Funds Plan of Henan University of Technology (2022ZKCJ13), National Natural Science Foundation of China (No. 62001215), Open project of scientific research platform of Henan University of Technology Grain Information Processing Center (No. KFJJ2023011).

a) E-mail: jackfu_zz@163.com

DOI: 10.1587/transinf.2023EDL8087

In recent years, research on cross-corpus SER has focused on solving the problem of sample space imbalance caused by multi corpus sources, and a variety of techniques such as subspace, adversarial and domain adaptive have been used to solve this problem. Zhang and Li et al. [2], [3] introduced the source domain and target domain features into a new subspace to eliminate the differences between different corpus. Gao et al. [4] used the domain adversarial training method to constrain the intra-class variation of the same emotion feature. In domain adaptation, Zhang et al. [5] used the adaptive method to learn the regression matrix by comprehensively considering the marginal probability distribution and conditional probability distribution between training and testing set. Zhuang et al. [6] adopted a deep domain adaptation method and combined with subdomain adaptation to achieve fine-grained feature distribution alignment. However, these methods still face challenges. Firstly, factors like differences in sample size and language lead to feature distribution is poorly aligned, restricting the effectiveness of transfer. Secondly, classical acoustic statistical features often incorporate substantial non-emotional information, diminishing the model's recognition capabilities.

To address the above challenges, this paper proposes a cross-corpus SER method based on causal emotion information representation (CEIR). In this paper, we attempt to separate the non-causal information of features [7], and then improve the transfer performance. Firstly, using the labeled information from the source domain and the reconstruction loss of the deep autoencoder network, causal features are separated. Secondly, a causal decomposition matrix is constructed, combined with the local maximum mean difference (LMMD) feature alignment techniques, ensuring the joint distribution independence of causal features' dimensions while addressing feature distribution discrepancies. Lastly, incorporating supervised fine-tuning with labeled data from the source domain ensures that the classification task, encompassing all causal information, is causally sufficient.

2. Methods

Existing studies have shown that causality can obtain more effective feature information. Based on the research of [8]–[10], causal features should meet three properties: separation of causal and non-causal features; causal feature dimension independence; the original features and labels must have a

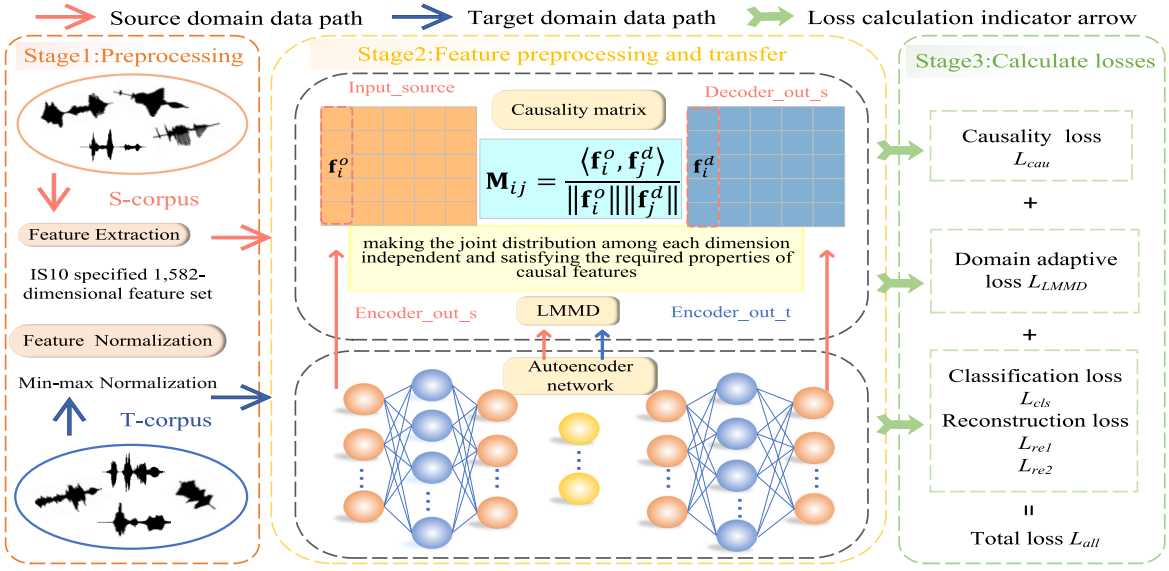


Fig. 1 Cross-corpus speech emotion recognition model framework based on CEIR.

causal relationship. Figure 1 shows the algorithmic process.

2.1 Feature Extraction and Normalization

The experiments extract 1,582-dimensional features based on the configure of IS10 [11] by the openSMILE [12] tool. It includes 34 basic low-level descriptors (LLDs), i.e., Mel-frequency cepstral coefficient (MFCC), line spectrum pair (LSP), loudness and 34 corresponding delta coefficients. Based on these LLDs, 21 statistical functions are applied to obtain 1,428 features. Additionally, applying 19 statistical functions to the 4 pitch-based LLDs and their corresponding delta coefficients results in 152 features. The onset of pitch and durations of utterances are added into the final two features, resulting in a total of 1,582 features. Then, these features are normalized to accelerate model convergence.

2.2 Feature Processing and Transfer

2.2.1 Separation of Causal and Non-Causal Features

Some domain-specific information such as individual differences and languages in SER, which cannot determine the emotional category of the input samples, are considered non-causal information. CEIR uses two deep autoencoders to reduce the dimensionality of features in the source and target domains and separate label and domain-related features to some extent. This process utilizes a symmetric encoder-decoder structure to achieve data reconstruction. The encoding and decoding process is as follows:

$$\mathbf{h} = f(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{1}$$

$$\mathbf{x}' = f(\mathbf{W}' \cdot \mathbf{h} + \mathbf{b}') \tag{2}$$

where $\mathbf{x}$ is the input sample feature, $f(\cdot)$ is the activate function, $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and bias used in the encoding process, while $\mathbf{W}'$ and $\mathbf{b}'$ correspond to the decoding process, $\mathbf{h}$ is the encoding output and $\mathbf{x}'$ is the decoded output.

The reconstruction error for this process uses the mean square error (MSE), and the loss is constructed as follows:

$$L_{re} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x} - \mathbf{x}'\|^2 \tag{3}$$

2.2.2 Causal Feature Dimensions Are Independent

After feature processing, each dimension of the features needs to be independent of each other, which is realized by using the causal correlation matrix and subdomain adaptation. The dimensions consistent with the original features are obtained by autoencoding the source domain features, and the correlation matrix is constructed to measure the correlation of the same dimension and the independence of different dimensions of the features before and after reconstruction. The correlation matrix is designed as follows:

$$\mathbf{M}_{ij} = \frac{\left\langle \mathbf{f}_i^o, \mathbf{f}_j^d \right\rangle}{\left\|\mathbf{f}_i^o\right\| \left\|\mathbf{f}_j^d\right\|}, \quad i, j \subseteq 1, 2, \ldots, D \tag{4}$$

where $\mathbf{f}^o$ and $\mathbf{f}^d$ represent the original feature and the decoded feature, respectively, and D represent the feature dimension. The correlation matrix is optimized into unit matrix $\mathbf{I}$, and the causal relationship between features and labels is obtained, causal loss is defined as follows:

$$L_{cau} = \|\mathbf{M} - \mathbf{I}\|_F^2 \tag{5}$$

By minimizing the causal loss, the diagonal elements of the correlation matrix can be 1, the non-diagonal elements can be 0, and isolating causal features while concurrently

ensuring the joint distribution independence among feature dimensions.

In order to improve the discrimination of target domain features, LMMD [13] is introduced on the basis of causality to learn more common features in the same sentiment category. LMMD is defined as follows:

$$L_{LMMD} = \frac{1}{c}\sum_{c=1}^{c}\left\|\sum w^{s}\Phi(\mathbf{x}_i^{s}) - \sum w^{t}\Phi(\mathbf{x}_j^{t})\right\|_H^2 \quad (6)$$

where H is the Reproducing Kernel Hilbert Space (RKHS), $\Phi(\cdot)$ indicates that the original sample is mapped to a certain feature map of the RKHS, $\mathbf{x}^s$ and $\mathbf{x}^t$ respectively represent the source domain features and target domain features output by the encoder, while w^s and w^t represent the corresponding weights.

2.2.3 The Original Features and Labels Must Have a Causal Relationship

In order to successfully perform the classification task, the representation should be causally sufficient. The simplest way to do this is to use the features generated by the decoder to perform emotion classification in the source domain, extract causal information, the loss for source domain classification is designed as follows:

$$L_{cls} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{5}\mathbf{y}_{ic}\log(\hat{\mathbf{y}}_{ic}) \quad (7)$$

where B represents the batch size in the training process, $\mathbf{y}_i$ is an one-hot vector indicating the label of $\mathbf{x}_i$, $\hat{\mathbf{y}}_{ic}$ denotes the predicted probability that the sample belongs to category c sentiment type.

2.3 Loss and Joint Optimization

Stochastic gradient descent (SGD) is used to optimize the loss function until the model converges, with the specific algorithm details provided in Table 1.

Table 1 Joint loss.

Algorithm: Causal emotion information representation (CEIR)
Input: $\mathbf{x}^s$, $\mathbf{x}^t$, $\mathbf{y}^s$, $\mathbf{y}^t$, parameter α, λ, β, η, γ;
Output: total loss L_{all};
1: Initialization: Set $i = 0$ and initialize parameters $\mathbf{W}$, $\mathbf{b}$;
2: **while** $i < 5,040$ **do**
3: Calculation Equation (3) (5) (6) (7);
4: $L_{all} = \alpha L_{cls} + \lambda L_{cau} + \beta L_{re1} + \eta L_{re2} + \gamma L_{LMMD}$;
5: **if** temp loss < total loss **then**
6: total loss = temp loss;
7: **else**
8: update parameter;
9: **end**
10: **end while**

3. Experiment

3.1 Experimental Setup

To evaluate the performance of the proposed model, three commonly used databases were selected for extensive experimentation, including Berlin [14], eNTERFACE [15] and CASIA database [16]. Six cross-corpus SER tasks are devised using three emotion corpus, the specific task settings are shown in Table 2.

In six tasks, the learning rate and batch size are set respectively to 0.001 and 32. The iteration round is set to 5,040. The unweighted average recall (UAR) is used as the evaluation index to assess the effects of different models. UAR is a popular evaluation index in the field of emotion recognition.

3.2 Analysis and Discussion of Results

3.2.1 Ablation Experiments and Visualized Analysis

In order to evaluate the rationality of the model, the ablation experiment in Fig. 2 and the visual analysis in Fig. 3 are established.

Intuitively from Fig. 2, CEIR performs optimally across six tasks, and ignoring any one of the algorithms will degrade the performance, which indicates that several algorithms combined in this paper are effective. In the visualization experiments using B-e as an example, where different colors represent distinct feature types. In (a), without causal and LMMD processing, features lack clear classification and exhibit unclear boundaries. Contrastingly, in (b), features undergo causal correlation analysis, resulting in more potent

Table 2 Task setting for cross-corpus speech emotion recognition.

Source domain	Target domain	Shared emotion types
eNTERFACE(e)	Berlin(B)	Anger, disgust, fear
Berlin(B)	eNTERFACE(e)	happiness, sadness
Berlin(B)	CASIA(C)	Anger, fear, happiness
CASIA(C)	Berlin(B)	neutrality, sadness
eNTERFACE(e)	CASIA(C)	Anger, fear, happiness
CASIA(C)	eNTERFACE(e)	sadness, surprise

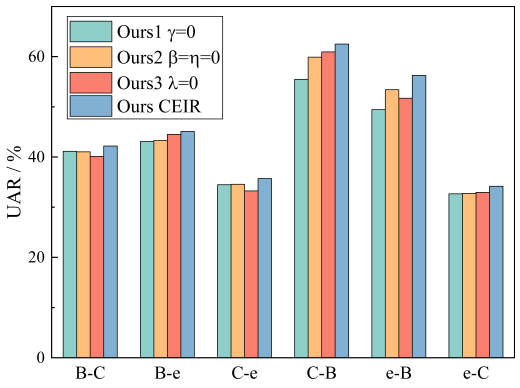
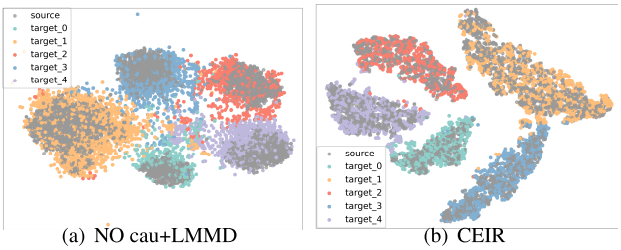


Fig. 2 Ablation experiments.



(a) NO cau+LMMD (b) CEIR

Fig. 3 Visualized analysis.

Table 3 The comparison of UAR for different algorithms.

Tasks	B-C	B-e	C-e	C-B	e-B	e-C	Average
TSDSL [2]	37.40	35.44	33.25	56.74	47.41	32.50	40.46
JDAR [5]	38.60	38.14	28.43	49.58	48.74	30.30	38.97
DANN [17]	**42.89**	36.53	29.17	57.64	52.67	**36.60**	42.58
DASA [6]	41.40	40.11	32.09	51.47	52.35	36.10	42.25
CEIR	42.18	**45.08**	**35.71**	**62.51**	**56.25**	34.17	**45.98**

representations and showcasing clear classification effects.

3.2.2 Comparative Experiment

In order to further evaluate the advancement of the algorithm, the proposed algorithm is compared with the most advanced algorithms in the field. These algorithms are transfer sparse discriminant subspace learning (TSDSL) [2], joint distribution adaptive regression (JDAR) [5], domain adversarial neural network (DANN) [17], and deep autoencoder subdomain adaptation (DASA) [6].

Observing Table 3, firstly, methods based on deep domain adaptation achieve better performance. Secondly, our proposed algorithm outperforms in multiple tasks, especially demonstrating significant improvement in the C-B task. Ultimately, across six tasks, CEIR surpasses other algorithms by 3.4% to 7.01% in average UAR.

4. Conclusion

To improve cross-corpus SER, this paper introduces a method based on CEIR. Using deep autoencoder networks, this approach separates causal features from domain and label features. The causal decomposition matrix ensures independent feature distributions, while LMMD feature alignment addresses distribution differences. In the final training phase, supervised fine-tuning maximizes the use of causal information. Experimental results show that causal learning enhances the ability of cross-corpus emotion recognition. This research has laid a solid foundation for subsequent work in related fields and provided valuable insights for the practical application of SER and cross-corpus recognition tasks.

References

[1] F.A.D. Rí, F.C. Ciardi, and N. Conci, "Speech emotion recognition and deep learning: an extensive validation using convolutional neural networks," IEEE Access, vol.11, pp.116638–116649, 2023.

[2] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," IEEE/ACM Trans. Audio Speech Lang. Process., vol.28, pp.307–318, 2020.

[3] S. Li, P. Song, L. Ji, Y. Jin, and W. Zheng, "A generalized subspace distribution adaptation framework for cross-corpus speech emotion recognition," ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp.1–5, 2023.

[4] Y. Gao, S. Okada, L. Wang, J. Liu, and J. Dang, "Domain-invariant feature learning for cross corpus speech emotion recognition," ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp.6427–6431, 2022.

[5] J. Zhang, L. Jiang, Y. Zong, W. Zheng, and L. Zhao, "Cross-corpus speech emotion recognition using joint distribution adaptive regression," ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, pp.3790–3794, 2021.

[6] Z. Zhuang, H. Fu, H. Tao, J. Yang, Y. Xie, and L. Zhao, "Cross-corpus speech emotion recognition based on deep autoencoder subdomain adaptation," Application Research of Computers, vol.38, no.11, pp.3279–3282+3348, 2021.

[7] F. Lv, J. Liang, S. Li, B. Zang, C.H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp.8036–8046, 2022.

[8] J. Peters, D. Janzing, and B. Schölkopf, Elements of causal inference: Foundations and learning algorithms, MIT Press, Cambridge, MA, USA, 2017.

[9] H. Reichenbach, The Direction of Time, University of California Press, Berkeley, 1991.

[10] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," ICML, pp.1255–1262, 2012.

[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S.S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," INTERSPEECH, Makuhari, Japan, pp.2794–2797, 2010.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," Proc. 18th ACM International Conference on Multimedia, pp.1459–1462, ACM, 2010.

[13] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, and Q. He, "Deep subdomain adaptation network for image classification," IEEE Trans. Neural Netw. Learn. Syst., vol.32, no.4, pp.1713–1722, 2021.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of German emotional speech," INTERSPEECH 2005, pp.1517–1520, 2005.

[15] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE '05 audiovisual emotion database," 22nd International Conference on Data Engineering Workshops (ICDEW '06), p.8, IEEE, 2006.

[16] J. Tao, F. Liu, M. Zhang, and H. Jia, "Design of speech corpus for mandarin text to speech," The Blizzard Challenge 2008 workshop, 2008.

[17] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," IEEE/ACM Trans. Audio Speech Lang. Process., vol.26, no.12, pp.2423–2435, 2018.