Interspeech 2024
1-5 September 2024, Kos, Greece

Hierarchical Distribution Adaptation for Unsupervised Cross-corpus Speech Emotion Recognition

Cheng Lu[1], *Yuan Zong*[1,*], *Yan Zhao*[1], *Hailun Lian*[1], *Tianhua Qi*[1], *Björn Schuller*[2], *Wenming Zheng*[1,*]

[1]Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, China [2]GLAM - the Group on Language, Audio, & Music, Imperial College London, UK

cheng.lu@seu.edu.cn, xhzongyuan@seu.edu.cn, wenming_zheng@seu.edu.cn

Abstract

The primary issue of unsupervised cross-corpus speech emotion recognition (SER) is that domain shift between the training and testing data undermines the SER model's ability to generalize on unknown testing datasets. In this paper, we propose a straightforward and effective strategy, called Hierarchical Distribution Adaptation (HDA), to address the domain bias issue. HDA leverages a hierarchical emotion representation module based on nested Transformers to extract speech emotion features at different levels (e. g., frame/segment/utterance-level), for capturing multiple-scale emotion correlations in speech. Furthermore, a hierarchical distribution adaptation module, including frame-level distribution adaptation (FDA), segment-level distribution adaptation (SDA), and utterance-level distribution adaptation (UDA), is developed to align the hierarchical-level emotion representations of the training and testing speech samples to effectively eliminate domain discrepancy. Extensive experimental results demonstrate the superiority of our proposed HDA over other state-of-the art (SOTA) methods.

Index Terms: Cross-corpus, speech emotion recognition, speech representation, hierarchical domain adaptation

1. Introduction

Speech Emotion Recognition (SER) enables machines to better understand and respond to human emotions, which plays a vital role in applications like human-computer interaction [1], affective computing [2], and digital health [3]. When the SER model is trained on one or multiple corpora, while tested on another corpus, called unsupervised cross-corpus SER task [4], the performance of the trained models tends to degradation on new data. This is primarily due to significant domain shift in feature distributions between the training data (source domain) and testing data (target domain) [5].

To address the domain bias issue, a practical solution is utilizing Domain Adaptation (DA) to eliminate the distribution disparities between the source and target domains [5]. These DA-based methods have achieved success in cross-corpus SER [6], [7]. The fundamental idea of DA is to seek a common feature subspace for the source and target domains, where the feature distributions between domains can be effectively measured and brought closer [8], [9]. This idea is straightforward and can be well-incorporated into linear or nonlinear subspace feature learning through linear projection [8] or kernel trick [10]), as well as into the high-dimensional deep feature learning through deep neural networks (e. g., CNN[11], RNN[12], and Transformer[13], [14]). Obvious, the advantage of DA-based method lies in its ability to effectively represent the feature distributions of the source and target data in the feature space of speech emotion, thus facilitating the measurement and elimination of distribution discrepancy between domains.

Nevertheless, these DA-based methods typically measure distribution discrepancy using the task-specific speech features, e. g., Low-level Descriptor (LLD) feature sets [8] and fully-connected layer features [11]. These features have the advantage of strong emotional relevance and are favorable for emotion classifiers. However, their weakness lies in potential obstruction to feature distribution alignment when updating network parameters to promote classification. This issue may be due to the high coupling between emotional features and other acoustic features (e. g., speaker information, noise) in speech signals [15], [16]. Therefore, capturing effective emotional patterns from speech signals while aligning inter-domain feature distributions and maintaining the emotion discriminability of speech features is crucial for cross-corpus SER.

Moreover, some SER studies have revealed that the emotional cues are distributed across multiple scales, e. g., word-/phrase-/utterance-level [13], [14]. These multi-scale emotional information can be aggregated into global emotional features that reflect the overall semantic information of the sentence, as well as correspond to the semantic gaps in emotional information at different levels (e. g., frame-/segment-/utterance-level). Thus, to accurately measure and mitigate the distribution discrepancy in emotion features across domains, it is necessary to explore hierarchical-level speech emotion features.

Inspired by above considerations, we propose a hierarchical distribution alignment (HDA) strategy for cross-corpus SER. HDA is implemented through a hierarchical emotion representation module based on nested Transformers and a hierarchical distribution adaptation module based on three levels of distribution measurement strategies. Its main advantages can be summarized as follows: (1) *Effectively excavating different-level emotion features from speech signals through a Transformer-based hierarchical network*; (2) *Precisely measuring and mitigating the distribution shift between domains using a hierarchical alignment strategy*, obtaining the speech representation that is discriminative to emotions and robust across domains.

2. Methodology

As shown in Fig.1, HDA inculdes the hierarchical emotion representation and hierarchical distribution adaptation modules, aiming to obtain different-scale emotional features and align feature distributions at different levels between domains.

2.1. Hierarchical Emotion Representation

Hierarchical Emotion Representation module aims to learn different-scale emotion representations (i. e., frame-/segment-

* Corresponding authors.

10.21437/Interspeech.2024-1948

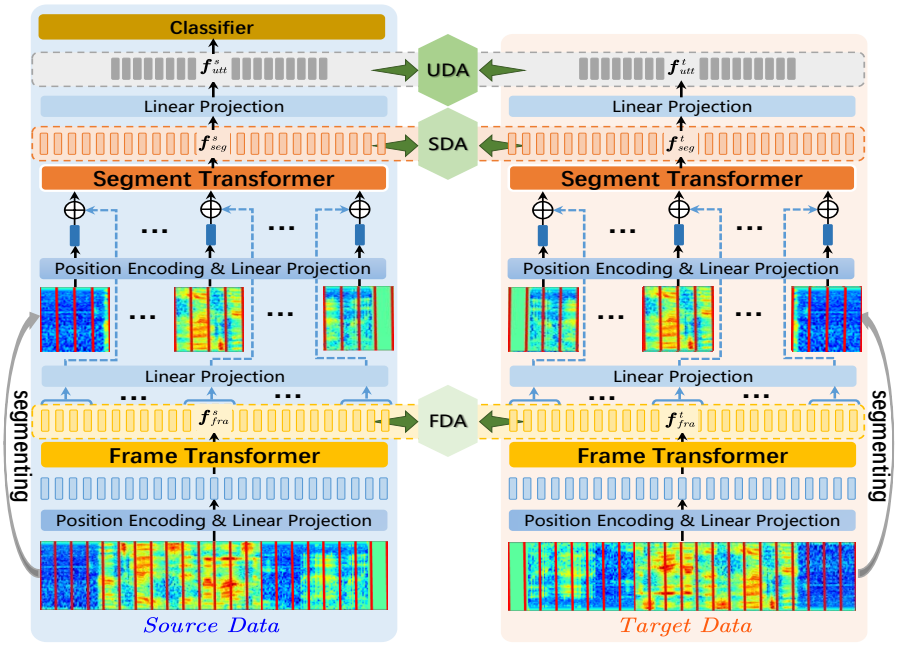


Figure 1: *Overall pipeline of Hierarchical Distribution Adaptation (HDA) for cross-corpus SER, where the backbone networks for source and target data share parameters.*

/utterance-level) of speech through the nested Transformers. To this end, we first formalize the input features of HDA. Herein, we utilize log-Mel-spectrogram features $\boldsymbol{x} \in \mathbb{R}^{F\times T\times C}$ as the model's input, where F, T, and C represent the numbers of Mel-filter banks, speech frames, and channels, respectively.

2.1.1. Frame-Level Emotion Representation

Frame-level emotion representation is obtained by operating on each frame using Frame Transformer. Specifically, we linearly project the i^{th} frame $\boldsymbol{x}_i \in \mathbb{R}^{F\times C}$ of the input feature $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{T}$ to obtain the frame-level embedding $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{1\times d_f}$, where d_f is the dimension of each frame embedding. This operation can be defined as

$$\hat{\boldsymbol{x}}_i = LP(\boldsymbol{x}_i), \tag{1}$$

where $LP(\cdot)$ represents the linear projection operation.

Furthermore, we add a learnable position encoding $\boldsymbol{p}_i \in \mathbb{R}^{1\times d}$ to $\hat{\boldsymbol{x}}_i$ as $\hat{\boldsymbol{x}}_i^f \in \mathbb{R}^{1\times d_f}$, which is inductive bias of frame Transformer according to [17]. The operation is denoted as

$$\hat{\boldsymbol{x}}_i^f = \hat{\boldsymbol{x}}_i + \boldsymbol{p}_i. \tag{2}$$

Then, we calculate the emotional correlation between speech frames through frame Transformer and obtain the frame-level feature $\boldsymbol{f}_{fra} \in \mathbb{R}^{T\times d_f}$, which can be defined as

$$\begin{aligned} \boldsymbol{m}^f &= MSA(LN(\hat{\boldsymbol{x}}^f)) + \hat{\boldsymbol{x}}^f, \\ \boldsymbol{f}_{fra} &= MLP(LN(\boldsymbol{m}^f)) + \boldsymbol{m}^f, \end{aligned} \tag{3}$$

where $\hat{\boldsymbol{x}}^f = \{\hat{\boldsymbol{x}}_i^f\}_{i=1}^{T} \in \mathbb{R}^{T\times d_f}$ is the frame embedding sequence after adding position coding and $\boldsymbol{m}^f$ is the output of the Multi-head Self-Attention (MSA) in frame Transformer. The $LN(\cdot)$, $MLP(\cdot)$, and $MSA(\cdot)$ represent the operations of layer normalization, multi-layer perceptron, and MSA [17].

2.1.2. Segment-Level Emotion Representation

To represent segment-level speech emotion, we need jointly input the segment-level log-Mel-spectrogram $\boldsymbol{s} = \{\boldsymbol{s}_j\}_{j=1}^{T/k}$ and frame-level feature $\boldsymbol{f}_{fra}$ by frame Transformer into segment Transformer to further learn emotional correlations between segments. $\boldsymbol{s}$ is the segment-level partition of input log-Mel-spectrogram $\boldsymbol{x}$ and $\boldsymbol{s}_j \in \mathbb{R}^{k\times F\times C}$ is composed of k frames log-Mel-spectrogram $\hat{\boldsymbol{x}}_i$ concatenation.

Then, we flatten the j^{th} segment feature $\boldsymbol{s}_j$ and its corresponding k frames feature $\boldsymbol{f}_j$ in $\boldsymbol{f}_{fra}$ to the same dimension $\mathbb{R}^{1\times(k\times d_f)}$. Then, we linearly project the flattened features and combine them to obtain the segment-level embedding $\hat{\boldsymbol{s}}_j \in \mathbb{R}^{1\times d_s}$, where d_s is the dimension of the segment-level embedding. The process can be represented as

$$\hat{\boldsymbol{s}}_j = LP(FL(\boldsymbol{s}_j)) + LP(FL(\boldsymbol{f}_j)), \tag{4}$$

where $FL(\cdot)$ represents flattening operation.

Similar to frame Transformer, segment Transformer is also stacked by L standard Transformer blocks. We add a learnable positional encoding $\boldsymbol{p}_j^s \in \mathbb{R}^{1\times d_s}$ to $\hat{\boldsymbol{s}}_j$ for input feature $\hat{\boldsymbol{s}}_j^s$ of segment Transformer, which can be denoted as

$$\hat{\boldsymbol{s}}_j^s = \hat{\boldsymbol{s}}_j + \boldsymbol{p}_j^s, \tag{5}$$

where $\boldsymbol{p}^s = \{\boldsymbol{p}_j^s\}_{j=1}^{T/k} \in \mathbb{R}^{(T/k)\times d_s}$ and $\hat{\boldsymbol{s}}^s = \{\hat{\boldsymbol{s}}_j^s\}_{j=1}^{T/k} \in \mathbb{R}^{(T/k)\times d_s}$.

Then, we calculate emotional correlations between segments using stacked segment Transformer and the operations of the first block can be defined as

$$\begin{aligned} \boldsymbol{m}^s &= MSA(LN(\hat{\boldsymbol{s}}^s)) + \hat{\boldsymbol{s}}^s, \\ \boldsymbol{f}_{seg} &= MLP(LN(\boldsymbol{m}^s)) + \boldsymbol{m}^s, \end{aligned} \tag{6}$$

where $\boldsymbol{m}^s$ is the output of the MSA unit and $\boldsymbol{f}_{seg} \in \mathbb{R}^{(T/k)\times d_s}$ is the output of the segment Transformer.

Note that frame and segment Transformers are both based on L stack blocks. Eq.3 and Eq.6 show the operations of one block, with others following a similar process.

2.1.3. Utterance-Level Emotion Representation

After obtaining segment emotion features by aggregating frame emotion features $\boldsymbol{f}_{fra}$ and segment log-Mel-spectrogram features $\boldsymbol{s}$, we linearly project the output features of the segment Transformer $\boldsymbol{f}_{seg}$ to obtain the utterance-level emotional feature $\boldsymbol{f}_{utt} \in \mathbb{R}^{1\times d_u}$, where d_u is dimension of utterance-level embedding. The operation can be represented as follows

$$\boldsymbol{f}_{utt} = \mathit{LP}(\boldsymbol{f}_{seg}). \tag{7}$$

Then, we can use $\boldsymbol{f}_{utt}$ to predict the emotion label y_{pre} through the classifier, then calculate the emotion classification loss $\mathcal{L}_{ce}$ between the predicted label y_{pre} and the ground-true label y by the cross-entropy loss function $\mathit{CE}(\cdot)$, denoted as

$$\mathcal{L}_{ce} = \mathit{CE}(y_{pre}, y). \tag{8}$$

2.2. Hierarchical Distribution Adaptation

Through the above hierarchical emotion representation module, source and target data can both obtain their different-scale speech emotion features. Note that the backbone networks for source and target domains share parameters. Then, we need accurately measure the feature distribution discrepancy between training and testing datasets, and effectively eliminate the domain shifts by the hierarchical distribution adaptation module.

Specifically, we utilize Multi-Kernel Maximum Mean Discrepancy (MK-MMD) to measure the distribution shift at different levels across domains [18], which is an effective method to evaluate the distance between two data distributions in a high-dimensional reproducing kernel Hilbert space (RKHS) [5] and has achieved success in SER tasks [15], [11].

Aiming to achieve feature distribution alignment across three scales in hierarchical emotion representation module, our first preference is to implement **Frame-Level Distribution Adaptation (FDA)**. Thus, FDA loss $\mathcal{L}_f$ can be generated through calculating the MK-MMD disttance $\mathit{MK\text{-}MMD}(\cdot)$ using frame-level emotion features of source domain $\boldsymbol{f}^s_{fra}$ and target domain $\boldsymbol{f}^t_{fra}$, which can be represented as

$$\mathcal{L}_f = \mathit{MK\text{-}MMD}(\boldsymbol{f}^s_{fra}, \boldsymbol{f}^t_{fra}). \tag{9}$$

Similarly, we can obtain the **Segment-Level Distribution Adaptation (SDA)** by segment-level features $\boldsymbol{f}^s_{seg}$ and $\boldsymbol{f}^t_{seg}$ of source and target domains, and FDA loss $\mathcal{L}_s$ can be defined as

$$\mathcal{L}_s = \mathit{MK\text{-}MMD}(\boldsymbol{f}^s_{seg}, \boldsymbol{f}^t_{seg}). \tag{10}$$

Also, **Utterance-Level Distribution Adaptation (UDA)** can be implemented through the utterance-level features $\boldsymbol{f}^s_{utt}$ and $\boldsymbol{f}^t_{utt}$ of two domains, and UDA loss $\mathcal{L}_u$ is denoted as

$$\mathcal{L}_u = \mathit{MK\text{-}MMD}(\boldsymbol{f}^s_{utt}, \boldsymbol{f}^t_{utt}). \tag{11}$$

Consequently, we can defined the total loss $\mathcal{L}_{\mathit{total}}$ of our proposed HDA model by jointing the emotion classification loss $\mathcal{L}_{ce}$ and three distribution adaptation loss of hierarchical distribution adaptation module as follow

$$\mathcal{L}_{\mathit{total}} = \mathcal{L}_{ce} + \lambda_f \cdot \mathcal{L}_f + \lambda_s \cdot \mathcal{L}_s + \lambda_u \cdot \mathcal{L}_u, \tag{12}$$

where λ_f, λ_s and λ_u are penalty coefficients.

3. Experiments

3.1. Experimental Database and setting

To evaluate the performance of the proposed HDA, we select three public speech emotion databases for experiments, i. e., *Emo-DB* [19], *eNTERFACE'05* [20], and *CASIA* [21]. **eNTERFACE** [20] is a English multi-modal emotion dataset, containing 1290 audio-visual samples with the sample rate 48 kHz. In this dataset, six emotions (i. e., *anger, disgust, fear, happiness, sadness,* and *surprise*) are induced by the pre-prepared performance contents. 43 volunteers coming from different countries participated in the recording of the dataset. **CASIA** [21] includes 7200 emotional speech sentences in Chinese language. Each sample is recorded with six emotions, i. e., *anger, fear, happiness, neutral, sadness,* and *surprise*, through some acting contents from four actors containing two males and two females. Note that we utilize 1200 public speech samples for the experiments. **Emo-DB** [19] is collected as a German emotional speech dataset with 535 speech samples by ten native speakers, including five males and five females. In Emo-DB, each sentence is recorded with 16 kHz under seven emotions, i. e., *anger, boredom, disgust, fear, happiness, neutral,* and *sadness*.

To perform the cross-corpus SER, we pick common emotion categories inside two datasets for the six cross-corpus task (source $\rightarrow$ target). And we use **e**, **c**, and **b** represent *eNTERFACE*, *CASIA*, and *Emo-DB*, respectively.

3.2. Comparison Methods

Subspace learning methods: Transfer Component Analysis (TCA) [22], Domain-adaptive Subspace Learning (DoSL) [6], Joint Distribution Adaptive Regression (JDAR) [23], Joint Distribution Implicitly Aligned Subspace Learning (JIASL) [7].
Deep learning methods: Deep Adaptation Network (DAN) [5], Deep Subdomain Adaptation Network (DSAN) [24], Deep Implicit Distribution Alignment Network (DIDAN) [25], Deep Transductive Transfer Regression Network (DTTRN) [26].

3.3. Results and Analysis

The experimental results of six cross-corpus SER tasks are reported in Table 1 with Weighted Average Recall (WAR, i. e., accuracy) and Unweighted Average Recall (UAR, class-wise recall added and divided by number of classes to compensate for class-imbalances). The comparison results reveal that our proposed HDA achieves the best performance versus the other SOTA methods. In detail, the DA-based methods are superior to the baseline methods for all six tasks of cross-corpus SER on the average accuracies. For each task, the DA-based methods also surpass the performance of most tasks. Considerably, the discrepancy-based methods (e. g., DTTRN), achieve comparable recognition rate with the adversarial-based method (e. g., DAN and DSAN), demonstrating that the distribution alignment strategy, either distance measurement or adversarial training, can promote corpus-invariant emotion features. Furthermore, our proposed method is beyond the mentioned DA-based methods. The reason is that our proposed HDA adapts hierarchial distribution adaptation to maintains the speech representation that is discriminative to emotions and robust across domains.

For the results in the Table 1, we can also observe that the tasks of **b** $\rightarrow$ **e**, **e** $\rightarrow$ **c**, and **c** $\rightarrow$ **e** have worse performances than other tasks (i. e., **e** $\rightarrow$ **b**, **b** $\rightarrow$ **c**, and **c** $\rightarrow$ **b**). This situation indicates that variations in training and test datasets may affect the generalization performance of all cross-corpus methods. In addition, it is also interesting to find that the actuaries of

Table 1: *WAR/UAR(%) results for cross-corpus SER tasks (source → target) on Emo-DB ("**b**"), eNTERFACE ("**e**"), and CASIA ("**c**").*

Method		**b** → **e**	**e** → **b**	**b** → **c**	**c** → **b**	**c** → **e**	**e** → **c**	Average
Subspace Learning	TCA[22]	30.51/30.52	45.07/44.03	33.40/33.40	42.65/45.07	32.32/32.32	31.10/31.10	35.84/36.07
	DoSL[6]	33.56/33.50	40.53/43.89	35.80/35.80	45.10/49.03	28.04/28.17	32.60/32.60	35.94/36.33
	JDAR[23]	36.41/36.33	40.27/39.97	31.10/31.10	43.63/46.29	31.56/31.50	32.40/32.40	35.90/36.27
	JIASL[7]	36.88/36.87	50.40/44.11	36.50/36.50	53.68/49.30	33.17/33.19	30.50/30.50	40.19/38.42
Deep Learning	DAN[5]	36.12/36.13	49.82/40.41	39.00/39.00	50.98/49.85	31.46/31.47	29.00/29.00	39.89/37.64
	DSAN[24]	36.29/36.25	52.16/46.90	40.30/40.30	51.81/50.69	32.61/32.61	29.70/29.70	40.47/39.41
	DIDAN[25]	33.12/33.05	50.13/47.11	38.90/38.90	58.58/56.22	34.14/34.06	31.10/31.10	41.00/40.07
	DTTRN[26]	**38.53/38.54**	52.27/47.46	39.10/39.10	**60.54/58.80**	31.30/31.30	36.85/36.83	43.10/42.01
	HDA (ours)	36.38/36.34	**57.07/52.70**	**39.40/39.40**	54.41/49.87	**37.50/37.46**	**39.90/39.90**	**44.11/42.61**

b → **e** are less than **e** → **b**, which may be because the database of Emo-DB is small such that it cannot support to obtain sufficiently robust speech emotion features. Furthermore, neither **c** → **e** nor **e** → **c** perform promisingly. This is very likely because the CASIA and eNTERFACE databases are based on different languages, in which CASIA is a Chinese dataset and eNTERFACE is an English one. The disparities across languages leads to the emotion variations in speech, which is also a research hotspot in the field of SER.

3.4. Ablation Study

We further evaluate the effectiveness of our proposed components. The ablation results on **e** → **b** is illustrated in Table 2, where backbone network serve as a baseline and "w/o" represents HDA without the corresponding module.

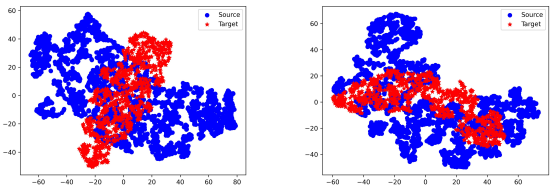
From the results in Table 2, we can easily draw a conclusion that the proposed UDA is the most crucial component since it greatly improves model performance compared to the baseline. It also reveal that the utterance-level information is effective in eliminating domain gap. Besides, the results in both WAR and UAR without SDA module performance worse than those without FDA, indicating that segment-level information is more important than frame-level information on the subtask **e** → **b**. Overall, the proposed HDA performs the best among all results and each component contributes to the final performance.

Table 2: *Ablation results of HDA on the subtask **e** → **b**.*

Ablation Components	Accuracy (%)	
	WAR	UAR
Backbone Network (BN)	35.73	28.05
HDA w/o SDA	40.53	33.99
HDA w/o UDA	37.33	28.39
HDA w/o FDA	52.80	46.63
HDA	**57.07**	**52.70**

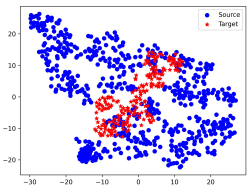
3.5. Discussions on Hierarchical Distribution Adaptation

To further evaluate our proposed HDA on cross-corpus SER, we visualize the feature distributions of source and target domains after incorporating HDA. For illustration purposes, herein, we take the subtask **e** → **b** as an example. From Fig. 2(a) and (b), we can observe that, with the inclusion of HDA, the data marginal distribution of the source and target domains (marked with red and blue) in Fig. 2(b) are closer in shape compared to Fig. 2(a), indicating a higher level domain confusion. Similarly, the results under class-wise conditional distributions in Fig. 2(c) and (d) (taking the data distribution of the *sad* emotion as an example) also demonstrate that CDA with HDA achieves a higher degree of overlap between the source and target domains. The visualizations in Fig. 2 indicate that our HDA effectively eliminates the distribution shifts between domains for learning domain-invariant and emotional discriminative speech features.

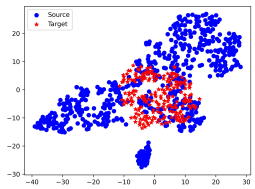


(a) Marginal distribution w/o HDA on **e** → **b**

(b) Marginal distribution w/ HDA on **e** → **b**



(c) Conditional distribution ('*sad*') w/o HDA on **e** → **b**



(d) Conditional distribution ('*sad*') w/ HDA on **e** → **b**

Figure 2: *Visualization of distribution adaptation based on our proposed HDA on the subtask **e** → **b**.*

4. Conclusion

In this paper, we introduced a straightforward and effective strategy, i. e., Hierarchical Distribution Adaptation (HDA), to address cross-corpus speech emotion recognition (SER). HDA leverages a hierarchical emotion representation module based on nested Transformers to extract speech emotion features at different scales (e. g., frame/segment/utterance-level), capturing multiple-level emotion correlations. Additionally, a hierarchical distribution adaptation module is designed to align the emotion representations of training and testing data at hierarchical levels for effectively eliminating domain shifts, including distribution adaptation on frame-level (FDA), segment-level (SDA), and utterance-level (UDA). Extensive experiments demonstrate that our proposed HDA outperforms the comparison methods. However, due to the high coupling between emotion and acous-

tic features, an appropriate strategy to measure cross-domain shifts in emotion space remains a challenge in future.

5. Acknowledgements

This work was supported in part by the National Key R&D Project under the Grant 2022YFC2405600, in part by the NSFC under the Grant U2003207 and 61921004, in part by the Jiangsu Frontier Technology Basic Research Project under the Grant BK20192004, in part by the YESS Program by CAST (Grant No. 2023QNRC001) and JSAST (Grant No. JSTJ-2023-XH033), in part by the ASFC under the Grant 2023Z071069003, in part by the China Postdoctoral Science Foundation under the Grant 2023M740600, and in part by the Jiangsu Province Excellent Postdoctoral Program.

6. References

[1] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] C. Tang, W. Zheng, Y. Zong, N. Qiu, C. Lu, X. Zhang, X. Ke, and C. Guan, "Automatic identification of high-risk autism spectrum disorder: a feasibility study using video and audio data under the still-face paradigm," *ieee transactions on neural systems and rehabilitation engineering*, vol. 28, no. 11, pp. 2401–2410, 2020.

[4] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, 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.

[5] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 12, pp. 3071–3085, 2018.

[6] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5144–5148.

[7] C. Lu, Y. Zong, C. Tang, H. Lian, H. Chang, J. Zhu, S. Li, and Y. Zhao, "Implicitly aligning joint distributions for cross-corpus speech emotion recognition," *Electronics*, vol. 11, no. 17, p. 2745, 2022.

[8] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE transactions on affective computing*, vol. 10, no. 2, pp. 265–275, 2017.

[9] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 307–318, 2019.

[10] Y. Zong, H. Lian, J. Zhang, E. Feng, C. Lu, H. Chang, and C. Tang, "Progressive distribution adapted neural networks for cross-corpus speech emotion recognition," *Frontiers in Neurorobotics*, vol. 16, p. 987146, 2022.

[11] C. Lu, C. Tang, J. Zhang, and Y. Zong, "Progressively discriminative transfer network for cross-corpus speech emotion recognition," *Entropy*, vol. 24, no. 8, p. 1046, 2022.

[12] S. Zhang, R. Liu, X. Tao, and X. Zhao, "Deep cross-corpus speech emotion recognition: Recent advances and perspectives," *Frontiers in neurorobotics*, vol. 15, p. 784514, 2021.

[13] C. Lu, H. Lian, W. Zheng, Y. Zong, Y. Zhao, and S. Li, "Learning local to global feature aggregation for speech emotion recognition," *arXiv preprint arXiv:2306.01491*, 2023.

[14] Y. Wang, C. Lu, Y. Zong, H. Lian, Y. Zhao, and S. Li, "Time-frequency transformer: A novel time frequency joint learning method for speech emotion recognition," *arXiv preprint arXiv:2308.14568*, 2023.

[15] C. Lu, Y. Zong, W. Zheng, Y. Li, C. Tang, and B. W. Schuller, "Domain invariant feature learning for speaker-independent speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2217–2230, 2022.

[16] C. Lu, W. Zheng, H. Lian, Y. Zong, C. Tang, S. Li, and Y. Zhao, "Speech emotion recognition via an attentive time–frequency neural network," *IEEE Transactions on Computational Social Systems*, 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

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

[20] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd international conference on data engineering workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[21] J. Zhang and H. Jia, "Design of speech corpus for mandarin text to speech," in *The blizzard challenge 2008 workshop*, 2008.

[22] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE transactions on neural networks*, vol. 22, no. 2, pp. 199–210, 2010.

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

[24] 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 transactions on neural networks and learning systems*, vol. 32, no. 4, pp. 1713–1722, 2020.

[25] Y. Zhao, J. Wang, Y. Zong, W. Zheng, H. Lian, and L. Zhao, "Deep implicit distribution alignment networks for cross-corpus speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[26] Y. Zhao, J. Wang, R. Ye, Y. Zong, W. Zheng, and L. Zhao, "Deep transductive transfer regression network for cross-corpus speech emotion recognition," 2022.