ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747129

DOMAIN-INVARIANT FEATURE LEARNING FOR CROSS CORPUS SPEECH EMOTION RECOGNITION

Yuan Gao[1,2]*, Shogo Okada*[2,*]*, Longbiao Wang*[1,*]*, Jiaxing Liu*[1]*, Jianwu Dang*[1,2]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{yuan_gao, longbiao_wang, jiaxingliu}@tju.edu.cn, {okada-s, jdang}@jaist.ac.jp

ABSTRACT

To deal with speech emotion recognition (SER) in real-life applications, researchers have to focus on cross corpus SER, where the feature distribution of source and target datasets are different. In this paper, we propose an efficient domain adversarial training method to cope with the non-affective information during feature extraction. Through the proposed domain-adversarial learning, we can reduce the domain divergence between train and test data. Furthermore, we incorporate center loss with the emotion classifier to reduce the intra-class variation of features learned from the same emotion. We conduct experiments on four emotional benchmark datasets to verify the performance of the proposed method. The experimental results demonstrate that our proposed model outperform the baseline system in both cross-corpus and multi-corpus evaluation.

Index Terms— Speech emotion recognition, domain adaptation, center loss

1. INTRODUCTION

Emotion recognition from speech signals is crucial for the development of artificial intelligence, and advanced SER systems can be applied in a wide range of applications, such as building empathetic chatbots which enable natural human-computer interaction and helping the manual service of call centers [1]. Therefore, this research topic has drawn growing attention in both industrial and academic communities. Previous studies empirically designed low-level descriptors (LLDs) for emotion classification. In recent years, some researchers presented that deep learning based models such as convolutional neural network (CNN) and recurrent neural network (RNN) show promising results on emotion recognition tasks [2, 3]. Compared with traditional hand crafted features, deep representation features can improve the performance of SER systems without expert knowledge.

Despite the recent developments in this field, most previous approaches are trained and tested on the same dataset [4, 5]. Since collecting large-scale annotated emotional utterances in a natural environment is time-consuming, the existing datasets contain a small number of speech samples, which is not enough to train robust deep learning models. Moreover, in real-world scenarios, the emotional information in speech is difficult to learn due to the variations in the domain information. Therefore, recognition performance usually decreases significantly when the system is applied to unseen datasets [6]. To deal with real-life applications, researchers have to evaluate the model using different datasets to validate the robustness of the emotion recognition systems [7]. In the initial work of cross corpus SER task, Schuller et al. [8] explored the feature selection strategy and defined the emotion annotations using six existing datasets. They also investigated several normalization methods to improve recognition performance. In [9], Hassan et al. try to solve the feature distribution mismatch between train and test data by importance-weighted support vector machine (IW-SVM). To generalize the model to unseen languages, Albornoz et al. [10] focus on decision level fusion to achieve better recognition accuracy of the SVM classifier. Their system can improve performance in real-life applications, with no available data from the target language to train the model. More recently, some researchers have also evaluated the performance of CNN, RNN, and attention in cross corpus SER [11, 12].

To further improve the generalization ability of emotion recognition system, we use the adversarial domain adaptation method to reduce the domain divergence between the training and test data. To be more specific, we incorporate adversarial training to eliminate the speaker, corpus, and other domain information of the latent representation. The domain adaptation is achieved by reversing the gradient between the feature extractor and the domain classifier; by doing this, our model can maximize the training loss of non-affective information. Moreover, in previous works, commonly used emotion classifiers use the softmax loss function to find a decision boundary and separate different emotions. To improve the discrim-

* Corresponding Authors

inability of feature representation, in this work, we incorporate center loss, which is trained to minimize the distances between the feature representations and their corresponding class centers, as joint supervision for the feature extractor.

The rest of this paper is organized as follows. We introduce the proposed approach in Section 2. In Section 3, we present the emotional datasets used in this study and describe the experimental setup. Section 4 analyses the results of comparative experiments. This paper is concluded in Section 5.

2. ADVERSARIAL DOMAIN ADAPTATION FOR FEATURE EXTRACTION

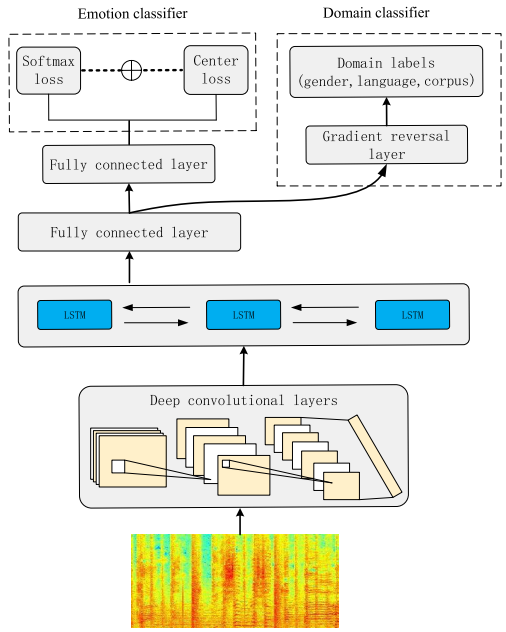

Fig. 1. Overview of the proposed method.

As shown in Figure 1, we use deep CNN and a BLSTM layer for feature extraction, whose parameters are similar to that used by Satt et al. [13]. We modified the feature extractor with a domain adversarial neural network (DANN). Furthermore, we use the center loss to reduce the intra-class variation of feature representation. Both DANN and center loss can address the domain divergence, and we describe the model details below.

2.1. Domain Adversarial Training

As the features extracted from different datasets contain speaker, recording condition, and other domain information, common deep learning based models show poor performance in cross-corpus tasks. In this study, we incorporate DANN with a feature extractor to eliminate the non-affective information. DANN is defined as a multi-task learning model, and the recognition targets of DANN are emotion classifier L_E and domain classifier L_D. In this work, the domain recognition targets of L_D are corpus, language, and gender. In order to achieve domain adaptation and feature representation learning within one training process, Ganin et al. [14] introduced a gradient reversal layer (GRL) between the domain classifier and the feature extractor. During backpropagation, the GRL can multiply a certain negative constant γ to the gradient of the domain classification task, and the DANN was trained to make the feature distribution learned from source and target domain indistinguishable to our model. Through the GRL, we can extract domain invariant representation and thus improve the generalization ability for cross-corpus emotion recognition. The objective function of our proposed feature extraction model is defined as:

$$L = L_E(G(x,\theta), y) + \gamma L_D \tag{1}$$

Where L_E is the loss function of the emotion classifier, which combines center loss and softmax loss, more detail of L_E can be found in section 2.2. In this specific task, we set γ as -0.3 to avoid our feature extractor $G(x,\theta)$ from learning the aforementioned non-affective information. Through this DANN, our model can eliminate the domain shift of feature distribution learned from the source and target datasets. The loss function of the domain classifier is defined as:

$$L_D = L_g(G(x,\theta), g) + L_l(G(x,\theta), l) + L_c(G(x,\theta), c) \tag{2}$$

Where L_g, L_l, and L_c are the loss functions of the gender, language, and corpus classification tasks in DANN. By finding a saddle point that minimizes the L_E and maximizes the L_D, our proposed feature extractor can significantly reduce the domain divergence in the input of the emotion classifier.

2.2. Center Loss

In addition to the proposed feature extractor, we further incorporate the softmax loss and center loss [15] as joint supervision to the emotion classifier L_E. The softmax loss function is commonly used in the emotion recognition system for finding a decision boundary in different emotions.

$$Softmax(G(x,\theta), y) = -\sum_{i=1}^{M} \frac{e^{(x,\theta)_i^T}}{\sum_{j=1}^{N} e^{(x,\theta)_j^T}} \tag{3}$$

Where M is the size of mini-batch and N is the number of emotion classes. In this study, although we defined the same emotion annotations for train and test samples, the feature distribution of different datasets shows no separable clusters, which makes cross corpus SER more difficult than common close-set identification tasks. To deal with this problem, we introduce center loss to learn a class center c for each emotion category and thus reduces the intra-class distance of feature distribution. This loss function is calculated as the

Table 1. Overview of the four emotion corpora. The emotion labels are: angry (A), boredom (B), disgust (D), fear (F), happy (H), joy (J), neutral (N), sad (SA), surprise (SU).

Corpus	Language	#m	#f	Type	Utterances	Valence		Arousal	
						negative	positive	low	high
IEMOCAP	English	5	5	Hybrid	5531	A,SA:3344	H,N:2187	N,SA:2792	A,H:2739
MSP-Improv	English	6	6	Acted	8438	A,SA:4546	H,N:3892	N,SA:3660	A,H:4778
SAVEE	English	4	0	Acted	480	A,D,F,SA:240	J,N,SU:240	D,N,SA:240	A,F,J,SU:240
Emo-DB	German	5	5	Acted	535	A,B,D,F,SA:385	H,N:150	B,D,N,SA:268	A,F,H:267

Euclidean distance between the input feature and the corresponding class center.

$$Center(x, c) = \frac{1}{M}\sum_{i=1}^{N} ||x_i - c_i||^2 \quad (4)$$

$$c_j = \frac{\sum_{i=1}^{m}(c_j - x_i)}{1 + m} \quad (5)$$

To update the class center c_i more efficiently, this loss function is trained on each mini-batch. In Equation 5, m is the number of samples of class i in the new mini-batch. The overall objective function of emotion classifier is defined as:

$$L_E(G(x,\theta), y) = \lambda Softmax(G(x,\theta), y) + (1-\lambda)Center(G(x,\theta), c) \quad (6)$$

We set λ as 0.5 to control the weight of each loss term. By combining the center loss with the softmax loss to jointly optimize our model, we can extract a more robust feature representation for the cross corpus SER task.

3. EXPERIMENTAL SETUP

3.1. Emotional Speech Datasets

Four emotional corpora are used to evaluate our model. We choose these four datasets because they are available to the community and have different feature distributions to meet real-life scenarios. The Interactive Emotional Dyadic Motion Capture (**IEMOCAP**) database [16] contains 12 hours of audiovisual data, including audio, video, and facial motion information, and textual transcriptions from 10 speakers. We use 5531 utterances from both scripted and improvised data for our experiments. The emotional states recorded are: Happiness, Sadness, Anger, and Neutral. The **MSP-IMPROV** database [17] is a multimodal emotional database recorded from actors interacting in dyadic sessions. The corpus consists of 8,438 utterances of emotional sentences recorded from 12 actors. The emotion categories in this dataset are also Happiness, Sadness, Anger, and Neutral. The Surrey audio-visual expressed emotion (**SAVEE**) dataset [18] contains audio-visual recordings of four male subjects. This dataset includes 480 native English utterances: 60 for each of six basic emotions (Happiness, Sadness, Disgust, Anger, Boredom, and Fear) and 120 utterances for Neutral. And the Berlin Emotional Speech Database (**Emo-DB**) [19] was performed by ten professional actors in a recording environment. The actors were asked to express each sentence in seven emotional states (Neutral, Bordom, Disgust, Sadness, Anger, Happiness, Fear). This corpus contains a total of 535 utterances.

As the emotional annotations of those datasets are different, in this work, we defined the emotion recognition task as binary classification of arousal and valence. We followed Schuller et al. [8] to map the categorical labels of each emotion into binary arousal and valence. The main attributes of each dataset are summarized in Table 1.

3.2. Experimental Settings

We use two validation schemes to evaluate our model: 1) $cross\text{-}corpus$ evaluation: the model is trained only on the IEMOCAP and tested on the other three corpora. 2) $multi\text{-}corpus$ evaluation: we split all four datasets into train set (80%) and test set (20%), and evaluate our model using test data from each corpus respectively. Note that there is no speaker-overlap between train and test data. We used adadelta as the optimizer and the mini-batch size was set as 128. During data-preprocessing, all the datasets are downsampled to 16 kHz. We use spectrogram as the input feature. The input utterances are split into 265-ms segments, and the input spectrogram was calculated for each segment, with a frame size of 25-ms. The time × frequency of the input spectrogram was 32 × 129.

4. RESULTS AND ANALYSIS

We choose unweighted accuracy (UA) as the evaluation criteria. The baseline is the combination of CNN and BLSTM. In this work, we compare two proposed DANN-based approaches: (1) in DANN_1, the recognition targets of domain classifier are speaker and corpus; (2) in DANN_2, the speaker classification is replaced by language and gender recognition. We use S and C to represent the softmax and center loss.

Table 2. Experiment results of multi-corpus evaluation. The domain recognition targets are speaker and corpus for DANN_1; and gender, language, and corpus for DANN_2. The loss functions are softmax (S) and center loss (C).

Model	Loss	Arousal (UA)					Valence (UA)				
		IEMOCAP	MSP	SAVEE	Emodb	Average	IEMOCAP	MSP	SAVEE	Emodb	Average
Baseline	S	75.90	62.05	85.83	90.95	78.68	69.51	59.44	64.16	53.03	61.54
DANN_1	S	**79.47**	63.15	82.55	90.48	78.91	70.32	60.40	63.83	56.47	62.76
DANN_2	S	77.62	62.74	**86.67**	88.01	78.76	69.44	61.15	65.33	58.85	63.69
DANN_2	S+C	78.59	**64.53**	84.16	**92.12**	**79.85**	**75.56**	**61.33**	**67.50**	**63.48**	**66.97**

Table 3. Arousal recognition in cross-corpus evaluation.

Model	Loss	MSP	SAVEE	Emodb	Average
Baseline	S	59.72	73.75	67.35	66.94
DANN_1	S	60.51	74.58	66.05	67.05
DANN_2	S	**63.57**	74.79	69.58	69.31
DANN_2	S+C	62.97	**75.20**	**71.64**	**69.94**

Table 4. Valence recognition in cross-corpus evaluation.

Model	Loss	MSP	SAVEE	Emodb	Average
Baseline	S	**59.52**	54.79	**49.73**	54.68
DANN_1	S	59.19	56.15	47.53	54.29
DANN_2	S	57.57	57.08	49.05	54.57
DANN_2	S+C	57.26	**58.12**	49.46	**54.95**

4.1. Multi-corpus Evaluation

We present the results of $multi\text{-}corpus$ evaluation in Table 2. For arousal recognition, our proposed DANN_2 has achieved the best performance for all four datasets, with a small but steady improvement than the comparative experiments. For valence, most comparative experiments show poor performance in Emodb. The train set of Emodb mainly consists of negative inputs. Furthermore, due to the language mismatch of Emodb and other datasets, the recognition performance of this dataset is relatively low. Despite this situation, our model shows relatively equal recognition accuracy on positive and negative, and thus improves the UA by 10.45%. Moreover, the proposed center loss helps our model to extract more discriminative feature representation and improve the average accuracy by 3.28%. The results show that our model can generalize emotion information across datasets.

4.2. Cross-corpus Evaluation

The experimental results of $cross\text{-}corpus$ evaluation demonstrated the effectiveness of our proposed model when dealing with unseen datasets. As shown in Table 3 and 4, the average performance of DANN based model show significant improvement over the CNN-BLSTM baseline in arousal recognition. Furthermore, due to the large number of speakers in these four datasets (which is also common in real-life scenarios), speaker recognition is difficult to achieve high accuracy in this work; thus, the DANN_2 generates better average performance than DANN_1. However, for the valence recognition, both DANN and baseline show relatively poor performance (below 60%). For valence recognition of Emo-DB, due to the language mismatch, the recognition performance of all four comparative experiments is below the chance level. These results indicate that domain invariant feature learning is more challenging to achieve for valence, which has also been reported in [20].

5. CONCLUSION

In this paper, we investigated adversarial domain adaption and center loss for increasing the generalization ability of cross-corpus SER systems. As a step towards domain invariant feature learning for the SER task, we modified the feature extractor as DANN and have reduced the domain divergence across different datasets. Furthermore, we incorporated center loss and softmax loss to learn discriminative feature representation for emotion recognition. Experimental results indicate that: 1) compared with arousal, the deep learning model is more difficult to generalize valence information to unseen datasets. 2) the proposed model achieves more promising average results than the traditional deep learning-based model, which demonstrates the effectiveness of our proposed approach.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 62176182, the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers 19H01120, 19H01719, and JST AIP Trilateral AI Research, Grant Number JPMJCR20G6, Japan.

7. REFERENCES

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

[2] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[3] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.

[4] Jianfeng Zhao, Xia Mao, and Lijiang Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[5] Jiaxing Liu, Zhilei Liu, Longbiao Wang, Lili Guo, and Jianwu Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7174–7178.

[6] Mohammed Abdelwahab and Carlos Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[7] Michael Neumann and Ngoc Thang Vu, "Cross-lingual and multilingual speech emotion recognition on english and french," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5769–5773.

[8] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[9] Ali Hassan, Robert Damper, and Mahesan 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.

[10] Enrique Marcelo Albornoz and Diego H Milone, "Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 43–53, 2015.

[11] Jack Parry, Dimitri Palaz, Georgia Clarke, Pauline Lecomte, Rebecca Mead, Michael Berger, and Gregor Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition.," in *INTERSPEECH*, 2019, pp. 1656–1660.

[12] Rosanna Milner, Md Asif Jalal, Raymond WM Ng, and Thomas Hain, "A cross-corpus study on speech emotion recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 304–311.

[13] Aharon Satt, Shai Rozenberg, and Ron Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms.," in *Interspeech*, 2017, pp. 1089–1093.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[15] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[16] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[17] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[18] P Jackson and S Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[19] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech," in *Ninth european conference on speech communication and technology*, 2005.

[20] Biqiao Zhang, Emily Mower Provost, and Georg Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 85–99, 2017.