TEMPORAL MODELING MATTERS: A NOVEL TEMPORAL EMOTIONAL MODELING APPROACH FOR SPEECH EMOTION RECOGNITION —SUPPLEMENTARY MATERIALS—

Jiaxin Ye[1], *Xincheng Wen*[2], *Yujie Wei*[1], *Yong Xu*[3], *Kunhong Liu*[4,†], *Hongming Shan*[1,†]

[1] Fudan University [2] Harbin Institute of Technology (Shenzhen)
[3] Fujian University of Technology [4] Xiamen University

1. DETAILS ABOUT DATASETS AND EVALUATION

1.1. Datasets

To sufficiently compare the performance of our methods with State-Of-The-Art (SOTA) methods, we conduct our experiments on the 6 benchmark SER corpora, including Chinese corpus CASIA [1], German corpus EMODB [2], Italian corpus EMOVO [3], English corpora IEMOCAP [4], RAVDESS [5], and SAVEE [6], as follows.

1. **CASIA** is collected from four Chinese native speakers (2 males and 2 females) in 6 emotions, *i.e. angry*, *fear*, *happy*, *neutral*, *sad* and *surprise*;

2. **EMODB** contains 10 sentences that cover 7 emotions (without *surprise* but adds *boredom* and *disgust* compared with CASIA) from daily communication by 10 German speakers (5 males and 5 females). They could be interpreted in all emotional contexts without semantic inconsistency;

3. **EMOVO** is recorded by 6 Italian speakers (3 males and 3 females) who played 14 sentences simulating 7 emotional (adds *disgust* compared with CASIA) states;

4. **IEMOCAP** comprises 12 hours of emotional speech performed by 10 American speakers (5 males and 5 females). We select 4 emotions *angry*, *happy*, *neutral* and *sad* and full type of dataset as same as previous studies [7, 8];

5. **RAVDESS** consists of 1440 utterances of 8 emotions (adds *disgust* compared with EMOVO) by 24 British speakers (12 males and 12 females);

6. **SAVEE** contains 480 British English utterances by 4 British speakers (4 males) in 7 emotions (the same as EMOVO).

The detailed information and sample statistics of each dataset are summarized in Tables S1 and S2, respectively.

1.2. Evaluation Metrics

Due to the class imbalance, we use two widely-used metrics, Weighted Average Recall (WAR) (*i.e.* accuracy) and Unweighted Average Recall (UAR) [9], to evaluate the performance of the single- and cross-corpus SER tasks. WAR uses the class probabilities to balance the recall metric of different classes while UAR treats each class equally. The two metrics are formally defined as follows:

$$\mathrm{WAR} = \sum_{k=1}^{K} \frac{M_k}{N} \times \frac{\sum_{i=1}^{M_k} \mathrm{TP}_{ki}}{\sum_{i=1}^{M_k} (\mathrm{TP}_{ki} + \mathrm{FN}_{ki})}, \tag{S1}$$

$$\mathrm{UAR} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{i=1}^{M_k} \mathrm{TP}_{ki}}{\sum_{i=1}^{M_k} (\mathrm{TP}_{ki} + \mathrm{FN}_{ki})}, \tag{S2}$$

where K, M_k, and N represent the number of classes, the number of speeches in class k, and the number of all speeches in the dataset, respectively. $\mathrm{TP}_{ki}, \mathrm{TN}_{ki}$, and FN_{ki} represent the true positive, true negative, and false negative values of class k for speeches i, respectively.

2. DETAILED RESULTS IN THE SINGLE-CORPUS SER TASKS

2.1. Confusion Matrices

Due to space limits, the confusion matrices of proposed TIM-Net on a single corpus is not given in the main paper. The confusion matrices for the highest WAR obtained by TIM-Net on five datasets are shown from Fig. S1 to Fig. S3. It can be observed that the TIM-Net achieves the excellent results with strong discriminability for almost all emotions.

2.2. Ablation Studies

We conduct ablation studies on all the corpus datasets, including the following variations of TIM-Net: **TCN**: the TIM-Net is replaced with TCN; **w/o BD**: the backward TABs are removed while keeping the forward TABs; **w/o MS**: the

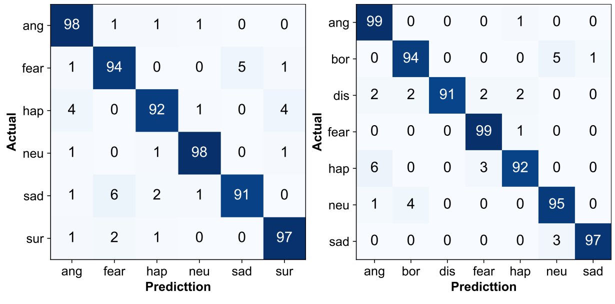
†: Corresponding author.

multi-scale fusion is removed and $\boldsymbol{g}_n$ is used as $\boldsymbol{g}_{\text{drf}}$ corresponding to max-scale receptive field; **w/o DF**: the average fusion is used to confirm the advantages of dynamic fusion. The detailed results over these ablation studies are shown in Table S3. The results show that our method outperforms the ablation methods in most cross-corpus cases, demonstrating that learning emotion-rich representation and corpus-invariant temporal patterns, and enhancing the discriminability of representation on the target corpus are beneficial.

3. DETAILED RESULTS IN THE CROSS-CORPUS SER TASKS

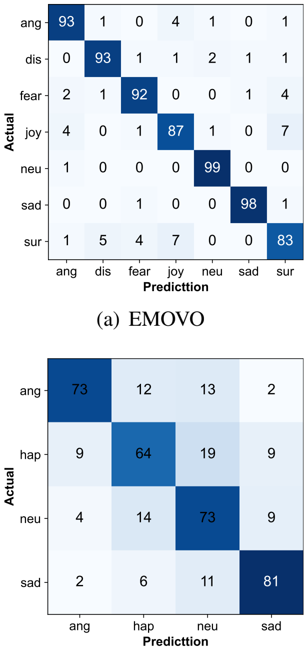
For the cross-corpus task, the source and target corpora have a considerably significant domain shift since these five datasets contain four languages and various speakers. We follow the same experimental setting as CAAM [10] except that TIM-Net does not have access to the target domain. Specifically, we likewise choose 5 emotional classes for a fair comparison, *i.e. angry*, *fear*, *happy*, *neutral*, and *sad*, shared among these 5 corpora (except for IEMOCAP, which has only 4 emotions). These 5 corpora form 20 combinations of cross-corpus; *e.g.* C→B indicates the model is trained with labeled data from dataset CASIA and unlabeled data from dataset EMODB, then tested on dataset EMODB.

In the main paper, Table 2 presents the average performance of 10 runs, each of which consists of 20 cross-corpus cases. We likewise show the best results of these 20 cross-corpus cases of each method in the Table S4. The results show that our method achieves UAR of 37.4% and WAR of 38.6% on average, which are considerably better than the CAAM, which is the task-specific domain adaptation method for the cross-corpus SER tasks. It suggests that our TIM-Net is more generalizable across different corpora.



(a) CASIA (b) EMODB

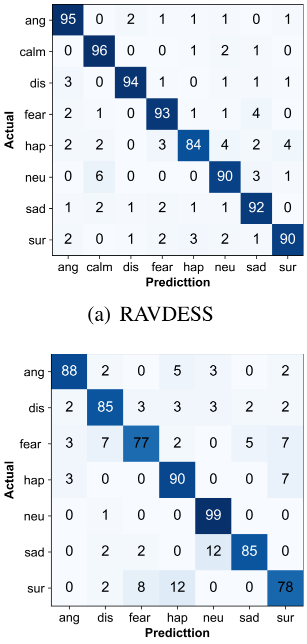
Fig. S1. The confusion matrices of the single-corpus task obtained by TIM-Net on the CASIA and EMODB.



(a) EMOVO

(b) IEMOCAP

Fig. S2. The confusion matrices of the single-corpus task obtained by TIM-Net on the EMOVO and IEMOCAP.



(a) RAVDESS

(b) SAVEE

Fig. S3. The confusion matrices of the single-corpus task obtained by TIM-Net on the RAVDESS and SAVEE.

Table S1. The detailed information of speech emotion datasets. The speakers are presented in the form of [No. of males / No. of females]. The unit for sampling frequency is KHz.

Dataset	Language	Speakers	Numbers	Emotion [No. : Category List]	Frequency
CASIA	Chinese	2/2	1,200	6: Angry, Fear, Happy, Neutral, Sad, Surprise	22.1
EMODB	German	5/5	535	7: Angry, Boredom, Disgust, Fear, Happy, Neutral, Sad	16.0
EMOVO	Italian	3/3	588	7: Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise	48.0
IEMOCAP	English	5/5	5,531	4: Angry, Happy, Neutral, Sad	48.0
RAVDESS	English	12/12	1,440	8: Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, Surprise	48.0
SAVEE	English	4/0	480	7: Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise	44.1

Table S2. The details of data distributions in five datasets.

Dataset	Angry	Boredom	Calm	Disgust	Fear	Happy	Neutral	Sad	Surprise
CASIA	200	-	-	-	200	200	200	200	200
EMODB	127	81	-	46	69	71	79	62	-
EMOVO	84	-	-	84	84	84	84	84	84
IEMOCAP	1103	-	-	-	-	1636	1708	1084	-
RAVDESS	192	-	192	192	192	192	96	192	192
SAVEE	60	-	-	60	60	60	120	60	60

Table S3. The average performance of ablation studies and TIM-Net under 10-fold CV on all six corpora. The 'w/o' means removing the component from TIM-Net.

Dataset	**TCN**	**w/o Bi**	**w/o MS**	**w/o DF**	**TIM-Net**
CASIA	86.92/86.92	91.33/91.33	90.50/90.50	91.83/91.83	**94.67/94.67**
EMODB	89.68/90.09	90.96/91.59	91.80/92.52	91.46/92.34	**95.17/95.70**
EMOVO	82.14/82.14	86.73/86.73	87.24/87.24	86.56/86.56	**92.00/92.00**
IEMOCAP	61.97/59.84	71.07/70.22	70.68/70.04	70.82/69.82	**72.50/71.65**
RAVDESS	83.20/83.75	88.48/88.68	89.97/90.21	87.50/87.99	**91.93/92.08**
SAVEE	78.81/80.63	80.95/83.33	82.50/84.38	80.95/82.92	**86.31/87.71**

Table S4. The overall results of different methods for cross-corpus SER on CASIA (C), EMODB (B), EMOVO (E), RAVDESS (R), and SAVEE (S). The average result is displayed in the last column. Left/Right: UAR(%)/WAR(%). (Bold is the best).

Method	**C → B**	**C → E**	**C → R**	**C → S**	**B → C**	**B → E**	**B → R**	**B → S**	**E → C**	**E → B**	**E → R**	**E → S**	**R → C**	**R → B**	**R → E**	**R → S**	**S → C**	**S → B**	**S → E**	**S → R**	***Average***
TCN	33.8/28.9	19.3/19.3	22.6/24.1	20.2/17.5	24.3/24.3	22.6/22.6	31.8/24.4	32.3/39.4	22.8/22.8	28.5/31.9	26.0/27.3	26.7/26.9	21.1/21.1	21.7/29.7	22.6/22.6	22.7/20.3	21.6/21.6	33.5/29.4	20.0/20.0	23.7/26.3	24.9/25.0
CAAM	**54.6/59.6**	32.1/32.1	29.3/27.0	40.2/**49.4**	39.6/39.6	**37.1/37.1**	28.0/**31.1**	34.3/29.7	39.0/39.0	47.9/49.3	31.0/26.6	27.3/25.8	25.9/25.9	27.0/33.1	**39.3/39.3**	32.2/35.3	33.2/33.2	36.1/36.5	31.0/31.0	24.6/24.4	34.5/35.3
TIM-Net	50.4/43.9	**36.4/36.4**	**30.7/32.4**	**42.0**/49.2	**42.4/42.4**	36.4/36.4	**33.9**/29.8	**35.2/42.5**	**39.8/39.8**	**47.0/53.4**	**33.1/33.2**	**38.0/42.8**	**28.3/28.3**	**32.5/38.2**	37.1/37.1	**37.7/40.8**	**34.0/34.0**	**43.6/37.3**	**36.2/36.2**	**33.2/36.9**	**37.4/38.6**

4. REFERENCES

[1] Jianhua Tao, Fangzhou Liu, Meng Zhang, and Huibin Jia, "Design of speech corpus for mandarin text to speech," in *The Blizzard Challenge 2008 workshop*, 2008.

[2] Felix Burkhardt, Astrid Paeschke, M. Rolfes, et al., "A database of german emotional speech," in *INTERSPEECH 2005, Lisbon, Portugal, September 4-8, 2005*, 2005, vol. 5, pp. 1517–1520.

[3] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco, "Emovo corpus: an italian emotional speech database," in *LREC 2014*, 2014, pp. 3501–3504.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[5] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. e0196391, 2018.

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

[7] Zixuan Peng, Yu Lu, Shengfeng Pan, et al., "Efficient speech emotion recognition using multi-scale CNN and attention," in *ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. 2021, pp. 3020–3024, IEEE.

[8] Arya Aftab, Alireza Morsali, Shahrokh Ghaemmaghami, et al., "LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition," in *ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. 2022, pp. 6912–6916, IEEE.

[9] Björn W. Schuller, Stefan Steidl, and Anton Batliner, "The INTERSPEECH 2009 emotion challenge," in *INTERSPEECH 2009, Brighton, United Kingdom, September 6-10, 2009*. 2009, pp. 312–315, ISCA.

[10] XinCheng Wen, JiaXin Ye, Yan Luo, et al., "CTL-MTNet: A novel capsnet and transfer learning-based mixed task net for single-corpus and cross-corpus speech emotion recognition," in *IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 2022, pp. 2305–2311.