ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096370

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

Jiaxin Ye[1], *Xin-cheng Wen*[2], *Yujie Wei*[1], *Yong Xu*[3], *Kunhong Liu*[4,†], *Hongming Shan*[1,5,†]

[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
[2]Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[3]School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou, China
[4]School of Film, Xiamen University, Xiamen, China
[5]Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

ABSTRACT

Speech emotion recognition (SER) plays a vital role in improving the interactions between humans and machines by inferring human emotion and affective states from speech signals. Whereas recent works primarily focus on mining spatiotemporal information from hand-crafted features, we explore how to model the temporal patterns of speech emotions from dynamic temporal scales. Towards that goal, we introduce a novel temporal emotional modeling approach for SER, termed **T**emporal-aware b**I**-direction **M**ulti-scale Network (**TIM-Net**), which learns multi-scale contextual affective representations from various time scales. Specifically, TIM-Net first employs temporal-aware blocks to learn temporal affective representation, then integrates complementary information from the past and the future to enrich contextual representations, and finally fuses multiple time scale features for better adaptation to the emotional variation. Extensive experimental results on six benchmark SER datasets demonstrate the superior performance of TIM-Net, gaining 2.34% and 2.61% improvements of the average UAR and WAR over the second-best on each corpus. The source code is available at https://github.com/Jiaxin-Ye/TIM-Net_SER.

***Index Terms*—** Speech emotion recognition, bi-direction, multi-scale, dynamic fusion, temporal modeling

1. INTRODUCTION

Speech emotion recognition (SER) is to automatically recognize human emotion and affective states from speech signals, enabling machines to communicate with humans emotionally [1]. It becomes increasingly important with the development of the human-computer interaction technique.

The key challenge in SER is *how to model emotional representations from speech signals*. Traditional methods [2, 3] focus on the efficient extraction of hand-crafted features, which are fed into conventional machine learning methods, such as Support Vector Machine (SVM). More recent methods based on deep learning techniques aim to learn the class-discriminative features in an end-to-end manner, which employ various architectures such as Convolutional Neural Network (CNN) [4, 5], Recurrent Neural Network (RNN) [6, 7], or the combination of CNN and RNN [8].

In particular, various temporal modeling approaches, such as Long Short-Term Memory (LSTM), Gate Recurrent Unit (GRU), and Temporal Convolution Network (TCN), are widely adopted in SER, aiming to capture dynamic temporal variations of speech signals. For example, Wang *et al*. [7] proposed a dual-level LSTM to harness temporal information from different time-frequency resolutions. Zhong *et al*. [9] used CNN with Bi-GRU and focal loss for learning integrated spatiotemporal features. Rajamani *et al*. [6] presented an attention-based ReLU within GRU to capture long-range interactions among the features. Zhao *et al*. [8] leveraged fully CNN and Bi-LSTM to learn the spatiotemporal features. However, these methods suffer from the following drawbacks: 1) they lack sufficient capacity to capture long-range dependencies for context modeling, where the capture of the context in speech is crucial for SER since human emotions are usually highly context-dependent; and 2) they do not explore the dynamic receptive field of the model, while learning dynamic instead of maximal ones can improve model generalization ability to unknown data or corpus.

To overcome these limitations in SER, we propose a **T**emporal-aware b**I**-direction **M**ulti-scale Network, termed **TIM-Net**, which is a novel temporal emotional modeling approach to learn multi-scale contextual affective representations from various time scales. The contributions are threefold. *First*, we propose a temporal-aware block based on the Dilated Causal Convolution (DC Conv) as a core unit in TIM-Net. The dilated convolution can enlarge and refine the

†: Co-corresponding author. This work was supported in part by National Natural Science Foundation of China (Nos. 62101136 and 61772023), Shanghai Municipal of Science and Technology Project (No. 20JC1419500), Fujian Science and Technology Plan Industry University Research Cooperation Project (No. 2021H6015), and Shanghai Center for Brain Science and Brain-inspired Technology.

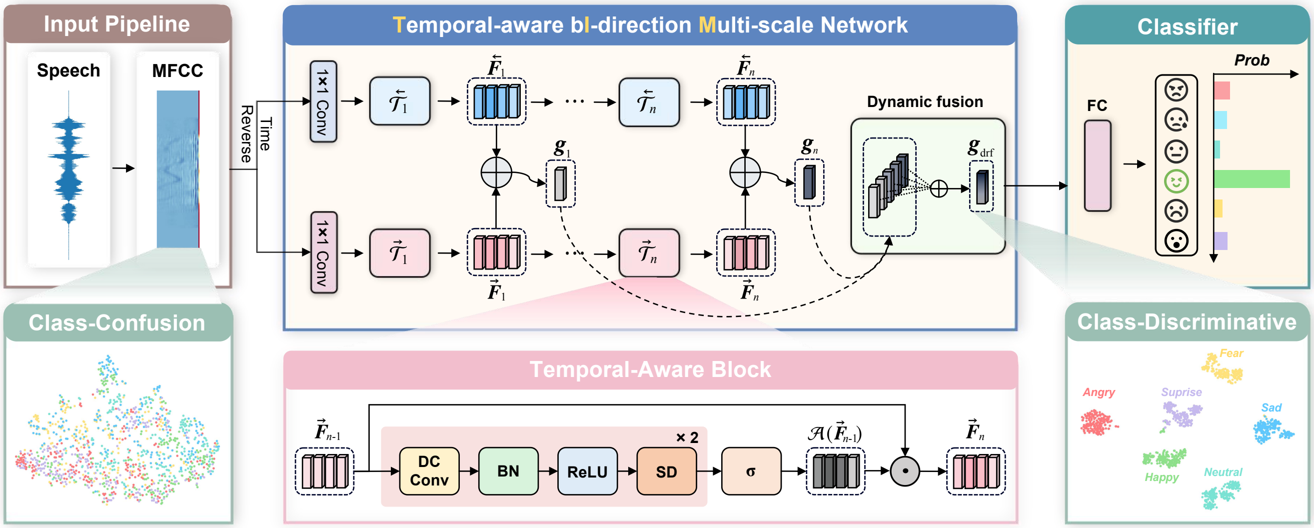

Fig. 1. The framework of the TIM-Net for learning affective features, whose feature extraction part is composed of a bi-direction module and a dynamic fusion module. Note that the forward $\vec{\mathcal{T}}_j$ and backward $\overleftarrow{\mathcal{T}}_j$ are the same structure with different inputs.

receptive field of temporal patterns. The causal convolution combined with dilated convolution can help model relax the assumption of first-order Markov property compared with RNNs [10]. In this way, we can incorporate an N-order (N denotes the number of all previous frames) connection into the network to aggregate information from different temporal locations. *Second*, we devise a novel bi-direction architecture integrating complementary information from the past and the future for modeling long-range temporal dependencies. To the best of our knowledge, TIM-Net is the first bi-direction temporal network by focusing on multi-scale fusion in the SER, rather than simply concatenating forward and backward hidden states. *Third*, we design a dynamic fusion module by combining dynamic receptive fields for learning the inter-dependencies at different temporal scales, so as to improve the model generalizability. Due to the articulation speed and pause time varying significantly across speakers, the speech requires different efficient receptive fields (*i.e.*, the time scale that reflects the affective characteristics) for each low-level feature (*e.g.*, MFCC).

2. PROPOSED METHOD

2.1. Input Pipeline

To illustrate the temporal modeling capacity of our TIM-Net, we use the most commonly-used Mel-Frequency Cepstral Coefficients (MFCCs) features [11] as the inputs to TIM-Net. We first set the sampling rate to the raw sampling rate of each corpus and apply framing operation and Hamming window to each speech signal with 50-ms frame length and 12.5-ms shift. Then, the speech signal undergoes a mel-scale triangular filter bank analysis after performing a 2,048-point fast Fourier transform to each frame. Finally, each frame of the MFCCs is processed by the discrete cosine transformation, where the first 39 coefficients are extracted to obtain the low-frequency envelope and high-frequency details.

2.2. Temporal-aware Bi-direction Multi-scale Network

We propose a novel temporal emotional modeling approach called TIM-Net, which learns long-range emotional dependencies from the forward and backward directions and captures multi-scale features at frame-level. Fig. 1 presents the detailed network architecture of TIM-Net. For learning multi-scale representations with long-range dependencies, the TIM-Net consists of n Temporal-Aware Blocks (TABs) in both forward and backward directions with different temporal receptive fields. Next, we detail each component.

Temporal-aware block. We design the TAB to capture dependencies between different frames and automatically select the affective frames, severing as a core unit of TIM-Net. As shown in Fig. 1, $\mathcal{T}$ denotes a TAB, each of which consists of two sub-blocks and a sigmoid function $\sigma(\cdot)$ to learn temporal attention maps $\mathcal{A}$, so as to produce the temporal-aware feature $\boldsymbol{F}$ by element-wise production of the input and $\mathcal{A}$. For the two identical sub-blocks of the j-th TAB $\mathcal{T}_j$, each sub-block starts by adding a DC Conv with the exponentially increasing dilated rate 2^{j-1} and causal constraint. The dilated convolution enlarges and refines the receptive field and the causal constraint ensures that the future information is not leaked to the past. The DC Conv is then followed by a batch normalization, a ReLU function, and a spatial dropout.

Bi-direction temporal modeling. To integrate complementary information from the past and the future for the judgement of emotion polarity and modeling long-range tempo-

Table 1. The overall results of different SOTA methods on 6 SER benchmark corpora. Evaluation measures are UAR(%) / WAR(%). The '-' implies the lack of this measure, and the best results are highlighted in bold.

Model	Year	CASIA	Model	Year	EMODB	Model	Year	EMOVO
DT-SVM [12]	2019	85.08 / 85.08	TSP+INCA [2]	2021	89.47 / 90.09	RM+CNN [4]	2021	68.93 / 68.93
TLFMRF [13]	2020	85.83 / 85.83	GM-TCN [14]	2022	90.48 / 91.39	SVM [15]	2021	73.30 / 73.30
GM-TCN [14]	2022	90.17 / 90.17	Light-SERNet [16]	2022	94.15 / 94.21	TSP+INCA [2]	2021	79.08 / 79.08
CPAC [17]	2022	92.75 / 92.75	CPAC [17]	2022	94.22 / 94.95	CPAC [17]	2022	85.40 / 85.40
TIM-Net	2023	**94.67 / 94.67**	**TIM-Net**	2023	**95.17 / 95.70**	**TIM-Net**	2023	**92.00 / 92.00**
Model	**Year**	**IEMOCAP**	**Model**	**Year**	**RAVDESS**	**Model**	**Year**	**SAVEE**
MHA+DRN [18]	2019	67.40 / -	INCA+TS-CNN [3]	2021	- / 85.00	3D CNN [19]	2019	- / 81.05
CNN+Bi-GRU [9]	2020	71.72 / 70.39	TSP+INCA [2]	2021	87.43 / 87.43	TSP+INCA [2]	2021	83.38 / 84.79
SPU+MSCNN [11]	2021	68.40 / 66.60	GM-TCN [14]	2022	87.64 / 87.35	CPAC [17]	2022	83.69 / 85.63
Light-SERNet [16]	2022	70.76 / 70.23	CPAC [17]	2022	88.41 / 89.03	GM-TCN [14]	2022	83.88 / 86.02
TIM-Net	2023	**72.50 / 71.65**	**TIM-Net**	2023	**91.93 / 92.08**	**TIM-Net**	2023	**86.07 / 87.71**

ral dependencies, we devise a novel bi-direction architecture based on the multi-scale features as shown in Fig. 1. Formally, for the $\vec{\mathcal{T}}_{j+1}$ in the forward direction with the input $\vec{\boldsymbol{F}}_j$ from previous TAB, the output $\vec{\boldsymbol{F}}_{j+1}$ is given by Eq. (1):

$$\vec{\boldsymbol{F}}_{j+1} = \mathcal{A}(\vec{\boldsymbol{F}}_j) \odot \vec{\boldsymbol{F}}_j, \tag{1}$$

$$\overleftarrow{\boldsymbol{F}}_{j+1} = \mathcal{A}(\overleftarrow{\boldsymbol{F}}_j) \odot \overleftarrow{\boldsymbol{F}}_j, \tag{2}$$

where $\vec{\boldsymbol{F}}_0$ comes from the output of the first 1×1 Conv layer and the backward direction can be defined similarly in Eq. (2).

We then combine bidirectional semantic dependencies and compact global contextual representation at utterance level to perceive context as follows:

$$\boldsymbol{g}_j = \mathcal{G}(\vec{\boldsymbol{F}}_j + \overleftarrow{\boldsymbol{F}}_j), \tag{3}$$

where the global temporal pooling operation $\mathcal{G}$ takes an average over temporal dimension, yielding a representation vector for one specific receptive field from the j-th TAB.

Multi-scale dynamic fusion. Furthermore, since the pronunciation habits (*e.g.*, speed or pause time) vary from speaker to speaker, the utterances have the characteristics of temporal scale variation. SER benefits from taking dynamic temporal receptive fields into consideration. We design the dynamic fusion module to adaptively process speech input at different scales, aiming to determine suitable temporal scale for the current input during the training phase. We adopt a weighted summation operation to fuse the features with Dynamic Receptive Fields (DRF) fusion weights $\boldsymbol{w}_{\text{drf}}$ from different TABs. The DRF fusion is defined as follows:

$$\boldsymbol{g}_{\text{drf}} = \sum_{j=1}^{n} w_j \boldsymbol{g}_j, \tag{4}$$

where $\boldsymbol{w}_{\text{drf}} = [w_1, w_2, \ldots, w_n]^{\mathrm{T}}$ are trainable parameters.

Once the emotional representation $\boldsymbol{w}_{\text{drf}}$ is generated with great discriminability, we can simply use one fully-connected layer with the softmax function for emotion classification.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. To demonstrate the effectiveness of the proposed TIM-Net, we compare TIM-Net with State-Of-The-Art (SOTA) methods on 6 benchmark SER corpora, including Chinese corpus CASIA [20], German corpus EMODB [21], Italian corpus EMOVO [22], English corpora IEMOCAP [23], RAVDESS [24], and SAVEE [25].

Implementation details. In the experiments, 39-dimensional MFCCs are extracted from the Librosa toolbox [26]. The cross-entropy criterion is used as the objective function. Adam algorithm is adopted to optimize the model with an initial learning rate $\alpha = 0.001$, and a batch size of 64. To avoid over-fitting during the training phase, we implement label smoothing with factor 0.1 as a form of regularization. For the j-th TAB $\mathcal{T}_j$, there are 39 kernels of size 2 in Conv layers, the dropout rate is 0.1, and the dilated rate is 2^{j-1}. To guarantee that the maximal receptive field covers the input sequences, we set the number of TAB n in both directions to 10 for IEMOCAP and 8 for others. For fair comparisons with the SOTA approaches, we perform 10-fold cross-validation (CV) as well as previous works [16, 17, 18] in experiments.

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), to evaluate the performance of each method. WAR uses the class probabilities to balance the recall metric of different classes while UAR treats each class equally.

3.2. Results and Analysis

Comparison with SOTA methods. Table 1 presents the overall results on 6 benchmark datasets, showing that our method significantly and consistently outperforms all these

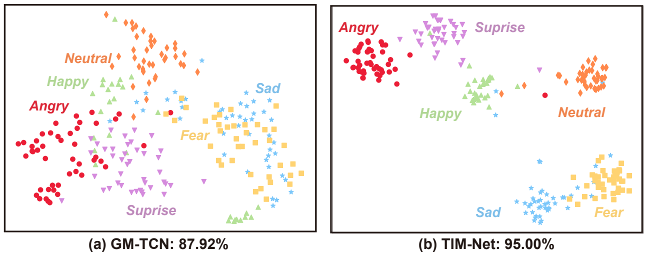

Fig. 2. t-SNE visualizations of features learned from SOTA method GM-TCN and TIM-Net. The score denotes WAR.

Table 2. UAR and WAR on the cross-corpus SER task with different methods. All values are the average ± std of 10 runs, each of which consists of 20 cross-corpus cases.

Method	TCN	CAAM [17]	TIM-Net
$\mathbf{UAR_{avg}}$ **± std**	24.47 ± 0.38	32.37 ± 0.27	**34.49 ± 0.43**
$\mathbf{WAR_{avg}}$ **± std**	24.39 ± 0.42	33.65 ± 0.41	**35.66 ± 0.32**

compared methods by a large margin. Remarkably, our approach gains 2.34% and 2.61% improvements of the average UAR and WAR scores than the second-best on each corpus.

Visualization of learned affective representation. To investigate the impact of TIM-Net on representation learning, we visualize the representations learned by TIM-Net and GM-TCN [14] through the t-SNE technique [27] in Fig. 2. For a fair comparison, we first use the same 8:2 hold-out validation on CASIA corpus for the two methods, and visualize the representations of the same test data after an identical training phase. Although GM-TCN also focuses on multi-scale and temporal modeling. Fig. 2(a) shows heavy overlapping between *Fear* and *Sad* or *Angry* and *Surprise*. In contrast, Fig. 2(b) shows that the different representations are clustered with clear classification boundaries. The results confirm that the TIM-Net provides more class-discriminative representations to support superior performance by capturing intra- and inter-dependencies at different temporal scales.

Domain generalization analysis. Due to various languages and speakers, the SER corpora, although sharing the same emotion, have considerably significant domain shifts. The generalization of the model to unseen domain/corpus is critically important for SER. Inspired by the domain-adaptation study in CAAM [17], we likewise validate the generalizability of TIM-Net on the cross-corpus SER task, following the same experimental setting as CAAM 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 cross-corpus combinations. And we report the average UAR and WAR, and their standard deviation from 10 random runs for each task in Table 2.

Table 3. 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

Method	TCN	w/o BD	w/o MS	w/o DF	TIM-Net
$\mathbf{UAR_{avg}}$	80.45	84.92	85.45	84.85	**88.76**
$\mathbf{WAR_{avg}}$	80.56	85.32	85.82	85.24	**88.97**

The performance of TCN over different corpora is close to random guessing with odds equal to 25%, and TIM-Net has a significant improvement over TCN. Surprisingly, TIM-Net outperforms CAAM, one latest task-specific domain-adaptation method. The results suggest that our TIM-Net is effective in modeling emotion with strong generalizability.

3.3. Ablation Study

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 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 results of ablation studies are shown in Table 3. We have the following observations.

First, all components contribute positively to the overall performance. *Second*, our method achieves 8.31% and 8.41% performance gains in UAR and WAR over TCN that also utilizes DC Conv. Since the inability of TCN to capture contextual multi-scale features, capturing intra- and inter-dependencies at different temporal scales is critical to SER. *Third*, when removing the backward TABs or multi-scale strategy, the results substantially drop due to the weaker capacity to model temporal dependencies and perceive the sentimental features with different scales. *Finally*, TIM-Net without dynamic fusion performs worse than TIM-Net, which verifies the benefits of deploying dynamic fusion to adjust the model adaptively.

4. CONCLUSIONS

In this paper, we propose a novel temporal emotional modeling approach, termed TIM-Net, to learn multi-scale contextual affective representations from various time scales. TIM-Net can capture long-range temporal dependency through bi-direction temporal modeling and fuse multi-scale information dynamically for better adaptation to temporal scale variation. Our experimental results indicate that learning representation from the context information with dynamic temporal scales is crucial for the SER task. The ablation studies, visualizations, and domain generalization analysis further confirm the advantages of TIM-Net.

5. REFERENCES

[1] Björn W Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] Türker Tuncer, Sengül Dogan, and U. Rajendra Acharya, “Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques,” *Knowl. Based Syst.*, vol. 211, pp. 106547, 2021.

[3] Mustaqeem and Soonil Kwon, “Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network,” *Int. J. Intell. Syst.*, vol. 36, no. 9, pp. 5116–5135, 2021.

[4] Ilyas Ozer, “Pseudo-colored rate map representation for speech emotion recognition,” *Biomed. Signal Process. Control.*, vol. 66, pp. 102502, 2021.

[5] Xin-Cheng Wen, Kun-Hong Liu, Wei-Ming Zhang, and Kai Jiang, “The application of Capsule neural network based CNN for speech emotion recognition,” in *in ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. 2020, pp. 9356–9362, IEEE.

[6] Srividya Tirunellai Rajamani, Kumar T. Rajamani, Adria Mallol-Ragolta, Shuo Liu, and Björn W. Schuller, “A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition,” in *ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. 2021, pp. 6294–6298, IEEE.

[7] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh, “Speech emotion recognition with dual-sequence LSTM architecture,” in *ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. 2020, pp. 6474–6478, IEEE.

[8] Ziping Zhao, Yu Zheng, Zixing Zhang, and othersi, “Exploring spatio-temporal representations by integrating attention-based Bi-directional-LSTM-RNNs and FCNs for speech emotion recognition,” in *Interspeech 2018, Hyderabad, India, 2-6 September 2018*. 2018, pp. 272–276, ISCA.

[9] Ying Zhong, Ying Hu, Hao Huang, and Wushour Silamu, “A lightweight model based on separable convolution for speech emotion recognition,” in *Interspeech 2020, Virtual Event, Shanghai, China, 25-29 October 2020*. 2020, pp. 3331–3335, ISCA.

[10] Steffen Jung, Isabel Schlangen, and Alexander Charlish, “A mnemonic Kalman filter for non-linear systems with extensive temporal dependencies,” *IEEE Signal Processing Letters*, vol. 27, pp. 1005–1009, 2020.

[11] Zixuan Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu, “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.

[12] Linhui Sun, Sheng Fu, and Fu Wang, “Decision tree SVM model with Fisher feature selection for speech emotion recognition,” *EURASIP J. Audio Speech Music. Process.*, vol. 2019, pp. 2, 2019.

[13] Luefeng Chen, Wanjuan Su, Yu Feng, Min Wu, Jinhua She, and Kaoru Hirota, “Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction,” *Inf. Sci.*, vol. 509, pp. 150–163, 2020.

[14] Jiaxin Ye, Xin-Cheng Wen, Xuan-Ze Wang, Yong Xu, Yan Luo, Chang-Li Wu, Li-Yan Chen, and Kunhong Liu, “GM-TCNet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition,” *Speech Commun.*, vol. 145, pp. 21–35, 2022.

[15] J Ancilin and A Milton, “Improved speech emotion recognition with Mel frequency magnitude coefficient,” *Applied Acoustics*, vol. 179, pp. 108046, 2021.

[16] 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.

[17] Xin-Cheng Wen, Jiaxin Ye, Yan Luo, Yong Xu, Xuan-Ze Wang, Chang-Li Wu, and Kun-Hong Liu, “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.

[18] Runnan Li, Zhiyong Wu, Jia Jia, et al., “Dilated residual network with multi-head self-attention for speech emotion recognition,” in *ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. 2019, pp. 6675–6679, IEEE.

[19] Noushin Hajarolasvadi and Hasan Demirel, “3D CNN-based speech emotion recognition using K-means clustering and spectrograms,” *Entropy*, vol. 21, no. 5, pp. 479, 2019.

[20] 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.

[21] 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.

[22] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco, “EMOVO corpus: an Italian emotional speech database,” in *LREC 2014*, 2014, pp. 3501–3504.

[23] 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.

[24] 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.

[25] Philip Jackson and SJUoSG Haq, “Surrey audio-visual expressed emotion (SAVEE) database,” *University of Surrey: Guildford, UK*, 2014.

[26] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th Python in Science Conference*, 2015, vol. 8, pp. 18–25.

[27] Laurens Van der Maaten and Geoffrey Hinton, “Visualizing data using t-SNE.,” *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.