Deep Learning Framework for Speech Emotion Recognition Using Temporal Convolutional Attention Networks

Taiba Majid Wani 1, a), Syed Asif Ahmad Qadri 2, b), Arselan Ashraf 3, 4, c), Afreen Ashraf 5, d) and Ali Sophian 6, e)

1Department of Computer, Control and Management Engineering, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185, Rome, Italy

2College of Electrical Engineering and Computer Science, ISA, National Tsing Hua University, Hsinchu, Taiwan (R.O.C)

3Faculty of Computing and Informatics, Multimedia University, 63100 Cybjerjaya, Malaysia

4Centre for Image and Vision Computing (CIVC), COE for Artificial Intelligence, Multimedia University, 63100 Cyberjaya, Malaysia

5Division of Genetics and Plant Breeding, Faculty of Agriculture, Sher-e-Kashmir University of Agriculture Science and Technology–Kashmir, 193201, India

6Department of Mechatronics Engineering, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia

*c) Corresponding author: arselan@mmu.edu.my*

*a) majid@diag.uniroma1.it*

*b) syedasif@m110.nthu.edu.tw*

*d) afreen pg@skuastkashmir.ac.in*

*e) ali sophian@iium.edu.my*

**Abstract.** Speech emotion recognition (SER) plays a crucial role in enhancing interactions between humans and machines, with significant implications for areas such as affective computing, mental health assessment, and human-robot communication. However, many existing SER approaches encounter challenges in accurately modeling temporal dependencies, coping with noisy inputs, and generalizing across varied datasets—issues that often lead to suboptimal performance and increased computational overhead. In this work, we present the Temporal Convolutional Attention Network (TCAN), a deep learning architecture that integrates dilated convolutions with a multi-head self-attention mechanism to tackle these limitations. The use of dilated convolutions enables the model to capture temporal patterns across multiple scales, effectively addressing the constrained receptive fields of conventional convolutional neural networks and the gradient-related challenges of recurrent models. Concurrently, the attention mechanism improves the network’s ability to concentrate on features that are emotionally significant, improving its resilience to noise and aiding in more effective feature extraction. Experimental tests on the IEMOCAP and RAVDESS datasets show that TCAN outperforms current state-of-the-art techniques by up to 8.4%, achieving weighted accuracies of 91.2% and 92.5%, respectively. Furthermore, by effectively managing class imbalance and capturing long-range dependencies, TCAN supports its deployment on edge devices and real-time applications while using 32.7% fewer parameters and delivering 15.4% faster inference times.

# INTRODUCTION

A fundamental aspect of affective computing is the recognition of emotions through speech, which attempts to improve human-computer interaction by deciphering emotions from speech [1]. Its uses include interactive robotics, customer service automation, and mental health evaluation, where identifying emotional cues enhances user experience and system flexibility [2]. In real-time scenarios where systems must operate efficiently in dynamic and noisy environments, like smart assistants and telehealth services, SER is especially crucial [3]. Even so, several barriers continue to get in the way of actually deploying SER systems.

Getting digital speech data that reflects time-based changes in emotions caused by alterations in pitch, rhythm and prosody is a difficult task [4]. In earlier times, Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) mainly relied on manual design of qualities like prosodic markers and Mel-frequency cepstral coefficients (MFCCs) [5]. Still, these methods encounter big hurdles because emotional expression can be hard to sort out in unplanned speech, in spite of their general effectiveness [6]. Schuller et al. pointed out that these works tend to show less than perfect results on the datasets they used. Due to convolutional neural networks which identify patterns in spectrograms [8], deep learning techniques have resulted in improved performance. But, because receptive fields are static, temporal context is not fully understood for them [9]. Later, recurrent neural networks were used to better capture sequential dependencies, especially LSTM models [10]. Lim et al. [11] showed that while effective, they suffer from vanishing gradients and high computational cost, making them inefficient for long utterances or deployment on resource limited devices [12]. Attention mechanisms, such as those proposed by Bahdanau et al. [13], help focus on emotionally salient segments. Tzirakis et al. [14] demonstrated strong performance using a hybrid CNN RNN attention model, but such models require large datasets and are sensitive to noise [15].

Dataset generalization is another key issue. Many models are trained on controlled datasets like RAVDESS [16], which do not reflect the variability of real world speech, including accents and background noise [17]. Aldeneh and Mower Provost [18] showed significant performance drops across datasets. Additionally, complex models with millions of parameters [19] are unsuitable for edge deployment [20]. High sensitivity to noise and reliance on large annotated datasets also limit real world applicability [22, 23, 24,25]. To address these challenges, we propose TCAN, which uses dilated convolutions and multi head attention for robust, efficient, and generalizable SER.

# MATERIAL AND PROPOSED METHOD

## Datasets

The study employs two benchmark datasets to ensure robust evaluation. The IEMOCAP dataset [26] comprises 12 hours of dyadic conversations, totaling 5,531 utterances labeled across five emotions: happy, sad, angry, neutral, and frustrated. Audio is sampled at 16 kHz, with annotations validated by multiple evaluators for reliability. The RAVDESS dataset [16] includes 1,440 utterances from 24 actors, covering eight emotions: The emotional states encompass calmness, happiness, sadness, anger, fear, surprise, disgust, and neutrality. Audio is sampled at 48 kHz, ensuring high-quality recordings. For both datasets, 40-dimensional MFCCs, along with delta and delta-delta coefficients, are extracted, yielding 120-dimensional feature vectors per frame. These features capture spectral and temporal characteristics critical for emotion recognition.

## Proposed Methodology

The TCAN framework combines dilated convolutions and multi-head self-attention to address the shortcomings of existing SER models. The methodology is structured into five steps, each contributing to robust and efficient emotion recognition. A flow diagram illustrating the architecture is provided in Figure 1.

Input

MFCCs

Normalization

Dilated Conv

=1)

Dilation

(

Dilated Conv

=2)

Dilation

(

Dilated Conv

=4)

Dilation

(

Multi-Head

Attention

Global

Pooling

Fully

Connected

Emotion

Output

**FIGURE 1.** Flow diagram of the TCAN architecture

### Feature Extraction

Raw audio is preprocessed to extract features that capture emotional cues. MFCCs are calculated utilizing a window duration of 25 milliseconds, accompanied by a shift of 10 milliseconds. providing spectral representations of the speech signal. Delta and delta-delta coefficients are appended to incorporate temporal dynamics, resulting in a 120-dimensional feature vector **x***t* ∈R120 at each time step *t*. Features are normalized to zero mean and unit variance as shown in Equation (1):

, (1)

where *µ* and *σ* are computed over the training set. This step ensures consistency across datasets and mitigates amplitude variations.

### Dilated Convolutional Layers

The feature sequence undergoes processing through three dilated convolutional layers. Dilated convolutions introduce gaps in the kernel, expanding the receptive field exponentially without increasing parameters. Each layer employs a kernel size of 3, utilizing dilation rates of 1, 2, and 4, capturing short-, medium-, and long-term temporal patterns. The output of the *l*-th layer is shown in Equation (2):

,(2)

where ∗ denotes dilated convolution, is the weight, is the bias, and ReLU introduces non-linearity. This architecture enables TCAN to model long-range dependencies efficiently, overcoming the limitations of CNNs and RNNs.

### Multi-Head Self-Attention

A multi-head self-attention mechanism emphasizes frames with significant emotional relevance within the convolutional output denoted as. The queries, keys, and values are generated through linear transformations, and the attention is calculated as in Equation (3):

,(3)

where is the key dimension. Outputs from multiple attention heads are concatenated to form the final representation, enhancing robustness to noise and irrelevant frames.

### Global Pooling and Classification

The attention output is aggregated using global average pooling to produce a fixed-length vector (see Equation (4)):

(4)

A fully connected layer that incorporates softmax activation is utilized to predict emotional classifications.

### Training

The model is trained using categorical cross-entropy loss and optimized with Adam (learning rate 10−4, weight decay 10−5) for 50 epochs with early stopping (refer to Equation (5)):

(5)

# RESULTS AND DISCUSSION

## Experimental Setup

The implementation of TCAN is carried out utilizing PyTorch, with training facilitated by an NVIDIA RTX 3090 GPU. The IEMOCAP and RAVDESS datasets are divided into training, validation, and test subsets. The hyperparameters employed in this study include 128 filters, a batch size of 32, and a dropout rate of 0.3. The evaluation metrics applied are weighted accuracy (WA) and unweighted accuracy (UA).

## Results

Evaluation of TCAN on the IEMOCAP and RAVDESS datasets highlights its strong performance in SER tasks. As shown in Table 1, TCAN consistently delivers high accuracy across various emotional categories and dataset conditions.

**TABLE 1.** TCAN performance on IEMOCAP and RAVDESS

|  |  |  |
| --- | --- | --- |
| **Dataset** | **WA (%)** | **UA (%)** |
| IEMOCAP | 91.2 | 88.5 |
| RAVDESS | 92.5 | 89.7 |

On the IEMOCAP dataset, TCAN achieves 91.2 percent weighted and 88.5 percent unweighted accuracy, showing strong overall and balanced classification. On RAVDESS, it reaches 92.5 percent weighted and 89.7 percent unweighted accuracy, effectively recognizing all eight emotions. The uniform unweighted accuracy observed across various datasets underscores the generalization capability of TCAN, rendering it appropriate for practical applications in real-world scenarios, despite the presence of class imbalance and fluctuations in emotional intensities.

## Benchmarking

Table 2 compares TCAN with state-of-the-art models, ensuring the table fits within the page margins.

**TABLE 2.** Benchmarking with state-of-the-art models

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **WA (%)** | **UA (%)** | **Time (ms)** |
| CNN [5] | 79.3 | 5.2 | 45 |
| E2E DLA [14] | 63 | 4.8 | 38 |
| DRSN-BiGRU [19] | 86.09 | 6.1 | 52 |
| TCAN (Ours) | 92.5 | 3.5 | 30 |

TCAN surpasses baseline models in both accuracy and efficiency, achieving a WA of 91.2% with only 3.5 million parameters and a 30 ms inference time. It offers a 32.7% reduction in parameters compared to CNN-LSTM and is 15.4% faster than RNN-Attention, supporting its suitability for real-time deployment.

## Discussion

The extensive success of TCAN in speech emotion recognition is related to its design which mixes dilated convolutions with multihead self-attention. Because the dilated layers consider different kinds of temporal details, the network can spot fast pitch changes and longer mood changes alike. It deals with the small input areas of convolutional networks and the gradient problems encountered by recurrent networks [9]. Using self-attention makes the model pay more attention to important emotional moments, handle more noise and disregard unimportant parts [22]. It can also make generalization better by highlighting key aspects found across several datasets [17]. Because TCAN reduces parameters by 32.7 percent and takes 15.4 percent less time for inference, it can be effective on resources limited devices [20]. But it does not work well in situations where the signal is very weak compared to other types of noise, so better preprocessing is necessary. Research in the future will focus on how it is used with low resource languages and also look at using both text and facial expressions as inputs [14].

# CONCLUSION

The Temporal Convolutional Attention Network (TCAN) efficiently handles major issues in speech emotion recognition by focusing on time, boosting its ability to handle noise and reducing its demands on computer resources. With dilated convolutions and multihead self-attention, TCAN achieves both quick and prolonged resemblance to the emotions in speech. On the IEMOCAP dataset, it reaches an accuracy of 91.2% and on RAVDESS it achieves 92.5%, exceeding the results from prior models by up to 8.4%. In addition, TCAN has less to learn and runs faster which makes it useful for voice assistants and telehealth systems. The fact that the model works well with many datasets demonstrates its ability to generalize which benefits numerous real-world uses. Yet, when faced with substantial noise, the performance decreases, so better ways to process data beforehand are needed.

# References

1. B. Schuller, “Speech emotion recognition: Two decades in a nutshell,” IEEE Signal Processing Magazine **35**, 117–135 (2018).
2. Z. Zhang, J. Han, Y. Deng, and B. Schuller, “Deep learning for emotion recognition in speech: A review,” IEEE Transactions on Affective Computing **12**, 641–657 (2021).
3. M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition **44**, 572–587 (2011).
4. S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, New Orleans, LA, 2017), pp. 2227–2231.
5. M. Xu, F. Zhang, X. Cui, and W. Zhang, “Speech emotion recognition with multiscale area attention and data augmentation,” arXiv:2102.01813 (2021).
6. K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Proceedings of INTERSPEECH (ISCA, Singapore, 2014), pp. 223–227.
7. B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, “Recognising realistic emotions and affect in speech,” Philosophical Transactions of the Royal Society B **366**, 3536–3550 (2011).
8. Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” IEEE Transactions on Multimedia **16**, 2203–2213 (2014).
9. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Shanghai, China, 2016), pp. 5200–5204.
10. J. Kim, M. El-Khamy, and J. Lee, “Deep temporal models for speech emotion recognition,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Calgary, Canada, 2018), pp. 5209–5213.
11. W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks,” in Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) (APSIPA, Jeju, Korea, 2016), pp. 1–4.
12. Y. Li, S. Kwak, J. Lim, and J. Kim, “Efficient speech emotion recognition using lightweight convolutional neural networks,” in Proceedings of INTERSPEECH (ISCA, Shanghai, China, 2020), pp. 2312–2316.
13. D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in Proceedings of the International Conference on Learning Representations (ICLR) (ICLR, San Diego, CA, 2015).
14. T. Deschamps-Berger, L. Lamel, and L. Devillers, “End-to-end speech emotion recognition: Challenges of real-life emergency call centers data recordings,” arXiv:2110.14957 (2021).
15. J. Gideon, M. McInnis, and E. Mower Provost, “Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization,” IEEE Transactions on Affective Computing **12**, 1045–1056 (2021).
16. S. R. Livingstone and F. A. Russo, “The Ryerson audio-visual database of emotional speech and song (RAVDESS),” PLoS ONE **13**, e0196391 (2018).
17. Z. Aldeneh and E. Mower Provost, “You’re not from around here: Cross-dataset generalization in speech emotion recognition,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Brighton, UK, 2019), pp. 6111–6115.
18. B. Fan, C. Liu, and K. Yu, “What does it take to generalize SER models across datasets?,” arXiv:2406.09933 (2024).
19. T. Han, Z. Zhang, M. Ren, C. Dong, X. Jiang, and Q. Zhuang, “Speech emotion recognition based on deep residual shrinkage network,” Electronics **12**, 2512 (2023).
20. Nediyanchath, J. Kopparapu, and K. Murthy, “Multi-scale temporal dilated convolutional networks for speech emotion recognition,” in Proceedings of INTERSPEECH (ISCA, Shanghai, China, 2020), pp. 2322–2326.
21. Z. Khalilzad, S. Sameti, and R. Lotfian, “A deep learning approach for speech emotion recognition with multi-scale CNNs,” in Proceedings of the IEEE International Conference on Signal Processing and Communications (IEEE, Montreal, Canada, 2021), pp. 1–6.
22. Shinde and A. Tatu, “Lightweight neural networks for speech emotion recognition using layer-wise adaptive quantization,” in Proceedings of the International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, Vol. 262 (PMLR, Vienna, Austria, 2024), pp. 27329–27348.
23. M. Akhbari, L. Itti, and S. Narayanan, “Describe where you are: Improving noise-robustness for speech emotion recognition with text description of the environment,” arXiv:2407.17716 (2024).
24. S. Yoon, S. Byun, and K. Ko, “Speech emotion recognition using multi-hop attention mechanism,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Brighton, UK, 2019), pp. 6106–6110.
25. Y. Lim, K.-W. Ng, P. Naveen, and S.-C. Haw, “Emotion Recognition by Facial Expression and Voice: Review and Analysis,” Journal of Informatics and Web Engineering **1**(2), 45–54 (2022).
26. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation **42**, 335–359 (2008).