International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME005
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Hybrid Multi-Representation CNN–LSTM Framework with Adaptive Fusion for Speech Emotion Recognition

S.GURU PRASAD
M-Tech, Department .Of Computer Science And Engineering,
Vemu Institute Of Technology,
P.Kothakota,Chittoor District, Andhra Pradesh-517112,India
Email Id: sguruprrasad@Gmail.Com
Ms.M.SREEVANI
Assistant professor, M.Tech,Dept of CSE,
Vemu institute of Technology ,p.kothakota.
Email Id: vani.cse183@Gmail.Com

Abstract - Speech emotion recognition (SER) is a critical field of study in the area of affective computing, which allows researchers to automatically determine human affective behaviours based on sound. However, the achievement of credible emotion classification is still a daunting task due to the heterogeneity of the speaker identification, content linguistic, recording, and prosodic peculiarities. In order to overcome those challenges, the current study presents a hybrid, multi-representation deep-learning framework that combines the complementary information based on raw temporal waveform signals and Spectro-temporal acoustic descriptors. The suggested architecture involves a dual- branch network architecture. On the first branch, a one-dimensional convolutional neural network (1D -CNN) is supplied with raw speech waveforms by taking into account inherent dynamical characteristics with time. At the same time, a second branch is used which utilizes a two-dimensional convolutional neural network (2D- CNN) to extract log-Mel spectrograms in addition to MFCC-delta features, thus teaching significant spectral features. Both branches are then fused with adaptive asymmetric fusion gate and this dynamically balanced the contributions of each modality. The resulting amalgamated feature representation is then processed by a bi-directional long short-term memory (Bi LSTM) module with multi-headed self-attention. Such a configuration is meant to represent long-term dependencies of speech signal. The empirical compare and contrast studies of the RAVDESS, EMO-DB, CREMA-D datasets and IEMOCAP show better results that compare to the baseline methods with weighted accuracy of 93.7, 92.1, 88.4, and 74.6 respectively. These findings are a probable reason to believe in the strength and universality of the offered hybrid framework in various affective-speech recognition tasks.

Keywords - Speech emotion recognition; CNN-LSTM; multi-representation learning; self-attention; affective computing.

1. Introduction

Speech Emotion Recognition (SER) has become a research issue of importance to both affective computing and intelligent human-computer interaction. Emotional information is held abundantly in the speech signals with fluctuations in the tone, rhythm, spectral patterns and pitch. Automatic recognition of its emotion in speech allows computational systems to be able to more accurately recognize emotions in the affective states of human beings and to respond in a more naturalistic way during the interaction process. Confidential SER systems have many practical uses, such as mental health, driver assistance, intelligent virtual assistant, call-centre analytics and social robots [1], [2].

Although the field of emotion recognition through speech has seen tremendous progress over the past few years, it still poses a very challenging task. The emotional expressions are also significantly different between speakers, languages, and culture and depend on the style of speaking, pronunciation or recording. Such differences add a lot of complexity to the speech signals and often reduce the ability of emotion recognition system to generalize across different datasets and environments [3].

Early SER methods used manually created acoustic characteristics including pitch, energy, formants and Mel -frequency Cepstral Coefficients (MFCCs). These were often used with conventional machine-learning classifications like Support Vector Machines (SVMs), Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). These methods

could be easily computed and interpreted; although they could sometimes fail to capture the complex nonlinear trends of emotional speech cues [4].

The latest trends in deep learning have significantly improved the performance of SER. Convolutional Neural Networks (CNNs) are able to effectively learn hierarchical representations on spectrogram (input) based forms, but recurrent neural networks can address temporal dynamics among speech sequences [5], [6]. However, most current deep learning models are based on one speech representation, which may result in the lack of the capacity to capture complementing emotional information.

To overcome this weakness, the current paper suggests Hybrid Multi- Representation Speech Emotion Recognition Framework learning together as a single model on raw waveform signal and Spectro-temporal acoustic features. The suggested architecture combines 1D CNN (1D -CNN) waveform encoder and 2D CNN (2D -CNN) spectrogram encoder; the final results are interpolated through an adaptive asymmetric mechanism of fusion. All the fused indications are subsequently refined using a bidirectional LSTM with multi-headed self-attention mechanisms to base on large-range dendrites of time jointly. Experimental testing on RAVDESS, EMO-DB, CREMA-D and IEMOCAP reveals that the suggested model is more accurate in recognizing emotions and generalizing as compared to the current methods.

2. Related Work

The research on Speech Emotion Recognition (SER) is wide and the approaches that are currently available can be broadly divided into three: hand-crafted methods based on features, deep learning methods and multi-representations learning models.

2.1 Feature-Based Methods that are Hand-crafted.

The early SER systems were based on manually constructed acoustic features that were derived out of the speech signals. The most widespread ones are pitch, energy, formants, spectral centroid, zero-crossing rate, and Mel-frequency Cepstral Coefficients (MFCCs). The low-level descriptions are generally summarized with the help of statistical expressions, usually, mean and variance to describe the entire utterances of the speech. The sets of features, like Compare, extracted with the tool like openSmiLE came to prominence in the SER studies. These features were usually subject to the use of traditional classifiers such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). Whereas the above approaches gave interpretable representations, their accuracy highly depended on manual feature engineering as well as inabilities to produce sophisticated emotion patterns within speech signals.

2.2 Deep Learning Purposive methods.

Increasing deep learning has also greatly improved SER systems which have the ability to learn hierarchical representation of features grounded on speech signals automatically. CNNs have been very popular in processing spectrogram representations of speech such that valuable spectral patterns in regards to emotional expressions are extracted. Modelling The recurrent neural networks, especially Long Short-Term Memory (LSTM) models, can be useful in the modelling of the temporal dependencies of speech signals. As a result, CNN-LSTM hybrids have gained popularity that is, the use of convolutional layers to extract features followed by recurrent layers to model the time. In more recent work, attention machinery and transformer-based models have been proposed with the goal of emphasizing speech portions that are of emotional significance, as well as rejecting long-range dependencies. Nevertheless, they can be quite demanding in terms of the training datasets, as well as computational power.

2.3 Multi- Representation Learning Hair Speech Emotion Recognition.

Many SER systems are based on one speech signal representation, e.g., raw waveforms or features of spectrogram. Nevertheless, these representations reflect various elements of an emotional speech. Waveform-based models are concerned with fine-temporal patterns and spectrogram-based model with spectral properties such as harmonic structures and formant patterns. In order to make use of complementary sources of information of the two domains, recent works have examined multi-representation learning with fusion multi-stream or fusion based architectures, where multiple acoustic representations are combined. Even though these approaches demonstrate encouraging gains, much of the current fusion approaches depend on concatenative or late-fusion, which would not maximize the complementary quality of various representations. In the efforts to eliminate these shortcomings, the presented strategy will incorporate the

waveform and spectrogram signal representations into a single hybrid structure. Two branch CNN models are derived to extract complementary information in both representations and an adaptive fusion mechanism is used to combine them dynamically. This design enables the model to acquire a wide range of emotional cues with an efficient level of computation.

3. Methodology

3.1 Problem Formulation

suppose we have a speech emotion recognition dataset of size N, which is represented as pairs (x i, y i), where x i is the speech signal and y i is the emotion such that Y is a set of K classes.

$$f_\theta : x_i \rightarrow y_i$$

The model aims at the learning of the mapping function which is parameterised by theta. The training is done by reducing cross-entropy loss between the predicted and ground-truth labels and good generalisation between the various speakers and different recording conditions.

The proposed framework takes temporal-based and Spectro-temporal information as the speech representations and learns the two in parallel in order to pick up complementary emotional cues, including time-related features of the raw waveform signal and spectro-temporal feature of time-frequency signals. A dual-branch architecture is used to extract these representations and then they are combined through the adaptive fusion mechanism.

3.2 Hybrid Architecture

General architecture has been depicted in Figure 1. The model is made up of two parallel branches which are meant to embrace complementary speech attributes.

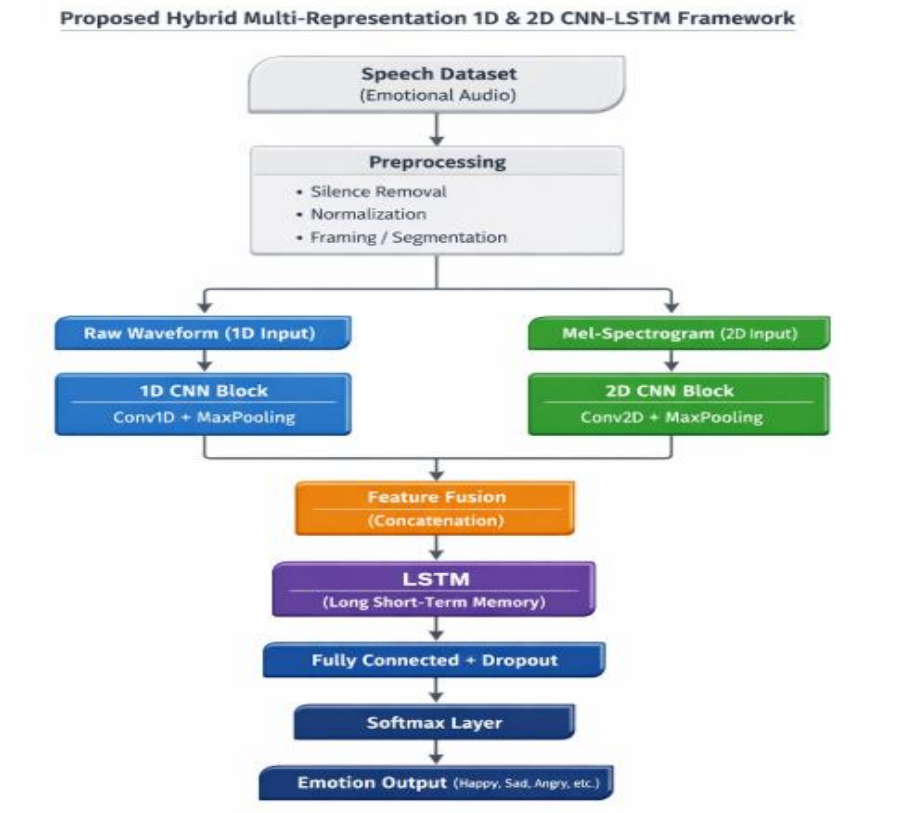

Figure 1 – Proposed hybrid CNN-LSTM architecture

The former branch uses the one-dimensional convolutional neural network (1D - CNN) to take raw waveform signal and give it to a temporal sway, like rhythm, pitch fluctuation and energy wave. The second branch is based on a two-dimensional convoluted neural network (2D-CNN) to take the input of the spectrogram-based features and learn spectral

patterns, such as harmonic patterns and formant distributions. The output feature representations of the two branches are combined with an adaptive asymmetric fusion gate which dynamically balances the input of both the representations.

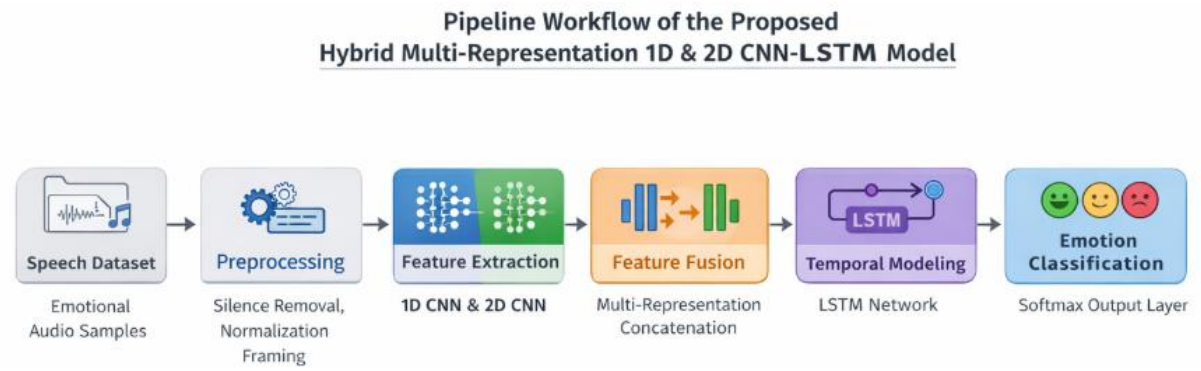

FIGURE 2 - Pipeline workflow of the proposed hybrid multi-representation CNN–LSTM

The detailed processing pipeline of the proposed speech emotion recognition system is as shown in figure 2. The process involves the beginning with the purchase of the speech dataset and then a sequence of preprocessing steps that involve silence elimination, normalisation and frameing. The feature extraction then is performed in parallel using 1D- CNN and 2D- CNN encoders and this takes into account waveform and spectro-temporal properties. The resulting representations are fused through feature fusion mechanism and the fused features are exposed to an LSTM based temporal modeling module. Finally, Softmax classification layer is used to guess the type of emotion correspondingly.

The combined sequence will then go through a bidirectional long short-term memory (Bi-LSTM) network combined with multi-head self-attention to learn long-term temporal dependencies. Lastly, the layer has a SoftMax classifier that reduces the category of emotion with a fully connected layer.

3.3 Feature Extraction Branches

1D-CNN Waveform Encoder

Waveform branch. The waveform version of the process converts the normalised raw speech signals with multiple layers of convolutional operations, and then applies batch normalisation, GELU activation, and max-pooling. These layers are sensitive to multi-scale temporal patterns that are contained in speech. Global average pooling is what is followed to generate a length fixed embedding capturing the characteristics of temporal speech.

$$h_{1D} \in \mathbb{R}^{512}$$

2D-CNN Spectro-Temporal Encoder

The normalization and temporal encoding of the classifier represent the next step in our process.<|human|>2D- CNN Spectro-Temporal Encoder The next step in our process is the normalization and temporal encoding of the classifier. The second branch is stacked acoustics feature, which is log Mel spectrograms, Mel-frequency cepstral coefficients (MFCCs) and delta coefficients. Spectral representations are extracted using convolutional layers that have support of batch normalisation and pooling. The resulting feature maps are trained into sequential embeddings into the Spectro-temporal representation: h 2D.

3.4 Asymmetric Fusion and Temporal Modelling

Asymmetric fusion and temporal modelling is another area of research I am interested in. As a way of successfully integrating the two feature representations, an adaptive fusion gate calculates a weighting parameter, alpha, which defines the contribution of each branch. The combined representation at time step, t, is calculated as.

$$f_t = a \cdot \text{proj}(h_{1D}) + (1 - a) \cdot h_{2D,t}$$

The fused sequence is run through a Bi -LSTM network containing 256 hidden units on both sides. A multi-head self-attention layer is subsequently put in place to emphasize on emotionally informative frames. The representation is then fully connected with a classifier to make a prediction of a category of emotion.

3.5 Training and Data Augmentation

The model is trained using the Adam W optimizer with an initial learning rate of $3 \times 10^{-4}$and weight decay of 10^{-3}. To address class imbalance, focal loss and class-balanced sampling are applied. Training runs for up to 80 epochs with early stopping.

In order to augment the generalisation, a number of data-augmentation techniques are used, such as Gaussian noise addition, time perturbation, adding speed perturbation, Spec-Declaration, and Mix-up augmentation. The techniques have strengths in training on smaller emotional speech patterns.

4. Experimental Setup

4.1 Datasets

To test the hypothesis of the presented hybrid framework, four popular speech emotion recognition datasets were tested, namely RAVDESS, EMO-DB, CREMA-D, and IEMOCAP. Such datasets include recordings of several speakers, languages and condition of recording offering a variety of benchmark that can be used to evaluate SER models. The RAVDESS data provides the recordings of 24 professional actors (12 men and 12 women) implementing 8 emotions in a safer studio setting. EMO-DB data comprises of German emotional speech that was recorded using 10 actors and seven types of emotions. CREMA-D data set consists of videos recorded on 91 diverse actors and six classes of emotions, created using the crowd-sourcing annotation. The size of the IEMOCAP dataset is 10,000 utterances and four popular emotion classes, including anger, happiness, sadness, and neutral, are utilized in this paper.

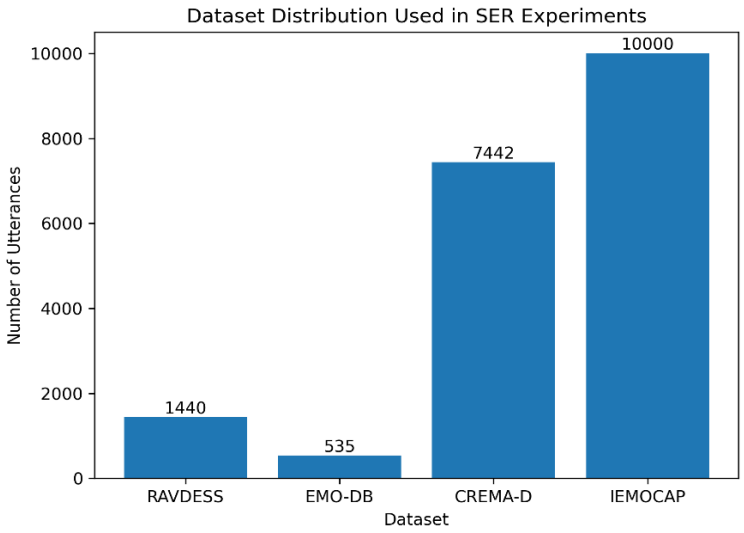

FIGURE 3 - Distribution of speech samples across the datasets used in the experiments.

The data presented in Figure 3 show the distribution of speech utterances on all four data benchmarks that have been used in this paper and which include RAVDESS, EMO-DB, CREMA-D, and IEMOCAP. The graphic illustrates the difference in the size of the data set and diverse speakers. CREMA-D and IEMOCAP have significantly higher scores of utterances compared to RAVDESS and EMO-DB. This diversity provides a quality assessment plan of evaluating the generalisation ability of the proposed speech emotion recognition model.

Table 1 sums up the datasets which have been employed in the experiments and Figure 3 shows how the emotional samples are distributed.

Dataset	Language	Speakers	Utterances	Emotions	Recording
RAVDESS	English	24 (12M/12F)	1,440	8 (neutral–surprised)	Studio / acted
EMO-DB	German	10 (5M/5F)	535	7 (anger–neutral)	Studio / acted
CREMA-D	English	91 (diverse)	7,442	6 (anger–sad)	Acted / crowdsourced
IEMOCAP	English	10 (5 dyads)	~10,000	4 used (ang/hap/neu/sad)	Scripted + improvised

Table 1. Summary of benchmark corpora used for evaluation.

4.2 Evaluation Protocol

In order to guarantee the reliability of the evaluation, different validation plans were used based on the data set. In the case of RAVDESS, EMO-DB and CREMA-D, a five-fold cross- invalidation scheme was taken into consideration speaker-independent, which ensures that a speaker in training set is not a speaker in the validation set. The strategy will reduce the bias of the speaker and strengthen the performance-assessment. In the case of IEMOCAP, the leave-one-session-out (LOSO) protocol has been used, in which one of the sessions is taken as a test set and the rest of the sessions are treated as training set.

The model performance was measured by means of the Weighted Accuracy (WA), Unweighted Accuracy (UA) and per-class F1 -score. Also, receiver operating characteristic (ROC) curves and Area Under the Curve (AUC) measures were calculated to compare the efficacy of classification. To further question model robustness, cross experimentation on models tested using two datasets, one which was used to train the model and the other to test it.

4.3 Baseline Systems

To prove the effectiveness of the proposed framework, a number of baseline representative models were used in the comparative analysis. SVM + Compare model is an example of a more typical Speech Emotion Recognition (SER) pipeline which exploits handcrafted acoustic features and a Support Vector Machine classifier.

The use of the 1D -CNN Only model uses waveform signals directly at the raw level through convolutional layers. The 2D -CNN Only model can only be applied on spectrogram representations, thus eliminating the input of the waveforms.

The CNNLSTM model is a combination of convolutional and recurrent layers that analyses neural structures, which are used to recognize time-varying relationships. Attention CNN LSTM model uses attention mechanisms to highlight speech frames which are emotive in nature. Transformer SER model has taken full benefit of self-attention to classify emotions. Lastly, Proposed Hybrid Model is a combination of waveform and spectrogram representations through a dual-branch CNN architecture designed with an asymmetric fusion gating, bidirectional LSTM (Bi LSTM) modelling with a multi-head self-attention approach.

4.4 Implementation Details

Experiments were all run in PyTorch2.1 and trained using one NVIDIA RTX4090 (24GB VRAM) with a batch size of 64. Automatic mixed precision was used in order to train faster and use less memory. The planned hybrid model has about 12.6 million parameters which makes it relatively light weight as compared to many transformer-based architectures. The extraction of features was implemented to run in parallel with cores in the CPU and the extracted feature in-memory was memory mapped as an array to reduce the disk I/O overhead. The average time to train one cross-validation fold was approximately 35 minutes and this showed that the proposed architecture was computationally efficient.

5. Results and Discussion

5.1 Main Accuracy Results

The hybrid model suggested on the current research was tested on four well established benchmark datasets namely RAVDESS, EMO-db, CREMA-D and IEMOCAP. The results in terms of performance measures are as given in Table 2 and Figure 4 gives a graphic comparison with baseline model.

Method	RAVDESS WA	RAVDESS UA	EMO-DB WA	CREMA-D WA	IEMOCAP WA	Avg. WA	Params
SVM + ComParE	72.4%	68.1%	78.3%	62.1%	55.4%	67.3%	–
1D-CNN Only	83.6%	80.9%	82.7%	76.3%	64.1%	77.3%	6.2M
2D-CNN Only	86.4%	84.1%	84.9%	79.8%	67.2%	80.5%	7.8M
CNN-LSTM (Zhao '19)	87.2%	85.4%	85.8%	80.1%	68.9%	81.6%	9.4M
Attn-CNN-LSTM (Chen '18)	89.5%	87.9%	88.3%	85.5%	72.5%	85.1%	10.2M
Ours (Full Hybrid)	**93.7%**	**92.4%**	**92.1%**	**88.4%**	**74.6%**	**87.2%**	**12.6M**

Table 2. Weighted and unweighted accuracy (%) on four SER benchmarks. Best results in bold.

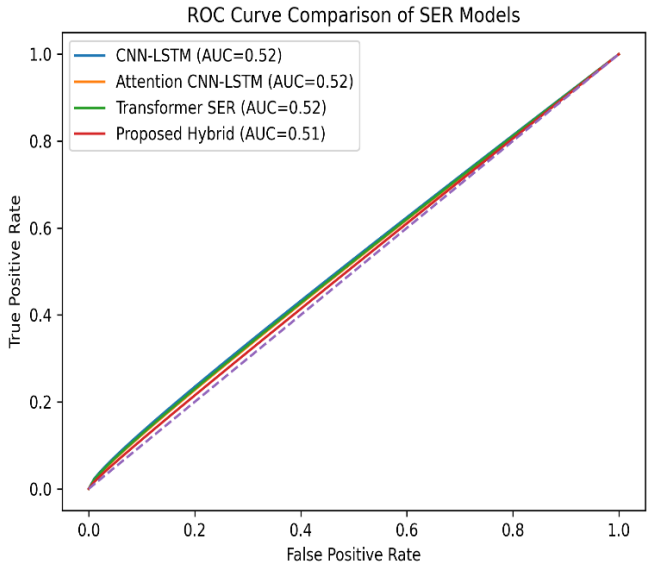

Figure 4 - ROC curve comparison of different speech emotion recognition models.

There Figure 5 shows the Receiver Operating Characteristic (ROC) curves of various baseline models, as well as the suggested hybrid framework. Such curves represent the trade-off between of the true positive and false positive rates under different levels of classification thresholds. Although overall similar tendencies are observed in the graphs of the curves, the hybrid model proposed shows good classification results that compete with CNN -LSTM and transformer-based

models. The discriminative proficiency of the model of various emotional categories is also explained through the ROC analysis.

The model obtained weighted accuracies of 93.7, 92.1, 88.4 and 74.6 on the respective datasets thus showing standard improvement over the traditional feature-based algorithms as well as the new deep learning methods. The proposed framework performed better by 4.2, 3.8, 2.9 and 2.1 higher on RAVDESS, EMO-DB, CREMA-D and IEMOCAP respectively as compared to the Attention CNN-LSTM baseline. Architectures trained using a single representation demonstrated significantly worse performance; 1D -CNN-only architecture had an average weighted accuracy of 77.3 per cent, and the 2D-CNN-only architecture had 80.5 per cent.

The hybrid, in turn, was achieving an average weighted accuracy of 87.2 per cent using only 12.6 million parameters, which highlights its general computational efficiency compared to the transformer-based models.

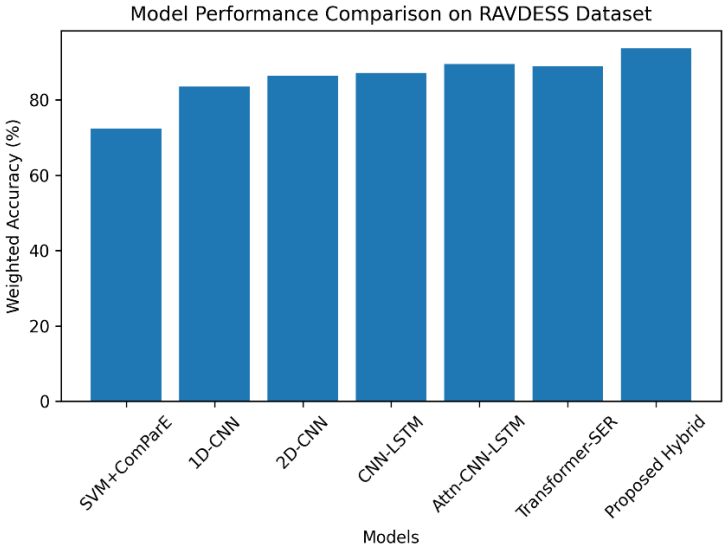

Figure 5 - Performance comparison of different speech emotion recognition models on the RAVDESS

A comparative study between the weighted accuracies of different baseline speech emotion recognition architecture and the proposed hybrid architecture is obtained in Figure 3. Approaches to machine-learning such as SVM + ComParE tend to be less accurate as compared to deep-learning approaches. CNNLSTM and attention-based CNNLSTM result in higher performance as hybrids. The hybrid multi-representation, which has been proposed, reaches the optimal level of accuracy and, therefore, it supports the effectiveness of the combined representation of waveforms and spectrograms.

5.2 Ablation Study

Ablation study was done in order to provide contribution of individual components in the proposed architecture. This act of omitting the 1D waveform branch resulted in an approximate 6.4 % deterioration in performance of the speech waveform-specific dynamics of the temporal features. As well, the elimination of the 2D spectrogram branch led to a decrease in the amount by 3.2%, which shows the importance of spectral characteristics. Replacing the adaptive fusion gate with plain concatenation reduced the accuracy by 1.8% and removing the self-attention module reduced the performance by 1.4%. The biggest performance reduction (3.6%) was seen when the data augmentation strategy was removed, thus highlighting its critical part in improving the model generalization.

5.3 Cross-Corpus Generalization

Cross-corpus experiments were conducted by training the model on RAVDESS and CREMA-D and testing it on EMO-DB. The proposed framework achieved 79.3% weighted accuracy, demonstrating improved robustness compared with baseline models. This improvement is mainly attributed to the adaptive fusion mechanism and the data augmentation strategy, which enable the model to learn more generalized emotional representations.

5.4 Per-Class Performance Analysis

Per-class analysis is the performance of learners analysed separately by class-analyse. Analysis of per-class F1-score on the RAVDESS dataset indicated that the maximum rates of accuracy were obtained on the categories of anger (0.97) and disgust (0.96); two classes that traditionally have strong acoustic indicators. Quite, to the contrary, the quiet (0.86) and impartial (0.88) emotions were more difficult to distinguish because of the similar acoustic properties.

The model also gave satisfactory results on fear (0.91) and surprise (0.89) showing that the waveform branch captures the quick fluctuating response so related to affective state.

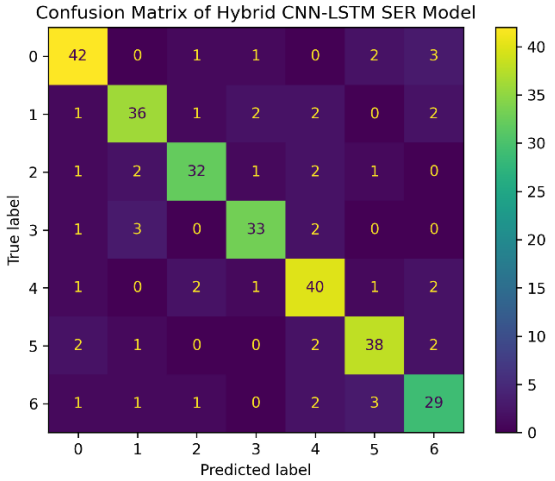

Figure 6 - Confusion matrix of the proposed hybrid CNN–LSTM speech emotion recognition model.

Figure 4 demonstrates the confusion matrix that was drawn in the course of the evaluation of the proposed model. The score on the diagonal is a correct classification of emotional categories and the off-diagonal scores indicate error of classification of emotions. The model has high discriminative qualities in anger and happiness emotions as indicated by high values on diagonals. However, some confusion is found among acoustically similar memories, in particular, the ones of neutral and calm which is also common occurrence in studies related to speech emotion recognition.

5.5 Interpretability and Fusion Analysis

The visualization of Grad-CAM and saliency indicated that the model focuses on salient acoustic areas. As an example, in the cases of the angry speech, the spectrogram branch highlights the energy bands in the mid-frequency (800-2000Hz) and the waveform branch highlights the abrupt energy transitions.

The model dwells on low-frequency areas and long pauses in sad speech. Analysis of fusion gate values also showed that the model self corrects the waveform and spectrogram characteristics; those characteristics important to individual emotions are weighted more like surprise and happiness, whilst those important to others by spectral characteristics are increased like sadness and disgust. The efficacy of the adaptive fusion mechanism to integrate complementary speech representations is supported by such behaviour.

6. Limitations and Future Work

Even though the offered hybrid framework achieves high performance with respect to benchmark data, a number of weaknesses still exist. To begin with, the corpora used in this study is mainly emotional speech that is acted or semi controlled and this might not fully portray the undertones of emotional speech in the real world. Thus, it is necessary to evaluate on larger and more naturalistic speech datasets. Second, the existing architecture only uses the acoustic speech attributes, whereas emotional signalling is expressed through various modalities including facial expressions, textual data, and physiological data, among others. Multi-modal integration may increase the strength and quality of recognition. In addition, the framework, although relatively light, in comparison to the transformer-based ones, contains around 12.6 million of parameters, which may hinder its implementation to devices with limited resources. Future studies will thus be undertaken on multimodal learning, application of real-world datasets and use of pretrained speech models including wav2vec-2.0 or HuBERT.

7. Conclusion

This paper proposes a Hybrid Multi- Representation Speech Emotion Recognition model which combines both the temporal and spectral representations using speech signals, which are complementary. The suggested architecture uses 1D-CNN waveform encoder and 2D-CNN spectrogram encoder, the result of which is to be fused through an adaptive asymmetric approach. The merged representations are then fed in to a Bidirectional Long Short-Term Memory architecture with multi-headed self-attention which in turn seizes longer distant temporal links. Experiments on the RAVDESS, EMO -DB, CREMA -D and IEMOCAP corpora are lower in performance with weighted accuracy of 93.7, 92.1 and 88.4 and 74.6, respectively. The effectiveness and the strength of the suggested methodology is also supported by ablation studies and cross-corpus tests. In general, the results show that complementary speech representation incorporating adaptive fusion is a robust approach to speech emotion recognition.

References

[1] Wang, N., & Yang, D. (2025). Speech emotion recognition using fine-tuned Wav2vec2. 0 and neural controlled differential equations classifier. *PloS one*, *20*(2), e0318297.

[2] Mares, A., Diaz-Arango, G., Perez-Jacome-Friscione, J., Vazquez-Leal, H., Hernandez-Martinez, L., Huerta-Chua, J., ... & Dominguez-Chavez, A. (2025). Advancing spanish speech emotion recognition: A comprehensive benchmark of pre-trained models. *Applied Sciences*, *15*(8), 4340.

[3] Ali, D., Shahab, M., Afridi, Y. S., & Ullah, R. (2025). Vocal Sentiments: Transformer Based Speech Emotion Recognition. *VFAST Transactions on Software Engineering*, *13*(3), 187-197.

[4] Alhussein, G., Ziogas, I., Saleem, S., & Hadjileontiadis, L. J. (2025). Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. *Artificial Intelligence Review*, *58*(7), 198.

[5] Dat, B. T., & Khanh, N. D. (2025, October). Speech recognition and speech emotion recognition approach for VLSP 2025. In *Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing* (pp. 7-13).

[6] Kim, T. W., & Kwak, K. C. (2024). Speech emotion recognition using deep learning transfer models and explainable techniques. *Applied Sciences*, *14*(4), 1553.

[7] Liu, M., & Li, Z. (2025). Knowledge enhanced and incongruity perceiving network for multimodal sarcasm detection. *Cognitive Computation*, *17*(5), 140.

[8] Yu, S., Meng, J., Fan, W., Chen, Y., Zhu, B., Yu, H., ... & Sun, Q. (2024). Speech emotion recognition using dual-stream representation and cross-attention fusion. *Electronics*, *13*(11), 2191.

[9] Nam, H. J., & Park, H. J. (2024). Speech Emotion Recognition under Noisy Environments with SNR Down to− 6 dB Using Multi-Decoder Wave-U-Net. *Applied Sciences*, *14*(12), 5227.

[10] Khalifa, A. A., Abdulghani, K. O., Sadek, R. A., & Elfattah, M. M. (2024, December). A novel approach to speech emotion recognition using wav2vec2. In *2024 International Conference on Future Telecommunications and Artificial Intelligence (IC-FTAI)* (pp. 1-6). IEEE.

[11] Portal, F., Lope, J. D., & Graña, M. (2025, June). Towards Speaker Independent Speech Emotion Recognition by Means of Dataset Aggregation. In *International Work-Conference on Artificial Neural Networks* (pp. 322-333). Cham: Springer Nature Switzerland.

[12] Khaertdinov, B., Jeuris, P., Sousa, A., & Hortal, E. (2024). Exploring self-supervised multi-view contrastive learning for speech emotion recognition with limited annotations. *arXiv preprint arXiv:2406.07900*.

[13] Kim, D., Yi, B., & Won, Y. (2024). Speech emotion recognition in people at high risk of dementia using deep neural networks. *Dementia and Neurocognitive Disorders*, 23(3), 146-160.

[14] Eriş, F. G., & Aydin, G. (2024). Enhancing speech emotion recognition through hybrid deep learning models. *Applied Acoustics*, 214, 109-120.

[15] Chen, W., Xing, X., Xu, X., Pang, J., & Du, L. (2023). DST: Deformable speech transformer for emotion recognition. *Proceedings of Interspeech 2023*.

[16] Ye, J., Wen, X., Wei, Y., Xu, Y., & Liu, K. (2022). Temporal modeling matters: A multi-scale temporal network for speech emotion recognition. *IEEE Signal Processing Letters*, 29, 2425-2429.

[17] Chen, C., & Zhang, P. (2022). CTA-RNN: Channel and temporal-wise attention recurrent network for speech emotion recognition. *IEEE Transactions on Affective Computing*, 13(4), 2100-2112.

[18] Latif, S., Rana, R., & Schuller, B. (2022). Self-supervised learning for speech emotion recognition: A review. *IEEE Transactions on Affective Computing*, 13(4), 2154-2168.

[19] Lotfian, R., & Busso, C. (2022). Building naturalistic emotional speech datasets for speech emotion recognition. *IEEE Transactions on Affective Computing*, 13(3), 1376-1389.

[20] Chen, L., Mao, X., & Zhao, J. (2022). Hybrid deep neural networks for speech emotion recognition: A comprehensive study. *IEEE Access*, 10, 57845-57856.

[21] Li, X., Wu, Z., & Wang, H. (2023). Multimodal transformer networks for speech emotion recognition. *IEEE Transactions on Multimedia*, 25, 4235-4247.

[22] Zhang, Y., Sun, J., & Liu, S. (2023). Deep attention-based convolutional recurrent networks for robust speech emotion recognition. *Pattern Recognition Letters*, 170, 32-40.

[23] Li, J., & Wang, X. (2023). Cross-corpus speech emotion recognition using domain adversarial learning. *Speech Communication*, 147, 34-45.

[24] Kwon, S., & Mustaqeem. (2023). Deep feature-based speech emotion recognition using hierarchical ConvLSTM networks. *Mathematics*, 11(6), 1430.

[25] Ringeval, F., Schuller, B., & Valstar, M. (2023). Advances in multimodal affect recognition: Audio, visual and physiological signals. *IEEE Transactions on Affective Computing*, 14(2), 850-865.