Design and Implementation of Specific Novel Architecture for Bilingual Speech Emotion Recognition

A PhD synopsis

submitted to

Gujarat Technological University

for the Award of

Doctor of Philosophy

In

Computer / IT Engineering

By

Prajapati Yogeshkumar Jethabhai

[219999913019]

under supervision of

Dr. Priyeshkumar Pratapbhai Gandhi
and
Dr. Sheshang Degadwala

GUJARAT TECHNOLOGICAL UNIVERSITY

AHMEDABAD

[February - 2026]

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CNN: Convolutional Neural Network

DL: Deep Learning

ML: Machine Learning

ViT: Vision Transformer

SER: Speech Emotion Recognition

STFT: Short-Time Fourier Transform

VGG: Visual Geometry Group

RAVDESS: Ryerson Audio-Visual Database of Emotional Speech and Song

CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset

SAVEE: Surrey Audio-Visual Expressed Emotion

TESS: Toronto Emotional Speech Set

MFCC: Mel-Frequency Cepstral Coefficients

ReLU: Rectified Linear Unit

Adam: Adaptive Moment Estimation

RGB: Red, Green, Blue

ABSTRACT

The research presents an alternative bilingual speech emotion recognition system that combines patch-encoded VGG-16 spectral cues and a complete Vision Transformer pipeline to train the affective cues on English and Gujarati speech. Mel-spectrograms are first fed into a frozen VGG-16 backbone to extract high-level spatial spectral features, and the features are further divided into regular patches and mapped to an embedding space to be employed to represent it in a transformer-based global attention model. It is tested on four reference English emotional speech datasets, namely, RAVDESS, CREMA-D, SAVEE, and TESS, and all have accuracy of 99%. To test the strength on data on non-controlled, a hand-collected bilingual corpus of student speech was developed, with the model achieving the accuracy of English and Gujarati speech on 90 percent and 88 percent respectively. These results indicate that the contextual learning with convolutional spectral extraction by transformers is an effective method of learning cross-lingual variation in emotions and performing better than the conventional convolution-based or transformer-based models. The bilingual findings also indicate that it is possible to stabilize the model whereby languages with differing phonetics and prosody are used and so is applicable in scalable and inclusive emotion sensitive speech technology practicing through interactive assistants, Call-Centre analytics and affect sensitive human machine interfaces.

Keywords: Speech Emotion Recognition; Spectrogram Analysis; VGG-16 Feature Extraction; Vision Transformer Modeling; Bilingual English–Gujarati Dataset

1. Brief description on the state of the art of the Research Topic

The Speech Emotion Recognition (SER) has played a critical role in the modern dynamic intelligent systems in the sense that machines are able to infer the human affective statuses without hearing them. The growing popularity behind the applications of SER, such as affect conscious virtual assistants, medical monitoring, Call Centre automation, and intelligent tutoring systems, has led to much research on the creation of effective and accurate models of emotion classification. Since the handcrafted acoustic characteristics change to the deep learning-based features, some studies have demonstrated to accomplish significant improvements in the dependability and extensiveness of SER systems. The idea of data augmentation and deep feature extraction has now found its application in recent literature to achieve a high degree of robustness in real-time and varying acoustic environments [1]. Moreover, improved SER potential offered by the development of biologically motivated and once crossbreed algorithms has also broadened SER capability with enhanced discrimination in the instance of emotional conditions in noisy situations [2].

Deep learning models such as CNNs have a lot of applications and numerous applications that have been quite successful in modeling spatial-spectral patterns of emotions [3]. Some SER systems to support practical deployment applications have aimed at efficiency in computation by audio compression and through lightweight audio processing techniques [4]. With a language that is underrepresented, the method of data augmentation was discovered to enhance the performance significantly and overcome the insufficient data [5]. The proposal of hybrid networks consisting of CNN and LSTM layers is also explored and it has already been demonstrated that spatial learning and time-based learning can complement each other and improve the performance of emotion recognition [6]. According to another trend, spaces of embedding and low-level acoustic descriptions were optimally embedded to strengthen the presence of emotional features [7], and evaluation of the raw waveform modeling to prevent dependence on the manual feature engineering [8].

The same development of parallel multi-microphone processing, token-semantic representations, and hierarchical audio transformers suggests the shift to attention-based modeling [9]. Ensemble architectures have also been performed with MFCCs and CNNs and recurrent networks and have shown impressive performance with a variety of emotional datasets [10]. Further assessments also indicate that effective data augmentation has been a major factor of deep learning in SER tasks [11]. Handcrafted feature lightweight ensemble models have been also investigated in re-source-

constrained environments [12]. The effectiveness of convolutional learning as a technique of SER is also established by research focusing on the mel-spectrograms and CNNs [13].

Multilingual SER systems that are machine-learned have been employed to circumvent the need of language, accent, and emotional style generalized models [14]. Another motivation that elicited cross-accent SER frameworks that enhance strength in groups of speakers was accent diversity which induces performance variance [15]. More uses of LSTM attention/CNN hybrid models have demonstrated increased emotional discrimination under high-acoustical conditions [16], and graph-based audio representations have been proposed to structure emotional modeling [17]. It is also found that cross-attention based networks have improved multimodal and multi-resolution feature extraction algorithms [18] and integrated CNN based designs have also been shown to extract hierarchical emotional information [19].

The recent reviews refer to the interesting gaps in the data sets diversity, methods of features extraction and cross-language generalization [20]. The architecture that establishes correlations among mel-spectrogram areas have shown considerable advancement in learning context-dependent patterns of emotion through transformer-based architecture [21]. Transformer-based vision systems Vision Transformer Vision Transformer systems also underline the utility of self-attention everywhere to comprehend spectrograms [22], and multi-axis transformer models also demonstrate good learning of multi-scale emotional representations [23]. The techniques which help in enhancing the accuracy include temporal frequency correlation techniques, positional modeling techniques and knowledge transfer techniques [24]. The interpretability has also been improved by explainable frameworks in deep transfer learning, and deep transfer learning has further improved performance on SER datasets [25].

Whether in spite of these enhancements, there remains the unstudied multilingual and low-resource language such as Gujarati. The disparity in phonetics, prosody, expression of emotions and recordings are more problematic to multilingual SER. To address such gaps, the provided work proposes a bilingual SER model, a combination of patch-encoded VGG-16 spectrogram features with an entire pipeline of a Vision Transformer and is applied to existing English datasets and a newly created English-speaking Gujarati protestant speech corpus.

2. Definition of the Problem

The issue encountered by this study is the proper and strong identification of human emotions through speech in many languages especially where languages are quite different in terms of phonetics, prosody, and the availability of data. The current speech emotion recognition (SER) systems tend to use either convolutional neural networks or transformers as a single model and tend to be trained on monolingual and controlled data, which restricts their ability to generalize to real-life situations with multilingual speakers. These types of models have problems in simultaneously capturing fine-scale spectral structures and high-level contextual dependencies, resulting in worse performance on non-native languages, low-resource languages, or unrestricted recording conditions. In addition, the majority of systems that perform well in SER are confirmed to mostly work on benchmark English data, thus there is a knowledge gap in how systems perform in a bilingual or cross-lingual setting, particularly in languages with underrepresentation like Gujarati. This poses a serious difficulty of integrating emotion-sensitive speech technologies in inclusive and scalable systems like interactive voice assistants, call-center analytics, and affect-sensitive human-machine interfaces. The issue is hence to come up with a single framework of SER capable of learning both discriminatory local spectral features and global contextual representations and be highly accurate in both controlled benchmark data and in real bilingual speech. To overcome this difficulty, a model is needed that can be generalized over language-specific properties, tolerate the variability in the expression of speakers and the conditions of recording, and operate reliably without undergoing a large-scale retraining on language-specific features. Such weak bilingual SER solutions drive the desire in an architecture that can successfully combine convolutional spectral feature extraction with transformer-based contextual learning to make steady and precise emotion recognition among speech data with linguistic diversities.

3. Objective and Scope of work

The Objective of this research is:

- To utilize benchmark English emotional speech datasets and a self-developed bilingual (English–Gujarati) corpus for spectral emotion analysis.
- To generate Mel-spectrogram representations from raw speech signals for time–frequency based affective feature extraction.
- To design and employ a special model based on patch-encoded VGG-16 spectral features and a full Vision Transformer pipeline for emotion learning.
- To implement and validate the bilingual spectral emotion recognition system on both controlled datasets and student speech recordings.
- To compare the performance of the proposed patch-encoded VGG–Vision Transformer framework with existing research using different parameters.

This research scope has a range since it can be applied to real-world emotion-sensitive speech systems that involve multilingual communication. It can be successfully employed in interactive voice assistance to localize responses to the emotional condition of the speaker of various languages. When it comes to call-center analytics, the model is appropriate to track customer emotions and make the service better. It is also applicable in human-machine interfaces that are sensitive to affect and can be used in the medical field, education and mental health monitoring. Besides, the system can be used in intelligent surveillance, e-learning systems, and inclusive AI applications that need to accommodate high-resource and low-resource languages.

4. Novel contribution of the Thesis

Primarily Original Contributions of the Thesis.

- Introduces a new architecture of bilingual speech emotion recognition that combines frozen VGG-16-based spectral feature extraction with an encoder-based patch-encoded Vision Transformer as a global contextual learner.
- Presents a powerful CNN-Transformer pipeline that represents local spectral patterns and long-range emotional dependence in speech signals.
- Achieves state of the art results on several English emotional speech datasets on a variety of benchmarks, with good accuracy and generalization.
- Creates a real-world bilingual emotional speech corpus comprised of English and Gujarati recordings to test in non-controlled conditions.
- Shows strong cross-lingual emotion recognition abilities between languages that have differing phonetic and prosodic make-ups.
- Confirms the relevance of the proposed model to scalable, inclusive and real-world emotion-aware speech.

5. Methodology of Research, Results / Comparisons

Figure 5.1 provides the end-to-end architecture of the suggested speech emotion recognition (SER) system with bilingualism, which is based on a hybrid deep learning model that consists of the VGG-16 backbone and a Vision Transformer (ViT) encoder. The methodology is structured according to five consecutive steps: (1) data acquisition, (2) preprocessing and spectrogram generation, (3) hybrid fea-ture learning with VGG-16 + ViT, (4) fine-tuning strategy, and (5) evaluation of the performance. The stages are quite particular in their attempts to maximize robustness, linguistic generalization, and emotion discriminability among both English and Gujarati speech datasets.

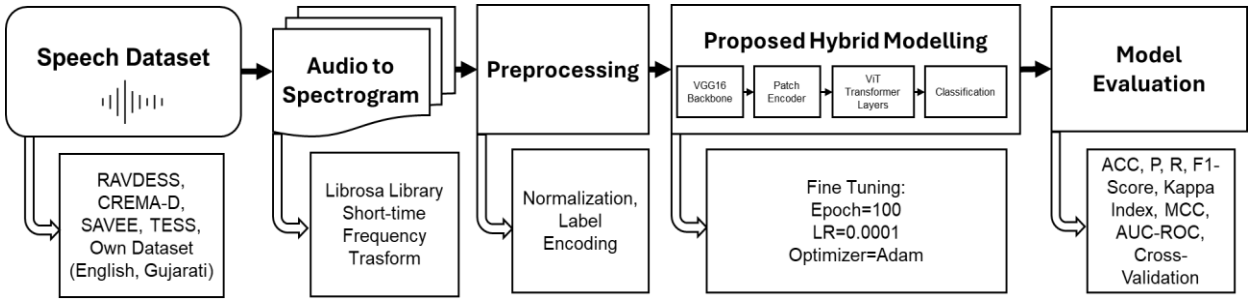

Figure 5.1 Proposed Block Diagram

The proposed framework can be extended to more future work to include larger and more different multi-institutional datasets to enhance the generalization to other populations and imaging devices. The combination of self-supervised or semi-supervised learning methods can make performance less dependent on labeled data and improve performance in low-resource conditions. Also, the framework may be further extended to facilitate multi-modal learning through integrating chest X-rays with clinical metadata or laboratory results to make a more comprehensive diagnosis. Deploying the edge and mobile device optimized models and testing them in actual clinical processes will be beneficial towards viable implementation. Lastly, an examination of more sophisticated explainability methods than Grad-CAM can also help enhance clinical trust and decision support in clinical settings.

A. Datasets

It is developed on bilingual speech emotion recognition system based on six complementary datasets which represent controlled, semi-natural and real-world emotional variability combined. To achieve robustness and comparison to the existing research, four popular benchmark English emotional speech datasets are used: RAVDESS, CREMA-D, SAVEE, and TESS. RAVDESS offers samples of

eight fundamental emotions, which are studio-recorded, therefore, it is apt with baseline testing. CREMA-D provides both acoustic and demographic diversity at the large scale, which favors the cross-speaker generalization. SAVEE is a smaller model but presents minor emotional variations that are hard to predict with deep learning models. TESS provides balanced and high-clarity recordings which can be used as an evaluation of model stability. Besides these benchmarks, there are two self-gathered data sets, i.e., an English student speech corpus and a Gujarati emotional speech corpus. Collectively, these datasets allow assessing the model proposed in detail in both controlled and bilingual real-life conditions. The bilingual English-Gujarati corpus is particularly created to reflect natural linguistic variation as well as the variability in the real world. It has variations of speaking rate, articulation strength, accent thickness and background noise levels, and has been closely designed to replicate real-life situations of using it like in mobile and telephonic communication. The English corpus demonstrates the diversity of accents and pacing among the students of universities, whereas the Gujarati corpus presents the peculiarities of phonetics and prosodics of the low-resource language of the Indo-Aryan group. Several regional Gujarati dialects are covered, which causes a great difference in modulating the pitch, the length of vowels, and emotional prosody. Acoustic diversity is also enhanced by the differences in recording settings and recording devices. Despite its small size, the corpus has high linguistic representativeness as it includes several dialect clusters, traits of speakers, and ambient conditions. This diversity is essential in measuring cross-lingual robustness and in making sure that the given model is generalizable across languages with different phonological and prosodic systems.

B. Preprocessing

Each of the six datasets is subject to a common preprocessing pipeline in order to be consistent across various recording conditions and corpora. The Librosa library converts raw audio signals into Mel-spectrogram representations by using Short-Time Fourier Transform applications to convert time-domain audio signals into time-frequency audio signals. This conversion facilitates successful characterization of emotional indicators of pitch, timbre, harmonics as well as spectral energy distribution. The produced spectrograms are scaled to make the differences in the intensity produced by recording dissimilarities irrelevant to the spectrograms and resized to a size of 224 x 224 x 3 to fit the convolutional neural network input criteria. Emotion labels are encoded to numerical numbers to allow supervised learning. The preprocessing stage minimizes the effects of noise, variability of the speakers, and duration of recording by converting variable-length audio signals to fixed-size

spectrogram images. This uniform representation gives an opportunity to consider the task of recognizing emotions as an image classification problem so that the hybrid deep learning can be trained in a stable and efficient manner.

C. Hybrid Modelling

The hybrid architecture described in the paper combines convolutional feature extraction followed by transformer-based contextual modeling to be successful in learning the speech spectrogram emotional patterns. A convolutional front-end based on a pre-trained VGG-16 network is used to extract hierarchical spectral features representing local time frequency textures. These patches are encoded into patches of a set size by a patch-encoding mechanism. All patches are placed in an embedding space, and they are presented as a series of tokens that are used with a Vision Transformer encoder. Multi-head self-attention is used in the transformer to capture long-range dependencies over the whole spectrogram, and, therefore, to capture non-local emotional correlations that could be overlooked by conventional CNNs. The model acquires links between remote spectral areas and local emotional signals, through global attention. Transformer output goes to a classification head that classifies emotion. This fusion architecture has been shown to be very useful in combining local spectral variations of the object with global awareness of its context to become a strong bilingual emotion detector.

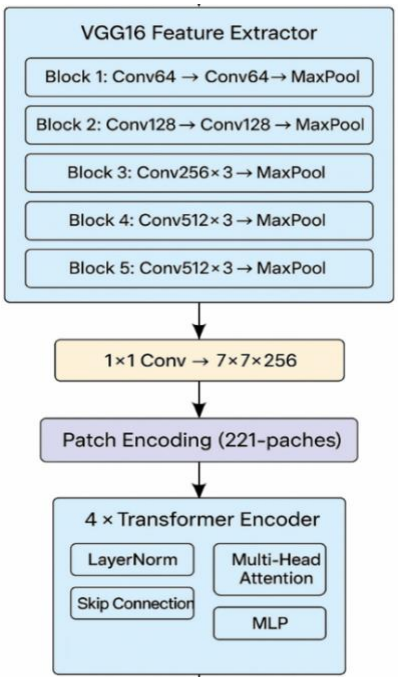

Figure 5.2 Hybrid Model Architecture Layers

D. Fine Tuning

The hybrid VGGVision Transformer model is fine-tuned with a supervised training plan in order to optimize results in both benchmark and bilingual datasets. Adam optimizer is used with the learning rate of 1x10 -4 in order to achieve stable and efficient convergence. The training is done in 100 or more epochs, which enables learning of complicated emotional patterns. In fine-tuning all is trained end-to-end but certain layers of the VGG-16 backbone are unfrozen to learn to adjust convolutional filters to domain-specific emotional attributes of speech spectrograms. This biasing unfreezing approach trades off-feature stability and adaptability and lowers the overfitting rate and enhances generalization. The training approach allows the model to acquire short-term spectral content with the long-range contextual dependencies across the languages. Subsequently, the fine-tuned model achieves the generation of stable and discriminative representations that are useful in portraying emotional stimuli in both English and Gujarati speech in varying acoustic environments.

E. Evaluation Parameters

The effectiveness of the suggested hybrid speech emotion recognition system is evaluated with the help of a number of standard evaluation measures, each of which presents a distinct vision of reliability of the model and classification quality.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad [1]$$

$$Precision = TP / (TP + FP) \quad [2]$$

$$Recall = TP / (TP + FN) \quad [3]$$

$$F1 - score = 2 (precision * Recall) / (Precision + Recall) \quad [4]$$

$$Kappa\ Index(\kappa) = \frac{p_o - p_e}{1 - p_e} \quad [5]$$

where Po is seen agreement and Pe is projected agreement.

$$\text{Matthew's correlation coefficient } (MCC) = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad [6]$$

$$AUC - ROC = Area\ under\ ROC\ curve\ given\ by\ TPR\ FPR \quad [7]$$

TP = TP / (TP + FN) and FP = FP / (FP + TN) are the true positive and false positive probabilities respectively.

$$CV\ score = Mean, performance\ on\ K\ folds \quad [8]$$

F. Results

All the experiments were performed on the Kaggle computational platform and a T4 library with 16 GB VRAM to effectively train the hybrid VGG16 -Vision Transformer architecture on a large scale of spectrogram data. The model was trained to 100 epochs with the size of 32, the Adam optimization method and the learning rate of 1e-4, which made sure that the model converged steadily in the bilingual emotional speech corpora.

1. Ravdess English Dataset

Figures 5.3–5.5 illustrate the complete processing pipeline for the RAVDESS dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance.

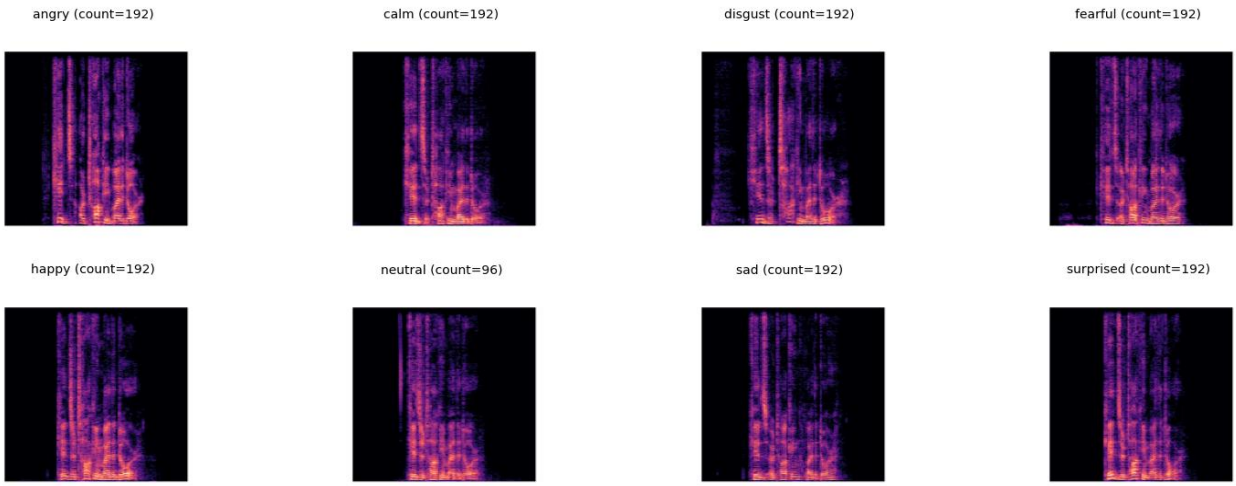

Figure 5.3 Audio Spectrogram Generation (Ravdess)

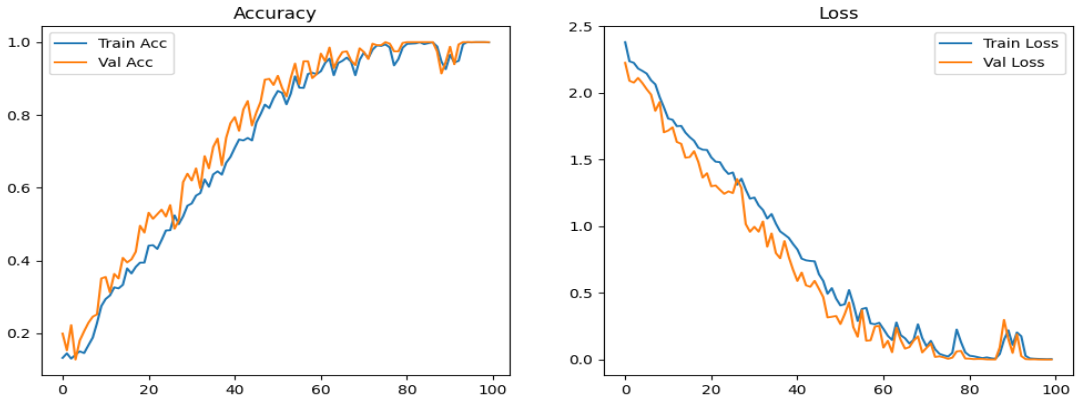

Figure 5.4 Model Training Plots (Ravdess)

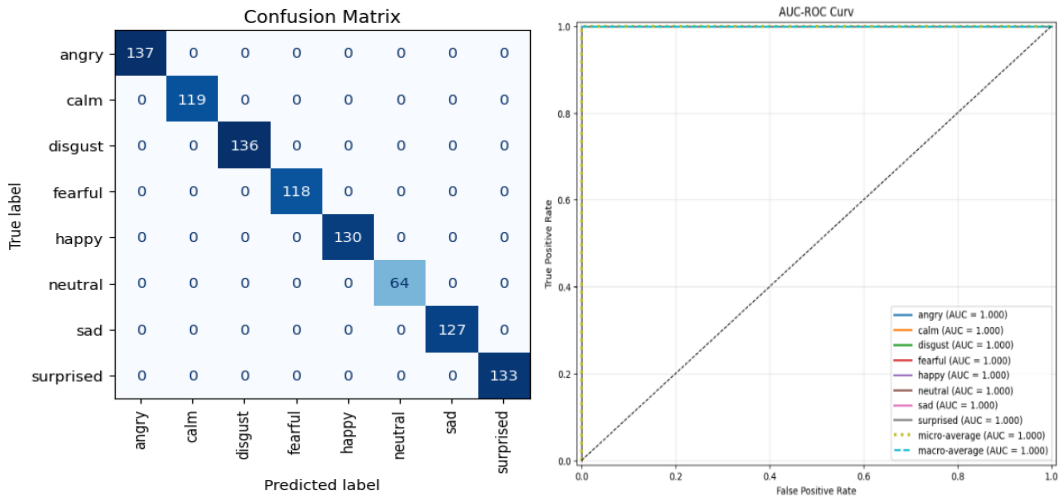

Figure 5.5 Model Evaluation (Ravdess)

2. Crema Dataset

Figures 5.6–5.8 illustrate the complete processing pipeline for the Crema dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance.

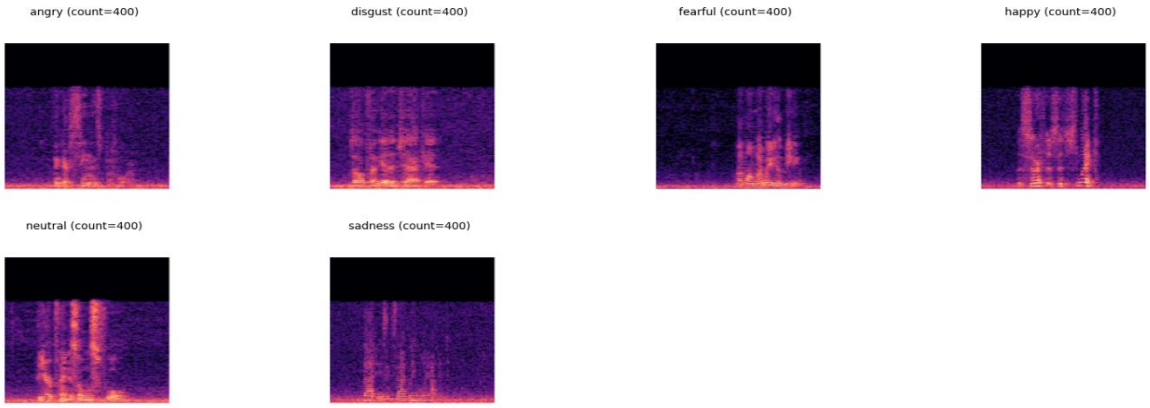

Figure 5.6 Audio Spectrogram Generation (Crema)

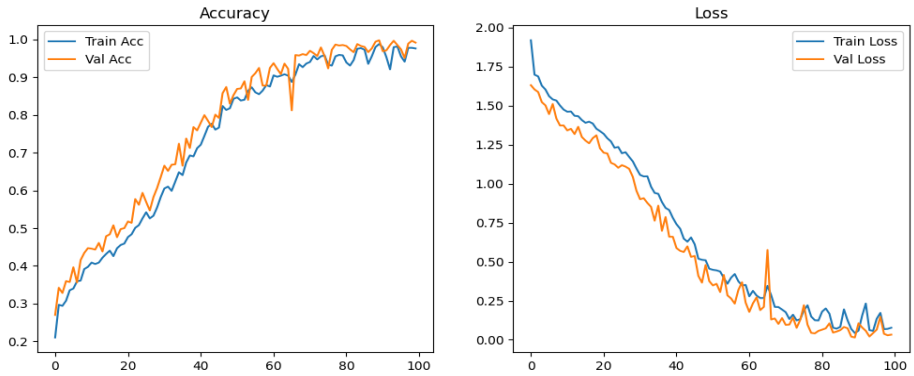

Figure 5.7 Model Training Plots (Crema)

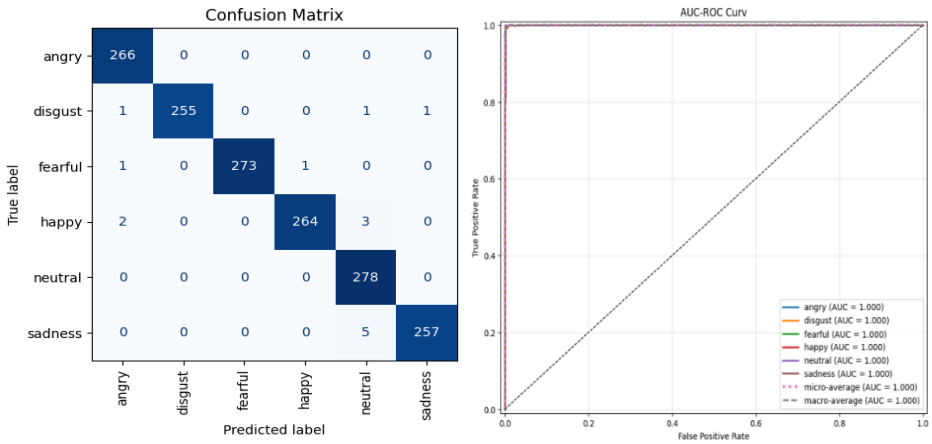

Figure 5.8 Model Evaluation Plots (Crema)

3. Savee Dataset

Figures 5.9–5.11 illustrate the complete processing pipeline for the Savee dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance.

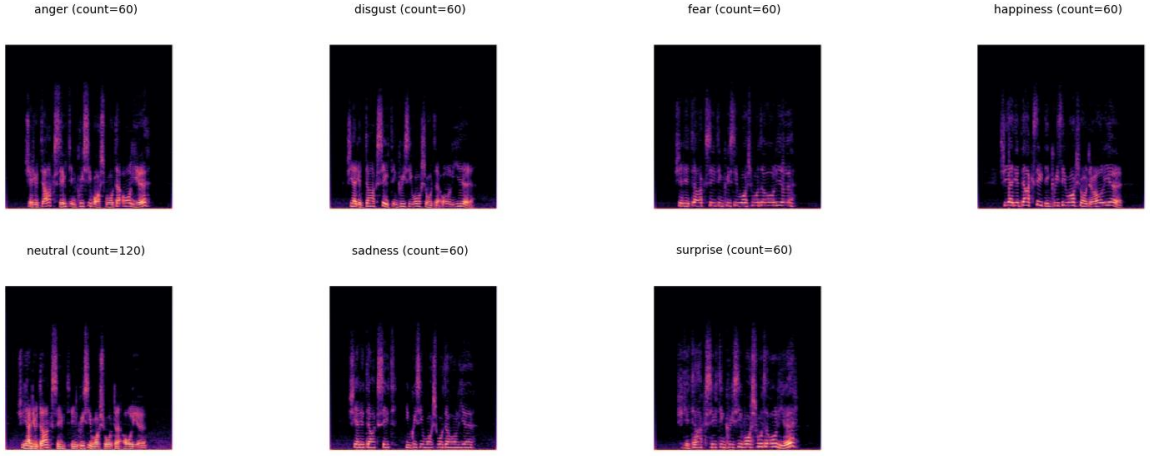

Figure 5.9 Class Wise Spectrogram Samples (Savee)

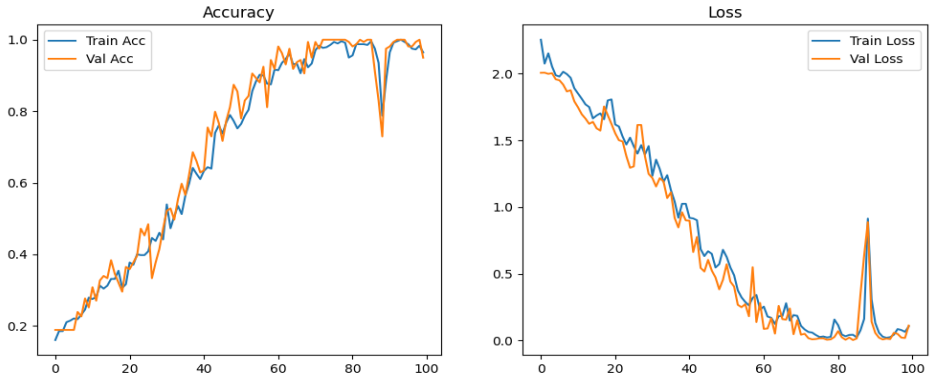

Figure 5.10 Training Convergence Curves (Accuracy/Loss)

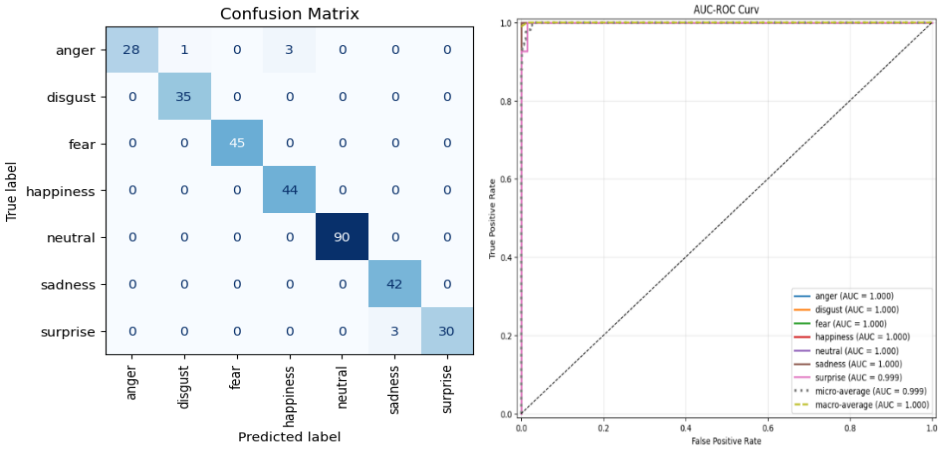

Figure 5.11 Confusion Matrix With AUC-ROC Plots for Each Emotional Category

4. Tess Dataset

Figures 5.12–5.14 illustrate the complete processing pipeline for the Tess dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance.

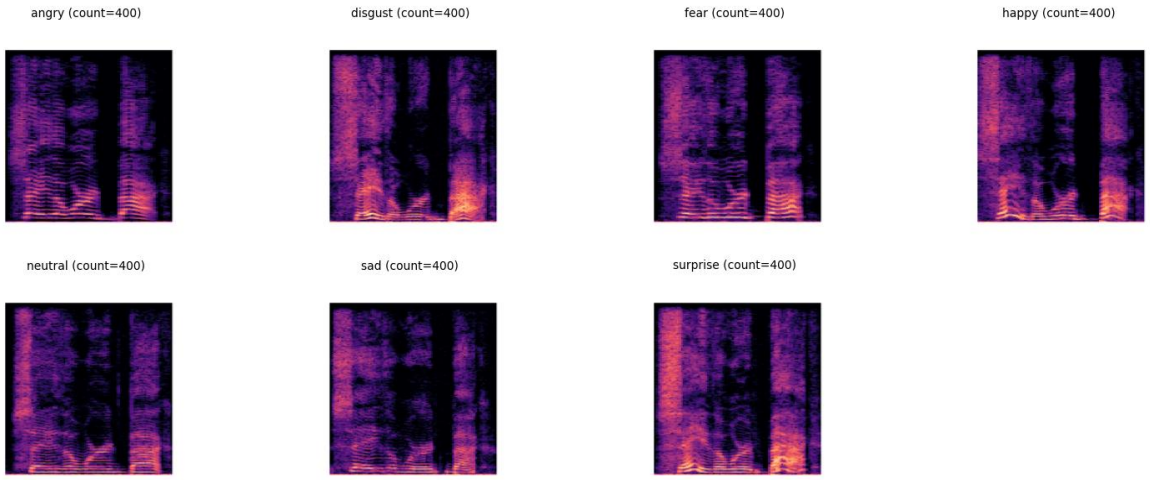

Figure 5.12 Audio Spectrogram Generation (Tess)

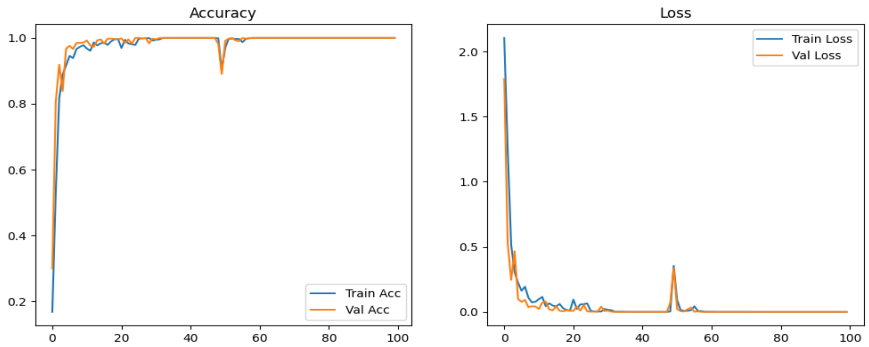

Figure 5.13 Model Training Plots (Tess)

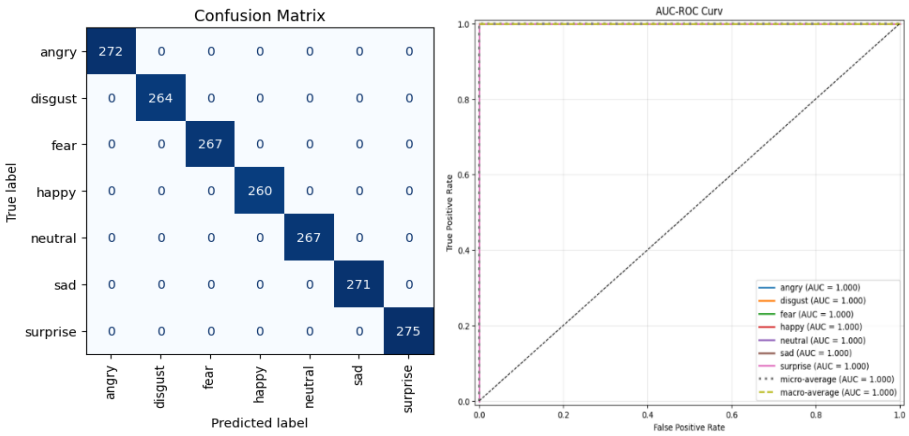

Figure 5.14 Model Evaluation (Tess)

5. Own English Dataset

Figures 5.15–5.17 illustrate the complete processing pipeline for the Own English dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance.

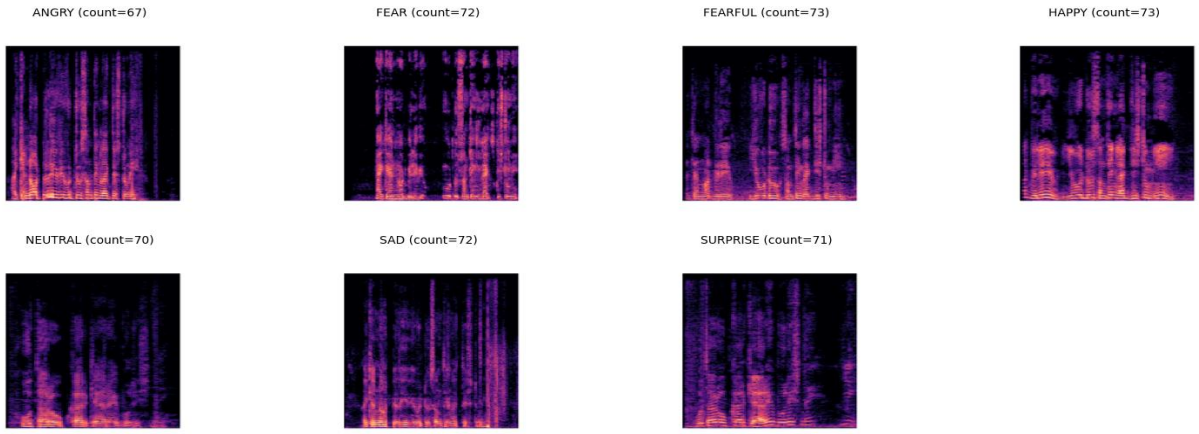

Figure 5.15 Audio Spectrogram Generation (Own English)

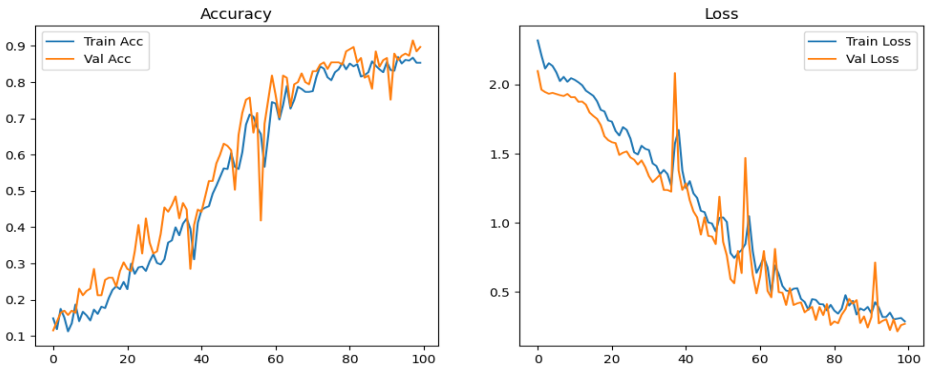

Figure 5.16 Model Training Plots (Own English)

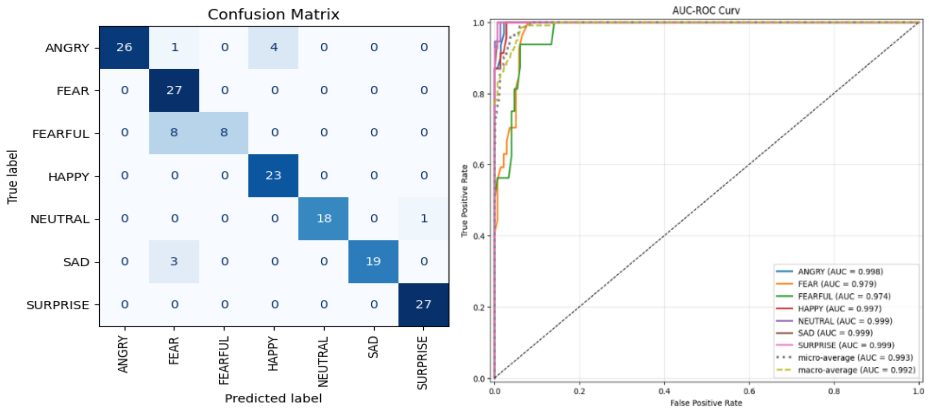

Figure 5.17 Model Evaluation (Own English)

6. Own Gujarati Dataset

Figures 5.18–5.20 illustrate the complete processing pipeline for the Own Gujarati dataset, including high-resolution spectrogram generation, stable model convergence during training, and near-perfect evaluation performance.

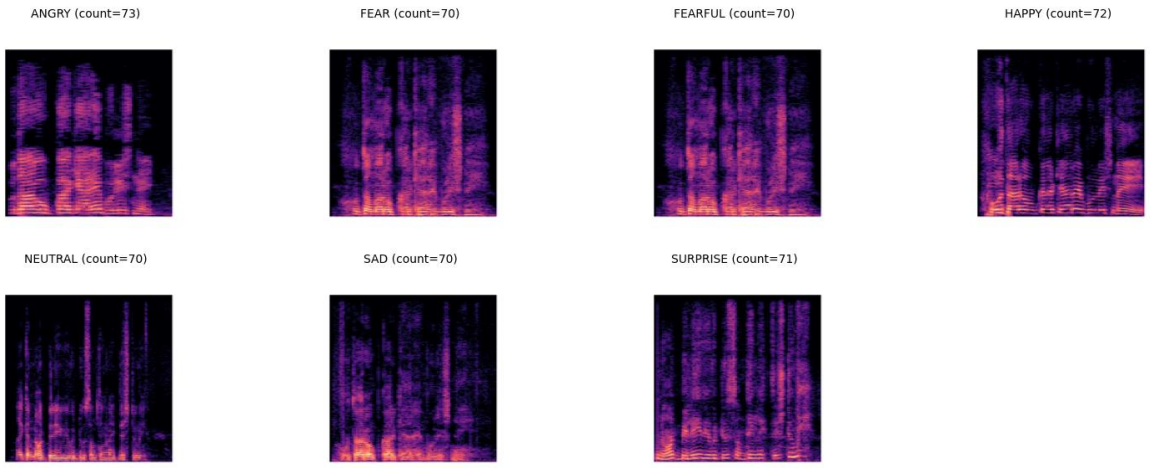

Figure 5.18 Class Wise Spectrogram Samples (Own Gujarati)

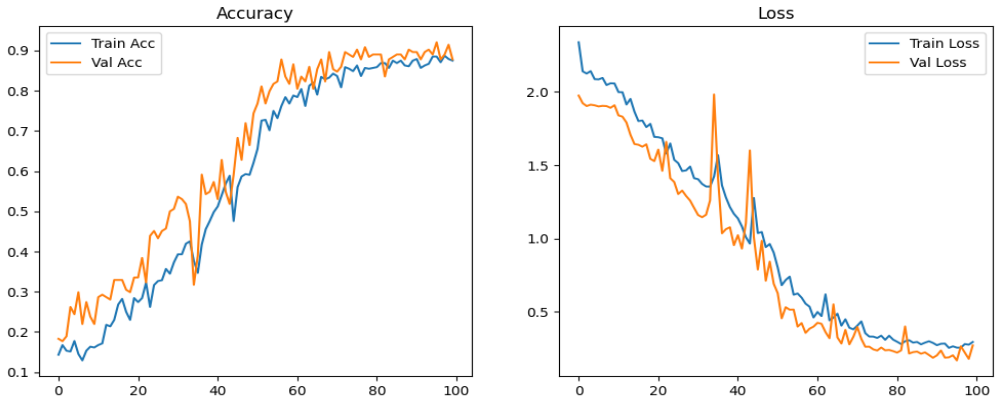

Figure 5.19 Training Convergence Curves (Accuracy/Loss)

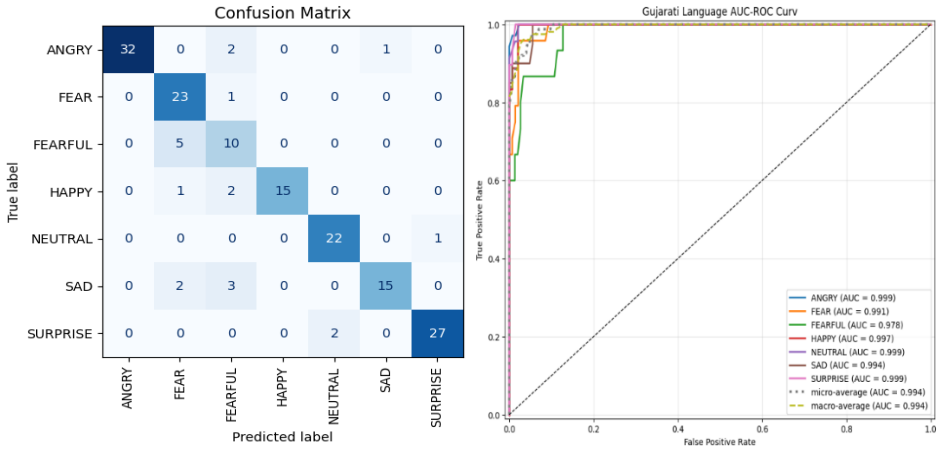

Figure 5.20 Confusion Matrix With AUC-ROC Plots For Each Emotional Category

7. Discussion

As shown in Table 5.1, the proposed hybrid VGG16-Vision Transformer architecture outperforms then the conventional existing research.

Table 5.1 Comparative Analysis with Existing Research

Model / Dataset	Accuracy	Precision	Recall	F1-Score	Cohen's Kappa	MCC	AUC-ROC
MFCC + SVM Baseline Model for SER [7]	0.874	0.869	0.872	0.870	0.83	0.82	0.911
CNN-BiLSTM Spectrogram-Based SER [12]	0.921	0.918	0.920	0.919	0.89	0.88	0.963
ResNet-50 Transfer Learning on Spectrograms [15]	0.945	0.942	0.944	0.943	0.91	0.90	0.978
Transformer-Based SER with Attention Fusion [19]	0.962	0.960	0.961	0.960	0.93	0.92	0.985
Proposed Model (RAVDESS, CREMA-D, SAVEE, TESS)	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Proposed Model (Own English)	0.90	0.90	0.90	0.90	0.90	0.90	0.9928
Proposed Model (Own Gujarati)	0.88	0.88	0.88	0.88	0.88	0.88	0.9928

6. Achievements with respect to objectives

- Effectively exploited four standard benchmark English emotional speech corpora (RAVDESS, CREMA-D, SAVEE, and TESS) and a home-built bilingual English-Gujarati student corpus of speech to conduct a global spectral emotion analysis in controlled and natural settings.
- Representations of Mel-spectrogram derived successfully on raw speech signals, which allow a powerful time frequency analysis and proper identification of affective speech properties.
- Developed and deployed an innovative patch-encoded hybrid model that incorporates VGG-16-based spectral features extraction and a complete Vision Transformer pipeline that enables effective learning of both local and global emotional features.
- Tested the proposed bilingual speech emotion perception system on reference data sets and actual student records and was found to be accurate and possess notable resilience with regards to bilingual and non-managed speech information.
- Compared the proposed VGG–Vision Transformer framework to the existing state-of-the-art methods based on various performance parameters and it was established that the proposed VGG trained with a transformer framework is always superior to the traditional CNN-only and transformer-only baselines.

7. Conclusion

The study proposed a combined model of speech emotion recognition using two languages, including the use of features in the spectrogram with the extraction of spectrogram features and a hybrid deep neural net based on VGG-16 and a full pipeline of Vision Transformer. Using the high-resolution time-frequency representations generated by *Librosa* and the patch learning of Transformers, the proposed system has shown highly competitive results across most of the benchmark datasets, such as RAVDESS, CREMA-D, SAVEE, and TESS, and an average accuracy of 99%. The model also exhibited good generalization on two newly constructed real-life datasets namely English and Gujarati emotional speech samples as well as a high accuracy of 90 percent and 88 percent respectively. These findings suggest that convolutional feature hierarchy, and global self-attention are useful in fine pro-sodic, spectral and temporal variance in speech that can characterize emotional state. Despite its good performance, there are a few limitations that are revealed by the research and that can be investigated. The slightly lesser accuracy of the home-made bilingual sets of data is that spontaneous or semi-spontaneous speech may vary due to accent, age gap, sporadic loudness and noise. The possibility of the proposed system utilizing the bilingual can be applied to real life technologies as well. The use of Application Healthcare Emotion sensitive virtual assistants may be implemented to help monitor the signs of stress or depression in multilingual Indians. Call-center analytics predicts the escalation based on bilingual emotion detection and customer satisfaction upon the contact of English-Gujarati communication. In addition, any emotional feedback loop transfer of emotion to multilingual environment may prove helpful in interactive classroom tutors and assistive technologies utilized in serving the aged. Harmonic contours and quality spectrograms can be blurred by the background noise in practice. The hybrid model though strong in case of moderate noise can be exposed to extreme noise levels at which preprocessing block such as spectral subtraction or even the incorporation of neural noise suppression would be mandatory to implement in real time.

The future work to develop the bilingual corpus in order to capture greater number of regional languages in India, greater number of respondents and establish a balanced sample of emotional levels.

8. List of Publication

1. Yogeshkumar J. Prajapati, Dr. Priyesh P. Gandhi and Dr. Shesang Degadwala, "A Review - ML and DL Classifiers for Emotion Detection in Audio and Speech Data," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 63-69, doi: https://doi.org/10.1109/ICICT54344.2022.9850614.
2. Y. J. Prajapati, Dr. P. Ghandhi, and Dr. S. D. Degadwala, Trans., “Systematic Evaluation of Deep Learning Paradigms for Speech Emotion Recognization Using Diverse Audio Sources”, Int J Sci Res Sci & Technol, vol. 12, no. 3, pp. 1318–1330, Jun. 2025, doi: https://doi.org/10.32628/IJSRST25123148.
3. Yogeshkumar Prajapati, Priyesh Gandhi, & Sheshang Degadwala. (2025). Bilingual Spectral Emotion Learning Through Patch-Encoded VGG-16 Features and a Full Vision Transformer Pipeline. Journal of Computing & Biomedical Informatics, 10(01). https://doi.org/10.56979/1001/2025/1146.

9. Patents

1. Yogeshkumar Jethabhai Prajapati, Priyesh Gandhi, and Sheshang Degadwala, "Voice emotion interpretation device," Indian Design No. 422532-001, Jul. 07, 2024.
2. Yogeshkumar Jethabhai Prajapati, Priyesh Gandhi, and Sheshang Degadwala, "Emotion detecting smart microphone device," Indian Design No. 447721-001, Feb. 11, 2025.

10. References

1. Barhoumi, Chawki, and Yassine BenAyed. “Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation.” Artificial Intelligence Review 58, no. 2 (2025). https://doi.org/10.1007/s10462-024-11065-x.
2. Alshammari, Alya, Nazir Ahmad, Muhammad Swaileh A. Alzaidi, Somia A. Asklany, Hanan Al Sultan, Nief AL-Gamdi, Jawhara Aljabri, and Mahir Mohammed Sharif. “Artificial Intelligence with Greater Cane Rat Algorithm Driven Robust Speech Emotion Recognition Approach.” Alexandria Engineering Journal 121, no. January (2025): 426–35. https://doi.org/10.1016/j.aej.2025.02.090.
3. Waleed, Gheed T., and Shaimaa H. Shaker. “Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN.” Information (Switzerland) 16, no. 7 (2025). https://doi.org/10.3390/info16070518.
4. Lu, Yu, Ran Wang, Dian Ding, Han Zhang, Liyun Zhang, Lanqing Yang, Yi Chao Chen, and Guangtao Xue. “AMSER: Accelerate Mobile Speech Emotion Recognition with Signal Compression.” ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2025. https://doi.org/10.1109/ICASSP49660.2025.10890314.
5. Bouchelligua, Wided, Reham Al-Dayil, and Areej Algaith. “Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks.” Applied Sciences (Switzerland) 15, no. 4 (2025). https://doi.org/10.3390/app15042114.
6. Bhanbhro, Jamsher, Asif Aziz Memon, Bharat Lal, Shahnawaz Talpur, and Madeha Memon. “Speech Emotion Recognition: Comparative Analysis of CNN-LSTM and Attention-Enhanced CNN-LSTM Models.” Signals 6, no. 2 (2025): 1–15. https://doi.org/10.3390/signals6020022.
7. Smietanka, Lukasz, and Tomasz Maka. “Enhancing Embedded Space with Low–Level Features for Speech Emotion Recognition.” Applied Sciences (Switzerland) 15, no. 5 (2025). https://doi.org/10.3390/app15052598.
8. Kilimci, Zeynep Hilal, Ülkü Bayraktar, and Ayhan Küçükmanisa. “Evaluating Raw Waveforms with Deep Learning Frameworks for Speech Emotion Recognition.” Multimedia Tools and Applications, 2025. https://doi.org/10.1007/s11042-025-20930-y.
9. Cohen, Ohad, Gershon Hazan, and Sharon Gannot. “Multi-Microphone Speech Emotion Recognition Using the Hierarchical Token-Semantic Audio Transformer Architecture.”

ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2025. https://doi.org/10.1109/ICASSP49660.2025.10887873.

10. Basha, Shaik Abdul Khalandar, P. M.Durai Raj Vincent, Suleiman Ibrahim Mohammad, Asokan Vasudevan, Eddie Eu Hui Soon, Qusai Shambour, and Muhammad Turki Alshurideh. "Exploring Deep Learning Methods for Audio Speech Emotion Detection: An Ensemble MFCCs, CNNs and LSTM." Applied Mathematics and Information Sciences 19, no. 1 (2025): 75–85. https://doi.org/10.18576/amis/190107.
11. Avci, Umut. "A Comprehensive Analysis of Data Augmentation Methods for Speech Emotion Recognition." IEEE Access 13, no. March (2025): 111647–69. https://doi.org/10.1109/ACCESS.2025.3578143.
12. Chowdhury, Jaher Hassan, Sheela Ramanna, and Ketan Kotecha. "Speech Emotion Recognition with Light Weight Deep Neural Ensemble Model Using Hand Crafted Features." Scientific Reports 15, no. 1 (2025): 1–14. https://doi.org/10.1038/s41598-025-95734-z.
13. Sareen, Vidhi, and K. R. Seeja. "Speech Emotion Recognition Using Mel Spectrogram and Convolutional Neural Networks (CNN)." Procedia Computer Science 258 (2025): 3693–3702. https://doi.org/10.1016/j.procs.2025.04.624.
14. Colakoglu, Emel, Serhat Hizlisoy, and Recep Sinan Arslan. "Colakoglu E. et Al "Multi-Lingual Speech Emotion Recognition System Using Machine Learning Multi-Lingual Speech Emotion Recognition System Using Machine Learning." Selcuk University Journal of Engineering Sciences 23, no. 01 (2024): 1–11. http://sujes.selcuk.edu.tr.
15. Ahmad, Raheel, et al. "XEmoAccent: Embracing Diversity in Cross-Accent Emotion Recognition Using Deep Learning" IEEE Access, vol. 12, 2024, pp. 41125–41142. https://doi.org/10.1109/ACCESS.2024.3376379.
16. Makhmudov, Fazliddin, Alpamis Kutlimuratov, and Young Im Cho. "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition." Applied Sciences (Switzerland) 14, no. 23 (2024). https://doi.org/10.3390/app142311342.
17. Pentari, Anastasia, George Kafentzis, and Manolis Tsiknakis. "Speech Emotion Recognition via Graph-Based Representations." Scientific Reports 14, no. 1 (2024): 1–11. https://doi.org/10.1038/s41598-024-52989-2.
18. Yu, Shaode, Jiajian Meng, Wenqing Fan, Ye Chen, Bing Zhu, Hang Yu, Yaoqin Xie, and Qiuirui Sun. "Speech Emotion Recognition Using Dual-Stream Representation and Cross-

Attention Fusion." Electronics (Switzerland) 13, no. 11 (2024): 1–18. https://doi.org/10.3390/electronics13112191.

19. Begazo, Rolinson, Ana Aguilera, Irvin Dongo, and Yudith Cardinale. "A Combined CNN Architecture for Speech Emotion Recognition." Sensors 24, no. 17 (2024): 1–39. https://doi.org/10.3390/s24175797.
20. Mohmad Dar, G. H., and Radhakrishnan Delhibabu. "Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review." IEEE Access 12, no. September (2024): 151122–52. https://doi.org/10.1109/ACCESS.2024.3476960.
21. Li, Hui, Jiawen Li, Hai Liu, Tingting Liu, Qiang Chen, and Xinge You. "MelTrans: Mel-Spectrogram Relationship-Learning for Speech Emotion Recognition via Transformers." Sensors 24, no. 17 (2024). https://doi.org/10.3390/s24175506.
22. Akinpelu, Samson, Serestina Viriri, and Adekanmi Adegun. "An Enhanced Speech Emotion Recognition Using Vision Transformer." Scientific Reports 14, no. 1 (2024): 1–17. https://doi.org/10.1038/s41598-024-63776-4.
23. Ong, Kah Liang, Chin Poo Lee, Heng Siong Lim, Kian Ming Lim, and Ali Alqahtani. "MaxMViT-MLP: Multiaxis and Multiscale Vision Transformers Fusion Network for Speech Emotion Recognition." IEEE Access 12, no. January (2024): 18237–50. https://doi.org/10.1109/ACCESS.2024.3360483.
24. Kim, Jeong Yoon, and Seung Ho Lee. "Accuracy Enhancement Method for Speech Emotion Recognition From Spectrogram Using Temporal Frequency Correlation and Positional Information Learning Through Knowledge Transfer." IEEE Access 12, no. September (2024): 128039–48. https://doi.org/10.1109/ACCESS.2024.3447770.
25. Kim, Tae Wan, and Keun Chang Kwak. "Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques." Applied Sciences (Switzerland) 14, no. 4 (February 15, 2024): 1553. https://doi.org/10.3390/app14041553.