Speech Emotion Recognition: Clustering and Deep Learning Approach to Detect Conflicting Emotions through Vocal Expressions

by

Nuhash Kabir Neeha
21301025
Nuzhat Rahman
21301538
Imtela Islam
21341018

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
Month Year.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nuhash Kabir Neeha
21301538

Nuzhat Rahman
21301538

Imtela Islam
21301538

Approval

The thesis/project titled "Speech Emotion Recognition: Clustering and Deep Learning Approach to Detect Conflicting Emotions through Vocal Expressions" submitted by

1. Nuhash Kabir Neeha (21301025)
2. Nuzhat Rahman (21301538)
3. Imtela Islam (21341018)

of Summer 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall 2024.

Examining Committee:

Supervisor:
(Member)

Dibyo Fabian Dofadar

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

____

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Given the increasing dependency to automated systems, it is essential that machines understand human emotions more effectively in order to improve user experience as well as prevent extreme damages in multiple areas, starting from customer service and virtual assistants to emergency hotlines and criminal interrogation. As such, Speech Emotion Recognition (SER) is a crucial part of human-computer interaction (HCI) as it aims to improve a machine's ability to understand human emotions through their vocal expressions. However, despite advancements, current SER models face challenges to accurately recognize emotions due to one having multiple emotions (conflicting emotions) at a time. This study aims to cluster and then train deep learning models to accurately recognize such emotional nuances. In conclusion, we aim to make machines have more natural and effective interactions with humans by making technology more responsive to emotional cues.

Keywords: Sound Emotion Recognition, Human-Computer Interface, Conflicting Emotions, Deep Learning Models, Clustering, Emotional Cues, Natural Interactions, Machine Responsiveness, Vocal Expressions.

Table of Contents

List of Figures

Chapter 1

Introduction

In this era of technological advancements, human-computer interactions have become a part of daily life. Along with improving user experience and others, HCI also aids in creating a natural and efficient interface for human-computer communication . Additionally, play a role in criminal interrogations particularly for analyzing behaviours, detection processes, improving communications and so on. Emotions are a core part of human interaction. A critical aspect of these interactions depends on machines' ability to understand human emotions and respond to it in real time. Sound Emotion Recognition (SER) can be used to bridge the emotional gap between humans and machines. SER enables systems to perceive and interpret emotions based on vocal expressions. While the existing SER models have been progressing significantly over time; there are areas in this field that are yet to be explored. Many models may struggle with the dynamic nature of human emotions, as one may be experiencing multiple emotions at the same time (i.e. fear and anxiety, fear and excitement, deep sadness and anxiousness and so on.). These real-life overlapping emotions make it challenging for the existing SER models to recognize and accurately classify emotions, which in turn leads to reduced model performance. This study aims to face these challenges by employing advanced clustering and deep learning techniques and models to accurately catch the nuanced emotions one might be experiencing through speech. This study will contribute in creating a more natural and emotionally intelligent interaction between humans and automated systems, by enabling machines to interpret such intricate emotions. Consequently, making these systems more accurate, sensitive and precise.

1.1 Background

Emotions are complex psychological and physiological states triggered by different stimuli. This section is concerned with defining what emotions are, what models are used to understand emotions and human expression of emotions.

1.1.1 Emotions, Moods, Feelings

Speech is one of the primary ways of conveying information, while emotions are fundamental human expressions of different experiences. Although closely related, the concepts of emotion, feeling, mood are distinct.

According to a study[14] emotions are intense, short-lived reactions triggered by internal or external stimuli (physiological and psychological). It does not occur consciously. In contrast, feelings are subjective experiences of an emotion. It is the interpretation of emotional responses. It occurs consciously. Lastly, moods are sustained emotional states with no specific triggers.

1.1.2 Emotional Models

Emotions can be categorised into 2 popular approaches. Firstly, the categorical (discrete) approach describes emotion as a discrete number of classes. One of the most popular theorists[2], listed 6 basic emotions: anger, happiness, surprise, sadness, fear and disgust. These are universal regardless of one's cultural and social background. Secondly, in Russel's 2D Model, the dimensional (continuous) approach describes emotions on a continuous scale of valence, which can be positive or negative, and arousal, which can be high or low. PAD (Russel's 3D Model, stands for Pleasure, Arousal, Dominance) is another widely used emotional state model, consisting of 3 dimensions: arousal, pleasure, dominance[1]. Discrete model distinctly categorises emotions giving them their own set of cognitive and psychological elements, whereas the dimensional model recognises the complexity of emotional experiences, influenced by many factors such as personal history, cultural background and so on. It uses 2D (arousal and valence) and 3D (arousal, valence and dominance) emotional space models.

1.1.3 Expression of Emotions

Humans use different channels to express emotions. According to another study[6], facial expressions can communicate multiple emotions without the need for verbalizing. Some of the universal nonverbal emotions expressed through facial features are anger, surprise, fear etc. In addition to facial expressions, vocal cues—including pitch, amplitude, and frequency of sound waves—play a crucial role in identifying an individual's emotional state, forming the foundation of SER. SER uses speech to recognize emotions; however, body language, consisting of gestures, postures, eyemovent, hand placement or movement further enhances the communication of emotions. Likewise, physiological signals such as skin conductance, electrocardiogram (ECG), blood volume pulse (BVP), heart rate and so on, can be used as modes of expression of emotions for individuals who suffer from mental or physical illnesses; where they are unable to communicate efficiently through facial expression, vocal cues or body language[17]. Figure 1.1 shows a breakdown of this.

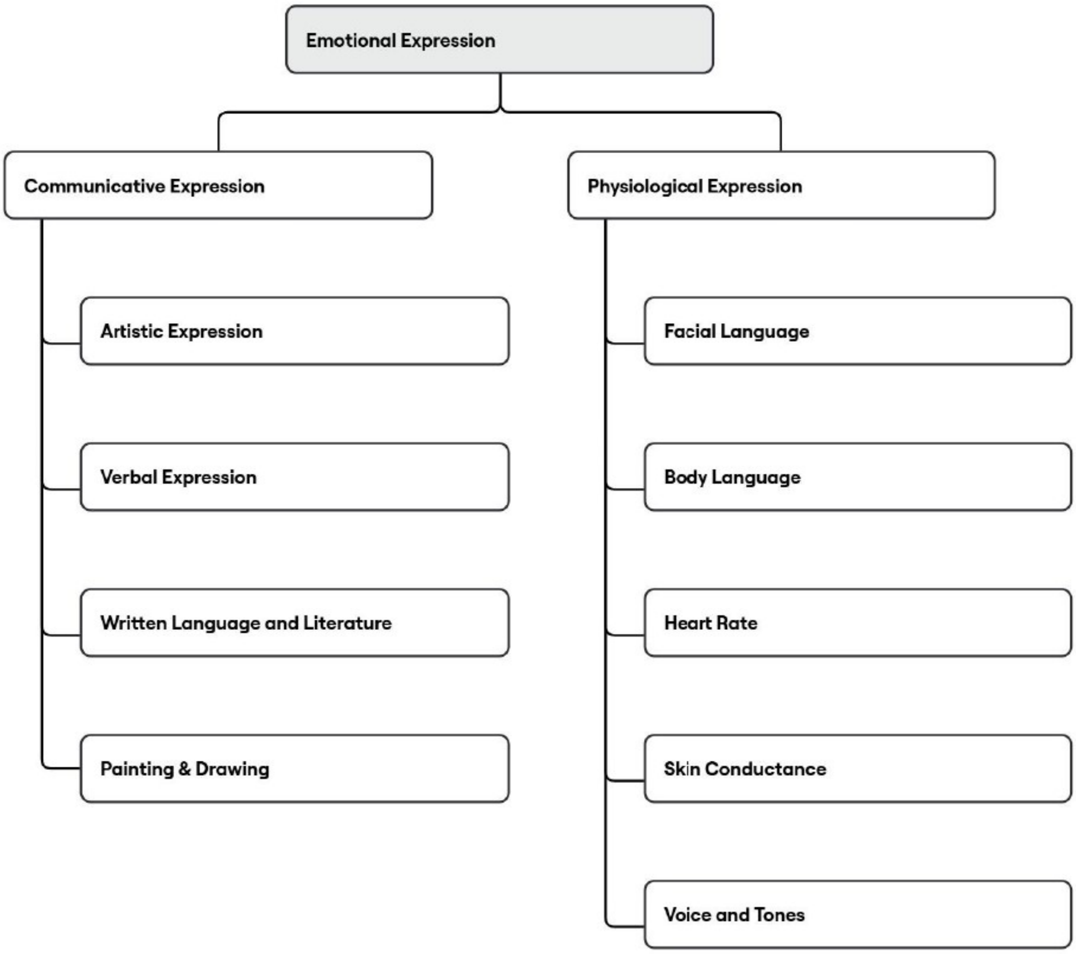

Figure 1.1: Human Emotional Expression

1.1.4 Emotion Recognition

There are two types of emotion recognition: non-automated recognition and automated recognition. Non-automated recognition refers to the natural human ability to perceive and interpret emotions. Researchers have developed structured techniques to assess emotions such as FACS (Facial Action Coding System), Geneva Emotion Wheel, SAM (Self-Assessment Manikin) along with psychological assessment questionnaires like PANAS (Positive and Negative Affect Schedule), BDI (Beck Depression Inventory), STAI (State-Trait Anxiety Inventory) and POMS (Profile of Mood States). In contrast, Automated recognition recognises emotions and interprets them with the help of machines. These systems can further be divided into 3 modes: unimodal, bimodal and multimodal. Unimodal takes into account 1 mode of expression, whereas Bimodal uses 2 modes of expression and Multimodal uses multiple modes of expression.[17]

1.2 Rational of the Study or Motivation

As technology advances and gets integrated into daily life, it becomes crucial for machines and humans to have natural and smooth communication, to improve user satisfaction, and ensure more intuitive and interactive computer interactions. Even though SER systems are high in demand and are useful in many aspects they still lack the ability to differentiate between the conflicting or overlapping emotions. Existing SER systems may face challenges classifying between frustration and determination or anxiety and excitement. A more adaptive and nuanced recognition technique is required to face said challenges.

1.3 Problem Statement

SER models can be indispensable in human-computer interaction, enabling human-machine communication to be emotional. However, traditional SER models struggle to identify conflicting and overlapping emotions, as most of the systems classify emotions into discrete categories such as happy, sad, angry. The models assume emotions to be static and independent. Although, in reality emotions are interdependent, dynamic, complex and multiple emotions can often be expressed simultaneously. For example, individuals may experience feeling both anxious and excited at the same time. Existing SER models, especially the models dependent on manually extracted features such as MFCCs, spectral features and so on fail to capture such emotional nuances, leading to misclassification and reduced performance. This research aims to use advanced clustering techniques in addition to deep learning models to identify said emotional nuances and create a model for better emotion recognition of conflicting emotions.

1.4 Objective

This study aims to solve the challenges of overlapping emotions faced by SER systems and optimizing the accuracy of said systems. The goal is to create a more robust and effective emotion recognition technique, in order to classify multiple emotions being expressed simultaneously. Key objective is analyzing limitations faced by SER models, developing multi-emotion classification approach, and evaluating its' effectiveness. Ultimately, this study aims to be able to engineer a more emotionally intelligent system based on SER for enhanced human-computer interaction.

1.5 Methodology in Brief

This research paper was based on reading many technical research articles and journals related to sound/speech emotion recognition. Firstly, the research focused on understanding the complexity of emotions. Also, how emotional nuances can be expressed using different channels. Moreover, different modes of emotional recognition. Secondly, review papers were analysed to look for research gaps and key challenges faced by currently existing SER systems. Lastly, different models were compared and contrasted to see which models performed better and what were their consequent drawbacks. The papers, articles and so on that were read for this study were chosen based on relevance, large number of citations and more recent publication. These strategically selected papers will help understand the recent developments and applications of SER. This will lay the foundation for future research and model development.

1.6 Work Plan

The work plan for this thesis paper has been divided into 3 distinct parts: pre-thesis 1, pre-thesis 2 and defense. In pre-thesis 1, a research gap is discovered. The problem is further researched and defined undertaking a comprehensive study of the problem. This paper comprises literature review which helps build a solid theoretical foundation in order to understand the problem and research objectives. Furthermore, pre-thesis 2 encompasses data collection, feature extraction and development of basic SER model in order to solve the challenge of detecting conflicting or overlapping emotions. Lastly, defense will involve the advancement and further development of the hybridized model made using clustering and deep learning models. The model will be evaluated depending on the suitable evaluation metric and compared with state of the art models.

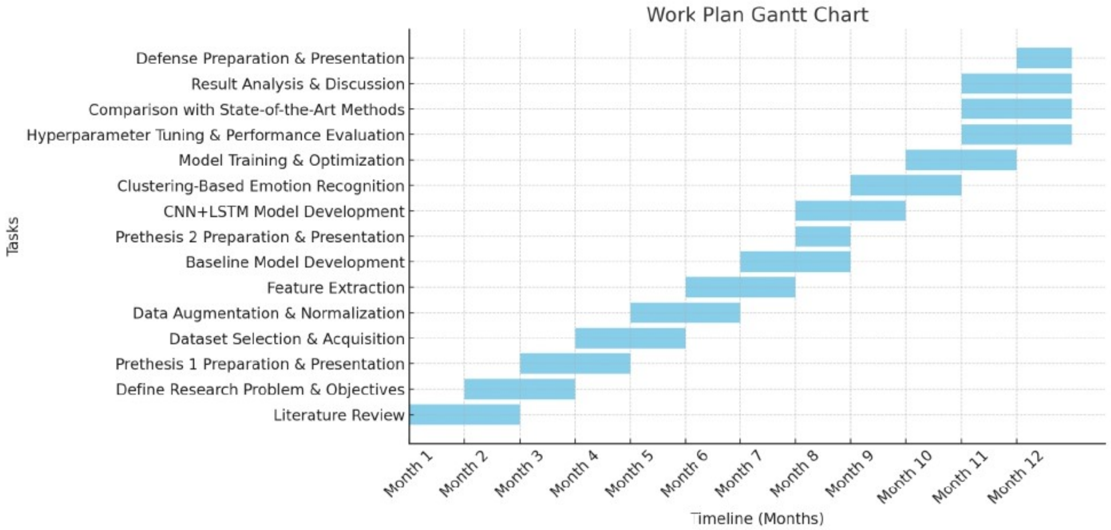

Figure 1.2: GANTT chart of work plan

Chapter 2

Literature Review

2.1 Sound/ Speech Emotion Recognition

A review on SER[15] states that it is a field of study that deduces human emotions from verbal expression. The system focuses on identifying and classifying voice input to various defined categories of emotion. It identifies the acoustic and linguistic features of a speech signal to capture these emotional states as such plays a crucial role in human-computer interaction (HCI) as well as behavioral analysis.

Definition of SER
It[16] is stated that speech emotion recognition is a part of speech processing which analyses speech signals and recognises patterns of speech such as prosody, pitch, frequency, rhythm in order to determine the emotional state of the speaker. Classifying emotions using sound/speech of an individual has various applications, majorly in human-computer interaction.

Relevance of Using SER
In today's day and age, increased dependency on technology gives rise to the need for better technical systems. People in recent years have become more reliant on technology, mostly human-computer interaction (HCI), artificial intelligence (AI) and so on, necessitating more emotionally-aware computer systems in order to enhance user satisfaction. SER could be applied in multiple fields, namely, emergency hotline, virtual assistants, criminal investigation etc..

2.2 Evolution of SER

2.2.1 Early Research and Traditional Methods

The research on sound emotion recognition had begun by using acoustic features and statistical classification models. Initially, models like support vector machine (SVM), hidden markov model (HMM) were used along with prosodic and spectral features (pitch, energy, mel-frequency cepstral coefficients). Although these methods did provide insights into emotion classification, they struggled to accurately identify the emotions in the presence of noise, speakers of different languages as well as in the cases of conflicting emotions.

2.2.2 Transition to Machine Learning

Due to the struggles faced by the traditional methods, Machine Learning (ML) approaches were introduced to SER. ML improved SER by allowing automated feature selection and classification. Models such as Random Forests and Gaussian Mixture Models (GMMs) demonstrated higher adaptability and accuracy. Even after these improvements, automatic feature selection is not completely reliable as it is highly dependent on feature engineering quality as well as dataset variability. As such, manually extracted features remain as a limitation.

2.2.3 Adoption of Deep Learning

The adoption of deep learning (DL) improved on the limitations previously faced by the earlier methods by eliminating the need for handcrafted features. Models such as Convolutional Neural Networks (CNNs) effectively extracted spatial patterns from spectrograms, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, modeled temporal dependencies in speech. Hybrid CNN-LSTM architectures improved the recognition accuracy as such improving generalization to real-world emotional speech. Compared to ML-based methods, DL models demonstrated a higher level of robustness and scalability, even though they required extensive data and computational resources.

2.2.4 Role of Benchmark Datasets

The availability of standardized datasets has been crucial in advancing SER. Corpora such as RAVDESS, IEMOCAP, Emo-DB, CREMA-D, SAVEE and TESS provide diverse linguistic and emotional variations, facilitating the development of more adaptable models. These datasets have enabled direct comparisons between ML and DL approaches, demonstrating the effectiveness of deep learning in handling complex emotional variations. SER has evolved from statistical methods to deep learning-driven approaches, significantly improving emotion recognition accuracy. However, challenges related to cross-corpus generalization, speaker variability, and computational efficiency remain key areas for further research.

2.3 Taxonomy of Methods in SER

The taxonomy in SER can be shown as follows:

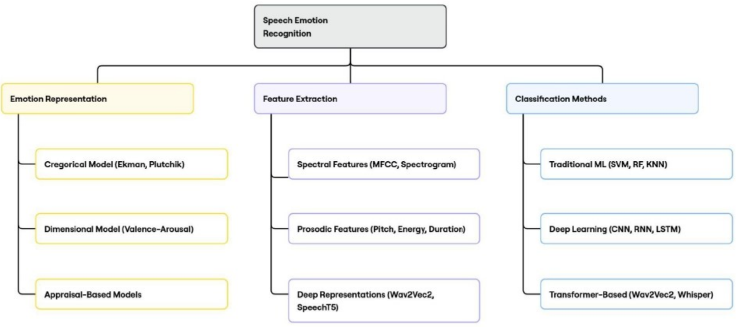
Speech Emotion Recognition
Emotion Representation
Feature Extraction
Classification Methods
Cregorical Model (Ekman, Plutchik)
Dimensional Model (Valence-Arousal)
Appraisal-Based Models
Spectral Features (MFCC, Spectrogram)
Prosodic Features (Pitch, Energy, Duration)
Deep Representations (Wav2Vec2, SpeechT5)
Traditional ML (SVM, RF, KNN)
Deep Learning (CNN, RNN, LSTM)
Transformer-Based (Wav2Vec2, Whisper)

Figure 2.1: Flow chart for SER taxonomy

2.4 Key Challenges in SER

The key challenges faced by SER systems, according to some studies[3]–[7], [13], were as follows:

1. **Dataset Limitations:** almost all the articles that were analysed in order to write this paper used the same datasets available. This can be addressed by data augmentation. A small dataset can lead to overfitting and limited generalizability.
2. **Noise and Distortions in Speech Data:** background noise and distortions reduce the quality of the data and introduce added complexities.
3. **Audio Feature Extractions:** extracting relevant audio features is crucial for effective SER
4. **Class Similarity and Misclassification:** overlapping emotions can cause 2 different emotional vocal cues to sound similar leading to an increased risk of misclassification.
5. **Scalability and Real-time Processing:** implementing operational SER systems that can process audio information remains highly challenging.
6. **Linguistic Variations:** Impact of cultural, personal, and contextual factors affect the differences between emotional tone and expression.
7. **Generalization:** Models performing poorly on unseen datasets or real-world scenarios
8. **Model Selection and performance:** different ML, DL models have different strengths and weaknesses. Model interpretability can also be a limiting factor.

2.5 Application of SER

According to multiple studies[4], [6], there are multiple applications for SER.
Practical applications:

1. **HCI:** SER enhances human computer interaction by allowing machines to identify and interpret human emotions for more natural communication.
2. **Healthcare:** SER can help with mental health assessments by using it during therapy session to monitor their emotional response
3. **Entertainment:** if SER is used in gaming or interactive media, understanding player emotions can lead to a higher user satisfaction.

Academic applications:

1. **Research & Development:** studying and using SER contributes in the field of interactive computing and emotional AI. Consequently, advancing machine learning, signal processing and cognitive psychology.

2. **Emotion Theory:** exploration of SER helps improve the existing theory on emotions, which in turn advances studies on psychological and linguistic studies depending on emotional states through vocal expression.

Societal impact:

1. **Communication improvement:** SER can improve social interactions by recognizing emotions.

2. **Educational enhancement:** SER can further help teachers understand students' emotional state and create a work plan that will best enhance their performance.

3. **Inclusion and accessibility:** SER can be used to aid individuals with disabilities by providing them with a smoother and better communication through emotionally aware systems.

SER has become very popular due to its multiple applications. However, many SER systems face challenges to accurately identify and interpret emotions due to conflicting and overlapping emotions at a time, limiting their effectiveness. This complexity makes it difficult for traditional categorical emotion classification making it insufficient and thus requiring more adaptive and nuanced recognition techniques.

2.6 Existing Model Architecture

One of the studies[13] uses a one-dimensional convolution network (Conv1D). It is a deep learning technique that processes audio features to learn the complex patterns using convolution layers. It is a good model for capturing local, temporal dependencies in audio signals. Additionally, they also use random forest (RF). It is a tree-based machine learning algorithm that works by constructing multiple decision trees. Using it alongside feature pruning increases performance, provides dimensionality reduction, and reduces computational cost. The comparative analysis of Conv1D and RF provides insight on the quality of performance of the two models. Consequently, becoming the highlight of the study.

In comparison, another study[5] uses an ASR model at its core for audio-to-text mappings, utilizing dilated convolution and gated convolutional units. Also, in this study the authors use the ASR model as a feature extractor for emotion recognition tasks. This is done by computing mean activations for different layers. Furthermore, a linear regression model is used to predict the arousal and valence values from the extracted features. This approach made their study novel by utilizing ASR as feature extractor, exploration of layer contributions, and end-to-end training considerations.

Conversely, another[6] uses deep learning models like RNN (Recurrent Neural Network) and LSTM (Long Short-Time Memory). RNN works by keeping a record of old information which makes it suitable for capturing temporal information or dependencies of sequential data such as speech signals. LSTMs are extended versions of RNN. LSTM allows for long-term dependency learning. In this study, the authors also used SVM (Support Vector Machine) for classification. Even though, SVM works with highly dimensional data it is a more traditional model used for SER.

In addition, CNN (Convolution Neural Network) can also be a good model to use when using automated feature extraction, i.e. spectrograms or mel-spectrograms. Spectrograms are computed using STFT (Short-Time Fourier Transform) and mel-spectrograms are computed by applying mel filter banks on spectrograms. Lastly, it was discovered that using a hybrid model enhances model performance. This hybrid approach introduces novelty in their study. Hybrid models can be created by integrating attention mechanisms in RNNs and LSTMs; consequently, aiding the model focus on emotionally rich parts of speech signal. It was also discovered that multimodal models classified emotions more accurately than unimodal or bimodal models as they considered multiple modes of expression

Similarly, a different source[7] also reviews various deep learning techniques, including Deep Belief Networks (DBNs), and Deep Boltzmann Machines (DBMs) along with more commonly used Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks. CNNs extracts spatial features directly from input data through convolutional layers. These deep learning techniques leverage automated feature extraction, which in turn helps RNNs and LSTMs identify temporal dependencies in speech signals. Consequently, enhances long-term dependency recognition. They introduce novelty through hybridization of multiple models. The implementation of their model was executed using Python, leveraging libraries such as TensorFlow, Keras, and Librosa, among others.

In contrast, another research[3] used Deep Neural Network (DNN) and Stochastic Gradient Descent (SGD). DNN uses a combination of convolution, pooling and fully connected layers; whereas, SGD is an optimization approach. This leads to a more efficient training and accurate classification, updating the model parameter in order to reduce the loss function. The use of automated feature extraction is the highlight of this study. To accomplish this they enhanced temporal resolution by using segmented 20 ms frames and implemented the model using python leveraging libraries such as Keras and Theano.

Subsequently, another study[4] also used a DNN. In contrast, the author hybridizes DNN (Deep Neural Networks) with KNN (K-Nearest Neighbor) rather than SGD (Stochastic Gradient Descent). As mentioned before, DNN uses convolution and multiple fully connected layers in order to solve complex problems. Whereas, KNN is a machine learning algorithm that classifies data depending on the distance between the data points of the number of neighbours chosen. KNN was used to improve accuracy, also, as a secondary validation method. Hybridization of ML and DL, where incorporating statistical features enhanced the performance of the model, and made it stand out amongst all the others.

On the other hand, a different study[8] used CNN and LSTM. They used 3 convolution layers for the CNN model for spectrograms and mel-spectrograms, with different feature maps and dropout layers each, and softmax function was used for output. Furthermore, they used a variant of LSTM, Bi-LSTM, where the layers are bidirectional and fully connected, including the softmax layers. This BLSTM model uses feature vectors for input and trains the models using sequence-to-one modelling. They have also used an ensemble of the two models where they used CNN followed by BLSTM, in order to automate feature extraction using CNN and feeding it directly to BLSTM. This is done to reduce the inefficiency of modelling the complexities of emotional nuances in a single utterance. Here, the CNN is used to extract high-level features, whereas the BLSTM is used to model the utterance-level long term dependencies subsequently. Finally, to compare and classify the models, a cross-entropy loss function was used along with the Adam optimiser to ensure the model is not too computationally expensive.

In a subsequent paper[10], supervised models - Regression and Artificial Neural Networks (ANN) - were used on the IADS dataset. The similarity of features in this dataset are compared using Pearson's r correlation coefficient, which were then used as inputs in the regression and ANN models. The emotional dimension of valence and arousal were predicted using the output from the following models. For regression analysis, RMSE was used, where the lower the RMSE value, the better the performance of the model. To ensure unbiasedness, principal component analysis (PCA) was used on data with and without dimension reduction in order to use five and ten fold cross validation. In contrast, they also used a two-layer feed forward ANN , consisting of one hidden layer. Additionally, the arousal and valence for this model were determined using two separate ANNs. The training-validation-testing ratio for the data set used was 70-15-15.

Furthermore in a differnt study[9], the authors study different models of SER including traditional as well as the ones created in recent times. Different feature extractors, for example, openSmile, Voice Quality Features, MFCCs, Pulse Code Modulation (PCM), Prosodic Features and more traditional Statistical Methods were also explored. And it was stated that the type of extractor used depends on the specific context it is to be used in. The feature extracted using these methods are used as inputs into models such as Hidden Markov Models (HMM) - which is a more traditional model , SVMs, ANNs, CNNs, RNNs - especially LSTM networks, and Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE) - which are more advanced architectures used in the more recent years. HMM is a statistical model which consists of a finite set of hidden states, transitions (probabilities of moving from one state to another), emissions (probabilities of an observable output), and initial probabilities (the starting probabilities in each of the hidden states). On the contrary, GANs and VAEs are both generative deep learning models. GAN works by training two neural networks (generator and discriminator) at the same time in a competitive manner, and VAEs consist of an encoder and a decoder -which are trained to minimise the difference between the input and the decoded output data.

2.7 Model Comparison

One of the studies[6] employed 2 models RNN and SVM and compared it against MLR. The features extracted were MFCC and MS (Modulation Spectral). Feature selection (FS) was used to find the most relevant feature subset. It was observed that the RNN model when run on the Spanish database obtained an accuracy of 94%, performing better than SVM or MLR. It was also observed that employing speaker normalisation improved model performance when the model was run on the Berlin database but had insignificant effect when run on the Spanish database. Moreover, feature selection enhanced performance by decreasing dimensionality. On the other hand, another[5] explored models like ASR model with transfer learning and Linear Regression for emotion estimation. The ASR system was used as a features extractor which in turn feeds the features into the linear regression model that estimates the valence-arousal values. Since, ASR- feature extractor was used instead of using manually extracted features which aided the model to outperform previously engineered models. This study was conducted on an IEMOCAP dataset, and the study also found that certain parts of the ASR system correspond differently to certain emotional expressions. From that fact, it can be inferred that an ASR based model is an unsupervised learner, where it naturally picks up on emotional trends in speech. In addition, a different paper[13] explored ML and DL models like Conv1D and RF. the extracted features used were MFCCs, chromograms, mel-spectrogram, tonnetz, ZCR and more. After feature selection was used, RF achieved better accuracy than Conv1D (69%). Moreover, it had a precision of 72% for fear and recall of 84% for calm. RF performed well whereas Conv1D misclassified emotions like anger, disgust and fear, with happy, neutral, sad respectively due to overlapping emotional features.

A different research[3] observed an accuracy of 96.97% after employing convolutional and fully connected layered DNN. Even though the model performed well it still faced challenges. The lack of feature selection and the model needing substantial hyperparameter tuning makes it dependent on larger datasets. Similarly, another study[4], also used a DNN model, although it was employed along with KNN as a hybrid model, in order to detect distress signals which consequently improves classification accuracy. Nevertheless, the hybridization introduced complexity which in turn raised computational cost. Alternatively, a different research[7], used deep learning techniques to automate feature extraction. In conclusion, it was observed that there was always a trade off between accuracy, precision computational cost and modular complexity. This highlights the challenge of balancing these factors in order to engineer an efficient system for optimal SER.

In a paper from a different set of studies, authors[8] found that a hybridized model of CNN and BLSTM outperformed the models on only CNN or BLSTM. An accuracy of 82.35% was obtained when MFCC features were applied on EmoDB dataset; on the other hand, 50.05% accuracy was obtained using mel-spectrogram for IEMOCAP dataset. For the former dataset, the confusion matrix showed that happy speech was not detected as accurately, which could be either because there were not enough instances of happy speech in the dataset or due to happy and angry being part of the same category (arousal), the subtle differences were not captured accurately. In contrast, the confusion matrix for the latter dataset showed that the model successfully recognised all the four basic emotions, despite the more natural instances in this dataset. They speculated that the staggering difference (82% vs 50%) in the success rates between the two datasets can be due to the naturalness of the emotions elicited by IEMOCAP.

A different research[10] showed that the regression model had unexpectedly poor outcomes in Audio Emotion Recognition, in contrast to the previous research findings which suggested regression models work well for MER. Their ANN model showed variance for 64.4% and 65.4% for arousal and valence prediction respectively.

Finally, an alternate paper[9] concludes with the finding that CNNs are the better at emotion recognition due to its higher low-level and short term discriminative capabilities. This, combined with the addition of LSTM networks not only allows the model to identify long-term paralinguistic patterns, it also has shown higher capabilities of speaker-independent emotion recognition.

2.8 Recent Trends

The analysis of review papers suggests that SER has made significant advancements in recent years. According to authors[12], traditional SER models use manually extracted features such as MFCCs, pitch, spectral features whereas modern approaches depend on deep representation learning to learn features directly from raw audio using neural networks reducing the need for feature extraction; consequently, reducing computational cost of a model. On the other hand, employment of self-supervised learning (SSL) for feature extraction have also been observed. Additionally, models with SSL (unsupervised learning) can easily work on unseen dataset without needing additional fine-tuning. VAEs (Variational Autoencoders) and GANs (Generative Adversarial Networks) are used to modify the dataset for better emotion recognition (i.e. data augmentation and emotion transformation). To introduce further generalization in recent models, transfer learning and domain adaptation have been used. A further study[11] found how incorporating attention mechanisms with SER models (i.e. RNN, LSTM, Transformer models) improves efficiency as it focuses on the more meaningful parts of speech.

2.9 Gaps in Research Field

While there has been progress, there are still some unresolved issues. Present models do not consistently recognize tonal variations across languages and accents. Present models also become less effective in noisy environments, limiting areas of real-world application. Deep learning-based SER systems also lack interpretability, leaving users with no explanations for model decisions and, as a result, making it difficult to earn users' trust[18]. Systems also fail to consistently detect conflicting emotions, and mixtures of emotional expressions have resulted in misclassification and inaccuracy. For SER to be applied practically, these limitations must be overcome.

Chapter 3

Proposed Methodology

3.1 Design Process or Methodology Overview

This paper focuses on building an objective-focused emotion classification model using an iterative methodology. The goal of this model is to be able to classify emotions into a primary set and identify secondary, tertiary or more conflicting emotions from an audio input. The foundational architecture utilizes a two-channel CNN for the extraction of spectral and temporal features, a GRU with attention mechanisms to highlight emotionally significant speech segments, and a BiLSTM to capture bidirectional temporal dependencies. The base model is based off of an ensemble model that uses the strength of all three architectures to classify the primary set of emotions. Additionally, this model is further adapted to include fuzzy c-means clustering in order to classify emotions into secondary or tertiary sets, proving the existence of conflicting emotions. It should be noted that alternate base models were also tested. The main constraint to consider are as follows:

- Dataset limitation:
  - Small sample sizes such the SAVEE dataset, leads to limited diversity
  - Acted emotions such as the ones in CREMA-D and RAVDESS lead to fewer nuanced real-life emotions.
  - Biased learning in most datasets leads to an imbalance of classes
  - Lack of demographic diversity such as TESS, which only includes older female speaker
- The cross-corpus setting the model was run in revealed speaker-dependent splits often leading to overfitting and poor performance
- Voice characteristics varying across datasets impact transfer learning
- Inconsistent emotion labels across the different datasets lead to decreased performance
- The framework increases resource requirements thus computational cost due to it being an ensemble model caused by multiple model branches.

The multiple tools used in this model making include:

- Python
- TensorFlow
- Librosa
- Matplotlib
- Scikit-fuzzy

3.2 Preliminary Design or Design (Model) Specification

To determine the most effective architecture, several models were tested, including independent CNNs (for extracting spatial features), BiLSTMs (for capturing bidirectional temporal dependencies), and a hybrid CNN-LSTM that integrates both methodologies. The standalone models failed to perform up to the mark and had low accuracy rates. The hybrid model turned out to be the next best base model with an accuracy rate of 41.62%, as it leverages the strengths of both the standalone models. Simulations for each model were conducted using Python and TensorFlow deep learning frameworks. Furthermore, Librosa was used to extract features like the MFCCs, Mel-spectrograms and so on. An 80-20 split was used to train and test the models. Moreover, stratified samples were used for validation for the preservation of distribution of emotional classes. In order to avoid overfitting due to the varying size of the datasets used, early stopping and learning rates schedulers were also used. The performance of the models were evaluated using accuracy, F1-scores, confusion matrices, and training-loss curves. The base model was selected based on the accuracy and performance rate of varying models. Furthermore, a novel conflict-resolution metric was introduced. The ensemble model used as the base model had the highest accuracy rate and hence was selected to be further adapted with fuzzy-c logic. The base model classifies the primary emotions and fuzzy-c means clustering is used to then find the conflicting emotions found under each emotion umbrella. Figure 3.1 shows an example of the output of said structure.

Metric	**Value**	**Description**
Primary Emotion	Neutral (0.453)	Detected as the strongest emotion with confidence score 0.453
Secondary Emotion	Sadness (0.218)	Detected as a secondary/conflicting emotion with confidence score 0.218
True Label	Neutral	Ground truth label assigned to the data
Conflict Score	0.547	Novel metric indicating the level of emotional conflict
Fuzzy Cluster	2	Cluster ID assigned by fuzzy c-means, denoting overlapping emotional regions

Table 3.1: Emotion Metrics and Their Interpretation

3.3 Data Collection

The datasets utilized were secondary data obtained from Kaggle. The standardized datasets include SAVEE, RAVDESS, CREMA-D, and TESS, which contain audio files (.wav files) categorized under emotionally distinct labels (such as sadness, fear, happiness, etc.). All datasets underwent preprocessing to convert them into numerical feature arrays (MFCCs, Mel-spectrograms, and Chroma). A stratified sampling technique was applied to enhance model robustness and minimize bias during both training and evaluation, thereby maintaining the emotional distribution. In cross-corpus evaluations, entire datasets were reserved for testing, while models were trained on a different dataset. This methodology emulates real-world deployment scenarios, where training and testing conditions differ. It highlighted the importance of strong generalization ability in the final model.

3.3.1 Data Cleaning

To maintain data integrity and quality, a thorough preprocessing procedure was implemented. Although missing values were minimal in the standardised datasets, failure during feature extraction of audio files led to their identification and removal. As invalid samples were completely excluded, there was no need for imputation. Similarly, duplicate samples were also excluded completely. Furthermore, Z-score analysis was performed for outlier detection. Samples were flagged as outliers if their feature values exceeded ±3 standard deviations. All the files were resampled to 16kHz to ensure compatibility during training.

3.3.2 Data Transformation

For effective model input features extracted were then normalised and encoded. Min-max scaling was performed to normalise values between 0 and 1, which lead to improvement in gradient stability. Even though it wasn't used in the final implementation, z-score standardization was also evaluated. Both one-hot encoding and label encoding were used for categorical labels.

3.3.3 Data Integration

Multiple datasets were used and processed individually but not merged into a single unified training dataset. Cross-corpus evaluations were also performed to test generalisation where TESS was kept as a constant testing set whereas others were used as training sets. Figures [start-end] demonstrates the generalization gaps caused by this approach. Consequently, no joint operations or data integrations were used. However, feature matrices being zero-padded or truncated to a fixed shape ensured alignment across different types of inputs which lead to better model compatibility. Naming conflicts between file structures, such as emotion labels like ”neutral” and ”calm,” were resolved by mapping to a consistent emotion label taxonomy in all datasets.

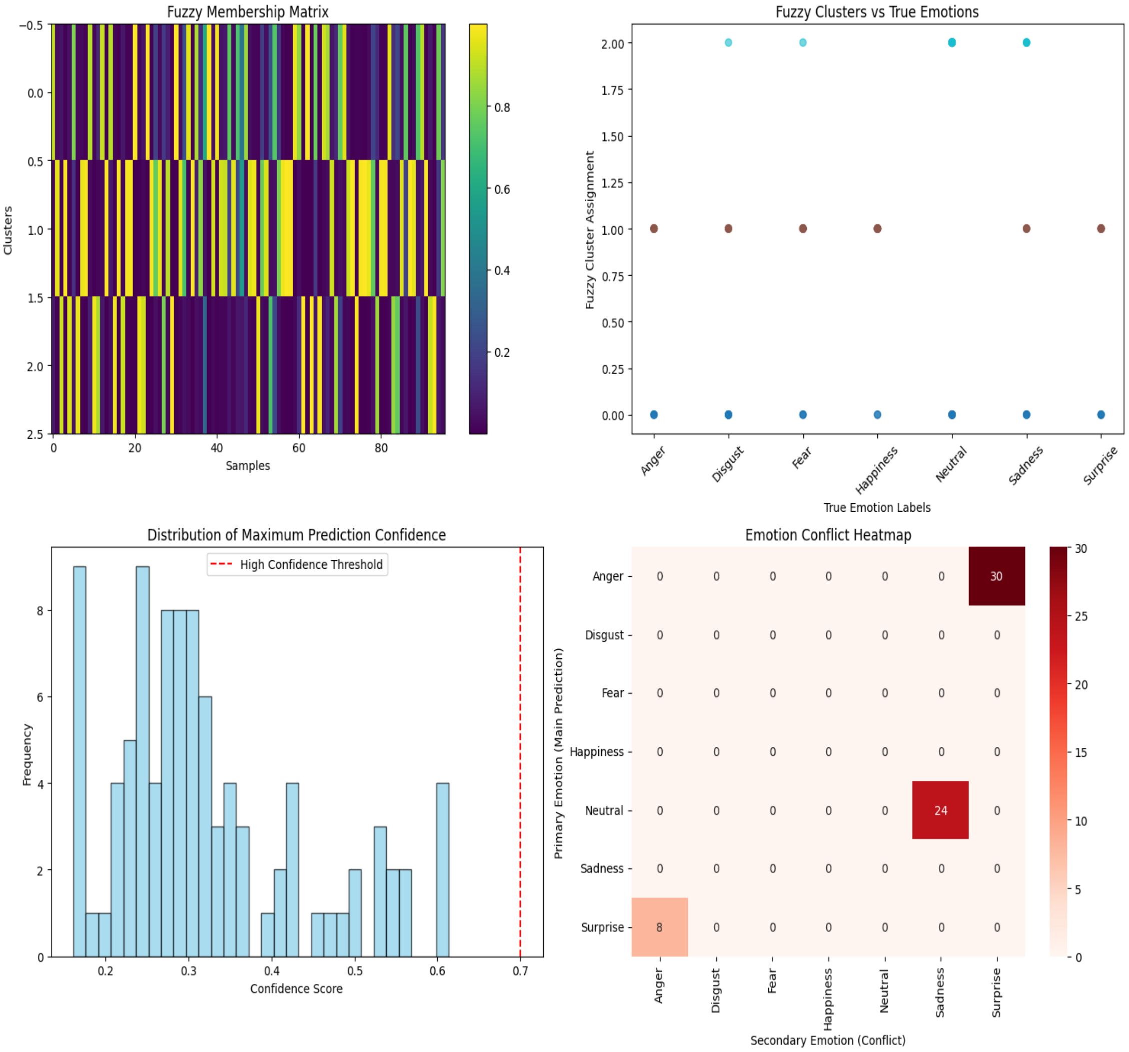

Figure 3.1: Confusion matrix, accuracy/loss curves, and fuzzy heatmap in preprocessing for SAVEE

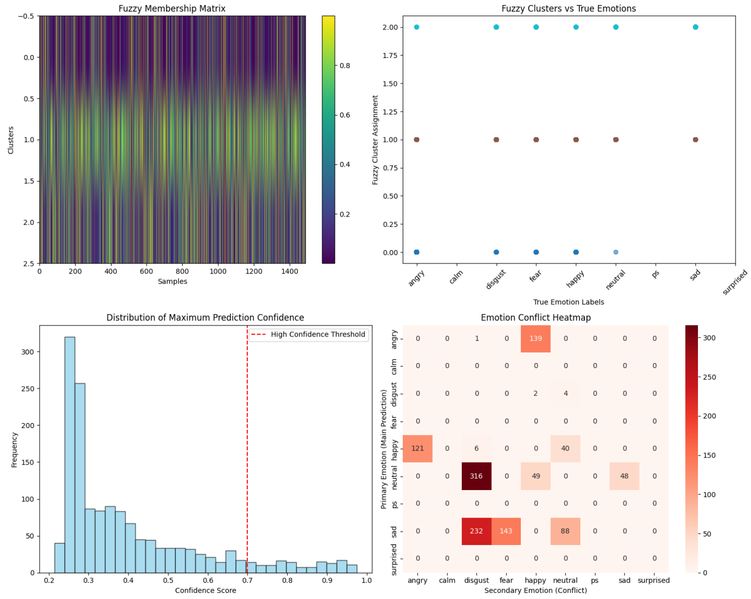

Figure 3.2: Confusion matrix, accuracy/loss curves, and fuzzy heatmap in preprocessing for CREMA-D

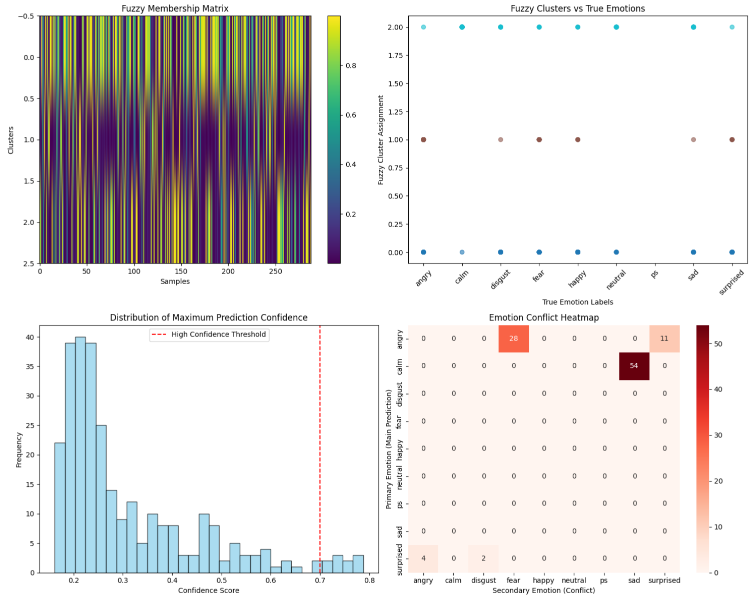

Figure 3.3: Confusion matrix, accuracy/loss curves, and fuzzy heatmap in preprocessing for RAVDESS

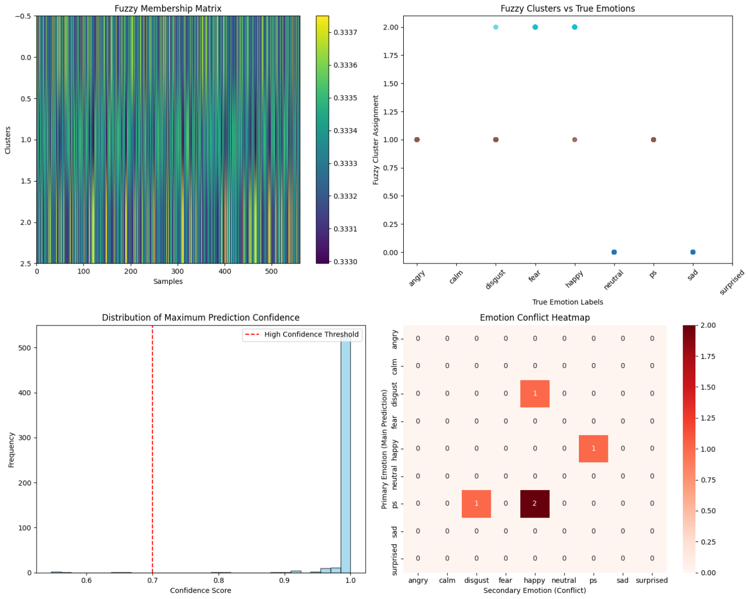

Figure 3.4: Confusion matrix, accuracy/loss curves, and fuzzy heatmap in preprocessing for TESS

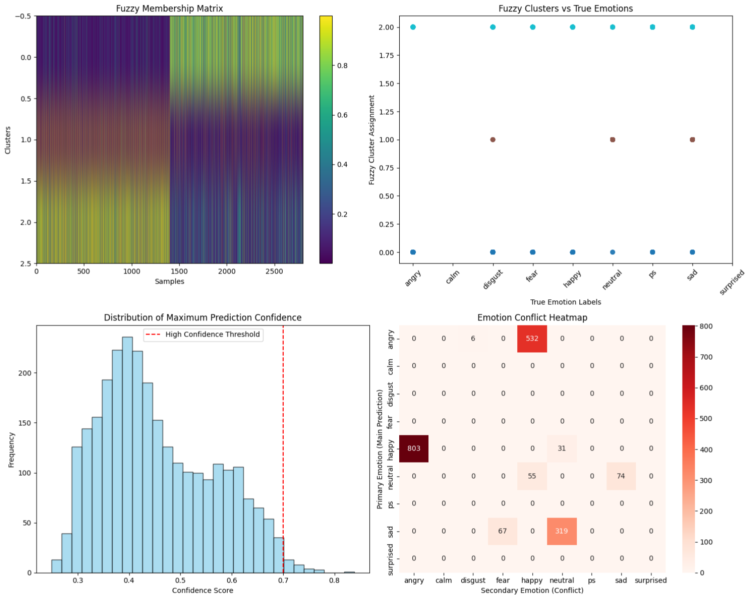

Figure 3.5: Confusion matrix, accuracy/loss curves, and fuzzy heatmap in preprocessing for cross-corpus CREMA-D with TESS

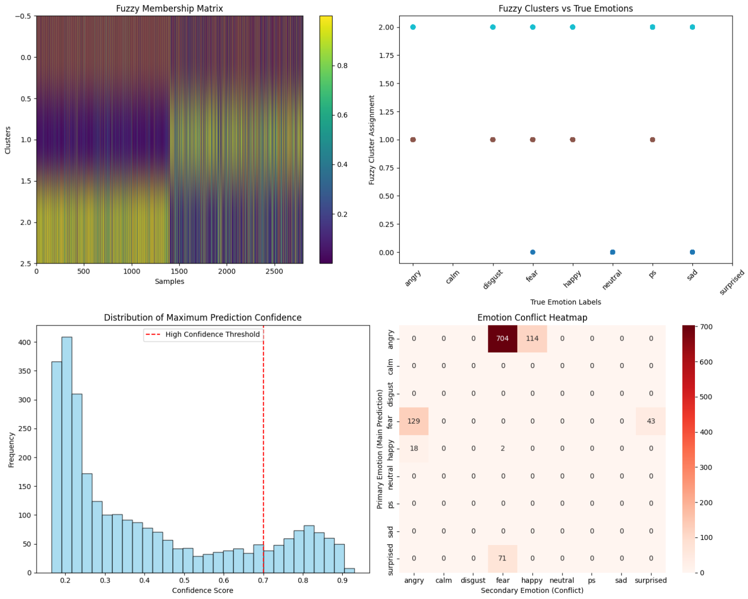

Figure 3.6: Confusion matrix, accuracy/loss curves, and fuzzy heatmap in preprocessing for cross-corpus RAVDESS with TESS

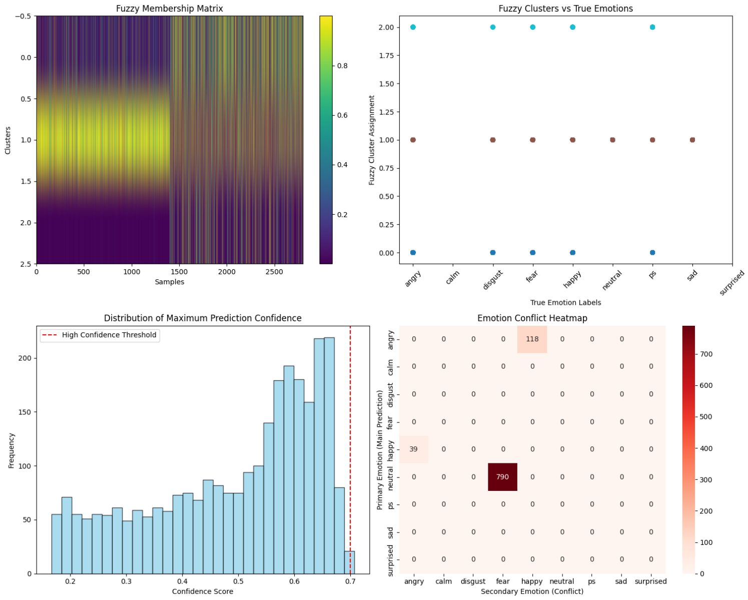

Figure 3.7: Confusion matrix, accuracy/loss curves, and fuzzy heatmap in preprocessing for cross-corpus SAVEE with TESS

3.3.4 Data Reduction

Preprocessing such as data reduction was done within the model to manage complexity and handle overfitting. Dimensionality reduction techniques like PCA, LDA and so on were explicitly employed. Although, model performance was compared while performing feature selection. For example, we compared MFCC-only inputs with multi-input configurations that included MFCC, Chroma, and Mel-Spec. This practical selection of inputs served as a way to manually reduce features. The use of global pooling and attention mechanism within the model reduced dimensionality during forward propagation, ensuring the model retains only the emotionally salient features. Consequently, improving generalisation and reducing training time.

3.3.5 Summary of Preprocessed Data

After preprocessing, the final input data for each sample included three types of features:

- MFCCs: 40 coefficients, 173 time steps
- Mel-Spectrograms: 128 bands, 53 frames
- Chroma: 12 dimensions, 173 frames

The sizes of the datasets were consistent with the original corpora, as no further filtering or augmentation was necessary beyond the initial preprocessing steps. The estimated sample counts for each dataset were as follows:

- SAVEE: 480 samples
- CREMA-D: 7,442 samples
- RAVDESS: 1,440 samples
- TESS: 2,800 samples

All data were processed into being fully numeric. Features were normalized and prepared for deep learning input. Categorical labels were encoded. Data readiness was ensured by validating all shapes, checking for label integrity, and confirming dataset splits.This prepared the dataset for modeling with high consistency and low redundancy.

3.4 Implementation of Selected Design

The final model was implemented in python using frameworks such as TensorFlow and Keras. Librosa was used to handle preprocessing and feature extraction while Matplotlib, Seaborn and scikit-learn were used for visualization and evaluation. The architecture contained 3 parallel branches 2-channel CNN was used for mel-spectrogram and chroma whereas GRU-BiLSTM with attention layer was used for MFCCs, which were then passed through dense layer for classification. Early stopping, dropout regularisation, and learning rate scheduling were used during training and a stratified 70-15-15 data split was maintained ensuring balance of classes and fair evaluation. Additionally, Fuzzy c-means clustering post initial classification in order to assess overlap and find conflicting emotions. Performance of the model was evaluated using accuracy, F1-score, loss curves and fuzzy entropy visualisation.

Chapter 4

Result Analysis

For all the datasets, a confusion matrix along with training accuracy and loss curves were used to assess the performance of the model. The confusion matrix illustrates the model's ability to differentiate between various emotion classes, with the diagonal entries representing accurate predictions. The accuracy and loss curves depict the model's learning trajectory across epochs, where stable and converging lines indicate successful training and generalization.

Fig 4.1 shows the final model evaluation for SAVEE. The performance of the SAVEE dataset is constrained by the limited size of the dataset and the insufficient diversity among speakers. The confusion matrix indicates a significant level of confusion between the categories of ”neutral” and ”sad,” which could be attributed to subtle acoustic similarities. Furthermore, the accuracy and loss curves imply an early convergence, revealing indications of overfitting as the validation loss stabilizes while the training accuracy persists in increasing.

Fig 4.2 shows the final model evaluation for CREMA-D. The dataset demonstrates a moderate level of performance attributed to its increased size and the range of emotional intensity it encompasses. Nevertheless, the confusion matrix reveals misclassifications among the emotions of ”fear,” ”disgust,” and ”sadness,” which can be attributed to overlapping nuanced expressions. The accuracy and loss curves indicate a consistent improvement along with a minor generalization gap, suggesting that while the model gains from the diversity present in the dataset, it continues to face challenges with emotions that are closely related.

Fig 4.3 shows the final model evaluation for RAVDESS.The model demonstrates strong performance on RAVDESS, especially in identifying high-arousal emotions such as ”anger” and ”happy.” There are instances of misclassification between ”calm” and ”neutral,” which is evident in the confusion matrix. The training curves exhibit a smooth and consistent pattern, suggesting effective learning. Nevertheless, the studio-controlled characteristics of the dataset may lead to the model's suboptimal performance in real-world, noisy environments.

Fig 4.4 shows the final model evaluation for TESS. The model demonstrates optimal performance on TESS, facilitated by distinct and exaggerated speech patterns from older female speakers. The confusion matrix indicates a minimal error rate, while the accuracy and loss curves reveal swift convergence and a low validation loss, implying a robust model fit. Nevertheless, the dataset's homogeneity and artificial clarity could restrict its applicability for generalization outside of controlled environments.

Fig 4.5 shows the final model evaluation for cross-corpus of CREMA-D with TESS. When the model is trained on CREMA-D and subsequently tested on TESS, it exhibits a significant decline in performance. The confusion matrix highlights misclassifications, especially between the categories of "happy" and "neutral," indicating that the model encounters challenges with the older female voices in TESS after being trained on the more varied and younger speaker demographic of CREMA-D. Furthermore, the fuzzy clustering analysis uncovers considerable emotional overlap, while the accuracy and loss curves demonstrate the model's struggle to adjust to the new vocal patterns.

Fig 4.6 shows the final model evaluation for cross-corpus of RAVDESS with TESS. This cross-corpus evaluation shows a moderate decline in performance. The confusion matrix reveals mistakes among the categories of "calm," "neutral," and "sad," which can be attributed to variations in tone and demographic factors. Although the training curves exhibit stability, the generalization gap increases, underscoring the inadequacy of studio-recorded training data in representing the expressive characteristics found in TESS's recordings of older females.

Fig 4.7 shows the final model evaluation for cross-corpus of SAVEE with TESS.This combination results in the least effective cross-corpus performance. The model, which is exclusively trained on the limited male speaker data from SAVEE, does not successfully generalize to the female-oriented emotional cues present in TESS. The confusion matrix exhibits significant dispersion, and the fuzzy heatmap reveals extensive emotional ambiguity. The training curves suggest overfitting to the source dataset, highlighting the necessity for a more diverse and balanced training dataset.

Table 4.1: Classification Report of the Model

Label	**Precision**	**Recall**	**F1-Score**	**Support**
angry	0.28	0.48	0.36	400
disgust	0.00	0.00	0.00	400
fear	0.00	0.00	0.00	400
happy	0.19	0.74	0.31	400
neutral	0.86	0.40	0.54	400
ps	0.00	0.00	0.00	400
sad	0.49	0.48	0.48	400
Accuracy			0.30	2800
Macro Avg	0.26	0.30	0.24	2800
Weighted Avg	0.26	0.30	0.24	2800

images/figures/Model1+Savee.png

Figure 4.1: Model Accuracy, Model loss, Confusion matrix and Classification report of SAVEE run on Final Model

images/figures/Model1+Crema-D.png

Figure 4.2: Model Accuracy, Model loss, Confusion matrix and Classification report of CREMA-D run on Final Model

images/figures/Model1+Ravdess.png

Figure 4.3: Model Accuracy, Model loss, Confusion matrix and Classification report of RAVDESS run on Final Model

images/figures/Model1+Tess.png

Figure 4.4: Model Accuracy, Model loss, Confusion matrix and Classification report of TESS run on Final Model

images/figures/Model1+crema-d+tess.png

Figure 4.5: Model Accuracy, Model loss, Confusion matrix and Classification report of cross-corpus CREMA-D with TESS run on Final Model

images/figures/Model1+ravdess+tess.png

Figure 4.6: Model Accuracy, Model loss, Confusion matrix and Classification report of cross-corpus RAVDESS with TESS run on Final Model

images/figures/Model1+savee+tess.png

Figure 4.7: Model Accuracy, Model loss, Confusion matrix and Classification report of cross-corpus SAVEE with TESS run on Final Model

Chapter 5

Conclusion

This study explored the evolution of SER, addressing key limitations in existing systems, particularly their struggle to detect overlapping and conflicting emotions. To overcome these challenges, a hybrid deep learning model was proposed that integrates CNNs, GRU-BiLSTM with attention mechanisms, and fuzzy c-means clustering. Through detailed preprocessing, strategic model design, and cross-corpus evaluations, the model demonstrated improved accuracy in recognizing both primary and nuanced emotional expressions. However, issues such as dataset imbalance, limited demographic representation, and reduced generalizability across diverse corpora remain. Despite these challenges, the proposed approach offers a meaningful step toward emotionally intelligent human-computer interaction. Future research can build on this foundation by expanding dataset diversity, incorporating multimodal signals, and enhancing model interpretability for real-world applications.

Bibliography

[1] J. A. Russell, “A circumplex model of affect,” *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[2] P. Ekman, “An argument for basic emotions,” *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992. DOI: 10.1080/02699939208411068. [Online]. Available: https://doi.org/10.1080/02699939208411068.

[3] P. Harar, R. Burget, and M. K. Dutta, “Speech emotion recognition with deep learning,” in *2017 IEEE International Conference on Signal Processing and Integrated Networks (SPIN)*, 2017, pp. 137–140. DOI: 10.1109/SPIN.2017.8049931.

[4] K. Tarunika, R. Pradeeba, and A. P, “Applying machine learning techniques for speech emotion recognition,” Jul. 2018, pp. 1–5. DOI: 10.1109/ICCCNT.2018.8494104.

[5] N. Tits, K. E. Haddad, and T. Dutoit, “Asr-based features for emotion recognition: A transfer learning approach,” in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 48–52. DOI: 10.18653/v1/W18-3307. [Online]. Available: https://aclanthology.org/W18-3307/.

[6] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cléder, “Automatic speech emotion recognition using machine learning,” in *Social Media and Machine Learning [Working Title]*, IntechOpen, 2019. DOI: 10.5772/intechopen.84856. [Online]. Available: https://hal.science/hal-02432557.

[7] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: A review,” *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019. DOI: 10.1109/ACCESS.2019.2936124.

[8] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, “Deep learning techniques for speech emotion recognition: A review,” in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2019, pp. 1–6. DOI: 10.1109/RADIOELEK.2019.8733432.

[9] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep learning techniques for speech emotion recognition, from databases to models,” *Sensors*, vol. 21, no. 4, p. 1249, 2021, ISSN: 1424-8220. DOI: 10.3390/s21041249. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1249.

[10] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, “Supervised machine learning for audio emotion recognition,” *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021. DOI: 10.xxxx/xxxx.

[11] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10101163. [Online]. Available: https://www.mdpi.com/2079-9292/10/10/1163.

[12] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, 2023. DOI: 10.1109/TAFFC.2021.3114365.

[13] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," *PLOS ONE*, vol. 18, no. 11, pp. 1–13, Nov. 2023. DOI: 10.1371/journal.pone.0291500. [Online]. Available: https://doi.org/10.1371/journal.pone.0291500.

[14] M. Vallejo, *Emotions vs. feelings vs. moods: Key differences*, Accessed: 2025-2-1, 2023. [Online]. Available: https://mentalhealthcenterkids.com/blogs/articles/emotions-vs-feelings-vs-moods.

[15] S. M. George and P. M. Ilyas, "A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise," *Neurocomputing*, vol. 568, 2024, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2023.127015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223011384.

[16] S. W. II, *What is speech emotion recognition?* Accessed: 2025-02-01, 2024. [Online]. Available: https://klu.ai/glossary/speech-emotion-recognition.

[17] S. Kalateh, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata, "A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges," *IEEE Access*, vol. 12, pp. 103 976–104 019, 2024. DOI: 10.1109/ACCESS.2024.3430850.

[18] M. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, Oct. 2024. DOI: 10.1002/widm.1563.