CSE 4000: Thesis/ Project

RECOGNITION OF EMOTION AND CLASSIFICATION USING DEEP LEARNING

By

Md. Tanvir Hossain Tamim

Roll: 1907060

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

October, 2024

Recognition of Emotion and Classification Using Deep learning

By

Md. Tanvir Hossain Tamim

Roll: 1907060

A thesis submitted in partial fulfillment of the requirements for the degree of

"Bachelor of Science in Computer Science & Engineering"

Supervisor:

Md. Abdus Salim Mollah

Assistant Professor

Department of Computer Science & Engineering

Khulna University of Engineering & Technology (KUET)

Signature

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

October 2024

Acknowledgment

I am profoundly grateful to the Almighty for granting me strength, perseverance, and wisdom throughout this academic journey. The grace and guidance I have received have been invaluable in shaping the path of my research. I extend my heartfelt appreciation to my thesis supervisor, Md. Abdus Salim Mollah, Assistant Professor, Department of Computer Science and Engineering, Khulna University of Engineering & Technology, for his unwavering support, insightful guidance, and constructive feedback. His expertise and dedication have played a pivotal role in shaping the trajectory of this research.

I am indebted to my parents for their boundless love, encouragement, and sacrifices. Their unwavering belief in my abilities has been a constant source of motivation.

Author

Abstract

In this study, we have seen that a novel approach for Speech Emotion Recognition (SER) that not only detects emotions from speech signals but also classifies the intensity of those emotions, such as normal or strong. Emotional intensity is crucial because high-intensity emotions like extreme sadness or anger can lead to severe or destructive behaviors, including self-harm or violence. Existing SER models largely ignore this aspect, focusing only on emotion categorization. To address this, the proposed method combines three speech signal transformation techniques—Mel-frequency Cepstral Coefficient (MFCC), Short-Time Fourier Transform (STFT), and Chroma STFT—into a 3D feature representation. This 3D feature is then processed using a deep learning model, which includes 3D Convolutional Neural Networks (CNN), a Time Distribution Flatten (TDF) layer, and Bidirectional Long Short-term Memory (Bi-LSTM) networks. Two deep learning frameworks are explored: a single framework for simultaneous emotion and intensity classification, and a cascaded framework, where emotion classification occurs first, followed by intensity classification. The method is evaluated on the RAVDESS dataset, a benchmark for emotional speech, and results show that the cascaded framework significantly outperforms the single framework, achieving superior accuracy in both emotion and intensity recognition compared to existing metho

Contents

List of Tables

List of Figures

CHAPTER I

Introduction

1.1 Introduction

The introduction addresses the significance of Speech Emotion Recognition (SER) in extracting emotions from speech, especially considering emotional intensity (e.g., normal, strong), which can influence behavior, including severe actions like self-harm. While existing deep learning models focus on categorizing emotions, they often ignore intensity, a critical factor in fully understanding emotional context. To fill this gap, the study proposes a novel method that combines three speech transformation techniques—MFCC, STFT, and Chroma STFT—into 3D features, processed by a deep learning model comprising 3D CNN, TDF, and Bi-LSTM, to recognize both emotion and its intensity.

1.2 Background Study

Emotion recognition from speech signals using deep learning has gained significant attention due to its applications in human-computer interaction, mental health monitoring, and personalized communication. By analyzing various acoustic features, such as pitch, tone, and tempo, deep learning models can classify emotions with high accuracy. Techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been particularly effective in capturing the temporal spatial characteristics of audio signals. Additionally, the integration of transfer learning has enhanced model performance by leveraging pre-trained networks on large datasets. As advancements in computational power and data availability continue, the potential for real-time emotion recognition systems in diverse fields, including telecommunication and education, becomes increasingly feasible, promising to enhance user experiences and interactions.

1.3 Motivation

The motivation behind this paper is rooted in the essential role that emotional communication plays across various domains, including healthcare, education, and customer service. Effective emotion recognition is crucial for understanding and responding to user needs, yet traditional methods often fall short in capturing the complexities of human emotions. As the demand for more sophisticated interaction increases, there is a pressing need for advanced systems that can accurately classify emotional states from speech signals, taking into account the subtleties that convey meaning in real-world scenarios.

This research aims to harness deep learning techniques to improve emotion recognition accuracy and reliability. Ultimately, this work seeks to create empathetic and responsive technology that enhances human-computer interaction, paving the way for applications in mental health monitoring, personalized education, and improved customer service, thereby fostering better communication and well-being in an increasingly digital landscape.

1.4 Problem Statement

The problem addressed in this paper is the inadequate accuracy and reliability of existing emotion recognition systems in classifying emotional states from speech signals. Traditional methods often struggle to capture the complexities and nuances of human emotions, particularly in real-world scenarios where emotional expression can vary significantly in intensity and context. This limitation hinders the effectiveness of applications in critical areas such as healthcare, customer service, and education, where understanding emotional states is essential for enhancing user experiences and responses.

Furthermore, many current models fail to account for the varying degrees of emotional intensity, this research proposes the use of advanced deep learning techniques, specifically convolutional and recurrent neural networks, to analyze acoustic features in speech. The goal is to develop a robust framework that accurately recognizes and classifies emotions, including their intensity, thereby improving the performance of emotion recognition systems and enabling more empathetic and responsive interactions in various applications.

1.5 Objectives

The objectives of the study as presented in the paper titled *"Recognition of Emotion with Intensity from Speech Signal Using 3D Transformed Feature and Deep Learning"* are as follows:

1. Enhance Emotion Recognition Accuracy: Develop a deep learning model that improves the accuracy of emotion recognition from speech signals by leveraging advanced algorithms and techniques.

2. Incorporate Emotional Intensity Classification: Design the model to classify not only discrete emotional states but also the intensity of emotions, providing a more nuanced understanding of emotional expression.

3. Implement Deep Learning Techniques: Utilize convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture temporal and spatial dependencies in speech data for improved classification performance.

4. Improving Recognition Accuracy: The paper aims to improve emotion recognition accuracy over existing models by combining 3D transformed features with a more sophisticated DL architecture that includes 3D CNN, Time Distribution Flatten (TDF) layers, and Bidirectional Long Short-Term Memory (Bi-LSTM) networks.

1.6 Scope and Required Tool

Scope:

1. Objectives: The research aims to develop a deep learning framework called "Recognition of Emotion with Intensity from Speech (REIS)," which can recognize emotions and their intensity (e.g., normal, strong) from speech signals using 3D transformed features.

2. **Application**: The model will classify emotions such as Happy, Sad, Angry, Fearful, and Neutral, as well as the intensity of these emotions (normal, strong).

3. **Evaluation**: The model will be trained and evaluated using the RAVDESS dataset to demonstrate its performance in emotion recognition.

Required Tools: To carry out the task a set of modern tools and technologies is required, some of them are mentioned below:

1. **Deep Learning Frameworks**: Keras and TensorFlow are used to implement the deep learning models.

2. **Convolutional Neural Networks (CNNs)**: The model utilizes 3D CNNs for extracting features from the speech signals.

3. **Bi-LSTM Networks**: Bidirectional LSTM networks are used to capture temporal patterns in the transformed speech features.

4. **3D Transformed Feature**: The integration of MFCC, STFT, and Chroma STFT features, combined into a 3D feature structure, is used to represent the speech signal.

5. **Dataset**: Some datasets can be used to implement the task. Such as RAVDESS,EMO-DB etc.

1.7 Unfamiliarity of the Problem

The unfamiliarity with the problem addressed in this paper can be highlighted As follows:

1. **Emotion Intensity Recognition**: While Speech Emotion Recognition (SER) has been widely researched, most existing models only focus on categorizing emotions.

2. **Lack of Integrated Approaches**: Traditional methods do not effectively integrate multiple signal transformation techniques (MFCC, STFT, Chroma STFT) into a unified 3D feature structure, which this paper proposes for more accurate recognition.

3. **DL Model Complexities**: Handling both emotion classification and intensity detection within a single deep learning framework introduces complexities that many existing approaches have not addressed.

1.8 Project Planning

1.8.1 Gantt Chart

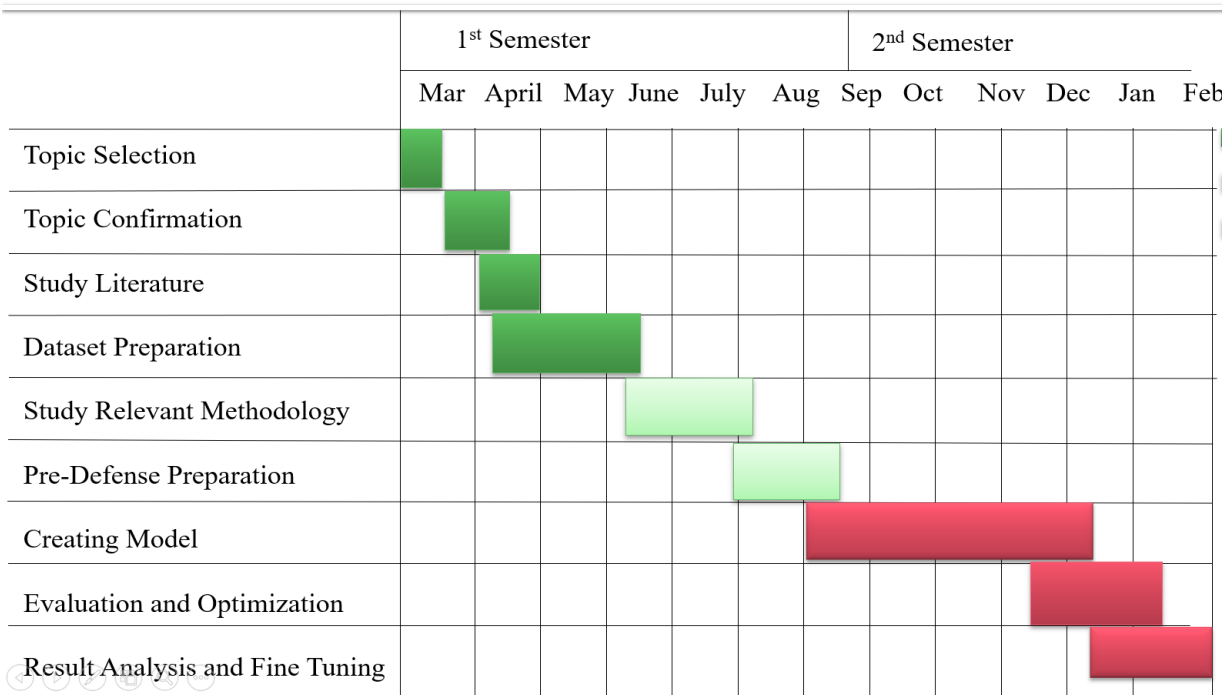

Fig1.8.1: Gantt chart of the thesis

The total thesis task is divided into two parts, carry out in two semesters. In the first semester, the main task is to generate an idea of how to approach the problem and study relevant papers for gathering information about how the problem can be solved. In the second semester, attention will be given to the implementation phase

1.8.2 Societal, health, and cultural issues

The societal, health, and cultural issues in the paper are:

1. **Societal**: Detecting intense emotions like anger or sadness can prevent harmful actions such as violence or s suicide, which are often shared on social media.
2. **Health**: Recognizing extreme emotional states can aid in mental health interventions, alerting professionals before dangerous behaviors occur.
3. **Cultural**: The system addresses the challenge of detecting emotions across different cultures, ensuring adaptability in cross-cultural communication.

1.9 Applications

The application of this field of research are below:

1. **Online Media Monitoring**: The REIS framework can be used to automatically detect intense emotions like anger or sadness in online media, allowing for real-time monitoring and filtering of potentially disruptive behaviors such as violence or suicidal actions.

2. **Mental Health**: This system can alert mental health professionals when individuals exhibit signs of extreme emotions, helping in early intervention to prevent self-harm or aggressive behavior.

3. **Human-Computer Interaction**: It can enhance emotion-aware AI systems in virtual assistants, gaming, and other interactive platforms, allowing more empathetic and effective communication.

1.10 Organization

The thesis report consists of in total of five chapters. Each of the chapters tells how the thesis problem was planned and how it will be executed.

Chapter I: Introduction to speech emotion recognition and the need for emotion intensity recognition.

Chapter II: Review of existing methods, their advantages, limitations, and gaps.

Chapter III: Description of the proposed REIS model and its methodology.

Chapter IV: Experimental setup, tools, dataset, and results.

Chapter V: Conclusion and future research direction

CHAPTER II

Literature Review

2.1 Introduction

The first paper represents the evolution of Speech Emotion Recognition (SER) from traditional machine learning (ML) methods, such as Random Forest and Support Vector Machines, to more advanced deep learning (DL) techniques like CNNs, LSTMs, and hybrid models. Existing DL models primarily focus on categorizing emotions but fail to integrate multiple signal transformation methods such as MFCC, STFT, and Chroma STFT into a unified system for intensity recognition. This study aims to bridge that gap by proposing a novel approach using 3D transformed features and cascaded DL frameworks to enhance the accuracy of both emotion and intensity recognition.

The second paper's literature review on speech emotion recognition (SER) explores different approaches to classifying emotions from speech signals. One paper focuses on the use of artificial neural networks (ANN) for SER, highlighting methods such as MFCC feature extraction, prosodic features like pitch and energy, and traditional classifiers like SVM and KNN, which have shown reasonable accuracy but still face challenges in speaker independence and emotion complexity.

The third paper delves into multimodal emotion recognition, emphasizing deep learning techniques such as CNN, LSTM, and hybrid models that combine facial and speech cues for better accuracy. This approach enhances emotion detection by integrating multiple modalities, offering improved performance compared to unimodal systems.

2.2 Related Terms

Related terms associated with the thesis are given below:

Prosodic Features: Acoustic features such as pitch, energy, and formant frequencies that reflect emotional states from speech signals.

Mel-frequency Cepstral Coefficients (MFCC): A common feature extraction method that represents the short-term power spectrum of sound, crucial for SER.

Preprocessing: Techniques such as normalization, sampling, and segmentation applied to speech signals before feature extraction.

Artificial Neural Networks (ANN): A machine learning algorithm used for emotion classification based on features extracted from speech.

Convolutional Neural Networks (CNN): Deep learning models used for both unimodal and multimodal emotion recognition, primarily in feature extraction from audio, images, and video.

Long Short-Term Memory (LSTM) and **Bidirectional LSTM (Bi-LSTM)**: Recurrent neural networks (RNN) designed to handle sequential data like speech and video, capturing temporal dependencies for improved emotion recognition.

Cross-modal Attention: A mechanism that focuses on important information across different modalities, enhancing the emotion recognition process.

Facial Expression Recognition (FER): Detecting emotions using visual cues from facial expressions, often combined with speech data for better accuracy in multimodal approaches.

Short-Time Fourier Transform (STFT) and **Chroma STFT**: Techniques used to transform speech signals into representations for emotion classification

2.3 Related Works

The **related works** across the three papers revolve around advancements in **Speech Emotion Recognition (SER)** and **Multimodal Emotion Recognition**, exploring different machine learning and deep learning techniques. Here are the related works from each paper:

2.3.1 Machine Learning for SER

Early SER systems relied on traditional machine learning techniques like SVM, which provided a foundation for later deep learning approaches.

2.3.2 Deep Learning Advancements: Related works highlight the shift towards deep learning, particularly CNNs and LSTMs, which have improved the accuracy of SER systems. Multimodal systems that combine data from different sources (audio, facial expressions, and text) also became more common.

2.3.3 Multimodal Emotion Recognition: Previous works in this field focused on combining audio and visual signals to improve emotion recognition, with a growing emphasis on using attention mechanisms and deep learning architectures like CNN and LSTM to handle complex, multimodal data sources.

2.3.4 Signal Transformation Methods: Previous research on signal transformations such as MFCC, STFT, and Chroma STFT has shown effectiveness in feature extraction for SER. This paper builds on that by integrating these features into a 3D structure for enhanced recognition.

2.3.5 Attention Mechanisms: Recent studies focused on using attention mechanisms in multimodal networks to emphasize key features from different modalities, leading to better emotion detection.

2.3.6 Use of Pretrained Models and Architectures: Work on using pretrained architectures like BERT for emotion recognition in text, which can be combined with speech and visual modalities, has shown promising results in improving recognition performance.

2.3.7 ANN for Emotion Classification: Neural networks were explored to improve the accuracy of emotion classification from speech, focusing on improving the recognition of basic emotions such as happiness, sadness, anger, and neutral states.

2.4 Observation of Relevant Papers

The relevant papers highlight several key observations in the field of emotion recognition. Traditional machine learning models like ANN and SVM have been widely used for speech emotion recognition (SER), relying on prosodic features such as pitch, energy, and MFCC. However, these models struggle with speaker independence and handling complex emotions, leading to limited performance. Deep learning approaches, particularly CNNs and LSTMs, have shown significant improvements in both unimodal and multimodal emotion recognition. Multimodal systems that combine data from various sources (e.g., audio, facial expressions, text) offer better accuracy, but they are computationally expensive and complex to implement in real-time. A notable advancement is the recognition of emotion intensity, introduced by models using 3D transformed features like MFCC and STFT, which enhance emotion detection accuracy. Despite these improvements, real-time emotion recognition and practical applications remain challenging due to the computational demands of these deep learning models.

Table 2.4: Statistical comparison of related methods

Paper	Objective	methodology	Dataset	Accuracy	Key findings	Future plan
1.	Propose robust automatic speech emotional-speech recognition	(BFN), CNN (CNA), and hybrid architecture (HBN)	RAVDESS	80.6% - 84.5%	Achieved precision between 81.5% and 85.5%	Explore more hybrid models and improve accuracy
2.	multimodal signals using deep learning. Compare applications	Multimodal affective computing systems and unimodal solutions	Various multimodal and unimodal datasets	74.9%-80.0%	Multimodal systems offer higher classification accuracy	integration of various signals for improved accuracy.
3.	deep attention-based dilated convolutional recurrent neural network	Uses hybrid data augmentation, dilated CNNs, and RNNs	EmoDB, ERC	88.03% (EmoDB), 66.56% (ERC)	Achieved highest unweighted recall rates of 88.03% (EmoDB) and 66.56% (ERC).	Refine the attention mechanisms and further optimize data .
4.	Enhance SER systems by combining MFCCs with time-domain features.	Hybrid features (MFCCT) fed into CNN.	Emo-DB, SAVEE, RAVDESS	97% (Emo-DB), 93% (SAVEE), 92% (RAVDESS)	Achieved accuracy of 97% (Emo-DB), 93% (SAVEE), and 92% (RAVDESS).	diverse datasets and explore other hybrid features.

The observations from the relevant papers reveal that while traditional machine learning models like ANN and SVM have been useful for basic speech emotion recognition (SER), they struggle with complexity and speaker-independence. Deep learning techniques, particularly CNNs and LSTMs, provide significant improvements by capturing both spatial and temporal features, especially in multimodal systems that combine audio, facial expressions, and text. Multimodal approaches enhance accuracy but come with challenges in computational cost and real-time application. A key advancement is the integration of emotion intensity detection using 3D transformed features, which improves the precision of emotion recognition. However, real-time implementation of these sophisticated models remains a challenge due to their high resource demands.

2.5 Conclusion

The conclusions of these three papers collectively highlight the progress made in the field of emotion recognition, with a focus on both speech and multimodal data, leveraging deep learning techniques. Initially, traditional machine learning approaches such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) were widely used for Speech Emotion Recognition (SER). These methods, while effective for recognizing basic emotions like happiness, sadness, and anger, showed limitations in handling more complex emotions, speaker variability, and intensity recognition.

A key innovation discussed in these papers is the introduction of emotion intensity recognition, which adds a new dimension to traditional emotion recognition systems. By using 3D transformed features like MFCC, STFT, and Chroma STFT, the papers offer models that are more precise in distinguishing between normal and intense emotional states, addressing an often-overlooked aspect of emotion detection.

While these models represent significant advancements in the field, practical challenges remain, particularly in optimizing them for real-time applications. High computational demands and the complexity of integrating multiple modalities or processing 3D features in real-time are key areas for future improvement. The papers collectively call for further research into making these systems more scalable and practical for widespread use in areas like mental health monitoring, customer service, and human-computer interaction.

In contrast, deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) have demonstrated significant advancements. CNNs effectively capture spatial features in visual data (e.g., facial expressions) and spectral features in speech, while LSTMs are crucial for understanding temporal dependencies in sequential data, such as speech signals. These models have not only improved the accuracy of emotion recognition but also enhanced the ability to recognize emotions across multiple modalities, including speech, facial expressions, and text. The combination of modalities (multimodal emotion recognition) allows for higher accuracy, as different signals complement each other. However, such systems come with higher computational costs and complexity, making real-time applications challenging.

CHAPTER III

Methodology

The methodology of this paper follows a systematic approach to develop a robust deep learning model for emotion recognition from speech signals with a focus on emotional intensity classification. The key steps in the methodology are as follows:

3.1 Data Collection

A large and diverse dataset of speech samples is collected from publicly available emotion recognition datasets such as IEMOCAP or RAVDESS, which contain labeled emotional states.The dataset is pre-processed to ensure uniformity, including tasks like noise reduction, normalization, and segmentation of speech signals.

3.2 Feature Extraction

A deep learning model is designed, incorporating convolutional neural networks (CNNs) for spatial feature extraction and recurrent neural networks (RNNs) or long short-term memory (LSTM) networks to capture temporal dependencies in speech data. The model includes layers specifically designed to classify both discrete emotions (e.g., happiness, sadness) and their intensity levels (e.g., low, medium, high).

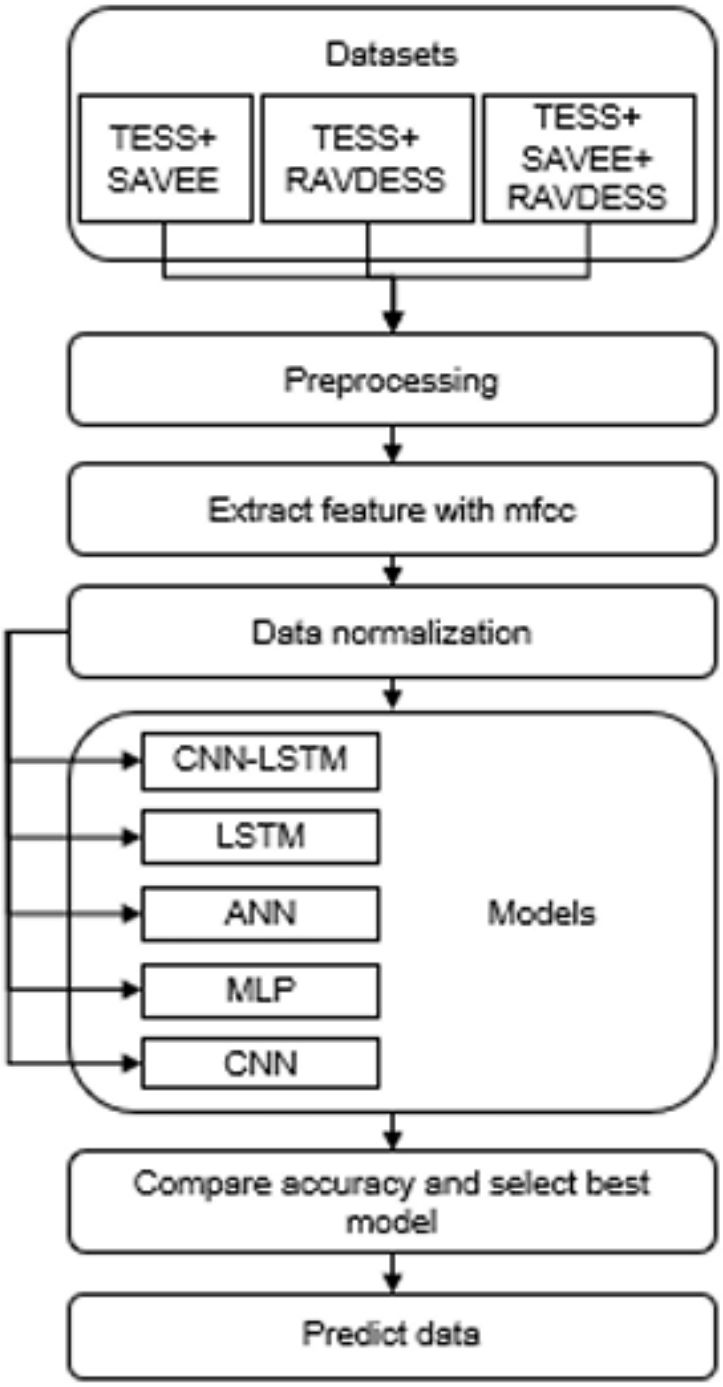

Figure 3.1:Flowchart of the SER system

3.3 Model Training

The model is trained on the pre-processed dataset using a supervised learning approach. Cross-validation techniques are employed to avoid overfitting. A loss function that considers both emotion classification and intensity is implemented, with optimization techniques such as Adam optimizer used to minimize this loss.

The flowchart outlines a speech emotion recognition system. It begins with datasets, followed by preprocessing and feature extraction using MFCC. Data normalization ensures consistency. Various models, including CNN-LSTM, LSTM, ANN, MLP, and CNN, are trained and evaluated to select the best. Finally, the chosen model predicts emotions on new data. This system effectively processes speech data, extracts relevant features, and accurately classifies emotions.

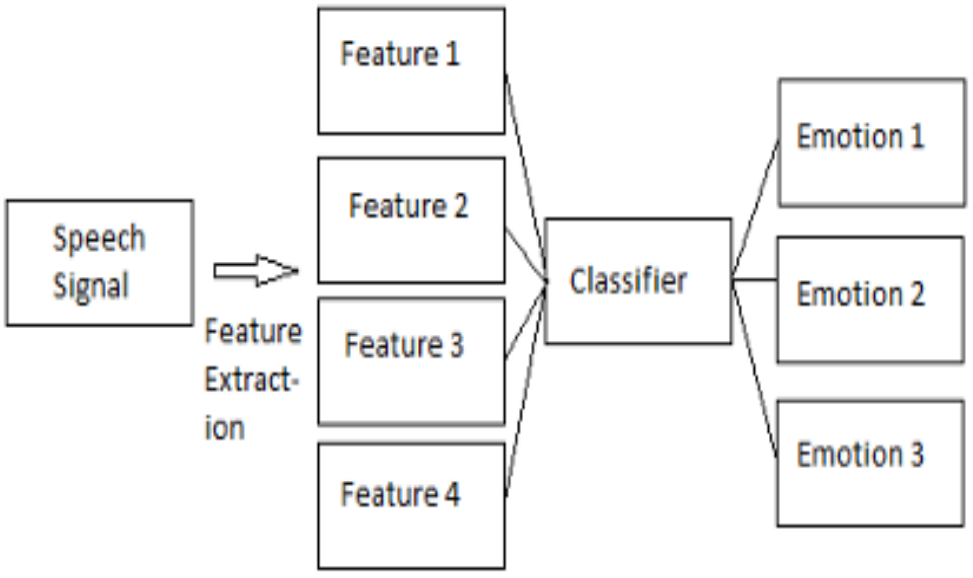

Figure 3.2: Flow of emotion recognition and classification

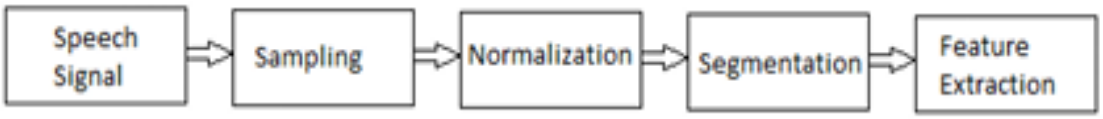

Figure 3.3: Pre-processing for emotion recognition

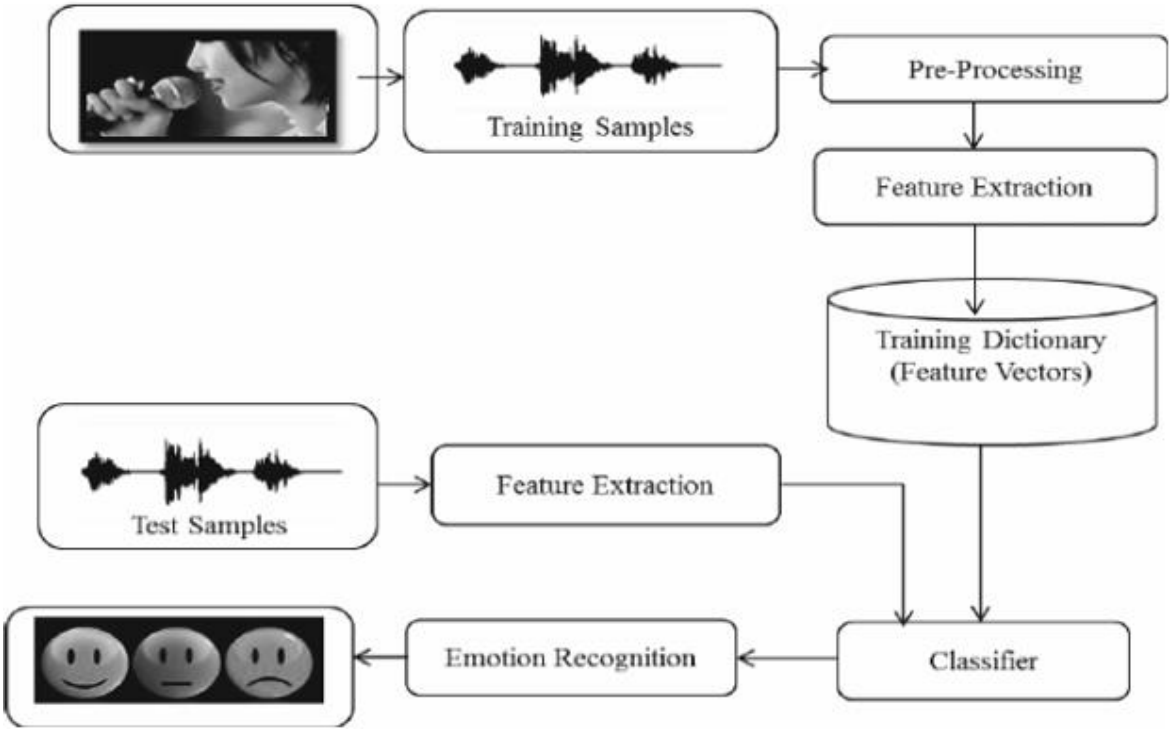

Figure 3.4:Basic process of SER system (Training and Testing)

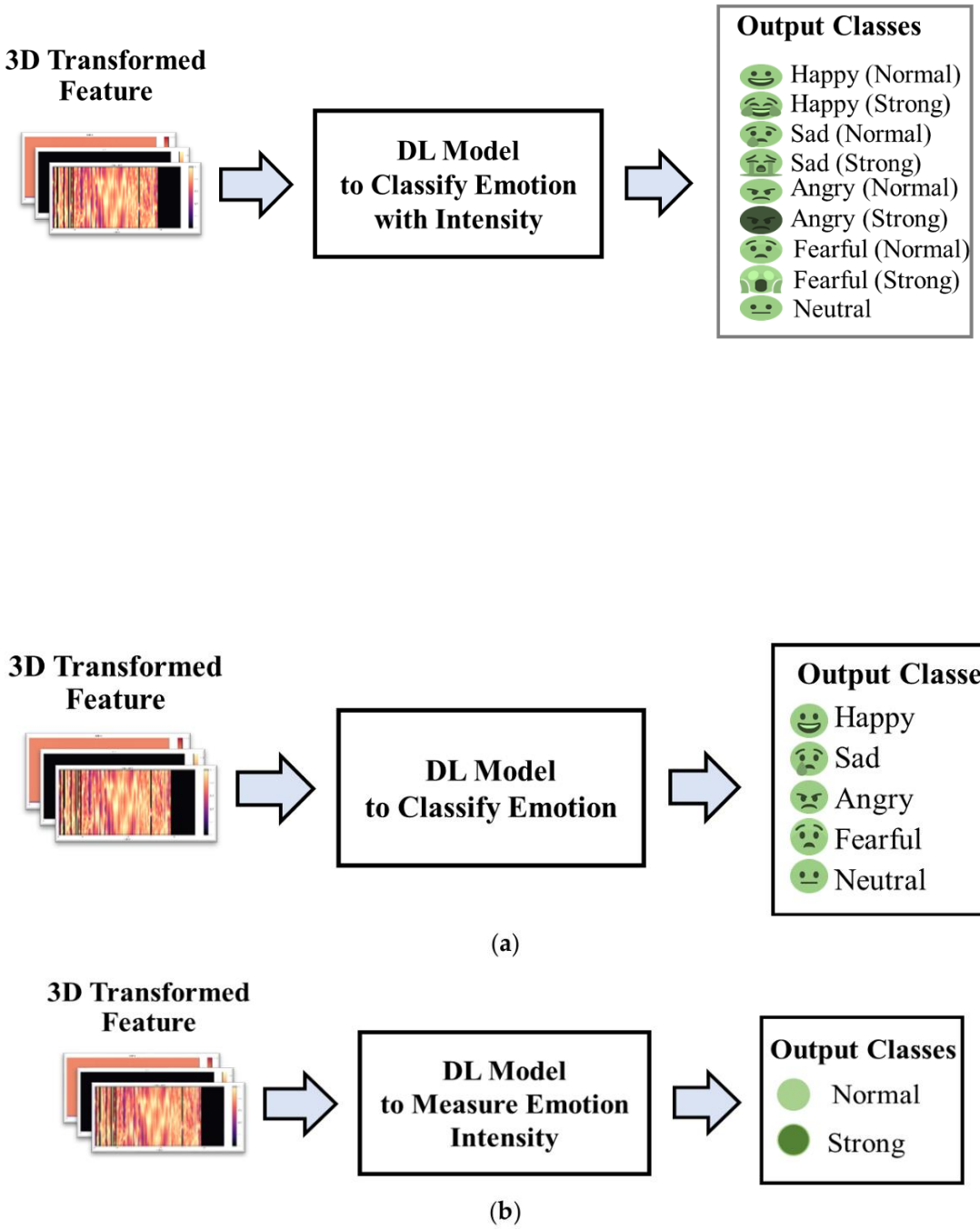

Figure 3.5: Single DL Framework vs Cascaded DL Framework

CHAPTER IV

Implementation and Results

4.1 Introduction

The implementation of this thesis involves several key steps to develop a deep learning model for emotion and intensity recognition from speech signals. First, relevant datasets, such as IEMOCAP or RAVDESS, are selected and pre-processed through noise reduction, normalization, and segmentation. Acoustic features, including Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and tempo, are extracted to represent emotional cues. A hybrid model architecture using convolutional neural networks (CNNs) for spatial feature extraction and recurrent neural networks (RNNs) or long short-term memory (LSTM) networks for temporal dependencies is designed. The model is trained to classify both emotion and intensity using a supervised learning approach and evaluated using metrics like accuracy, precision, and recall. Hyperparameter tuning is conducted to optimize performance. The model's results are compared with traditional techniques, and future improvements, such as real-time processing and multilingual support, are suggested for broader applications in fields like mental health and human-computer interaction.

4.2 Morality and Ethical Issues

The production and consumption of paper present significant moral and ethical challenges, particularly regarding environmental sustainability and social responsibility. Deforestation for paper manufacturing leads to biodiversity loss and contributes to climate change, while poor labor practices in the industry raise concerns about exploitation and injustice. As society increasingly turns to digital alternatives, it is essential to evaluate the ethical implications of both paper and digital consumption, ensuring that sustainability and equity guide our choices for a more just future.

4.3 Socio-Economic Impact and Sustainability

The socio-economic impact of this research lies in its potential to revolutionize fields like healthcare, customer service, and education by enabling more empathetic and responsive systems. In healthcare, emotion recognition from speech can support mental health monitoring, helping professionals detect emotional distress early, which can lead to improved mental health outcomes. In customer service, systems that understand emotional states can provide more personalized and effective interactions, leading to higher customer satisfaction and potentially increasing business revenue. In education, adaptive learning platforms that recognize student emotions can foster better engagement and tailored support, improving learning outcomes. From a sustainability perspective, the technology developed in this paper promotes more inclusive and accessible services, supporting the well-being of individuals across different demographics. The development of real-time, emotion-aware systems could also reduce the need for human intervention in various processes, contributing to efficient resource utilization. Moreover, such technology aligns with the growing emphasis on AI-driven solutions that foster human-centered, sustainable innovation.

4.4 Financial Analysis and Budget

The financial analysis of this research includes several key components. Data acquisition costs range from 55,000 to 165,000 BDT for licensing public emotion datasets. Hardware expenses, particularly for GPU purchases or cloud-based computational services, are estimated at 220,000–550,000 BDT for local GPUs or 22,000–110,000 BDT per month for cloud services. Software costs are minimal due to open-source deep learning frameworks, while personnel costs for researchers can range from 330,000 to 660,000 BDT per month. Cloud storage for datasets adds 11,000–33,000 BDT per month, and publication fees for journals may cost between 55,000 and 220,000 BDT. Miscellaneous expenses, such as conference travel, could add another 110,000–330,000 BDT. The total budget is estimated at 660,000–2,200,000 BDT, depending on resources used.

CHAPTER V

Conclusion

5.1 Conclusion

The conclusion of this paper highlights the successful development and implementation of a deep learning model for emotion recognition and intensity classification from speech signals. By leveraging advanced techniques such as convolutional and recurrent neural networks, the research demonstrates improved accuracy in identifying emotional states and their intensity levels compared to traditional methods. The proposed model effectively extracts acoustic features, enabling nuanced recognition of emotions, which is crucial for real-world applications in areas like mental health, customer service, and education. The findings suggest that integrating emotion intensity into classification enhances the system's responsiveness, making it more applicable to human-computer interaction.

The paper also emphasizes the potential for further advancements, such as optimizing real-time processing capabilities and expanding to multilingual datasets. Overall, the research contributes to the growing field of emotion recognition and sets the foundation for future innovations in empathetic and adaptive AI systems that improve communication and user experiences across various sectors.

5.2 Future Works

Future work in the field of emotion recognition from speech signals offers several promising directions:

1. **Real-time Emotion Recognition**: Future research can focus on optimizing models for real-time emotion recognition, reducing processing latency and enabling seamless integration into interactive systems like virtual assistants, customer service bots, and telehealth platforms.
2. **Context-Aware Emotion Detection**: Future systems can integrate semantic analysis of speech content, enabling the model to understand context and improve the accuracy of emotion recognition, particularly in complex or ambiguous speech.
3. **Transfer Learning and Pre-trained Models**: Leveraging pre-trained models on larger datasets or using transfer learning could reduce the need for extensive data collection and make emotion recognition systems more accessible to developers with limited resources.
4. **Personalized Emotion Models**: Developing models that adapt to individual emotional patterns and voice characteristics can significantly improve accuracy, especially in personalized services like mental health monitoring or personalized learning environments.

REFERENCES

[1] M. R. Islam, M. A. H. Akhand, M. A. S. Kamal, and K. Yamada, "Recognition of Emotion with Intensity from Speech Signal Using 3D Transformed Feature and Deep Learning," *Electronics*, vol. 11, no. 15, p. 2362, 2022, doi: 10.3390/electronics11152362.

[2] A. V. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal Emotion Recognition with Deep Learning: Advancements, Challenges, and Future Directions," *Information Fusion*, vol. 105, p. 102218, 2024, doi: 10.1016/j.inffus.2023.102218.

[3] R. Vardhan, "Emotion recognition and classification in speech using Artificial Neural Networks," *International Journal of Computer Applications*, vol. 140, no. 8, pp. 1-5, 2016.

[4] M. Ezz-Eldin, A. A. M. Khalaf, H. Hamed, and A. Hussein, "Efficient Feature-Aware Hybrid Model of Deep Learning Architectures for Speech Emotion Recognition," *IEEE Access*, vol. PP, pp. 1-1, 2021, doi: 10.1109/ACCESS.2021.3054345.

[5] D. Bharti and P. Kukana, "A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals," in *Proceedings of the International Conference on Security and Privacy (ICOSEC)*, 2020, pp. 491-496, doi: 10.1109/ICOSEC49089.2020.9215376.

[6] M. N. Adnan, R. Ema, S. Galib, S. K. Kabir, and S. K. Hazra, "Emotion recognition of human speech using deep learning method and MFCC features," *Radioelectronic and Computer Systems*, vol. 2022, pp. 161-172, 2023, doi: 10.32620/reks.2022.4.13.