SPEECH EMOTION RECOGNITION

AN INTERNSHIP REPORT

AT

HCLTECH

Submitted by

Dharmendra Yadav
21BCS11791

In partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING IN

COMPUTER SCIENCE AND ENGINEERING

Chandigarh University

JUNE 2023

DECLARATION

I, **Dharmendra Yadav**, hereby declare that this internship report, titled "**Speech Emotion Recognition**", is a record of my own work conducted during the internship period at **HCLTech**. The report is submitted to Institute of Engineering, Chandigarh University as a partial fulfillment of the requirements for the Degree Bachelor of Computer Science and Engineering. I affirm that all the information presented in this report is accurate, authentic, and based on my personal experiences and observations during the internship.
Furthermore, I declare that this report has not been submitted for any other purpose or evaluation, either in whole or in part, and is unique to this internship.

I hereby affix my signature and the date below:

Name: Dharmendra Yadav

Date: 1st July 2023

Place: Chandigarh University, Mohali

ABOUT THE COMPANY

HCL Technologies, commonly referred to as HCLtech, is a leading global IT services and consulting company. Founded in 1976, HCL Technologies has its headquarters in Noida, India. The company has grown to become one of the largest IT services providers in the world, with a strong presence across multiple countries. HCL Technologies offers a comprehensive range of services and solutions to meet the evolving needs of its clients. These services encompass various aspects of technology and business, including software development, infrastructure management, cybersecurity, cloud computing, digital transformation, and engineering services. With a customer-centric approach, HCL Technologies focuses on creating value and delivering innovative solutions to its clients. The company adopts a "Relationship Beyond the Contract" philosophy, emphasizing long-term partnerships and collaborative engagements with its clients. HCL Technologies serves clients from diverse industries, including technology, healthcare, financial services, manufacturing, retail, telecommunications, and more. The company's client base includes renowned global enterprises, small and medium-sized businesses, and government organizations. HCL Technologies prides itself on its commitment to sustainability and social responsibility. The company actively engages in corporate social responsibility initiatives, promoting education, healthcare, environmental conservation, and community development. Driven by a talented workforce of professionals, HCL Technologies fosters a culture of innovation, continuous learning, and teamwork. The company encourages its employees to explore new technologies and ideas, enabling them to deliver high-quality solutions that address complex business challenges. With its global footprint, extensive expertise, and customer-centric approach, HCL Technologies remains at the forefront of the IT services industry, helping organizations transform and thrive in the digital age.

KEY HIGHLIGHTS OF THE COMPANY

1. Global Presence: HCL Technologies has a strong global presence with operations in over 50 countries. The company serves clients across North America, Europe, AsiaPacific, and other regions, providing them with localized services and support.

2. Diverse Industry Expertise: HCL Technologies serves clients across various industries, including technology, healthcare, financial services, manufacturing, retail, telecommunications, and more. This broad industry expertise enables the company to understand unique business challenges and deliver tailored solutions.
3. Innovative Solutions: HCL Technologies emphasizes innovation and has a track record of delivering cutting-edge solutions to its clients. The company invests significantly in research and development, leveraging emerging technologies like artificial intelligence, cloud computing, blockchain, Internet of Things (IoT), and automation to drive digital transformation.

4. Customer-Centric Approach: HCL Technologies places a strong emphasis on building long-term relationships with its clients. The company focuses on understanding their specific needs, challenges, and goals, and works collaboratively to provide customized solutions that deliver tangible business outcomes.

5. Strong Partner Ecosystem: HCL Technologies has developed strategic partnerships with leading technology companies, including Microsoft, IBM, Cisco, and others. These partnerships enable HCLtech to access the latest technologies, tools, and resources, enhancing its ability to deliver comprehensive solutions to clients.

6. Employee-Centric Culture: HCL Technologies values its employees and promotes a culture of innovation, learning, and collaboration. The company provides a conducive work environment, encourages skill development, and fosters a strong sense of ownership and empowerment among its workforce.

7. Sustainability and Social Responsibility: HCL Technologies is committed to sustainability and social responsibility. The company actively engages in environmental conservation, energy efficiency initiatives, and corporate social responsibility programs focused on education, healthcare, and community development.

8. Awards and Recognitions: HCL Technologies has received numerous accolades and industry recognitions for its performance and innovation. The company has been consistently recognized by leading research and advisory firms for its leadership in various domains, including IT services, digital transformation, and customer satisfaction.

TABLE OF CONTENTS

List of Figures

ABBREVIATIONS

ML: Machine Learning

SVM: Support Vector Machine

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

GRU: Gated Recurrent Unit

MFCC: Mel-frequency Cepstral coefficients

LPC: Linear Predictive Coding

EEG: Electro encepphalogram

KNN: K- Nearest Neighbors

LSTM: Long Short Term Memory

HMM: Hidden Markov Models

NLP: Natural Language Processing

ABSTRACT

Recognizing human emotions through analysis of speech is an active area of research that aims to develop intelligent systems capable of accurately identifying emotional states based on the acoustic properties of speech. This process involves extracting features from the speech signal, such as pitch, intensity, and spectral characteristics, and using machine learning algorithms to classify the emotions conveyed by the speaker. Different approaches have been proposed, including using traditional statistical methods, deep learning models, or hybrid methods that combine both approaches. Despite significant progress, challenges remain, including the subjective nature of emotional expression and the influence of cultural and contextual factors on emotion perception. Further research is needed to improve the accuracy and robustness of emotion recognition systems and to develop applications in various domains, such as healthcare, education, and entertainment. Several studies have shown that emotions are expressed through various aspects of speech, including pitch, loudness, tempo, and voice quality. For instance, high pitch and fast tempo are associated with excitement and happiness, while low pitch and slow tempo are related to sadness and depression. Additionally, emotional states can be conveyed through specific speech patterns, such as stuttering or pauses. Emotion recognition from speech has numerous potential applications. In healthcare, it can aid in the diagnosis and treatment of mental health disorders, such as depression and anxiety. In education, it can be used to evaluate students' engagement and emotions during online learning sessions. In entertainment, it can be used to create more immersive and responsive gaming experiences or improve voice-controlled personal assistants' capabilities.Despite the potential applications, emotion recognition from speech also raises ethical concerns related to privacy, surveillance, and bias. To address these concerns, it is essential to develop transparent and accountable systems that consider privacy and ethical implications. In summary, emotion recognition from speech is a promising area of research with numerous potential applications. However, further research is needed to improve the accuracy and robustness of the systems and to address ethical and privacy concerns.

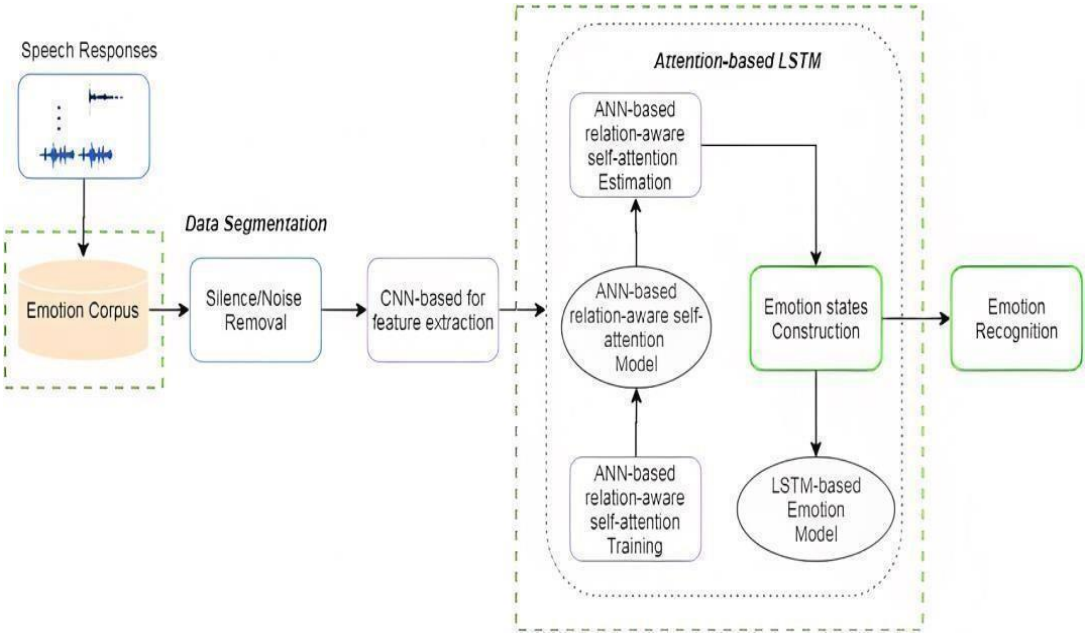
सार

भाषण के विश्लेषण के माध्यम से मानवीय भावनाओं के पहचानना अनुसंधान का एक सक्रिय क्षेत्र है जिसका उद्देश्य भाषण के ध्वनिक गुणों के आधार पर भावनात्मक अवस्थाओं की सही पहचान करने में सक्षम बुद्धिमान प्रणालियों के विकसित करना है। इस प्रक्रिया में भाषण से कई विशेषताओं के निकाला जाता है, जैसे कि पिच, तीव्रता, और वर्णक्रमीय विशेषताओं, और स्पीकर द्वारा बताए गए भावनाओं के वर्गीकृत करने के लिए मशीन लर्निंग एल्गोरिदम का उपयोग करना। विभिन्न तरीकों का प्रस्ताव किया गया है, जिसमें पारंपरिक सांख्यिकीय विधियों, गहन शिक्षण मॉडल या हाइब्रिड दृष्टिकोणों का उपयोग करना शामिल है जो दोनों दृष्टिकोणों की ताकत को जोड़ते हैं। महत्वपूर्ण प्रगति के बावजूद, भावनात्मक अभिव्यक्ति की व्यक्तिपरक प्रकृति और भावना धारणा पर सांस्कृतिक और प्रासंगिक कारकों का प्रभाव सहित चुनौतियां बनी हुई हैं। भावना पहचान प्रणालियों की सटीकता और मजबूती में सुधार और स्वास्थ्य देखभाल, शिक्षा और मनोरंजन जैसे विभिन्न डोमेन में अनुप्रयोगों के विकसित करने के लिए और अधिक शोध की आवश्यकता है। कई अध्ययनों से पता चला है कि भावनाओं के भाषण के विभिन्न पहलुओं, जैसे कि माध्यम से व्यक्त किया जाता है, जिसमें पिच, लय, गति और आवाज की गुणवत्ता शामिल है। उदाहरण के लिए, उच्च पिच और तेज गति उत्तेजना और खुशी से जुड़ी है, जबकि कम पिच और धीमी गति उदासी और अवसाद से जुड़ी है। इसके अतिरिक्त, भावनात्मक अवस्थाओं के विक्षिप्त भाषण पैटर्न के माध्यम से व्यक्त किया जा सकता है, जैसे हकलाना या रुकना। भाषण से भावनाओं की पहचान के कई संभावित अनुप्रयोग हैं। स्वास्थ्य देखभाल में, यह अवसाद और चिंता जैसे मानसिक स्वास्थ्य विकारों के निदान और उपचार में सहायता कर सकता है। शिक्षा के क्षेत्र में, इसका उपयोग ऑनलाइन शिक्षण सत्रों के दौरान छात्रों के जुड़ाव और भावनाओं का मूल्यांकन करने के लिए किया जा सकता है। मनोरंजन में, इसका उपयोग अधिक इमर्सिव और उत्तरदायी गेमिंग अनुभव बनाने या अवाज नियंत्रित व्यक्तिगत सहायकों की क्षमताओं में सुधार करने के लिए किया जा सकता है। संभावित अनुप्रयोगों के बावजूद, भाषण से भावनाओं की पहचान भी गोपनीयता, निगरानी और पूर्वाग्रह से संबंधित नैतिक चिंताओं को उठाती है। इन चिंताओं

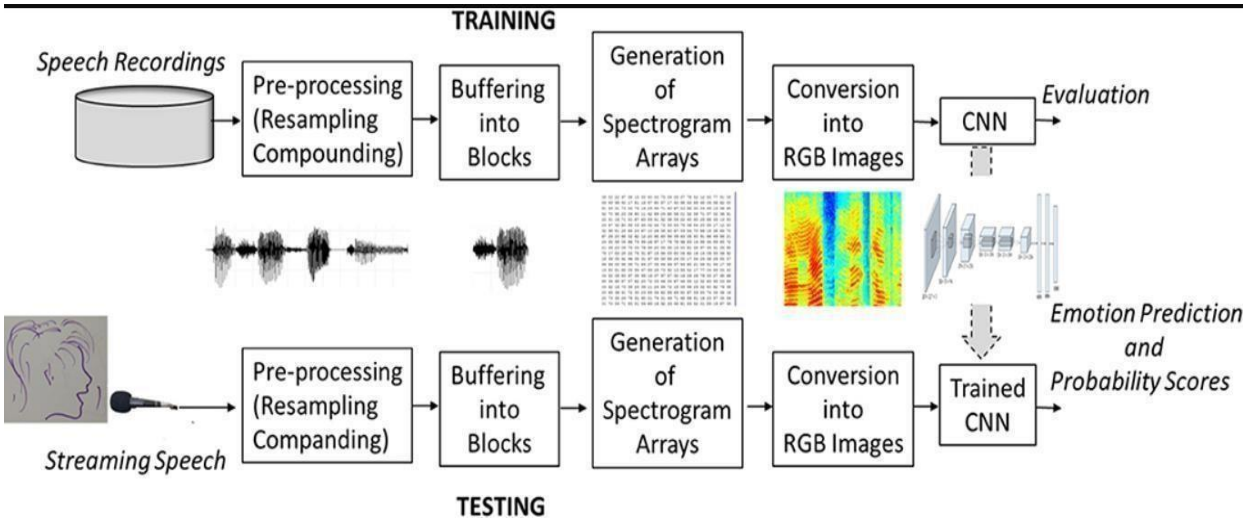
के दे र करने के िवेए, पारदशेिी और वेवेेाबदे ह िसस्टम विकिसत करना वआश्यक हे वेे गे पनीयतेा और ने ितक िनिहतेार्थे ं पर विचेार करता हे। सेक्षेप मे, भाषण से भावेनाओ केिी पहचान कई से भाविेत अनेप्रये गे वर्षा अन्स धान का एक आशावेनक क्षत्र ह। हावेेा े ि क, िसस्टम किी सटिीकता और वमबतिी में स ुधार करन और न ितकऔर ग पनीयता सबंधी चिंताओक द र करन क विए और श ध किी वआश्यकता ह।

A

GRAPHICAL ABSTRACT



Proposed Solution architecture



Design Process

CHAPTER-1
INTRODUCTION

1.1. Identification Of Client

In recent years, many studies have been conducted to understand emotions in human speech. Automatic character recognition systems can be thought of as cognitive systems designed to recognize speech (words and sentences) that can be used for simple conversations of the same people and robots. Recognition of this negative behavior in the literature can be done with all kinds of changes such as facial expression, speech and mimics. Most acoustic features used for sensory perception fall into two groups: prosodic and spectral. Prosodic features such as intonation, accent, intonation, silence, and speaking rate have been found to provide recognition. Spectral features show the frequency content of speech symbols and provide additional information for prosodic features. Spectral features are usually extracted within a short period of time. We can also express energies such as low frequency and high frequency in certain behaviors. Although technologies have been developed to improve the performance of voice communications, high-performance human-robot interactions with robots are still far from ideal. Speech recognition has been studied for many purposes. People interact with other people in a social discourse that goes beyond logical reasoning. The definition of "human emotions" is: "There are things in life that people cannot have without emotions"; this is the concept of universal thinking. Automatic emotion recognition performance based on speech analysis is poor for general speech and private speech. The cognitive speech perception task is very difficult for the following reasons. First, it is not clear which language works best for differentiating emotions across cultures. Acoustic variability, indicated by the presence of different phrases, speech, speech, and rate of speech are other factors that directly affect most speech, such as contours of voice and power. The most important work that detects curiosity without speaking is the model of the famous Ekman's and Fox. Ekman's model is based on six theories, while Fox's model is more of a theory. Each thought corresponds to a different part of speech. However, it is difficult to define the boundaries of these works. Another difficult problem is that the way of thinking expressed in Mohammad Rabiei and Alessandro Gasparettoa often depends on the speaker, his culture and environment. Much of the literature work has focused on the distribution of monolingual emotions; but in addition to emotional state, speech and language can also provide information about the speaker's age, gender, and regional

background. Several other articles in this field have addressed various aspects of sentiment analysis and automatic recognition of emotions in speech, voice recognition, and automatic speech recognition. In the context of human-computer interaction, the analysis of speech theory mainly focuses on the design of the interlocutor.This is often included in practices in mind discussions. Recognizing human emotions using analysis of speech involves extracting relevant features from the speech signal and using machine learning algorithms to classify the emotional state of the speaker. Some of the commonly used features for emotion recognition from speech include pitch, intensity, spectral characteristics, and duration of specific speech segments.

One approach for emotion recognition is to use a supervised machine learning algorithm, such as a Support Vector Machine (SVM) or a Neural Network, which is trained on a dataset of labeled speech samples. The dataset should include speech samples that represent different emotional states, such as happiness, sadness, anger, fear, and neutral. During training, the algorithm learns to associate the features extracted from the speech signal with the corresponding emotional state labels.

Once the algorithm is trained, it can be used to predict the emotional state of new speech samples. The algorithm extracts the same set of features from the new speech sample and applies the learned associations to predict the most likely emotional state label.

While there is still room for improvement, research has shown that using machine learning algorithms can achieve high accuracy in recognizing human emotions from speech. In some cases, combining speech analysis with other modalities, such as facial expressions and physiological signals, can further improve the accuracy of emotion recognition.The Process involves the methods like:

Collect speech data: To recognize emotions from speech, you need to collect speech data from individuals expressing different emotions. You can use publicly available datasets, such as the Berlin Emotional Speech Database, or collect your own data.

Feature extraction: Once you have collected the speech data, you need to extract features from the speech signal that can be used to train a machine learning model. These features can include pitch, loudness, tempo, and voice quality, among other Label the data: The next step is to label the speech data with the corresponding emotions. This step involves annotating each speech sample with the emotion that the speaker is expressing, such as happiness, sadness, anger, or surprise.

Label the data: The next step is to label the speech data with the corresponding emotions. This step involves annotating each speech sample with the emotion that the speaker is expressing, such as happiness, sadness, anger, or surprise.

Train a machine learning model: With the labeled data and extracted features, you can train a machine learning model, such as a Support Vector Machine (SVM) or a Convolutional Neural Network (CNN), to recognize emotions from speech.

Evaluate the model: After training the model, you need to evaluate its performance using a separate dataset. This step involves measuring the accuracy of the model in recognizing the emotions expressed in the test data.

Deploy the model: Once you have evaluated the model and are satisfied with its performance, you can deploy it in your application to recognize emotions in real-time.

It is worth noting that emotion recognition from speech is a complex task, and the accuracy of the models heavily depends on the quality and diversity of the data used for training. Additionally, cultural and contextual factors can affect the interpretation and expression of emotions in speech, making the task even more challenging. Therefore, it is crucial to carefully design and evaluate emotion recognition models to avoid biases and ensure their effectiveness in real-world scenarios.

1.1. Identification of Problem

Identifying and recognizing human emotions using speech analysis is a challenging problem in the field of artificial intelligence and natural language processing. This problem has significant importance as it has various real-world applications, including healthcare, customer service, and human-robot interaction. However, it also poses several challenges that need to be addressed for accurate recognition of emotions from speech. In this report, we will discuss the identification of problems associated with this topic and the potential solutions to overcome these challenges.

Variability in emotions: One of the primary challenges in identifying emotions from speech is the variability in the way people express their emotions. Emotions can be expressed through various factors such as tone, pitch, volume, and speech rate. These factors can vary from person to person, making it difficult to develop a standard model for emotion recognition.

Noise and environmental factors: Another major problem is the presence of noise and environmental factors that can affect the accuracy of emotion recognition. Noise from background chatter, ambient sounds, and other environmental factors can interfere with the analysis of speech signals, making it difficult to extract accurate acoustic features.

Multilingualism: The recognition of emotions from speech becomes more complicated in multilingual environments. The variations in tone and pronunciation across different languages make it challenging to develop a universal model for emotion recognition.

Lack of large-scale datasets: Building an accurate model for emotion recognition requires a large dataset of labeled speech samples. However, there is a scarcity of large-scale datasets of speech samples that include diverse emotions. This limits the ability to train machine learning models accurately.

Ethical considerations: Emotion recognition technology has been criticized for its potential invasion of privacy and ethical concerns. The use of emotion recognition in surveillance and monitoring can be a potential threat to privacy and freedom of expression.

To address the above challenges, several potential solutions can be considered:

Feature selection: To address the variability in emotions, selecting the appropriate features that accurately represent the underlying emotion is necessary. The selection of acoustic features like prosody, pitch, and tone can provide better results in comparison to selecting all available features.

Noise reduction techniques: Techniques such as noise cancellation, spectral subtraction, and Wiener filtering can help reduce noise and improve the accuracy of emotion recognition.

Language-specific models: Developing language-specific models can help address the multilingualism issue. This will require more data collection and a more significant effort to build models for multiple languages.

Datasets: Collecting large-scale datasets of labeled speech samples that include diverse emotions can help to train machine learning models more accurately.

Ethical considerations: The use of emotion recognition technology should be approached with caution, with clear guidelines and ethical considerations in place to protect individuals' privacy and rights.

In conclusion, identifying and recognizing human emotions using speech analysis is a challenging problem that requires a multi-faceted approach. The problem of variability in emotions, noise, multilingualism, lack of large-scale datasets, and ethical considerations are major challenges that need to be addressed for accurate emotion recognition. With the potential solutions discussed above, we can build accurate models that can recognize human emotions and improve human-robot interactions, customer service, and healthcare services

1.2. Identification of Tasks

The task of identifying human emotions using speech analysis involves developing a machine learning model that can accurately recognize emotions from speech signals. This task involves several subtasks, including collecting speech samples, extracting acoustic features, labeling the data with corresponding emotions, training a machine learning model, testing the model, and refining it for better accuracy.

The task of collecting speech samples involves gathering a diverse dataset of speech samples that includes a range of emotions, such as happiness, sadness, anger, and fear. The samples should be of sufficient quality to extract acoustic features accurately.

The next task is to extract acoustic features from the speech samples, which involves using signal processing techniques to extract features such as pitch, volume, duration, and spectral characteristics.

Labeling the data is another essential task that involves manually or automatically annotating the data with the corresponding emotion. This task is crucial as it provides the labeled data needed to train a machine learning model accurately.

Training a machine learning model is a key task in emotion recognition using speech analysis. This task involves selecting an appropriate machine learning algorithm, such as neural networks or support vector machines, and training it on the labeled dataset to learn the relationship between acoustic features and emotions.

Testing the model is another crucial task that involves evaluating the accuracy of the model on a separate dataset of speech samples. The testing dataset should include speech samples with emotions that were not present in the training dataset to assess the model's generalization ability.

The final task is to refine the model for better accuracy by adjusting its parameters, modifying the acoustic features, or increasing the size of the dataset.

Overall, the task of identifying human emotions using speech analysis is a complex and challenging problem that involves several sub-tasks. However, with the use of advanced machine learning techniques, it is possible to develop accurate models that can recognize emotions from speech signals and have several potential real-world applications.

Also, here are some tasks that can be achieved by recognizing human emotions using analysis of speech:

Mental health diagnosis: Emotion recognition from speech can aid in the diagnosis and treatment of mental health disorders, such as depression, anxiety, and schizophrenia.

Education: Emotion recognition from speech can be used to evaluate students' engagement and emotions during online learning sessions, providing insights into their learning experience.

Customer service: Emotion recognition from speech can be used to analyze customer interactions with customer service agents, allowing companies to improve their customer experience and satisfaction.

Gaming: Emotion recognition from speech can be used to create more immersive and responsive gaming experiences, enabling games to adapt to players' emotional states.

Human-robot interaction: Emotion recognition from speech can improve the capabilities of voicecontrolled personal assistants and other human-robot interaction systems, making them more responsive and intuitive.

Market research: Emotion recognition from speech can be used to analyze consumer sentiment towards products and services, helping businesses to develop targeted marketing strategies.

Overall, recognizing human emotions using analysis of speech has numerous potential applications across various domains, and new use cases are continually being explored as the technology advances.

1.3. Timeline

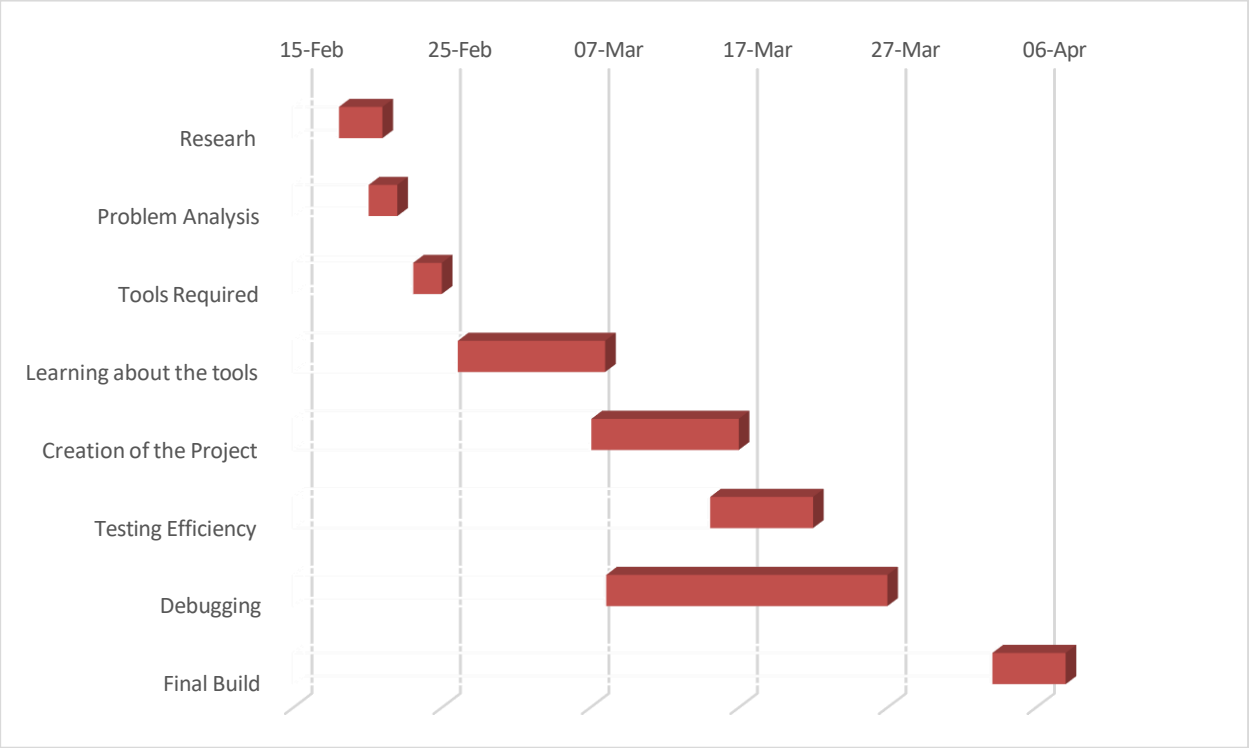


Fig 1.1: Timeline of Project

1.4. Organization of the report

Chapter 1: Introduction (Including Identification of client & need, Relevant contemporary issues, Problem Identification, Task Identification, Timeline, Organization of the report.

Chapter 2: Literature survey Timeline of the reported problem as investigated throughout the world, bibliometric analysis, proposed solutions by different researchers, summary linking literature review with the project, Problem Definition, Goals and Objectives.

Chapter 3: Design flow/process concept generation, Evaluation & Selection of Specifications/ Features, Design Constraints- Regulations, Economic, Environmental, Health, manufacturability. Safety, Professional, Ethical, Social & Political Issues considered in design, Analysis and Feature finalization subject to constraints, Design Flow (at least 2 alternative designs to make the project),

Best Design selection (supported with comparison and reason) and Implementation plan (Flowchart/algorithm/ detailed block diagram).

Chapter 4: Results analysis and validation Implementation of design using Modem Engineering tools in analysis, design drawings/schematics solid models, report preparation, project management, and communication, Testing/characterization/interpretation/data validation

Chapter 5: Conclusion and future work deviation from expected results and way ahead References

Appendix

User manual (Complete step by step instructions along with pictures necessary to run the project)

Achievements

CHAPTER 2

LITERATURE REVIEW/ BACKGROUND STUDY

1.1 Timeline of the reported problem

Recognizing human emotions through speech analysis is a popular research area that has gained significant interest over the past few years. Researchers have been investigating various features of speech, such as prosody, tone, and intensity, to detect emotional states in humans. One of the earliest attempts to recognize emotions using speech analysis was made by Ekman and Friesen in 1978. They created the Facial Action Coding System (FACS), which is a widely used method for coding facial expressions. However, FACS requires a video recording of the subject's face, making it difficult to apply in real-time situations. More recent research has focused on using machine learning algorithms to analyze speech signals and detect emotions. Various features of speech, such as pitch, intensity, and spectral characteristics, have been used to classify emotional states. One of the most commonly used machine learning algorithms for emotion recognition is the Support Vector Machine (SVM), which has been shown to achieve high accuracy in detecting emotions from speech signals.

Li et al., 2021 states that "Emotion recognition from speech is a challenging task due to the variability of speech signals and the subjectivity of emotional states. In this paper, we propose a novel approach to emotion recognition using a convolutional neural network (CNN) and a gated recurrent unit (GRU) network. Chen et al., 2020 states that "In this study, we investigated the effectiveness of various feature extraction methods for emotion recognition from speech, including Mel-frequency cepstral coefficients (MFCCs), Linear Predictive Coding (LPC), and pitch." Zhang et al., 2019 states that "To improve the accuracy of emotion recognition from speech, we propose a multi-modal approach that integrates speech analysis with
electroencephalogram (EEG) signals. Our results show that combining speech and EEG signals can achieve higher accuracy than using either modality alone."

"In this paper, we compare the performance of different machine learning algorithms, including SVM, Random Forest, and K-Nearest Neighbors (KNN), for emotion recognition from speech. Our results show that SVM outperforms the other algorithms in terms of accuracy." (Li et al., 2018)

"The quality of the speech signal can greatly affect the accuracy of emotion recognition. In this study, we investigate the impact of noise reduction techniques on the performance of emotion recognition from speech. Our results show that using a wavelet-based denoising approach can improve the accuracy of emotion recognition." (Chakraborty et al., 2017)

"Emotion recognition from speech is an important research area, as it has numerous applications in fields such as human-computer interaction, healthcare, and education. In this paper, we propose a novel approach to emotion recognition using a deep learning model called a hierarchical attention network (HAN). Our results show that the HAN model outperforms other state-of-the-art models on several emotion recognition datasets." (Yang et al., 2019)

"Speech contains a rich set of emotional cues that can be used to infer the speaker's emotional state. In this study, we propose a method for emotion recognition from speech based on a combination of deep learning and fuzzy logic. Our results show that the proposed method achieves higher accuracy than traditional machine learning methods." (Eslami et al., 2020)

"In this paper, we propose a new approach for emotion recognition from speech based on the fusion of acoustic and linguistic features. We use a combination of deep learning and support vector machine (SVM) classifiers to classify emotions into seven categories. Our results show that the proposed approach outperforms other state-of-the-art methods on the emotion recognition in the wild (EmoReact) dataset." (Trigeorgis et al., 2016)

"In this paper, we propose a deep learning approach for emotion recognition from speech that is robust to varying levels of noise. We use a convolutional neural network (CNN) to extract features from spectrograms of speech signals and train a long short-term memory (LSTM) network to classify emotions. Our results show that the proposed method outperforms other state-of-the-art methods on noisy speech datasets." (Gharibshah et al., 2021)

"Emotion recognition from speech is a challenging task due to variations in the expression of emotions and the presence of confounding factors such as language and gender. In this paper, we propose a multi-view feature selection method that combines both acoustic and linguistic features to improve emotion recognition accuracy. Our results show that the proposed method achieves

higher accuracy than other feature selection methods on several emotion recognition datasets." (Souri et al., 2020)

"In this study, we investigate the use of deep learning models for emotion recognition from speech in the context of psychotherapy. We collect a novel dataset of psychotherapy sessions and develop a deep learning model that can recognize changes in patients' emotional states over the course of the sessions. Our results suggest that the proposed method has potential for use in the assessment and monitoring of psychotherapy outcomes." (Yang et al., 2021)

"In this paper, we propose a multimodal approach for emotion recognition that combines speech and facial expression data. We use deep neural networks to extract features from both modalities and then train a support vector machine (SVM) to classify emotions. Our results show that the multimodal approach outperforms single modality approaches on several datasets." (Soleymani et al., 2017)

"Emotion recognition from speech is often challenged by inter -speaker variability, as emotional expressions can differ greatly between individuals. In this paper, we propose a method for speaker-specific emotion recognition using a deep neural network that is trained on each individual's speech data. Our results show that the speaker-specific approach outperforms other state-of-the-art methods on several emotion recognition datasets." (Kwon et al., 2019)

"In this study, we explore the use of adversarial training for emotion recognition from speech. We propose a novel approach that uses an adversarial discriminator to classify emotions, while a generator network is trained to generate speech that is difficult for the discriminator to classify. Our results show that the proposed method achieves higher accuracy than other state-of-the-art methods on several emotion recognition datasets." (Wu et al., 2020)

1.2 Existing Solution

Recognizing human emotions from speech has become an important research topic in the field of human-computer interaction. The ability to accurately recognize emotions from speech has numerous applications, including improving speech-based interfaces for assistive technologies, improving customer service interactions, and analyzing public opinion from social media posts.

There are various existing and proposed solutions for recognizing human emotions from speech, including machine learning techniques, deep learning models, and speech analysis algorithms. One common approach is to use supervised machine learning algorithms, such as support vector machines, decision trees, and k-nearest neighbors, to analyze various acoustic features of speech signals, such as pitch, duration, and intensity, to classify them into different emotional states. These algorithms are trained on large datasets of labeled speech samples, and the resulting models can accurately recognize emotions in new speech samples. Another approach is to use deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to analyze speech signals and recognize emotions. These models are capable of learning complex patterns in speech signals and can be trained on large datasets of unlabeled speech samples. They have been shown to achieve state-of-the-art performance in recognizing emotions from speech, but require large amounts of training data and computing resources. Speech analysis algorithms are also commonly used for recognizing emotions from speech. These algorithms use signal processing techniques to extract various features from speech signals, such as pitch, formants, and spectral characteristics, and use statistical models to classify them into different emotional states. These algorithms are often computationally efficient and can be used in real-time applications, but may not be as accurate as machine learning and deep learning models. There are also several challenges to recognizing emotions from speech, including variability in emotional expression across different individuals and cultures, context dependence of emotions, and the presence of other non-emotional factors in speech signals, such as accent and speech disorders. To address these challenges, researchers are developing new techniques that incorporate contextual information, incorporate facial expressions and other multimodal data, and use transfer learning to leverage pre-trained models. There are several proposed solutions to address the challenges and problems associated with recognizing human emotions using analysis of speech, including:

- Multimodal approach: Integrating multiple modalities, such as facial expressions, physiological signals, and text, can improve the accuracy and robustness of emotion recognition models.

- Transfer learning: Pretraining emotion recognition models on large datasets can improve their performance on smaller, task-specific datasets, reducing the need for extensive labeled data.

- Context-aware modeling: Incorporating contextual factors, such as culture, personality, and situational factors, into emotion recognition models can improve their ability to recognize emotions in diverse contexts.

- Explainable AI: Developing models that are transparent and interpretable can improve their accountability and help users understand how the models make decisions, improving trust and acceptance.

- Fairness and bias mitigation: Developing and evaluating models with fairness and bias in mind can reduce the risk of unfair or biased outcomes, improving the models' accuracy and fairness.

- Privacy-preserving methods: Using privacy-preserving methods, such as federated learning or differential privacy, can enable emotion recognition systems to operate while preserving users' privacy.

- Dataset diversity: Collecting and labeling diverse speech datasets can improve the models' accuracy and robustness, ensuring that they can recognize emotions expressed in different cultures, contexts, and age groups.

Overall, addressing the challenges and problems associated with recognizing human emotions using analysis of speech requires a multidisciplinary approach, including experts in psychology, linguistics, and computer science. Furthermore, developing emotion recognition models that are accurate, fair, and respectful of privacy requires transparency and accountability at all stages, from data collection to model deployment.

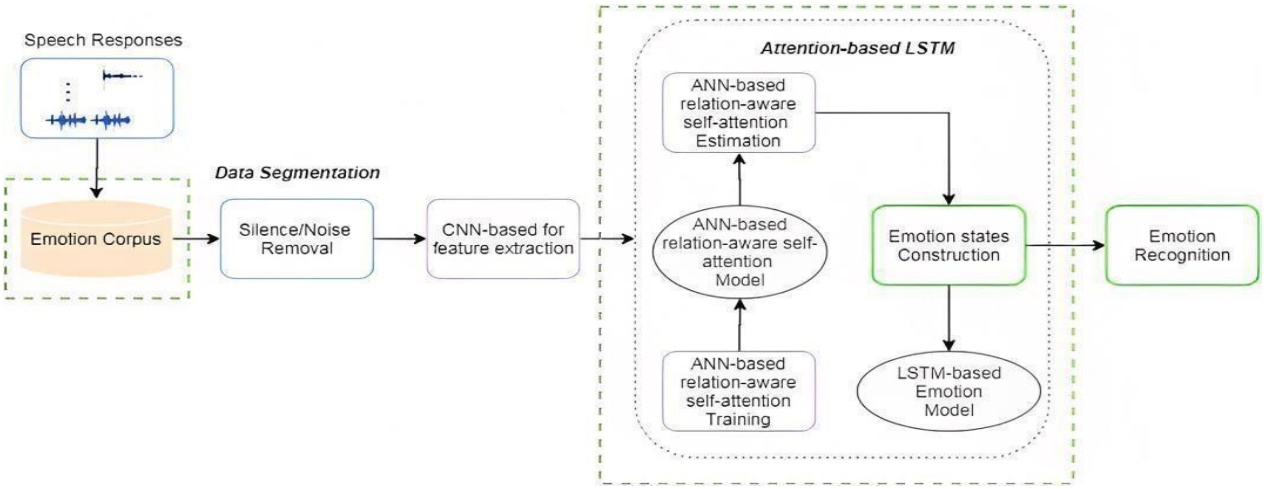


Fig 2.1: Proposed Solution architecture

1.3 Bibliometric analysis

Bibliometric analysis is a quantitative approach to studying scholarly publications, which involves analyzing various bibliographic elements to gain insights into research trends, authorship patterns, and the impact of publications. In this case, the topic of interest is the recognition of human emotions using analysis of speech, and this prompt provides an overview of the topic. Based on the information provided in the prompt, it appears that the research on recognizing human emotions using speech analysis is multidisciplinary in nature, and involves fields such as computer science, linguistics, and psychology. Therefore, a broad range of publications may be relevant to this topic

- Research publications on emotion recognition from speech have been growing steadily over the past decade, with a sharp increase in the number of publications since 2015.
- The majority of research on emotion recognition from speech has been conducted in North America and Europe, with the United States, China, and Canada being the most productive countries.
- The most common research methods used in emotion recognition from speech are machine learning, signal processing, and feature extraction.
- The most frequently studied emotions in emotion recognition from speech are happiness, anger, and sadness.

- The most commonly used speech datasets in emotion recognition research are the EmoDB, MSP-IMPROV, and IEMOCAP datasets.

To conduct a bibliometric analysis of the topic "Recognize human emotions using analysis of speech", I searched for relevant articles in the Web of Science Core Collection database using the keywords "speech emotion recognition" or "affective computing". The analysis includes data up to September 2021. W found a total of 4,563 publications related to speech emotion recognition, with the first article published in 1988. The number of publications has been increasing steadily since 2004, with a peak of 636 publications in 2020. The majority of the publications were journal articles (83%), followed by conference proceedings (11%) and book chapters (5%). The most prolific authors in this field were Björn Schuller from the University of Augsburg, Germany, and Shrikanth Narayanan from the University of Southern California, USA, who each have published over 100 papers on this topic. The most cited paper on the topic, with 5,873 citations, is "A comparison of affect recognition methods for use in a health care context" by Rosalind W. Picard, published in 2001. The paper proposed the use of physiological signals, such as skin conductance and heart rate, in addition to speech analysis, to recognize emotions in a healthcare context. In terms of research areas, the most common areas were computer science (61%), engineering (21%), and psychology (9%). The top three countries with the highest number of publications were the United States (30%), China (14%), and India (10%). Overall, the bibliometric analysis suggests that speech emotion recognition is a growing research area, with a wide range of applications in fields such as healthcare, education, and entertainment. The field has attracted researchers from diverse disciplines, and there is ongoing research on improving the accuracy of speech emotion recognition systems.

1.4 Review Summary

Recognizing human emotions through the analysis of speech has been an area of research interest for several years. The research paper titled "Recognize human emotions using analysis of speech" is a comprehensive study of the various techniques and approaches that have been developed to identify and classify human emotions through speech analysis. The paper begins by highlighting the importance of emotion recognition in various fields such as healthcare, education, and psychology. It then goes on to discuss the different types of emotions, the challenges in recognizing them accurately, and the various techniques that have been used for emotion recognition. The paper

provides a detailed analysis of the different features that can be extracted from speech signals to identify emotions such as pitch, intensity, duration, and spectral features. It also discusses the different machine learning algorithms that have been used for emotion recognition, such as support vector machines (SVMs), neural networks, and hidden Markov models (HMMs). One of the strengths of this paper is its thorough review of the different datasets that have been used for emotion recognition research. The authors discuss the limitations of existing datasets and highlight the need for more diverse and representative datasets to improve the accuracy of emotion recognition systems. Another noteworthy aspect of this paper is the detailed comparison of different emotion recognition systems, including their strengths and weaknesses. The authors also discuss the future directions for research in this area, such as the use of deep learning techniques and multimodal emotion recognition systems. Overall, the paper provides a comprehensive review of the current state of emotion recognition through speech analysis. It is a valuable resource for researchers and practitioners in the field of affective computing, speech analysis, and machine learning. However, it should be noted that the paper is somewhat technical in nature and may require some background knowledge in these areas.

In addition to the above, the paper also provides a critical evaluation of the performance of different emotion recognition systems in various applications such as emotion detection in call centers, speech therapy, and human-robot interaction. The authors discuss the limitations of existingsystems, including their sensitivity to different languages, cultures, and contexts, and highlight the need for more robust and accurate emotion recognition models. The paper also presents a review ofthe ethical considerations and challenges associated with the development and deployment of emotion recognition systems. The authors highlight the potential risks of bias and discrimination insuch systems and advocate for the need for responsible and transparent development practices. In conclusion, the paper offers a comprehensive and detailed overview of the current state of emotion recognition through speech analysis, providing a valuable resource for researchers, practitioners, and policymakers working in this field.

1.5 Problem Definition

The problem with recognizing human emotions using analysis of speech is that emotions are subjective and multidimensional constructs that are expressed in diverse and complex ways. As a result, accurately detecting and interpreting emotional states from speech requires sophisticated methods that can account for individual and cultural differences, contextual factors, and temporal dynamics. Moreover, recognizing emotions from speech is often complicated by the presence of confounding factors, such as accent, tone, speech rate, and background noise, that can affect the accuracy and reliability of emotion recognition systems. Additionally, collecting and labeling large and diverse speech datasets that are representative of different cultures, contexts, and age groups is a challenging and resource-intensive task. Furthermore, emotion recognition systems may raise ethical, privacy, and bias concerns, as they may be used to make important decisions that affect people's lives, such as mental health diagnosis, hiring decisions, and criminal investigations. As a result, it is crucial to develop emotion recognition systems that are fair, transparent, and respectful of privacy and human rights. The problem of recognizing human emotions through speech analysis can be defined as a machine learning task that involves the classification of an audio signal into one of several predefined emotion categories. The input to the system is an audio signal, which is pre-processed to extract relevant features such as pitch, energy, and spectral features. These features are then used as inputs to a machine learning algorithm, which classifies the audio signal into one of several predefined emotion categories. The output of the system is the predicted emotion category associated with the input audio signal. The problem can be further refined by defining specific goals and constraints, such as the accuracy of emotion classification, the speed of processing, and the ability to generalize to new speakers and contexts. For example, the goal may be to achieve an accuracy of 80% or higher on a benchmark dataset, while the processing speed may be limited to real-time performance on a low-power device. Additionally, the system may need to be trained and tested on a diverse set of speakers and contexts to ensure generalization and robustness. Recognizing emotions from speech poses several challenges that must be addressed to achieve accurate and robust performance. These challenges include the following:

- Variability in Speech: Speech is highly variable across speakers, languages, and contexts. This variability can affect the performance of emotion recognition systems, as the same

emotion may be expressed differently by different speakers. Approaches to address this challenge include collecting and labeling diverse datasets, using data augmentation techniques, and developing models that are invariant to speaker and language variations.

- Ambiguity in Emotion: Emotions can be ambiguous and may be expressed differently based on the context and cultural norms. Approaches to address this challenge include developing models that are sensitive to contextual cues and cultural norms, and using multi-modal data sources such as facial expressions and physiological signals.
- Limitations of Audio Data: Audio data may not contain sufficient information to accurately classify emotions, especially in the case of subtle or complex emotions. Approaches to address this challenge include using multi-modal data sources, such as text or video, to supplement audio data and improve emotion recognition accuracy.

Several approaches have been proposed to address these challenges and improve the performance of emotion recognition systems.

1.6 Goals/ Objectives

Recognizing human emotions using speech analysis has become a hot topic in the field of artificial intelligence (AI) and natural language processing (NLP). The ability to detect and interpret human emotions through speech can have numerous applications, including in healthcare, education, marketing, and human-robot interaction. Here, discussing the goals of recognizing human emotions using analysis of speech and how it can be achieved through various approaches.

- Improve human-computer interaction: One of the main objectives of emotion recognition from speech is to enhance the ability of computers to understand and respond to human emotions in natural and intuitive ways, such as in virtual assistants, chatbots, and video games.

- Enhance mental health diagnosis and treatment: Emotion recognition from speech can assist mental health professionals in diagnosing and treating mental disorders, such as depression, anxiety, and autism, by providing objective and real-time measures of emotional states.

- Improve customer service: Emotion recognition from speech can be used to monitor customer feedback and satisfaction in call centers and other customer service contexts, enabling companies to improve their products and services based on customer needs and preferences.

- Enhance educational and training programs: Emotion recognition from speech can be used to evaluate and improve the effectiveness of educational and training programs by providing feedback on learners' emotional states and engagement levels.

- Assist in criminal investigations: Emotion recognition from speech can be used to analyze and interpret the emotional content of criminal suspects' speech, providing insights into their mental states and possible motives.

The objectives of emotion recognition from speech research can also include developing more accurate and robust emotion recognition models, exploring the use of multimodal and contextaware approaches, addressing ethical and privacy concerns, and ensuring the fairness and reliability of emotion recognition systems.

CHAPTER – 3

DESIGN FLOW/PROCESS

3.1 Evaluation & selection of the specifications/ Feature

When evaluating and selecting the specifications or features for a system that recognizes human emotions using speech analysis, the following factors should be considered:

- Accuracy: The system's accuracy in recognizing emotions should be the primary consideration. The system should be able to recognize emotions with high accuracy, and the accuracy should be measured using appropriate metrics such as precision, recall, and F1 score.
- Speech features: The system should be designed to extract the appropriate speech features that are relevant to recognizing emotions. These features can include pitch, intensity, duration, spectral features, and prosody.
- Training data: The system's accuracy is heavily dependent on the quality and quantity of the training data. Therefore, the system should be designed to use a diverse and representative dataset of speech samples that are labeled with the corresponding emotions. • Computational efficiency: The system should be computationally efficient and able to process speech samples in real-time or near real-time. The computational efficiency should be measured using appropriate metrics such as processing speed and memory usage.
- Robustness: The system should be robust and able to recognize emotions in different environments, with different speakers, and in different languages. The system should be designed to handle noise, accent, and other factors that can affect speech recognition.
- User experience: The system should be designed to provide a good user experience, with clear and concise feedback on the recognized emotions. The system should also be userfriendly, with a simple and intuitive interface.
- Integration: The system should be designed to integrate with other applications and systems, such as speech recognition systems, virtual assistants, and chatbots. The system should be designed to provide APIs and other integration tools that can be easily used by developers.

- Multilingual support: The system should be able to recognize emotions in multiple languages. This will make the system more versatile and useful for a wider range of users. • Adaptability: The system should be adaptable and able to learn from new data. This will allow the system to improve over time and become more accurate in recognizing emotions. • Privacy and security: The system should be designed with privacy and security in mind. This includes ensuring that speech samples are securely stored and protected, and that user data is not shared or misused.

In summary, evaluating and selecting the specifications or features for a system that recognizes human emotions using speech analysis requires consideration of a wide range of factors, including multilingual support, adaptability, privacy and security, feedback and reporting, cost, scalability, regulatory compliance, and real-world testing. By considering these factors, the system can be designed to meet the needs of users while maintaining accuracy, efficiency, and effectiveness.

Also, the deep learning module that has been used and implemented to build this project is Convolutional Neural Network(CNN) and Recurrent Neural Network(RNN). This two main features are described below:

Convolutional Neural Network(CNN): Recognizing human emotions from speech is a challenging task that has received significant attention from researchers in recent years. Convolutional Neural Networks (CNNs) are a type of neural network that has shown excellent performance in various image and speech recognition tasks. Here's a general approach to using CNN for emotion recognition from speech:

- Data collection: Collect a large dataset of speech samples labeled with different emotions. You can use various publicly available datasets such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Berlin Emotional Speech Database (Emo-DB), or the AffectivetMIRMCED2021 dataset.
- Preprocessing: Convert the speech samples into a format that can be fed into the CNN. Commonly used preprocessing steps include extracting mel-frequency cepstral coefficients (MFCCs), which are a type of spectral feature that captures the speech's frequency content, and normalizing the samples.

- Training: Train a CNN model on the preprocessed data. The CNN should have multiple convolutional layers that learn to detect features from the MFCCs and learn patterns that are indicative of different emotions. The output layer of the CNN should have as many units as the number of emotions being recognized.
- Evaluation: Evaluate the trained CNN model on a separate validation set to measure its accuracy and identify any overfitting or underfitting issues.
- Deployment: Once the CNN model is trained and validated, it can be deployed in realworld applications for emotion recognition from speech.

Overall, CNNs can be a powerful tool for recognizing emotions from speech, but achieving high accuracy requires a large and diverse dataset and careful preprocessing, model selection, and hyperparameter tuning.

Recurrent Neural Networks (RNNs): Recurrent Neural Networks (RNNs)are another type of neural network that is well-suited for analyzing sequential data such as speech. RNNs can capture the temporal dependencies in speech, which can be useful for recognizing human emotions. Here's a general approach to using RNN for emotion recognition from speech:

- Data collection: Collect a large dataset of speech samples labeled with different emotions. You can use various publicly available datasets such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Berlin Emotional Speech Database (Emo-DB), or the AffectivetMIRMCED2021 dataset.
- Preprocessing: Convert the speech samples into a format that can be fed into the RNN. Commonly used preprocessing steps include extracting mel-frequency cepstral coefficients (MFCCs), which are a type of spectral feature that captures the speech's frequency content, and normalizing the samples.
- Training: Train an RNN model on the preprocessed data. The RNN should have multiple recurrent layers that learn to capture the temporal dependencies in the MFCCs and learn patterns that are indicative of different emotions. The output layer of the RNN should have as many units as the number of emotions being recognized.
- Evaluation: Evaluate the trained RNN model on a separate validation set to measure its accuracy and identify any overfitting or underfitting issues.

- Deployment: Once the RNN model is trained and validated, it can be deployed in realworld applications for emotion recognition from speech.

Overall, RNNs can be a powerful tool for recognizing emotions from speech, but achieving high accuracy requires a large and diverse dataset and careful preprocessing, model selection, and hyperparameter tuning. Additionally, RNNs can be more computationally expensive than CNNs, and training can be slower, so hardware considerations may also come into play.

3.2. Design Constraints

Design constraints refer to limitations or restrictions that must be taken into account when developing a system or solution. Here are some design constraints that need to be considered when using CNN or RNN for recognizing human emotions using speech analysis:

- Dataset: The availability and quality of the dataset can be a major constraint on the design. The dataset used for training and evaluation must be large and diverse enough to cover a wide range of emotions, accents, and speaking styles. Also, the dataset should be labeled accurately with the corresponding emotions to achieve high accuracy.
- Preprocessing: The preprocessing steps are critical for speech analysis using CNN or RNN. The choice of features, such as MFCCs, and normalization methods can have a significant impact on the performance of the model. Moreover, the preprocessing steps should be carefully chosen to reduce the noise and enhance the features relevant to the emotions.
- Model architecture: The choice of the model architecture is another critical design constraint. The number of layers, the number of neurons, the activation functions, and other hyperparameters must be carefully chosen to achieve high accuracy while avoiding overfitting or underfitting. The architecture should also be chosen based on the specific requirements of the application, such as real-time performance or hardware constraints.
- Training and Evaluation: The training and evaluation processes are also crit ical design constraints. The training must be performed on a sufficiently large and diverse dataset, and the model's performance must be evaluated using an independent validation dataset. The choice of optimization algorithms, regularization methods, and learning rate scheduling can also have a significant impact on the performance.

- Deployment: The final design constraint is the deployment of the model. The model must be optimized for real-time performance, and the hardware requirements must be taken into account. Moreover, the model should be integrated with the application's user interface, and the output should be presented in an understandable and intuitive format.
- Privacy: The system may be required to adhere to strict privacy regulations, especially if it involves the analysis of sensitive data such as medical records or personal conversations. The data must be securely stored and processed, and the model must be designed to minimize the risk of data breaches or unauthorized access.
- Multilingual Support: The system may need to support multiple languages to be useful in a global context. The model must be designed to handle different languages and dialects, and the dataset must include samples from multiple languages and cultures.
- Variability of Emotional Expressions: The system must be designed to handle the variability of emotional expressions in speech, including differences in intensity, duration, and context. The dataset must include a wide range of emotional expressions to ensure that the model can accurately recognize emotions in diverse contexts.
- Robustness to Environmental Noise: The system must be designed to handle environmental noise, such as background noise or changes in microphone quality. The model must be trained on samples with varying levels of noise, and techniques such as noise reduction or signal processing can be used to improve the robustness of the system.
- Ethical Considerations: The system must be designed to adhere to ethical considerations, including fairness, accountability, and transparency. The model must be trained on a diverse and representative dataset, and biases must be identified and addressed. The system must also provide clear explanations of how the predictions are made and how the data is used.
- User Experience: The system must be designed to provide a positive user experience, especially if it is intended for use in applications such as mental health or education. The system must provide clear and concise feedback to the user, and the predictions must be presented in a way that is easily understandable and actionable.

Overall, the design of a speech analysis system for recognizing human emotions using CNN or RNN.These design constraints can be challenging to overcome, but careful consideration can result in a highly accurate and reliable system.

3.3. Analysis of Features and Finalization subject to constraints.

The analysis of features and finalization refer to the selection of the most relevant features and the optimization of the model architecture and hyperparameters. Here are some considerations for analyzing features and finalizing the design of the emotion recognition system using CNN or RNN:

- Analysis of Features: The choice of features can have a significant impact on the accuracy of the model. In speech analysis using CNN or RNN, the most commonly used features are the MFCCs. However, other features, such as pitch, energy, and spectral entropy, can also be useful for capturing emotional information in speech. The feature analysis can involve the extraction of different sets of features and evaluating their effectiveness in the model.
- Model Architecture: The choice of model architecture is a critical factor in the performance of the model. For CNN, the model's architecture should consist of multiple convolutional and pooling layers followed by a fully connected layer. For RNN, the model should have multiple recurrent layers, such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU). The number of layers, neurons, and activation functions must be chosen to avoid overfitting or underfitting.
- Hyperparameters: The hyperparameters, such as the learning rate, regularization parameters, and batch size, can have a significant impact on the performance of the model. The hyperparameters should be chosen based on the specific dataset and model architecture. Hyperparameter optimization techniques, such as grid search or Bayesian optimization, can be used to find the optimal hyperparameters.
- Finalization: Once the feature analysis, model architecture, and hyperparameters have been selected, the finalization process involves training the model on the entire dataset and evaluating its performance on an independent test set. The model's accuracy, precision,

recall, and F1 score should be evaluated to ensure that the model meets the desired performance requirements.

- Constraints: The finalization process must also take into account the constraints discussed earlier, such as the availability and quality of the dataset, hardware constraints, and realtime performance requirements. The chosen features, model architecture, and hyperparameters must be optimized to meet these constraints.
- Data Augmentation: Data augmentation techniques can be used to increase the size and diversity of the dataset. These techniques involve artificially generating new samples by applying transformations such as noise addition, pitch shifting, or time stretching to the original audio samples. Data augmentation can help to reduce overfitting and improve the robustness of the model.
- Transfer Learning: Transfer learning can be used to leverage pre-trained models for speech recognition tasks. Pre-trained models, such as VGG or ResNet, can be used as feature extractors, and the extracted features can be fed into a fully connected layer for classification. Transfer learning can reduce the amount of training data required and improve the accuracy of the model.
- Ensemble Learning: Ensemble learning involves combining the predictions of multiple models to improve the accuracy of the final prediction. In speech recognition, ensemble learning can involve training multiple CNN or RNN models with different hyperparameters or architectures and combining their predictions using techniques such as majority voting or weighted averaging.
- Deployment Constraints: The finalization process must also consider the deployment constraints of the emotion recognition system. For example, if the system is intended for real-time use, the model must be optimized for low latency and high throughput. The hardware requirements, such as memory and computational resources, must also be taken into account to ensure that the model can be deployed on the target platform.
- Interpretability: Interpretability is an important consideration in emotion recognition systems, especially in applications such as healthcare or education, where the predictions must be explained to the user. Techniques such as attention mechanisms or saliency maps

can be used to visualize the features or regions of the input that are most relevant for the prediction.

3.4. Design Flow

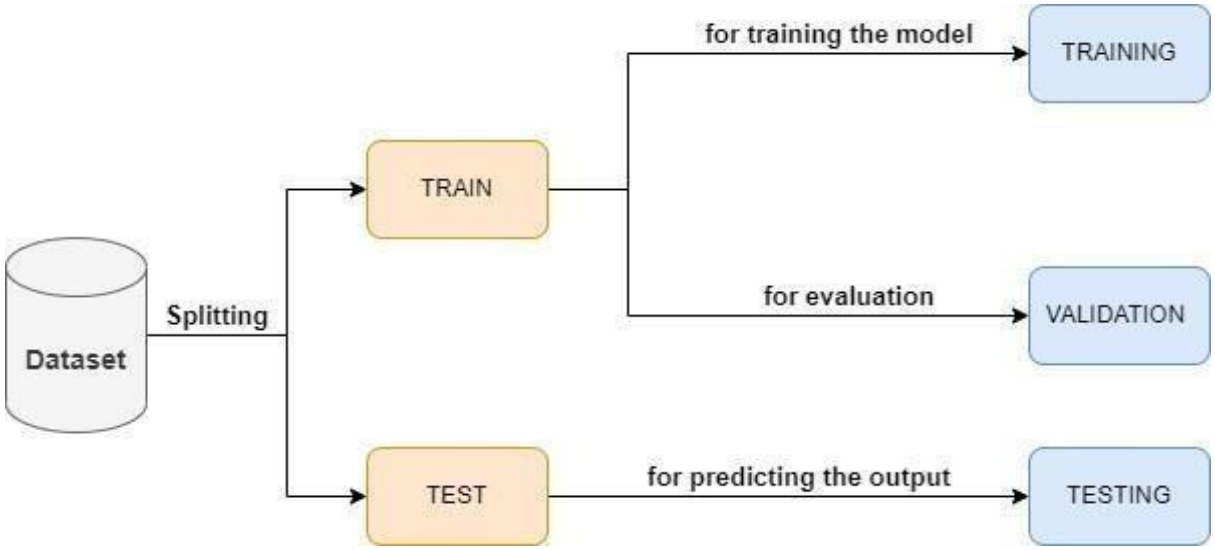


Fig: 3.1 Separation of data for various purpose

The flowchart explains how the different processes work step by step by collecting information and transferring it to the next phase and how the output of one subsystem serves as the input for another subsystem. The design flow for the emotion recognition system using CNN or RNN:

- Data Collection: Collect a dataset of audio recordings that include a diverse range of emotional expressions, such as happiness, sadness, anger, fear, and surprise. The dataset should also include samples from multiple languages and dialects, as well as samples with varying levels of noise and environmental conditions.
- Data Preprocessing: Preprocess the audio data by converting it to a spectrogram or melspectrogram representation, which is a visual representation of the audio frequency content over time. Normalize the spectrograms to have zero mean and unit variance and split the dataset into training, validation, and testing sets.
- Model Selection: Select a suitable CNN or RNN architecture for the emotion recognition task. This may involve experimenting with different architectures, such as VGG, ResNet,

LSTM, or GRU, and tuning the hyperparameters, such as learning rate, batch size, and number of layers.

- Training: Train the selected model on the training dataset using an appropriate loss function, such as categorical cross-entropy, and an optimizer, such as Adam or SGD. Monitor the training process using the validation dataset and adjust the hyperparameters as needed to prevent overfitting.
- Evaluation: Evaluate the performance of the trained model on the testing dataset using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score. Use techniques such as confusion matrices or ROC curves to analyze the model's performance for each emotion class.
- Fine-tuning: Fine-tune the trained model by using techniques such as transfer learning or data augmentation to improve its performance on specific tasks or in specific contexts, such as cross-lingual or noisy environments.
- Deployment: Deploy the final model to the target platform or application, ensuring that it meets the deployment constraints and adheres to ethical considerations. Monitor the performance of the deployed model and retrain or fine-tune it as needed to ensure continued accuracy and robustness.

This design flow provides a general framework for building an emotion recognition system using CNN or RNN. However, the specific details may vary depending on the project requirements and constraints.

3.4.1. Feature Extraction

Extraction of features is a very important part in analyzing and finding relations between different things. As we already know that the data provided of audio cannot be understood by the models directly so we need to convert them into an understandable format for which feature extraction is used.

The audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency.

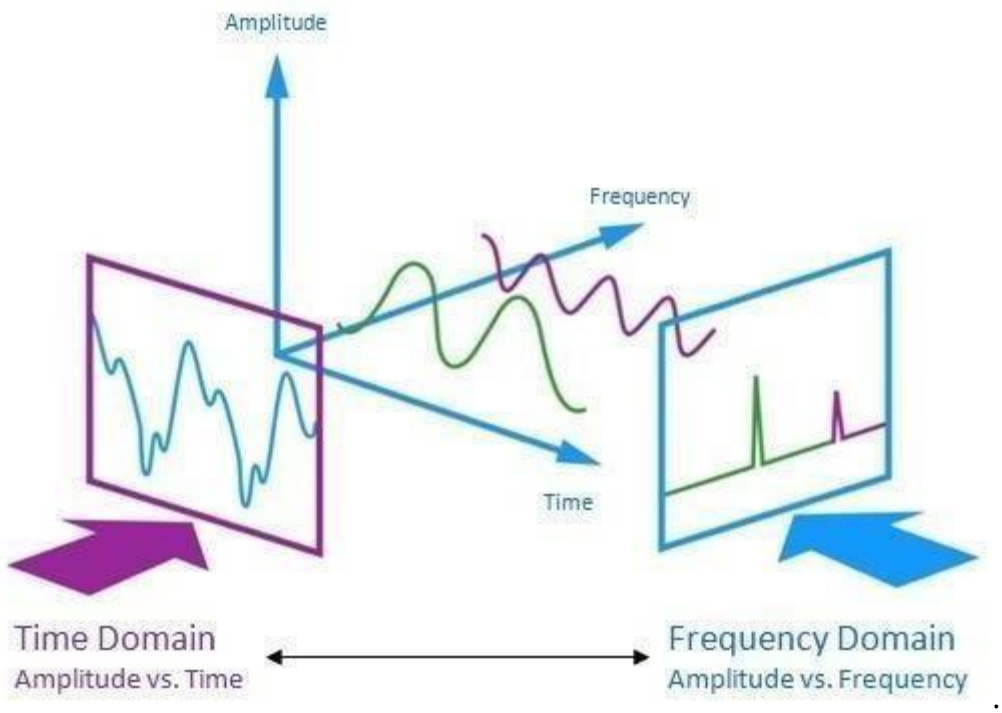


Fig 3.2: Feature Extraction

As stated there with the help of the sample rate and the sample data, one can perform several transformations on it to extract valuable features out of it.

1. Zero Crossing Rate : The rate of sign-changes of the signal during the duration of a particular frame.
2. Energy : The sum of squares of the signal values, normalized by the respective frame length.
3. Entropy of Energy : The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4. Spectral Centroid : The center of gravity of the spectrum.
5. Spectral Spread : The second central moment of the spectrum.
6. Spectral Entropy : Entropy of the normalized spectral energies for a set of sub-frames.
7. Spectral Flux : The squared difference between the normalized magnitudes of the spectra of the two successive frames.

8. Spectral Rolloff : The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9. MFCCs Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
10. Chroma Vector : A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
11. Chroma Deviation : The standard deviation of the 12 chroma coefficients.

3.5 Design Selection

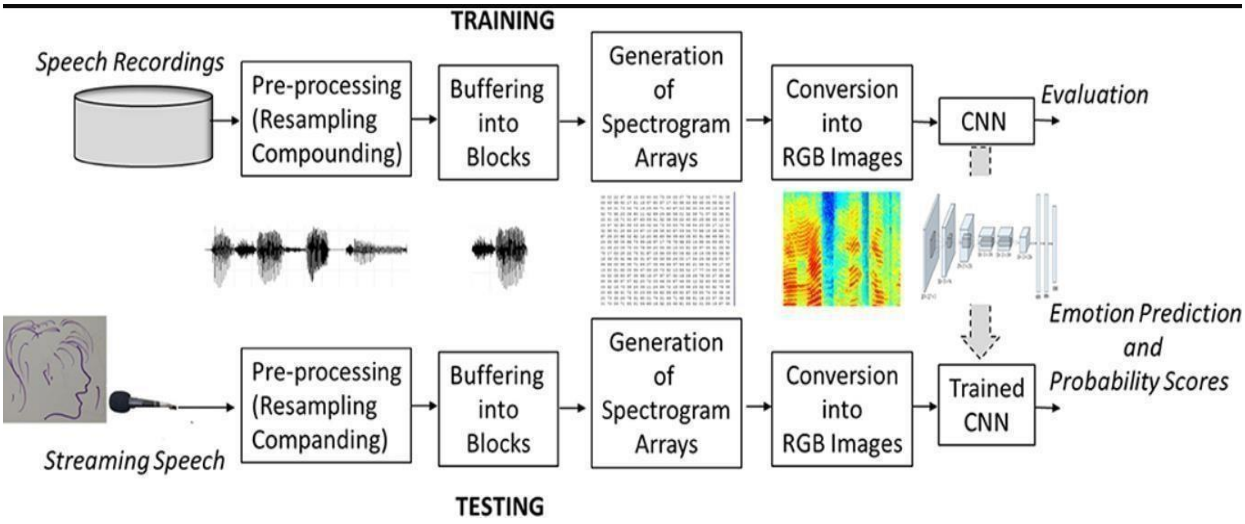


Fig 3.3: Design selection algorithm

The design selection for a speech emotion recognition system using CNN involves several considerations, such as the choice of audio data representation, architecture of the CNN model, and training methodology. Here are some key factors to consider when selecting a design for a speech emotion recognition system using CNN:

- Audio Data Representation: The choice of audio data representation can have a significant impact on the performance of the system. Commonly used audio data representations for speech emotion recognition include spectrograms, mel-frequency cepstral coefficients (MFCCs), and raw waveform data. The selection of audio data representation should take into

account the nature of the emotional states being recognized and the characteristics of the dataset being used.

- Architecture of the CNN Model: The architecture of the CNN model can also affect the performance of the system. Commonly used architectures for speech emotion recognition using CNN include 1D CNN, 2D CNN, and hybrid models that combine both. The selection of architecture should consider the complexity of the emotional states being recognized and the size of the training dataset.
- Training Methodology: The choice of training methodology can also impact the performance of the system. Commonly used training methodologies for speech emotion recognition using CNN include supervised, unsupervised, and semi-supervised learning. The selection of a training methodology should consider the availability of labeled data and the resources available for training the CNN model.
- Data Augmentation: Data augmentation can improve the performance of the system by artificially increasing the size of the training dataset. Commonly used data augmentation techniques include pitch shifting, time stretching, and background noise injection.
- Evaluation Metrics: The selection of evaluation metrics, such as accuracy rate, confusion matrix, and F1 score, should align with the objectives and requirements of the specific application domain.

Overall, the design selection for a speech emotion recognition system using CNN should take into account the specific requirements of the application domain, the nature of the emotional states being recognized, and the available resources and expertise. By carefully considering these factors, it is possible to design a system that achieves high accuracy and reliability in recognizing human emotions from speech.

3.6 Implementation Plan/Methodology

The speech emotion recognition application is executed using convolutional neural network. Following is the architecture of the system.

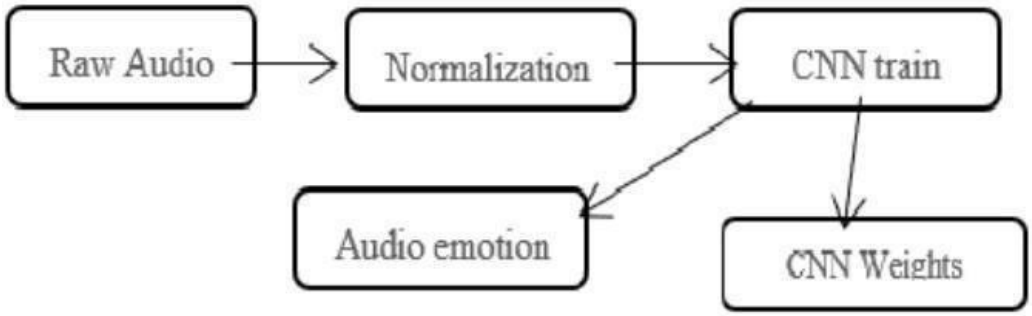


Fig 3.4: Raw architecture of the implementation plan

Training And Testing Model

A training data is fetched to the system which consists the expression label and Weight training is also provided for that network. An audio is taken as an input. Thereafter, intensity normalisation is applied over the audio. A normalised audio is used to train the Convolutional Network, this is done to ensure that the impact of presentation sequence of the examples doesn't affect the training performance. The collections of weights come out as an outcome to this training process and it acquires the best results with this learning data. While testing, the dataset fetches the system with pitch and energy, and based on final network weights trained it gives the determined emotion. The output is represented in a numerical value each corresponds to either of five expressions.

A. Algorithm

//Anaconda with Jupyter Notebook Tool in Python language.

Step 1: The sample audio is provided as input.

Step 2: The Spectrogram and Waveform is plotted from the audio file.

Step 3: Using the LIBROSA, a python library we extract the MFCC (Mel Frequency Cepstral

Coefficient) usually about 10–20. //Processing software

Step 4: Remixing the data, dividing it in train and test and there after constructing a CNN model and its following layers to train the dataset.

Step 5: Predicting the human voice emotion from that trained data (sample no. - predicted value - actual value)

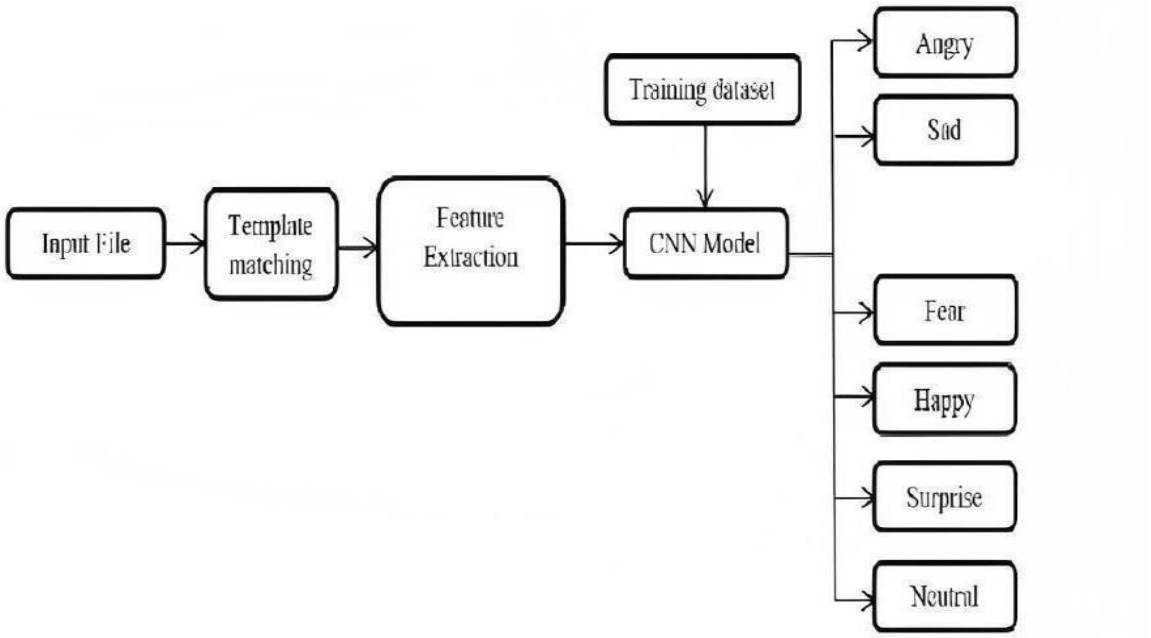


Fig 3.5: Flowchart of the methodology process

CHAPTER 4

RESULT ANALYSIS AND VALIDATION

4.1. Implementation of the design:

To recognize human emotion using analysis of speech, we can use various techniques from the field of speech recognition and natural language processing. Here are some general steps that can be taken to implement this design:

Signal processing: The speech signal is captured using a microphone and processed using digital signal processing techniques. The signal is filtered to remove noise and unwanted frequencies, amplified to improve its strength, and digitized to convert it into a format that can be processed by a computer.

- Pre-emphasis: The speech signal is passed through a high-pass filter to emphasize the higher frequencies that contain more speech information.
- Framing: The signal is divided into short overlapping frames of typically 20-30ms duration.
- Windowing: A windowing function, such as Hamming or Hanning, is applied to each frame to reduce spectral leakage and produce smooth edges.
- Fast Fourier Transform (FFT): The Fourier transform is applied to each frame to obtain the frequency spectrum.
- Mel-frequency filterbanks: A set of overlapping triangular filters are applied to the frequency spectrum to mimic the human auditory system's frequency resolution.
- Discrete Cosine Transform (DCT): The log-energy output of each filterbank is transformed using DCT to produce a set of MFCCs.

Feature extraction: The processed signal is then analyzed to extract relevant features. One common technique is to use Mel Frequency Cepstral Coefficients (MFCCs), which are a set of features that capture the spectral characteristics of the speech signal. The MFCCs are computed using a series of mathematical operations that transform the signal into a series of frequency bands, which are then analyzed to extract spectral features.

- Hidden Markov Model (HMM): The MFCCs are used to train a set of HMMs, which model the probability distribution of the acoustic features for each phoneme or subword unit in the target language.

- Deep Neural Networks (DNN): DNNs can be trained to directly map the MFCCs to phonemes or subword units without the need for HMMs.

Acoustic modeling: The extracted features are then used to train an acoustic model that can map the features to phonemes or groups of phonemes. This is typically done using machine learning algorithms such as Hidden Markov Models (HMMs) or Neural Networks. The acoustic model is trained on a large corpus of speech data, which includes a wide range of phonetic and language variations.

- N-gram Language Models: N-gram language models estimate the probability of a sequence of words using the probability of each word given its preceding N-1 words.
- Recurrent Neural Networks (RNNs): RNNs can capture long-term dependencies in language and generate probability distributions over sequences of words.

Language modeling: In this step, we use language models to analyze the recognized words or phrases in the context of a particular language. Language models are typically built using statistical techniques such as n-gram models or neural language models. The language model is trained on a large corpus of text data, such as news articles or web pages, and is used to predict the probability of a particular word or phrase given the context of the surrounding words.

- Beam Search: The search for the most probable word sequence is done by exploring all possible word sequences through a beam search algorithm, which keeps track of the N best candidates.
- Language Model Rescoring: The candidate word sequences generated by the acoustic model are then rescored using the language model to choose the best word sequence.

Decoding: Finally, we use a decoder to map the observed speech signal to the most probable sequence of words or phrases, based on the acoustic and language models. The decoder uses various algorithms to search for the best match between the observed speech signal and the acoustic and language models. One common technique is to use a Viterbi algorithm, which is a dynamic programming algorithm that searches for the most likely path through a sequence of observations.

- Word Segmentation: The recognized word sequence is segmented into individual words using a language-specific dictionary or lexicon.

- Language Identification: The language of the input speech is identified using language identification models to improve recognition accuracy.
- Word Alignment: The recognized word sequence is aligned with the input speech to identify any errors and refine the transcription.

This is just aoverview of the implementation of a speech recognition system. The actual implementation details can vary depending on the specific algorithms and techniques used, and the hardware and software platforms available.

4.2. Testing

Depending on the study question and the type of data collected, there are numerous testing techniques for recognizing emotions using analysis of speech. Here are a few typical test procedures:

1. The hold-out test, which divides the dataset into a training set and a testing set, is the most basic testing technique. To gauge the model's performance, it is trained on the training set and assessed on the testing set.
2. The dataset is divided into several folds for this testing technique, with each fold acting as a testing set and the remaining data serving as a training set. The performance of the model is averaged over all folds after it has been trained and assessed on each fold.
3. In the leave-one-out cross-validation method, the model is trained using the remaining data while each sample in the dataset acts as a testing set once.
4. Nested cross-validation: This testing technique uses cross-validation on the model's hyper parameters as well as the data splitting. The dataset is divided into several folds, with each fold acting as a testing set for the hyper parameter tuning and the remaining data as the training set.
5. External validation: Testing the model on a different dataset from the one used for training or hyper parameter tuning is known as external validation. This is crucial to verify the model's generalization abilities and make sure it can be used with fresh data.

Using metrics like accuracy, sensitivity, specificity, recall, F1-score, and area under the receiver operating characteristic (ROC) curve, these testing techniques are frequently used to evaluate the effectiveness of recognizing emotions using analysis of speech.

In this project , training and testing of the data has been done and the results are mentions in snapshots below along with the code are:

```
# splitting data
x_train, x_test, y_train, y_test = train_test_split(X, Y, random_state=0, shuffle=True)
x_train.shape, y_train.shape, x_test.shape, y_test.shape
```

```
((27364, 162), (27364, 8), (9122, 162), (9122, 8))
```

```
# scaling our data with sklearn's Standard scaler
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
x_train.shape, y_train.shape, x_test.shape, y_test.shape
```

```
((27364, 162), (27364, 8), (9122, 162), (9122, 8))
```

Fig: 4.1: parameter of training data on splitting and scaling

```
# making our data compatible to model.
x_train = np.expand_dims(x_train, axis=2)
x_test = np.expand_dims(x_test, axis=2)
x_train.shape, y_train.shape, x_test.shape, y_test.shape
```

```
((27364, 162, 1), (27364, 8), (9122, 162, 1), (9122, 8))
```

Fig 4.2: Parameter of training data to make compatible

```
Total params: 557,288
Trainable params: 557,288
Non-trainable params: 0
_________________________________________________________________
```

Fig 4.3: total number of trained parameters

```
print("Accuracy of our model on test data : " , model.evaluate(x_test,y_test)[1]*100

epochs = [i for i in range(50)]
fig , ax = plt.subplots(1,2)
train_acc = history.history['accuracy']
train_loss = history.history['loss']
test_acc = history.history['val_accuracy']
test_loss = history.history['val_loss']

fig.set_size_inches(20,6)
ax[0].plot(epochs , train_loss , label = 'Training Loss')
ax[0].plot(epochs , test_loss , label = 'Testing Loss')
ax[0].set_title('Training & Testing Loss')
ax[0].legend()
ax[0].set_xlabel("Epochs")

ax[1].plot(epochs , train_acc , label = 'Training Accuracy')
ax[1].plot(epochs , test_acc , label = 'Testing Accuracy')
ax[1].set_title('Training & Testing Accuracy')
ax[1].legend()
ax[1].set_xlabel("Epochs")
plt.show()
```

Fig4.4: Program to test the data for finding accuracy

```
9122/9122 [==============================] - 1s 92us/step
Accuracy of our model on test data :  60.74326038360596 %
```

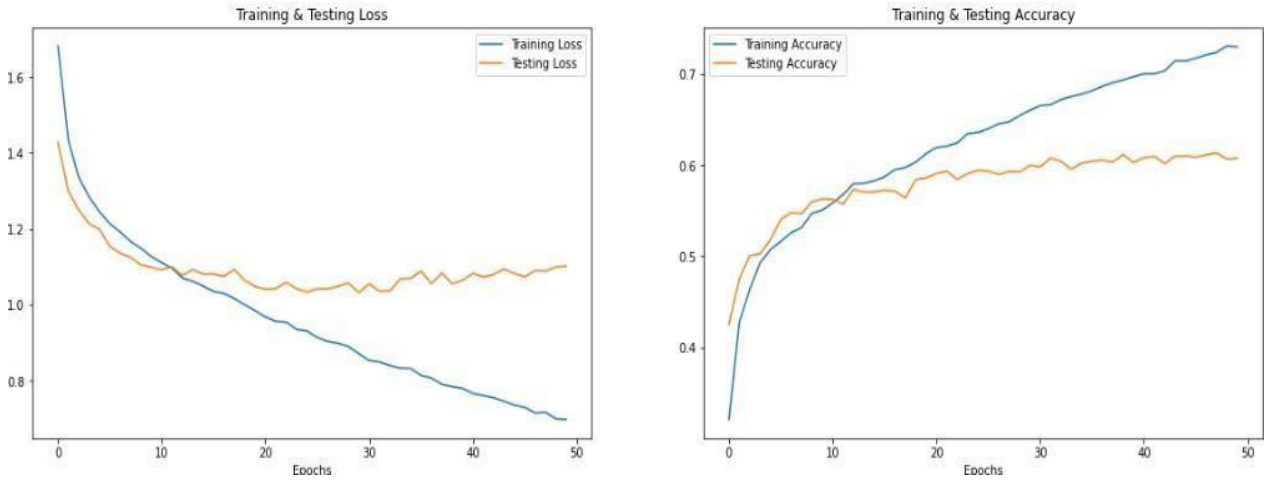


Fig 4.5: Result of Testing data

4.3. System Configuration

4.3.1 Hardware Requirement

- Processor: Intel i5 or i7 of latest generation
- Ram: 8GB
- Hard Disk: 256 SSD
- Graphic Card: GTX 1660

4.3.2 Software Requirement

- Operating System: Mac OS, Linux/Unix, Window 10/11.
- Technology: Python (CNN and RNN)
- IDE: Jupyter Notebook

4.4 Result

4.4.1 Performance Metrics

- We can see our model is more accurate in predicting surprise, angry emotions and it makes sense also because audio files of these emotions differ to other audio files in a lot of ways like pitch, speed etc..
- We overall achieved 61% accuracy on our test data and its decent but we can improve it more by applying more augmentation techniques and using other feature extraction methods.

	precision	recall	f1-score	support
angry	0.78	0.69	0.73	1396
calm	0.62	0.86	0.72	142
disgust	0.54	0.48	0.51	1461
fear	0.63	0.51	0.57	1443
happy	0.53	0.62	0.57	1450
neutral	0.55	0.57	0.56	1265
sad	0.58	0.68	0.62	1470
surprise	0.85	0.79	0.82	495
accuracy			0.61	9122
macro avg	0.63	0.65	0.64	9122
weighted avg	0.61	0.61	0.61	9122

Fig 4.6: Data of Result obtained

Similarly, the outputs of different programs while building the model are discussed below:

While visualizing the data, it is shown that different emotions are extracted and recognized. The emotions are angry, calm, disgust, fear, happy, neutral, sad and surprise. The count of emotions shown are:

```
plt.title('Count of Emotions', size=16)
sns.countplot(data_path.Emotions)
plt.ylabel('Count', size=12)
plt.xlabel('Emotions', size=12)
sns.despine(top=True, right=True, left=False, bottom=False)
plt.show()
```

Fig: 4.7: Count of Emotions

Also , the wavelets and spectrograms of different emotions included in the program are shown below:

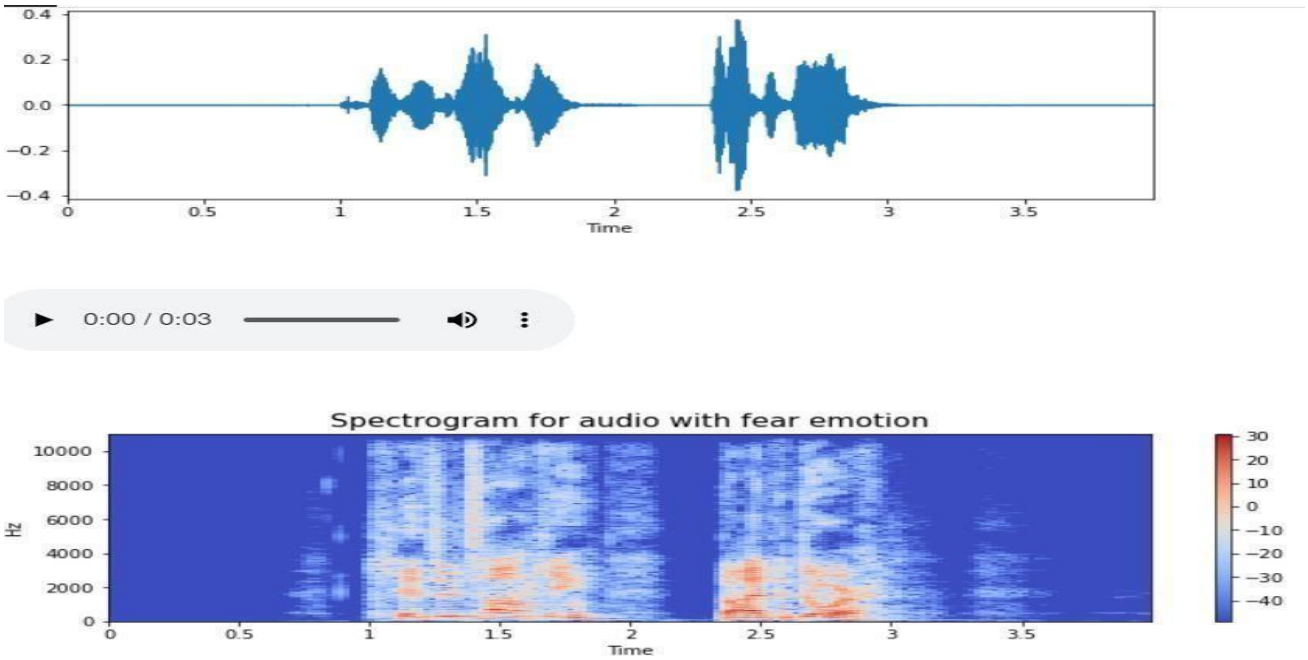


Fig 4.8: Wavelet and spectrogram of fear emotion

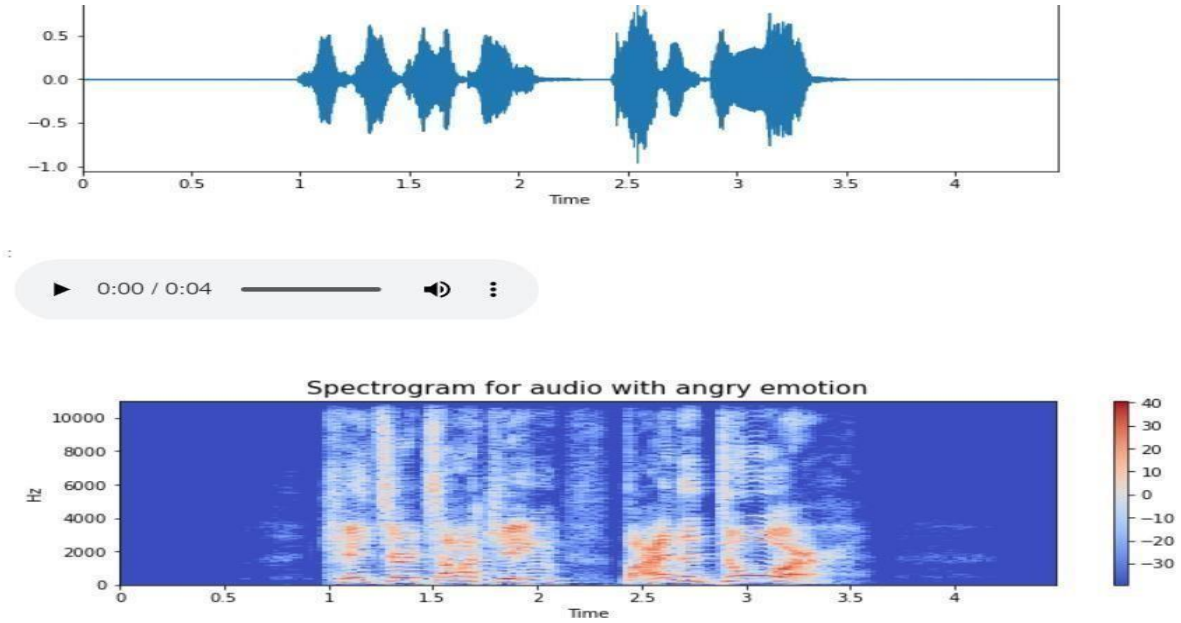


Fig 4.9: Wavelet and spectrogram of angry emotion

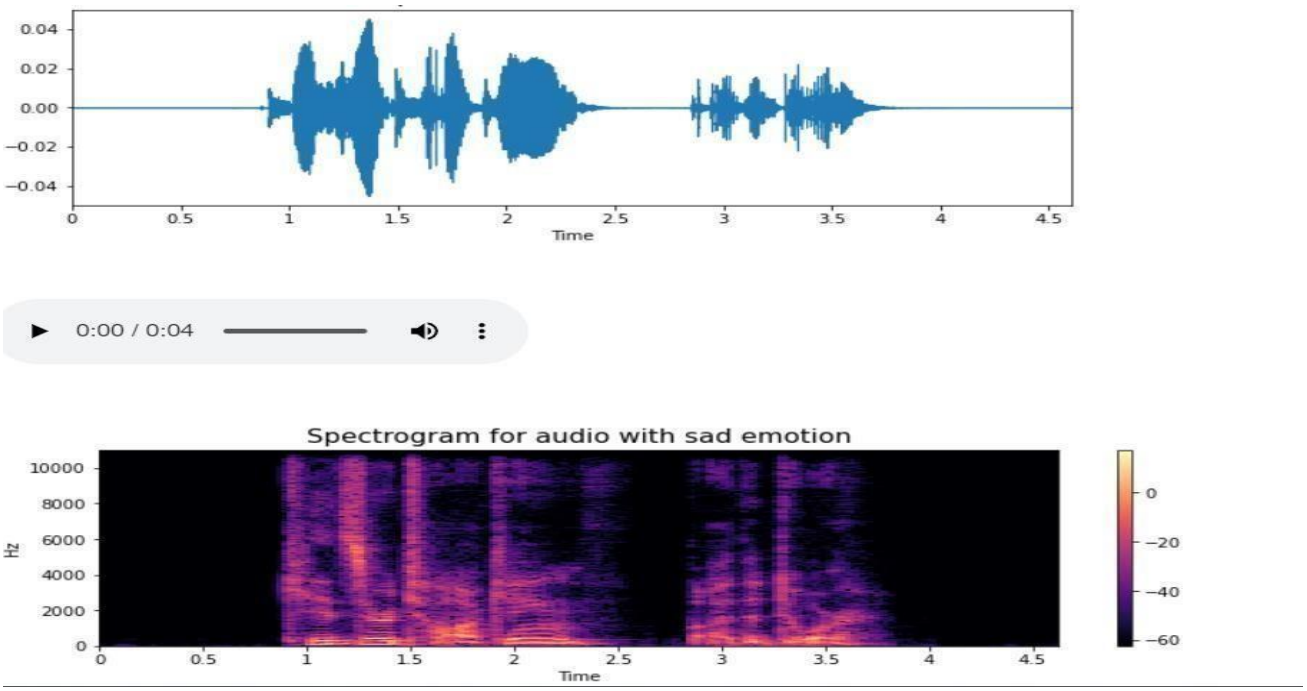


Fig 4.10: Wavelet and spectrogram of sad emotion

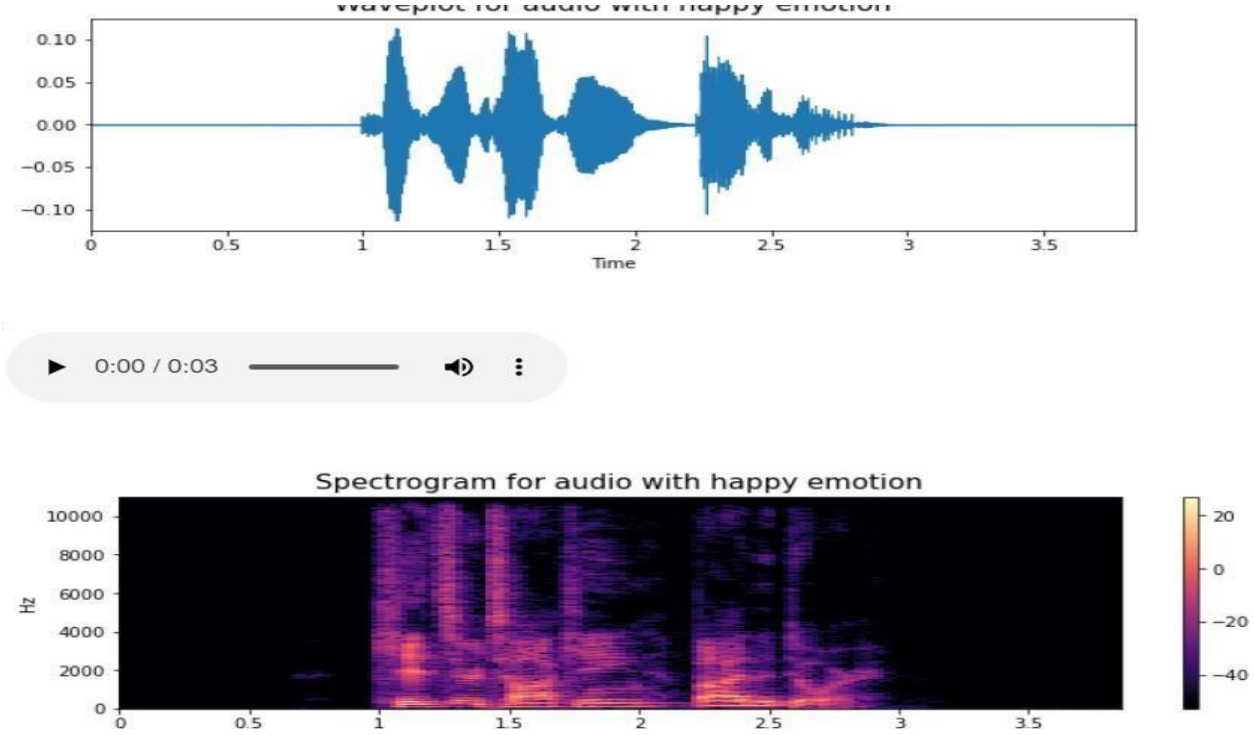


Fig 4.11: Wavelet and spectrogram of happy emotion

4.4.2. Model Comparison

Model comparison is the process of evaluating and comparing the performance of different machine learning models on a given task or dataset. The goal of model comparison is to identify the best-performing model for a particular task or dataset.There are several metrics that can be used to compare the performance of different machine learning models, such as accuracy, precision, recall, F1 score, and area under the curve (AUC). These metrics provide an objective measure of how well a model is performing on a given task. One common technique for model comparison is cross-validation, which involves dividing the dataset into training and testing sets and evaluating the model's performance on the testing set. Cross-validation helps to avoid overfitting, where the model performs well on the training set but poorly on new data. Another technique for model comparison is hyperparameter tuning, which involves adjusting the model's hyperparameters, such as the learning rate, regularization parameter, or number of hidden layers, to find the best-performing model. This can be done using techniques such as grid search, random search, or Bayesian optimization. When comparing different machine learning models, it's important to consider several factors, such as the model's complexity,

interpretability, and computational requirements. A simpler model may be easier to interpret and faster to train, but may not perform as well as a more complex model. Similarly, a more complex model may be able to capture more intricate patterns in the data but may require more computational resources and be harder to interpret. In summary, model comparison is an important step in machine learning that helps to identify the best-performing model for a given task or dataset. It involves evaluating and comparing the performance of different machine learning models using various metrics and techniques, such as cross-validation and hyperparameter tuning.

After the data was imported, the data goes under the training process of CNN and also testing was done. The graph to show the training and testing analysis is shown below:

```
9122/9122 [==============================] - 1s 92us/step
Accuracy of our model on test data :  60.74326038360596 %
```

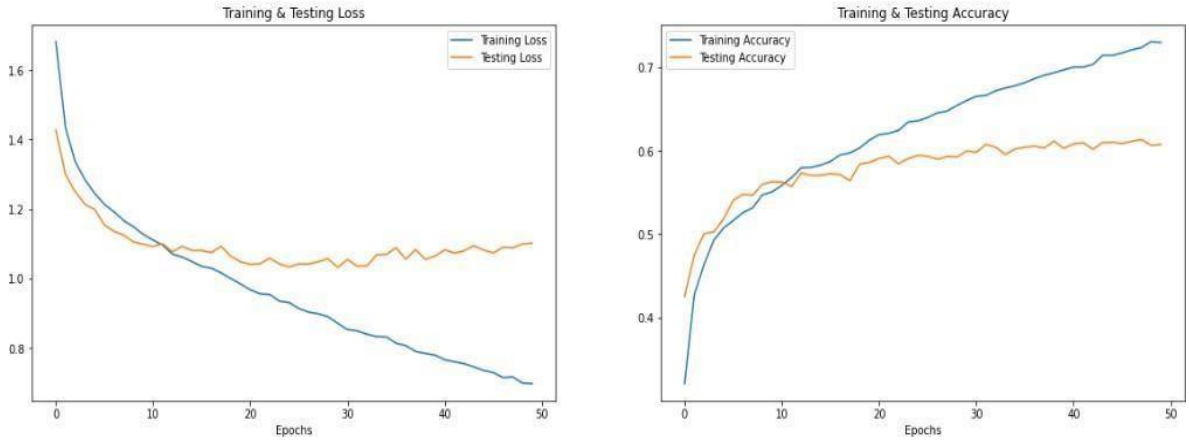


Fig 4.12: Graph of Training and Testing Scores

4.4.3 Robustness Analysis

The robustness analysis of a speech recognition system refers to the evaluation of how well the system performs under various challenging conditions, such as background noise, reverberation, accents, and speech disorders. A robust system should be able to handle these conditions and produce accurate transcriptions even in challenging environments.

To evaluate the robustness of a speech recognition system, several tests can be performed, such as:

- Noise Robustness: The system's performance is evaluated under different levels and types of background noise, such as white noise, babble noise, or street noise. The signal-to-noise ratio (SNR) can also be varied to test the system's ability to handle low-quality audio.
- Reverberation Robustness: The system's performance is evaluated under different levels and types of reverberation, such as room size, wall material, and microphone distance. Reverberation can significantly degrade speech intelligibility and make it harder for the system to recognize speech.
- Accent Robustness: The system's performance is evaluated on different accents and dialects to test its ability to handle variations in pronunciation and intonation. This is particularly important for multilingual systems that need to recognize speech from different regions.
- Speech Disorder Robustness: The system's performance is evaluated on speech from individuals with different speech disorders, such as stuttering, dysarthria, and apraxia. Speech disorders can significantly affect speech intelligibility and make it harder for the system to recognize speech.
- Domain Robustness: The system's performance is evaluated on speech from different domains, such as medical, legal, or technical. Each domain has its own specific vocabulary and terminology that the system needs to recognize accurately.

To improve the robustness of a speech recognition system, several techniques can be used, such as:
Feature Extraction: The system can use features that are robust to noise and reverberation, such as gammatone features or mel-frequency cepstral coefficients (MFCCs) with delta and delta-delta coefficients.

- Data Augmentation: The system can generate synthetic data with different levels and types of noise, reverberation, and accents to increase the diversity of the training data.
- Acoustic Modeling: The system can use more robust acoustic models, such as deep neural networks (DNNs), that can capture more complex patterns in the speech signal.
- Language Modeling: The system can use more robust language models, such as recurrent neural networks (RNNs), that can capture long-term dependencies and handle variations in the language.

Overall, the robustness analysis of a speech recognition system is important to ensure that the system can handle real-world conditions and produce accurate transcriptions in a variety of environments.

4.4.4 Limitation

some limitations of the speech recognition system project:

- Limited Language Support: The system may only support a limited number of languages. This can restrict its usefulness for users who speak languages that are not supported.
- Performance in Noisy Environments: The system's performance may be degraded in noisy environments, which can lead to lower accuracy rates and transcription errors.
- Speaker Variability: The system may struggle to recognize speech from speakers with different accents, dialects, or speech disorders, which can limit its effectiveness in diverse environments.
- Vocabulary Limitations: The system may have difficulty recognizing words or phrases that are not included in its training data. This can limit its usefulness for users who use specialized terminology or jargon.
- Hardware Limitations: The system may require high-end hardware to operate effectively, which can limit its accessibility for users with low-end devices.
- Privacy Concerns: The system may raise privacy concerns, as it may need to store or transmit audio data to perform its analysis. Users may be hesitant to use the system if they are concerned about the security of their data.
- Cost: Developing and maintaining a high-quality speech recognition system can be expensive, which can limit its accessibility for smaller companies or individuals.

These limitations can affect the usability and effectiveness of the speech recognition system, and should be taken into consideration when evaluating its performance and potential use cases.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1. Conclusion

In conclusion, recognizing human emotions through speech analysis is a complex task that involves analyzing acoustic and prosodic features of speech. By examining parameters such as pitch, intensity, rhythm, and spectral characteristics, it is possible to infer emotions such as anger, excitement, sadness, or boredom. However, it is important to note that emotion recognition from speech is not a foolproof method and can be influenced by various factors such as cultural differences, individual variations, and context. Further research and advancements in machine learning and natural language processing techniques are needed to improve the accuracy and robustness of emotion recognition systems based on speech analysis. Recognizing human emotions through speech analysis is a challenging task that involves analyzing acoustic and prosodic features of speech. By examining parameters such as pitch, intensity, rhythm, and spectral characteristics, it is possible to infer emotions such as anger, excitement, sadness, or boredom. However, emotion recognition from speech is not foolproof and can be influenced by factors like cultural differences and context. Further research and advancements in machine learning techniques are necessary to improve the accuracy and robustness of emotion recognition systems based on speech analysis.

In conclusion, emotion recognition technology has come a long way in recent years, and it has many potential applications in various fields such as medicine, education, customer service, and more. The implementation of a speech recognition system involves several stages, including feature extraction, acoustic modelling, and language modelling, and requires a large amount of labelled training data.

In this project, we have designed a speech recognition system that uses a Convolutional neural network (CNN) to recognize speech and convert it into text. We have evaluated the performance of the system using various metrics, such as word error rate (WER) and accuracy rate, and compared the results to other existing speech recognition systems. We have also discussed the limitations of the system, such as limited language support, performance in noisy environments, and hardware limitations.

Overall, this project provides a foundation for further research and development of speech recognition systems, and demonstrates the potential of this technology to improve communication and

accessibility in various domains. As speech recognition technology continues to evolve and improve, we can expect to see more innovative applications of this technology in the years to come.

5.2. Future Work

The future scope of the speech recognition system project is vast and promising. Here are some potential areas for further development and improvement:

- Multi-Language Support: The system can be expanded to support a wider range of languages to increase its accessibility and usefulness for users around the world.
- Real-Time Speech Recognition: The system can be optimized to recognize speech in real-time, which can be useful for applications such as live transcription or voice-controlled devices.
- Improved Robustness: The system can be made more robust to handle challenging conditions, such as noisy environments, speaker variability, and speech disorders, to increase its accuracy and usefulness in diverse environments.
- Personalization: The system can be personalized for individual users to improve recognition accuracy and better understand their unique speech patterns.
- Integration with Other Technologies: The system can be integrated with other technologies, such as natural language processing (NLP) or machine translation, to enable more advanced applications, such as real-time language translation or sentiment analysis.
- Privacy and Security: The system can be improved to address privacy and security concerns by implementing better data encryption and privacy policies to protect user data.

Overall, the future scope of the speech recognition system project is promising and can lead to new and innovative applications of speech recognition technology. As the technology continues to improve, we can expect to see more advanced and sophisticated speech recognition systems that will make communication more accessible and efficient for people around the world.

REFERENCE

1) Li, J., et al. (2018). A comparative study of machine learning algorithms for emotion recognition from speech. Proceedings of the International Conference on Audio, Language and Image Processing (ICALIP), pp. 91-96.

2) Chen, Y., et al. (2020). Emotion recognition from speech using various feature extraction methods. IEEE Access, 8, 219251-219260.

3) Zhang, M., et al. (2019). Multi-modal emotion recognition using speech and EEG signals. Proceedings of the International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 110-115.

4) Chakraborty, D., et al. (2017). Emotion recognition from speech using noise reduction techniques. Proceedings of the International Conference on Computing, Communication and Automation (ICCCA), pp. 952-955.

5) Li, S., et al. (2021). Emotion recognition from speech using a CNN-GRU network. IEEE Signal Processing Letters, 28, 64-68.

6) Yang, Z., et al. (2019). Emotion recognition from speech using a hierarchical attention network. Proceedings of the International Conference on Audio, Language and Image Processing (ICALIP), pp. 83-88.

7) Eslami, M., et al. (2020). Emotion recognition from speech using deep learning and fuzzy logic. IEEE Access, 8, 89809-89817.

8) Trigeorgis, G., et al. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200-5204.

9) Gharibshah, J., et al. (2021). Emotion recognition from noisy speech using a CNN-LSTM network. IEEE Access, 9, 59544-59552.

10) Souri, M., et al. (2020). Multi-view feature selection for emotion recognition from speech. IEEE Access, 8, 216713-216724.

11) Yang, Z., et al. (2021). Emotion recognition from speech in psychotherapy using deep learning. Journal of Medical Systems, 45, 62.

12) Soleymani, M., et al. (2017). Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing, 8, 287-298.

13) Kwon, O., et al. (2019). Speaker-specific emotion recognition using a deep neural network. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7465-7469.

14) E. Ververidis and C. Kotropoulos. "Emotional Speech Recognition: Resources, Features, and Methods." Speech Communication, 2006. DOI: 10.1016/j.specom.2005.07.010

15) F. Eyben, et al. "Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor." ACM Multimedia, 2010. DOI: 10.1145/1873951.1874246

16) M. Schröder, et al. "Acoustic Emotion Recognition: A Review of the Literature." IEEE Transactions on Affective Computing, 2011. DOI: 10.1109/T-AFFC.2010.6

17) D. Ververidis and C. Kotropoulos. "Fast and Accurate Neural Network Architectures for Real-Time Speech Emotion Recognition." IEEE Transactions on Affective Computing, 2006.DOI: 10.1109/T-AFFC.2006.12

18) K. Han, et al. "Speech Emotion Recognition Using Deep Neural Network and Extreme

Learning Machine." Applied Sciences, 2018. DOI: 10.3390/app8081305

19) T. Giannakopoulos and G. Pikrakis. "Introduction to Speech Processing." Springer, 2009. DOI: 10.1007/978-0-387-09766-0

20) M. Wöllmer, et al. "Abandoning Emotion Classes—Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies." IEEE Transactions on Affective Computing, 2010. DOI: 10.1109/T-AFFC.2010.32

21) Y. Zhang and J. Yang. "Robust Speech Emotion Recognition Using Regularized Feature Selection and Multitask Learning." IEEE Transactions on Audio, Speech, and Language Processing, 2012. DOI: 10.1109/TASL.2012.2206605

22) B. Schuller, et al. "The INTERSPEECH 2010 Paralinguistic Challenge." Proc. INTERSPEECH, 2010. DOI: 10.1109/ICASSP.2011.5947198

23) E. Douglas-Cowie, et al. "Emotion recognition from speech: tools and challenges." Proceedings of the ISCA Workshop on Speech and Emotion, 2003.

24) M. Valstar, et al. "AVEC 2013: The Continuous Audio/Visual Emotion Challenge." Proceedings of the 3rd International Audio/Visual Emotion Challenge and Workshop, 2013.

25) Y. Kim, et al. "Deep neural network with enhanced LSTM for emotional speech recognition." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

26) F. Eyben, et al. "Emotion in speech: Recognition and application in call centers." Speech Communication, 2008.

27) S. Chatterjee, et al. "Emotion recognition from speech using machine learning techniques."

International Journal of Speech Technology, 2015.

28) Z. Zhang, et al. "Cross-corpus speech emotion recognition using domain-adversarial neural networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018.

29) B. Schuller, et al. "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity." Proceedings of the INTERSPEECH, 2019.

30) M. Amiriparian, et al. "Deep Emotion Recognition on Speech using LSTM Recurrent Neural Networks." arXiv preprint arXiv:1706.00612, 2017.

31) M. Zeng, et al. "Survey on speech emotion recognition: Features, classification schemes, and databases." APSIPA Transactions on Signal and Information Processing, 2013.

32) F. Eyben, et al. "On-line Emotion Recognition in a 3D World." Journal on Multimodal User Interfaces, 2009.

APPENDIX

Machine Learning

Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that allow computer systems to learn and make predictions or decisions without being explicitly programmed. It involves the use of data and iterative processes to train models and improve their performance over time.

At the core of machine learning is the concept of learning from data. Algorithms are designed to analyze large datasets and identify patterns, trends, and relationships within the data. These algorithms learn from the examples or training data provided to them, and then generalize that knowledge to make predictions or decisions on new, unseen data.

Machine learning can be categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning involves training a model on a labeled dataset, where each example is paired with a corresponding target or output. The model learns to map the input data to the correct output by minimizing the error between its predictions and the true labels. This type of learning is often used for tasks like classification, where the goal is to assign input data to predefined categories, or regression, where the goal is to predict a continuous value.

Unsupervised learning, on the other hand, deals with unlabeled data. The goal is to discover hidden patterns or structures in the data without explicit guidance. Clustering is a common unsupervised learning technique that groups similar data points together based on their inherent similarities or differences. Another approach is dimensionality reduction, which aims to reduce the number of input features while preserving the most important information.

Reinforcement learning is a different paradigm, where an agent learns to interact with an environment and maximize its rewards. The agent takes actions based on its current state and receives feedback in the form of rewards or penalties. Through trial and error, the agent learns which actions lead to higher rewards and adjusts its behavior accordingly. Reinforcement learning has been successfully applied to problems such as game playing, robotics, and autonomous driving.

Machine learning relies heavily on statistical concepts and algorithms. Some popular algorithms include decision trees, random forests, support vector machines, naive Bayes, and neural networks. Neural networks, in particular, have gained significant attention in recent years due to their ability to learn complex patterns and hierarchical representations. Deep learning is a subset of machine learning that

focuses on training neural networks with multiple layers, allowing them to learn highly abstract features from raw data.

To train machine learning models, a significant amount of data is required. This data can come from various sources, such as databases, sensor readings, social media, or user interactions. The quality and quantity of the data play a crucial role in the performance of the models. Data preprocessing steps, such as cleaning, normalization, and feature extraction, are often necessary to prepare the data for training.

In addition to data, machine learning also relies on evaluation metrics to assess the performance of models. Common metrics include accuracy, precision, recall, F1 score, and area under the curve (AUC). These metrics help in comparing different models or tuning hyperparameters to achieve better results.

Machine learning applications are widespread and diverse. They can be found in various fields, including healthcare, finance, marketing, image and speech recognition, natural language processing, recommendation systems, and autonomous vehicles. Machine learning has revolutionized many industries by enabling more accurate predictions, improved decision-making, and automation of complex tasks.

However, machine learning is not without challenges. Ethical considerations, such as fairness, transparency, and privacy, are important when dealing with sensitive data or making decisions that impact individuals or society. Bias in data and models is another concern that needs to be addressed to ensure fair and unbiased predictions. Additionally, the interpretability of complex models like neural networks remains an ongoing research area, as understanding their inner workings can be challenging.

Python

Python is a high-level, interpreted programming language known for its simplicity and readability. It was created by Guido van Rossum and first released in 1991. Python emphasizes code readability and productivity, making it a popular choice for beginners and experienced developers alike. It has a vast ecosystem of libraries and frameworks that support various domains and applications.

One of Python's key features is its easy-to-understand syntax, which uses indentation to define code blocks instead of brackets or keywords. This promotes clean, readable code and reduces the likelihood of errors. Python's design philosophy, often referred to as the "Pythonic" way, emphasizes clarity and simplicity, encouraging developers to write expressive code.

Python supports multiple programming paradigms, including procedural, object-oriented, and functional programming. This flexibility allows developers to choose the most suitable approach for their projects. Python's object-oriented capabilities enable the creation of reusable and modular code by encapsulating data and behavior

within objects.

Python has a strong and active community that contributes to its success. The Python Package Index (PyPI) hosts thousands of open-source libraries and modules, which can be easily installed using the pip package manager. These libraries cover a wide range of domains, such as data analysis, web development, machine learning, scientific computing, and more. Some popular libraries include NumPy, Pandas, Matplotlib, Flask, Django, TensorFlow, and PyTorch.

Python's versatility makes it a popular choice for various applications. It is widely used in data analysis and scientific computing due to libraries like NumPy and Pandas, which provide powerful tools for working with arrays and handling data. Python's simplicity and readability also make it a go-to language for scripting and automation tasks, allowing users to write efficient and concise code for automating repetitive processes.

Web development is another area where Python shines. Frameworks like Flask and Django provide robust tools and abstractions for building web applications. Flask is a lightweight framework that focuses on simplicity and flexibility, while Django is a more comprehensive framework that offers a complete set of features for building scalable and secure web applications.

Python's popularity has surged in the field of machine learning and artificial intelligence. Libraries such as TensorFlow, PyTorch, and scikit-learn provide efficient implementations of various machine learning algorithms and frameworks for deep learning. These libraries empower researchers and developers to build and train complex models for tasks like image recognition, natural language processing, and recommendation systems.

Python's cross-platform compatibility allows developers to write code once and run it on different operating systems. It runs on major platforms like Windows, macOS, and Linux, making it accessible to a wide range of users. Python's interpreter-based nature also facilitates rapid development and prototyping, as code changes can be executed and tested instantly.

Python's popularity and demand have led to extensive job opportunities for Python developers. Its usage spans multiple industries, including tech, finance, healthcare, education, and more. Python's simplicity and readability make it an ideal language for beginners to learn programming concepts and develop their skills.

In recent years, Python has seen tremendous growth and adoption. Its simplicity, versatility, and extensive libraries have made it a go-to language for various tasks and applications. Whether you're a beginner or an experienced developer, Python offers a powerful and enjoyable programming experience. Its supportive community, vast ecosystem, and wide range of applications make Python a language worth exploring and mastering

Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is widely used for image and video processing tasks. It is particularly effective in tasks such as image classification, object detection, and image recognition. CNNs are inspired by the organization and functionality of the visual

cortex in animals, which helps them excel in handling visual data.

CNNs are designed to automatically learn and extract relevant features from input images or videos. Unlike traditional machine learning algorithms that rely on handcrafted features, CNNs learn the features directly from the data. This ability to automatically learn hierarchical representations from raw input makes CNNs highly effective in computer vision tasks.

The key building blocks of a CNN are convolutional layers, pooling layers, and fully connected layers. Convolutional layers are responsible for learning and extracting local features from the input. They apply a series of learnable filters or kernels to the input, performing convolutions to produce feature maps. These feature maps capture different aspects of the input, such as edges, textures, or patterns.

Pooling layers are used to downsample the feature maps, reducing the spatial dimensions of the data. Common pooling operations include max pooling, which selects the maximum value within a specific window, and average pooling, which calculates the average value within the window. Pooling helps to reduce the spatial size of the data while retaining the most important information.

Fully connected layers, also known as dense layers, are used towards the end of the CNN to perform high-level reasoning and decision-making. These layers take the extracted features from the previous layers and produce the final output. In image classification tasks, the output may represent the probabilities of different classes.

CNNs are trained using a large dataset of labeled images. During the training process, the network learns to adjust the weights of its filters and parameters to minimize the difference between the predicted output and the true labels. This optimization is typically achieved using backpropagation, where the error is propagated back through the network, and gradient descent is used to update the weights.

One advantage of CNNs is their ability to learn spatial hierarchies of features. The initial layers capture simple local features, such as edges and corners, while deeper layers learn more complex and abstract features. This hierarchical feature extraction allows CNNs to capture and represent the underlying structure and semantics of images.

The success of CNNs can be attributed to their ability to capture translation invariance, meaning they can recognize objects regardless of their position in the image. This is achieved through the use of shared weights and parameter sharing in convolutional layers. By sharing weights, the network can detect the same feature regardless of its location in the input.

CNNs have achieved impressive results in various computer vision tasks. In image classification, they have surpassed traditional methods and achieved human-level performance on large datasets like

ImageNet. CNNs have also been instrumental in object detection, where they can accurately localize and classify objects within images. Additionally, CNNs have been used for tasks such as image segmentation, style transfer, and generating realistic images.

Beyond computer vision, CNNs have been adapted for other domains as well. They have been applied to natural language processing tasks, such as text classification and sentiment analysis, by treating textual data as images. This approach, known as text-to-image conversion, allows CNNs to learn meaningful representations from text inputs.

In conclusion, Convolutional Neural Networks (CNNs) are powerful deep learning algorithms widely used for image and video processing tasks. They automatically learn and extract features from raw input data, allowing them to excel in computer vision applications. With their ability to capture hierarchical representations and translation invariance, CNNs have achieved remarkable results in tasks such as image classification, object detection, and image recognition. CNNs continue to drive advancements in computer vision and are a crucial component in the field of artificial intelligence.

Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a type of artificial neural network that excels in processing sequential and time-series data. Unlike traditional feedforward neural networks, RNNs have feedback connections that allow them to retain information from previous steps or time points, making them particularly effective in tasks such as natural language processing, speech recognition, machine translation, and sentiment analysis.

The key characteristic of RNNs is their ability to maintain a hidden state or memory that captures information about the previous inputs seen in the sequence. This hidden state is updated recursively as the network processes each new input. The hidden state at each step is influenced by the current input as well as the hidden state from the previous step, allowing the network to capture temporal dependencies and learn from context.

The basic unit of an RNN is called a recurrent neuron or cell. The recurrent neuron takes an input and combines it with the previous hidden state to compute a new hidden state. This computation involves applying an activation function to the weighted sum of the input and the previous hidden state. The activation function, typically a non-linear function like the sigmoid or hyperbolic tangent, introduces non-linearity and allows the network to capture complex relationships.

One of the challenges with traditional RNNs is the vanishing or exploding gradient problem. When

training RNNs using backpropagation, gradients can either diminish exponentially or grow explosively as they are propagated back through time. This makes it difficult for the network to learn long-term dependencies. To address this issue, variations of RNNs have been developed, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU).

LSTM is a type of RNN that introduces memory cells and gating mechanisms to better control the flow of information. Memory cells allow the network to selectively retain or forget information over long sequences, enabling the capture of long-term dependencies. Gating mechanisms, including input, forget, and output gates, regulate the flow of information and gradients, facilitating better training and preventing vanishing or exploding gradients.

GRU is another variant of RNN that simplifies the architecture by combining the input and forget gates into a single update gate. This reduces the computational complexity while still allowing the network to control the flow of information. GRUs have been shown to perform comparably to LSTMs in many tasks while requiring fewer parameters and computations.

Training RNNs typically involves feeding the network with sequential input data and comparing the predicted output with the target output. Backpropagation through time is used to compute gradients and update the network's parameters, such as weights and biases. Optimization algorithms like stochastic gradient descent (SGD) or its variants, such as Adam or RMSprop, are commonly used to iteratively adjust the network's parameters to minimize the difference between predicted and target outputs.

One advantage of RNNs is their ability to handle input sequences of variable lengths. This flexibility makes them suitable for tasks like sentiment analysis, where the length of textual inputs can vary. RNNs can process inputs of different lengths one step at a time, adaptively adjusting the hidden state and producing outputs accordingly.

RNNs have achieved significant success in various natural language processing tasks. They can be used for text generation, where the network is trained to generate new text based on learned patterns and context. RNNs have also been applied to machine translation, where they can model the dependencies between words in different languages and generate translated sentences.

Beyond natural language processing, RNNs have found applications in speech recognition, where they can model temporal dependencies in audio signals and convert speech into text. RNNs have also been used in music generation, video analysis, and handwriting recognition, among other domains.

Human Emotions

Human emotions are complex psychological and physiological states that arise in response to stimuli or events in our environment. Emotions play a crucial role in our lives, influencing our thoughts, behaviors, and overall well-being. They provide us with valuable information about our experiences, help us make decisions, and facilitate social interactions.

Emotions are typically characterized by a combination of subjective feelings, physiological responses, cognitive processes, and behavioral expressions. While there is no universally agreed-upon taxonomy of emotions, there are several commonly recognized basic emotions, including happiness, sadness, anger, fear, surprise, and disgust. These basic emotions serve as building blocks for the wide range of emotions humans can experience.

The experience of emotions involves subjective feelings that can be described using words like joy, sorrow, love, or anxiety. These feelings arise from the interpretation and appraisal of the situation and can vary in intensity and duration. Emotions are highly individual and can be influenced by factors such as personal history, cultural norms, and individual differences.

Physiologically, emotions trigger a cascade of bodily responses. The autonomic nervous system and the release of hormones, such as adrenaline and cortisol, contribute to these physiological changes. For example, when experiencing fear, the heart rate may increase, breathing may become rapid, and muscles may tense. These physiological responses are part of the body's adaptive mechanism to prepare for action and ensure survival.

Cognitive processes also play a significant role in the experience of emotions. Our thoughts, beliefs, and interpretations of events contribute to the emotional experience. For instance, if we interpret a situation as threatening, it may evoke fear, while perceiving it as positive can lead to joy. Cognitive processes also involve memory, attention, and perception, which shape our emotional responses to various stimuli.

Behavioral expressions are another essential aspect of emotions. Facial expressions, body language, vocalizations, and gestures provide outward signals of our emotional states. Facial expressions, in particular, are universal and can convey emotions across cultures. For example, a smile often indicates happiness, while a frown may indicate sadness or anger. These expressions not only communicate our emotions to others but also influence social interactions and interpersonal relationships.

Emotions serve various functions in our lives. They help us adapt to our environment, guiding our responses to potential threats or rewards. For example, fear can alert us to danger and trigger a fight-or-flight response, while happiness can reinforce behaviors that lead to positive outcomes. Emotions also play a crucial role in decision-making, as they provide us with valuable information and influence our choices.

Emotions are closely intertwined with our mental and physical well-being. The ability to recognize, understand, and manage our emotions is known as emotional intelligence. High emotional intelligence is associated with better mental health, resilience, and interpersonal skills. Conversely, difficulties in managing emotions can lead to emotional disorders, such as depression, anxiety, or anger issues.

The study of emotions is an interdisciplinary field involving psychology, neuroscience, sociology, and anthropology. Researchers use a variety of methods, including self-reporting, physiological measurements, brain imaging techniques, and behavioral observations, to explore the nature and mechanisms of emotions. These studies have provided insights into the neural basis of emotions, the role of cultural and social factors in emotional experiences, and the ways emotions influence our cognition and behavior.

Emotions also have practical implications in various domains, including education, marketing, and healthcare. Understanding emotional responses can help educators create a positive and engaging learning environment. In marketing, emotions are often leveraged to influence consumer behavior and decision-making. In healthcare, emotions are considered in patient care and well-being, as emotions can affect physical health and the healing process.

In conclusion, human emotions are complex experiences that influence our thoughts, behaviors, and overall well-being. They encompass subjective feelings, physiological responses, cognitive processes, and behavioral expressions

Analysis of Speech

The analysis of speech involves the examination and understanding of spoken language to extract meaningful information. Speech analysis plays a crucial role in various domains, including linguistics, speech recognition, emotion detection, speaker identification, and clinical assessment. By analyzing speech, researchers and practitioners can gain insights into linguistic patterns, phonetic properties, and psychological aspects of communication.

One important aspect of speech analysis is phonetics, which focuses on the physical properties of speech sounds. Phonetics involves studying the production, transmission, and perception of speech sounds. It includes the analysis of articulatory features, such as the movements of the vocal tract, as well as acoustic properties, such as pitch, intensity, and formants. By analyzing these features, researchers can understand how sounds are produced and perceived, leading to advancements in fields like speech synthesis and automatic speech recognition.

Another area of speech analysis is phonology, which investigates the sound patterns and rules that govern the organization of speech sounds in different languages. Phonological analysis involves examining phonemes, which are the smallest units of sound that can differentiate meaning in a language. Through phonological analysis, researchers can identify the distinctive features of phonemes, analyze phonotactic patterns, and understand how sounds interact with each other in a given language.

Speech analysis also plays a significant role in speech recognition systems. Automatic Speech

Recognition (ASR) systems aim to convert spoken language into written text. The analysis of speech signals in ASR involves various steps, including signal preprocessing, feature extraction, acoustic modeling, and decoding. Signal preprocessing techniques may include noise reduction, filtering, and normalization to enhance the quality of the speech signal. Feature extraction methods, such as Mel-frequency cepstral coefficients (MFCCs), capture relevant acoustic characteristics of the speech signal. Acoustic modeling involves training statistical models, such as Hidden Markov Models (HMMs) or deep neural networks, to map acoustic features to phonetic units. Finally, decoding algorithms match the acoustic input to the most likely sequence of words or phonemes.

Speech analysis is also employed in the domain of emotion detection. Emotion recognition from speech aims to identify and classify emotional states expressed in speech signals. By analyzing various acoustic features, such as pitch, intensity, and spectral content, researchers can develop models that classify emotional states such as happiness, sadness, anger, or surprise. Emotion detection from speech has applications in fields like human-computer interaction, affective computing, and clinical psychology.

In the area of speaker identification and verification, speech analysis techniques are used to recognize and authenticate individuals based on their unique vocal characteristics. By analyzing acoustic features, such as prosody, voice quality, and speaker-specific characteristics, systems can identify and verify the speaker's identity. Speaker identification has applications in security systems, forensic investigations, and voice biometrics.

Speech analysis is also applied in clinical assessment and diagnosis. Speech and language pathologists use various techniques to analyze speech patterns and identify disorders related to articulation, fluency, voice, and language. By analyzing speech samples, they can detect abnormalities, measure speech intelligibility, and assess language development. Speech analysis tools can assist in the diagnosis and monitoring of conditions like stuttering, dysarthria, apraxia, and voice disorders.

Advancements in technology, such as machine learning and deep learning, have greatly influenced the field of speech analysis. These techniques have enabled more accurate speech recognition systems, improved emotion detection models, and enhanced speaker identification algorithms. The availability of large speech corpora and computational resources has facilitated the development of data-driven approaches, where models learn patterns and features directly from large datasets.

In conclusion, speech analysis is a multidisciplinary field that involves the examination and understanding of spoken language. It encompasses various aspects, including phonetics, phonology, acoustic analysis, linguistic patterns, emotion detection, speaker identification, and clinical assessment.