IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Emotion Speech Recognition – Bridging Human and AI Communication

[1]Mrs.P.Divya, [2]Ms.G.Harini, [3]Ms.T.S.Suruthi

[1]Assistant Professor, [2]Student , [3]Student
[1]Artificial Intelligence and Data Science,
[1]IFET College of Engineering, Villupuram , India

Abstract : Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Deep Learning techniques have been recently proposed as an alternative to traditional techniques in SER. Speech emotion recognition is the task of automatically detecting the emotional state of a speaker from their spoken words. It is a growing area of research that has applications in various fields such as human computer interaction, education, and psychology. There are several approaches to speech emotion recognition, including the use of machine learning algorithms, which can be trained on large datasets of annotated speech samples to recognize patterns associated with different emotions. Other approaches include the use of linguistic features, prosodic features, and physiological signals such as facial expressions and heart rate. One challenge in speech emotion recognition is the variability in the expression of emotions across individuals and cultural groups. Another challenge is the need to accurately identify the underlying emotion, as opposed to simply recognizing the presence or absence of an emotion. Overall, speech emotion recognition has the potential to improve communication between humans and machines, and to provide insights into the emotional states of individuals.

Keywords : Human-Computer Interaction (HCI), Deep Learning, linguistic features, emotional understanding

CHPATER 1
1.INTRODUCTION

Emotion recognition from speech has evolved from being a niche to an important component for Human-Computer Interaction (HCI). These systems aim to facilitate the natural interaction with machines by direct voice interaction instead of using traditional devices as input to understand verbal content and make it easy for human listeners to react. Some applications include dialogue systems for spoken languages such as call center conversations, onboard vehicle driving system and utilization of emotion patterns from the speech in medical applications. Nonetheless, there are many problems in HCI systems that still need to be properly addressed, particularly as these systems move from lab testing to real-world application. Hence, efforts are required to effectively solve such problems and achieve better emotion recognition by machines.

Determining the emotional state of humans is an idiosyncratic task and may be used as a standard for any emotion recognition model.

The approach for speech emotion recognition (SER) primarily comprises two phases known as feature extraction and features classification phase. In the field of speech processing, researchers have derived several features such as source-based excitation features, prosodic features, vocal traction factors, and other hybrid features. The second phase includes feature classification using linear and non-linear classifiers.

The most commonly used linear classifiers for emotion recognition include Bayesian Networks (BN) or the Maximum Likelihood Principle (MLP) and Support Vector Machine (SVM). Usually, the speech signal is considered to be non-stationary. Hence, it is considered that non-linear classifiers work effectively for SER. There are many non-linear classifiers available for SER, including Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). These are widely used for classification of information that is derived from basic level features. Energy-based features such as Linear Predictor Coefficients (LPC), Mel Energy-spectrum Dynamic Coefficients (MEDC), Mel-Frequency Cepstrum Coefficients (MFCC) and Perceptual Linear Prediction cepstrum coefficients (PLP) are often used for effective emotion recognition from speech. Other classifiers including K-Nearest Neighbor (KNN), Principal Component Analysis (PCA) and Decision trees are also applied for emotion recognition.

This research work considers the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) and TESS dataset (Toronto Emotional Speech Set).Here, the three key features such as MFCC (Mel Frequency Cepstral Coefficients), Mel Spectrogram and chroma are extracted.

1.2 Goal : The goal of speech emotion recognition (SER) is to develop algorithms and systems that can automatically detect and classify the emotional states conveyed by speakers through their speech signals.

1.3Domain Overview

Emotion speech recognition involves the technology that enables machines to interpret and respond to human emotions expressed through speech. This domain aims to enhance communication between humans and AI by enabling systems to understand not just the words spoken but also the emotional context. Emotion speech recognition is a fascinating field at the intersection of artificial intelligence and human-computer interaction. In this domain, the goal is to develop technologies that allow machines to not only comprehend the literal content of spoken words but also discern the emotional nuances embedded within human speech. It has applications in customer service, mental health support, and human-computer interaction, fostering more natural and empathetic exchanges between humans and AI. This technology holds significant promise in various applications, ranging from customer service interactions to mental health support.

Deep Learning has been considered as an emerging research field in machine learning and has gained more attention in recent years. Deep Learning techniques for SER have several advantages over traditional methods, including their capability to detect the complex structure and features without the need for manual feature extraction and tuning; tendency toward extraction of low-level features from the given raw data, and ability to deal with un-labeled data.

Deep Neural Networks (DNNs) are based on feed-forward structures comprised of one or more underlying hidden layers between inputs and outputs. The feed-forward architectures such as Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) provides efficient results for image and video processing. On the other hand, recurrent architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) are much effective in speech-based classification such as natural language processing (NLP) and SER. Apart from their effective way of classification these models do have some limitations. For instance, the positive aspect of CNNs is to learn features from high-dimensional input data, but on the other hand, it also learns features from small variations and distortion occurrence and hence, requires large storage capability. Similarly, LSTM based RNN's are able to handle variable input data and model long-range sequential text data.

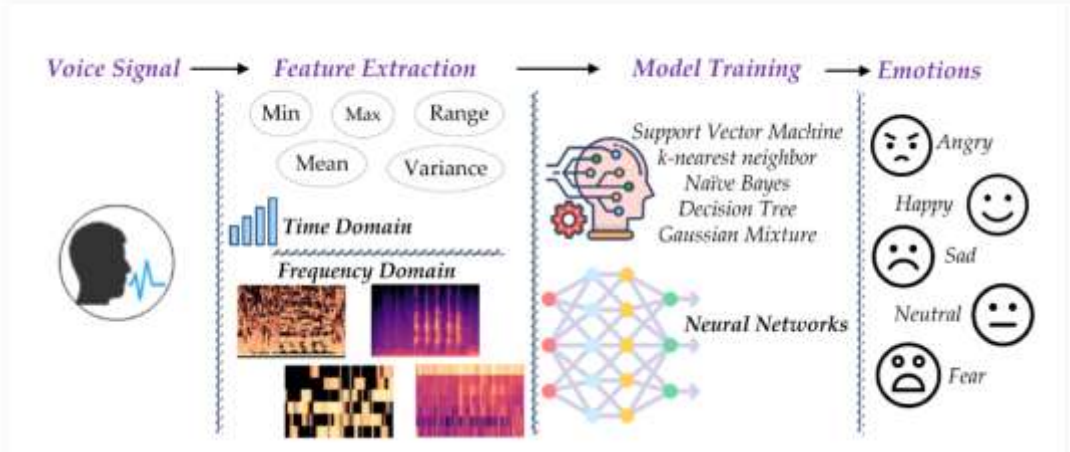

FIGURE 1: SPEECH EMOTION RECOGNITION

CHAPTER 2

2. LITERATURE REVIEW

Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Deep Learning techniques have been recently proposed as an alternative to traditional techniques in SER. This paper presents an overview of Deep Learning techniques and discusses some recent literature where these methods are utilized for speech-based emotion recognition. The review covers databases used, emotions extracted, contributions made toward speech emotion recognition and limitations related to it.

Raw audio data were standardized so every audio file has zero mean and unit variance. Every file was split into 20 millisecond segments without overlap. We used Voice Activity Detection (VAD) algorithm to eliminate silent segments and divided all data into TRAIN (80%) VALIDATION (10%) and TESTING (10%) sets. DNN is optimized using Stochastic Gradient Descent. As input we used raw data without and feature selection. Our trained model achieved overall test accuracy of 96.97% on whole-file classification.One challenge in speech emotion recognition is the variability in the expression of emotions across individuals and cultural groups. Another challenge is the need to accurately identify the underlying emotion, as opposed to simply recognizing the presence or absence of an emotion.

CHAPTER 3

3. PROJECT METHODOLOGY:

3.1 General

The existing work in this area reveals that most of the present work relies on lexical analysis for emotion recognition, that have been used for the purpose of classification of emotions into many categories, i.e.., Angry, Happy, Neutral, Fear, Disgust, Sad, Pleasant Surprise. The maximum cross-correlation between the discrete time sequences of the audio signals is computed and the highest degree of correlation between the testing audio file and the training audio file is used as an integral parameter for identification of a particular emotion type. The second technique is used with the feature extraction of discriminatory features with the Cubic SVM Classifier for recognition of Angry, Happy and Neutral emotion segments only.

3.2 Proposed System

Speech emotion recognition, the best example of it can be seen at call centers. If you ever noticed, call centers employees never talk in the same manner, their way of pitching/talking to the customers changes with customers. SER is tough because emotions are subjective and annotating audio is challenging. Human voice is given as the input. Then the input is converted into frames of frame size 60ms for every 50ms which means overlapping of data for 10ms. This is because for no missing of data. Fundamental frequencies are calculated based on pitch autocorrelation function. Nowadays, speech is only used for converting voice into text for searching purposes, but in our system, it can also be used to gather feedback and recognize emotions, among other applications.

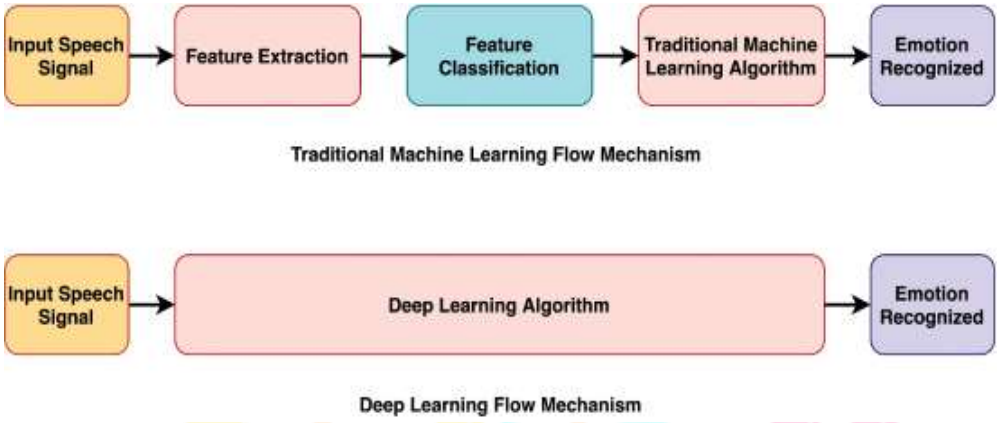

FIGURE 2: DEEP LEARNING FLOW MECHANISM

It can be used in an automatic remote call center, a car board system, the field of E-learning, and the emotions of students during the lecture.

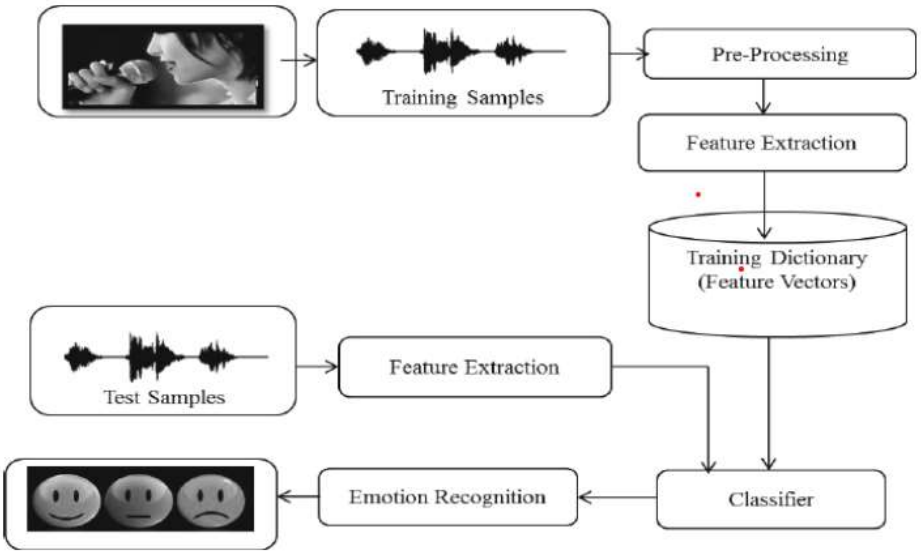

FIGURE 3: ARCHITECTURE

3.3 ADVANTAGES OF PROPOSED SYSTEM

Enhanced Human-Computer Interaction (HCI):

SER enables machines to understand and respond to human emotions conveyed through speech, leading to more intuitive and effective interactions between humans and computers.

Personalized User Experiences:

By recognizing users' emotional states, SER can tailor responses and recommendations to individual preferences and needs, leading to more personalized user experiences.

Improved Customer Service:

In customer service applications, SER can analyze customer sentiment in real-time, allowing businesses to address issues promptly and provide better support.

Healthcare Applications:

In healthcare, SER can be used to monitor patients' emotional well-being, detect signs of distress or depression, and provide early intervention when needed.

Educational Tools:

SER can enhance educational tools by providing feedback on students' emotional engagement and comprehension, helping educators tailor instruction to individual learning styles.

Market Research and Sentiment Analysis:
SER can be applied in market research to analyze consumer sentiment and preferences, helping businesses make informed decisions about product development and marketing strategies.

CHAPTER 4
4. MODULE DESCRIPTION

Flask: Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier. It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single Python file.
Python: Python is commonly used for developing websites and software, task automation, data analysis, and data visualisation. Since it's relatively easy to learn, Python has been adopted by many non-programmers, such as accountants and scientists, for a variety of everyday tasks, like organising finances.
HTML: HTML (Hyper Text Markup Language) is the code that is used to structure a web page and its content. For example, content could be structured within a set of paragraphs, a list of bulleted points, or using images and data tables.
Librosa: Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.
Pandas: Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way towards this goal.
Numpy: Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

CHAPTER 5
5. CONCLUSION

Emotion Speech Recognition – Bridging Human and AI Communication has successfully explored the intersection of human emotions and artificial intelligence. Through advanced speech recognition technologies, we've demonstrated the potential to enhance communication by enabling machines to discern and respond to human emotions. This not only marks a significant step forward in the field of AI but also opens avenues for more empathetic and context-aware human-machine interactions. The challenges faced and overcome during this project, from fine-tuning algorithms to handling diverse emotional expressions, have contributed to the evolution of a technology that can bridge the communication gap between humans and AI. Looking ahead, the ongoing refinement of our system and its integration into real-world applications hold promise for a future where AI is not just intelligent but emotionally intelligent. As we continue to refine and expand upon these capabilities, the impact on various sectors, including customer service, mental health, and beyond, holds promising prospects. The fusion of emotional intelligence with AI communication is a pivotal development, fostering a deeper connection between humans and machines.

5.1 Future Goal

Enhanced Accuracy and Robustness:
Continued research aims to improve the accuracy and robustness of SER systems, particularly in recognizing subtle nuances and variations in emotional expression across different contexts, languages, and speakers.
Multimodal Integration:
Future SER systems may integrate multiple modalities such as facial expressions, gestures, physiological signals, and contextual information to improve emotion recognition accuracy and provide a more holistic understanding of human emotions.
Cross-Cultural and Cross-Lingual Recognition: Future research aims to address the challenges of cross-cultural and cross-lingual emotion recognition, developing models that are sensitive to cultural norms, linguistic differences, and diverse emotional expression patterns.
Emotionally Intelligent Assistive Technologies:
SER can be integrated into assistive technologies to provide support for individuals with disabilities, elderly populations, and those in need of emotional assistance, fostering greater independence, well-being, and social connection.
Applications in Healthcare and Mental Health:
SER holds promise for applications in healthcare and mental health, including the early detection of emotional distress, monitoring of mood fluctuations, and personalized interventions for mental health disorders.

5.2 References

[1].R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Vocal Emotion Recognition Using Deep Learning Techniques: A Review", IEEE Access, vol. 7, no. 7, pp. 117327-117345, 2019
[2].P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2017, pp. 137-140, doi: 10.1109/SPIN.2017.8049931.
[3].K. S. Chintalapudi, I. A. K. Patan, H. V. Sontineni, V. S. K. Muvvala, S. V. Gangashetty and A. K. Dubey, "Speech Emotion Recognition Using Deep Learning," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-5, doi: 10.1109/ICCCI56745.2023.10128612.
[4].R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma and N. Mukesh, "Speech Emotion Recognition using Machine Learning," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1608-1612, doi: 10.1109/ICOEI51242.2021.9453028
[5].K. V. Krishna, N. Sainath and A. M. Posonia, "Speech Emotion Recognition using Machine Learning," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1014-1018, doi: 10.1109/ICCMC53470.2022.9753976.

[6].R. Y. Cherif, A. Moussaoui, N. Frahta and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430224.
[7].S. Ullah, Q. A. Sahib, Faizullah, S. Ullahh, I. U. Haq and I. Ullah, "Speech Emotion Recognition Using Deep Neural Networks," 2022 International Conference on IT and Industrial Technologies (ICIT), Chiniot, Pakistan, 2022, pp. 1-6, doi: 10.1109/ICIT56493.2022.9989197.
[8].A. Kumar, V. Kumar and P. Rajakumar, "Speech Emotion Recognition Using Machine Learning," 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-6, doi: 10.1109/ICIPTM57143.2023.10118251.
[9].S. Suganya and E. Y. A. Charles, "Speech Emotion Recognition Using Deep Learning on audio recordings," 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019, pp. 1-6, doi: 10.1109/ICTer48817.2019.9023737.