Speech Emotion Recognition Using Machine Learning

Apeksha G[1], Sreedevi S[2], Ganavi H Patel[3], Likhitha R[4], Roja M V[5]

Department of CSE JNN College of Engineering,

Apeksha G, Ganavi H Patel, Likhitha R, Roja M V Department of CSE JNN College ofEngineering,

-- ***

Abstract - **The speech is the most effective means of communication, to recognize the emotions in speech is the most crucial task. In this paper we are using the Artificial Neural Network to recognize the emotions in speech. Hence, providing an efficient and accurate technique for speech based emotion recognition is also an important task. This study is focused on seven basic human emotions (angry, disgust, fear, happy, neutral, surprise, sad). The training and validating accuracy and also lose can be seen in a graph while training the dataset. According to it confusion matrix for model is created. The seven features Frequency, Pitch, Amplitude and formant of speech is used to recognize seven basic emotions from speech.**

Keywords: Emotion detection, Emotion classification, verbal emotion classification.

1. INTRODUCTION

Speech Emotion Recognition (SER) using Artificial Neural Networks (ANN) with Long Short-Term Memory (LSTM) and Mel Frequency Cepstral Coefficients (MFCC) is a powerful approach to analyze emotional content in speech. MFCCs convert raw audio into feature vectors, capturing essential frequency characteristics. LSTM networks then model temporal dependencies within these features, enabling the recognition of nuanced emotional patterns over time. ANN complements LSTM by learning complex patterns and mapping MFCC-LSTM representations to specific emotion categories. This integrated approach enhances accuracy and generalization in SER tasks, facilitating applications like human-computer interaction and mental health monitoring. By combining feature richness, temporal modeling, and pattern recognition, ANN with LSTM and MFCCs provides a robust framework for understanding and interpreting emotional cues embedded in speech signals.

2. LITERATURE SURVEY

In this section, various papers have been presented with various Machine Learning techniques.

In [1] employs a self-attention-based deep learning model combining CNN and LSTM networks for speech emotion recognition. Techniques include incorporating multiple data sources, experimenting with different acoustic features. The methodology involves training models on multiple datasets to enhance performance and robustness, utilizing advanced techniques like CNNs and attention models to distinguish subtle emotions. The study compares results with traditional RNNs and CNNs on various datasets, conducting feature selection experiments and applying normalization and augmentation techniques for improved accuracy.

In [2] paper employs a hybrid approach combining MFCCs and time-domain features for speech emotion recognition (SER). It utilizes a lightweight 1D CNN model with multiple layers for feature extraction and classification. The methodology involves data collection, feature extraction, model training, and prediction stages, achieving high accuracy rates compared to baseline methods. The study focuses on optimizing classification algorithms to enhance the robustness of the SER system.

In [3] paper employs a combination of acoustic features like MFCC and deep learning through a 1-D DCNN to enhance Speech Emotion Recognition (SER) accuracy. The methodology involves feature extraction, model training, and evaluation on datasets like EMODB and RAVDESS. Results show significant improvements in SER performance.

In [4] paper utilizes a two-way approach for feature extraction from speech data, employing deep neural networks for emotion recognition. It compares the proposed models with existing state-of-the-art approaches, focusing on evaluation metrics like accuracy score, loss, classification report, and confusion matrix for performance analysis.

In [5] paper integrates a hybrid data augmentation approach with a modified mADCRNN model to enhance speech emotion recognition. It combines traditional and generative adversarial network methods to generate additional data samples and extract utterance-level features for improved classification accuracy.

In [6] The paper proposes a hybrid model for Speech Emotion Recognition (SER) by combining LSTM and Transformer architectures to improve recognition by leveraging long-term dependencies in speech signals and using Multi Head Attention on the Transformer encoder layer with MFCC feature vectors.

In [7] The paper proposes an end-to-end multi-speaker emotional text-to-speech system with a condition encoder to modulate speaker voice features and emotions in the output. It then utilizes the synthesized emotional speech to augment speech emotion recognition systems, showing improvements in performance.

In [8] The paper utilizes a Wavelet Multiresolution Analysis Based Speech Emotion Recognition System using 1D CNN LSTM networks. Data preprocessing involves trimming leading and trailing silence and data augmentation with Additive White Gaussian Noise to enhance robustness.

In [9] The paper proposes a novel CNN-based multistage fusion architecture that utilizes summary linguistic embeddings from a pre-trained language model to condition multiple intermediate layers of a CNN operating on log Mel spectrograms. It contrasts this with single stage DNN architecture and late fusion methods, and includes an in-depth analysis of the impact of dialogue acts and differences between scripted vs. improvised text on an acted emotional dataset.

In [10] The paper proposes a speech emotion recognition (SER) model using parallel CNNs with multi-head self-attention layers for enhanced accuracy. It also incorporates Transformer encoders to capture robust temporal features in speech spectrograms, aiming to achieve state-of-the-art results.

Authors	**Research focus**	**Remarks**
Jagjeet Singh, Lakshmi Babu Saheer, Oliver Faust [1], 2023	Involves self-attention-based deep learning model combining CNN and LSTM networks.	Limitations of rhythmic features, and potential for multi-modal integration.
Ala Saleh Alluhaidan, Oumaima Saidani. [2], 2023	Hybrid approach combining MFCCs and time-domain features. It utilizes a lightweight 1D CNN model with multiple layers for feature extraction and classification	Difficulty of applying proposed technique to diverse SER datasets due to variations in expressions and environmental factors.
Kishor Bhangale and Mohanaprasad Kothandaraman [3],2023	Combination of acoustic features like MFCC and deep learning through a 1-D DCNN.	The need for further exploration on how to effectively handle class imbalances in the dataset.
Apeksha Aggarwal, Akshat Srivastava, Ajay Agarwal. [4],2023	It utilizes two-way approach for feature extraction, employing deep neural networks for emotion.	Dataset's bias towards North American speakers, potentially leading to reduced.
Nhat Truong Pham, Duc Ngoc Minh Dang [5],2023	A hybrid data augmentation approach with a modified mADCRNN model.	Issue in identifying representative features and addressing imbalanced labeling.
Felicia Andayani, Lau Bee Thengi, Mark Teekit Tsun. [6],2020	Hybrid approach combining LSTM and Transformer architectures using Multi Head Attention on the Transformer encoder layer with MFCC feature vectors.	Differentiation between emotions such as happiness and fear in different languages.

Authors	Research focus	Remarks
Abdullah Shahid, Siddique Latif, and Junaid Qadir. [7],2023	End-to-end multi-speaker emotional text-to-speech system with a condition encoder to modulate speaker voice features and emotions in the output.	Rhythmic features have limited control and struggle to synthesize high-quality emotional speech compared to baseline models.
Aditya Dutt and Paul Gader [8],2023	It utilizes a Wavelet Multiresolution Analysis using 1D CNN LSTM networks.	It requires improved feature extraction and overcoming severe overfitting.
Andreas Triantafyllopoulos, Uwe Reichel, Shuo Liu, Stephaet al. [9],2023	The architecture combines linguistic embeddings from a pre-trained model with a CNN operating on log Mel spectrograms.	Linguistic complexity and processing.
Rizwan Ullah, Muhammad Asif, Wahab Ali Shah et al [10],2023	It uses parallel CNNs with multi-head self-attention layers	It require high-quality datasets and context-aware models

3. METHODOLOGY

The Speech Emotion Recognition application is executed using the below methodology.

A. Pre-processing

1) Sampling: Signals which we used normally, are all analog signals i.e., continuous time signals. Therefore, for processing purpose in computer, discrete signals are better. In order to convert these continuous time signals to discrete time signals, sampling is used.

2) Pre-emphasis: The input signal often has certain low frequency components which will result in samples similar to their adjacent samples. Therefore, we are performing pre-emphasizing by applying a high pass filter on the signal in order to emphasize the high frequency components which represent the rapidly changing signal.

3) De-silencing: Audio signals often contain regions of absolute silence occurring at the beginning or at the end and sometimes in between higher frequencies. It is required to remove this unwanted part from the signal and hence de-silencing is performed.

4) Framing: For the purpose of analysis, observable stationary signal is preferable. We divide the input signal into small constituent frames of a specific time interval. Generally, for speech processing, it was observed that frame duration of 20-30 ms is implemented.

5) Windowing: Most of the digital signals are large and infinite that they cannot be analyzed entirely at the same time. In order to convert large digital signal into a small set for processing and analyzing, and for smoothing ends of signal, windowing is performed.

B. Feature Extraction

1) MFCC feature vector extraction for each frame: MFCCs represent the short-term power spectrum envelope. This envelope in turn is representative of the shape of the human vocal tract which determines the sound characteristics. Therefore, MFCC is a vital feature for speech analysis.

a) Fourier transform of windowed signal (FFT): MFCC is a spectral feature which is not extractable in the time domain. Hence conversion to frequency domain is required. In order to obtain the periodogram estimate, we have to convert the signal into frequency domain using FFT.

b) Mel filter bank: The Mel scale relates the perceived sound frequency to its actual frequency. An upper and lower limiting frequency is determined.

c) Logarithm of obtained frame energies: Logarithm of each of the filter bank energies is performed. This is because loudness is not perceived linearly. For loud sound, variations in energy may sound the same. Hence such compression or normalization is performed to bring perceived signal nearer to actual signal.

f) Discrete Cosine Transform to obtain real MFCC coefficients: After taking logarithm of obtained filter banks, DCT is applied on each of them. Due to the overlapping filter banks, these energies are correlated with each other. In order to decorrelate these energies, DCT is performed.

g) MFCC Feature vector extraction: DCT coefficients are related to energies of filter bank. High DCT values represents high rate of change of filter bank energies. For better performance and extraction, only 13 values of DCT coefficients are kept and rest are discarded.

C. Prediction: The prediction step in the system design process of a Speech Emotion Recognition (SER) model involves utilizing the trained model to make inferences or predictions based on new input data. Once the model has been trained on a dataset, it is then tested and validated to ensure its accuracy and reliability. The prediction step occurs when the model is deployed to make emotion predictions based on audio signals that it hasn't seen during training. In the context of the provided document, the prediction step would involve applying the proposed SER model to new or unseen audio data to recognize and classify emotions. This step would evaluate the model's effectiveness in accurately identifying emotions from speech signals by using the knowledge gained during the training phase.

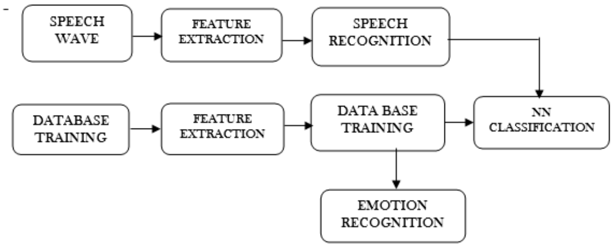

Fig 3.1 Methodology of SER Model

4. CONCLUSION

The conclusion emphasizes Artificial Neural Networks (ANN) as a cornerstone in speech emotion recognition, leveraging their prowess in deep learning and feature extraction from audio signals. Despite their capacity to handle high-dimensional data like spectrograms or MFCCs, ANN models encounter challenges such as limited labeled training data and cultural biases. Enhancing ANN architectures with techniques like transfer learning and domain adaptation can bolster their adaptability and generalization capabilities. Real-time feedback mechanisms and interdisciplinary collaborations play pivotal roles in refining ANN-based emotion recognition systems for interactive applications and nuanced emotion analysis. Through these efforts, ANN's potential in deciphering human emotions from speech signals can be maximized across diverse domains and scenarios.

REFERENCES

[1] Jagjeet Singh, Lakshmi Babu Saheer and Oliver Faust, “Speech Emotion Recognition Using Attention Model”, Vol-20 Issue-6 2023.

[2] Ala Saleh Alluhaidan, Oumaima Saidani, Rashid Jahangir, “Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network”, Department of Information Systems, College of Computer and Information Sciences, Saudi Arabia, Vol-13 Issue-8 2023.

[3] Kishor Bhangale and Mohanaprasad Kothandaraman, “Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network”, School of Electronics Engineering (SENSE), Vellore Institute of Technology, Chennai, India, Vol-12 Issue-4 2023.

[4] Apeksha Aggarwal, Akshat Srivastava, Ajay Agarwal, Nidhi Chahal, Dilbag Singh, Abeer Ali Alnuaim, Aseel Alhadlaq and Heung-No Lee, “Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning”, Department of Computer Science Engineering & Information Technology, Vol-22 Issue-6 2022.

[5] Nhat Truong Pham, Duc Ngoc Minh Dang, Ngoc Duy Nguyen, Thanh Thi Nguyen, “Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition”, FPT University, Viet Nam, Inveto Research, Brisbane,Australia, Khoury College of Computer Sciences, Boston, USA, Vol-230 2023.

[6] Felicia Andayani, Lau Bee Thengi, Mark Teekit Tsun, Caslon Chua, “Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files”, Faculty of Engineering, Computing, and Science, Swinburne University of Technology Sarawak Campus, Malaysia, Vol-10 2022.

[7] Abdullah Shahida, Siddique Latifb, and Junaid Qadir, “Generative Emotional AI for Speech Emotion Recognition: The Case for Synthetic Emotional Speech Augmentation”, Information Technology University (ITU), Punjab, Pakistan, University of Southern Queensland, Australia, 2023.

[8] Aditya Dutt, Graduate Student Member, IEEE, and Paul Gader, Fellow, IEEE, “Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks”, Vol-21 2023.

[9] Andreas Triantafyllopoulos, Uwe Reichel, Shuo Liu, Stephan Huber. “Multistage Linguistic Conditioning of Convolutional Layers For Speech Emotion Recognition”, University of Augsberg, Germany, Imperial College, London, UK, Vol-5 2023.

[10] Rizwan Ullah, Muhammad Asif, Wahab Ali Shah, Fakhar Anjam, Ibrar Ullah, Tahir Khurshaid, “Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer”, Department of Electrical Engineering, Chulalongkorn University, Thailand, Department of Physics and Astronomy, College of Science, King Saud University, Saudi Arabia, 2023.