WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Recognition using Machine Learning

Majji Ankitha

GMR Institute of Technology

ABSTRACT:

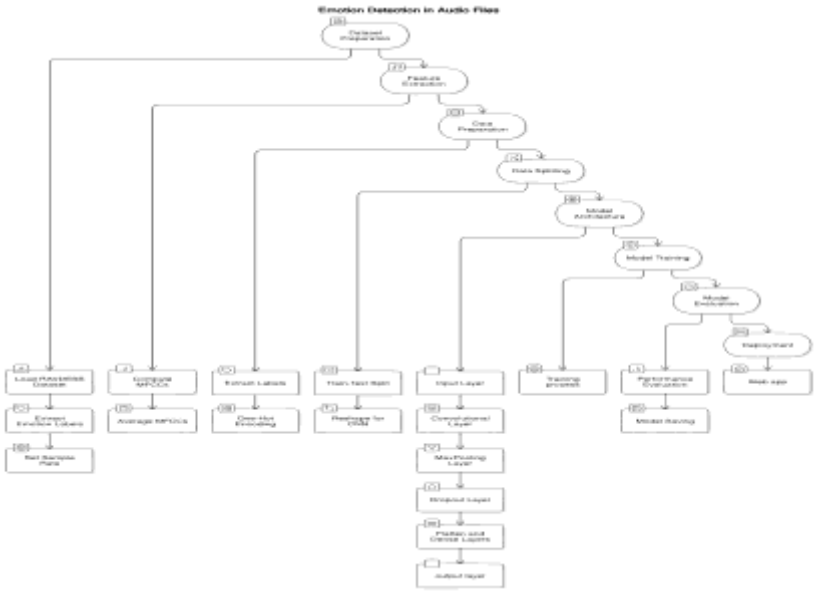
Speech emotion recognition (SER) is the process of analysing a person's voice and identifying the emotions expressed by him/her based on the pitch, volume, and tone. As speech enunciation is a primary human communication method, SER has become a more significant area of research in both academic and business settings. This study adopts the use of machine learning algorithms to investigate vocal features and classify different emotions such as joy, sadness, fear, thanksgiving, and neutrality. These feelings can be measured in order to extract valuable knowledge of human behavior, create better communication, and to help in services like education, mental health counseling, and customer service. Nevertheless, SER has serious obstacles mainly the subjective nature of emotions and the fact of efficient auditory input identification. The differences in voice caused by the languages, cultures, and personal expressions make the situation even more complicated. Despite these challenges, the applications of SER are numerous, such as enhancing human-computer interaction, emotional state monitoring, and creating personalized user experiences. The progress in large datasets and machine learning methods has notably improved the efficiency and effectiveness of SER making it an essential and dynamic research area that uses speech analysis to offer deeper understanding of human emotions and behavior.

Keywords: speech recognition, Emotion Recognition, K-nearest algorithm, Deep neural network, Machine learning.

Introduction:

Speech is a natural and powerful means of communication, which carries not only linguistic information but also emotional context. The ability to recognize emotions from speech has been an important task in numerous areas including human-computer interaction, customer service, and healthcare. Speech Emotion Recognition (SER) is the method that aims to detect and categorize emotional states including happiness, anger, sadness, and fear based on vocal characteristics. Due to the increasing availability of data and computing power, machine learning has become a dominant technology used to enhance the accuracy and efficiency of SER systems. Such systems can be trained to identify different emotional cues through various algorithms and can then excel in giving back more personalized and adaptive responses in real-time applications.

Methodology:



Dataset selection:

For this project, the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) database was selected. This dataset has the recordings of actors expressing and feeling the mixture of emotions from speaking and song. The name of each file is based on the emotion expressed, the speaker, and other details, hence, we can obtain emotion descriptions directly from the file name. For example, "03" in the file name means happiness, which we changed it to a number for processing. The sample rate was set to 16 kHz, which is a typical audio setting that gives a nice compromise between quality and speed of processing.

Eg: 03-01-06-01-02-01-12.wav

03: Modality = Speech

01: Vocal Channel = Speech

06: Emotion = Fearful

01: Emotional Intensity = Normal

02: Statement = “Dogs are sitting by the door”

01: First Repetition

12: Actor ID = Actor 12 (Male)

.wav: Audio file format

Feature selection:

It is an important step in SER, covering both model performance and computational efficiency, since it identifies only relevant features from the audio information that can contribute to accurate emotion classification while removing others irrelevant or redundant features.

Many of the features commonly used in SER are prosodic-like pitch, energy, and duration; spectral features such as MFCCs; and voice quality characteristics including jitter and shimmer. It is likely that the choice of an appropriate subset of these features could significantly improve recognition rate while reducing overfitting.

MFCCs, one of the most popular features in speech processing including Speech Emotion Recognition, is considered as a way to capture the characteristics of the human auditory system as regard how it considers human perception of sound frequency and, in this case, speech signals.

Extraction of MFCC involves the following steps:

1.Pre-Emphasis: Amplifies higher frequencies in the speech signal to balance the energy across the spectrum.

2. Framing and windowing: The signal is partitioned into small overlapping frames that are then windowed to minimize signal discontinuities at frame edges using Hamming windows.

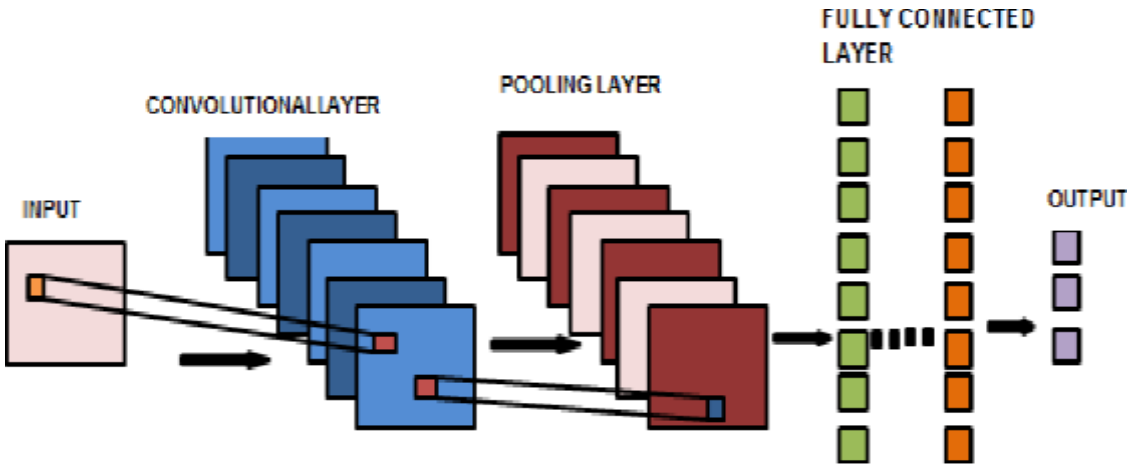
3. FFT: This converts the signal from the time domain to the frequency domain. It gives a spectrum representation.

4. Mel-Scale Filter Bank: This applies a number of triangular filters spaced at intervals given by the Mel scale, thereby stressing more the frequencies perceptible to human ears.

5. Log-Discosine Transform Hybrid: The DCT is taken on the logarithmically filtered outputs to yield the MFCC that summarizes the signal in compact representation of its frequency content.

Since these are very efficient, these codes focus on the perceptual properties of speech and then remove redundancy in feature representation. Typically, 12 to 13 MFCCs along with their first (delta) and second (delta-delta) derivatives are used to capture time dynamics in the speech signal.

Convolutional Neural Networks (CNN) in Speech Emotion Recognition:



Convolutional neural networks are a favorite deep learning architecture that works very well in feature extraction and, particularly in classification tasks. The CNN of SER is used to handle spectrograms or such other 2D representations of audio signals like MFCCs.

- **Convolution Layers:**

These layers, applying filters to the input data, look to extract local patterns-such as changes in pitch, tone, and energy-patterns all strong predictors of emotions in speech. Convolutional layers generate feature maps, highlighting these patterns.

•Activation Functions:

For instance, using ReLU gives the introduction of non-linearity, thus enabling the model to learn complex relationships in the data.

•Pooling Layers:

Pooling layers reduce the dimension of feature maps by maintaining the most important features while discarding redundant information, which helps reduce computational complexity and avoid overfitting.

• Flattening and Fully Connected Layers:

Features extracted are squashed to a single vector and fed into dense layers used for classification of emotions, like anger, happiness, or sadness.

• Output Layer:

A softmax or sigmoid activation function is used in the final layer for the classification of emotions.

Model Training:

Compilation:

- We used the Adam optimizer to help the model learn efficiently by adjusting learning rates automatically.
- The loss function chosen was categorical cross-entropy, which is ideal for our multi-class classification task.

Training Process:

- The model was trained for 30 epochs, meaning it went through all the training data 30 times.
- A batch size of 32 was used, so the model learned from small groups of 32 samples at a time.
- Validation data was used during training to check the model's accuracy and make adjustments as needed.

Applications:

- **Virtual Assistants:** Enhances interaction by enabling empathetic responses from devices like Alexa or Siri.
- **Healthcare:** Helps find out stress, anxiety, or depression to monitor mental health.
- **Call Centres:** Analyses customer emotions to improve service quality.
- **Education:** Adapts e-learning content based on students' emotions.
- **Entertainment:** Gaming or media content is customized according to users' moods.

- **Security:** It detects stress or aggression in observation systems.
- **Human Resources:** Assesses candidates' emotions during interviews.
- **Market Research:** Explores consumer sentiment to inform strategies.

Results & Analysis:

The CNN-based SER model performed well in the RAVDESS dataset. Using MFCC features, it really extracted good emotional information like pitch, tone, and energy variation successfully and classified emotions accurately, such as happiness, sadness, and anger. Good classification with minimum confusion between various categories of emotional data was represented by precision, recall, and F1-score measures. However, the subtle feelings and the characteristics overlapping posed problems in classification that demonstrated the intricacies involved in emotion recognition. The visual results, such as confusion matrices and accuracy plots, further validated the effectiveness of the model and emphasized MFCCs and CNN architecture as the reasons behind these results.

Conclusion:

This research succeeded with a CNN-based approach for speech emotion classification with MFCC features. It underscored the feature selection and deep learning techniques in improving the performance of the SER systems. Application areas of SER in virtual assistants, healthcare, education, and customer services came to be realized areas where emotions need to be understood. Future work could focus on addressing challenges like multilingual datasets and exploring advanced models to improve generalization and accuracy. This research reinforces the importance of SER in bridging the gap between technology and human emotional understanding.

References:

1. Taiba Majid Wani; Teddy Surya Gunawan; Syed Asif Ahmad, "A Comprehensive Review of Speech Emotion Recognition Systems", IEEE Access (Volume: 9),pp. 47795 – 47814, 2021.
2. K. Tarunika; R.B Pradeeba; P. Aruna."Applying Machine Learning Techniques for Speech Emotion Recognition", 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018.
3. Prof. Kinjal S. Raja, & Prof. Disha D. Sanghani. (2024). Speech Emotion Recognition Using Machine Learning. *Educational Administration: Theory and Practice*, *30*(6(S), pp.118–124.
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
5. and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1608-1612.
6. Hadhami Aouani, Yassine Ben Ayed, Speech Emotion Recognition with deep learning, Procedia **Computer Science, Volume 176,2020,Pages 251-260, ISSN 1877-0509**