WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Speech Emotion Recognition using Machine Learning

Nagalaxmi Bogem[1, a)], K. Little Flower[1, b)] and Thayyaba Khatoon Mohammad [1, c)]

[1]*Department of ML-DL, School of Engineering, Malla Reddy University, Maisammaguda, Dhulapally, Hyderabad, INDIA.*

[a)] *bogem.nagalaxmi@gmail.com*, [b)] *littleflower@mallareddyuniversity.ac.in*, [c)] *hodaiml@mallareddyuniversity.ac.in*, *thayyaba.khatoon16@gmail.com*

ABSTRACT.

"Speech is our natural way of communicating and it conveys not only words but also emotions. Speech Emotion Recognition (SER) aims to understand and interpret these emotions, helping to bridge the gap between humans and machines. Although it can be challenging to annotate emotions and capture their subjectivity, SER has the potential to greatly benefit us. Our research introduces an Artificial Neural Network (ANN)-based system designed to extract emotions from speech by focusing on important acoustic features. By utilizing existing datasets and models, we trained our ANN to recognize seven emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. Our proposed model achieved an impressive accuracy, reaching 100% during training and 99% during validation. This work significantly contributes to the advancement of SER technology, thereby paving the way for more empathetic human-computer interactions."

Keywords — Detection, Speech Input, Feature Extraction, SVM.

Introduction

The human voice is the primary mode of communication, offering a rich conduit for information exchange and capturing the nuanced emotions that arise in response to various stimuli. Research on automating the analysis of voice and facial expressions to detect emotions has seen a notable increase in recent years [1-3]. The broad potential applications of these systems in fields such as education, automotive technology, security, communication, and healthcare are fueling this growing interest.

Due to their accessibility and abundance of emotional information, auditory signals have replaced facial expressions as the primary modality for emotion identification, even though facial expressions still offer useful hints. In order to assess these acoustic cues, researchers have investigated a variety of classification algorithms, including well-known ones like Support Vector Machines (SVM), Hidden Markov Models, Gaussian Mixture Models, Neural Networks, and k-Nearest Neighbors (KNN) [4]. A number of techniques have been developed to identify human emotions from speech. In order to identify and categorize emotions using acoustic characteristics taken from emotional speech, these techniques rely on training data sets. A significant amount of study looks on the process of identifying emotional cues in audio data extraction. Usually, this procedure entails choosing or creating an emotive speech corpus, then painstakingly identifying its innate traits. Then, emotion classification is based on these extracted data, which can include prosodic and spectral features or both (refer to Figure 1). The precision of this categorization is predominantly dependent on the effectiveness of feature extraction, prompting scholars to investigate diverse methodologies such as the evaluation of spectral, prosodic, or their cooperative amalgamation. To accomplish accurate emotion categorization, for example, several research have merged prosodic energy characteristics with Mel-frequency cepstral coefficients (MFCCs) in a combined manner.

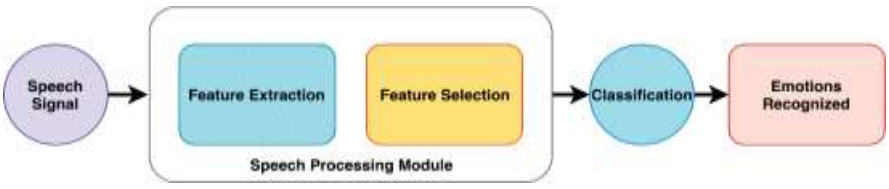

Figure 1. Model to extract Speech signal for Emotion Recognition.

Machines may struggle to recognize human emotions. The emotional data is critical in strengthening the link between the machine and the user. Cultural and environmental factors can influence hate speech. Emotions can be classified into two types: transitory and persistent. Modern machine learning methods, such as Convolutional Neural Networks (CNNs), K-Nearest Neighbors (KNNs), and Random Forests, have shown promise in detecting emotional states. CNNs are a sophisticated deep learning approach known for its feed-forward design and outstanding pattern recognition capabilities

(Figure 2). This architecture efficiently classifies data by employing layers of interconnected neurons, each performing localized analysis through "receptive fields." Figure 2 provides a visual representation of this layered structure in a basic CNN model.

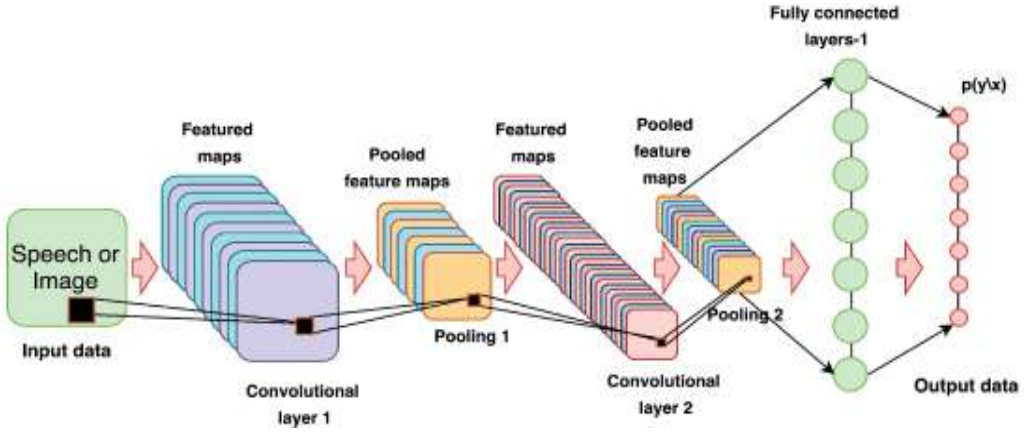

Figure 2. Layer-wise architecture for CNN network

Speech enhancement begins by segmenting it into smaller, meaningful units called phonemes or subword units. These segments are then meticulously analyzed for specific acoustic features, such as spectral energy, formant frequencies, and pitch variations, providing valuable clues about their underlying meaning and emotional content. Classification, based on extracted features, falls into two categories: short-term, focusing on brief temporal features, and long-term, utilizing statistics like mean and standard deviation. Additionally, prosodic features like intensity, pitch contour, speech rate, and their variability play a crucial role in deciphering emotional states expressed through speech. Table 1 illustrates some prominent acoustic features associated with various emotions.

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

Table 1. Characteristics of Emotions of Speech

Literature Review

The intricate connection between emotions and behaviors holds the key to unlocking emotion detection mechanisms within our physiological systems. Understanding how emotions arise within us could pave the way for their accurate identification [5]. Viewed from an evolutionary perspective, emotions can be interpreted as adaptive biological responses to external challenges and stimuli, playing a crucial role in human survival [5]. This realization has fueled extensive research within affective science, leading to the development of various theories and models for emotion detection and classification. According to Imran et al. [6], there are two primary categories of emotion classification models: discrete and dimensional. Discrete models categorize emotions into distinct, mutually exclusive classes. A prominent example is Plutchik's Wheel of Emotions [7], which identifies eight fundamental emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. Similarly, Ekman's model [8] proposes six "basic emotions" considered universally recognizable across cultures: anger, disgust, fear, happiness, sadness, and surprise [6].

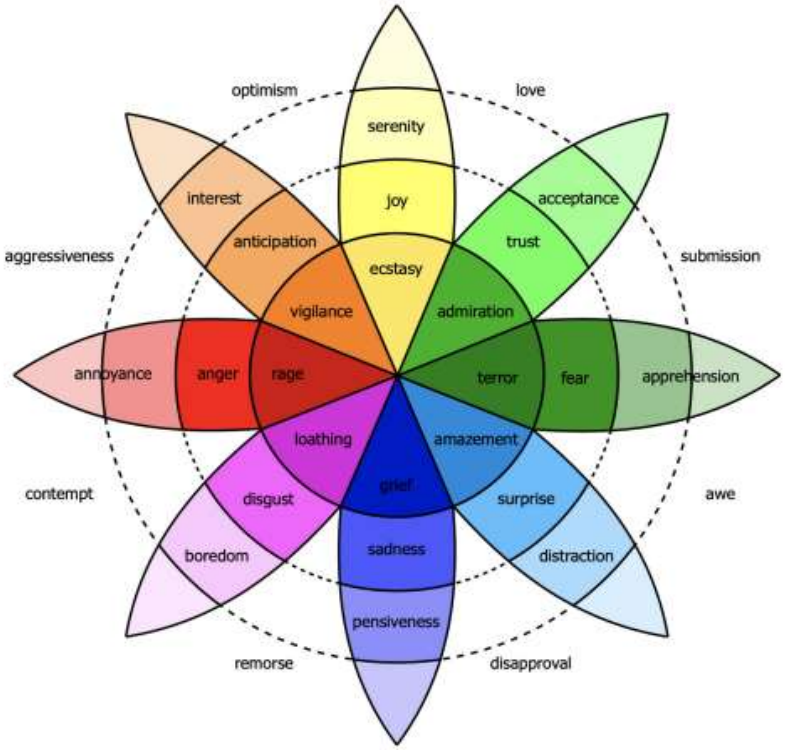

Figure 3. Wheel of emotion. Source[7]

In contrast to discrete models, dimensional models represent emotions along continuous axes or dimensions, typically including arousal, valence (positivity/negativity), and intensity. Several notable examples exist, such as the PANA model by Watson and Tellegen [9], the Circumplex model by Russell, and the PAD emotional state model.

Interestingly, Plutchik's model [7] bridges the gap, offering a well-known three-dimensional approach that incorporates both basic and complex emotions. This unique structure allows for combining emotions of varying intensities, resulting in complex "emotional dyads." As depicted in Figure 3, his model depicts eight core emotions like joy, trust, fear, and sadness on a wheel, which can then be combined into 24 distinct primary, secondary, and tertiary dyads.

Methodology

Before diving into the heart of our research, let's take a closer look at how we prepared the audio data for analysis. This crucial step, known as preprocessing, often requires specific techniques to address challenges inherent in the data. As highlighted in Table 2, researchers have successfully employed various supporting methods during this stage to enhance the performance of machine learning-based Speech Emotion Recognition (SER) systems.

One particularly noteworthy approach involves data augmentation. Imagine expanding your library of books for better understanding: data augmentation works similarly for audio data. By employing specific techniques, we can artificially generate additional audio samples based on the existing data. This not only increases the available dataset but also helps address potential biases or limitations within the original data, ultimately leading to a more robust and generalizable SER system.

Table 2. Challenges in Implementing ML-based SER systems

Challenges	**Solutions**
“Silence and noise removal”	“AFFT and adaptive filtering; Google WebRTC VAD.”
“Audio data insufficiency”	“Data augmentation”
“Data imbalance”	“Data augmentation; Skew-robust technique.”
“Speaker differences in audio data”	“Cross-Speaker Histogram Equalization (CSHE)”

"Signal trend issues"	"Removal signal trend by zero-crossing rate detection method"
"Perturbations in audio signal"	"Reducing the impact of perturbations by using down-sampling to decompose the signal."
"Solve the non-stationary and non-linearity issues"	"EMD-TKEO technique"
"Redundant and irrelevant information in LLD and HSF features"	"Feature selection methods. E.g., PCA, LDA, correlation analysis."
"Speaker-independent challenges"	"Attention model; Local Feature Learning block; A combination of loss function of centre-loss and softmax to train neural network"

Dataset

The initial module of our system development focuses on acquiring the input dataset for both training and testing purposes. We have sourced a readily available dataset containing 2,800 speech emotion audio recordings from the Kaggle platform (URL reference provided). This comprehensive dataset categorizes audio samples into seven distinct emotional classes: anger, disgust, fear, happiness, neutral, sadness, and surprise. Leveraging this well-established and diverse dataset serves as the foundation for effectively training and evaluating our system, ultimately enabling it to accurately recognize emotions within spoken language.

Kaggle Link: https://www.kaggle.com/datasets/jayaprakashpondy/speech-emotion-dataset

Data Augmentation

Data augmentation enhances the effectiveness of machine learning models by creating artificial variants of existing data samples. For audio data, this approach involves introducing modifications such as adding noise, altering pitch and tempo, or shifting the data along the time axis. These variations help the model become more robust to real-world changes in audio recordings, ultimately improving its generalization capabilities. Crucially, the emotional label assigned to each sample must remain unchanged throughout the augmentation process to ensure the model learns the correct associations. In the domain of image data, augmentation techniques often involve transformations like shifting, zooming, and rotating images, all while preserving the original label.

Importing the necessary libraries

Module two delves into the heart of our speech emotion detection system by incorporating two essential libraries. Librosa, renowned for its expertise in audio and music analysis, serves as our primary tool for processing and extracting meaningful features from the speech recordings. As we move towards the deep learning aspect, TensorFlow takes center stage, providing the robust computational framework necessary for building and training our powerful emotion recognition model.

Exploratory Data Analysis of Audio data

Before analysis, it's essential to understand how to handle raw audio data through preprocessing. This step involves loading and visualizing audio files using the IPython library and leveraging the Librosa library for efficient processing. Librosa normalizes the dataset, presents audio files in a unified sample rate, and offers a streamlined and standardized method for handling audio data. The increasing popularity of Librosa in audio signal processing can be attributed to its features: mono conversion, normalization, and standardized sample rate.

Discussions: Emerging techniques in SER

Researchers are continually refining Speech Emotion Recognition (SER) technology to improve performance, reliability, and complexity. Privacy protection is a primary focus. Federated Learning (FL) is a promising method for collaborative model training that safeguards user data. Challenges include communication and resource limitations on user devices, and the scarcity of labeled data in SER applications. Recent research explores semi-supervised learning within FL settings to address these challenges. These efforts demonstrate the potential of FL for privacy-preserving, data-efficient SER development.

CONCLUSION and future directions

This summary provides an overview of recent developments in speech emotion recognition (SER) systems, highlighting the potential applications such as reducing driver fatigue and aiding medical diagnoses. The study emphasizes the need for ongoing research to improve existing techniques and to develop new models that can enhance SER performance. It also points out the importance of addressing common obstacles in the SER process to effectively use machine learning for emotion detection.

References

Chavan, V. M., & Gohokar, V. V. (2012). Speech emotion recognition by using SVM-classifier. International Journal of Engineering and Advanced Technology (IJEAT), 1(5), 11-15.

Rao K. S., Kumar T. P., Anusha K., Leela B., Bhavana I. And Gowtham S.V.S.K., "Emotion Recognition from Speech" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2), pages: 3603-3607,2012.

Das, A., Nair, K., & Bandi, Y. (2022). Emotion Detection Using Natural Language Processing and ConvNets. In Data Science and Security: Proceedings of IDSCS 2022 (pp. 127-135). Singapore: Springer Nature Singapore.

Milton, A., Roy, S. S., & Selvi, S. T. (2013). SVM scheme for speech emotion recognition using MFCC feature. International Journal of Computer Applications, 69(9).

Gladys, A. A., & Vetriselvi, V. (2023). Survey on Multimodal Approaches to Emotion Recognition. Neurocomputing, 126693.

Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. Ieee Access, 8, 181074-181090.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Theories of emotion (pp. 3-33). Academic press.

Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.