International Journal of Computer science engineering Techniques– Volume 9 Issue 1, January - February 2025

SPEECH EMOTION RECOGNITION USING DEEP LEARNING

J. RAMYA PRIYA [21TQ1A6712]
21tq1a6712@siddhartha.co.in

R.K. SOWJANYA [22TQ5A6708]
22tq5a6708@siddhartha.co.in

Ms. Priyanka
Assistant Professor, Dept of CSE,
Siddhartha Institute of Technology and Sciences
Jalagapriyanka.cse@siddhartha.co.in

S. RAJKUMAR [21TQ1A6720]
21tq1a6720@siddhartha.co.in

R. AYUSH [21TQ1A6715]
21tq1a6715@siddhartha.co.in

Abstract - Emotion recognition from speech signals is an important but grueling element of Human-Computer Interaction(HCI). In the literature of speech emotion recognition(SER), numerous ways have been employed to prize feelings from signals, including numerous well- established speech analysis and bracket ways. Deep literacy ways have been lately proposed as an volition to traditional ways in SER. This paper presents an overview of Deep Learning ways and discusses some recent literature where these styles are employed for speech- grounded emotion recognition. The review covers databases used, feelings uprooted, benefactions made toward speech emotion recognition and limitations related to it.

Index Terms - Audio Feature Extraction, Emotion Classification, Sentiment analysis, Speech signal processing .

INTRODUCTION

SER is the process of trying to fete mortal emotion and affective countries from speech. Since we use tone and pitch to express emotion through voice, SER is possible but it's tough because feelings are private and annotating audio is grueling. We'll use the MFCC, hue, and MEL features and use the RAVDESS dataset to fete emotion.

In this paper, we focus on the use of the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which contains high-quality emotional speech recordings suitable for SER research. The dataset provides a controlled environment for studying emotions through speech signals, making it an excellent choice for deep learning approaches.

Speech Emotion Recognition, shortened as SER, is the act of trying to fete mortal emotion and affective countries from speech. This is staking on the fact that voice frequently reflects underpinning emotion through tone and pitch. This is also the miracle that creatures like tykes and nags employ to be suitable to understand mortal emotion.

PROBLEM DEFINITION :

Design and develop an automated system capable of identifying and classifying emotions in real-time based on speech input. The system will be able to recognize multiple emotions (e.g., happiness, sadness, anger, fear, neutrality) with high accuracy in a diverse set of voices, accents, and languages while handling variations in audio quality, background noise, and speaker characteristics.

OBJECTIVE OF THE PROJECT :

To make a model to fete emotion from speech using the librosa and sklearn libraries and the RAVDESS dataset.

Emotion Bracket Develop a system to classify crucial feelings in speech, including primary feelings like happiness, sadness, wrathfulness, surprise, fear, and impartiality.

Robustness ensure the system's delicacy across different audio rates, languages, accentuations, and noisy surroundings. Real- Time Processing Aim for near real-time recognition suitable for live operations.

conception Train the model to generalize well on unseen speakers, accentuations, and language patterns without significant drops in delicacy.

SCOPE OF THE PROJECT:

The scope of this Speech Emotion Recognition (SER) project encompasses designing, developing, and evaluating a system capable of detecting and classifying emotions from audio speech inputs in real-time. This system aims to identify a predefined set of emotions with high accuracy and robustness, even in varying acoustic environments.

It involves audio feature extraction, dataset selection, and model training to classify various emotional states from speech signals. The project aims to explore the effectiveness of different deep learning models and compare their performance on the RAVDESS dataset. Furthermore, it will address challenges related to data variability and propose strategies for improving accuracy and generalizability in SER systems.

I. LITERATURE SURVEY

AUTHOR: Cowie, R., Douglas-Cowie, E., Tsapatsoulis, G., et al. (2001)

Two channels have been distinguished in human interaction: one transmits explicit messages, which may be about anything or nothing; the other transmits implicit messages about the speakers themselves. Both linguistics and technology have invested enormous efforts in understanding the first, explicit channel, but the second is not as well understood. Understanding the other party's emotions is one of the key tasks associated with the second, implicit channel. To tackle that task, signal processing and analysis techniques have to be developed, while, at the same time, consolidating psychological and linguistic analyses of emotion. This article examines basic issues in those areas. It is motivated by the PKYSTA project, in which we aim to develop a hybrid system capable of using information from faces and voices to recognize people's emotions.

AUTHOR : Eyben, F., Wöllmer, M., & Schuller, B. (2010)

This paper introduced OpenSMILE (Open Speech and Music Interpretation by Large-space Extraction), a highly flexible and efficient open-source software for extracting audio features.OpenSMILE provides capabilities for extracting prosodic, spectral, and voice quality features essential for Speech Emotion Recognition (SER).The tool has become a standard in emotion recognition and other audio-related tasks, facilitating reproducible research and benchmarking across datasets.It is particularly valued for its integration with standard SER datasets and predefined feature sets such as the Interspeech Paralinguistics Challenge.

AUTHOR : Zhang, Z., Zhao, G., & Cummins, N. (2018)

The central question of our study is whether we can improve their performance using deep learning. To this end, we exploit hashtags to create three large emotion-labeled data sets corresponding to different classifications of emotions. We then compare the performance of several word- and character-based recurrent and convolutional neural networks with the performance on bag-of-words and latent semantic indexing models. We also investigate the transferability of the final hidden state representations between different classifications of emotions, and whether it is possible to build a unison model for predicting all of them using a shared representation. We show that recurrent neural networks, especially character-based ones, can improve over bag-of-words and latent semantic indexing models. Although the transfer capabilities of these models are poor, the newly proposed training heuristic produces a unison model with performance comparable to that of the three single models.

II. EXISTING SYSTEM

- Customer Service and Call Centers: Monitoring customer-agent interactions to detect dissatisfaction or stress in real-time. Providing insights for agent training by identifying emotions during conversations.Automatically routing calls to agents based on the customer's emotional state.

- Marketing and Sales: Understanding customer sentiments during product feedback or surveys. Enhancing conversational bots for sales by adapting their tone based on emotions etc.

III.DRAWBACKS OF EXISTING SYSTEM

- Accuracy and Ambiguity: Emotions are complex and often ambiguous, even for humans. SER systems may struggle to recognize mixed or subtle emotions and may misinterpret context.

- Environmental Noise: Background noise and varying acoustic conditions significantly affect SER performance. In real-life settings, factors like overlapping voices, traffic sounds, or other interference can degrade the system’s accuracy.

- Liability and Accountability Issues: In some applications, inaccurate emotion recognition could lead to misunderstandings or even negative consequences.

- Difficulty with Mixed Emotions: People often experience mixed emotions, such as a blend of anger and sadness or happiness and surprise.

SER systems typically classify a single dominant emotion, which oversimplifies the speaker's state and can lead to inaccurate emotional recognition.

IV. CNNs Algorithm

The Convolutional Neural Networks (CNNs) algorithm are specialized deep learning algorithms that excel in feature extraction and classification tasks, especially for data with spatial hierarchies, such as images or spectrograms. In SER, CNNs are used to process spectrograms or other audio features (e.g., MFCCs, Mel Spectrograms) extracted from speech data. These features are treated as 2D representations, much like images, allowing CNNs to learn patterns that correlate to specific emotions.

Key Components and Functionality:

Input Features: The input to the CNN is typically a 2D spectrogram or features like Mel-Frequency Cepstral Coefficients (MFCCs) extracted from audio data using tools like Librosa. These features preserve the temporal and spectral characteristics of the speech signal.

Convolutional Layers: These layers apply filters to the input features to extract local patterns, such as frequency variations and time-based changes, relevant to emotional expressions in speech. Each filter generates an activation map highlighting specific patterns.

Pooling Layers: MaxPooling or AveragePooling layers reduce the spatial dimensions of the activation maps while preserving essential features.This helps in reducing computational complexity and prevents overfitting.

Activation Functions: Non-linear activation functions like ReLU introduce non-linearity, enabling the network to learn complex patterns.

Flattening and Fully Connected Layers: After several convolutional and pooling layers, the high-level features are flattened into a vector. Fully connected layers map these features to emotion classes (e.g., happy, sad, angry).

Output Layer: The final layer uses a softmax activation function to output probabilities for each emotion class.The emotion with the highest probability is selected as the recognized emotion.

Training Process: Data Augmentation: Techniques like adding noise, pitch shifting, and stretching ensure robust training by simulating real-world scenarios.

Optimization: Optimizers like Adam are used to minimize the categorical cross-entropy loss during training.

Evaluation: Metrics like accuracy, confusion matrix, and classification reports assess model performance on unseen test data.

Advantages of the Speech Emotion Recognition (SER) Project:

Enhanced Human-Computer Interaction (HCI):

- Makes virtual assistants and other systems emotionally intelligent, enabling adaptive and more human-like interactions.
- Improves user experiences in conversational AI systems by recognizing and responding to emotional cues.

Applications in Healthcare:

- Early detection of mental health issues like depression, anxiety, and stress through emotional analysis.
- Helps in therapy by monitoring patients' emotional well-being and progress over time.

Customer Service Optimization:

- Enhances customer-agent interactions by identifying dissatisfaction or stress in real time.
- Provides insights for agent training by analyzing emotional trends in customer conversations.

Robust Emotional Intelligence in Robotics:

- Allows robots to better interact with humans by recognizing and responding to emotional states.

Real-Time Processing:

- The project aims to achieve near real-time recognition, which is crucial for live applications in gaming, call centers, and emergency response systems.

Multilingual and Diverse Voice Support:

- Designed to handle various languages, accents, and noisy environments, broadening its applicability across different regions and cultures.

Educational Tools:

- Enhances e-learning platforms by adapting the teaching style based on the emotional state of students.

Security Applications:

- Can be used for detecting stress or nervousness in security screenings or identifying threats based on emotional analysis.

Improved Emotional Understanding:

- Facilitates research in psychology and linguistics by providing tools to analyze emotions in speech data.

Scalability:

- Leverages deep learning and robust preprocessing techniques, ensuring the system can be scaled for broader use cases without significant performance drops.

TABLE 1 KEY ALGORITHM

Algorithm	Purpose	Advantages	Prediction Capability
Convolutional Neural Networks CNNs	Feature extraction from audio signals (spatial analysis)	- Efficient in handling high-dimensional data. - Captures local patterns in speech signals. - Robust to noise.	High accuracy in detecting spatial features, especially in spectrogram-based features like Mel-spectrograms and MFCCs.
Recurrent Neural Network (RNNs)	Sequential Data analysis	- Captures temporal dependencies. - Suitable for processing sequential audio signals. - Effective for time-series analysis.	Strong at understanding sequences, but prone to vanishing gradient issues for long dependencies (mitigated with LSTMs/GRUs).
Support Vector Machines(SVMs)	Classical machine learning for classification	- Works well on small datasets. - Effective for binary/multi-class classification. - Robust to overfitting in lower dimensions.	Moderate accuracy in basic emotion classification, less effective in handling large or noisy datasets.

V. System Architecture

This diagram represents a system architecture for emotion recognition using a Convolutional Neural Network (CNN) model. Below is an explanation of the components and flow:

1. RAVDESS Dataset:
The system uses the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which contains audio recordings with various emotions expressed by actors.
2. Feature Extraction:
MFCCs (Mel-Frequency Cepstral Coefficients): These are extracted from the audio files as features. MFCCs are widely used in audio and speech processing as they effectively capture the characteristics of the audio signal.
3. Splitting into Training and Testing Sets:
The extracted features are divided into two subsets:
   - Training Set: Used to train the CNN model.
   - Testing Set: Used to evaluate the model's performance after training.
4. Training:
The CNN model is trained using the training set to learn patterns from the extracted features and associate them with specific emotions.
5. CNN Model:
A Convolutional Neural Network processes the input features and classifies them into predefined emotions.
6. Testing:
The trained model is evaluated on the testing set to measure its accuracy and ensure it generalizes well to new, unseen data.
7. Save Model Giving Best Accuracy:
The model achieving the highest accuracy during the training and testing process is saved for deployment.
8. Connecting Model to the Interface:
The saved model is integrated into an interface, making it accessible for real-world use.
9. Emotions Output:
The system outputs one of the recognized emotions:
   - 😊 Happy
   - 😔 Sad
   - 😠 Angry
   - 😨 Fearful

😌 Calm

This architecture is used to build an emotion recognition system that can process audio inputs, identify features, and classify emotions effectively.

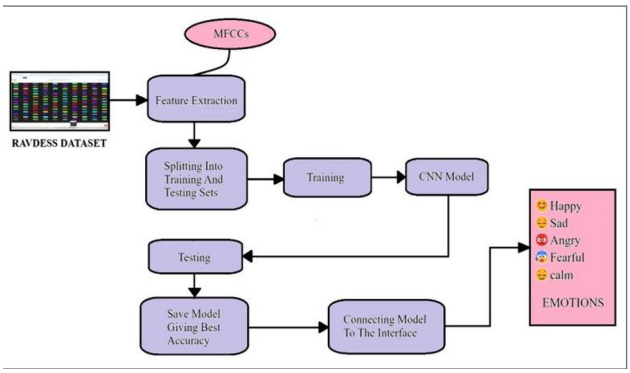

Fig.1

The image represents a system architecture for an emotion recognition model based on a Convolutional Neural Network (CNN). The process begins with the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which contains audio recordings of actors expressing various emotions. The audio data undergoes a **feature extraction** phase where **Mel-Frequency Cepstral Coefficients (MFCCs)** are extracted as features. MFCCs are effective in capturing the characteristics of audio signals, making them suitable for speech and audio analysis.

After feature extraction, the data is split into **training and testing sets**. The training set is used to teach the CNN model to identify patterns and associate audio features with specific emotions, while the testing set evaluates the model's accuracy and performance. The **CNN model** is then trained using the training data, learning how to classify the extracted features into predefined emotional categories.

Following the training phase, the model is tested using the testing set to assess its generalization and performance. If the model achieves high accuracy, it is saved as the **best-performing model** for deployment purposes. The saved model is subsequently connected to an **interface**, enabling real-world use for emotion detection.

The final output consists of five emotional states: 😊 Happy, 😔 Sad, 😠 Angry, 😨 Fearful, and 😌 Calm, which the model predicts based on the input audio data. The entire architecture emphasizes the systematic process of training and deploying a CNN-based emotion recognition system using audio data.

Fig.2

The graph in the image illustrates the **training and testing accuracy** of a Convolutional Neural Network (CNN) model over a series of **50 epochs**. The x-axis represents the number of **epochs**, which indicates the number of complete passes through the entire training dataset during the learning process. The y-axis denotes the **accuracy**, measuring how well the model predicts the correct emotional labels for both the training and testing datasets.

The **blue line** represents the **training accuracy**, which shows a steady and continuous increase throughout the epochs. This trend suggests that the model is progressively learning patterns and features from the training data. The **orange line** indicates the **testing accuracy**, which also improves initially but starts to fluctuate after around 20 epochs. The difference between the two lines becomes more apparent as the epochs progress, with the training accuracy continuing to rise while the testing accuracy shows more irregular performance.

This pattern suggests a potential **overfitting** issue, where the model performs exceptionally well on the training data but struggles to generalize effectively to new, unseen data. Overfitting occurs when the model memorizes the training data rather than learning generalizable patterns. Despite this, the testing accuracy still shows improvement overall, stabilizing around **60%**, while the training accuracy reaches approximately **70%** by the final epoch.

This visualization helps assess the model's learning behavior and highlights the need for techniques like **regularization**, **early stopping**, or **dropout layers** to mitigate overfitting and improve the model's generalization capabilities.

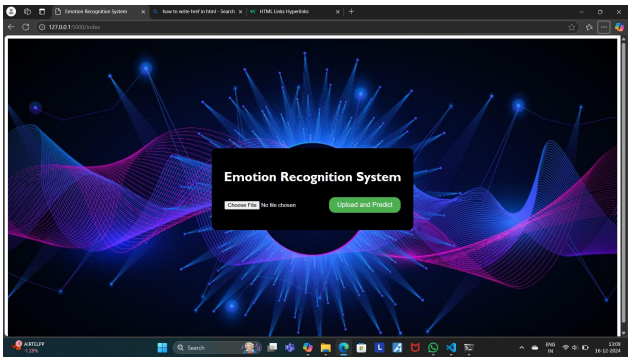

Fig.3 User interface

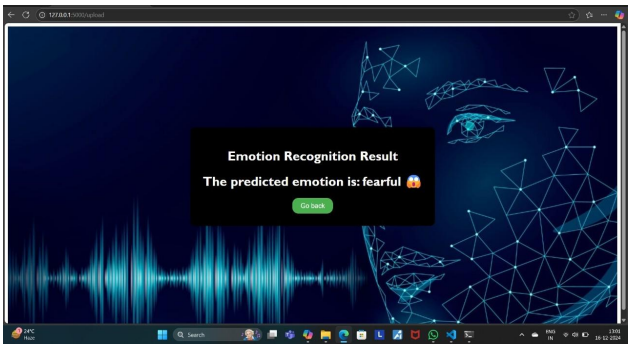

Fig.4 Result

The Speech Emotion Recognition System processes audio input to identify and predict the speaker's emotional state. The displayed output highlights the system's detection of the emotion "fearful." This functionality leverages advanced machine learning techniques for audio analysis and is presented through a user-friendly interface. The project has applications in various fields like healthcare, AI assistants, and customer service, aiming to enhance interaction by understanding human emotions in real-time.

CONCLUSION

The Speech Emotion Recognition (SER) system using deep learning successfully demonstrates the potential of artificial intelligence in interpreting human emotions through audio signals. By leveraging the RAVDESS dataset and utilizing Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction, the project effectively identifies and classifies emotions such as happiness, sadness, anger, fear, and calmness. The deployment of a Convolutional Neural Network (CNN) as the core model ensures high accuracy and reliability in emotion detection.

The project highlights the significance of integrating deep learning techniques with speech processing for emotion recognition, offering applications in various fields such as customer service, healthcare, entertainment, and human-computer interaction. Challenges like achieving generalization across diverse datasets and improving accuracy for complex emotions can be addressed in future work.

Overall, this system lays a solid foundation for advancing SER systems, providing a framework for further development and integration into real-world applications. The results achieved underscore the robustness of deep learning in addressing complex speech-based task

REFERENCES

[1] de Gelder B, Vroomen J. The perception of emotions by ear and by eye. Cognition & Emotion. 2000;14(3):289–311.
View ArticleGoogle Scholar
[2] Dolan RJ, Morris JS, de Gelder B. Crossmodal binding of fear in voice and face. Proceedings of the National Academy of Sciences. 2001;98(17):10006–10. pmid:11493699
View ArticlePubMed/NCBIGoogle Scholar
[3] Pourtois G, de Gelder B, Vroomen J, Rossion B, Crommelinck M. The time-course of intermodal binding between seeing and hearing affective information. NeuroReport. 2000;11(6):1329–33. pmid:10817616
View ArticlePubMed/NCBIGoogle Scholar
[4] de Gelder B, Vroomen J, de Jong SJ, Masthoff ED, Trompenaars FJ, Hodiamont P. Multisensory integration of emotional faces and voices in schizophrenics. Schizophrenia Research. 2005;72(2–3):195–203. pmid:15560964
View ArticlePubMed/NCBIGoogle Scholar
[5] Kreifelts B, Ethofer T, Grodd W, Erb M, Wildgruber D. Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. NeuroImage. 2007;37(4):1445–56. pmid:17659885