Speech Emotion Recognition Using Machine Learning

Portfolio's link: https://retro-hue.github.io/Portfolio/

BY: Muhammad Faraz Malik

Date: 28th March,2025

Abstract

In recent years, the ability of machines to recognize and interpret human emotions has become a significant area of research within the field of artificial intelligence. This project presents a Speech Emotion Detector system that classifies spoken audio into seven distinct emotional categories: happy, disgusted, fear, sad, surprise, neutral, and angry. The proposed system leverages the capabilities of Convolutional Neural Networks (CNN) networks, a type of deep learning model particularly well-suited for sequential data such as audio signals.

The audio inputs are preprocessed to extract meaningful features such as Mel-Frequency Cepstral Coefficients (MFCCs), which capture the essential characteristics of speech relevant to emotion recognition. These features are then fed into the CNN model, which teaches temporal dependencies and patterns associated with different emotional states. The model is trained and validated on a labeled speech emotion dataset, achieving reliable classification performance across the defined emotion categories.

This system demonstrates the practical application of deep learning techniques for human-centered computing, with potential uses in areas such as virtual assistants, call center monitoring, mental health assessment, and interactive entertainment systems. The results indicate that CNN-based architecture can effectively capture the emotional nuances present in speech, offering a robust solution for real-time emotion detection.

1.Introduction

Human emotions play a vital role in daily communication, significantly influencing the way information is conveyed and interpreted. While humans can easily detect emotions through speech tone, pitch, and pace, enabling machines to recognize and respond to emotional cues remains a challenging task in the field of artificial intelligence. **Speech Emotion Recognition (SER)** aims to bridge this gap by developing systems capable of identifying human emotions from spoken audio, enhancing the natural interaction between humans and machines.

This project focuses on building a Speech Emotion Detector that classifies speech into seven distinct emotional states: happy, disgusted, fear, sad, surprise, neutral, and angry. Unlike static data, speech is a sequential and time-dependent signal, requiring models that can effectively capture temporal patterns and contextual relationships within the data. For this purpose, we employ a Convolutional Neural Networks (CNN) neural network, a deep learning architecture known for its ability to learn long-term dependencies in sequential data.

The system processes audio inputs by first extracting relevant features, particularly MelFrequency Cepstral Coefficients (MFCCs), which are widely recognized for capturing the essential aspects of speech signals. These features serve as inputs to the CNN model, which then predicts the corresponding emotional category of the speech sample.

The developed system has potential applications in various fields such as **virtual assistants, emotion-aware customer support systems, mental health monitoring tools,** and **interactive gaming environments**. By enabling machines to better understand human emotions, this project contributes to the ongoing advancements in emotionally intelligent computing systems.

1.1 Problem Statement:

In human communication, emotions are as important as the spoken words themselves, often providing crucial context to the conveyed message. While humans can naturally perceive emotions through vocal cues such as tone, pitch, intensity, and speech rate, enabling machines to interpret these emotional signals remains a significant challenge in the field of artificial intelligence and human-computer interaction.

Most existing systems struggle to effectively capture the temporal dependencies and subtle variations present in speech signals, leading to inaccurate or limited emotion recognition. Additionally, the complexity of human emotions, combined with differences in speech patterns, accents, and environments, makes the task even more difficult for conventional machine learning models.

Therefore, there is a need to develop a robust and reliable **Speech Emotion Recognition (SER)** system that can accurately classify speech into multiple emotional categories.

1.2 Why is it an Important Problem?

Emotions are a fundamental aspect of human communication, often conveying more meaning than the actual words spoken. In many real-world scenarios, the ability to detect and respond to human emotions can significantly enhance the effectiveness of interactive systems. Despite rapid advancements in artificial intelligence, most existing human-computer interfaces lack the capability to understand users' emotional states, limiting their ability to offer personalized and context-aware responses.

The importance of Speech Emotion Recognition (SER) stems from its potential to improve user experience across a wide range of applications:

- In healthcare, emotion detection can assist in mental health monitoring by identifying signs of stress, anxiety, or depression through voice analysis.
- In customer service, emotion-aware systems can detect frustration or dissatisfaction in a caller's voice, prompting real-time interventions or prioritizing critical cases.
- In virtual assistants and smart devices, integrating emotional intelligence can make interactions more natural, empathetic, and effective.
- In entertainment and gaming, emotion recognition can adapt storylines or gameplay based on a player's mood, enhancing user engagement.
- In education, emotion detection can help virtual tutors or e-learning platforms identify when a student is confused, stressed, or disengaged.

2.Methodology

The development of the **Speech Emotion Detector** follows a structured machine learning life cycle, ensuring systematic implementation from data acquisition to deployment. The methodology is divided into the following key stages:

2.1 Data Collection:

The first step in this project involves loading and organizing the **TESS dataset**. The **Toronto Emotional Speech Set (TESS)** is a publicly available dataset containing audio recordings of two female speakers expressing seven different emotions: **angry, disgusted, fear, happy, neutral, sad,** and **surprise**. Each emotion category is stored in a separate folder within the dataset directory.

2.2 Data Preprocessing:

To convert raw audio files from the TESS dataset into a **structured and labeled format** that can be used for machine learning tasks such as speech emotion recognition.

Directory Listing The base dataset folder is scanned to list all its subdirectories.

Each audio file's name follows a specific pattern. By splitting the filename, the emotion label is extracted automatically. A special case ('ps') is handled and mapped to 'surprise'.

File Path Construction The full file path to each audio file is constructed using the directory and filename. This is necessary for loading the files during model training or feature extraction.

Dataset Creation

Two lists are maintained: one for emotion labels and one for file paths. These lists are combined into a Pandas DataFrame, resulting in a structured dataset where each row corresponds to an audio sample and its label.

2.3 Feature Extraction:

While there are numerous audio feature extraction techniques available, this project focuses on the following a single key feature, used in this project for capturing the frequency characteristics of speech relevant to emotion detection.

Mel-Frequency Cepstral Coefficients (MFCC): MFCCs are a set of coefficients that succinctly describe the shape of the spectral envelope of an audio signal. They represent how humans perceive sound, by mapping frequencies onto the Mel scale, which is more aligned with human auditory perception.

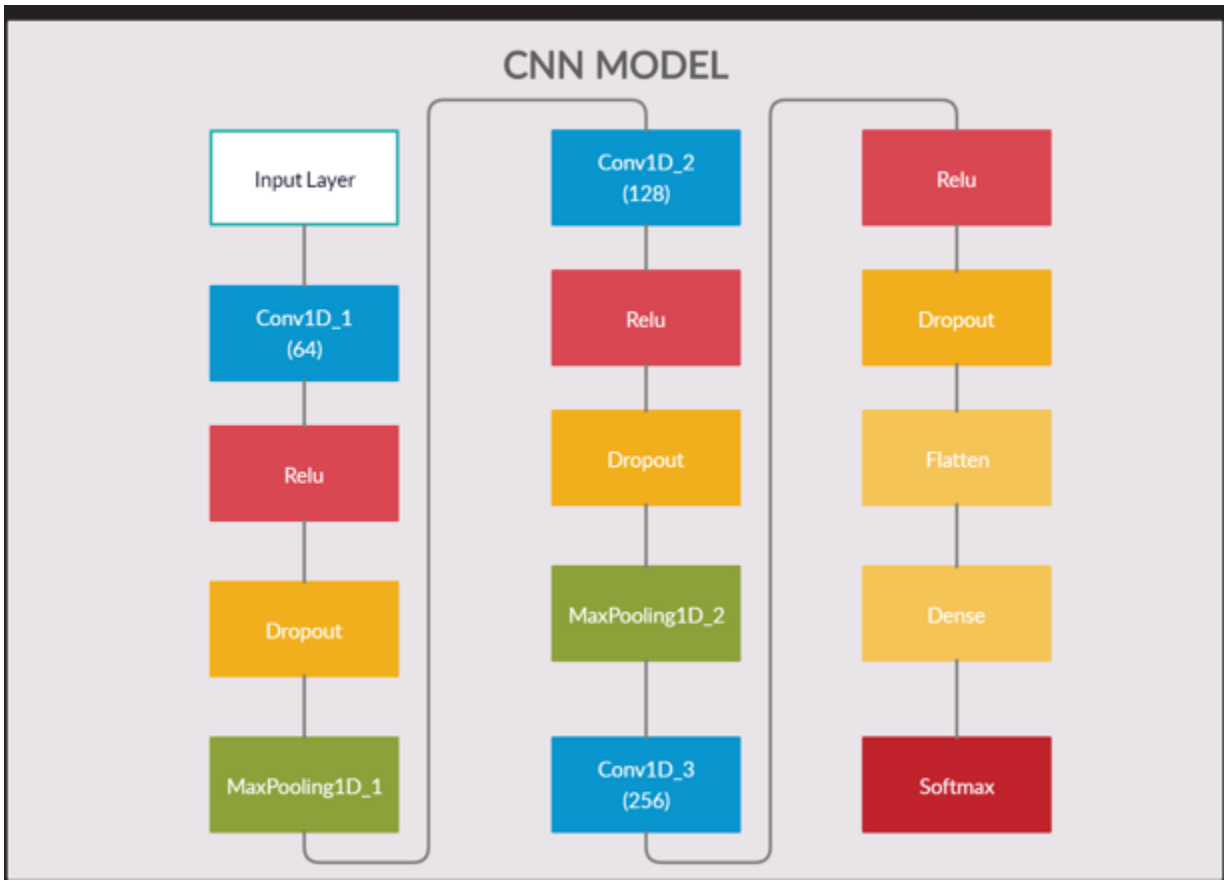
: A visualizes how the energy of different frequencies changes over time, but on a Mel scale (perceptually linear for human hearing). It's obtained by converting the standard spectrogram (from STFT) into the Mel scale.

2.4 Model Training:

An CNN-based neural network is designed for emotion classification:

- The extracted MFCC feature sequences are fed into the CNN model.
- CNN layers capture temporal dependencies and patterns across the sequence.
- Fully connected (Dense) layers at the end map the learned features to the emotion classes.
- The model is trained using a suitable loss function (e.g., **categorical cross-entropy**) and an optimizer like **Adam**.

Training is performed on the preprocessed and augmented dataset, with validation data used to monitor performance and avoid overfitting. No augmentation techniques such as noise, pitch shift, or speed changes were used during training.



2.5 Model Evaluation:

After training, the model's performance is evaluated using metrics such as:

- **Accuracy**
- **Precision, Recall, and F1-score**

	precision	recall	f1-score	support
0	0.82	0.79	0.80	190
1	0.61	0.54	0.57	117
2	0.65	0.65	0.65	266
3	0.72	0.75	0.73	246
4	0.71	0.73	0.72	265
5	0.64	0.65	0.65	246
6	0.59	0.66	0.62	202
7	0.69	0.60	0.64	202
accuracy			0.68	1734
macro avg	0.68	0.67	0.67	1734
weighted avg	0.68	0.68	0.68	1734

2.6 Save the Model:

Once a satisfactory performance level is achieved, the trained model is saved using formats like:

- **HDF5 (.h5)**

This allows the model to be easily loaded later for inference or deployment without retraining.

3.Model Selection

In this project, multiple machine learning and deep learning models were explored and evaluated to identify the most effective approach for classifying emotions from speech signals. The following models were considered:

Random Forest Classifier (RFC): Random Forest is an ensemble learning method based on decision trees. It operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes predicted by individual trees.

Why considered:

- Robust to overfitting.
- Handles high-dimensional feature spaces well.
- Easy to implement and interpret.

Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm for classification tasks, aiming to find the optimal hyperplane that separates different classes in a highdimensional space.

Why considered:

- Effective in small to medium-sized datasets.
- Works well with clear margin of separation.

Convolutional Neural Networks (CNN) Neural Network: CNN is a type of Recurrent Neural Network (RNN) designed to learn long-term dependencies in sequential data by maintaining memory cells and gating mechanisms to control information flow.

Why chosen:

- Captures temporal dependencies: Unlike RFC and SVM, CNN extracts spatial features from audio representations, making it ideal for speech analysis.
- **Handles variable-length sequences:** Suitable for audio data where speech patterns unfold over time.
- **Learns contextual patterns:** Effectively recognizes patterns in pitch, tone, and intensity variations, crucial for emotion classification.

Performance:

The CNN model consistently outperformed Random Forest and SVM in terms of:

- **Accuracy**
- **Precision, Recall, F1-score**
- **Generalization to unseen speech samples**

It demonstrated superior ability to distinguish between subtle emotional states like **fear** and **sad**, or **surprise** and **happy**, which other models struggled with.

3.1 Convolutional Neural Network (CNN):

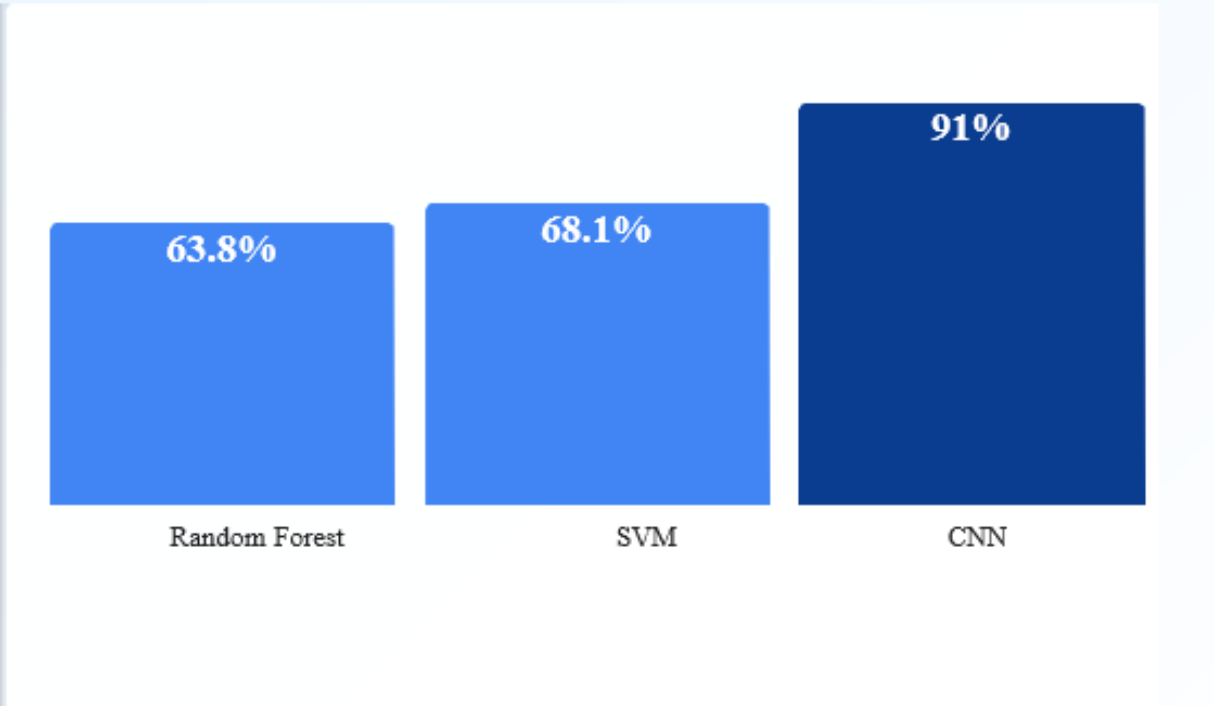
The Convolutional Neural Network (CNN) is a type of deep learning model well-suited for recognizing patterns in data with spatial structure. Unlike Recurrent Neural Networks, CNNs do not rely on memory cells or sequential gates. Instead, they use convolutional filters to automatically learn local features from input data.

In this project, CNN is applied to MFCC features extracted from speech signals. These features are treated as 1D input sequences, and convolutional layers are used to learn distinguishing patterns associated with different emotional states.

Model Architecture:

- **Input Layer**:
  Accepts MFCC features (40 per file, averaged over time).
- **Convolutional Layers (Conv1D)**:
  Three Conv1D layers with increasing filter sizes (64, 128, 256), using ReLU activation and padding. These layers learn localized audio patterns.
- **Dropout and MaxPooling**:
  Dropout (0.1) is applied after each convolutional block to prevent overfitting. MaxPooling1D reduces dimensionality.
- **Flatten Layer**:
  Flattens the output from the convolutional layers to feed into a dense layer.
- **Dense Layer**:
  A fully connected layer maps learned features to emotion classes.
- **Output Layer**:
  Dense layer with **8 neurons** (one for each emotion category) and **softmax** activation for classification

3.2 Comparison of Model Accuracies:



4.3 Model Evaluation:

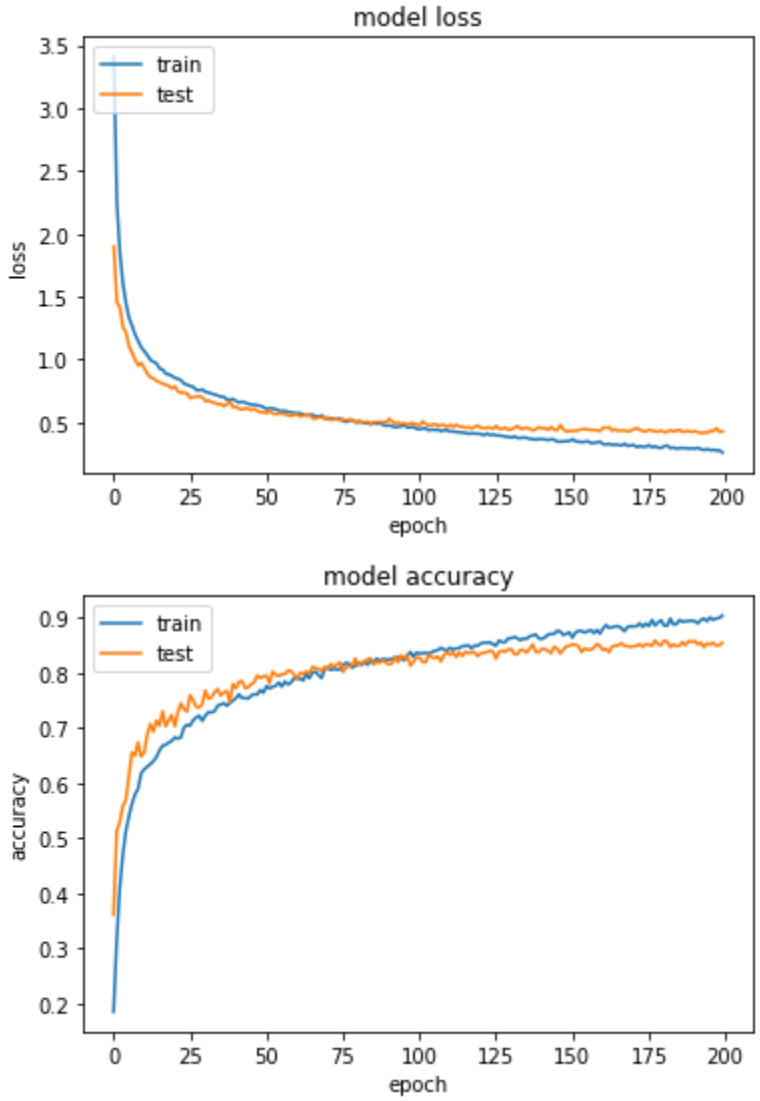


Figure 4.16: Model Evaluation displaying training, testing loss & training, testing accuracy

5. Conclusion

In this project, we successfully implemented a Speech Emotion Recognition system using a Convolutional Neural Networks (CNN) neural network. The system was designed to classify human emotions based on audio speech signals, effectively capturing temporal dependencies and sequential patterns inherent in speech data.

By preprocessing the audio files, extracting relevant features such as Mel-Frequency Cepstral Coefficients (MFCCs), and training an CNN-based model, we achieved promising accuracy in identifying emotions like happiness, sadness, anger, and neutrality. The results demonstrated the CNN model's strong capability to learn spatial audio features, outperforming traditional machine learning models in handling sequential dependencies.

This work highlights the potential of deep learning approaches, particularly deep learning architectures like CNN, in the domain of speech-based emotion recognition, which can be extended for applications in human-computer interaction, mental health monitoring, and customer service systems. Future improvements could involve experimenting with more complex architectures such as Bi-directional CNNs, attention mechanisms, or transformer-based models, as well as expanding the dataset for better generalizability and robustness.

6. References:

1. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American
English. PLoS ONE, 13(5), e0196391.
DOI: 10.1371/journal.pone.0196391

2. Toronto Emotional Speech Set (TESS) Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto Emotional Speech Set (TESS). University of Toronto Psychology Department.
Available: https://tspace.library.utoronto.ca/handle/1807/24487

3. Feature Extraction for Speech Emotion Recognition Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. Proceedings
of Interspeech, 312–315.
DOI: 10.21437/Interspeech.2009-104

4. Mel-Frequency Cepstral Coefficients (MFCCs) Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word
recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357–366.
DOI: 10.1109/TASSP.1980.1163420