EMODEPICTOR: HARNESSING EMOTIONS FROM AUDIO

Team Members: Soumya Shanigarapu, Tarun Kunkunuri, Teja Vineeth Reddy Yeramareddy

BACKGROUND:

Historically, human communication was heavily reliant on vocal cues to convey emotions, long before written language was established. As technology advanced, our modes of communication evolved, connecting us on a global scale but sometimes lacking the personal, emotional touch inherent in face-to-face interactions. In this digital age, understanding not just the words, but the emotions behind them, is pivotal. As we engage more with digital platforms and devices, there's a burgeoning need for them to resonate with our emotions, ensuring a deeper connection. The field of Speech Emotion Recognition (SER) stems from this necessity, striving to bridge the gap between human emotional expression and machine understanding. To achieve this, various tools, including RNNs, SVMs, and CNNs, have come to the forefront, offering potential pathways to more empathetic human-machine interactions.

INTRODUCTION:

In the digital world of today, personalizing experiences for users is becoming more and more important. This means not just showing people what they might like, but also understanding how they feel. With advances in machine learning and artificial intelligence, we're now looking into a new area: figuring out emotions from the way people talk.

Every day, we all chat and share our feelings and ideas. Imagine if our devices could not only hear our words but also understand the feelings behind them. This is what Speech Emotion Recognition (SER) is all about. While we often show our feelings through our faces, our voices can also tell a lot about how we feel.

Speech Emotion Recognition (SER) operates at the intersection of linguistics, psychology, and computer science. It's a discipline that goes beyond just transcribing spoken words into text; it delves into the nuances that often lie hidden beneath the surface of what we say. Every inflection in our voice, the pace at which we speak, and even the brief pauses we take, can convey a plethora of emotions. SER aims to tap into these subtle cues to paint a clearer picture of our emotional state.

The human voice is a powerful instrument, capable of expressing a vast range of emotions. From the exhilarating heights of joy and excitement to the somber tones of sadness and despair, our voice reflects our innermost feelings, often more accurately than our words. SER's challenge, therefore, is to decode these intricate vocal patterns and translate them into recognizable emotions.

The importance of SER extends beyond mere academic interest. In a world that's becoming increasingly digitized, having machines that can discern our emotional state can lead to more empathetic and responsive technology.

`

Imagine a virtual assistant that can sense when you're stressed and offers soothing music, or a customer service chatbot that can detect frustration in a caller's voice and escalate the call accordingly.

To dive deeper into this, we're using three main tools or models: Recurrent Neural Networks (RNN), Support Vector Machine (SVM), and Convolutional Neural Network (CNN). Each of these models has its strengths. Our main aim is to see which one is the best at figuring out emotions from speech.

PROJECT OBJECTIVES:

DATA COLLECTION:

Gathered data from Kaggle, ensuring data integrity and consistency.

CHARACTERISTICS OF DATASET:

- **Source & Accessibility:** The dataset is sourced from Kaggle and is known as RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song). Reference.
- **Volume:** It comprises over 1,400 audio clips.
- **Contributors**: 24 professional actors have provided both speech and song recordings for this dataset.
- **Diversity in Emotion:** The dataset encapsulates a broad spectrum of emotions, such as neutral, calm, happy, sad, angry, fearful, disgusted, and surprised, making it versatile for emotion recognition research.
- **Data Identification:** Each audio file in the dataset possesses a unique 7-part numerical identifier, which can be deciphered to understand details about the recording, including the emotion conveyed.
- **Emotional Categories:** RAVDESS classifies emotions into eight distinct categories. This categorization facilitates an in-depth study of vocal emotional expression nuances.
- **Audio Features:** Key audio features embedded in the clips include pitch, intensity, and spectrogram data, pivotal for emotion recognition modeling

FURTHER STEPS:

- **Audio File Visualization:**

  In the coming days, we'll embark on our dataset exploration phase. Our aim is to visualize the inherent patterns and structures by plotting waveforms and spectrograms of the audio samples. This visualization will provide clarity on the data we're working with.
- **Feature Extraction Using LibROSA:**

  We've earmarked the LibROSA library for our feature extraction, given its sterling reputation in audio analysis.

`

An essential step we're planning is to standardize the duration of all audio files to 3 seconds. This will ensure we extract a consistent number of features across the board.

Additionally, to enhance our feature set, especially given the dataset's size constraints, we're contemplating doubling the sampling rate for each file while ensuring the sampling frequency remains consistent.

- **Review of Extracted Features:**

  Once we proceed with and complete the feature extraction, we anticipate having arrays rich in pertinent data. Each array will be representative of an audio file, encompassing vital features and paired with its respective label. This curated data will set the stage for our subsequent analytical endeavors.

TRAINING THE MODELS:

Within the scope of our classification project, we will first train and then evaluate three promising models: the Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Support Vector Machine (SVM). Each of these models offers distinct advantages for various tasks, and it will be an intriguing endeavor to discern their potential efficacy for Speech Emotion Recognition (SER)."

METHODOLOGY:

We'll first collect audio samples and then visualize them using waveforms and spectrograms. Using the LibROSA library, we'll extract key features from each standardized 3-second audio file. With this processed data, we'll train models like CNN, RNN, and SVM and then compare their performance to identify the best approach for Speech Emotion Recognition.

PREDICTED OUTCOMES:

After fine-tuning the model, we will proceed to assess its predictive capabilities using the test data. This evaluation will not only revolve around a single model but will encompass all three - the Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Support Vector Machine (SVM). By comparing the accuracy and performance metrics of each model, we aim to derive a comprehensive understanding of their respective strengths and weaknesses. Our ultimate goal is to determine the most suitable model for the task at hand, ensuring that our Speech Emotion Recognition (SER) system is built upon the foundation of the best-performing analytical tool. This rigorous assessment will be instrumental in ensuring the reliability and effectiveness of our emotion recognition endeavors.

`