**Emotion recognition through speech**

Recognizing emotions holds significant significance within contemporary marketing strategies. The ability to tailor experiences uniquely to an individual's preferences is paramount. With this motivation, we embarked on a project aiming to discern emotions solely from a person's voice. This advancement can revolutionize various AI applications, such as implementing music in call centers when anger is detected or enabling smart cars to adjust speed based on emotions like anger or fear. Consequently, these applications possess substantial potential to offer advantages to companies and enhance consumer safety on a global scale.

**INTRODUCTION**

This report has explored the application of deep learning techniques using the TESS (Toronto Emotional Speech Set) dataset. It throws light on the process of extracting meaningful features from speech data and building a machine learning model to classify emotions. To this end, the project has leveraged the power of TensorFlow, an open-source machine learning framework, to build Long Short-Term Memory (LSTM) models for emotion recognition. Additionally, the Mel-Frequency Cepstral Coefficients (MFCC) has been employed as feature extraction techniques to preprocess the speech data effectively.

**About the data**

The TESS dataset consists of 2,800 recordings of seven different emotional expressions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. Each recording is a short clip of a single spoken word expressing the emotion. The dataset is divided into separate folders, each representing an emotion category. The librosa library was used to analyze sound files and IPython library to display and hear audios.

**EXPLORATORY DATA ANALYSIS**

After the TESS data was formatted into a usable format, each individual label (emotion) was plotted, and it turned out that all classes are equal in distribution.

There are several types of plots that can be used to visualize and understand the emotional characteristics of speech data. Some common plots used in voice-based emotion recognition are waveplots and spectrograms, for which two functions were defined.

* **waveplot() to view the waveform of the audio file**
* **spectogram() to view the frequency levels of the audio file**

Thus, each emotion’s audio file was tested using these two functions. It was observed that lower pitched voices have darker colours whereas higher pitched voices have more brighter colours.

**Feature extraction**

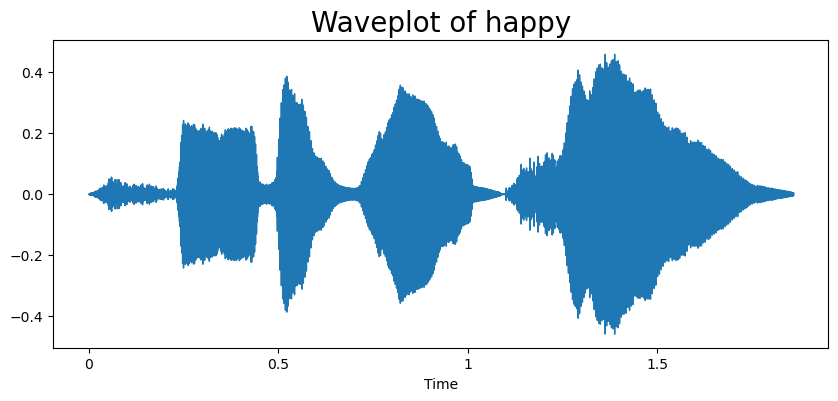
This project has used **MFCC** for feature extraction: they are like special features that help the computer understand sounds. They capture the spectral characteristics of audio signals and are widely used in tasks like speech recognition, speaker identification, music genre classification, and more. Imagine breaking down the sound into different frequency parts and then figuring out how much energy is in each part. We then make these energy values look more like how our ears hear sounds. After that, we arrange and compress these values using math. The resulting set of compressed values, called MFCCs, become a kind of special code that describes the important parts of the sound. Here a function has been defined for the same which will extract the MFCC features with the limit of 40 and take the mean as the final feature.

This preprocessing step is crucial for providing meaningful inputs to the LSTM model. OneHotEncoder was used to convert the labels into categorical form.

A purple and black image

Description automatically generatedA blue graph with text

Description automatically generatedA blue and orange striped background

Description automatically generated with medium confidence

**MODEL ARCHITECTURE**

LSTM, a type of recurrent neural network (RNN), is chosen for its ability to handle sequence data effectively and capture long-term dependencies. An LSTM learns to remember and use information from different parts of a sequence to make predictions or understand patterns in the data. A sequential model has been designed using TensorFlow's Keras API. The input to the model consists of the extracted MFCC features, and the output is a probability distribution over the emotional categories.

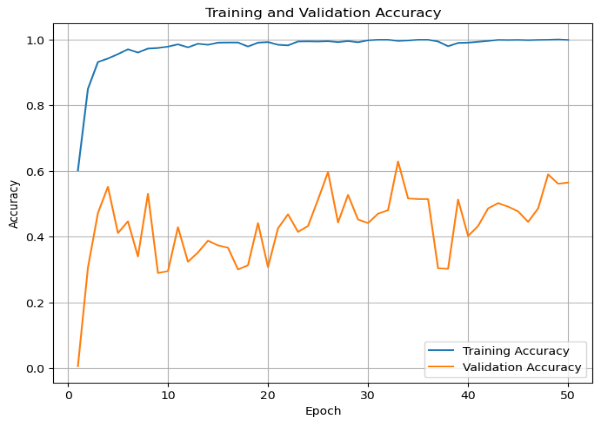
**Model Training and Evaluation:** The TESS dataset is divided into training and validation sets. The model's performance is evaluated using metrics such as accuracy, validation accuracy, and training loss and validation loss.

**RESULTS**

The training accuracy and validation accuracy increases with each iteration. We can see that the accuracy reaches nearly 99%. Similarly, the validation accuracy is also mostly increasing. Therefore, we can say that this is a good model. Here number of iterations used were 50.

The graph for accuracy shows that accuracy for training increases steadily and stabilizes near 0.9. For the validation accuracy there are more jumps but there is a slight increasing trend. Overall, it's a moderately good model. We can use more data for better results.For the loss graph: In an ideal scenario, both losses decrease as the model learns from the data, with the validation loss trailing slightly behind the training loss. This indicates that the model is learning meaningful patterns without overfitting. However, if the validation loss starts to increase or remains stagnant while the training loss drops, it suggests that the model might not be generalizing well, and adjustments might be needed to prevent overfitting. In this case, the training loss decreases however, A graph with orange lines and blue lines

Description automatically generatedvalidation loss is not decreasing simultaneously, nor is it remaining stagnant. It's a moderate accuracy overall.

In conclusion, more training data will get you better accuracy. This report showcases the successful implementation of deep learning techniques for emotion recognition through speech, demonstrating the fusion of powerful tools like TensorFlow, LSTM models, and MFCC feature extraction methods in the field of data science and audio analysis.