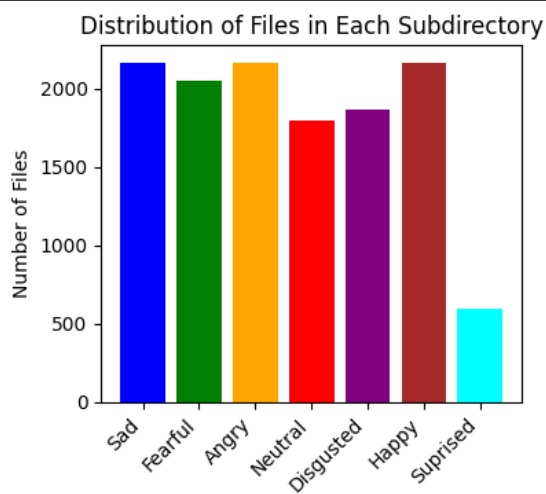
**Introduction**

Audio emotion recognition technology is transformative for customer service and mental health care. In customer service, it interprets emotions from vocal interactions, enhancing satisfaction and personalizing services.   
For mental health, it monitors emotional states through speech, supporting early distress detection and therapy. This technology offers significant advancements in understanding and responding to emotional cues, benefiting both sectors.

**Dataset**

Our project utilized a comprehensive dataset sourced from Kaggle, comprising roughly 12,500 audio samples, categorically divided across seven distinct emotions. The distribution of samples per emotion is as follows:

****

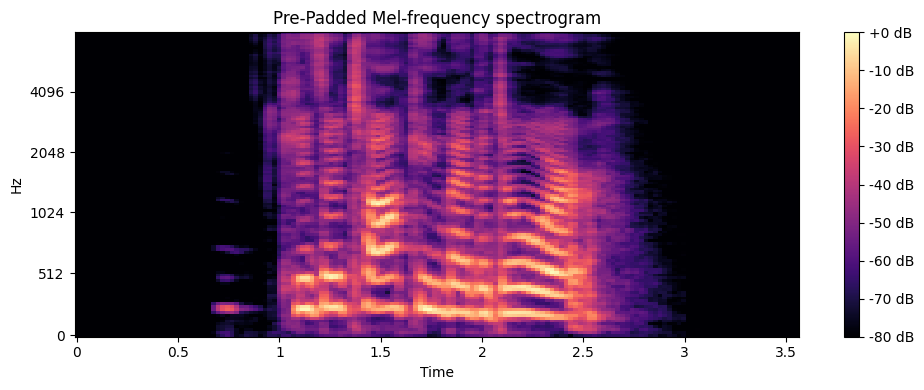
During our data review, we encountered numerous recordings with ambiguous emotional expressions, challenging even for our human judgement. This raised concerns about the potential impact on our models' performance, given the complexity of accurately classifying such nuanced samples. Here are two examples:

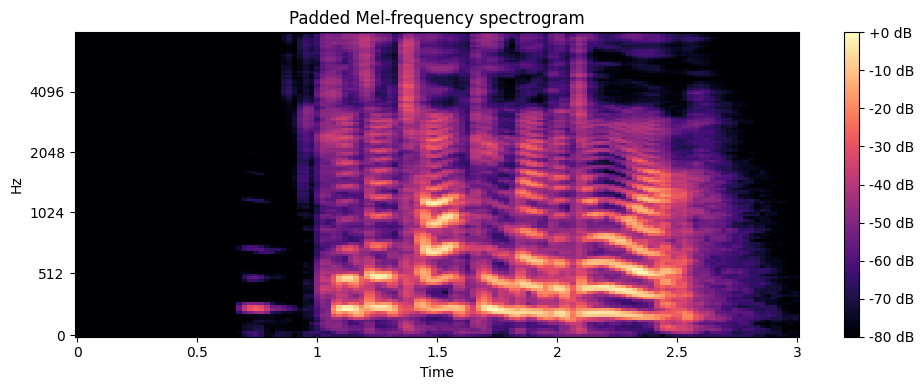
****

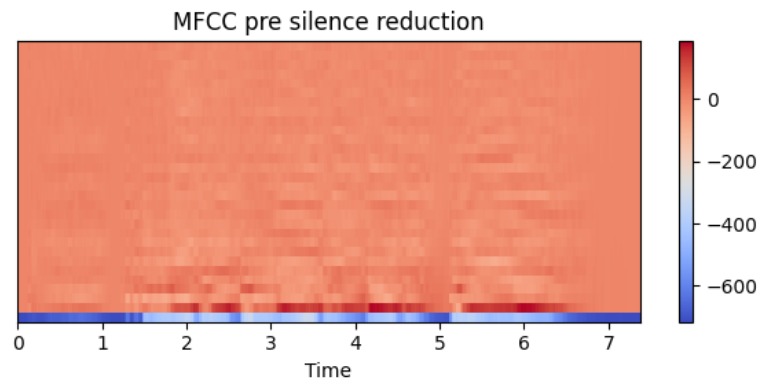
The implications of these findings on model efficiency will be discussed further in the subsequent sections.

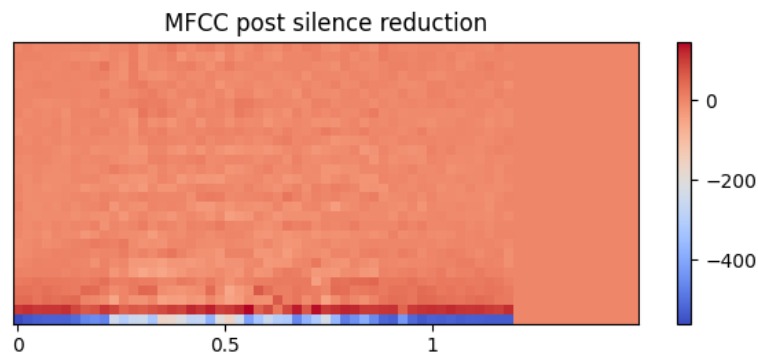
Following the acquisition of audio files from the dataset, our preprocessing approach bifurcated to cater to the distinct requirements of our Spectogram and MFCC-based models:

**For the CNN Model**: Each audio file was transformed into a Mel Spectrogram, encapsulating the crucial frequency information over time. This conversion facilitated the extraction of spatial features relevant for our CNN architecture. Observing the length distribution of the spectrograms, with a median of 107 and an average of approximately 118, we standardized the input by padding all spectrograms to a consistent length of 130. This length was chosen to accommodate the majority of the data while maintaining computational efficiency.





**For the MFCC Model**: Building on the preprocessing steps for the CNN model, we further refined the Mel Spectrograms by trimming silence from both the start and end of each file. This additional step aimed to concentrate the model's focus on the substantive audio content, potentially enhancing the MFCC model's ability to discern and classify emotional nuances effectively  




The resulting MFCCs are a compact representation of the spectral characteristics of the audio signal, capturing important features such as the distribution of energy in different frequency bands, timbral information, and speaker characteristics.

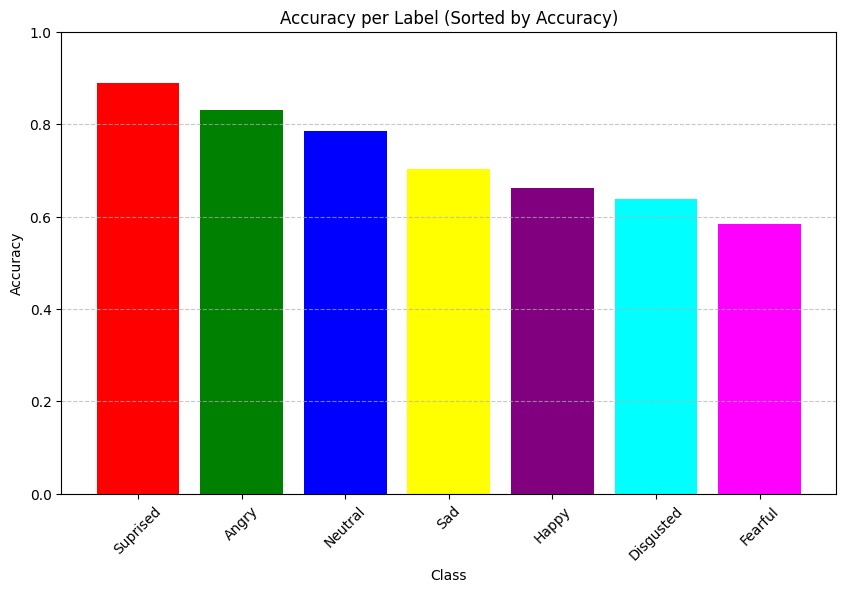
**Spectogram CNN Model**   
Our Spectogram model, tailored for spectrogram analysis, features a sequential architecture beginning with a Conv2D layer tailored to the input shape. It includes multiple convolutional and pooling layers for feature extraction, followed by flattening and dense layers for classification. The model incorporates batch normalization and L2 regularization to enhance performance and stability. The output layer, designed with softmax activation, classifies inputs into five emotion categories. Optimization is managed by an Adam optimizer with an adaptive learning rate, complemented by early stopping to prevent overfitting. This streamlined structure effectively processes audio spectrograms for emotion recognition.

**MFCC CNN Model**

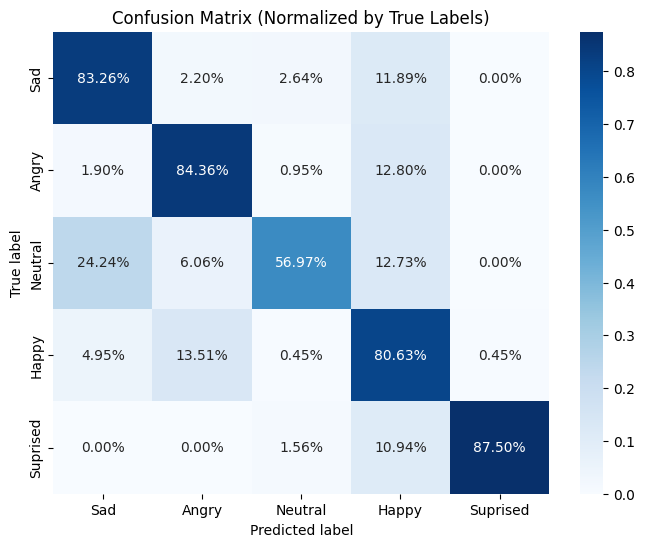
Our MFCC CNN model layers extract features from MFCC inputs, scaling up from 64 to 256 filters. SpatialDropout2D and GlobalAveragePooling2D enhance the model's efficiency and generalization. Dense layers with dropout manage complexity, leading to a softmax output for multi-classification. Optimized with Adam and safeguarded by EarlyStopping, it's finely tuned for nuanced emotion recognition from MFCCs. By trimming silent parts and downsampling the images, we were able to reduce the input shape of the model to 30\*65.

**Training Process**

We divided our dataset into training, testing, and validation sets with an 80:10:10 split to ensure a balanced representation for model evaluation. As we feared, during training we observed lower accuracy for the "Fearful" and "Disgusted" emotions. We decided to exclude both emotions, in order to enhance overall model performance. This adjustment aimed at refining our models' ability to more accurately classify the remaining emotions.

****

After reprocessing the data to focus on five emotions—excluding "Fearful" and "Disgusted"—we refitted the models. This strategic refinement led to a significant improvement in accuracy, with results surging by nearly 15%. This enhancement underscores the importance of targeted data preprocessing in optimizing model performance.

****

Ultimately, the performance of both the Spectogram and MFCC models converged aroung 78 percent overall accuracy. This outcome highlights the versatility and capability of both architectures in emotion recognition from audio data, despite the initial variations in their design and preprocessing approaches.