SpeakFeel: Emotion Recognition In Speech

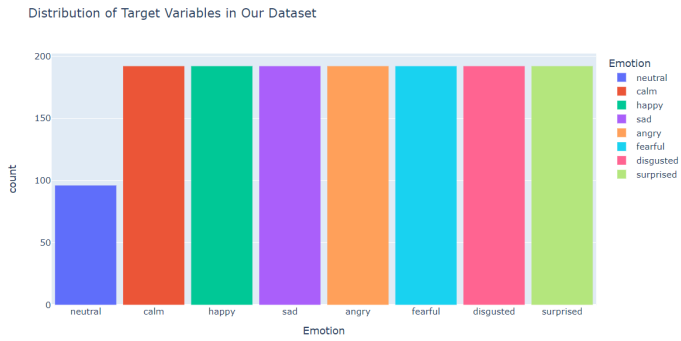
Milestone Two: Data Quality

Bach Le, Kien Tran, Sauryanshu Khanal and Sike Ogieva.

Distribution of Parameters in Our Dataset

We use the audio speech-only portion of the Ryerson Audio-Visual Database of Emotional Speech and Song [2]. Upon investigation [1], we found it to be a beautiful, carefully curated dataset, with data points generally well distributed amongst our target variables (eight different emotions).

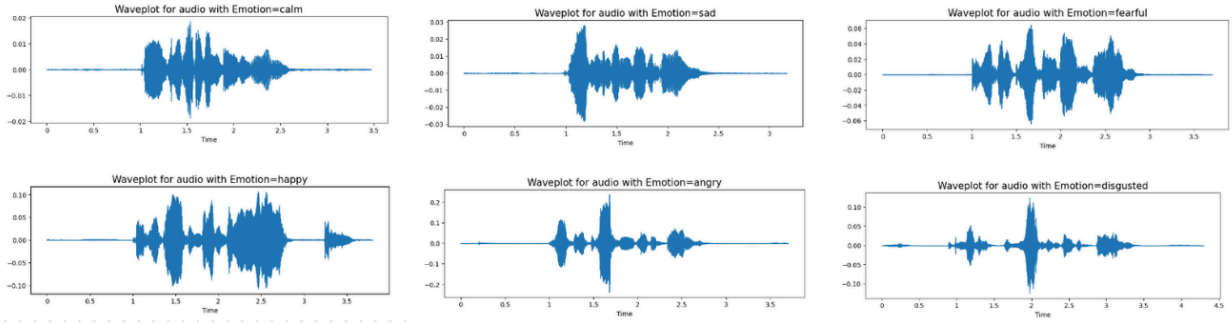
The exception is the neutral emotion which only has entries for a normal intensity, while the other labels have data for both normal and strong intensities. This is a reasonable property, but we will take note of if and how this affects our model.



Additionally, the other features (the gender of the speaker, the statement being spoken, the speech intensity, and the speaking individual of which there are twenty four persons) were perfectly distributed within the larger data set and also within each target variable [1].

Quality of the Audio Files

Having dealt with the dataset distribution, we go on to examine the audio data itself– the .wav files. We want to know whether there are enough differences in the audio such as frequency, amplitude, troughs and crests etc, across different labels such that a good model can find features to differentiate on. We use spectrograms and wave plots to answer this question. While spectrograms plot the discrete values in the audiofile against time, for the wave plots (which are continuous in form), we use a "sampling-rate" parameter to obtain those discrete values.



Our spectrograms and wave plots suggest that there is enough variability within our labels and attributes in terms of audio information–frequency, amplitude etc [1]–that given a proper modeling implementation, we should be able to make good predictions.

SpeakFeel: Emotion Recognition In Speech

Milestone Two: Data Quality

Bach Le, Kien Tran, Sauryanshu Khanal and Sike Ogieva.

Data Augmentation

There were 1440 audio files in the set and we applied audio augmentation techniques (adding noise, stretching, time shifting and changing the pitch) to bring the number up to 4_896 files [1]. Existing project work indicates that this is large enough to train a model to achieve over 90% test accuracy [3].

Our accompanying code where we analyze the feature distributions, create the spectrograms and wave plots and implement the data augmentation process is linked below, as first reference [1]

Relevant Related Works

This section reviews three related studies employing deep learning for speech emotion recognition (SER).

- Model 1: This work proposes a deep learning architecture based on Long Short-Term Memory (LSTM) networks. The model utilizes Mel-frequency cepstral coefficients (MFCCs) extracted from audio files as input. A single unidirectional LSTM layer captures temporal dependencies within the speech sequence. The LSTM output feeds into a dense layer for final classification into emotional categories. This model achieves exceptionally high 100% validation and test accuracy, demonstrating strong performance. However, we think further investigation is necessary to assess generalizability on unseen data [3].

- Model 2: The work presented in Model< 2 explores a convolutional neural network (CNN) based approach for SER. The model leverages various features extracted from the speech data, including zero-crossing rate (ZCR), chroma features, MFCCs, root mean square (RMS) value, and mel spectrogram. A series of 1D convolutional layers aim to detect patterns within the speech spectrogram. Similar to Model 1, the convolutional output feeds into a dense layer for emotion classification. However, the reported test accuracy of 59.7% suggests the need for further optimization on this approach [4].

- Model 3: The Model 3 investigates a deep learning architecture that combines convolutional and recurrent neural networks. The model utilizes MFCC-related data as input and employs 1D convolutional layers for initial feature extraction. It is then followed by a sequence of bidirectional LSTMs. This combination allows the model to capture dependencies in both past and future contexts within the speech sequence. The final output is processed by a dense layer for emotion classification. While not achieving the exceptional accuracy of Model 1, the model demonstrates promising results with a test accuracy of 85.53%. Thus, it opens a new strategy in combining RNN and CNN [5].

Next Steps

Moving forward, we will extract the Mel-Frequency Cepstral Coefficients (numerical data which mimics human hearing) from the audio files, and begin building both the kNN baseline and actual neural network model.

SpeakFeel: Emotion Recognition In Speech

Milestone Two: Data Quality

Bach Le, Kien Tran, Sauryanshu Khanal and Sike Ogieva.

References

1. Speak-Feel Data Exploration Notebook 01 (accompanying code)
2. RAVDESS Emotional Speech Dataset on Kaggle
3. SER Notebook - AbdelRahman on Kaggle
4. SER Notebook - Aditya on Kaggle
5. SER Notebook - Farneet Singh on Kaggle
6. Ravdess + Crema + Tess + Savee Speech Emotion Datasets on Kaggle
7. Audiomentations Documentation