Predicting Emotional OutBERSt

CMPT 419 Final Project
David Xiong, Danny Nguyen, Tegnoor Gill

OUR TASK

- **Offline Recognition**
  - Create offline social signal recognition machine learning models that will analyze and predict the basic emotion from voice recordings
  - Train models on 3000+ audio files from the BERSt dataset and label test audio
  - Use neural networks, which are a class of machine learning models that uses multiple layers to transform input data into meaningful output
- **Data Handling**
  - We are to label a set of 100 emotion speech audio files with perceived emotions to test against model

OUR METHODS

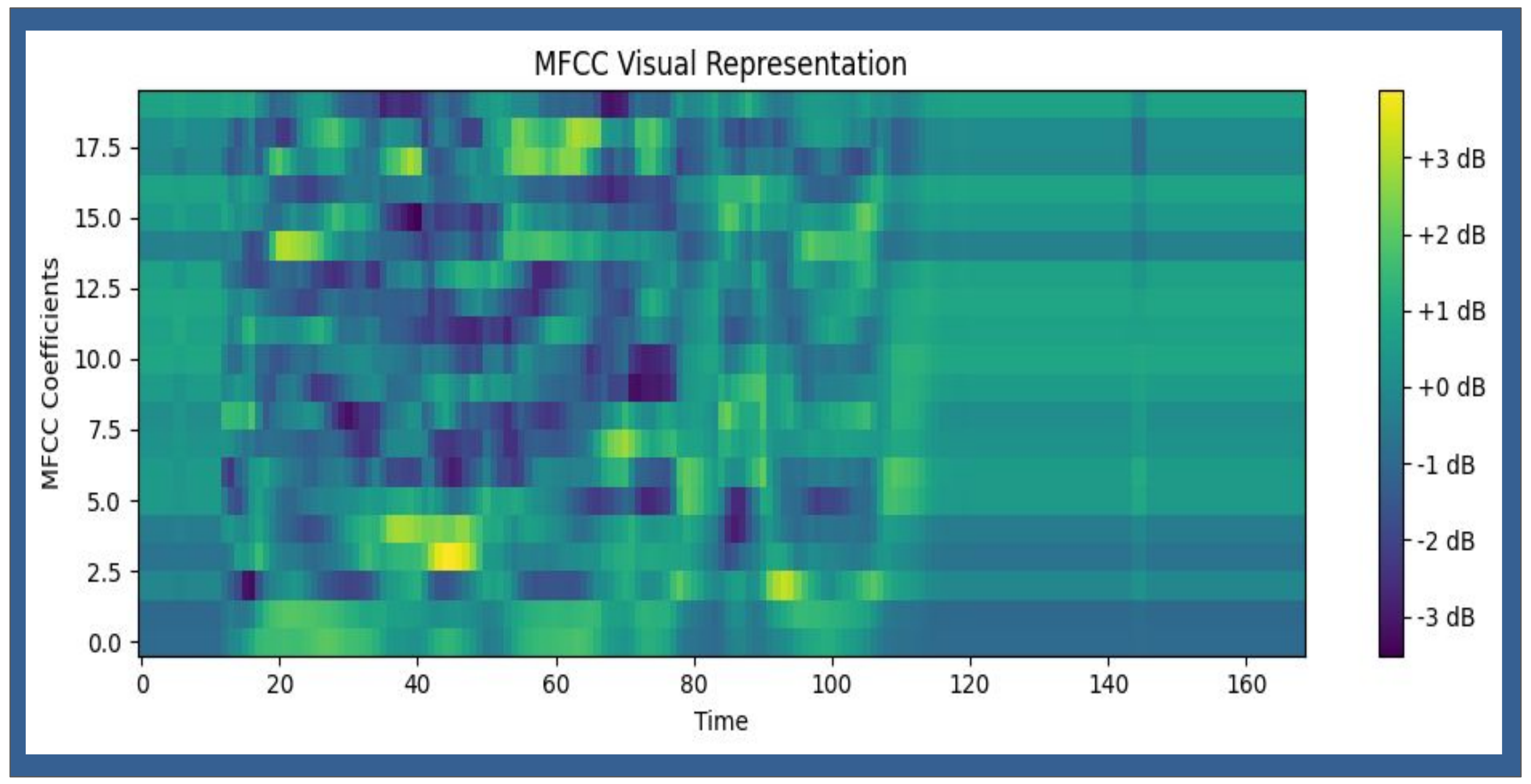
- **Mel-Frequency Cepstrum Coefficients**
  - MFCC is a feature extraction technique commonly used in audio and speech processing as a compact representation of sound
  - Derived from the Mel-frequency scale that maps the frequency range of human hearing into space
  - Used as primary training feature to represent the audio data for training the models
- **Convolution Neural Network**
  - Known for its ability to extract local features from input data, we use conv2d layers to apply convolution operations on the input data with multiple filters
  - Allow model to learn to detect more complex patterns
- **Recurrent Neural Network**
  - Recognize patterns across time and can apply previous inputs to help predict the outcome
  - Uses LSTM(Long-Short Term Memory), capable of learning long-term dependencies in data, useful for sequence prediction problem
- **Convolutional Recurrent NN**
  - Combines the strength of both CNN and RNN model, aiming to improve accuracy

REFERENCES

- BERSt Paper
- BERSt Data Collection Powerpoint
- CNN for Audio Classification
- RNN for Audio Classification
- CRNN for Recognition

BACKGROUND

- **BERSt Project**
  - People express intense emotions in different ways, and someone could scream out of excitement, anger, frustration, or surprise.
  - The goal is to determine distress from speech on smartphones with varying distances for a more 'naturalistic' setting



DISCUSSION

- CNN out performed the other two by a good margin, which could be a result of rearranging nodes in increasing order with different drop out rates improving the overall test accuracy
- Increasing the frequency of layer flattening after convolution and recurrence decreased computational times significantly at a marginal cost on accuracy
- Labeled "perceived emotion" only matched 42% of matching prompt affect
- Discrepancies in "perceived emotions" labeling of test data - our group had differing perception of the emotions being displayed
- Despite the difference in perceived emotions from the audio chunks, our perceived valence and arousal scores were, for the most part, very similar
- High validation loss and low validation accuracy on models indicate potential overfitting of the model

DATA SET

- **Audio Files**
  - 3000+ audio chunk files provided by BERSt team
  - Each chunk features a spoken script that is 1-3 seconds long and has an assigned affect recorded by participants
- **Training Data**
  - A CSV file containing information about the feature labels of each audio chunk that gives context to the sounds
  - Each audio recorded has the participant's age range, gender, phone model, affect, script, phone position, language
- **Labeled Test Data**
  - A Subset of 100 audio files randomly selected from the training dataset to be used for testing
  - Introduced a new feature of 'perceived emotion' labeled by us with using inter-rater agreement
  - Valence and arousal also labeled with inter-rater agreement

RESULTS