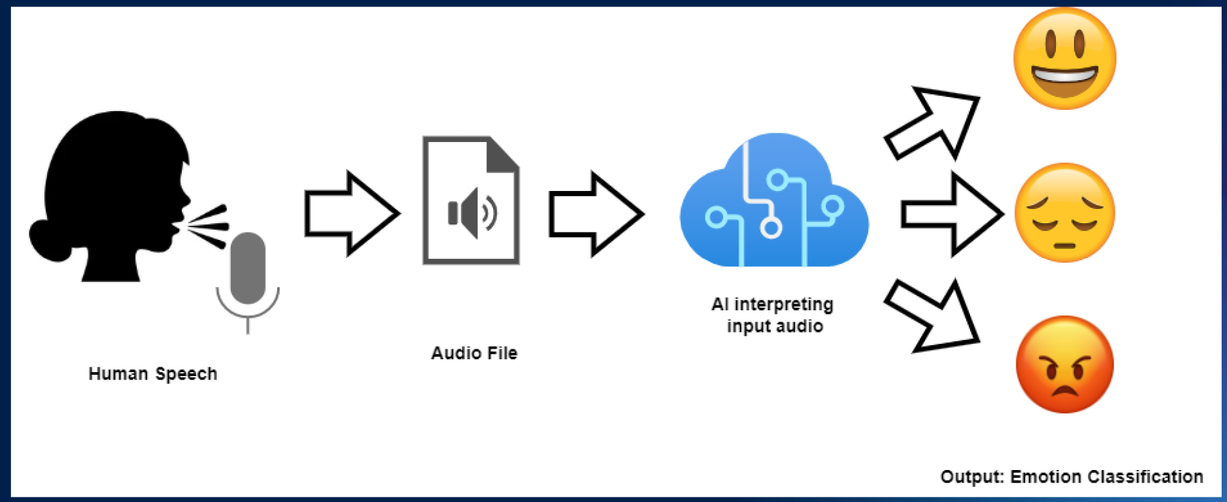
Voice Sculptor

Unlocking the hidden dimensions of Speaker Identity

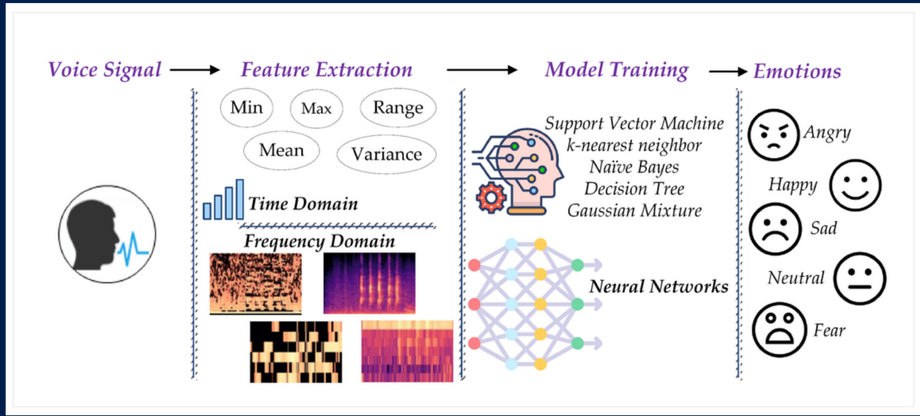
Introduction

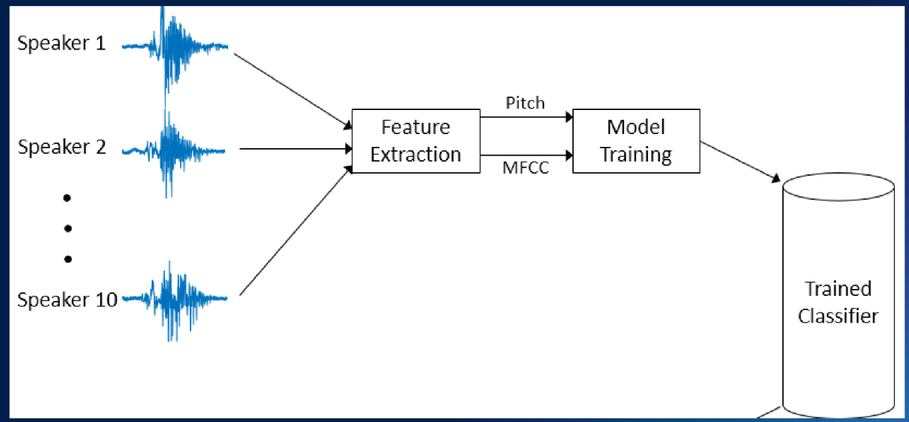
- The human voice is a unique and powerful tool of expression.
- It carries not only the identity but also their emotion, creating a rich tapestry of communication.
- The ability to detect both the speaker's identity and their emotions from their voice has become a critical technology.



Objective

- **Accurate Speaker Identification:** Develop robust model that can identify the speaker with a high degree of accuracy, contributing to enhanced security and personalization.
- **Precise Emotion Recognition:** Create another model which is capable of recognizing the emotion from speaker's voice.

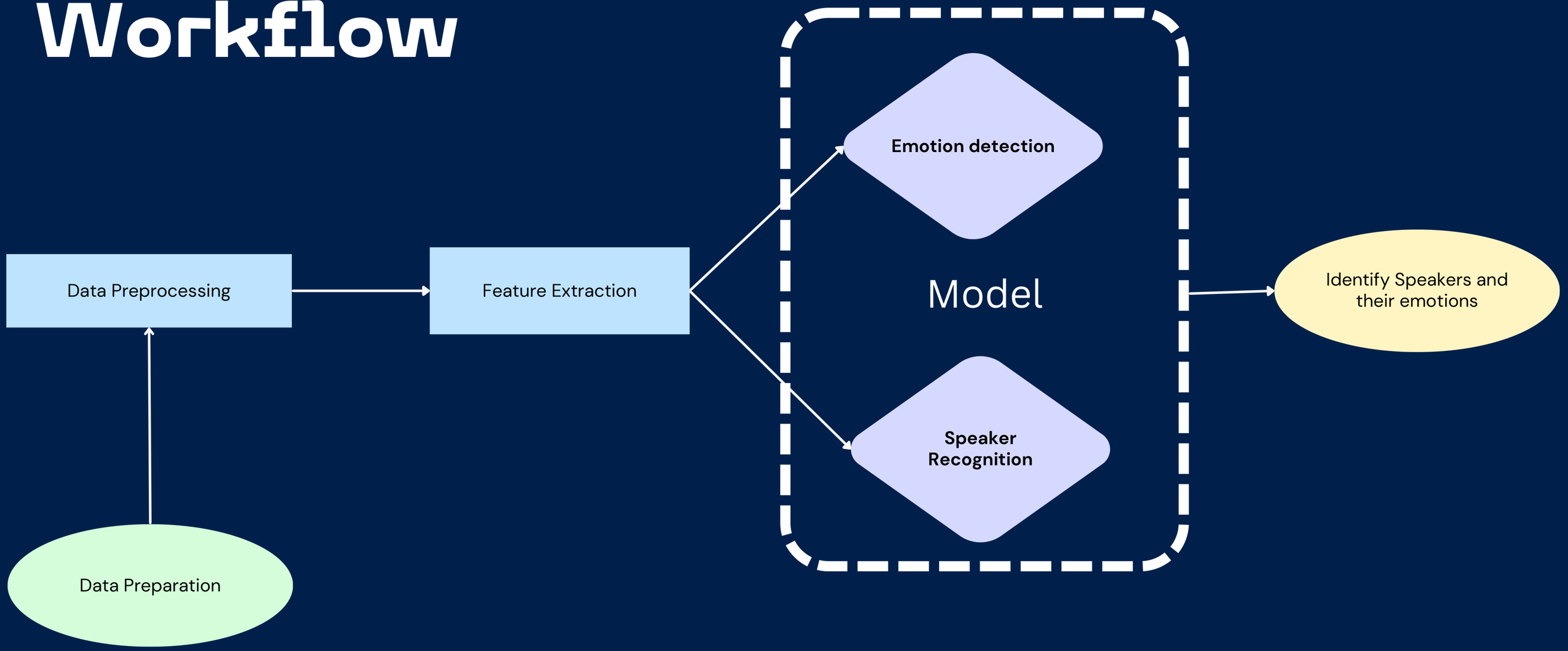




Datasets

- **Toronto Emotional Speech Set(TESS):** Consists of 2800 audio files with 7 emotions spoken by 2 speakers. Each emotion contain 400 audio files.
- **Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) Dataset:** Consists of 1440 audio files with 24 speakers. Each speaker contains 60 audio files.

Workflow


Data Preprocessing
Feature Extraction
Emotion detection
Model
Speaker Recognition
Identify Speakers and their emotions
Data Preparation

Feature Extraction

The following features have been extracted in each audio file:

- **Mel-Frequency Cepstral Coefficients(MFCCs):** Capture the spectral(timbral and pitch-related) characteristics of the audio signal.
- **Chroma Feature:** Represents the pitch content of an audio signal. They are useful for capturing the harmonic content and chord progressions in music.
- **Spectral Contrast:** Describes the difference in amplitude between peaks and valleys in the spectrum of an audio signal. Measures the variation in loudness.

Models

Emotion Detection

- Normalization- StandardScalar()
- Shape:
  - (2880, 32, 129) -TESS dataset
  - (1440, 32, 129) -RAVDESS dataset
- Model:
  - LSTM -95.36%
  - CNN - 96.82%

- RAVDESS data performed Poorly for emotion detection -- 55%

Speaker Recognition

- Normalization- StandardScalar()
- Shape: (1440, 495, 13) - RAVDESS dataset
- Model:
  - LSTM -48.56%
  - CNN - 76.04%
- Hyperparameter Tuning - Keras_tuner()

Results – Accuracy table

Dataset	CNN		LSTM	
	Val Acc	Test Acc		
TESS Emotion Detection	98.64%	96.82%	96.42%	95.36%
RAVDESS Emotion Detection	62.45%	55.01%	38.98%	39.24%
RAVDESS Speaker Recognition	82.78	76.04%	55.23%	48.56%

Conclusion

- Achieved better predictions for a range of emotions and successfully identified different speakers.
- Acknowledge the need for continuous improvement in model performance.
- Plan to integrate emotion detection and speaker recognition into unified system.

Thank You

K Muni Sai – 20bds028
Peddisetty Venkata Sai Pranay – 20bds038