High-Quality Speech Understanding Using Feature Engineering with LSTM and Wav2Vec

Dhruv Kumar Patel,Arti Gupta,Kshitij Kumar Soni

Problem Statement and Data Processing

Accurate understanding of speech audio is a fundamental problem in modern AI systems such as virtual assistants, transcription tools, and emotion-aware applications. While recent Automatic Speech Recognition (ASR) models provide strong transcription performance, they often underutilize important acoustic properties like pitch, volume, and spectral variations.

Dataset

- Audio datasets: Emotion and speech datasets (e.g., RAVDESS / CREMA-D)
- Format: WAV, mono channel
- Sampling rate: 16 kHz
- Duration: Short utterances (1–5 seconds)

Preprocessing Steps Implemented

- Audio loading and resampling
- Silence trimming
- Amplitude normalization
- Conversion to log-Mel spectrograms (for Whisper)
- Raw waveform preparation (for Wav2Vec)

This ensures consistent and model-ready audio input.

Methodology

We compare and integrate two modern speech models: Whisper, LSTM and Wav2Vec 2.0, each serving a distinct role.

Step 1: Data Preprocessing

- Convert all audio to mono WAV format
- Resample to 16 kHz
- Normalize amplitude and remove silence

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

Step 2: Feature Engineering

We extract acoustic features including:

- Pitch (Fundamental Frequency, F_0)
- Energy (signal power)
- MFCC (Mel Frequency Cepstral Coefficients)
- Spectral features (centroid, bandwidth)

Step 3: Model Processing

- Whisper processes log-Mel spectrograms for transcription
- Wav2Vec learns latent speech representations directly from raw audio

Step 4: Evaluation

- Outputs are evaluated using transcription accuracy and Word Error Rate (WER).

Pitch (Fundamental Frequency)

$$F_0 = \frac{1}{T}$$

Where:

- F0 = pitch (Hz)
- T = time period of vocal fold vibration

Mel Frequency Mapping

$$\mathrm{Mel}(f) = 2595 \cdot \log_{10}\left(1+\frac{f}{700}\right)$$

Where:

- f= frequency in Hz

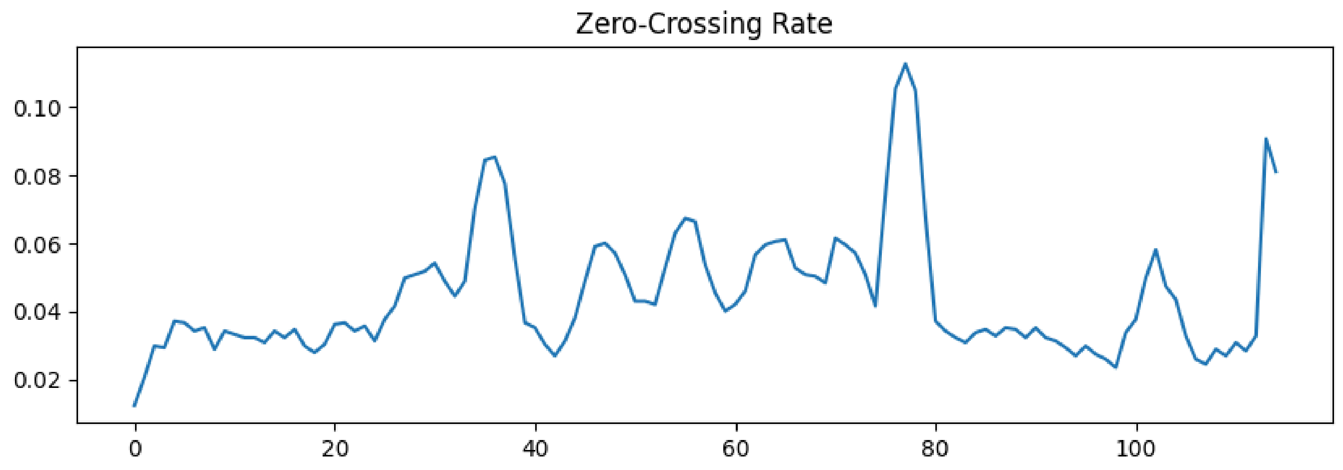
MFCC Pipeline

1. Short-Time Fourier Transform (STFT)
2. Mel Filterbank
3. Log Energy
4. Discrete Cosine Transform (DCT)

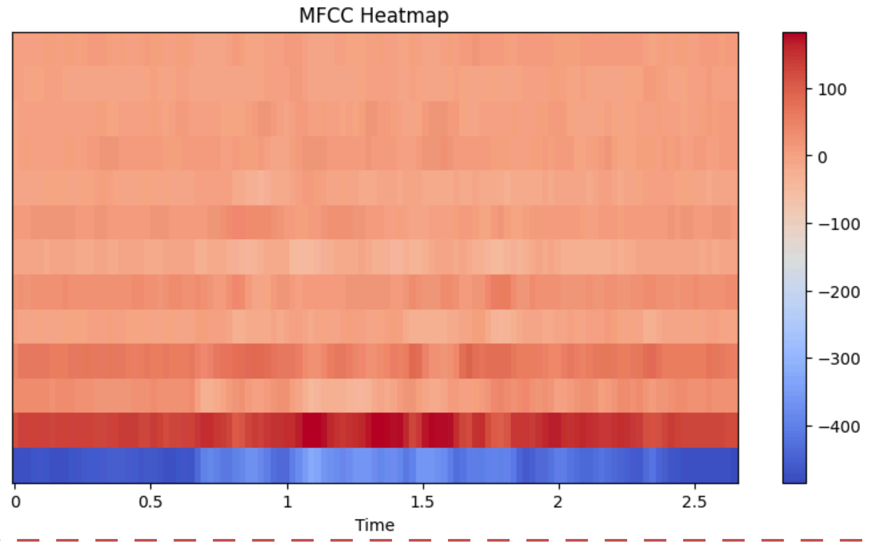
As both Whisper and Wav2Vec models were implemented, the dataset was divided into training and testing subsets. Each model was trained for multiple epochs, where one epoch represents a complete pass through the training audio data. After training across the defined number of epochs, random unseen audio samples were selected as test data. The models were then evaluated by generating transcription outputs and visualizations such as spectrograms and MFCCs. Performance across epochs was analyzed using Word Error Rate (WER) to compare transcription accuracy and model convergence.

Feature Engineering Results

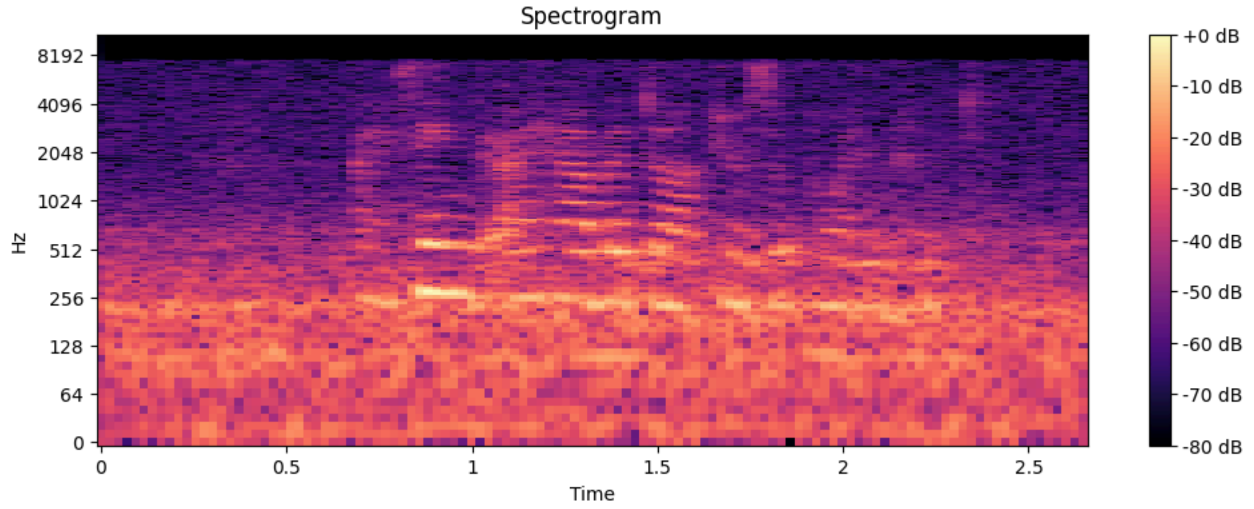
1



2

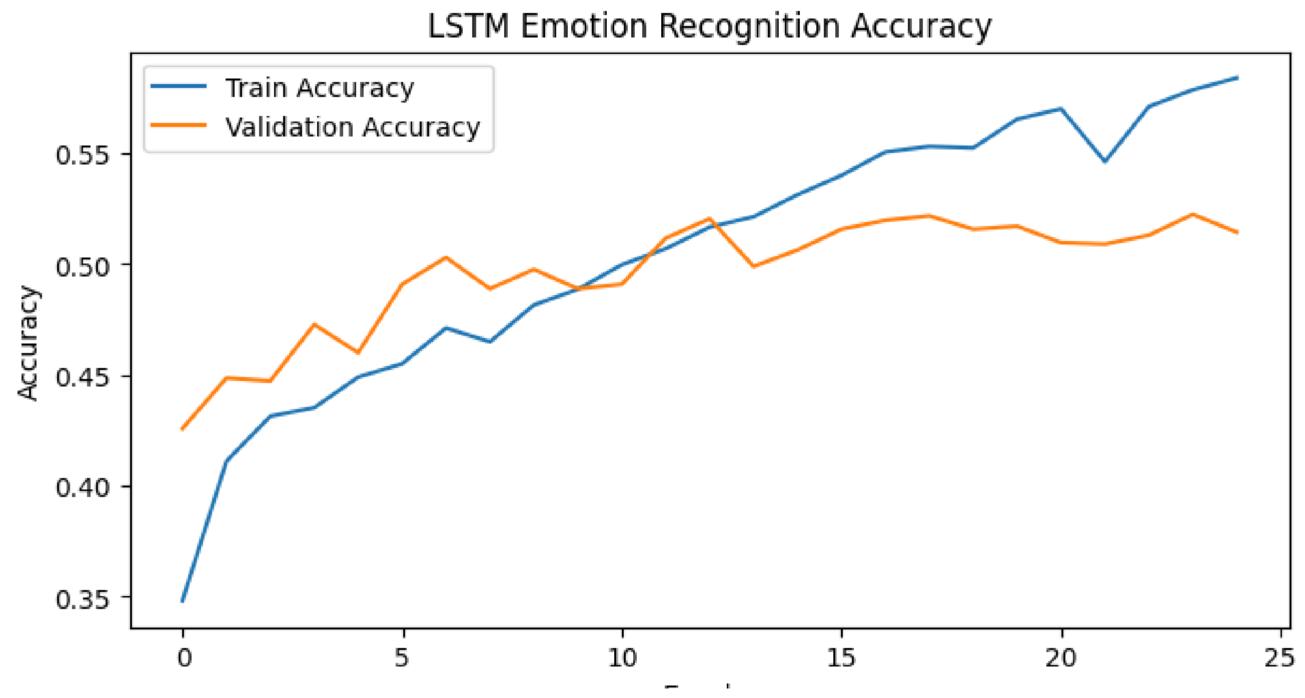


3

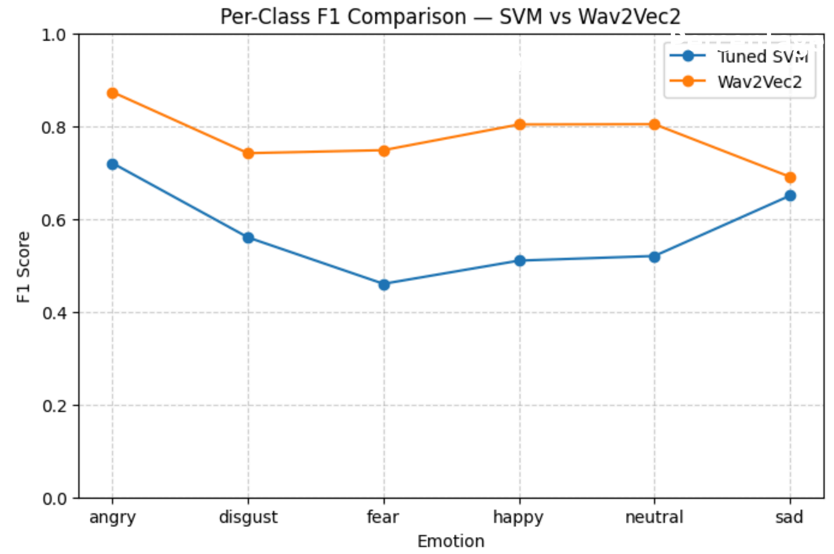


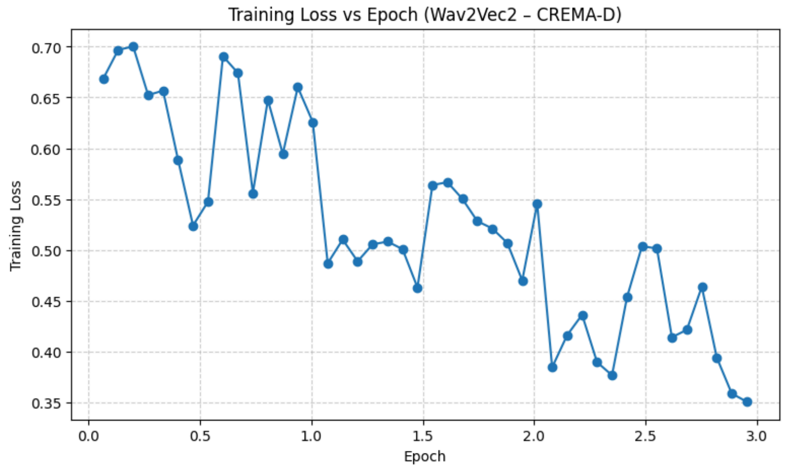
Result And Analysis

1



Scenarios and Comparison





Future Works And Conclusion

FUTURE WORK

- Emotion recognition from speech
- Deepfake audio detection
- Multilingual and low-resource language support
- Real-time speech analysis for wearable devices
- Integration with health and mental-state monitoring systems

CONCLUSION

This project demonstrates that combining feature engineering with modern speech models significantly improves speech understanding quality. Whisper and Wav2Vec together provide a powerful framework for accurate, robust, and interpretable audio analysis.

References

[1] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, MD, USA, 2022, pp. 28492–28518.
doi: 10.48550/arXiv.2212.04356

(Whisper – core paper, mandatory reference)

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 12449–12460, 2020.
doi: 10.48550/arXiv.2006.11477

(Wav2Vec 2.0 – representation learning)

Acknowledgements

Dr. Rahul Dubey