Affexion

Gurkirat Singh (301566100), Karanbir Singh (301566213), Leo Tsai (301391409)

Contact info: gsa136@sfu.com, ksa205@sfu.ca, lta67@sfu.ca

Background

AI-generated speech is increasingly realistic, but can it train social signals recognition models that generalize to real-world data? This project explores that question using synthetic and human-recorded audio.

Goal

Testing if the audio data generated by AI(ChatGPT) is good enough to train models and test real-world audio data. Social Signals: **Uncertainty, Boredom, Panic, Excitement.**

Dataset

OpenAI generated Data Using the model gpt-4o-audio-preview (Samples: 70 x 4). Validation data using YouTube (Samples: 13 x 4).

Method

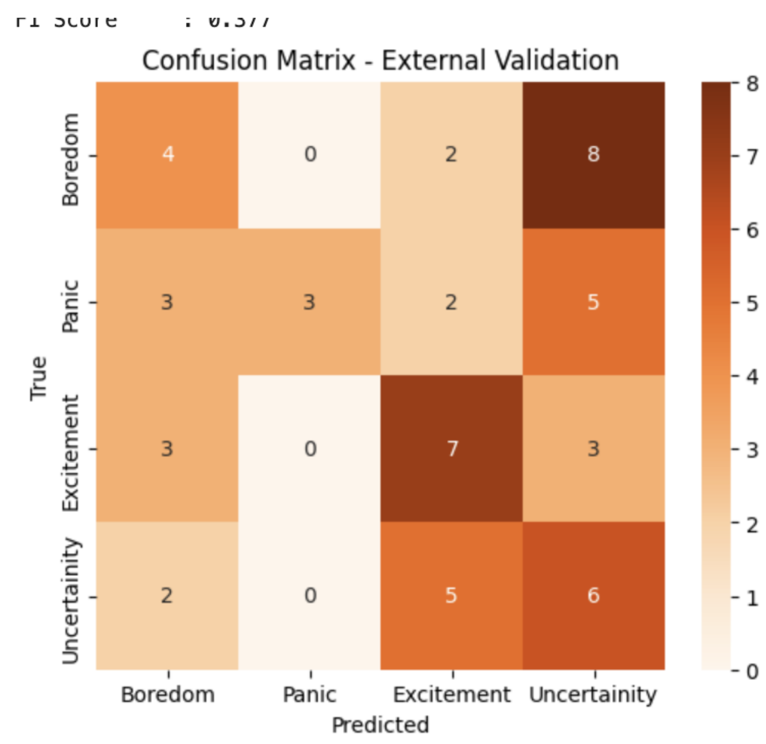
Extracted MFCCs + Delta + Delta^2 features from audio (shape: 200 × 39 per clip).

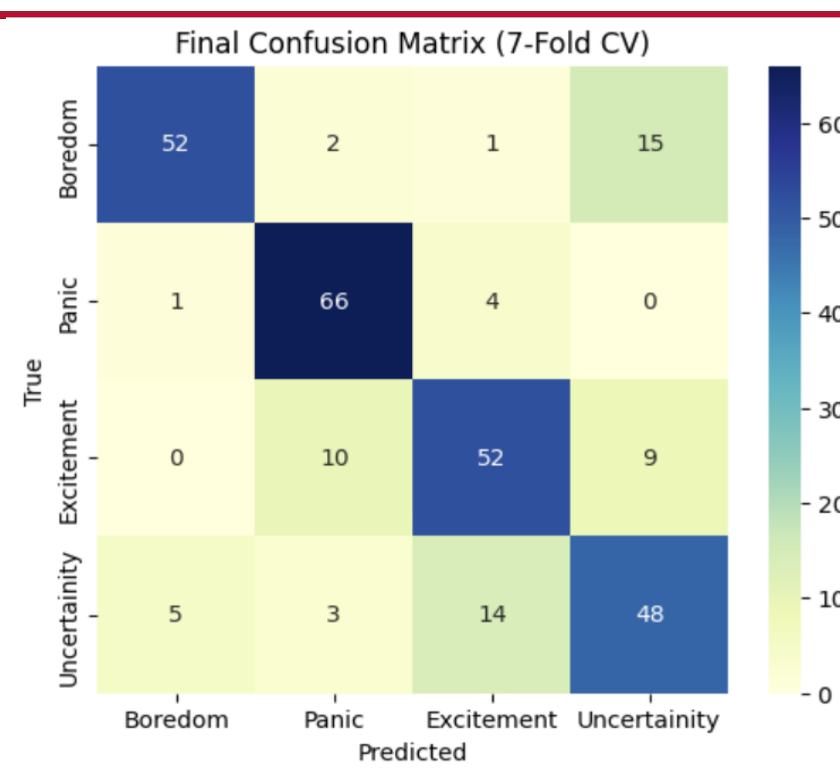
Applied Z-score normalization and padded/truncated all inputs.

Designed a deep model inspired by wav2vec 2.0:

- We applied three parallel 1D convolution layers with kernel sizes 3, 5, and 7 to capture short, medium, and long-range speech patterns
- Bi-Directional LSTM (2 layers, 128 hidden units) for sequential modeling. This helps the model to understand how earlier and later frames influence emotion, improving temporal context awareness.
- A self-attention mechanism learns which time steps in the audio carry the most emotionally relevant information and amplifies them, while less important frames are down-weighted.
- Final Dense Layers classify into 4 emotions

Trained with Adam optimizer, 10 epochs, and learning rate scheduler.
Evaluated using:
External validation set (real human voices – gathered from YouTube). 7-fold cross-validation to assess generalization.





Results

- Model trained on **AI-generated audio** (ChatGPT + OpenAI Audio-4o) performed well on synthetic data (7-fold CV).
- When evaluated on **real-world YouTube clips**, performance dropped significantly:
- **Accuracy**: 37.7%
- **Precision**: 51.1%
- **Recall**: 37.9%
- **F1 Score**: 37.7%

Learning: Models trained on generated speech **do not generalize well to natural human speech**.

- Variations in tone, background noise, and expression highlight the need for diverse real-world data in social signal recognition.

Challenges and Lessons Learned

Collecting accurate YouTube data for specific social signals like Excitement, Panic, Uncertainty and Boredom was a challenging task.

We also learned that length of the audio samples collected influenced the learning. Longer audio samples affected the accuracy of the model.

Future Work

Next, we plan to gather larger audio samples for each social signal to improve the model's accuracy. Given more time, we would divide the real-world data into 7–8 distinct groups based on these signals. Each group would serve as a separate validation set. By evaluating the model on each set, we could identify whether its performance issues are general or specific, helping us better understand and refine its capabilities.

References

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (arXiv:2006.11477). arXiv. https://doi.org/10.48550/arXiv.2006.11477

Efficient Speech Emotion Recognition using Multi-scale CNN and Attention. (n.d.). Ar5iv. Retrieved April 4, 2025, from https://ar5iv.labs.arxiv.org/html/2106.04133

Marafioti, A. (2023, November 12). *Answer to "How do I implement this attention layer in PyTorch?"* [Online post]. Stack Overflow. https://stackoverflow.com/a/77469335

Chen, S., Zhang, M., Yang, X., Zhao, Z., Zou, T., & Sun, X. (2021). The Impact of Attention Mechanisms on Speech Emotion Recognition. *Sensors (Basel, Switzerland)*, *21*(22), 7530. https://doi.org/10.3390/s21227530