Affexion: Social Signal Processing from Speech Using Deep Learning on Synthetic and Real-World Data

Gurkirat Singh
SFU
Burnaby, Canada
gsa136@sfu.ca

Karanbir Singh
SFU
Burnaby, Canada
ksa205@sfu.ca

Leo Tsai
SFU
Burnaby, Canada
lta65@sfu.ca

ABSTRACT

Speech-based social signal recognition plays a vital role in human-computer interaction, but building effective models requires large amounts of high-quality emotional speech data, which are costly and difficult to obtain. A common workaround is to use synthetic speech generated by AI models, yet it remains unclear whether such data can substitute real-world recordings in training robust emotion classifiers. This raises the core question: Can models trained solely on AI-generated speech generalize to natural emotional speech in real-world settings? In this paper, we investigate this question by training a deep learning model by combining CNNs, BiLSTMs, and self-attention layers on synthetic emotional speech generated using OpenAI's GPT-4o-audio and evaluating it on real-world YouTube audio clips by classifying four emotional states, *Boredom*, *Panic*, *Excitement*, and *Uncertainty*. Our results show that while the model achieves high accuracy on synthetic data, its performance drops significantly when tested on real speech, highlighting a gap in generalizability. This work contributes an analysis of the limitations of synthetic data in social signal recognition and emphasizes the continued need for diverse real-world datasets.

1 INTRODUCTION

Communication involves not just the content of speech but also the vocal cues that convey emotion. The increasing use of chatbots and virtual assistants requires the ability to interpret not only the semantic content but also the affective cues in speech. Our work is motivated by the growing ability of AI to generate synthetic emotional speech using large language models with voice synthesis. While this synthetic data is high-quality and emotion-rich, it remains uncertain whether it can effectively train models that generalize to real-world speech. To explore this, we trained our model entirely on AI-generated audio and validated it against real-world samples from YouTube. This setup allowed us to evaluate the domain gap and assess how well deep learning architectures—specifically our multi-scale CNN, BiLSTM, and self-attention framework—can bridge that divide and extract emotionally important features.

In the literature, techniques based on Mel-Frequency Cepstral Coefficients (MFCCs) have been popular; however, modern approaches benefit from learning representations directly from audio signals [2]. Recent studies, such as those on wav2vec 2.0 [2] and efficient speech social signal processing using multi-scale CNN and attention [1], have motivated our design choices. We build upon these works by incorporating insights from attention mechanisms as detailed in [3, 4].Our work stands out by creating a model that combines multi-scale CNNs, BiLSTM, and self-attention based on recent research, and by carefully selecting both training and validation data. We systematically explored the impact of tuning essential hyperparameters, such as the maximum length of MFCC sequences, dropout rates, MFCC feature dimensionality, and the number of convolutional output channels to significantly enhance the model's recognition accuracy and robustness.

In our approach, the dataset is composed of both synthetically generated audio using OpenAI's Audio model and audio clips extracted from YouTube. The synthetic data, designed to model social signals like *Boredom*, *Panic*, *Excitement*, and *Uncertainty*, is generated with AI, which produces high-quality data comparable to lab-controlled data, whereas the real-world dataset naturally presents additional variability such as background noise and speaker differences. We employed a 7-fold cross-validation scheme on the synthetic dataset and validated on the real-world data.

2 APPROACH

2.1 Overview and Model Architecture

Our model is designed to extract detailed temporal features from speech and to emphasize relevant information via self-attention. The architecture consists of the following principal components:

(1) **Multi-Scale CNN Block:** This block applies three parallel one-dimensional convolutional layers with kernel sizes of 3, 5, and 7. Each convolution captures features at different time resolutions, allowing the extraction of short-term, medium-term, and long-term acoustic patterns [1]. The outputs of these layers are concatenated and normalized via batch normalization, then passed through a dropout layer (20%) for regularization.
(2) **Bi-Directional LSTM:** The concatenated features from the multi-scale CNN block are fed into a two-layer bidirectional LSTM. The BiLSTM enables the model to learn temporal dependencies in both forward and backward directions, thereby capturing the influence of past and future frames on the current state.
(3) **Self-Attention Mechanism:** After the BiLSTM, a self-attention layer is applied to dynamically reweight the temporal features. This layer learns a set of attention weights that highlight time steps with higher emotional relevance while down-weighting less informative frames. The self-attention formulation follows the query–key–value framework, where linear transformations are applied to compute queries, keys, and values. The mechanism is inspired by both the work in [3] and insights from [4], ensuring proper alignment of emotional features.

(4) **Fully Connected Layers:** Finally, the pooled output from the self-attention module is processed by two dense layers. The first dense layer uses a ReLU activation and dropout (30%) for further feature refinement, while the second layer maps the features to the four target emotion classes.

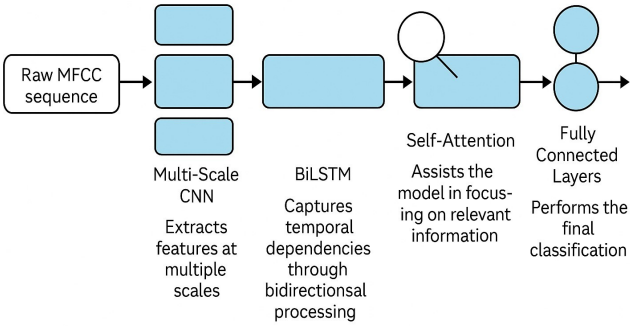

Figure 1: Schematic of the proposed model architecture showing the multi-scale CNN block, BiLSTM, self-attention mechanism, and fully connected layers.

2.2 Feature Extraction and Pre-Processing

The model begins with audio pre-processing using Librosa to extract 25 MFCC features per frame along with their first and second derivatives, resulting in an aggregated feature vector of 75 (25 MFCC + 25 for first derivative + 25 for second derivative) dimensions per frame. Each audio clip is resampled at 16 kHz, and sequences are either padded or truncated to a maximum length (MAX_LEN) of 1,000 time steps (about 10 seconds of audio) to standardize input dimensions. Z-score normalization is applied channel-wise to mitigate variability and ensure numerical stability.

The primary hyperparameters that were tuned include:

- **MAX_LEN:** Initially set to 200, then increased to 1200, with the best performance observed at 1000.
- **Dropout Rate:** Experimentation with dropout rates ranging from 10% to 40% determined that a dropout rate of 30% offered the best trade-off between underfitting and overfitting.
- **MFCC Features:** Variations from 7 to 30 MFCC coefficients were tested; 25 was selected to balance model complexity and overfitting.
- **CNN Output Channels and LSTM Hidden Size:** The number of output channels in the convolutional layers was varied between 64 and 128; 64 channels provided optimum performance while a hidden dimension of 128 in the LSTM layers ensured adequate sequential representation.

2.3 Model Training and Optimization

The training process is divided into two main phases: an initial training phase and a cross-validation phase. In the training phase, the model is optimized using the Adam optimizer with a learning rate of 0.001. A learning rate scheduler reduces the learning rate by a factor of 0.5 every 10 epochs. The loss function used is cross-entropy loss, and performance is evaluated using accuracy, precision, recall, and F1 score.

In the cross-validation phase, a 7-fold cross-validation scheme is adopted. The full synthetic dataset is partitioned into 7 equal parts, with each fold being used in turn as a validation set and the remaining folds as training data. This method provides robust estimates of model generalization capability. In each fold, the model is trained for 20 epochs. The cross-validation results indicate an average accuracy of 75.1%, mean precision of 77.9%, mean recall of 76.0%, and mean F1 score of 74.4%.

The model's training loop involves the following steps:

(1) Data loading from the directory structure where each folder corresponds to an emotion.
(2) Extraction and normalization of MFCC features.
(3) Forward propagation through the multi-scale CNN, followed by the BiLSTM and self-attention layers.
(4) Loss computation via cross-entropy.
(5) Back-propagation and weight updates.
(6) Scheduler step and epoch-wise logging of training loss and accuracy.

3 DATASET

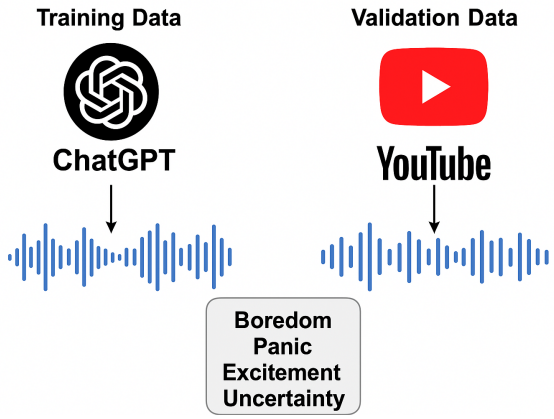

Figure 2: Training overview using synthetic audio generated with ChatGPT to capture Boredom, Panic, Excitement, and Uncertainty.

The dataset for our experiments comes from two sources. The training data consists of 280 synthetic audio clips (around 70 per emotion), which we generated by first using ChatGPT-4o to create sentences tailored to express specific emotions—Boredom, Panic, Excitement, and Uncertainty. These sentences were then input into OpenAI's Audio-4o-preview model to produce realistic emotion-based audio samples. The validation data include 52 real-world samples (about 13 per emotion) collected manually from YouTube. This dual-source approach is designed to evaluate the generalizability of the model.

Synthetic audio was pre-processed using MFCC features. The real-world validation set from YouTube presented greater variability and noise, offering a more challenging test. Due to cultural

differences, some labeling bias may have occurred, but team members cross-checked all samples individually and removed unclear ones, finalizing 13 reliable samples per emotion.

More about dataset collection and dataset in Appendix.

4 EXPERIMENTS AND RESULTS

4.1 Experiment

The central hypothesis of our project was that a model trained exclusively on AI-generated emotional speech samples could generalize effectively to real-world speech and accurately classify natural expressions of social emotions. Specifically, we aimed to test whether the synthetic training data produced using consistent prompts via ChatGPT's voice synthesis contained enough emotional variability and realism to support social signal processing in human speech.

4.2 Training Results

The training loop was executed for 15 epochs for the initial model. At epoch 15, the model reached a training loss of 3.7168 and a training accuracy of 98.3%. These promising training metrics, however, were contrasted by a lower validation performance on the external real-world set, where the accuracy was 45.3%, precision 49.4%, recall 45.5%, and F1 score 43.6%. The discrepancy between training and validation results highlights the difficulty of generalizing from synthetic data to natural human speech.

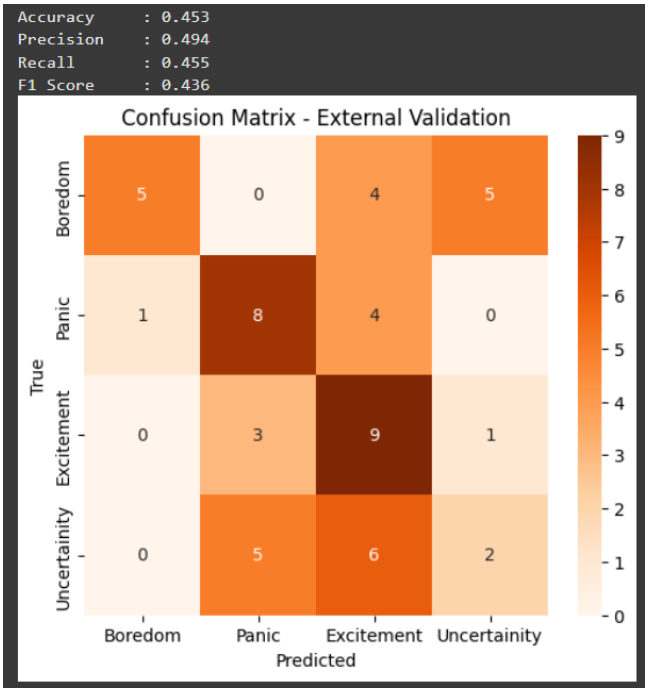

Figure 3: Training performance metrics over 15 epochs illustrating training loss and accuracy.

In our cross-validation experiments using the 7-fold method, the fold-wise accuracies ranged from 67.5% to 87.8%, with average scores of 75.1% accuracy, 77.9% precision, 76.0% recall, and 74.4% F1 score. Each fold's metrics were recorded and analyzed to understand the variance introduced by different data splits.

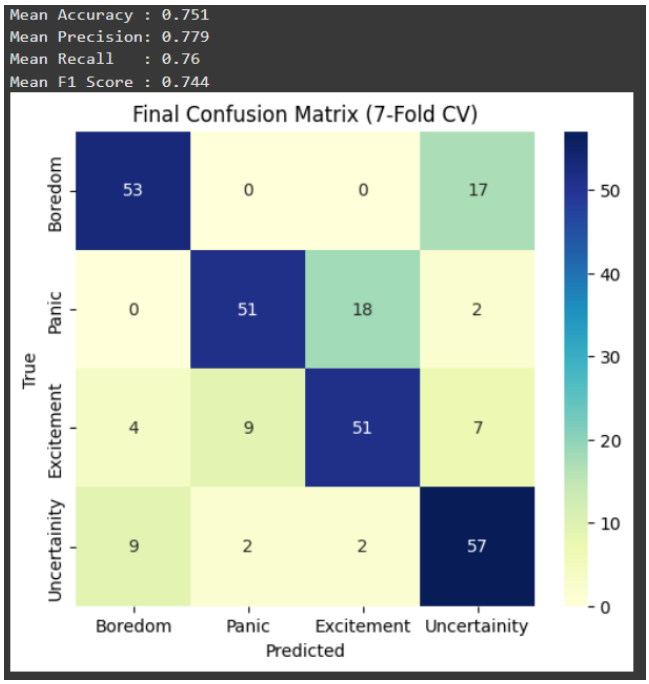

Figure 4: Cross-validation results showcasing the performance metrics across 7-fold splits.

Our results reveal a notable gap between training and validation performance. While the model demonstrated strong learning capabilities on the synthetic dataset, its accuracy dropped when evaluated on real-world speech. This contrast highlights the limitations of synthetic data in capturing the full complexity and variability of human emotional expression. However, the model retained a meaningful predictive ability, suggesting that AI-generated data is a promising starting point, but not a complete substitute for natural emotional speech in training affective computing systems.

4.3 Hyperparameter Tuning

A series of experiments were conducted to fine-tune the hyperparameters. Specifically, we adjusted:

- **Number of Epochs:** We initially experimented with training the model for 7, 15, and 20 epochs. At 7 epochs, the model underfit the data and did not converge sufficiently. While 20 epochs yielded higher training accuracy, validation accuracy decreased to 28%, and signs of overfitting began to emerge. The most consistent and optimal performance, for validation sets, was observed at 15 epochs, which we selected as the default for final training.
- **Sequence Length (MAX_LEN):** We experimented with frame lengths from 200 to 1200 and found that 1000 frames (10 seconds) gave the best validation performance. Shorter lengths (e.g., 300) often missed key emotional cues in longer clips, resulting in poor validation (23%) and training (66%)

accuracy. Increasing MAX_LEN beyond 1000 led to excessive padding for shorter clips, which introduced noise and did not improve results.

- **Dropout Percentage:** Reducing dropout below 30% led to overfitting, while increasing beyond 30% caused the network to underfit.
- **MFCC Dimensionality:** The model was tested with MFCC feature dimensions at 7, 13, 25, 30 and at other different points. Lower dimensions like 7 and 13 resulted in validation accuracies of 23.5% and 31.4%, respectively, indicating underfitting. Increasing to 25 coefficients significantly improved performance, achieving the highest accuracy. However, further increasing to 30 led to a drop in accuracy (28.33%), likely due to overfitting or the inclusion of redundant information. Thus, 25 MFCCs provided the best balance to capture emotional nuances.
- **CNN and LSTM Dimensions:** Adjusting the output channels of the CNN layers and the hidden size of the LSTM layers showed that 64 channels and a hidden size of 128 achieved superior performance compared to other configurations.

Each of these adjustments affected the model by either enhancing its ability to capture local features (in the case of CNN adjustments) or improving the temporal modeling (with the BiLSTM and self-attention layers). The final configuration represents the best compromise between learning capacity and generalizability.

5 DISCUSSION

The experiments demonstrate a clear trade-off between training accuracy and real-world performance. Although the model was able to learn the features from synthetic data quite effectively, as evidenced by the near-perfect training accuracy, the drop in validation performance indicates issues with generalizability. This discrepancy is likely due to the differences between the controlled, AI-generated audio and the more variable, natural human speech found in YouTube clips. Additionally, some labeling bias may have been introduced in the validation set due to cultural differences in how emotions are expressed and interpreted. Although we minimized this by cross-validating and filtering unclear samples, subtle misalignments may still have affected the model's performance.

The multi-scale CNN component effectively captured local spectro-temporal features, as each parallel convolution layer was able to focus on different resolution levels. The subsequent BiLSTM component modeled the temporal dependencies well, yet the introduction of the self-attention mechanism was the critical factor in enabling the model to reweight the features and focus on the most emotionally relevant sections of the audio. In our implementation, the self-attention layer, inspired by [3] and refined using insights from [4], assigned weights that helped mitigate the effects of irrelevant background information.

The hyperparameter tuning process also revealed several important findings. For instance, setting the maximum sequence length to 1000 frames allowed the model to work with consistent input sizes even when audio durations varied, ensuring that key emotional information was neither truncated nor excessively padded. Similarly, the choice of a 30% dropout rate struck a balance between overfitting and information loss. By experimenting with the number of MFCC features, we found that using 25 features provided sufficient discriminative power without overwhelming the network with extraneous data.

Another significant observation is related to the training methodology. The use of 7-fold cross-validation provided a more robust estimate of the model's performance, as variability across folds highlighted potential inconsistencies in the synthetic data distribution. These results underscore the necessity of supplementing synthetic data with real-world examples to build more generalizable models.

Limitations of this study include the relatively small size of the synthetic training dataset and the inherent variability in real-world audio quality. Moreover, although the self-attention mechanism improved performance, its placement and configuration could be further explored. Future work should focus on expanding the dataset to include a broader range of voices and environmental conditions, as well as adding noise to AI-generated data using online tools. Experimenting with alternative attention formulations and deeper network architectures is another area of future improvement.

6 CONCLUSION

In conclusion, our study demonstrates that a deep neural network architecture combining multi-scale CNNs, BiLSTM, and self-attention can effectively capture the nuanced emotional states from speech. The proposed model achieves high training accuracy on synthetic data and, despite some performance degradation on real-world samples, offers valuable insights into feature extraction and temporal modeling for speech social signal processing. The comprehensive hyperparameter analysis reveals that parameter choices such as sequence length, dropout rate, and MFCC dimensionality significantly impact performance. Although challenges remain in generalizing from synthetic to natural audio, the findings indicate that integrating these advanced network components leads to better emphasis on emotionally relevant segments. Future work includes expanding the dataset, adding noise to synthetic data, refining attention, and exploring deeper architectures.

REFERENCES

[1] [n. d.]. Efficient Speech Social Signal Processing using Multi-scale CNN and Attention. https://ar5iv.labs.arxiv.org/html/2106.04133. Apr. 2025.
[2] Alex Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv preprint arXiv:2006.11477* (2020). https://doi.org/10.48550/arXiv.2006.11477
[3] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun. 2021. The Impact of Attention Mechanisms on Speech Social Signal Processing. *Sensors* 21, 22 (2021), 7530. https://doi.org/10.3390/s21227530
[4] A. Marafioti. [n. d.]. Answer to 'How do I implement this attention layer in PyTorch?'. https://stackoverflow.com/a/77469335. Nov. 2023.

APPENDIX: DATASHEET FOR THE AFFEXION DATASET

A.1 Motivation

Purpose: The dataset was created to evaluate whether AI-generated speech can effectively train deep learning models for recognizing social signal states in speech. The task focused on classifying four specific social signals: *Boredom*, *Uncertainty*, *Panic*, and *Excitement*.

Creators: The dataset was created by the Affexion project team — Gurkirat Singh, Karanbir Singh, and Leo Tsai — as part of a course project of CMPT 419 at Simon Fraser University.

Funding: No external funding or grants were used in the creation of this dataset. All resources used (e.g., OpenAI API credits) were obtained through personal student accounts, and it cost us 20 CAD.

Other Comments: The work was conducted as part of a research course on machine learning and affective computing with an intention for academic exploration rather than commercial deployment.

A.2 Composition

Instance Types: Each instance in the dataset represents short audio clips (approximately 3–4 seconds) as well as long audio clips (approximately 10–12 seconds) expressing one of the four emotions. There are two types of instances: synthetic audio (generated by AI) and real-world audio (collected from YouTube).

Instance Count:

- **Synthetic Audio:** 280 samples (approximately 70 per class)
- **Real-world Audio:** 52 samples (approximately 13 per class)

Sampling & Representativeness: The dataset is a curated sample rather than a statistically representative one. It was selected to reflect diverse examples of the target emotions, but does not cover the full spectrum of demographic or acoustic diversity. The training data set includes different male (2) and female (3) voices. There is a difference in pitch, loudness, and voice quality which makes it similar to lab-controlled data collection.

Data Content: Each instance consists of raw audio in WAV format. For training purposes, features such as MFCCs and delta coefficients were extracted. Label annotations were assigned by the team based on the intended social signal (synthetic) or by observation (YouTube samples).

Labels: Yes, each instance is labeled with one of the four target emotions: *boredom*, *uncertainty*, *panic*, or *excitement*.

Missing Information: No samples had missing audio. Some real-world clips lacked consistent quality due to background noise.

Explicit Relationships: In real-world data collected from YouTube, there were samples of *Panic* that were similar to those of *Excitement*.

Data Splits: No predefined train-test split was provided. A 7-fold cross-validation was used on synthetic data; real-world data was used solely for external evaluation.

Errors/Noise/Redundancy: Real-world samples include natural background noise and speaker variability. In contrast, synthetic data is noiseless and very consistent in each sample.

External Resources: Real-world clips were sourced from YouTube videos. There is no guarantee of their long-term availability. Links were not archived, but excerpts were downloaded locally for fair academic use.

Confidentiality: No confidential data was used. All content was publicly available online.

Offensive or Harmful Content: No instances of harmful or offensive language were observed in either set of data.

Subpopulations: No formal demographic or speaker metadata is included.

Identifiability: No individual is directly or indirectly identifiable from the dataset.

Sensitive Data: No sensitive personal data is included.

A.3 Collection Process

Acquisition Method:

- **Synthetic:** Generated using prompts through OpenAI's GPT-4o-audio-preview model. The prompt used was:
  "You are a professional voice actor, for a given sentence, please speak it with extreme 'Boredom' feeling."
- **Real-world:** Manually downloaded from YouTube based on visual and contextual cues indicating the target emotions.

Mechanisms Used:

- **Synthetic:** OpenAI Playground feature on their official website.
- **Real-world:** Manual curation using screen recording and FFmpeg for clip extraction.

Sampling Strategy: The synthetic set was balanced by design (equal samples per class). Real-world samples were selected opportunistically based on availability and clarity of emotional signal.

Personnel & Compensation: No external contractors were used. All data collection and labeling were done voluntarily by the student team. Data collection cost us 20 CAD.

Timeframe: All data was collected over two weeks in March–April 2025.

Ethical Review: No formal ethics review was required, as no data collection from human participants was conducted directly.

Data Origin:

- **Synthetic Data:** Created by the team.
- **Real-world Data:** Publicly available and not collected directly from subjects.

Consent & Notification: For real-world data, all clips were used under fair use from publicly accessible platforms. No direct interaction or consent from individuals was required.

Impact Analysis: The dataset poses minimal risk. It does not contain personal, sensitive, or identifiable information. No negative social or ethical impact is expected.

A.4 Preprocessing, Cleaning, and Labeling

Preprocessing: Real-world audio samples were initially selected, and those that did not clearly convey the intended emotions were subsequently removed. Audio was converted to 16 kHz mono-channel WAV. MFCCs, delta, and delta^2 features were extracted using Librosa. Clips were padded or truncated to a fixed length of 200 frames.

Raw Data Retention: Original WAV files for both synthetic and real-world clips were stored separately from the processed feature arrays.

Software Used: Librosa for audio processing, NumPy and PyTorch for feature extraction and modeling. Code is available upon request.

A.5 Uses

Prior Uses: The dataset has not been used prior to its creation.

Repository: No public repository exists yet. The dataset and models are currently hosted privately and can be shared upon request for academic purposes.

Potential Tasks: Social signal processing, synthetic vs. real data classification, speech feature extraction benchmarking, and multimodal emotion modeling.

Limitations & Risk Mitigation: The limited size and diversity of the dataset restrict its generalizability. Users should be cautious when deploying models trained on this data in real-world applications without further validation. All YouTube audio clips were collected under section 107 of the Copyright Act of 1976, allowing for FAIR USE. All future use of the dataset should comply with FAIR USE.

Unsuitable Tasks: The dataset should not be used for commercial deployment, speaker identification, or demographic profiling, as it lacks the necessary consent, diversity, and scale.

TEAM CONTRIBUTIONS

Karanbir Singh: Led the model architecture design and implementation, including integration of the multi-scale CNN, BiLSTM, and self-attention components. Collected the synthetic training data using ChatGPT, and contributed to code development, hyperparameter tuning, training pipeline, and results analysis.

Gurkirat Singh: Assisted in model design and algorithm development. Contributed to code implementation, hyperparameter tuning, and training pipeline. Collected synthetic training data from ChatGPT and analyzed results.

Leo Tsai: Collected the real-world validation dataset from YouTube across all four emotion categories. Also supported evaluation, result interpretation, and contributed to dataset documentation.