# A Multimodal Stress Estimation System Integrating Social Signal Processing and LLM for Enhanced Mental Health Support

Hossain Syeda Tanzina  Shun Shiramatsu

Department of Information Science, Nagoya Institute of Technology

## 1 Introduction

Increased stress is a growing concern worldwide, with significant impacts on individual mental health and well-being. Traditional stress detection methods often rely on single-modal data, such as text or audio, which limits their effectiveness and accuracy.[1] This research proposes a multimodal stress detection system that combines Social Signal Processing (SSP) techniques and voice feature analysis with a LLM such as GPT-4. The system analyzes user speech to interpret emotional states and stress indicators through both linguistic (text-based) and paralinguistic (audio-based) cues. The system utilizes LSTM networks to analyze temporal patterns in vocal features for enhanced stress detection. By leveraging the multimodal capabilities of GPT-4 alongside LSTM models trained on audio spectrograms, the proposed system estimates stress levels with enhanced accuracy and reliability.

## 2 Social Signal Processing and Mental Health

The prevalence of stress in modern society has highlighted the urgent need for effective and accessible mental health support tools. While traditional methods of stress assessment often require in-person clinical evaluations, advancements in Artificial Intelligence (AI) have paved the way for automated stress detection systems.However, many current systems focus on single-modal inputs, such as analyzing either the text or audio of user interactions, limiting their ability to capture the full spectrum of stress indicators. This research presents a multimodal approach that combines Social Signal Processing (SSP)[2] and voice feature analysis with the power of Large Language Models (LLMs), such as GPT-4, to create a more comprehensive and accurate stress detection system. Social Signal Processing provides a framework for interpreting nonverbal cues, such as tone, pitch, speech patterns, while LLMs enable nuanced understanding of linguistic content. By integrating these techniques, we aim to develop a system that not only detects stress but also offers insights to support users' metacognitive awareness and mental health. The proposed system's novelty lies in its ability to fuse data from multiple sources—linguistic and paralinguistic cues—allowing for a more robust estimation of stress levels.

## 3 Methodology

The proposed system comprises three primary modules:

### 3.1 Text Processing with GPT-4:

This module transcribes user speech into text and performs language analysis to detect stress-related emotions, such as sadness, fear. It identifies specific words and phrases associated with stress, using GPT-4's capabilities to interpret context and emotional nuance. The system capture the audio. A raw audio is passed into speech recognition API. OpenAI's Whisper to convert into text. This is where the spoken input ("The place is on fire...") is converted into a textual transcript. For text formatting the transcript may be cleaned or standardized (removing unnecessary pauses or artifacts).Once the transcript is ready, a prompt is formulated. The prompt include instructions to analyse the transcript for emotional content or specific stress indicators. This can be done using

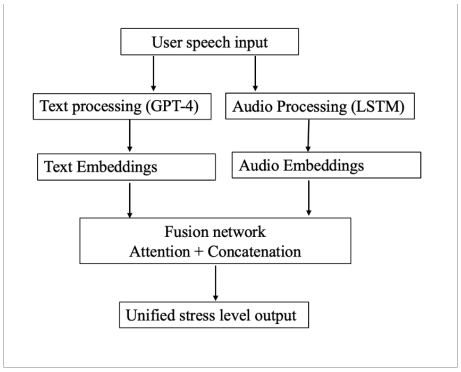


Figure 1: System Architecture for Multimodal Stress Detection

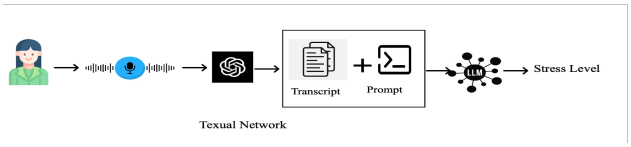


Figure 2: Textual Network

a template-based approach or dynamic prompt creation, where certain keywords (example: "fire," "help") trigger stress-related prompts. The transcript and prompt are fed into an LLM (GPT-4o). This step involves packaging the input as a query that the LLM understands, potentially formatted as "prompt": "from now you are expert to find any stress in the following voice transcript. ".Then, LLM processes the prompt and transcript, using its pre-trained knowledge of language patterns, semantics, and emotional cues to analyse the text. The LLM evaluates both the content ("fire," "help") and the context (the urgency in the sentence structure).

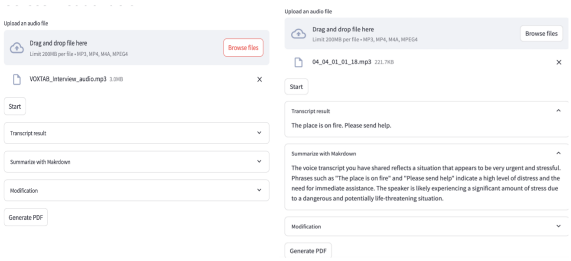

Figure 3: Textual analysis

### 3.2 Voice Analysis Using LSTM and Social Signal Processing:

This module processes vocal features such as frequency, amplitude. Model trained on spectrograms are used to classify emotions based on audio characteristics, and SSP

techniques provide additional insights into vocal stress indicators, such as pitch variations and vocal energy.

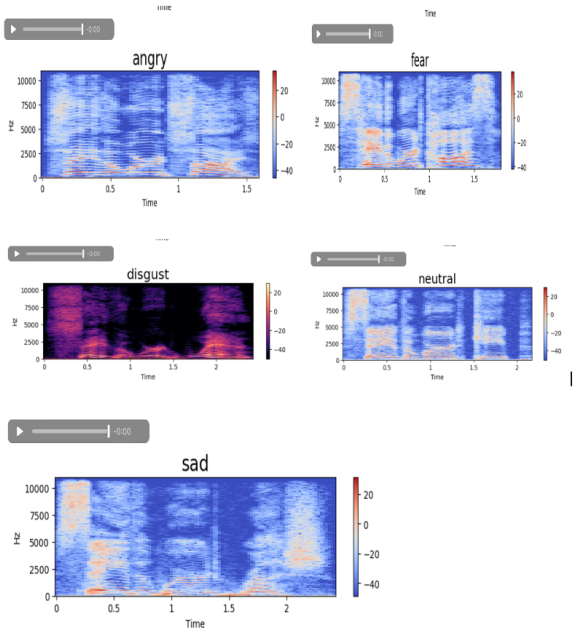


Figure 4: Frequency and amplitude spectogram

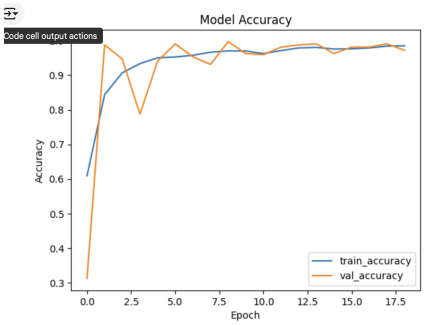


Figure 5: Model Accuracy

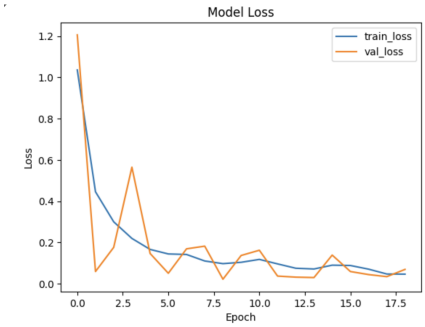


Figure 6: Model Loss

### 3.3 Stress Level Estimation:

Outputs from both text and audio analyses are combined using a fusion network that integrates linguistic and paralinguistic data. The system applies attention mechanisms to focus on most relevant features for stress estimation, producing a unified stress level. Concatenation merges two embeddings into one multimodal vector, capturing both text, audio features.

## 4 Result and Discussion

Data collection involves recording voice samples across a range of emotions and stress levels. Speech transcripts are generated for each recording, capturing linguistic information, emotional context. The audio data is processed to create spectrograms, both text and audio data undergo preprocessing for model training. GPT-4's embeddings are used to represent linguistic features, focusing on emotional context and stress-related keywords. The LSTM model processes sequential audio data to capture temporal dependencies, making it effective in detecting patterns of stress in user's voice over time.

The model accuracy plot shows the training and validation accuracy for LSTM model over several epochs in figure 6 and 7. I compared with CNN model and both models perform well in terms of accuracy, the LSTM model's stability and suitability for sequential data give it an edge over CNN for stress detection task.

Table 1: Experiment results for proposed model

| Type | Training Accuracy | Validation Accuracy |
|---|---|---|
| Model Accuracy | 0.9861 | 0.9962 |
| Model Loss | 0.0426 | 0.0695 |

## 5 Conclusion

This research presents a multimodal stress detection system that integrates Social Signal Processing, LSTM-based voice analysis, and GPT-4-based text analysis. By combining linguistic and paralinguistic cues, the system provides an assessment of stress levels. Future work will explore further personalization of the dialogue system and the inclusion of physiological data for enhanced stress detection accuracy.

References

[1] Schneeberger, Tanja, et al. "The deep method: Towards computational modeling of the social emotion shame driven by theory, introspection, and social signals." IEEE Transactions on Affective Computing (2023).

[2] Yongsatianchot, Nutchanon, Parisa Ghanad Torshizi, and Stacy Marsella. "Investigating large language models' perception of emotion using appraisal theory." 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE, 2023.

[3] Jaiswal, Mimansa, and Cristian-Paul Bara. "Muse: a multimodal dataset of stressed emotion." Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.

[4] Sun, Congshan, Haifeng Li, and Lin Ma. "Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network." Frontiers in Psychology 13 (2023): 1075624.