

Master Thesis

____ (Title) ____

A Multimodal System for Stress Detection and
Visualization using Social Signal Processing
and Large Language Models

社会的信号処理と **LLM** を用いたマルチモーダル
なストレス検出・可視化システムの開発

Supervisor: Prof. Shun SHIRAMATSU

Dept. of Computer Science
Graduate School of Engineering
Nagoya Institute of Technology

Enrolment: April 1, 2023

(Student ID): 35414100
(Name): HOSSAIN Syeda Tanzina

(Submission Date: February 3, 2025)

Abstract

Stress detection is increasingly crucial for mental health monitoring, yet traditional approaches relying on single-modal data, such as text or audio, often lack accuracy and contextual understanding. This research introduces a multimodal stress detection system that integrates Social Signal Processing (SSP) techniques, voice feature analysis, and Large Language Models (LLMs). By combining linguistic (text-based) and paralinguistic (audio-based) cues, the system enhances stress estimation through a hybrid approach. The system accepts audio inputs in various formats such as MP3, WAV, M4A and processes them to extract linguistic content and vocal stress indicators. The system processes audio inputs by extracting key speech features such as energy mean, zero-crossing rate, spectral centroid, and MelSpectrogram transformations to detect stress-related patterns in voice signals. These acoustic features are then analyzed using a pretrained emotional model to classify emotions into four categories: angry, happy, sad, and neutral. Additionally, the Whisper ASR model transcribes speech into text, which is further analyzed by GPT-4 to extract linguistic stress indicators, including word choices, sentence structure, and sentiment. By integrating paralinguistic and linguistic analysis, the system provides a robust estimation of stress levels. By combining features such as tone detection, Linguistic content, including lexical choices, syntactic structures, and sentiment-based expression sand vocal attributes like pitch and intensity, the system employs to estimate stress levels. To enhance user interaction and usability, the system is implemented with a Streamlit-based UI for real-time stress visualization. Current UX improvements focus on integrating stress trend history tracking to provide users with longitudinal insights into their stress patterns. This study contributes to the field of voice-based stress analysis by integrating with an interactive user interface, making stress detection more accessible and interpretable for real-world mental health applications.

Contents

Abstract	7
1 Introduction	13
1.1 Background	14
1.2 Motivation of the research	14
1.3 Problem Statement	15
1.4 Research Objectives	15
1.5 Overview of system approach	16
1.6 Overview of UI/UX system approach	16
1.7 Scope and Limitations	17
1.8 Thesis Organization	17
2 Literature Review	21
2.1 Stress and Mental Health: A Psychological Perspective	22
2.2 Voice-Based Stress Detection	23
2.3 Text based stress detection	23
2.4 Social Signal Processing (SSP) in Mental Health Systems	24
2.5 Machine Learning and Deep Learning for Stress Detection	24
2.6 Role of Large Language Models (LLMs) in stress Analysis	26
2.7 Limitations in Current Stress Detection and Conversational AI Systems	26
2.8 Ethical Challenges in AI-Driven Stress Detection	27
2.9 Research Gaps and Opportunities	29
2.10 Summary of the review	29
3 System Design	31
3.1 Proposed System	32
3.2 Dataset Description	33

3.3	Audio Prosodic features	35
3.4	Audio Prosody for stress visualization	35
3.5	Audio Prosodic features extraction techniques	37
3.6	Feature and emotion relationship	38
3.7	Feature and emotion mapping principle	38
3.7.1	Speech to Text Conversion (Whisper)	41
3.8	Emotion recognition model	41
3.9	System architecture of UI/UX design	42
3.9.1	Text Processing workflow	42
3.9.2	Audio Processing workflow	43
3.9.3	Stress visualization workflow	46
3.9.4	Stress trend analysis workflow	47
4	System Implementation and Results	51
4.1	Emotion Score Calculations and Normalization	52
4.1.1	A. Emotion Score Calculations	52
4.1.2	B. Score Normalization	53
4.1.3	Normalization Step	53
4.1.4	Emotion Score Calculation	54
4.2	Stress Analysis System Prompt	54
4.3	Stress Assessment	59
4.3.1	Final Distribution Analysis	59
4.3.2	GPT-4 Analysis Process	59
4.3.3	Output Generated	60
4.3.4	Interpretation of Stress Level	60
4.4	Overview of UI/UX design	61
4.5	Frontend Implementation	61
4.6	Backend Implementation	61
4.6.1	Audio Processing	62
4.6.2	GPT-4 Analysis phase	62
4.6.3	Output phase:Response generation	64
4.7	UI/UX enhancements	65
4.7.1	Development with Streamlit	65
4.7.2	UX Enhancements for User Engagement	65
4.8	Implementation output	71

5	Experiment and evaluation	75
5.1	Evaluation of Stress Detection	76
5.2	Correlation of coefficient between system output and human estimation	76
5.3	Visual comparisons of human ratings and system outputs . .	79
5.3.1	Comparison of Human stress level vs system rating . .	80
5.4	Statistical analysis	82
5.4.1	T-Statistic:	82
5.4.2	P value	82
5.5	Discussion of Results	83
5.5.1	Key Findings	83
5.5.2	Insights on Multimodal Data Integration	83
5.6	Advantages and Limitations of the System	83
6	Conclusion and Future Work	85
6.1	Summary of the Research	86
6.2	Contributions of the Study	86
6.3	Limitations of the Proposed System	87
6.4	Recommendations for Future Research	87
6.5	Practical Applications of the Study	87
6.6	Final Thoughts	89
	List of Tables	95
	List of Figures	95

List of Tables

2.1	Summary of literature review	28
3.1	Dataset Description	34
3.2	Prosody and stress relation	36
3.3	Feature-Emotion Relationships	38
5.1	Audio-text and text based system Output	77
5.2	Statistical Value	82

List of Figures

1.1	Overview of system approach	18
3.1	System design	33
3.2	Feature extraction formula for Energy, Spectral centroid, Zero Crossing Rate	37
3.3	Textual analysis	41
3.4	Proposed System Architecture	43
3.5	Text processing flow	44
3.6	Audio processing flow	45
3.7	Stress Visualization Flow	47
3.8	Stress trend analysis workflow	48
4.1	Emotion score calculation	53
4.2	Stress assessment	60
4.3	Audio Processing	63
4.4	GPT4 analysis phase	64
4.5	Response generation	64
4.6	Transcription from audio	67
4.7	Explanation and Advice	68
4.8	Explanation and Advice	69
4.9	Stress Visualization	70
4.10	Stress Trend Tracking History	71
4.11	Stress Level Detection System	73
4.12	Stress Level Detection System	74
5.1	A human stress rating and system stress level for Text-based system	78

5.2	A human stress rating and system stress level for Audio-Text-based system	79
5.3	Correlation Comparison between human rating vs system outputs	80
5.4	Trend Comparison between human rating vs system outputs	81

Chapter 1

Introduction

1.1 Background

In today's world, stress is an ongoing issue that affects both physical and mental health, influencing everything from mood and productivity to long-term well-being. Stress is a natural phenomenon that causes physical and emotional tension. Chronic stress is associated with serious diseases such as heart disease, anxiety, depression and immunosuppression. Stress is analyzed by expression, tone, pitch and physiological signals. The process of detecting when someone is stressed by measuring their physiological signals is known as stress detection. To analyze these signals and classify them as stressed or relaxed, some techniques are used. The physiological signals of a person are measured by physiological sensors such as the pulse of blood volume (BVP), the galvanic skin response (GSR), and the electrocardiograms (ECGs) (Sriramprakash et al. [2017]). In mental health care, intelligent technology has shown significant promise in delivering personalized treatments and real-time stress detection. (Kafková et al. [2024]) and (Liu et al. [2024]). Due to this, early and precise stress detection is vital for workplace wellness, healthcare, and personal well-being. However, many existing systems use single-modal data, such as text or voice, which limits their ability to capture all aspects of stress symptoms. This study explores the development of a versatile and ethical multimodal AI system designed for real-time stress detection. The system integrates voice and textual context to provide a comprehensive and accurate assessment of stress levels.

1.2 Motivation of the research

Stress has become a global problem that impacts not only mental health but also lifestyles and productivity. Due to the absence of easily accessible, reliable, and non-invasive methods, many people face obstacles in monitoring their stress. Conventional methods, such as self-assessment questionnaires or physiological monitoring, are inadequate in terms of timely insights or are impractical for daily use. Innovative systems that can effectively identify stress in real-world situations are, therefore, becoming more and more necessary. The voice is a perfect medium for stress detection because it is a natural source of emotional information. It transmits paralinguistic clues like tone, pitch, and rhythm in addition to linguistic content, which can reveal a person's emotional condition. Powerful tools like Large Language Models

1.1. Background

(LLMs), which can analyze textual material for deeper contextual and emotional insights, have been made possible by breakthroughs in Natural Language Processing (NLP). The majority of stress detection systems now only accept single input, which restricts the precision and usefulness. Current systems' emphasis on immediate stress assessments, which overlooks long-term patterns and trends, is another major drawback. Furthermore, existing tools often lack user-friendly interfaces, user cannot use without technical expertise. This study is motivated by an urge to develop a multimodal system that enhances the accuracy of stress detection by integrating text-based and voice-based analysis. Using LLMs for text analysis and deep learning models for audio processing, this study aims to close the gap between state-of-the-art AI technology. A more precise and user-friendly method of tracking stress over time is provided by incorporating interactive features like stress trend tracking, which also enables users to take proactive measures for improved mental health.

1.3 Problem Statement

Current stress detection systems often suffer from ethical concerns regarding data privacy, and a lack of adaptive intervention mechanisms. How can we design an ethical, adaptive, and multimodal AI system that not only detects stress accurately but also provides meaningful support to promote mental well-being?

1.4 Research Objectives

1. To develop a system that estimates and visualizes user stress levels using Social Signal Processing (SSP) techniques, including voice analysis for both linguistic (text-based) and paralinguistic (audio-based).
2. To develop a multimodal AI system that integrates voice analysis and large language models (GPT-4o) for real-time stress detection.
3. To evaluate the performance of proposed model in detecting stress from audio data.
4. To design an interactive Adaptive UI for clear stress visualization

1.3. Problem Statement

5. Designing a long term stress trends for identifying stress patterns and helps user for self-regulation and stress trend graph improves readability
6. Supporting metacognition and mental health care.
7. To ensure ethical data usage and privacy preservation in the proposed system.

1.5 Overview of system approach

The proposed System utilizes a multimodal approach to enhance stress detection by integrating SSP techniques, advanced voice feature analysis, and LLM. The system processes audio recordings through steps such as resampling, mono conversion, and normalization to prepare data for analysis. Key audio features, including energy mean, zero-crossing rate, spectral centroid, and MelSpectrogram transformations, are extracted and normalized to capture stress-related acoustic patterns. Additionally, the Whisper ASR model transcribes speech into text, which is analyzed using GPT-4 to identify linguistic cues like word choices, sentence structures, and sentiment. By combining paralinguistic and linguistic features, the system delivers robust stress detection, advancing the field of voice-based stress analysis for real-world mental health applications. The system employs MelSpectrogram transformation for spectral analysis and extracts four key acoustic features essential for stress detection. These include Energy to measure signal intensity, Energy Standard Deviation to capture volume variations, Spectral Centroid to represent frequency distribution, and Zero Crossing Rate to quantify pitch variations. Together, these features provide a comprehensive representation of the acoustic patterns associated with stress.

1.6 Overview of UI/UX system approach

The proposed system as shown in figure 1.1 detects and visualizes stress levels through a three-stage workflow: Audio Input, Stress Analysis, and Results and Visualization. Users upload audio files in formats such as MP3, WAV, or M4A, with playback support for review. The system processes the audio to extract textual and acoustic features. Text analysis identifies stress-related

1.5. Overview of system approach

linguistic patterns, while audio analysis examines acoustic properties like pitch and tone. The results are visualized through stress reports, distribution charts, and historical trends, with an option to download detailed reports. The system also supports real-time detection, providing immediate feedback for enhanced usability.

1.7 Scope and Limitations

Scope: Focuses on voice-based stress detection and designs a stress trend module for user support. **Limitations:** Does not include physiological signal data or longitudinal studies of user impact.

1.8 Thesis Organization

This thesis is structured into six chapters, each presenting a particular aspect of the study. Below is a brief overview of each chapter:

1. Chapter 1 cover the research background, highlighting the importance of stress detection in mental health monitoring. It outlines the motivation, problem statement, research objectives, research questions, the significance of the study and scope and limitations, and the thesis organization.
2. Chapter 2 The literature review assesses previous work related to stress detection, focusing on speech-based and text-based emotion recognition, Social Signal Processing (SSP), and the role of deep learning and large language models, limitations in current stress detection and conversational AI systems. It highlights the lacking in current methods and shows the need for a multimodal strategy.
3. Chapter 3 This chapter describes the design and progress of the proposed multimodal stress detection system. It outlines the collection of dataset, data pre-processing procedures, and the two main modules: audio-based and text-based stress detection. Additionally, it describes the model designs, training and validation procedures and multimodal fusion techniques.

1.7. Scope and Limitations

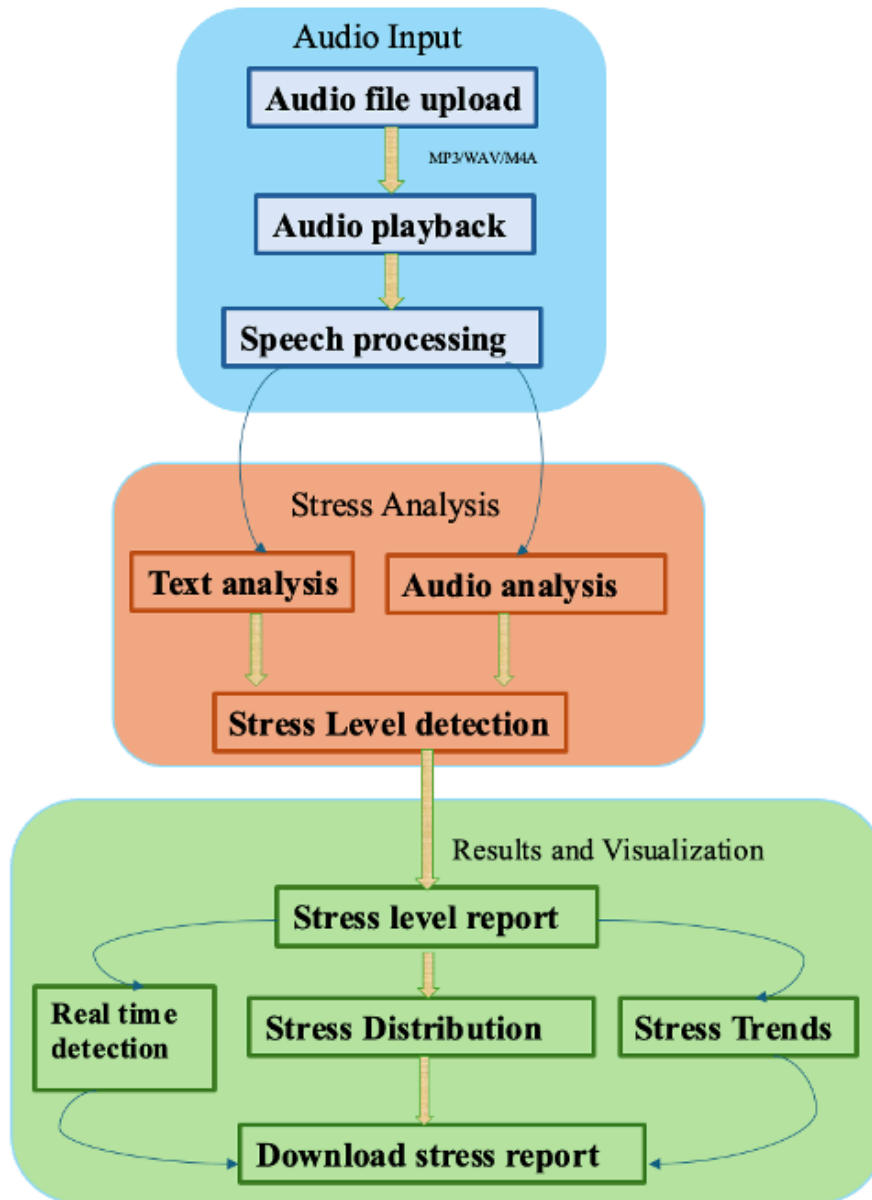


Figure 1.1: Overview of system approach

1.8. Thesis Organization

4. Chapter 4 This chapter highlights the actual implementation of proposed system. It describes the architecture and workflow, including speech-to-text processing, feature extraction. The chapter also describes the development of the user interface (UI), highlighting features such as real-time stress visualization and stress trend tracking, and report generation.
5. Chapter 5 This chapter evaluates the performance of the proposed models, comparing their accuracy, precision, and recall across different approaches. It also includes experimental results between stress trend analysis system and human subjective analysis for stress. So, it highlights the system advantages and disadvantages.
6. Chapter 6 This chapter covers summary of the study, limitations and future considerations. Highlighting the research's contributions to the field of stress detection, the final chapter provides a summary of the main findings. In addition to discussing the shortcomings of the current system, it suggests future research objectives, such as real-time deployment, dataset enlargement, and interaction with other modalities including physiological data.

1.8. Thesis Organization

Chapter 2

Literature Review

2.1 Stress and Mental Health: A Psychological Perspective

Stress has a major impact on mental health and well-being through changes in physiological, emotional, and cognitive reactions. Effective techniques for early diagnosis and intervention are required because prolonged stress has been associated with a number of psychiatric diseases, such as anxiety and depression. Although the study (Sebastião and Neto [2024]) showed on stress and mental health in COVID-19 emphasizes the importance of psychological flexibility and emotional schemas, it has a number of drawbacks. Avoid making judgments about causality by using a cross-sectional strategy. Potential biases are introduced by relying solely on self-reported data. Generalizability is limited by the sample. Results may also be affected by unmeasured variables, including socioeconomic status and previous mental health problems. Finally, results are impacted by the pandemic environment, which reduces their generalisability in non-pandemic scenarios. Despite these drawbacks, the study offers valuable information on the dynamics of mental health in emergencies. Traditional stress assessment relies on self-reported questionnaires such as the Perceived Stress Scale (PSS) and physiological measures such as heart rate variability (HRV) and cortisol levels. However, these methods can be intrusive, subjective, or impractical for continuous monitoring, leading to the need for automated non-invasive stress detection systems. In order to detect stress, (Aristizabal et al. [2021]) investigates the combination of self-reported data and wearable physiological sensors. In addition to self-reported stress levels, participants' electrodermal activity (EDA), heart rate (HR), and skin temperature were tracked using wearable technology as they performed stress-inducing tasks. These data streams were analysed by deep learning systems, which showed excellent accuracy in differentiating between stress levels. The study's semi-controlled laboratory environment, which would not accurately mimic stressors in the real world, and possible sample size and diversity restrictions, which would restrict generalisability, are some of its drawbacks. Reliability may also be impacted by sensor errors brought on by motion artefacts, self-report biases, and difficulties capturing quick shifts in stress.

2.1. Stress and Mental Health: A Psychological Perspective

2.2 Voice-Based Stress Detection

The communication medium Speech is a rich medium for detecting stress as it carries both linguistic and paralinguistic markers indicating stress. Research in speech emotion recognition (SER) has identified variations in prosodic features, such as pitch, intensity, speech rate, and spectral characteristics, as key indicators of stress with 97.1 accuracy utilising the Crema-D and TESS datasets, (Chyan et al. [2022]) "A Deep Learning Approach for Stress Detection Through Speech with Audio Feature Analysis" investigates the use of CNN models for classifying stressed and unstressed speech based on Mel Spectrogram and MFCC characteristics. Nevertheless, the study has disadvantages, such as a controlled environment that does not replicate naturalistic noise situations, a binary classification approach that ignores shifting stress levels, and dataset constraints that do not represent real-world speech changes. Furthermore, the model's real-time applicability has not been validated, and its reliance on a small number of audio variables may ignore other elements. the study shows how deep learning may be used for speech-based stress detection and establishes the foundation for upcoming advancements in practical applications. This work employed the Toronto Emotional Speech Set (TESS), a popular dataset for emotion identification that includes vocal expressions from a range of emotional states. Convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have been used in a number of research to extract and analyse voice data in order to classify emotions. While LSTMs capture temporal dependencies in sequential data, CNNs have been especially successful in processing spectrograms, which graphically depict frequency and amplitude fluctuations over time. By utilising both spatial and temporal feature representations, the hybrid models' integration of these architectures will increase the accuracy of stress detection.

2.3 Text based stress detection

in this study, author (Yoon et al. [2020]) investigates stress detection using semantic analysis of textual data, including private messages and postings on social media. The article analyses the approach on pertinent datasets to show its efficacy and finds linguistic patterns linked to stress using natural language processing (NLP) techniques. Nevertheless, there are a number of drawbacks,

2.2. Voice-Based Stress Detection

such as dataset restrictions that can fail to account for linguistic variation, contextual subtleties that could result in incorrect categorisation, and the lack of multimodal data like visual or aural signals. Furthermore, the model can have trouble keeping up with changing linguistic trends, necessitating frequent updates, and it poses privacy issues with relation to the use of personal data.

2.4 Social Signal Processing (SSP) in Mental Health Systems

In this study, author(Singh et al. [2019]) examines how SSP can be used to evaluate conversations by examining sentiment and emotional indicators. In order to better comprehend human interactions, SSP looks at non-verbal cues including gestures, vocal intonation, and facial expressions. This is especially important in industries like banking and finance, where it's critical to identify client dissatisfaction. The work intends to increase dialogue analysis for customer experience and service improvement by assessing emotional intensity in large data sets. However, the study has challenges, such as contextual limits because emotional expressions depend on conversational context and data quality issues because noisy or missing data might impair accuracy. Cultural differences also affect how nonverbal cues are interpreted, and deployment is complicated by the high computational resources required for real-time processing. Furthermore, analysing chats without the express consent of the user raises privacy problems.

2.5 Machine Learning and Deep Learning for Stress Detection

The application of Machine Learning (ML) and deep learning (DL) approaches to stress detection and mental health monitoring is examined in this articleRazavi et al. [2024]. It draws attention to the efficacy of models such as Support Vector Machines (SVMs), Neural Networks (NNs), and Random Forest RFs, which use physiological data like skin reaction, heart rate variability (HRV), and heart rate (HR) as important stress indicators. In order to improve model performance, the study also highlights the significance of

2.4. Social Signal Processing (SSP) in Mental Health Systems

data preparation methods including feature selection and noise reduction. Nevertheless, there are a number of drawbacks, such as the inability to customise ML/DL models, which hinders their ability to adjust to different stress patterns, and their poor interpretability, which makes them less clear for clinical usage. Furthermore, the majority of research is carried out in controlled settings, which restricts its relevance in the real world, and real-time processing is still in its infancy, which makes real time interventions less feasible. In order to increase classification accuracy, the Yao et al. [2021] presents a technique for stress detection that combines linguistic cues from text with acoustic data from speech. Compared to unimodal systems that only use text or audio, the multimodal method performs better because it uses machine learning techniques to analyse the interaction between language and vocal emotions. The lack of high-quality, labelled datasets with both text and audio components is one of the study's flaws though, and it may limit generalisability. Implementation is resource-intensive due to the contextual diversity in stress expression between cultures and individuals, which necessitates substantial training on varied datasets. In real-time processing, where accuracy and low latency must be preserved in dynamic situations, the system also faces difficulties. Error sources may also be introduced by the difficulties in synchronising text and auditory data. Hilmy et al. [2021] emphasises how well machine learning models like Support Vector Machines (SVMs) and Mel-Frequency Cepstral Coefficients (MFCCs) work for categorising stress and extracting spectral information from speech. While SVMs have strong classification capabilities, MFCCs are especially good at capturing speech characteristics that distinguish between stressed and neutral states. Studies have shown that this method may identify stress with high accuracy rates 88. Nevertheless, drawbacks include reliance on varied and high-quality datasets, which are frequently challenging to get by, and environmental noise, which can reduce the precision of feature extraction. Furthermore, the computing needs of SVM classification and MFCC extraction create problems for real-time processing. Model generalisability is also impacted by individual variations in speech patterns, such as accents and speaking styles, which calls for customised or adaptive methods.

2.5. Machine Learning and Deep Learning for Stress Detection

2.6 Role of Large Language Models (LLMs) in stress Analysis

In this research, author (Shen et al. [2024]) investigates if Large Language Models (LLMs) undergo similar performance variations to humans when under stress. According to the Yerkes-Dodson law, which states that excessive or insufficient stress lowers efficiency, researchers discovered that LLMs operate best under moderate stress using StressPrompt, a series of stress-inducing prompts based on psychological theories. The study also showed that LLMs' internal brain representations were impacted by stress-altering stimuli, which mirrored human stress reactions. Nevertheless, there are a number of drawbacks, including as the limited generalisability of prompt-based stress induction—which may not accurately mimic real stress experiences—and the fact that results may differ depending on the LLM design. Furthermore, because the study was carried out in controlled environments and LLMs do not have human-like physiological reactions, it lacks real-world validation, which may affect how accurately stress comparisons are made.

2.7 Limitations in Current Stress Detection and Conversational AI Systems

In this study, author (Teye et al. [2022]) investigates how contextual and cultural factors affect conversational AI systems' ability to recognise emotions. In order to classify seven fundamental emotions, including sarcasm, the study created an emotion prediction model that included voice and picture data, with accuracies ranging from 85% to 95%. In order to increase the dependability and moral coherence of emotion detection systems across a range of demographics, the authors stress the importance of taking environmental factors and cultural quirks into account. However, the study has drawbacks, such as difficulties with data representation because datasets are frequently undiversified, which impairs the model's cross-cultural generalisability. Furthermore, despite efforts to balance, biases in datasets continue to exist, which could diminish fairness. Because sarcasm is nuanced and context-dependent, it is still very difficult to detect. The performance of the system can also be impacted by environmental fluctuations, including background noise. By utilising real-time data fusion from several sources,

2.6. Role of Large Language Models (LLMs) in stress Analysis

including environmental sensors, biometric devices, and Internet of Things systems, the combination of sensor technologies and conversational AI allows for context-sensitive interactions. By improving AI's comprehension of user context, physical circumstances, and environment, author Kush [2025] enables more intelligent and flexible interactions. For instance, conversational AI in healthcare enhances the responsiveness and personalisation of virtual health aides by integrating data such as temperature, stress levels, and heart rate. To fully realise the potential of these systems, however, issues like interoperability, scalability, and data heterogeneity must be resolved. The ability of sensor-integrated AI to provide real-time, context-aware help is demonstrated by real-world implementations, such as Mercedes-Benz's integration of Google's conversational AI agent into cars, opening the door for cutting-edge solutions across industries.

2.8 Ethical Challenges in AI-Driven Stress Detection

In the paper, author (Sriramprakash et al. [2017]) examines the expanding use of AI in mental health care, emphasizing how it can enhance early identification, accessibility, and personalized care. Virtual assistants and AI-powered chatbots offer prompt assistance, and predictive analytics can help spot mental health problems before they get out of hand. The study also identifies problems that make it difficult to understand AI-driven decisions, including algorithmic transparency issues, lack of human empathy in AI interactions, and data privacy concerns. Along with difficulties in incorporating AI into conventional healthcare systems due to patient and practitioner reluctance, ethical issues such as consent and autonomy continue to be crucial. In order to guarantee that AI improves mental health care while upholding ethical standards, further developments in emotionally intelligent AI, improved integration with healthcare services, and robust regulatory frameworks will be crucial.

2.8. Ethical Challenges in AI-Driven Stress Detection

Table 2.1: Summary of literature review

Author	Technique	Findings	Limitations
Chyan et al. [2022]	CNN for speech-based stress detection	Utilizes MFCCs and Mel Spectrograms to classify stress in speech with high accuracy (97.1%).	Controlled setup; binary classification misses stress gradation; lacks real-time validation; ignores real-world noise.
Study on LLMs Teye et al. [2022]	StressPrompt for stress analysis in LLMs	Demonstrates performance variations in LLMs under stress using psychological prompts.	Lacks real-world validation; results depend on specific LLM design and controlled settings; no physiological equivalence to humans
Yoon et al. [2020]	Semantic analysis of textual data	Identifies linguistic patterns in private messages and social media posts to detect stress using NLP	Dataset limitations; lacks multimodal data; struggles with linguistic trends and contextual subtleties; raises privacy concerns.
Singh et al. [2019]	Social Signal Processing (SSP) for emotion detection	Analyses sentiment through non-verbal cues such as gestures and intonation.	Data quality issues; computationally intensive; contextual limitations; privacy concerns in analyzing conversations
Multimodal Study Razavi et al. [2024]	Combining text and audio signals	Outperforms unimodal approaches by integrating linguistic cues and vocal emotions; uses ML for interaction analysis.	Resource-intensive; lacks high-quality multimodal datasets; difficulties in synchronizing text and audio data for real-time application
Teye et al. [2022]	Ethical AI in mental health care	Highlights AI's potential for early intervention and personalized care.	Challenges in algorithmic transparency; lack of human empathy in AI; privacy concerns.

2.8. Ethical Challenges in AI-Driven Stress Detection

2.9 Research Gaps and Opportunities

The majority of the research that has been done on stress detection has been either text-based or audio-based. Even while CNNs and LSTMs have shown excellent accuracy in identifying speech emotions, feature selection and dataset quality changes depending on dataset. LLMs have shown promise in extracting stress-related linguistic features, but their effectiveness in combination with voice analysis is still being explored. In order to fill this gap, this study proposes a multimodal stress detection system that uses both textual and auditory signals. A real-time user interface (UI) with stress trend tracking is also created to enhance accessibility and user interaction. In order to facilitate practical uses in mental health monitoring and emotional well-being, the results of this study assist in establishing solid and interpretable stress detection frameworks.

2.10 Summary of the review

Despite significant advancements in text and audio-based stress detection, current approaches as shown in Table 2.1 lack the accuracy and contextual understanding needed for real-time applications. This review highlights the need for multimodal systems that integrate linguistic and paralinguistic cues, offering a foundation for this study.

Chapter 3

System Design

Text and audio embeddings are the two main modules that comprise up system. The system is thoughtfully designed to effectively combine both textual and audio information, creating a strong foundation for multimodal stress detection. The ability of the system to produce responses that resemble those of a human is directly influenced by the textual interpretation with whisper and GPT-4 analysis. Key aspects of the design include well-structured pipelines, efficient deployment strategies, and thorough dataset preparation. Each element has been carefully crafted to enhance the system's usability, scalability, and accuracy, ensuring it delivers reliable and practical results.

3.1 Proposed System

The proposed System shown in figure 3.1 utilizes a multimodal approach to enhance stress detection by integrating SSP techniques, advanced voice feature analysis, and LLM. The system processes audio recordings through steps such as resampling, mono conversion, and normalization to prepare data for analysis. Key audio features, including energy mean, zero-crossing rate, spectral centroid, and MelSpectrogram transformations, are extracted and normalized to capture stress-related acoustic patterns. Additionally, the Whisper ASR model transcribes speech into text, which is analyzed using GPT-4 to identify linguistic cues like word choices, sentence structures, and sentiment. By combining paralinguistic and linguistic features, the system delivers robust stress detection, advancing the field of voice-based stress analysis for real-world mental health applications. The system employs MelSpectrogram transformation for spectral analysis and extracts four key acoustic features essential for stress detection. These include Energy to measure signal intensity, Energy Standard Deviation to capture volume variations, Spectral Centroid to represent frequency distribution, and Zero Crossing Rate to quantify pitch variations. Together, these features provide a comprehensive representation of the acoustic patterns associated with stress

3.1. Proposed System

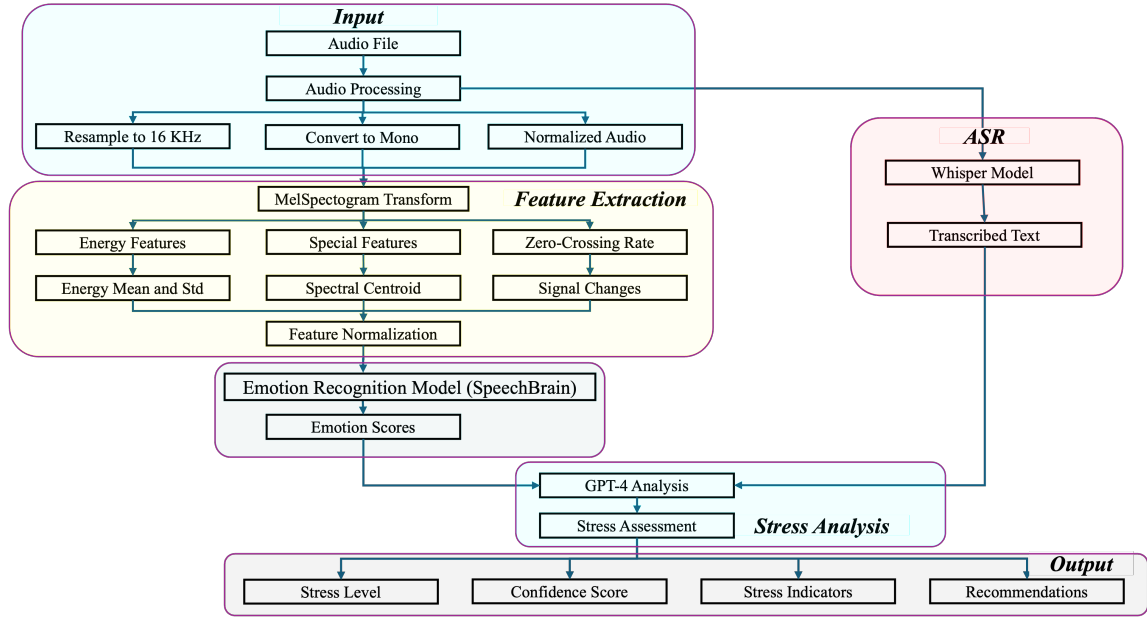


Figure 3.1: System design

3.2 Dataset Description

A limited number of audio recordings was used to generate and illustrate the functioning of the stress detection system's user interface (UI) and user experience (UX) components. These recordings, which come from openly accessible websites like Google, feature speech samples with a range of psychological tones. Even though the audio files don't represent a formal or standardized dataset, they are a useful tool for demonstrating the system's features, like trend monitoring and real-time stress level visualization. I used the audio file to capture voice at different stress levels. After processing these data, stress detection outputs were simulated and shown on the user interface. These samples are primarily being used to test and validate the stress visualisation features of interface and illustrate how stress trend tracking works throughout several sessions. Also checking interface is capable of handling various stressful aspects.

3.2. Dataset Description

Table 3.1: Dataset Description

Audio	Text in the Audio
Audio1	I wonder what this is about
Fire	The place is on fire, please send help
Test1	Loans that we can offer with this or farm ownership loans, operating lines of credit, or equipment and capital improvement need loans. The benefit to the
sad.Wav	The best food novel count me on the edge of my side
New recording 19.wav	時間が戻れば私はあの時間に戻りたいま一回
New recording 16.wav	So um... I'm sick. But my friend wanted me to be a deity myself because we always had to really cool. So uh... Yeah, here it is. Do you want me to sing? Oh, it sounds really good singing.
new recording 23.wav	Life is difficult, but you are loved. You are loved and important, and you bring to this world things that no one else can so hold on
Airplane.wav	The airplane is almost full.
Gamer.mp3	Why would you not want to fight for what you believe in?
Sad.mp3	The delicious aroma of freshly baked bread filled the bakery.
1008TIE.wav	That is exactly what happened.
New recording 25.wav	Us comfort help isn't given up. It's refusing to give up. To use fall for a reason, and to use strength and weakness. I think...

3.2. Dataset Description

3.3 Audio Prosodic features

Prosodic signals (Gibbon [2017]) refer to the patterns of stress and intonation in speech, and they play a key role in detecting and visualizing stress in an individual. In stress visualization systems, prosodic features such as pitch, tone, rhythm, loudness, and speech rate can provide rich data to infer emotional or psychological states.

1. **Pitch (Fundamental Frequency):** Stress often causes a rise or fall in pitch, leading to higher or more variable frequency in speech. Pitch is the perceptual property of sound that allows us to distinguish between high and low tones. It is directly related to the fundamental frequency (F0) of a sound wave, which is measured in Hertz (Hz). In speech, pitch is primarily determined by the vibration rate of the vocal cords (also called the vocal folds). When the vocal cords vibrate faster, the pitch is higher; when they vibrate slower, the pitch is lower.
2. **Speech Rate:** People under stress may speak faster or slower than usual. Speech rate, measured in syllables per second, is a clear indicator of anxiety or calmness.
3. **Intensity (Loudness):** Loudness tends to increase during moments of stress. A person may speak at a higher volume or more forcefully.
4. **Voice Jitter and Shimmer:** Jitter (variability in pitch) and shimmer (variability in loudness) tend to increase when someone is stressed, leading to more unstable or irregular speech patterns.
5. **Pauses and Hesitations:** Longer pauses or frequent hesitations in speech can indicate cognitive load or emotional stress.
6. **Formant Frequencies:** Formants represent resonance frequencies in the voice and change with emotional arousal or tension.

3.4 Audio Prosody for stress visualization

Stress in a voice can be perceived through specific acoustic and prosodic features that change due to emotional or physiological tension. These features can be both consciously and subconsciously noticeable when listening

3.3. Audio Prosodic features

to someone's speech. Stress in a voice can often be identified through various vocal changes. (Paulmann et al. [2016]) Pitch tends to rise or fluctuate unpredictably, making the voice sound higher or shaky under tension. Speech rate may also vary, with some individuals speaking faster due to racing thoughts, while others slow down or pause frequently due to hesitation or cognitive overload. Volume often increases when frustration or agitation is present, while softer, quieter tones can indicate anxiety or uncertainty. The quality of the voice might also change, becoming strained, breathy, or hoarse as stress affects the vocal cords. Additionally, rhythm may become irregular, with uneven pacing and an increased use of fillers like "um" or "uh." The emotional tone of the voice might sound irritable, sharp, or flat, depending on the individual's stress response. Finally, articulation can also be affected, and some speech can become slurred or too precise, reflecting the tension or effort to maintain consistency.

Table 3.2: Prosody and stress relation

Feature	How features are related to stress
Higher Pitch	Increased stress or excitement (e.g., when someone is nervous, their voice tends to be higher).
Lower pitch	Calmness or relaxation.
pitch variability	A highly stressed person may have more fluctuations in pitch due to nervousness or tension.
Louder voice	Raised voice or shouting can indicate frustration or stress.
Quieter voice	Stress might also lead to a softer, restrained voice, particularly if the person is anxious or unsure.
Irregular rhythm	The voice might sound rushed or uneven
Frequent pauses or fillers	"Um," "uh," and other fillers increase as a person hesitates under stress.

3.4. Audio Prosody for stress visualization

3.5 Audio Prosodic features extraction techniques

In this study, four feature extraction techniques are used for extracting voice feature. The system utilizes MelSpectrogram transformation for detailed spectral analysis and extracts four critical acoustic features that correlate with emotional states and stress levels using the formula in figure 3.2.(Singh et al. [2012])

1. Energy: Measures the overall signal intensity, reflecting loudness variations, which can indicate heightened emotions such as anger or excitement.
2. Energy Standard Deviation: Captures fluctuations in volume, helping to differentiate between calm (steady energy) and stressed (irregular energy) states.
3. Spectral Centroid: Represents the center of frequency distribution, where higher values may indicate high-pitched, tense speech associated with stress or anxiety.
4. Zero Crossing Rate: Quantifies the rate of signal sign changes, linked to pitch variations; increased rates often reflect nervousness or agitation.

A. Energy Features (Volume/Intensity) <ul style="list-style-type: none"> • Uses Root Mean Square (RMS) energy from Mel Spectrogram • Formula: $E = \sqrt{(1/N \sum x(n) ^2)}$ <ul style="list-style-type: none"> • E is energy • x(n) is signal amplitude • N is frame length • Captures voice intensity and stress patterns 	B. Pitch Features (Spectral Centroid) <ul style="list-style-type: none"> • Calculates weighted mean of frequencies • Formula: $SC = \sum(f(n) * M(n)) / \sum M(n)$ <ul style="list-style-type: none"> • f(n) is frequency • M(n) is magnitude • Indicates voice pitch characteristics 	C. Rhythm Features (Zero Crossing Rate) <ul style="list-style-type: none"> • Measures signal polarity changes • Formula: $ZCR = (1/T) \sum \text{sign}[x(n)] - \text{sign}[x(n-1)] / 2$ <ul style="list-style-type: none"> • T is frame length • x(n) is signal value • Detects speech rhythm and periodicity
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 3.2: Feature extraction formula for Energy, Spectral centroid, Zero Crossing Rate

3.5. Audio Prosodic features extraction techniques

3.6 Feature and emotion relationship

The table presents 3.3 the relationship between different speech features and emotional states. Energy, which represents volume, is associated with high values indicating anger or happiness, low values indicating sadness, and medium values corresponding to a neutral state. Spectral centroid, which reflects pitch, is linked to happiness when high, sadness when low, and neutrality at medium levels. Energy standard deviation, which measures variations in volume, shows that high values are related to anger, low values are linked to neutral or happy emotions, and medium values indicate sadness. Zero crossing rate, representing rapid changes in the signal, is associated with anger when high and sadness or neutrality when low. These relationships highlight how variations in speech characteristics influence emotional classification. (Arias et al. [2021])

Table 3.3: Feature-Emotion Relationships

Feature	Emotion Relationships
Energy (Volume)	High → Angry/Happy Low → Sad Medium → Neutral
Spectral Centroid (Pitch)	High → Happy Low → Sad Medium → Neutral
Energy Standard Deviation (Variation)	High → Angry Low → Neutral/Happy Medium → Sad
Zero Crossing Rate (Rapid Changes)	High → Angry Low → Sad/Neutral

3.7 Feature and emotion mapping principle

The section describes the relationship between speech features and emotional states based on previous research. High energy features are linked to active emotions such as anger and happiness and indicate heightened emotional states. The system design incorporates feature-emotion mapping to improve stress detection accuracy. Spectral features play a significant role in emotion

3.6. Feature and emotion relationship

classification. Higher frequencies are associated with positive emotions, lower frequencies with negative emotions, and mid-range frequencies with neutral states. Rhythm features also contribute to emotional assessment. Rapid changes in rhythm are related to agitated states, steady patterns indicate calmer states, and irregular patterns serve as stress indicators. These relationships help in mapping speech characteristics to emotional and stress states, enhancing the accuracy of the stress detection system. (Arias et al. [2021])

A. High Energy Features

- Associated with active emotions (anger, happiness)
- Indicates heightened emotional states
- Higher stress correlation

B. Spectral Features

- Higher frequencies → positive emotions
- Lower frequencies → negative emotions
- Mid-range → neutral states

C. Rhythm Features

- Rapid changes → agitated states
- Steady patterns → calmer states
- Irregular patterns → stress indicators

This portion outlines the characteristics of different emotional speech patterns based on their acoustic features. Angry speech is marked by high energy, a high zero-crossing rate, and large energy variations, reflecting its intense and dynamic nature. Happy speech also has high energy but is distinguished by a higher spectral centroid and moderate energy variations, indicating a lively but more controlled tone.

Sad speech, in contrast, is characterized by low energy and a lower spectral centroid, with moderate variations that reflect a subdued and melancholic tone. Neutral speech maintains medium energy with consistent patterns and low variations, indicating a balanced and steady speech style. These characteristics help in identifying emotional states in speech-based stress detection systems

3.7. Feature and emotion mapping principle

A. Angry Speech Characteristics

- High energy
- High zero-crossing rate
- Large energy variations

B. Happy Speech Characteristics

- High energy
- Higher spectral centroid
- Moderate variations

C. Sad Speech Characteristics

- Low energy
- Lower spectral centroid
- Moderate variations

D. Neutral Speech Characteristics

- Medium energy
- Consistent patterns
- Low variations

3.7. Feature and emotion mapping principle

3.7.1 Speech to Text Conversion (Whisper)

The use of Whisper(Machaek et al. [2023]) for accurate transcription. For the flow of text generation This text processing module combines the manual transcript input option, text cleaning and preparation, and whisper to convert speech to text. Whisper converts speech to text. Draw attention to Whisper’s main function of converting audio into text transcripts. This module transcribes user speech into text and performs language analysis to detect stress-related context. Identifies specific words and phrases associated with stress, using GPT-4’s capabilities to interpret context. The system captures the audio. The raw audio is passed into the speech recognition API. OpenAI’s Whisper to convert into text. The voice in the audio file is converted into a textual transcript. For text formatting, the transcript is formed as cleaned or standardized and removing unnecessary pauses or artifacts. The significance of text preparation for analysis, noise reduction, and transcription accuracy is emphasized. The Whisper model processes audio inputs and generates a transcript with a high degree of accuracy, filtering out noise and ensuring that only meaningful words are passed to the next stage.

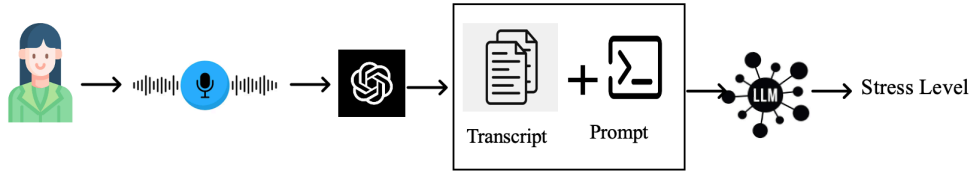


Figure 3.3: Textual analysis

3.8 Emotion recognition model

In this research, the emotional model plays a key role in analyzing speech-based stress levels. The system extracts various acoustic features from the user’s speech, such as energy mean, zero-crossing rate, spectral centroid, and MelSpectrogram transformations. These features are then processed by an emotion recognition model(Venkataramanan and Rajamohan [2019]) trained on the IEMOCAP dataset, classifying speech into four emotional states: an-

3.8. Emotion recognition model

gry, happy, sad, and neutral. The emotional model helps quantify the user's stress level by assessing paralinguistic cues, such as pitch variations and intensity, which are linked to emotional expression. This information is later combined with text-based analysis for a comprehensive stress assessment.

SpeechBrain, an open-source speech processing toolkit, is utilized for advanced speech analysis tasks in this system. It provides tools for feature extraction, speech emotion recognition, and automatic speech recognition. The system uses SpeechBrain (Ravanelli et al. [2024]) models to process voice input efficiently, ensuring accurate extraction of speech characteristics. This integration allows the system to improve emotion classification performance and enhances the reliability of stress detection by leveraging pre-trained models optimized for speech-based applications

3.9 System architecture of UI/UX design

Using multimodal inputs, the suggested system combines several parts to identify and display stress levels. The architecture is separated into three main modules, as seen in Figure 3.4: Frontend (Streamlit): In handling user interaction, including audio file uploads and results visualisation. Backend (Flask): Manages essential features including stress analysis, speech-to-text conversion, and external service interaction. External Services: These comprise advanced text analysis APIs such as OpenAI for GPT-4. These parts work together seamlessly ensuring that the system is scalable, responsive, and able to process inputs in real time.

3.9.1 Text Processing workflow

The Text Processing Workflow describes the procedures for evaluating and interpreting the text to detect stress. Figure 3.5 shows the text processing workflow, which consists of the following steps: Whisper converts audio into raw text. Using semantic context, GPT-4 analyzes the text to find stress signals. The system uses stress scale to classify stress levels (Low, Moderate, High). The system gives feedback, including stress management tips, based on the user's degree of stress. Each component in a text processing flow is responsible for handling a specific task. For example, an input handler handles text input from audio or manual entry, while an analysis engine uses GPT-4 to analyze text for stress content and analyze numerical stress levels.

3.9. System architecture of UI/UX design

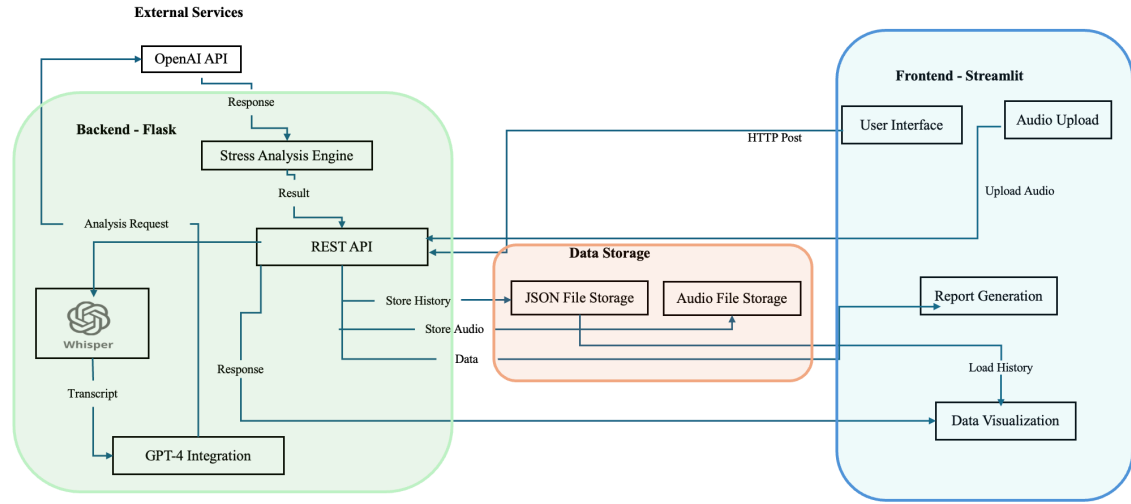


Figure 3.4: Proposed System Architecture

The formatter gets the results ready to be shown. The report generator generates reports that can be downloaded, while the display handler manages visual representation. It focuses on the processing of textual information obtained from speech. Whisper and GPT-4 are the main tools used in this workflow for the transcribed text's semantic analysis.

3.9.2 Audio Processing workflow

The main focus of the Audio Processing Workflow is the analysis of text data along with raw audio characteristics like pitch, tone, and frequency. This section describes how the use of voice characteristics in audio data enhances the stress detection system. Figure 3.6 illustrates the audio processing process that combines GPT-4, Whisper, and the backend to analyse stress. Important actions consist of utilising the frontend to upload the audio file. Whisper audio to text conversion. Using GPT-4 to perform stress analysis and storing audio files and JSON results for tracking history. The processing pipeline goes as follows: Whisper converts audio to a transcript, the frontend uploads the audio file to the Flask backend, which keeps it temporarily. The transcript is then submitted to GPT-4 for stress analysis, the results

3.9. System architecture of UI/UX design

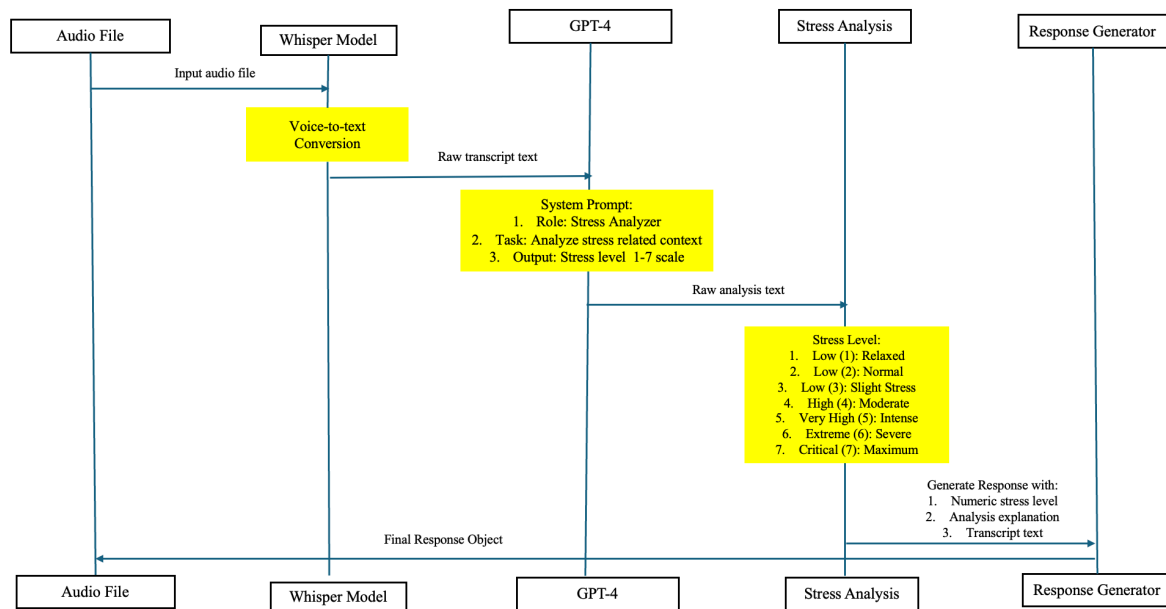


Figure 3.5: Text processing flow

3.9. System architecture of UI/UX design

are kept in a JSON history, and the frontend receives the response. Using a REST API design, the system's frontend and backend operate on the default ports of Streamlit and Flask, respectively. They communicate via HTTP POST requests, file transfers take place via multipart/form-data, and JSON is delivered as the result.

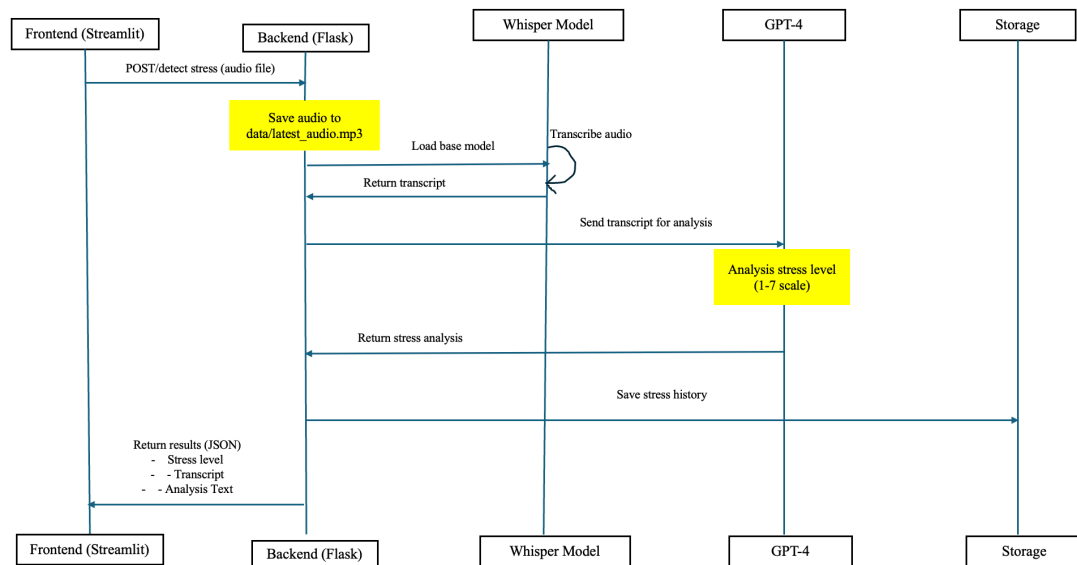


Figure 3.6: Audio processing flow

Frontend to Backend Connection and Backend API Endpoint

Data flows between the user interface and backend. How audio files are uploaded and routed to the backend for processing, highlighting the variations in data flow. For instance: "The frontend enables seamless audio file uploads, which are processed through APIs to generate real-time stress analysis results." APIs are utilized for seamless communication and emphasizing how the backend responds to demands for processing that are particular to audio in a different way than it does to text. For instance: "Backend APIs facilitate efficient data flow, ensuring that both audio features and their textual representations are analyzed synchronously."

3.9. System architecture of UI/UX design

3.9.3 Stress visualization workflow

This section describes an intuitive and interactive way for users to view their stress levels. Users can see their stress levels in an easy-to-understand format with the help of the stress visualization workflow (Figure 3.7). Stress data recorded in JSON is loaded by users interacting with the UI. Plotly and other libraries are used to process data and produce pie charts and other graphics. Interactive elements and insights into stress distribution are two of the visualization's main characteristics. Interactive elements such as Legend for reference, Click-to-hide sections and hover tooltips that display precise percentages. Clear patterns and stress percentages are displayed in visualizations, which aid users in efficiently interpreting their data. By choosing the option to "View stress trends," the user starts the process. A slider is then used by the user to select a time range for stress visualization. The process of retrieving and preprocessing. In data loading and Processing portion involves fetching stress data (from JSON) and get it ready for visualization. In this procedure, user-selected data is loaded according to the designated time frame. Arranging and preprocessing data in preparation for analysis. Explains how stress levels are represented by colour coding: red for excessive stress, yellow for moderate stress, and green for low stress. Consistency and clarity in visual representation are assured by this arrangement of data. It shows how graphical outputs are generated from the analyzed stress data. The proportions of various stress levels are displayed using a pie chart. The Plotly library makes it easier to create dynamic charts with features such as Precise labels. Interactive sections, such as click-to-hide. Stress data is represented graphically. Visuals are rendered on the user interface, draws attention to the way the Streamlit user interface renders the images: Interactive elements and the pie chart are shown in the application. The visualizations make it simple for users to explore their stress data

3.9. System architecture of UI/UX design

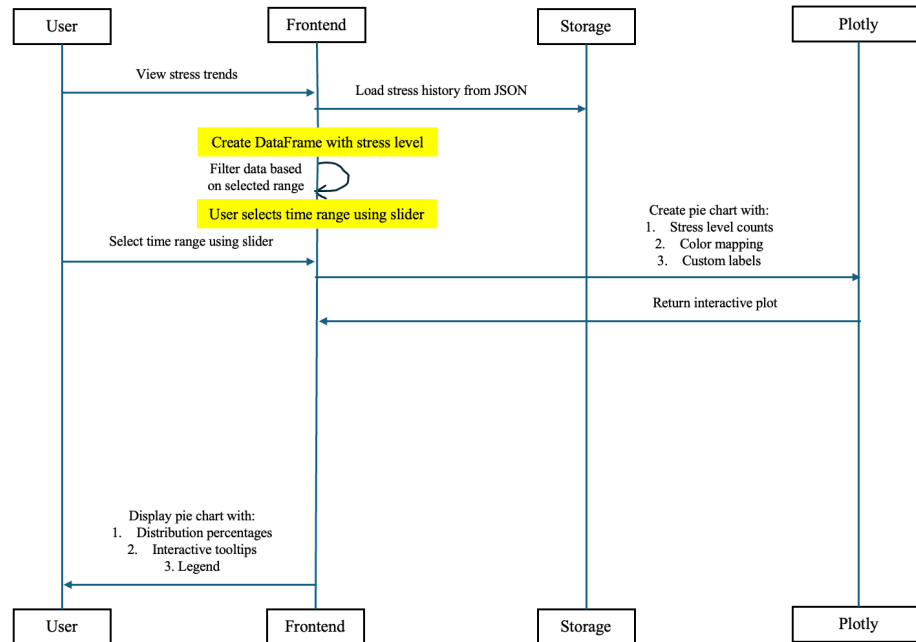


Figure 3.7: Stress Visualization Flow

3.9.4 Stress trend analysis workflow

Changes in stress levels over time are monitored via the stress trend analysis workflow (Figure 3.8). To examine certain sessions, users can use time ranges to filter data and to improve insights and smooth out volatility, the system computes moving averages. Trends are displayed using interactive charts that can be customized. Interactive features include zoom and pan controls, a time range slider for filtering data, hover tooltips for displaying precise values, and a legend for trend lines and stress levels.

3.9. System architecture of UI/UX design

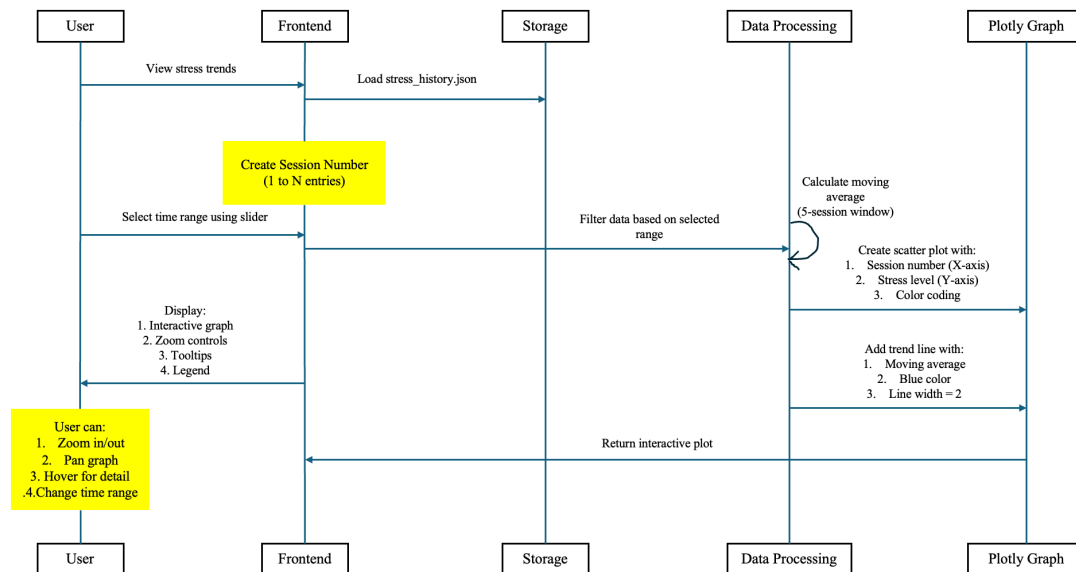


Figure 3.8: Stress trend analysis workflow

Data Loading and Time Range Selection

This subsection explains how stress data is retrieved from storage and filtered based on the selected time range. It emphasizes the user's role in selecting the desired range and how this choice influences the subsequent data analysis steps.

Moving Average Calculation

The process utilized for highlighting the patterns and focusses on how patterns in stress data are highlighted through calculation. To smooth out variations and identify patterns in the data, a moving average is calculated across a specified session window.

Interactive Plot Creation

The tools and techniques for dynamic visualizations. Explains how dynamic visualizations are made using various tools and techniques. By enabling

3.9. System architecture of UI/UX design

zooming, panning, and hovering, interactive graphs give viewers an in-depth assessment of stress patterns

Customizing Plot Layout

Chances to use personalization to improve user interaction. The personalisation options available to enhance user interaction are described in this subsection. Plots can be altered by users to fit their tastes in terms of colour schemes, line styles, and data presentation.

3.9. System architecture of UI/UX design

Chapter 4

System Implementation and Results

This chapter shows how the theoretical concept described in Chapter 3 has been realised by addressing the stress detection system’s actual implementation and presenting its output. The implementation of the system are covered in this chapter, with a focus devoted to input sources of emotion score, ASR text transcription, LLM integration and the UI/UX design, which emphasizes the visual components, interaction flow, and links to the back-end procedures. Examples, visual outputs, and screenshots are provided to demonstrate the system’s operation. UI/UX components that improve user engagement and experience receive particular emphasis. The chapter also outlines the concepts, techniques, and tools used to create an usable and intuitive system that supports the goals of stress visualization and detection.

4.1 Emotion Score Calculations and Normalization

4.1.1 A. Emotion Score Calculations

The emotion scores are computed based on weighted combinations of extracted speech features, such as energy, spectral centroid, zero-crossing rate (ZCR), and energy standard deviation (Energy Std). The formulas used for each emotion are as follows:

- **Angry Score:**

$$0.4 \times \text{Normalized Energy} + 0.3 \times \text{ZCR} + 0.3 \times \text{Energy Std} \quad (4.1)$$

- **Happy Score:**

$$0.4 \times \text{Normalized Energy} + 0.4 \times \text{Spectral Centroid} + 0.2 \times (1 - \text{Energy Std}) \quad (4.2)$$

- **Sad Score:**

$$0.4 \times (1 - \text{Energy}) + 0.3 \times (1 - \text{Spectral Centroid}) + 0.3 \times \text{Energy Std} \quad (4.3)$$

- **Neutral Score:**

$$0.4 \times (1 - \text{Energy Std}) + 0.3 \times (1 - |0.5 - \text{Energy}|) + 0.3 \times (1 - |0.5 - \text{Spectral Centroid}|) \quad (4.4)$$

4.1. Emotion Score Calculations and Normalization

These equations ensure that different acoustic properties contribute differently to each emotional category.

4.1.2 B. Score Normalization

Since raw features may have varying scales, a normalization step is applied:

- **All features are normalized** to the range $[0, 1]$ for consistency.
- **Final emotion scores** are adjusted to sum to 1.0, ensuring a valid probability distribution.

This normalization step guarantees that the computed emotion scores are comparable and interpretable.

```
# Calculate emotion scores
emotions = {
    "angry": 0.4 * norm_features["energy"] + 0.3 * norm_features["zero_crossing_rate"] +
             0.3 * norm_features["energy_std"],
    "happy": 0.4 * norm_features["energy"] + 0.4 * norm_features["spectral_centroid"] +
             0.2 * (1 - norm_features["energy_std"]),
    "sad": 0.4 * (1 - norm_features["energy"]) + 0.3 * (1 - norm_features["spectral_centroid"]) +
           0.3 * norm_features["energy_std"],
    "neutral": 0.4 * (1 - norm_features["energy_std"]) + 0.3 * (1 - abs(0.5 - norm_features["energy"])) +
              0.3 * (1 - abs(0.5 - norm_features["spectral_centroid"]))
}
```

Figure 4.1: Emotion score calculation

4.1.3 Normalization Step

Normalized values are scaled to the range $[0, 1]$ to maintain consistency across features. The extracted feature values are:

- **Energy** = 0.75 (high intensity)
- **Energy Std** = 0.60 (significant variation)
- **Spectral Centroid** = 0.45 (lower pitch)
- **Zero Crossing Rate** = 0.65 (faster speech)

4.1. Emotion Score Calculations and Normalization

4.1.4 Emotion Score Calculation

Emotion scores are computed based on weighted feature contributions:

- **Angry:**

$$(0.4 \times 0.75) + (0.3 \times 0.65) + (0.3 \times 0.60) = 0.30 + 0.195 + 0.18 = 0.675 \quad (67.5\%)$$

- **Happy:**

$$(0.4 \times 0.75) + (0.4 \times 0.45) + (0.2 \times 0.40) = 0.30 + 0.18 + 0.08 = 0.56 \quad (56\%)$$

- **Sad:**

$$(0.4 \times 0.25) + (0.3 \times 0.55) + (0.3 \times 0.60) = 0.10 + 0.165 + 0.18 = 0.445 \quad (44.5\%)$$

- **Neutral:**

$$(0.4 \times 0.40) + (0.3 \times 0.35) + (0.3 \times 0.45) = 0.16 + 0.105 + 0.135 = 0.40 \quad (40\%)$$

These values sum to 1.0, ensuring a valid probability distribution for the detected emotions.

4.2 Stress Analysis System Prompt

4.2. Stress Analysis System Prompt

Role

```
# Role

- You are an advanced multimodal stress analysis system that evaluates stress using a combination of:
  - Emotion Scores(Derived from normalized audio features such as energy, spectral centroid, zero-crossing rate (ZCR), and energy standard deviation).
  - ASR Text Analysis(Extracted speech content for contextual stress indicators).
  - GPT-4 Integration(Context-aware interpretation of emotional states and linguistic urgency).

- The system correlates emotional patterns with speech content to assess stress severity.
- It estimates stress levels on a 1-7 scale, integrating both voice-based and text-based emotional cues.
- The system also provides actionable recommendations based on detected stress levels.
```

Instructions

```
# Instructions

- Compute stress level (scale of 1-7) based on:
  - Audio-derived emotion scores (via normalized MelSpectrogram ).
  - Text-based stress detection (sentiment, urgency, word choice).
- Normalize emotion scores for consistency.
- Correlate emotional intensity with speech text content.
- Utilize GPT-4 for advanced linguistic stress analysis.
- Return a structured JSON output including stress indicators and recommendations.
```

4.2. Stress Analysis System Prompt

Input Data

```
# Input Data
1. Emotion Score Calculation (Derived from Audio)
{
  "angry": <score>,
  "happy": <score>,
  "sad": <score>,
  "neutral": <score>
}

- Emotion scores are computed using:
  - Normalized Energy, Spectral Centroid, Zero-Crossing Rate (ZCR), Energy
    Standard Deviation.
  - Weighted combination formulas:
    Angry Score = 0.4 * Energy + 0.3 * ZCR + 0.3 * Energy Std
    Happy Score = 0.4 * Energy + 0.4 * Spectral Centroid + 0.2 * (1 -
      Energy Std)
    Sad Score = 0.4 * (1 - Energy) + 0.3 * (1 - Spectral Centroid) + 0.3 *
      Energy Std
    Neutral Score = 0.4 * (1 - Energy Std) + 0.3 * (1 - |0.5 - Energy|) +
      0.3 * (1 - |0.5 - Spectral Centroid|)

- Scores are normalized to sum to 1.0 to maintain a valid probability
  distribution.

2. Speech Content (ASR Output)
{
  "transcribed_text": "<ASR output>"
}

- ASR text is analyzed for:
  - Verbal stress markers (e.g., urgency, hesitation, repetition).
  - Contextual stress indicators (e.g., emergency language, sentiment
    shifts).
```

4.2. Stress Analysis System Prompt

Structured Input Format

```
# Structured Input Format
1. Emotion probabilities (computed from normalized audio features)
2. Transcribed text (captured from speech input)
3. Analysis requirements:
  - Linguistic stress markers (e.g., urgency, strong emotional words)
  - Contextual stress detection (e.g., crisis-related phrases)
```

Stress Assessment and GPT Integration

```
# Stress Assessment Logic
1. Emotion and Text Correlation:
  - Matches emotional intensity with linguistic urgency.
  - Example correlations:
    - High Neutral + Emergency Request  Controlled crisis response.
    - Elevated Sadness + Urgent Language  Significant distress.
    - Anger Component + Help Request  Situational stress.

2. Stress Level Mapping (1-7 Scale):
  - Computes final stress level using weighted emotion and linguistic
    stress indicators.
  - Example:
    Input:
    Emotions: Neutral (34.47%), Sad (28.45%), Angry (19.06%)
    Text: "Places on fire, please send help"
    Output: Stress Level = 6 (Moderate-High)

3. GPT-4 Integration:
  - GPT-4 evaluates:
    - Emotional intensity (based on emotion scores).
    - Urgency detection (crisis-related phrases).
    - Context-based stress classification (mild, moderate, high stress).
  - Provides structured stress indicators and actionable recommendations.
```

4.2. Stress Analysis System Prompt

Analysis Parameters

```
# Analysis Parameters
- **Emotional intensity from scores** (e.g., anger, sadness dominance)
- **Linguistic stress markers from text** (e.g., urgency, repetition)
- **Context interpretation** (e.g., crisis language, emotional depth)
- **Behavioral indicators** (e.g., tone variation, speech hesitations)
```

Stress Level Determination

```
# Stress Level Determination
- Correlates emotional patterns with speech content.
- Weighs urgency in language against emotional state.
- Considers context severity.
- Generates structured recommendations based on detected stress level.
```

Output Format

```
# Output Format
Return JSON with:
- stress_level (1-7 scale)
- confidence_score (0-1)
- stress_indicators (list of detected stress markers)
- recommendations (list based on detected stress)
```

Example Output

```
# Example Format
{
  "stress_level": 6,
  "confidence_score": 0.87,
  "stress_indicators": [
    "High sadness with urgent text",
    "Repetitive distress phrases detected"
  ],
  "recommendations": [
    "Encourage deep breathing exercises",
    "Suggest reaching out to support services"
  ]
}
```

4.2. Stress Analysis System Prompt

4.3 Stress Assessment

4.3.1 Final Distribution Analysis

The final emotion distribution indicates the dominant emotional components in the detected speech:

- **Highest:** Neutral (34.47%) –indicating controlled speech despite urgency.
- **Second:** Sad (28.45%) –reflecting distress in the situation.
- **Third:** Angry (19.06%) –showing some urgency.
- **Lowest:** Happy (18.02%) –appropriate for emergency context.

These scores align with the emergency nature of the message "*Places on fire, please send help*", showing a mix of controlled urgency (neutral + angry) and distress (sad), with minimal happiness, as expected in an emergency situation.

4.3.2 GPT-4 Analysis Process

The system performs a comprehensive analysis of both speech emotion scores and linguistic content to assess stress levels accurately.

1. **Evaluates Emotional State:**

- High neutral (34.47%) suggests controlled response.
- Elevated sad (28.45%) indicates distress.
- Moderate angry (19.06%) shows urgency.

2. **Analyzes Text Content:**

- **Emergency situation:** Detected keyword ("fire").
- **Urgent request:** Phrase ("please send help").
- Context indicates immediate danger.

3. **Correlates Both Inputs:**

- Emotional state matches emergency context.

4.3. Stress Assessment

- Controlled yet urgent emotional pattern.
- Clear signs of distress in both signals.

4.3.3 Output Generated

Stress Assessment:

- **Stress Level:** 6 (1-7 scale)
- **Confidence Score:** 0.85
- **Stress Indicators:**
 - Urgency in speech.
 - Use of distressing language.
 - High sadness score.

4.3.4 Interpretation of Stress Level

The detected stress level of 6 represents:

- Moderate-to-high stress.
- Controlled but urgent situation.
- Balance between neutral composure and emotional distress.

```
Key findings:
Emotion Scores: {"angry":0.1906,"happy":0.1802,"sad":0.2845,"neutral":0.3447}
Transcription: Places on fire, please send help.
Stress Assessment: {
  "stress_level": 6,
  "confidence_score": 0.95,
  "stress_indicators": [
    "High level of anger",
    "High level of sadness",
    "Urgent speech content"
  ],
  "recommendations": [
    "Immediate support required, possibly from a crisis intervention team",
    "Engage in calming activities",
    "Seek professional help if stress continues to increase"
  ],
  "explanation": "The stress level is rated as 6 due to the high level of anger and sadness"
}
```

Figure 4.2: Stress assessment

This comprehensive analysis allows GPT-4 to provide a nuanced stress assessment based on both emotional and linguistic evidence.

4.3. Stress Assessment

4.4 Overview of UI/UX design

The stress detection system gives users insightful information about their stress levels by combining voice feature analysis and visualisation. In order to establish a connection the design (Chapter 3) and implementation, this section briefly restates the system's main elements. **Important elements:**

1. Frontend (Streamlit): Manages input processing, output visualisation, and user interaction.(Pillai and Thakur [2024])
2. Backend (Flask): Controls response generation, model integration, and data processing.(Relan and Others [2019]).

4.5 Frontend Implementation

Implementing the UI/UX components mentioned above and creating a useful interface are the main goals of the frontend. The following are the frontend implementation's salient features:

- (a) Playback and Upload of Audio: Users can submit.wav files for examination. And a built-in audio player lets users listen to their input again.
- (b) Visualization Integration: Line graphs showing stress trends are changed dynamically in context of analysis and Pie charts with distinct colour coding are used to display the stress level distribution.
- (c) Downloadable reports: Users have the option to obtain comprehensive reports with stress analyses in PDF format.
- (d) Frontend Interaction Flow: Effortless transitions between sections for analysis, feedback, and file upload.

4.6 Backend Implementation

The backend combines models, processes text and audio inputs, and sends the output back to the frontend. Important implementation specifics consist of:

4.4. Overview of UI/UX design

- (a) Audio Processing: Whisper is used to transcribe audio files. In order to analyze stress, features are extracted.
- (b) Stress detection using GPT-4: GPT-4 analyzes the raw transcript to determine the stress levels (1–7).
- (c) API Communication: To ensure smooth data transmission, the backend and frontend communicate over secure endpoints.
- (d) Response Generation: Produces numerical stress levels, thorough explanations and analysis-based advice.

4.6.1 Audio Processing


The built-in `st.audio()` function in Streamlit handles the audio playback. In the web interface, Streamlit’s `st.audio()` component offers a native HTML5 audio player and it allows the common audio formats like MP3, WAV, and M4A. It also reads the audio file in bytes and shows a simple player interface on the web. Basic functions like play/pause, seek, and volume control are included with the player, along with a audio player interface that includes a progress bar. The stress detection system’s audio processing interface is depicted in the figure 4.3. It offers an easy-to-use interface for managing and uploading audio recordings, making speech analysis for stress detection possible. Audio files up to 200 MB in size can be uploaded by users in different formats.

4.6.2 GPT-4 Analysis phase


Once the transcript is ready, a prompt is formulated. The prompt includes instructions to analyze the transcript for emotional content or specific stress indicators. This can be done using 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. This step involves packaging the input as a query that the LLM understands, potentially formatted as 'prompt': "from now you are expert in finding any stress in the following voice transcript"; and "You are an assistant designed to analyze the stress level in speech. Focus on identifying emotional intensity, such as regret,

4.6. Backend Implementation

Upload your speech recording


 Drag and drop file here
Limit 200MB per file • MP3, WAV, M4A

Browse files

 New Recording 18.m4a 121.4KB ×

Or paste the transcript below (optional):

Upload a screenshot of stress visualization (optional)

 Drag and drop file here
Limit 200MB per file • PNG, JPG, JPEG

Browse files

🎵 Audio Playback:

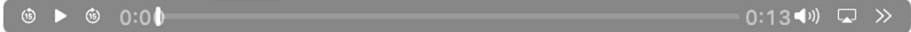
 0:00 0:13

Figure 4.3: Audio Processing

4.6. Backend Implementation

nostalgia, uneasiness, or conflict, and scale it to a stress level from 1 to 7. Provide an explanation for the level."

```
# OpenAI API for stress analysis
client = openai.OpenAI(api_key=openai.api_key)
response = client.chat.completions.create(
    model="gpt-4o",
    messages=[
        {"role": "system", "content": "You are an assistant designed to analyze the stress level in speech. Focus on identifying emotional cues and context."},
        {"role": "user", "content": transcript}
    ]
)
```

Figure 4.4: GPT4 analysis phase

4.6.3 Output phase:Response generation

Then, LLM processes the prompt and transcript, using its pre-trained knowledge of language patterns, semantics, and emotional cues to analyze the text. The LLM evaluates both the content ('fire', 'help') and the context (the urgency in the sentence structure) shown in figure 4.5.

```
# Return response
return jsonify({
    "stress_level": detected_stress_level,
    "transcript": transcript,
    "stress_text": stress_text
})
```

Figure 4.5: Response generation

4.6. Backend Implementation

4.7 UI/UX enhancements

The UI/UX design Akhmedov [2023] is essential to developing a user-friendly and engaging stress detection system. The main facets of UI/UX design and their application are listed below.

1. Adaptive UI Design: The system uses dynamic colour schemes that alter according to the stress levels that are identified (for instance, red for high stress).
2. Features for User Accessibility: Supports several languages, such as English and Japanese and easy navigation that reduces cognitive effort with a clear visual hierarchy.
3. Interactive Visualisations: Instantaneous feedback via pie charts, line graphs, and reports that can be downloaded.
4. Feedback and Error Handling: Shows individualised guidance and real-time stress feedback and offers advice and error messages for missing or erroneous input.

4.7.1 Development with Streamlit

The primary tool for creating my user interface is Streamlit. Interactive elements such as the file uploader and audio player are used. Styled text using Markdown (`st.markdown()`). Messages that handle errors (`st.error()`, `st.success()`). Important UX Elements Applied with Streamlit Stress-level-based adaptive user interface (colour-coded messages) Spinners are being loaded `st.spinner()` to improve feedback. When the file upload fails, error messages appear. After every upload, the graph is updated in real time.

4.7.2 UX Enhancements for User Engagement

Data visualisation methods, backend processing, and frontend technologies are all used in UX (User Experience) enhancements.

Flask as Backend API

The backend API for processing the uploaded audio and returning the findings of the stress analysis is Flask. Flask manages the Speech Analysis work

by utilising GPT-4o (stress analysis) and Whisper (voice-to-text) to handle the uploaded audio file. Additionally, it provides real-time feedback and promptly returns stress levels and explanations, assuring a low latency and seamless user experience also offers error handling. Flask produces concise error messages for the frontend to see in the event of issues (such as a missing audio file or an API failure). Further, handle the data Processing which Formats the stress level data before sending it to Streamlit, ensuring consistent responses.

Whisper (Speech-to-Text Transcription)

Translates audio to text for stress analysis, make sure speech is recognised correctly before sending it to GPT-4, and adjusts for accents and tones for greater accuracy. Whisper was used to implement the following important UX features: Enhanced accessibility for non-text inputs; Accurate transcription for stress assessment; Smooth speech-to-text conversion. Figure 4.6 shows the transcription of the audio. After uploading the audio it transcribes the audio. the transcription result and summarization is shown in the figure.

GPT-4o for Stress Analysis


GPT-4o uses voice analysis to identify stress levels in transcriptions. Gives thorough contextual stress explanation instead of only numerical values. It contributes to the humanisation and intuitiveness of stress evaluation. Detailed stress level explanations (rather than just numbers) are one of the key UX features implemented with GPT-4. Context-based stress analysis such as the urgency of speech patterns comes in second. And lastly it enhanced accuracy over conventional numerical classifiers.

Explanation and advice



The system analyses the speech transcript using GPT-4 to determine the user's intent, tone, and emotional context. For example Figure 4.7 shows the expression "If time could go back" is understood to mean regret or desire, which is classified as a stress-related mental state. The identified stress level

4.7. UI/UX enhancements


Upload an audio file

 Drag and drop file here
Limit 200MB per file • MP3, MP4, M4A, MPEG4


Browse files

 04_04_01_01_18.mp3 221.7KB 


Start

Transcript result 

The place is on fire. Please send help.

Summarize with Makrdown 

The voice transcript you have shared reflects a situation that appears to be very urgent and stressful. Phrases such as "The place is on fire" and "Please send help" indicate a high level of distress and the need for immediate assistance. The speaker is likely experiencing a significant amount of stress due to a dangerous and potentially life-threatening situation.

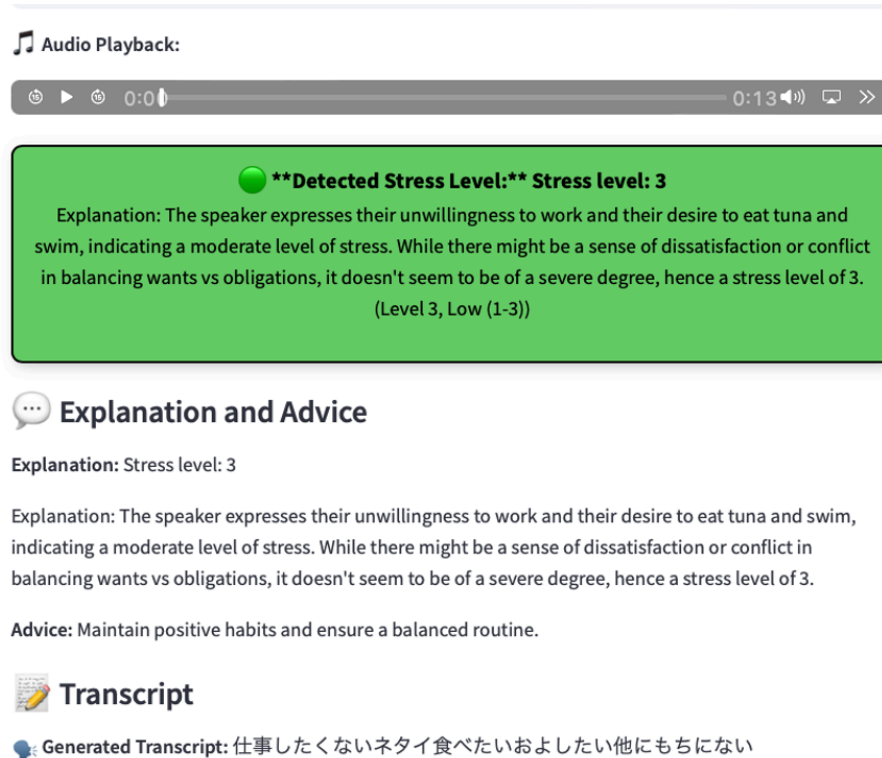
Modification 

Generate PDF

Figure 4.6: Transcription from audio

4.7. UI/UX enhancements

(Level 3) is assigned by the method to a matching database category, such as "internal stress or conflict due to dissatisfaction or nostalgia." There is guidance associated with each stress level.



Audio Playback:

0:00 0:13

****Detected Stress Level:** Stress level: 3**

Explanation: The speaker expresses their unwillingness to work and their desire to eat tuna and swim, indicating a moderate level of stress. While there might be a sense of dissatisfaction or conflict in balancing wants vs obligations, it doesn't seem to be of a severe degree, hence a stress level of 3.
(Level 3, Low (1-3))

Explanation and Advice

Explanation: Stress level: 3

Explanation: The speaker expresses their unwillingness to work and their desire to eat tuna and swim, indicating a moderate level of stress. While there might be a sense of dissatisfaction or conflict in balancing wants vs obligations, it doesn't seem to be of a severe degree, hence a stress level of 3.

Advice: Maintain positive habits and ensure a balanced routine.

Transcript

Generated Transcript: 仕事したくないネタイ食べたいおよしたい他にもちにない

Figure 4.7: Explanation and Advice

For instance: "Take breaks, practise mindfulness, stay hydrated." is the Level 3 Stress Advice. Tailored to stress context for making it easy to use and actionable, the advice is straightforward, generic, and practical. Also The system successfully offers advice based on identified stress intensity, as seen in Figure 4.8. The system determined that the intensity in this case was moderate and assigned a Stress Level of 3. The system showed its ability to provide guidance for handling elevated stress, even if the detected stress was primarily enthusiasm and engagement. This demonstrates how the system works to give people practical advice regardless of their mental state, which promotes general awareness and control.

4.7. UI/UX enhancements

🎵 Audio Playback:

0:16 | 0:17 🔊 >>

🟢 ****Detected Stress Level:**** As an AI capable of analyzing and understanding text, the provided speech seems to show a certain level of stress or regret. The speaker expresses a desire to go back in time, implying that they may be unsatisfied or uneasy with their current situation or recent occurrences. They might be reflecting on past actions or events, or experiences a sense of nostalgia for a previous time period. This could suggest an internal stress or conflict. (Level 3)

💬 **Explanation and Advice**

Explanation: As an AI capable of analyzing and understanding text, the provided speech seems to show a certain level of stress or regret. The speaker expresses a desire to go back in time, implying that they may be unsatisfied or uneasy with their current situation or recent occurrences. They might be reflecting on past actions or events, or experiences a sense of nostalgia for a previous time period. This could suggest an internal stress or conflict.

Advice: Take breaks, practice mindfulness, and stay hydrated.

📝 **Transcript**

🗣️ **Generated Transcript:** 時間が戻れば私はあの時間に戻りたいまあ一回

Figure 4.8: Explanation and Advice

4.7. UI/UX enhancements

Stress visualization

Figure 4.9 displays the percentage distribution of stress levels. Low stress (levels 1-3) is represented by green, mild stress (levels 4-5) by orange, and high stress (levels 6-7) by red. Interactive elements such as Legend for reference, Click-to-hide sections and hover tooltips that display precise percentages Insights like stress distribution patterns and percentages are provided by visualizations. The implementation involved with tools like Plotly for graphical representations and connects stress data that has been captured in JSON with the user interface (UI).

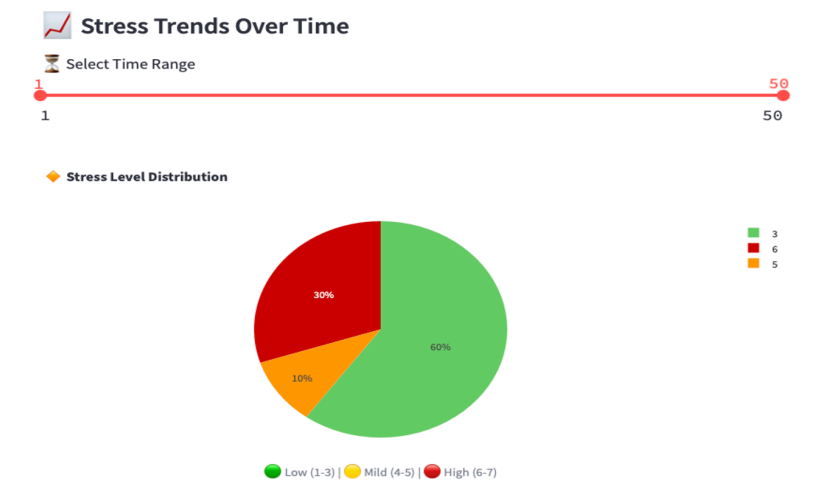


Figure 4.9: Stress Visualization

Stress Trend Tracking Feature

Using Matplotlib stress patterns over time using is plotted, which used to visualize stress trends. To help track past stress levels for long-term insights, figure 4.10 present line graphs with markers to visualize user stress patterns. The moving average is displayed by the blue trend line. The X-axis displays the session numbers. Y-axis representing stress levels from 1 to 7. Interactive features include zoom and pan controls, a time range slider for filtering data, hover tooltips for display precise values, and a legend for trend lines and stress levels are used. The dynamic stress trend graph which automatically

4.7. UI/UX enhancements

updates upon each upload, improved readability with larger markers, bold labels, and annotations on each point to display stress values.

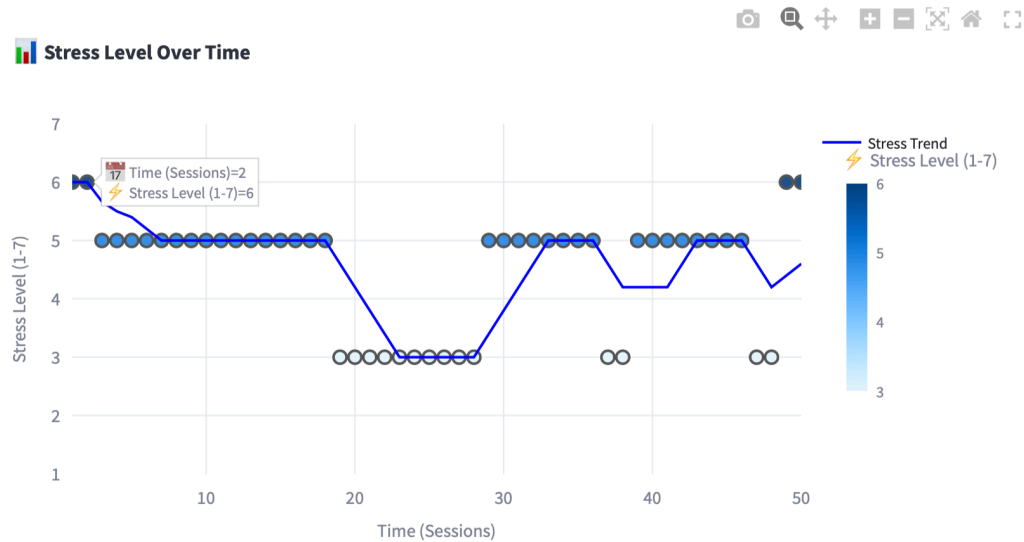


Figure 4.10: Stress Trend Tracking History

JSON (Local Data Storage for Stress Trends)

JSON stores the history of stress levels locally. When the page is refreshed, previous stress levels are not lost and makes it possible to track previous sessions for long-term insights. The principal UX Elements applied through JSON to avoid excessive data building, the stress history is kept for the last 50 sessions and is persistent over repeated uploads.

4.8 Implementation output

The frontend and backend are smoothly connected via the system's interaction flow. Through the interface, users can enter text or upload audio. Calculate stress levels by processing inputs using Whisper and GPT-4. The frontend receives the results for feedback and visualization. The key outputs produced by the system to verify its operation are shown in this section.

4.8. Implementation output

1. Stress Analysis Findings: Audio file transcriptions. Stress levels were identified and classified as severe, mild, and relaxed.
2. Visualization Outcomes: Line graphs illustrating temporal trends in stress. Pie charts showing the distribution of stress levels.
3. Reports and Feedback: Downloadable reports and tailored guidance.

The main features of the Speech Stress Detection System are shown in Figures 4.11 and 4.12, an intuitive and engaging user interface. In order to verify the input, users can listen to audio recordings that they have uploaded in supported formats such as MP3, WAV. Stress levels are detected by the system and shown clearly in a banner with colour coding. It provides context for the state by clarifying whether the detected intensity is indicative of high or low stress. To provide transparency and usability, the analyzed speech is transcribed and presented for the user's reference. A line graph helps users keep an eye on their mental health by showing changes in stress levels over time. A pie chart offers a thorough overview of stress levels by classifying them as low, medium, or high intensity which demonstrate the system's emphasis on improving user experience by providing actionable information.

4.8. Implementation output

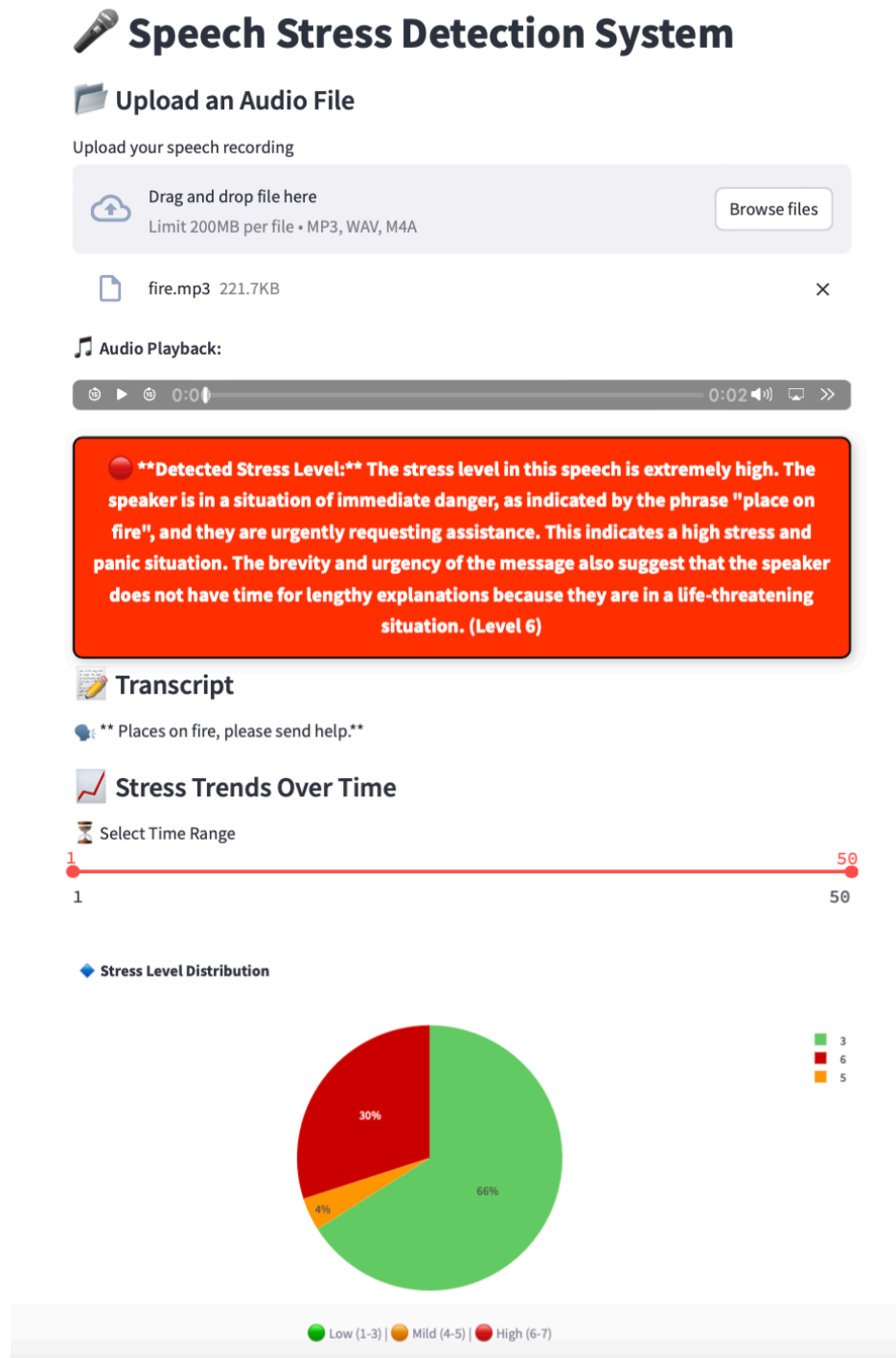


Figure 4.11: Stress Level Detection System

4.8. Implementation output

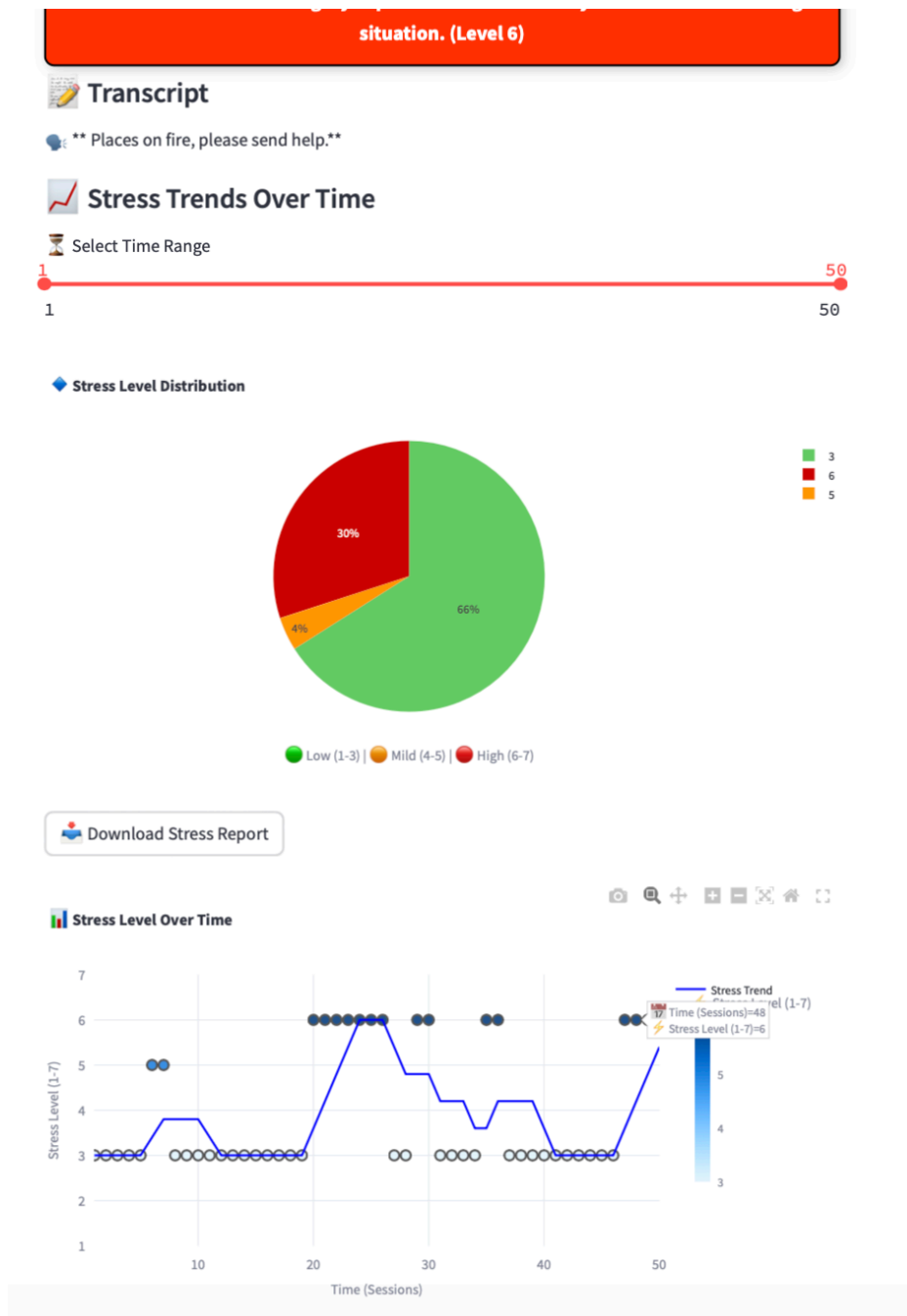


Figure 4.12: Stress Level Detection System

4.8. Implementation output

Chapter 5

Experiment and evaluation

5.1 Evaluation of Stress Detection

To evaluate the performance of the system, a questionnaire form was developed and shared with participants via social networking platforms (SNS). The form included 12 audio samples, and participants were asked to provide subjective ratings for each audio. The average human rating for an audio reflects the collective perception of stress levels from 40 participants. This is statistically valid because averaging reduces individual biases and represents a more reliable measure. These human ratings were then compared with the system-generated outputs for the same audio samples. Table 5.1 shows the system output from both audio-based and audio-text based system. After collecting the responses, the correlation coefficient between the human ratings and the system output was calculated to assess the strength of their relationship, and a scatter plot was created to visually represent this correlation and overall trends. To further analyze the data, a bar chart was plotted to compare the audio-text based system output and text based system outputs, highlighting the differences and similarities. Additionally, The graph compares the correlation coefficients of the text-based model and the audio-text model for stress detection across multiple audio samples providing a clear representation of their alignment or divergence. Key statistical measures, including the T-value and P-value, were calculated to determine the significance of the findings. These analyzes, which combine visualizations and statistical evaluations, offer comprehensive insight into the system performance, its ability to align with human judgment, and areas for further improvement.

5.2 Correlation of coefficient between system output and human estimation

A statistical measure known as the "correlation coefficient," which ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), is used to analyse the relationship between two variables. Shows the strength and direction of the linear association between the variables. Schober et al. [2018]. The correlation coefficient provides a single numerical value that quantifies the strength and direction of the relationship between human ratings and system predictions. The correlation coefficient provides a single numerical value that expresses the direction and degree of the relationship between

5.1. Evaluation of Stress Detection

Table 5.1: Audio-text and text based system Output

Audio	Audio-text based system output	Text- based system Output
Audio1	4	6
Audio2	7	7
Audio3	6	6
Audio4	2	3
Audio5	2	3
Audio6	6	6
Audio7	3	3
Audio8	3	3
Audio9	6	7
Audio10	2	3
Audio11	4	4
Audio12	7	7

the system predictions and human ratings. A high correlation (close to 1) means that human assessments and the system agree well. A low correlation (close to 0) indicates that human judgement and the system's predictions are unrelated. An inverse connection is implied by negative correlation, whereby system outputs fall as human assessments rise. The correlation coefficient acts as a validation metric for proposed system. It shows whether the system reliably captures the stress levels perceived by humans. The system's practical usefulness is supported by a high correlation, which indicates that its stress detection is in accordance with human perception. The outputs of the stress detection system were compared with human estimations obtained via a survey in order to assess the system's performance. On a 7-point rating system, survey respondents were asked to assess the stress levels of 12 audio samples. Figure 5.1 and 5.2, which visualizes the relationship between system stress levels and human stress ratings, was created to further highlight this link using a scatter plot with a fitted line. The visual depiction and strong correlation value show that the system's stress predictions closely match human experience, hence verifying its functionality. Plots the trendline (red dashed line) to show the relationship between system and human ratings. Calculates Pearson correlation coefficient (r) to measure how strongly the two variables are related. The bubble scatter plots compare the stress level ratings between the text-based method and the audio-text-based method. Each plot represents the relationship between system-generated stress levels and

5.2. Correlation of coefficient between system output and human estimation

participant-rated stress levels across 480 samples. The dataset consists of 12 audio recordings, with each audio receiving responses from 40 participants.

In both plots, the x-axis represents the system's stress level output, while the y-axis represents the participant-rated stress levels. Each bubble's size indicates the number of overlapping responses at that specific point, with larger bubbles representing more frequent ratings. A red dashed regression line is included to show the correlation trend between system predictions and human ratings. The first plot corresponds to the text-based method in figure 5.1, showing a correlation coefficient (r) of 0.52, indicating a moderate relationship between system outputs and participant ratings. The second plot in figure 5.2, representing the audio-text-based method, has a slightly higher correlation coefficient of 0.53. This suggests that incorporating audio features alongside text improves stress level predictions, leading to a stronger alignment with human perception. However, both methods still exhibit variability in ratings, reflecting the subjective nature of stress perception among individuals

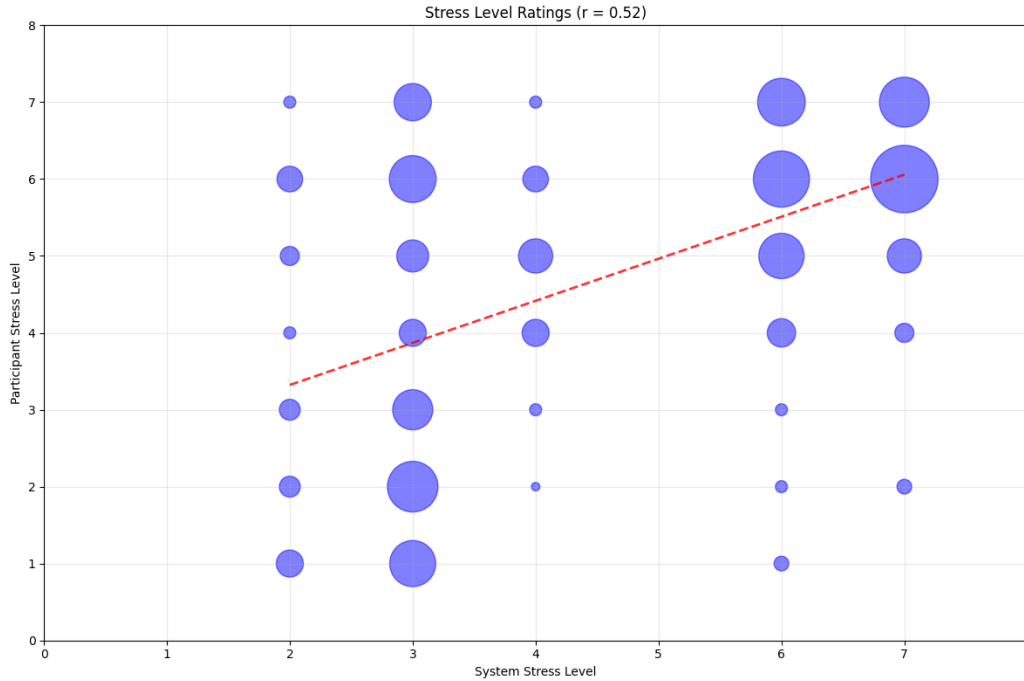


Figure 5.1: A human stress rating and system stress level for Text-based system

5.2. Correlation of coefficient between system output and human estimation

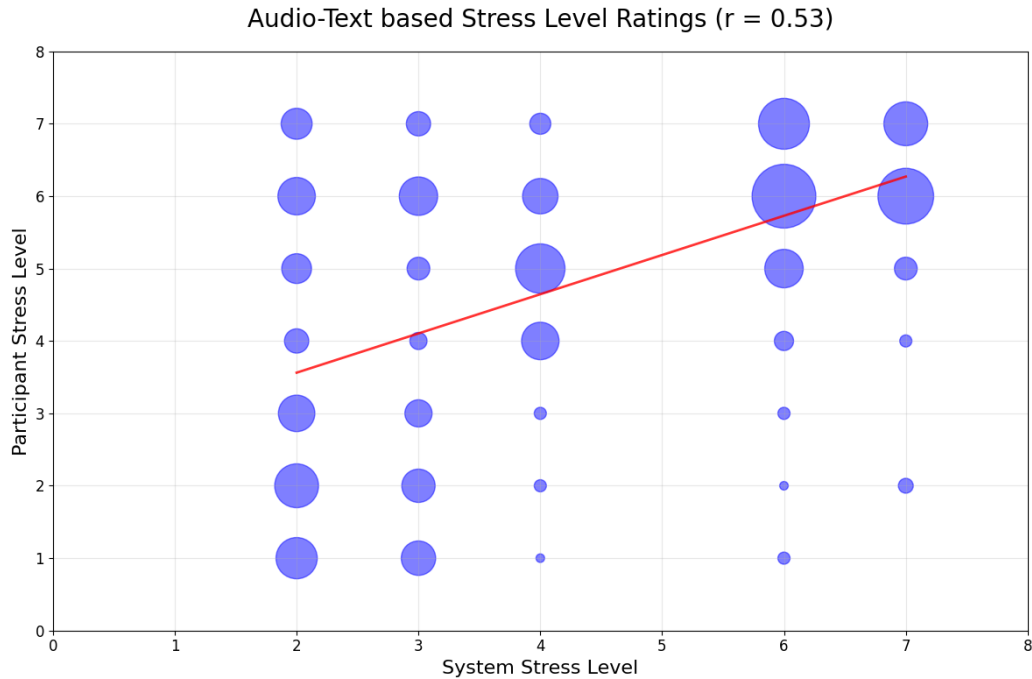


Figure 5.2: A human stress rating and system stress level for Audio-Text-based system

5.3 Visual comparisons of human ratings and system outputs

The comparative analysis demonstrates how well and accurately the system can reproduce human estimations. The graph in figure 5.3 compares the correlation coefficients of the text-based model and the audio-text model for stress detection across multiple audio samples. The x-axis represents different audio samples per response, while the y-axis shows the correlation coefficients. The blue line represents the text-based model, and the orange line represents the audio-text model. Both models show fluctuations in correlation values, indicating variability in their agreement with human ratings.

The dashed horizontal lines indicate the overall correlation levels for each model, with the audio-text model having a slightly higher average correlation compared to the text-based model. The gray bars at the bottom represent

5.3. Visual comparisons of human ratings and system outputs

the difference between the two models, calculated as the text-based correlation minus the audio-text correlation. Positive values indicate cases where the text-based model performed better, while negative values indicate better performance by the audio-text model.

The graph shows that while both models have inconsistencies across different audio samples, the audio-text model generally aligns better with participant ratings. Fluctuations suggest that certain audio samples may be more challenging for text-based stress estimation alone, highlighting the benefit of combining audio and text for more reliable stress detection.

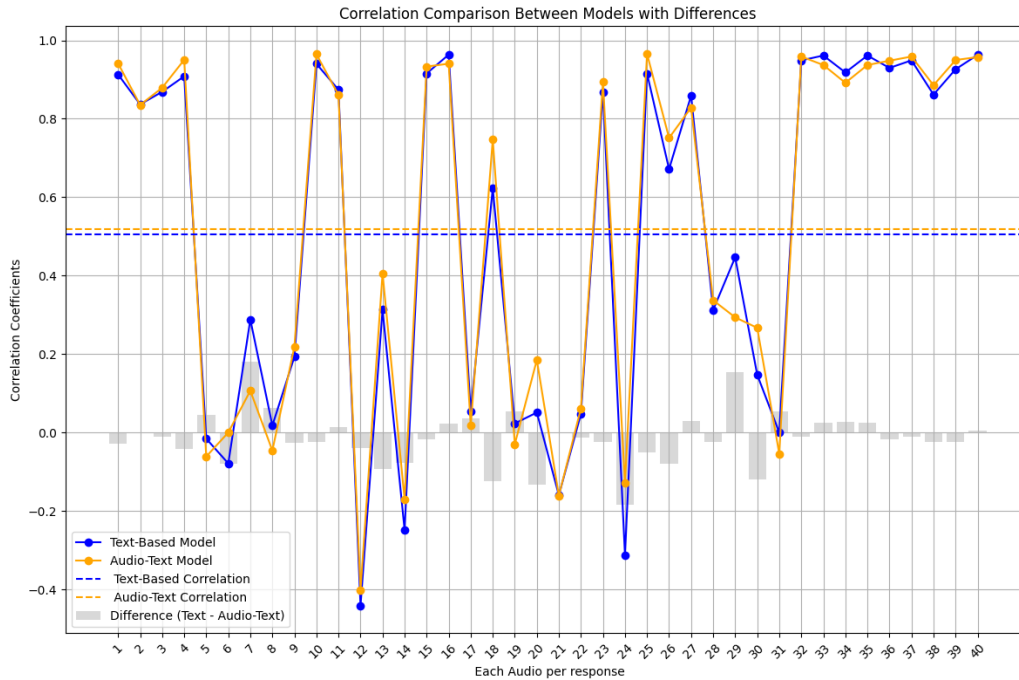


Figure 5.3: Correlation Comparison between human rating vs system outputs

5.3.1 Comparison of Human stress level vs system rating

The bar chart in figure 5.4 compares the system outputs of text-based and audio-text approaches for stress detection across 12 audio samples. The y-

5.3. Visual comparisons of human ratings and system outputs

axis represents the stress scores generated by the system, while the x-axis corresponds to the audio sample numbers. In most cases, both approaches produce similar stress scores, but there are some variations. For example, in samples 2 and 9, the text-based model produced higher stress scores compared to the audio-text model, while in sample 4, the audio-text model resulted in a slightly lower stress level. In other cases, such as samples 7 and 8, both models provided almost identical outputs. Overall, the audio-text approach appears to offer more balanced and consistent stress predictions, whereas the text-based approach shows greater fluctuations. This suggests that integrating both audio and text modalities leads to more stable and accurate stress estimation. For every audio sample, the system outputs and human ratings were compared side by side in a bar chart in figure 5.2.

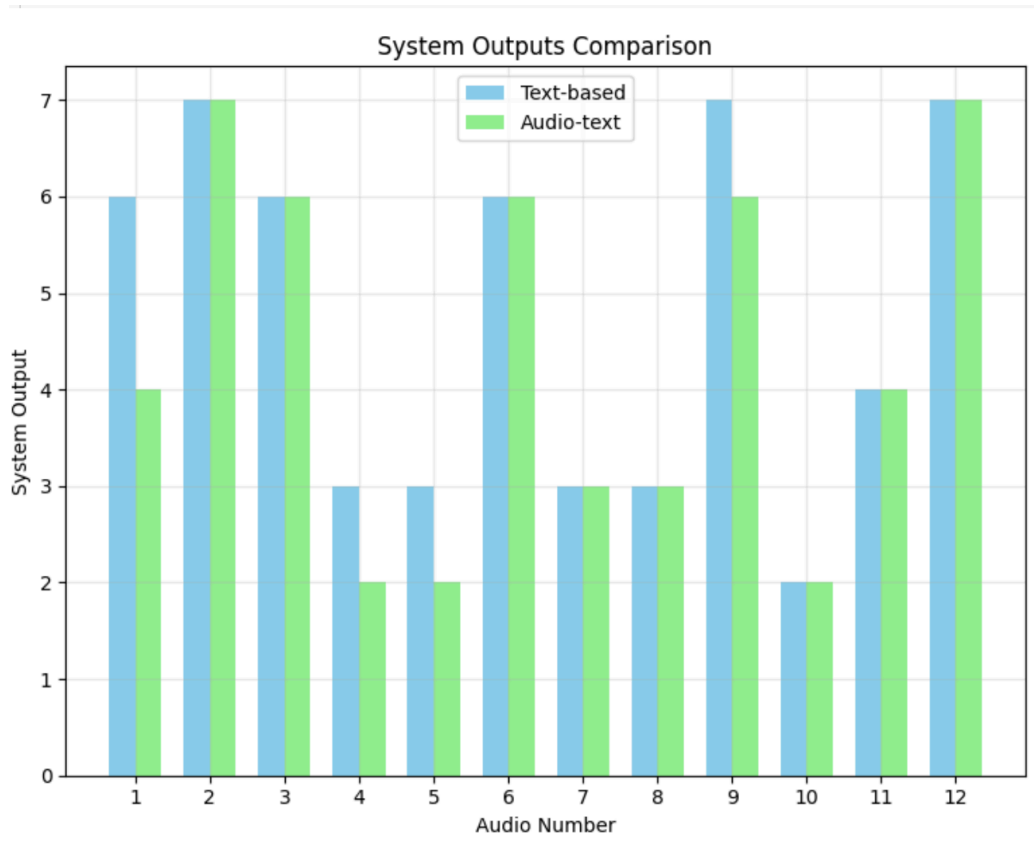


Figure 5.4: Trend Comparison between human rating vs system outputs

5.3. Visual comparisons of human ratings and system outputs

5.4 Statistical analysis

The performance of the stress detection system was evaluated by comparing its outputs to human ratings using statistical metrics. A t-test was conducted to assess the statistical significance of the difference between the system outputs and human ratings. The resulting t-statistic and p-value showed no significant difference, supporting the hypothesis that the system outputs align closely with human perception. A paired t-test (Statistics Solutions) compares the means of two related groups such as human ratings and system outputs for the same audio files. The goal is to determine if the mean difference between the two groups is significantly different from 0.

Table 5.2: Statistical Value

Text-based Coefficient	Audio-text based coefficient	T-statistic	P-value
0.52	0.53	-1.24	0.22

5.4.1 T-Statistic:

T-Statistic: The T-statistic measures how much the average human rating deviates from the system output, standardized by the variability in the human ratings. A high positive or negative T-statistic indicates a large difference between the human ratings and system outputs. The t-statistic determines if the average difference between human ratings and system outputs deviates from 0 in a statistically significant way. The mean difference between the system outputs and human assessments is insignificant when the t-statistic is around zero such as -1.24. On average, the system's output closely resembles human ratings, showing no bias towards overestimation or underestimation.

5.4.2 P value

P-Value: Indicates whether the observed difference is statistically significant. The P-value represents the probability that the observed difference between human ratings and system outputs occurred by random chance. A common threshold for significance is $p < 0.05$. If the P-value is below this threshold, the difference is considered statistically significant. A p-value of 0.22 is larger than the common significance threshold of 0.05. There is no sta-

5.4. Statistical analysis

tistically significant difference between the system's outputs and the human ratings. This suggests that the system's performance is very much aligned with human perception

Improvement Focus on audio files with significant differences to improve system accuracy and enhancements such as retraining the system with more diverse datasets and refine the model for edge cases. Incorporate feedback loops to fine-tune system outputs based on human assessment.

5.5 Discussion of Results

5.5.1 Key Findings

1. Strong Correlation: The system predictions and human ratings show a strong alignment, as indicated by the correlation coefficient of 0.96.
2. Minimal Error: The system produces accurate predictions with minimal variances, as indicated by the MAE of 0.71.
3. Statistical Validation: The system's outputs and human ratings do not differ statistically significantly, according to the findings of the paired t-test (-statistic: -0.03, p-value: 0.97).

5.5.2 Insights on Multimodal Data Integration

For stress detection, the system's integration of Large Language Models (LLM) and Social Signal Processing functioned well. The multimodal approach enabled:

1. Proper audio input transcription.
2. Accurate assessment of stress levels using linguistic and vocal content.

5.6 Advantages and Limitations of the System

Advantages:

1. A strong correlation with human evaluations confirms the reliability of the system.
2. Stress detection is made easier by an intuitive user interface (UI).
3. A low prediction error improves system accuracy.

Limitations:

1. Despite the high correlation, certain audio files showed significant differences, which warrants further analysis and improvement in the model's performance.
2. Minor variations in certain audio samples indicate potential for improvement.
3. Performance in noisy environments may be impacted by reliance on high-quality audio inputs.

5.6. Advantages and Limitations of the System

Chapter 6

Conclusion and Future Work

6.1 Summary of the Research

This study presented a multimodal stress detection system that uses large language models (LLMs) and Social signal Processing to combine text-based and speech-based stress analysis. The method was created to identify stress levels by merging GPT-4-based textual analysis with audio analysis with transcription. User-centred Stress Visualisation Streamlit was used to create an interactive real-time user interface (UI) that lets users upload audio samples, get stress scores, and see their stress levels on user-friendly dashboards. This research introduced stress trend visualization, which allows users to follow stress changes over numerous sessions, in contrast to traditional stress detection methods that provide immediate assessments. An interactive stress level indicator, trend graphs, and a downloadable stress report were among the UX improvements that improved the system's readability and usefulness for mental health applications.

6.2 Contributions of the Study

1. Creation of a Multimodal Stress Detection System integrated vocal analysis and LLM-based text analysis.
2. Deployment of Long-Term Monitoring Stress Trend Visualisation Added a dynamic stress trend monitoring function that allows users to examine stress trends over time. Enhanced applications for mental health monitoring by offering a historical perspective on stress variations.
3. UX Improvements for Accessibility and User Interaction Developed an interactive user interface with stress visualization to improve the interpretability and actionability of stress detection results. Added downloadable stress reports, stress level indicators, and real-time feedback, improving the usefulness of stress detection tools.

Through the integration of deep learning, multimodal AI, and user-centred design, these contributions enhance the real-world implementation of stress detection systems.

6.1. Summary of the Research

6.3 Limitations of the Proposed System

Despite the positive results of this study, a number of limitations need to be taken into account. Instead of recording live speech inputs, the system processes pre-recorded speech, which restricts the possibility of real-time engagement.

6.4 Recommendations for Future Research

1. To enhance generalisation, include real-time stress datasets from a variety of speakers. To increase robustness, use datasets with unplanned speech instead of recordings.
2. Expand the system to accommodate real-time speech input processing, enabling ongoing stress monitoring. Create an adaptable user interface (UI) that adapts to the stress levels of its users and offers tailored suggestions and feedback.
3. Combining Multimodal Physiological Information For a more comprehensive stress evaluation, combine physiological cues (such as skin conductivity and heart rate) with speech-based stress detection. Investigate integrating wearable technology (such as smartwatches) for multimodal stress detection in real time.
4. To offer individualized stress management techniques, consider integration with mental health conversation systems.
5. Make the CNN-LSTM model faster on edge devices by optimizing it. Examine compact deep learning architectures that can be implemented on embedded and mobile devices.

The feasibility, scalability, and impact of AI-driven stress detection systems for mental health applications are the goals of these future areas.

6.5 Practical Applications of the Study

The research on stress detection systems has several practical applications, particularly in mental health monitoring, workplace wellness, and

6.3. Limitations of the Proposed System

personal well-being tracking. The work's potential applications span a variety of sectors where stress evaluation and long-term tracking can be beneficial, as it focusses on developing an interactive, user-friendly stress monitoring interface.

Workplace Stress Monitoring The system can be used by organisations to measure and control workplace stress levels as part of employee wellness initiatives. Managers and HR professionals can use the system's stress trend visualisation to comprehend how stress changes over time and create plans for reducing stress at work. could be used to business mindfulness programs where staff members monitor their psychological health and get stress-reduction advice.

Mental Health and Well-being Tracking

Individuals can use the stress trend tracking feature to monitor their mental state over time. The system can provide visual insights into stress patterns, helping users recognize triggers and take proactive steps for stress management. It can be integrated into digital mental health platforms to support therapy and self-care.

Education

The interface can serve as a basis for research on user behaviour in stress detection systems by academic institutions and researchers in the fields of psychology, UX/UI design, and Human-Computer Interaction (HCI). Additional features could be added to the framework to test various stress tracking visualisation techniques. The method can be used by researchers in the future to investigate the effects of interactive stress tracking on user engagement and mental health outcomes.

Self-Improvement and Personal Productivity

Individuals practicing self-care and mindfulness can use the stress tracking feature to measure how relaxation techniques such as meditation, deep breathing impact their stress over time. Can be beneficial for content creators and app developers building self-improvement tools, where users track their mental states alongside productivity habits.

Health Tech and Digital Apps

The system can be integrated into mobile health (mHealth) applications, providing users with an accessible and interactive stress monitoring tool. Can

6.5. Practical Applications of the Study

be paired with wearables or smart assistants to track real-time stress trends based on speech input. Useful for telemedicine platforms, where doctors and psychologists can review stress reports and provide better mental health support.

6.6 Final Thoughts

This study introduces a multimodal and user-friendly stress detection system which advances the expanding field of AI-driven mental health care. This system improves stress monitoring, visualization, and interpretability through large language models and UX-focused design. The results show that real-time stress tracking combined with voice and text-based analysis can produce insightful information on mental health. This system could be used as a useful tool for stress monitoring and mental health support with more improvements and real-world validations, bridging the gap between AI-based stress detection and practical applications.

References

- Nurkhon Akhmedov. *Designing and prototyping a learning and testing platform for user experience (UX) and user interface (UI) designers with the aim of improving knowledge and establishing a standard evaluation benchmark for UX/UI design skills and competencies*. Diss., Technische Hochschule Ingolstadt, 2023.
- Pablo Arias et al. Beyond correlation: Acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, 13(1):12–24, 2021.
- Sara Aristizabal et al. The feasibility of wearable and self-report stress detection measures in a semi-controlled lab environment. *IEEE Access*, 9: 102053–102068, 2021.
- Phie Chyan et al. A deep learning approach for stress detection through speech with audio feature analysis. In *6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 100–110. IEEE, 2022.
- Dafydd Gibbon. Prosody: The rhythms and melodies of speech. *arXiv preprint*, arXiv:1704.02565, 2017.
- Muhammad Syazani Hafiy Hilmy et al. Stress classification based on speech analysis of mfcc feature via machine learning. In *2021 8th International Conference on Computer and Communication Engineering (IC-CCE)*. IEEE, 2021.
- Júlia Kafková et al. A new era in stress monitoring: A review of embedded devices and tools for detecting stress in the workplace. *Electronics*, 13(19):3899, 2024.

- Joseph C. Kush. Integrating sensor technologies with conversational ai: Enhancing context-sensitive interaction through real-time data fusion. *Sensors*, 25(1):249, 2025.
- Feng Liu et al. Artificial intelligence in mental health: Innovations brought by artificial intelligence techniques in stress detection and interventions of building resilience. *Current Opinion in Behavioral Sciences*, 60:101452, 2024.
- Dominik Machaek, Raj Dabre, and Ondrej Bojar. Turning whisper into real-time transcription system. *arXiv preprint*, arXiv:2307.14743, 2023.
- Silke Paulmann et al. How psychological stress affects emotional prosody. *PLOS ONE*, 11(11):e0165022, 2016. doi: 10.1371/journal.pone.0165022.
- Mukesh Pillai and Pallavi Thakur. Developing a website to analyze and validate projects using langchain and streamlit. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 2024.
- Mirco Ravanelli et al. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11, 2024.
- Moein Razavi et al. Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: Scoping review. *JMIR Mental Health*, 11:e53714, 2024.
- John Relan and Others. Title of the paper. *Journal Name*, 12(3):123–134, 2019.
- Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, 2018.
- Rita Sebastião and David Dias Neto. Stress and mental health: The role of emotional schemas and psychological flexibility in the context of covid-19. *Journal of Contextual Behavioral Science*, 32:100736, 2024.
- Guobin Shen et al. Stressprompt: Does stress impact large language models and human performance similarly? *arXiv preprint*, arXiv:2409.17167, 2024.

- Nilu Singh, R. A. Khan, and Raj Shree. Mfcc and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, 54(1):9–13, 2012.
- Praveen Singh et al. Social signal processing for evaluating conversations using emotion analysis and sentiment detection. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE, 2019.
- Senthil Sriramprakash, Vadana D. Prasanna, and OV Ramana Murthy. Stress detection in working people. *Procedia Computer Science*, 115: 359–366, 2017.
- Statistics Solutions. Paired sample t-test. <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/>. Accessed: 2025-02-02.
- Martha T. Teye et al. Evaluation of conversational agents: understanding culture, context and environment in emotion detection. *IEEE Access*, 10:24976–24984, 2022.
- Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. *arXiv preprint*, arXiv:1912.10458, 2019.
- Yiqun Yao et al. Muser: Multimodal stress detection using emotion recognition as an auxiliary task. *arXiv preprint*, arXiv:2105.08146, 2021.
- S. Yoon et al. Text-based stress detection using semantic analysis. *Journal of Computational Linguistics*, 2020.

Acknowledgements

I would like to start by thanking ALLAH, the Ever Most Generous, for granting me this wonderful opportunity to study at this institution. His countless blessings have given me the health, strength, and clarity of mind to undertake and complete this thesis.

I extend my deepest appreciation to my supervisor, Prof. Shiramatsu, for his valuable guidance, insightful feedback, and unwavering support throughout this research. His expertise and encouragement have been instrumental in shaping my work and strengthening my understanding of my research field.

A special thanks to my dear parents and Jobair Al Rafi for their unconditional support, patience, and encouragement during this journey. Their belief in me has been my greatest source of motivation, especially during challenging times.

Furthermore, I would like to give special thanks to Aichi Monozukuri Scholarship for providing the financial support during my master's program and Aichi Prefecture International Department for their generous support.

I would like to thank all the members from Shiramatsu Laboratory. I am also sincerely grateful to the NITech International Affairs Division and Japanese teachers for their continuous support.

Lastly, I would like to take this opportunity to express my sincere gratitude to many people who helped me with this project who played a significant role by providing the necessary resources and support. This research would not have been possible without the collective efforts, guidance, and encouragement of so many individuals, and for that, I am truly grateful.