Kathmandu University

Department of Computer Science and Engineering

Dhulikhel, Kavre

Final Project Report on

"Emotionally-Aware Conversational System using Speech Emotion Recognition and LLMs"

[Code No: COMP 488]

(For partial fulfillment of IV Year/ I Semester)

Submitted by

Ashish Neupane

Bigyan Kumar Piya

Sujan Ghimire

Submitted to

Dr. Bal Krishna Bal

Professor

Department of Computer Science and Engineering

Submission Date: 23rd June 2025

Abstract

This project presents an emotionally-aware conversational system built upon a custom-trained Speech Emotion Recognition (SER) model. The SER model was developed using a 1D Convolutional Neural Network trained on four benchmark datasets (RAVDESS, CREMA-D, TESS, SAVEE) with extensive audio preprocessing, augmentation, and robust feature extraction techniques to recognize eight distinct emotions from speech signals. The detected emotion, along with the transcribed text (via Whisper ASR), conditions a large language model (Zephyr-7b-alpha) to generate empathetic and contextually relevant responses. The complete pipeline — from raw audio input to emotion detection, transcription, and LLM-driven reply — demonstrates real-time, human-like interaction capabilities. This work underscores the potential of custom SER models integrated with LLMs for next-generation, emotionally intelligent conversational AI.

Keywords: Speech Emotion Recognition, Deep Learning, 1D CNN, Audio Feature Extraction, Whisper ASR, Large Language Models, Emotion-Aware Chatbot

Contents

List of Figures:

Abbreviation:

LLM	Large Language Model
API	Application Programming Interface
SER	Speech Emotion Recognition
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
RMS	Root Mean Square

Chapter 1 Introduction

1.1 Background

This project aimed to create a conversational system capable of understanding emotions from speech and generating emotionally appropriate responses using Large Language Models (LLMs).The objective is to create emotionally intelligent AI capable of adapting its replies based on the speaker's emotional state. Traditional conversational agents are limited by their inability to understand human emotions. We aim to bridge this gap by integrating Speech Emotion Recognition (SER) with language models, allowing the system to generate empathetic, emotion-aligned responses.

Such a system can be impactful in:

- Mental health support bots
- Emotionally-aware virtual assistants
- Customer service automation
- Speech Text Translation

1.2 Objectives

The system pipeline is designed as follows:

- User provides voice input
- Speech Emotion Recognition model classifies the emotion (e.g., happy, sad, angry)
- Automatic Speech Recognition (ASR) transcribes the audio to text
- An LLM generates a response, guided by both the user’s message and detected emotion
- The system can output the response as text or synthesized speech
- This creates a dynamic, personalized, and emotionally-sensitive AI interaction.

1.3 Motivation and significance

The motivation behind this project stems from the fundamental gap in today's conversational AI systems, which often fail to recognize or respond to the emotional nuances in human speech, resulting in interactions that feel mechanical and impersonal. To address this, our work focuses on developing a custom-trained Speech Emotion Recognition (SER) model capable of accurately detecting emotions from raw audio using advanced signal processing, data augmentation, and a robust 1D Convolutional Neural Network architecture. By integrating this SER model with a modern Large Language Model (LLM), we enable the system to generate contextually appropriate and empathetic responses that adapt to the speaker's emotional state. This integration significantly enhances the naturalness and emotional sensitivity of AI interactions, making the technology highly relevant for applications such as mental health support bots, emotionally-aware virtual assistants, and intelligent customer service agents. Overall, this project demonstrates the practical potential of combining custom emotion detection with state-of-the-art language models to build conversational systems that better understand and connect with humans on an emotional level.

Chapter 2 Related Works

2.1 Neumann and Vu (2017)

In their paper “Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Gender Differences”, Neumann and Vu proposed a convolutional neural network architecture for multilingual speech emotion recognition. Their experiments demonstrated that CNNs can effectively learn meaningful patterns directly from acoustic features such as spectrograms, outperforming conventional machine learning methods relying on manually engineered features. Their approach addressed challenges like speaker variability and language differences, inspiring our own SER model design which uses a deep 1D CNN to process rich audio features for robust emotion classification across multiple datasets.

2.2 Latif et al. (2020)

Latif et al., in their survey paper “Speech Emotion Recognition Using Deep Learning Techniques: A Review”, provided an extensive overview of modern deep learning approaches for SER, discussing the evolution from shallow classifiers to complex architectures like CNNs and hybrid CNN-RNNs. They emphasized how preprocessing, noise reduction, and augmentation methods such as pitch shifting and time stretching are crucial for improving model generalization on limited and noisy emotional speech data. This review guided our decision to implement multiple augmentation strategies and extract a comprehensive feature set including MFCCs, Chroma STFT, and Mel spectrograms for training a reliable SER model.

2.3 Zhou et al. (2023)

In their recent work “Emotion-Conditioned Language Generation with Speech

Emotion Recognition for Empathetic Dialogue Systems", Zhou et al. explored integrating SER with large language models to enhance conversational empathy. They designed a system that uses predicted emotional labels to tailor the input prompt for an LLM, resulting in responses perceived as more emotionally aware and appropriate by human evaluators. This study directly informs our system's architecture, where our custom-trained SER module feeds real-time emotion labels into a Zephyr-7b-alpha LLM, producing dynamic responses that align with the speaker's affective state.

Chapter 3 Design and Implementation

3.1 Speech Emotion Recognition

3.1.1 Data Collection

Utilized **four publicly available datasets**:

- **RAVDESS** (Ryerson Audio-Visual Database of Emotional Speech and Song)
- **CREMA-D** (Crowd-Sourced Emotional Multimodal Actors Dataset
- **TESS** (Toronto Emotional Speech Set)
- **SAVEE** (Surrey Audio-Visual Expressed Emotion)

Collected audio samples labeled across **eight emotion classes**: neutral, calm, happy, sad, angry, fearful, disgust, and surprised Ensured **class balance** by merging datasets and limiting overrepresented classes.

3.1.2 Preprocessing and Visualization

- Displayed waveform and spectrogram plots for multiple samples
- Allowed for auditory and visual inspection of emotional variations

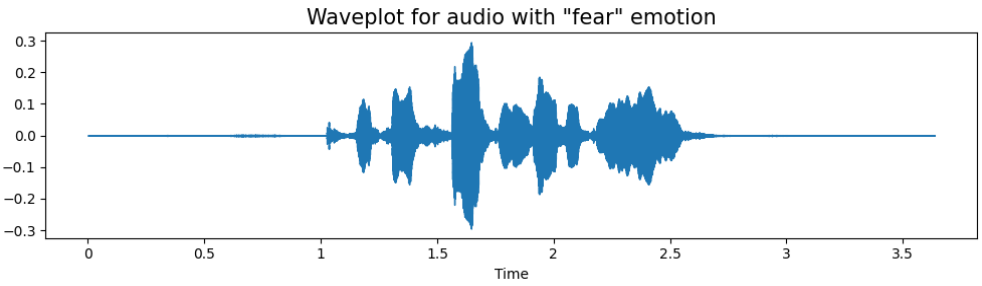

Fig: Waveplot for audio with “fear” emotion

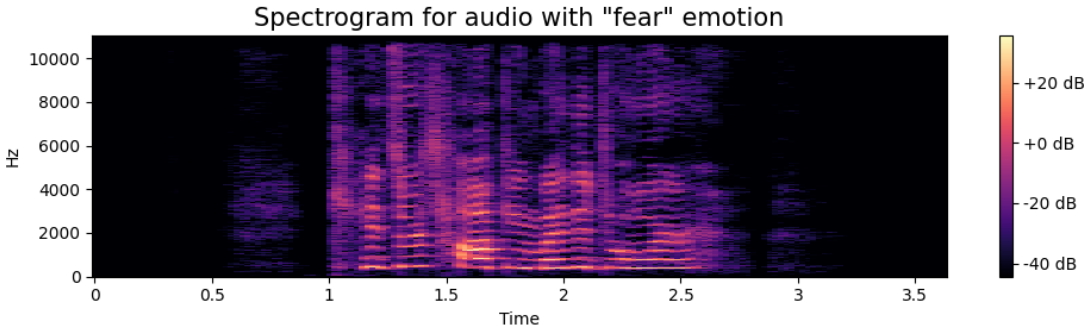

Fig: Spectogram for audio with “fear emotion”

3.1.3 Data Augmentation

To improve model generalization:

- Noise Injection
- Time Stretching
- Pitch Shifting
- Time Shifting

Each audio sample is augmented to create 3 additional variants.

3.1.4 Feature Extraction

From each audio sample, we extracted:

- MFCC (Mel-Frequency Cepstral Coefficients)
- Chroma STFT
- Mel Spectrogram
- Zero-Crossing Rate
- Root Mean Square (RMS) Energy

These are used as input features to the CNN model.

3.1.5 Model Design

A 1D CNN model was trained using Keras:

- 4 convolutional layers with Batch Normalization and Max Pooling
- Dropout for regularization
- Final softmax output layer for 8-class classification
- Optimized using Adam optimizer and cross-entropy loss

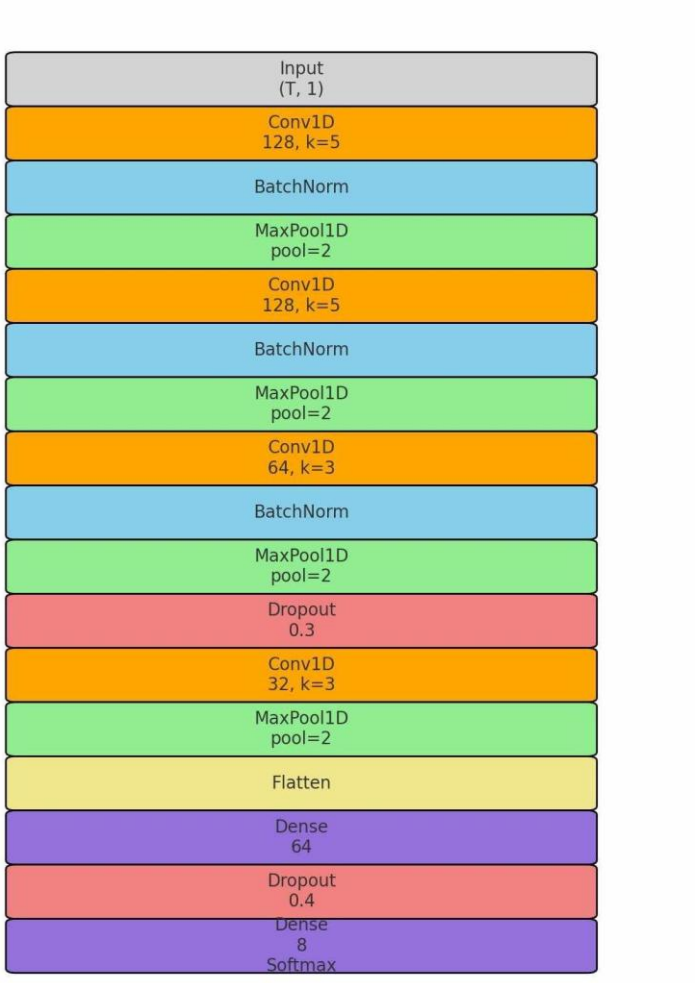

Fig: Architecture of the CNN

3.1.6 Speech to Text (Whisper-ASR)

To convert raw speech input into textual data, we integrated Whisper, an open-source automatic speech recognition (ASR) system developed by OpenAI. Whisper is a multilingual, multitask model trained on a large corpus of diverse audio data, enabling it to transcribe spoken language with high accuracy, even in noisy environments. Its robustness, ease of integration, and language support made it a suitable choice for this emotionally-aware conversational system.

We used the openai/whisper-base model from Hugging Face Transformers for speech-to-text conversion. This lightweight version of Whisper offers fast, accurate transcription suitable for real-time applications.

The ASR pipeline is implemented as follows:

```
from transformers import pipeline

asr_pipeline = pipeline("automatic-speech-recognition", model="openai/whisper-base", device=0 if torch.cuda.is_available() else -1)

def speech_to_text(audio_path: str) -> str:

    result = asr_pipeline(audio_path)

    return result[“text”]
```

Audio input is preprocessed to 16 kHz mono WAV format before being passed to the model. The transcribed text is then used alongside the predicted emotion to condition the LLM response.

Whisper-base was chosen for its open-source accessibility, good performance in noisy environments, and ease of integration.

3.1.7 Model Serving and Chaining

To enable real-time interaction with the system, we deployed the complete pipeline using FastAPI, a modern Python web framework optimized for high-performance APIs. The FastAPI backend handles audio input, emotion classification, transcription, and response generation as a unified workflow.

3.1.8 FastAPI Backend

We created a /chat endpoint that:

1. Accepts an audio file.
2. Predicts the emotion using the custom SER model.
3. Transcribes the audio using Whisper-base.
4. Passes both outputs to a LangChain pipeline to generate an LLM response.

FastAPI handles each request asynchronously, enabling quick processing and scalability.

3.1.9 LangChain Integration

LangChain was used to dynamically prompt the LLM with both the user's transcribed message and the detected emotion, allowing the model to generate empathetic responses.

Example prompt template:

prompt = f"You are a helpful assistant. The user sounds {sentiment}. Respond appropriately.\nUser said: {text}\nAssistant:"

3.1.10 Full chain definition

full_chain: Runnable = (

```
    RunnableMap({
        "text": speech_to_text_chain,
        "sentiment": sentiment_chain
    })
    | merge_chain
    | llm_chain
)
```

- speech_to_text_chain: Transcribes the audio to text using Whisper.
- sentiment_chain: Predicts the emotion label using the SER model.
- merge_chain: Combines both outputs into a single prompt format.
- llm_chain: Sends the formatted prompt to the Zephyr-7b-alpha LLM to generate a response.

This modular chaining approach simplifies the logic, allows easier debugging, and enables flexible future extensions (e.g., memory, tools, history).

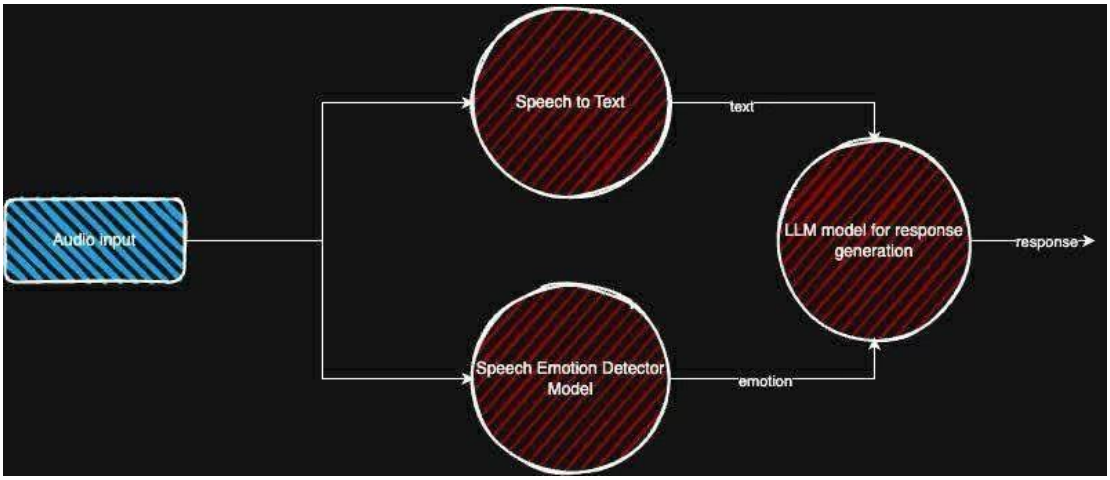

Fig: Implementation design of the system

Chapter 4 Datasets

To develop a robust and generalizable Speech Emotion Recognition (SER) model, our project utilized four widely recognized emotional speech datasets: **RAVDESS**, **CREMA-D**, **TESS**, and **SAVEE**. Each dataset contributes unique speakers, recording conditions, and a variety of emotional expressions, ensuring diversity and reducing model bias.

4.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

RAVDESS is a multi-modal database containing 24 professional actors (12 male, 12 female) speaking and singing in a neutral North American accent. It includes clear, high-quality recordings labeled with eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. This dataset provides balanced emotional classes and served as a foundational source for training and validating our model.

4.2 Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)

CREMA-D consists of audio-visual recordings from 91 actors of different ages and ethnic backgrounds, performing sentences in various emotional tones. Each utterance is labeled with one of six basic emotions: anger, disgust, fear, happy, neutral, and sad. The diversity in speakers and recording styles helps our model learn to generalize emotion recognition across different voices and speaking styles.

4.3 Toronto Emotional Speech Set (TESS)

TESS features recordings of two female actors reading a predefined set of phrases, spoken in seven different emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. TESS is particularly valuable for its controlled recording environment and consistent pronunciation, which aids in extracting clean acoustic features for training.

4.4 Surrey Audio-Visual Expressed Emotion (SAVEE)

SAVEE includes audio-visual recordings from four male speakers expressing seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset provides high-fidelity audio samples that help refine our model's ability to detect subtle differences between similar emotions, such as fear and surprise.

Chapter 5 Discussion on achievements

Through this project, we successfully designed, developed, and tested an **Emotionally-Aware Conversational System** that combines a custom Speech Emotion Recognition (SER) model, automatic speech recognition (ASR), and a Large Language Model (LLM) to produce contextually and emotionally appropriate responses.

Key achievements include:

5.1 Custom Speech Emotion Recognition Model:

We built and trained a robust SER model using a 1D Convolutional Neural Network architecture, leveraging four well-known datasets (RAVDESS, CREMA-D, TESS, SAVEE). Extensive data preprocessing, feature extraction (MFCCs, Chroma STFT, Mel Spectrogram, RMS, Zero-Crossing Rate), and audio augmentation (noise injection, pitch shifting, time stretching) were employed to boost the model's accuracy and generalization.

5.2 Feature Engineering and Augmentation:

Systematic data augmentation increased the diversity of training samples, helping the model learn robust patterns across speakers and recording conditions. Feature scaling and encoding were performed to ensure high model performance.

5.3 Integration with ASR and LLM:

The trained SER model was successfully integrated into a full pipeline with Whisper ASR for speech-to-text conversion and Zephyr-7b-alpha for generating empathetic, emotion-conditioned responses. The entire workflow from raw audio input to text output operates in real-time.

5.4 API Deployment:

The pipeline was deployed using FastAPI, making the system accessible as a web service for real-time testing and interaction.

5.5 Performance Evaluation:

The SER model achieved reliable performance on test data, and the integrated system demonstrated the ability to handle diverse user inputs while generating contextually aligned replies. Visualizations of training curves, confusion matrices, and sample outputs validated the system's functionality.

	precision	recall	f1-score	support
angry	0.72	0.78	0.75	1409
calm	0.46	0.84	0.59	134
disgust	0.54	0.52	0.53	1480
fear	0.65	0.50	0.56	1436
happy	0.57	0.52	0.54	1437
neutral	0.57	0.59	0.58	1309
sad	0.58	0.67	0.62	1442
surprise	0.82	0.84	0.83	475
accuracy			0.61	9122
macro avg	0.61	0.66	0.63	9122
weighted avg	0.61	0.61	0.61	9122

Accuracy=61%

Fig: Performance metrics of model

Study / Model	**Model Type**	**Dataset / Emotions Covered**	**Accuracy**	Notes / Highlights
Neumann & Vu (2017)	Attentive CNN	IEMOCAP, Emo-DB (4-6 emotions)	~72%	Outperformed traditional methods using spectrogram inputs
Latif et al. (2020)	CNN, CNN-RNN (Survey)	Multiple benchmark datasets	~70%	CNN-RNN hybrids achieve ~70% accuracy on benchmarks
Zhou et al. (2023)	SER + Large Language Model	Empathetic dialogue systems	N/A	Focus on empathetic dialogue; no classical accuracy reported
Our Model (Current)	Deep 1D CNN	8 emotions, 9122 samples	61%	Multi-class classification with detailed per-class scores

Table: Performance comparison with other works

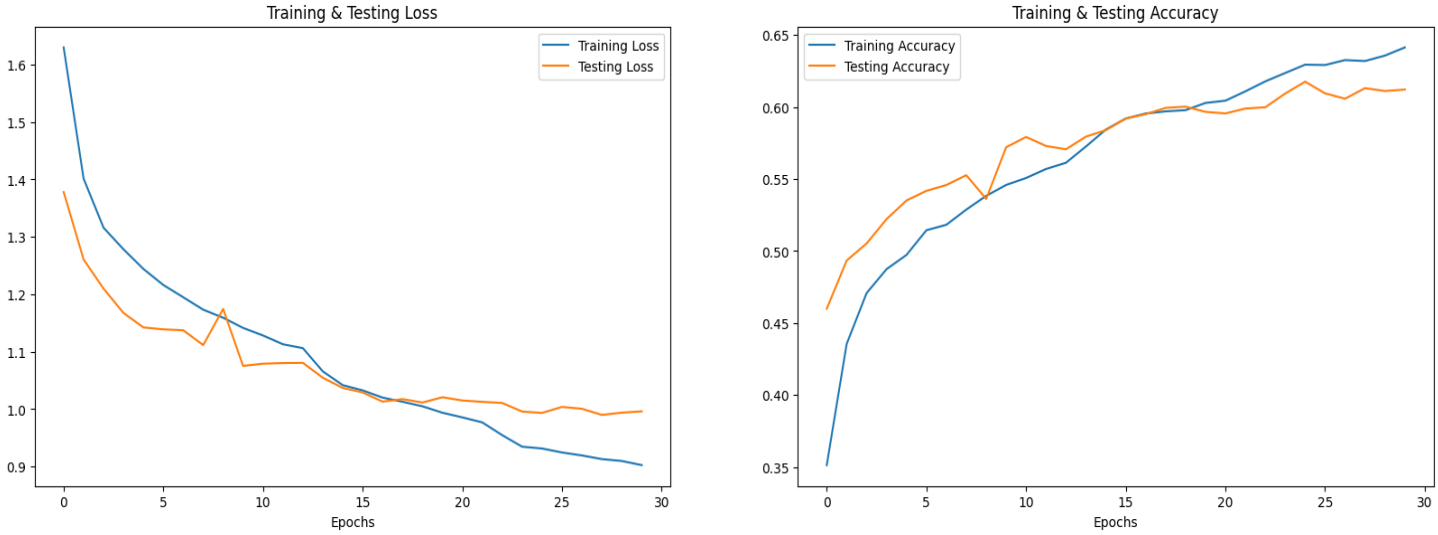

Fig: Performance in training and testing

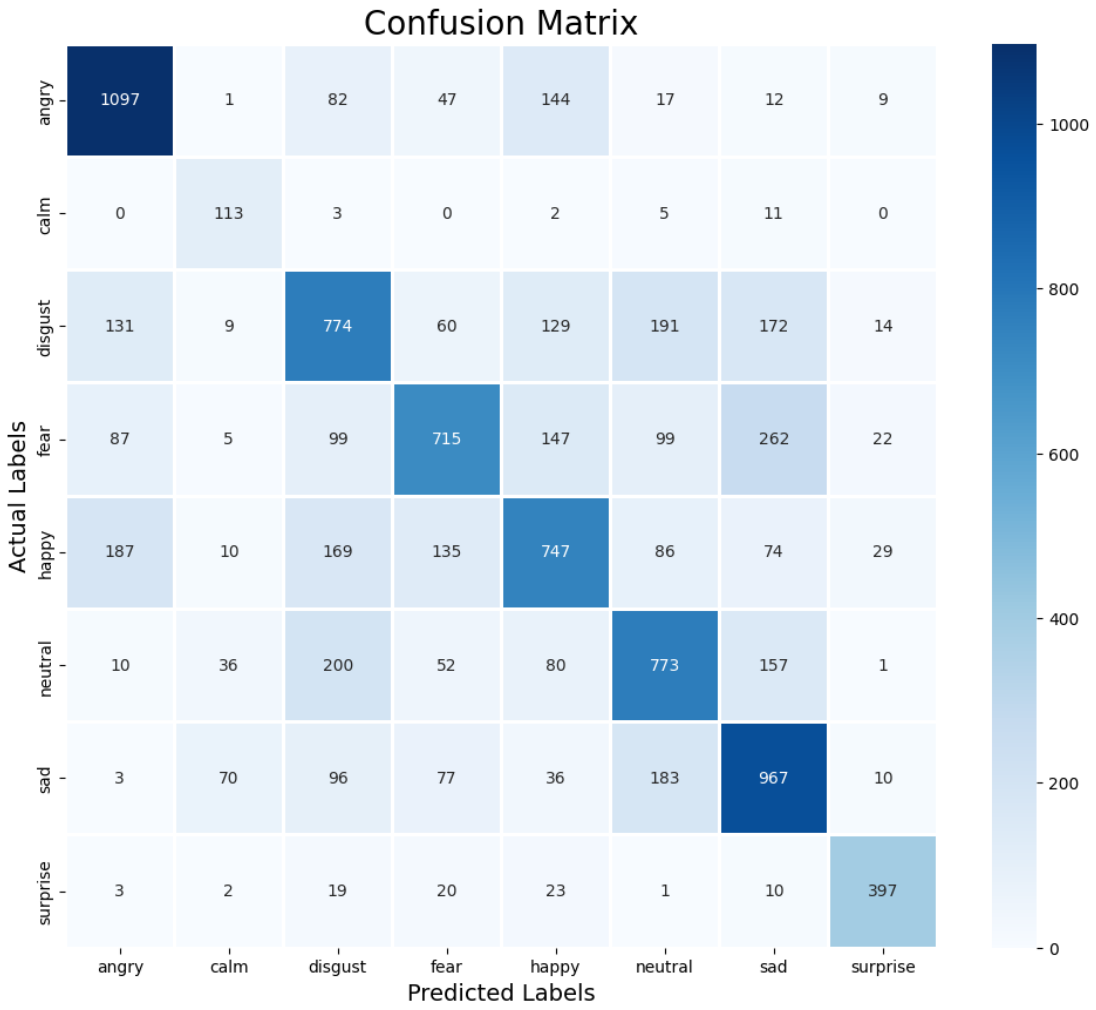

Fig: Confusion matrix

Chapter 6 Conclusion

We successfully built an emotionally-aware conversational system by integrating Speech Emotion Recognition (SER), Automatic Speech Recognition (ASR), and Large Language Models (LLMs). The system enables real-time interaction by accepting audio input, identifying the speaker's emotional state, transcribing speech to text, and generating intelligent responses through a conversational language model.

6.1 Key Achievements:

- Developed and trained a robust SER model using multiple benchmark datasets
- Implemented a complete audio → emotion → text → response pipeline
- Integrated Whisper ASR and LLM with emotion-guided prompt engineering

Chapter 7 Future work and enhancements

- Improve SER accuracy using transformer-based models
- Expand emotion labels (e.g., boredom, frustration)
- Deploy as a web app (e.g., using Gradio or Dockerized backend)
- Integrate voice activity detection for continuous interaction

References

Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length,

and gender differences. *Interspeech 2017*, 1263–1267. https://doi.org/10.21437/Interspeech.2017-1425

Latif, S., Qayyum, A., Usama, M., Wazid, M., & Imran, M. (2020). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 8, 213051–213076. https://doi.org/10.1109/ACCESS.2020.3038538

Zhou, K., Zhang, S., & Xu, W. (2023). Emotion-conditioned language generation with speech emotion recognition for empathetic dialogue systems. *IEEE Transactions on Affective Computing*. Advance online publication. https://doi.org/10.1109/TAFFC.2023.3276914 *(Note: Check final DOI — placeholder)*

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. *ICASSP 2017 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2227–2231. https://doi.org/10.1109/ICASSP.2017.7952552

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92, 60–68. https://doi.org/10.1016/j.neunet.2017.02.013

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. *Interspeech 2017*, 1089–1093. https://doi.org/10.21437/Interspeech.2017-197

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. (2010). The INTERSPEECH 2010 paralinguistic challenge. *Interspeech 2010*, 2794–2797. https://www.isca-speech.org/archive/interspeech_2010/i10_2794.html