Voice2sentiment: An End-to-End System for Speech Emotion Recognition and Textual Sentiment Analysis

UNDER THE GUIDANCE OF

Prof. Sonali Das

Department of B.Tech in Computer Science Engineering

(Specialization in Artificial Intelligence and Machine Learning)

Supreme Knowledge Foundation Group of Institutions, Hooghly, West Bengal

Under
Maulana Abul Kalam Azad University of Technology
Kolkata, West Bengal, India.

SUBJECT : PROJECT -I

SUBJECT CODE : PROJ-AIML781

TEAM MEMBER	UNIVERSITY ROLL NUMBER
Uma Saha	25330822021
Jamima Khatun	25330822009
Debolina Samanta	25330822008
Aditya Choudhary	25330822026

Introduction

Brief Overview of the Project

- An AI system that analyzes voice emotion and text sentiment together
- It converts speech to text, extracts audio features, and detects real-time emotions
- It provides a more accurate understanding of emotions by combining tone and meaning

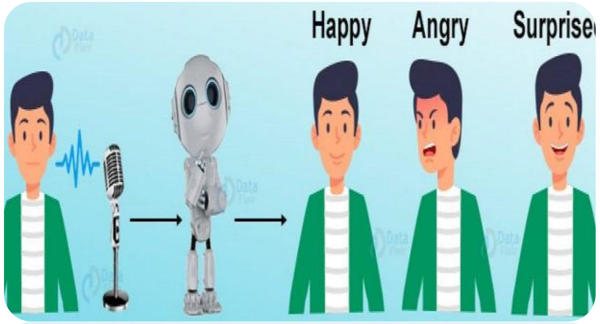
Why This Project Was Chosen

- Current systems look at only one form (voice or text), which causes mistakes
- Human communication needs AI that understands emotions for better interactions
- There is a high demand in customer care, mental health, education, and virtual assistants
- The aim is to build a multilingual Indian emotion-sentiment system

Problem Statement

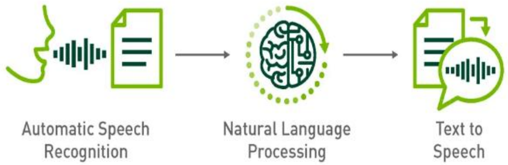
What problem exists in the current scenario

- Current systems analyze either speech emotion or text sentiment, not both at the same time
- Emotional meaning is often lost when tone and spoken words disagree
- Noise, different accents, and multilingual speech reduce accuracy



Gaps/limitations in existing systems

- There is a lack of unified real-time emotion and sentiment systems
- They perform poorly in noisy environments , support for Indian languages is limited
- These systems struggle to detect sarcasm, stress, or hidden emotions



Why the problem needs a solution

- Understanding emotions accurately is vital for human-like AI interaction
- It helps improve customer service, mental health monitoring, and digital communication
- A combined voice and text system provides a truer and more reliable emotional interpretation

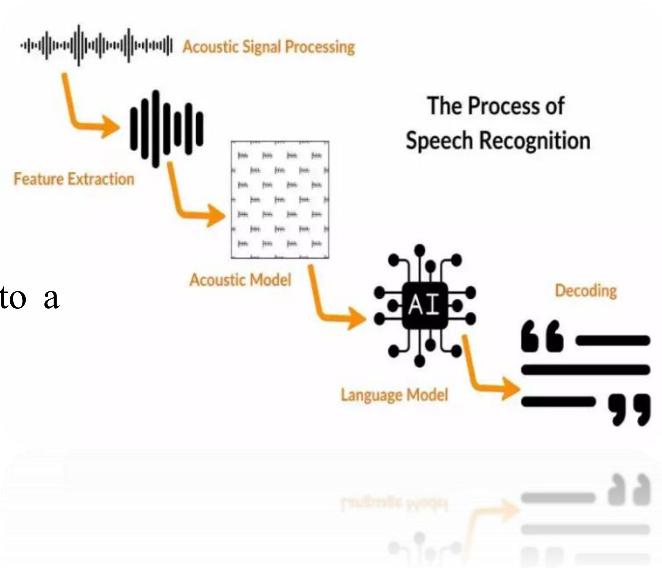
Objectives of the Project

Main objectives

- To build an AI system that analyzes speech emotion and text sentiment together
- To develop a real-time, accurate, and multilingual emotion and sentiment detection model

Specific goals to be achieved

- Convert speech to text using ASR with high accuracy
- Extract audio features such as MFCC, pitch, and chroma for emotion classification
- Detect emotions like anger, neutral, disgust, happiness, sadness, and fear
- Compare tone and meaning to identify the true sentiment. Integrate both models into a single end-to-end pipeline



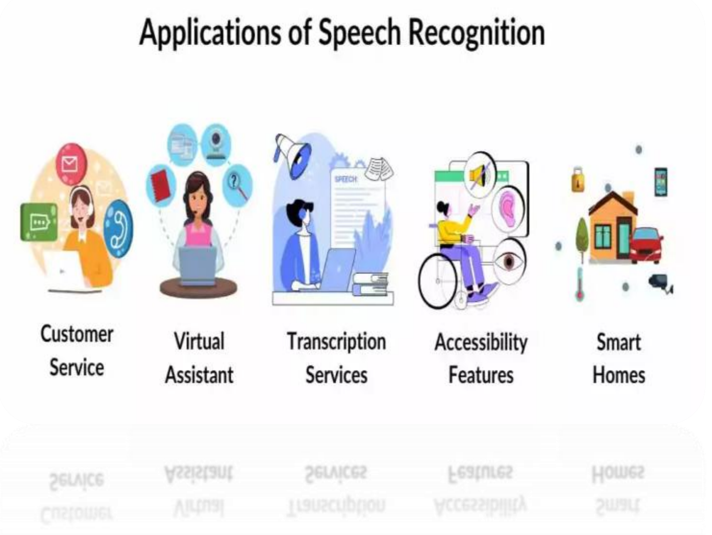
Expected outcomes

- A working system that provides real-time emotion and sentiment output
- Improved accuracy through combined analysis of voice and text
- Multilingual support for Indian languages including Hindi, Bengali, and English
- A deployable prototype for use in customer care, mental health support, and AI assistants

Relevance in real-world application

How your solution overcomes in the current scenario?

- It combines voice emotion and text sentiment, unlike existing systems that rely on only one method
- It performs better in noisy, real-life settings by using audio preprocessing
- It supports multilingual speech in Hindi, Bengali, and English, overcoming language barriers
- It detects deeper emotions like sarcasm, stress, and frustration, which current models often overlook
- It provides real-time results, making it suitable for practical use



Feasibility Study

Social Feasibility

- Improves communication, mental health support, and user experience
- Voice-based AI tools are becoming common

Economic Feasibility

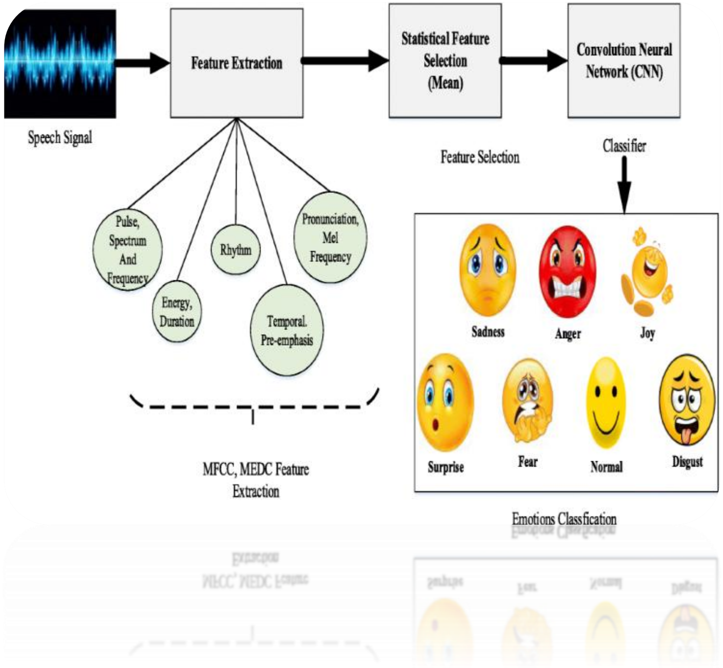
- Primarily uses open-source tools, which lowers development costs
- The cost of cloud and GPU resources is manageable and needed only during training

Technical Feasibility

- Built with proven technologies: Whisper, Tensorflow, HuggingFace, and NLP tools
- Real-time performance is possible with current hardware and cloud options

Organizational Feasibility

- Can be used by call centers, healthcare, education, and AI industries
- Deployment is simple through web apps like Flask and Streamlit, making it easy to implement



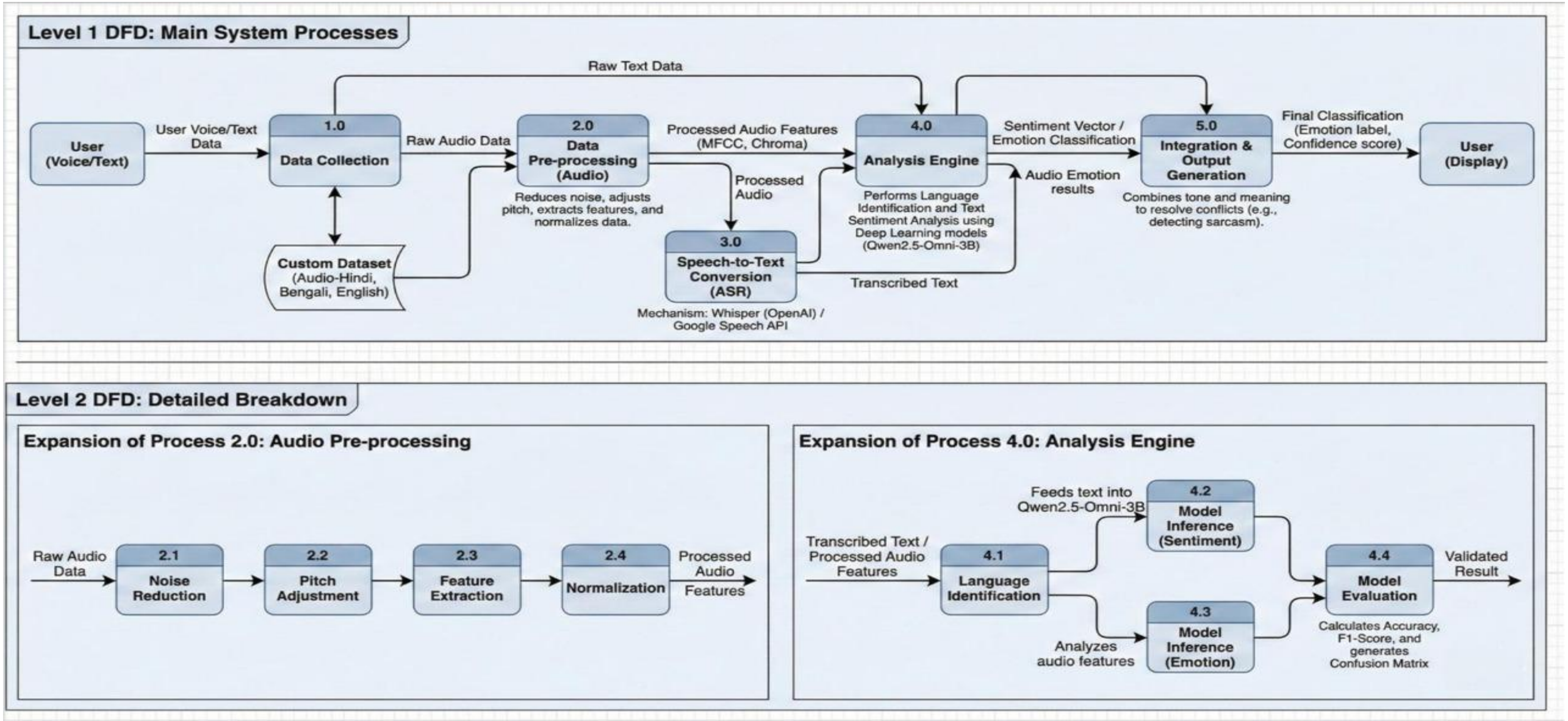
Gantt Chart (Work Done)

Particulars of the Task	*week*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Estimated date of completion	*1-4th July*	*21-27th July*	*28th July-3rd Aug*	*4-10th Aug*	*11-17th Aug*	*18-24th Aug*	*25-31st Aug*	*1-7th Sept*	*8-12th Sept*	*15-21st Sept*	*22nd Sept-5th Nov*
Topic selection & group Formation Problem definition & objectives +Literature review+ PPT1 &report submission											
Custom Dataset collection(audio-Hindi, Bengali, English language) + PPT2 submission											
DL Model design + model training(emotion +sentiment)+hyperparameter tuning											
Actual date of completion	4th July	Ongoin g	Ongoing	6th Aug	Ongoin g	Ongoing	Ongoin g	Ongoi ng	Ongoin g	Ongoing	Ongoing (dataset) & 5th Nov PPT2 submit

Gantt Chart (work will be done)

Particulars of the Task	*week*										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Estimated date of completion	***6-12 Jan***	***13-19 Jan***	***20-26 Jan***	***27 Jan - 2 Feb***	***3-9 Feb***	***10-16 Feb***	***17-23 Feb***	***24 Feb-2 Mar***	***3-9 Mar***	***10-16 Mar***	***17Mar- 30 April***
DL Model design + model training(emotion +sentiment)+hyperparameter tuning +text transcription											
System integration (voice + text pipeline)+evaluation metrics + improvement error analysis + deployment											
Documentation + testing on unseen data + final refinement											
Actual date of completion	–	–	–	–	–	–	–	–	–	–	–

Proposed System



Technologies Used

Technologies to be used :

•***Languages***: Python (core), JavaScript (backend/frontend),Django or Flask(backend),HTML&CSS(frontend)

•***Frameworks/Libraries***:

•PyTorch / TensorFlow (deep learning)

•Hugging Face

•Whisper Speech Recognition (speech-to-text)

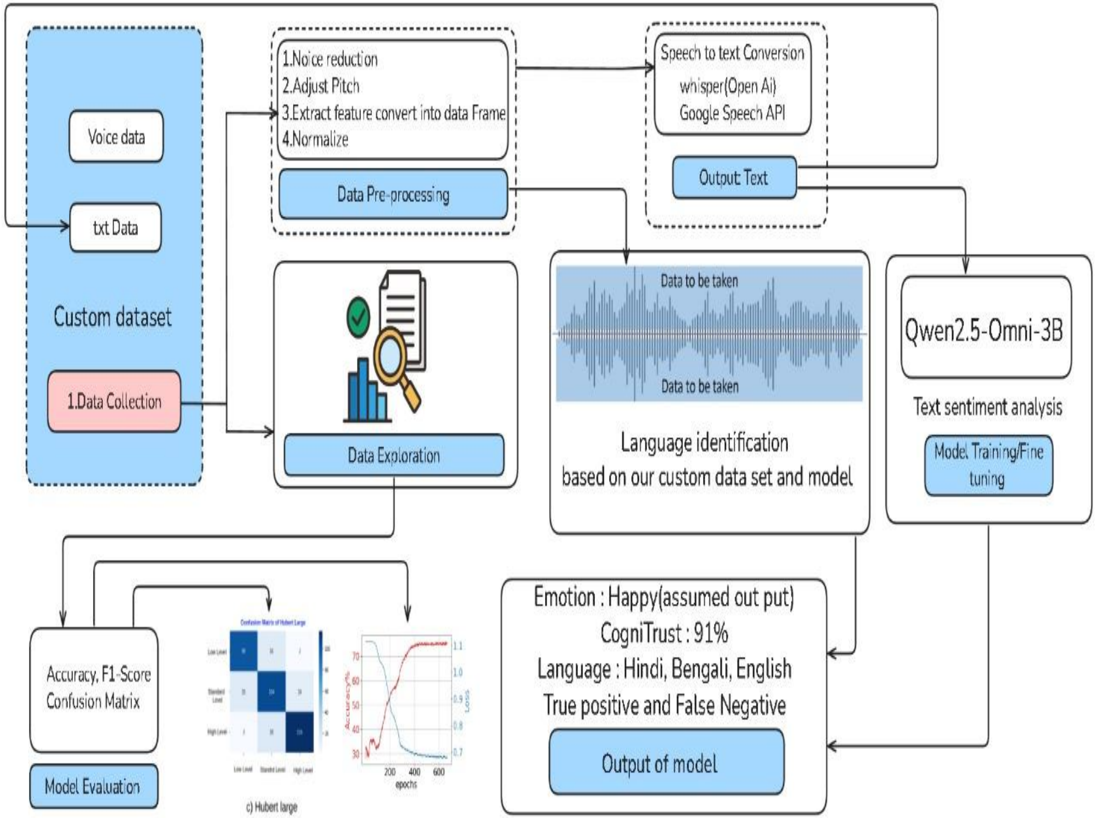
•Lib ROSA / PyDub (audio processing)

•VADER / TextBlob (sentiment analysis)

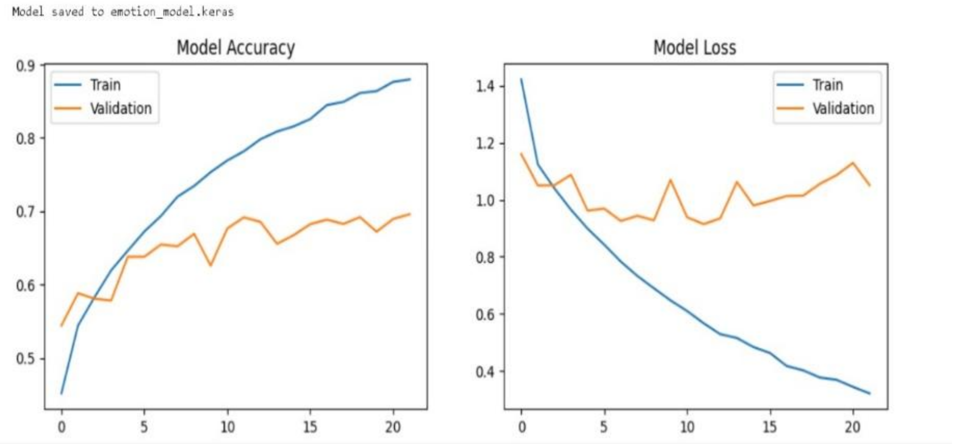
•***Tools***: Streamlit, Flask, Google Colab, GitHub

•***Hardware***: 8GB+ RAM system, GPU recommended, or cloud (AWS/GCP),render

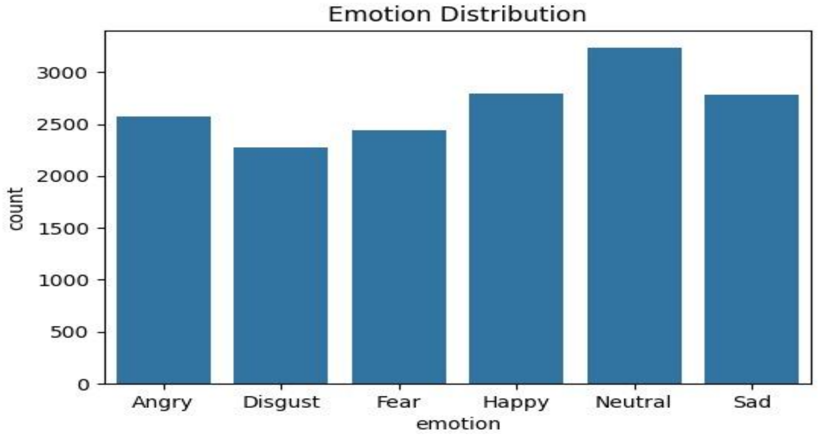
Methodology and process for implementation:



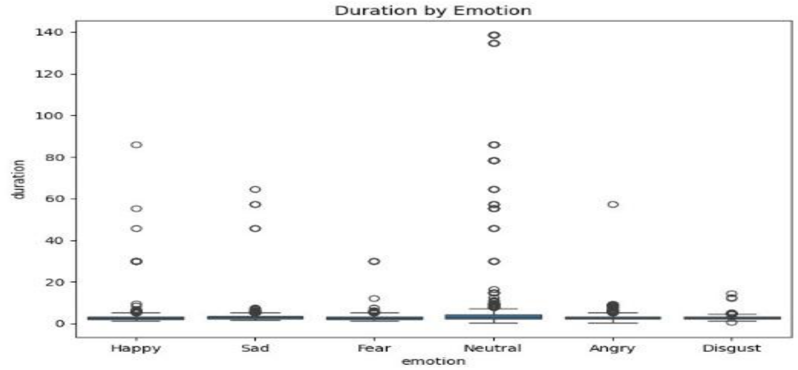
Training accuracy output



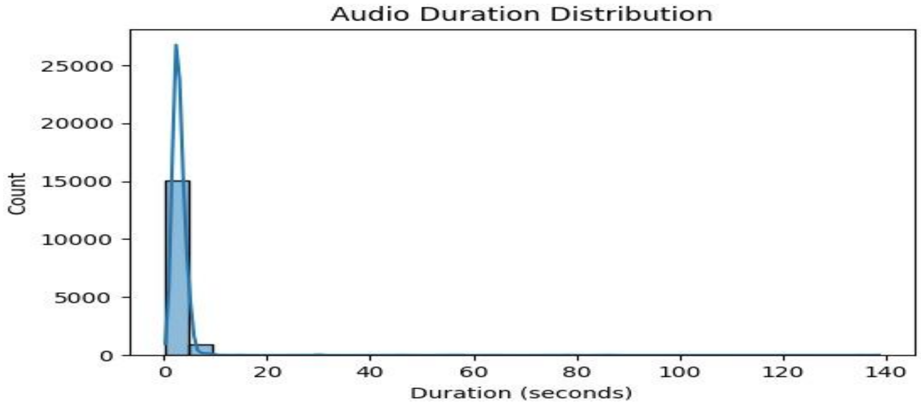
Training Performance – Accuracy & Loss Curves



Emotion Distribution (Dataset Analysis)



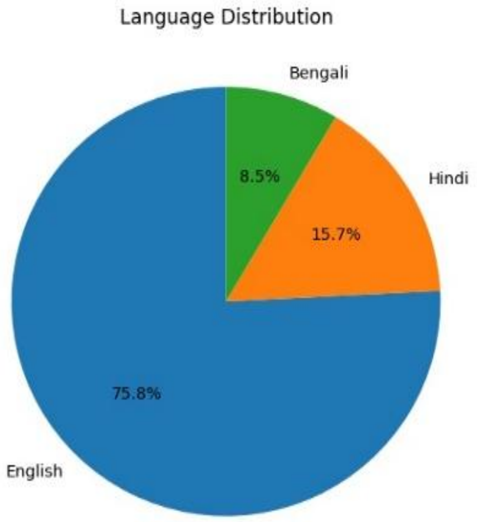
Duration by Emotion (Scatter Plot)



Audio Duration Distribution (Histogram)

Dataset analysis

SUPREME KNOWLEDGE FOUNDATION

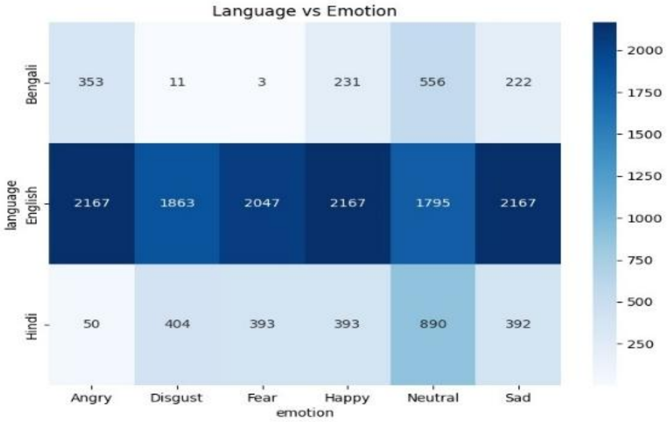


Language Distribution (Dataset Analysis)

```
==================================================
OVERALL CLASSIFICATION REPORT
==================================================
Total Accuracy:  90.12%
Weighted Precision: 90.28%
--------------------------------------------------
              precision    recall  f1-score   support

       angry       0.94      0.94      0.94      2570
     disgust       0.87      0.91      0.89      2278
        fear       0.94      0.82      0.87      2443
       happy       0.92      0.88      0.90      2791
     neutral       0.90      0.94      0.92      3241
         sad       0.85      0.91      0.88      2781

    accuracy                           0.90     16104
   macro avg       0.90      0.90      0.90     16104
weighted avg       0.90      0.90      0.90     16104

Confusion Matrix:
[[2415   52   16   40   33   14]
 [  28 2062   18   37   58   75]
 [  38   95 2000   54   42  214]
 [  83   68   48 2449  101   42]
 [  12   28    6   40 3059   96]
 [   4   72   46   40   91 2528]]
```

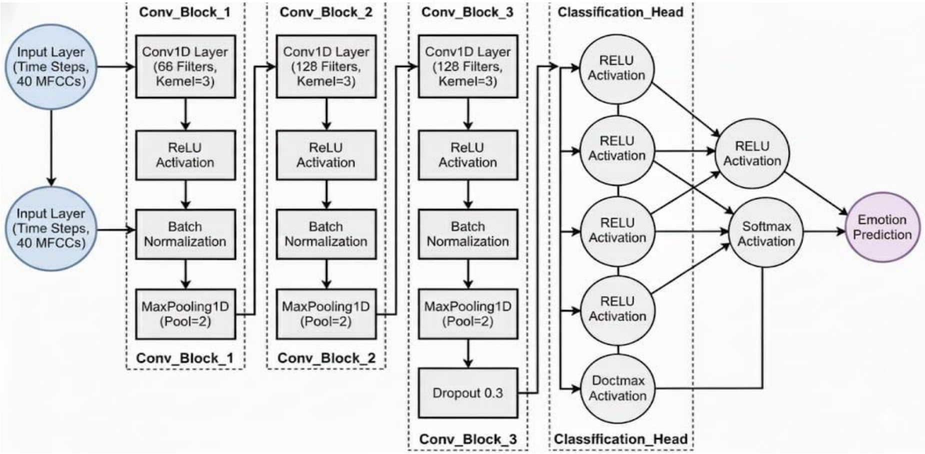
Classification Report (Precision, Recall, F1-Score)

Confusion Matrix (Emotion Recognition Model)

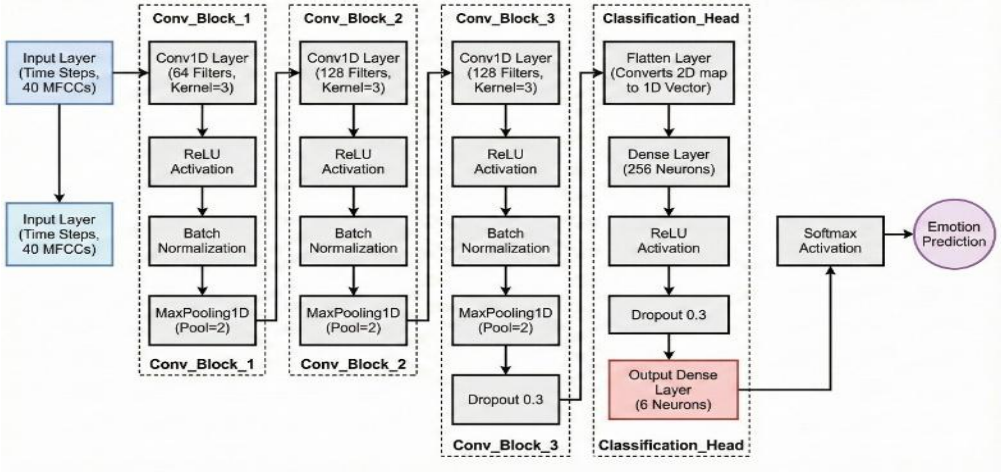


Language vs Emotion Heatmap

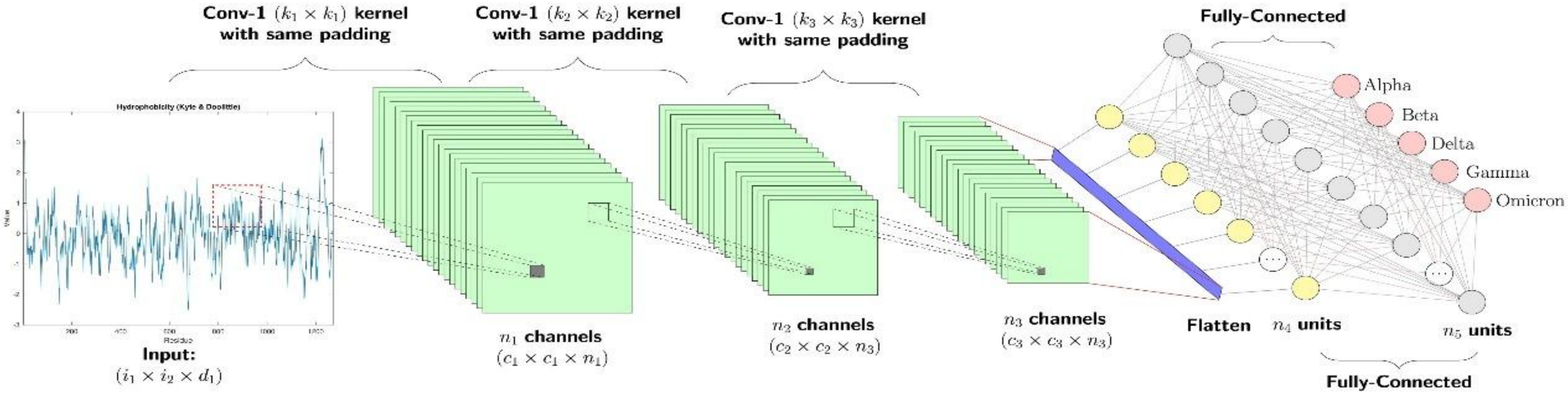
General model analysis



Neural Network Flow Diagram (General Model Flow)

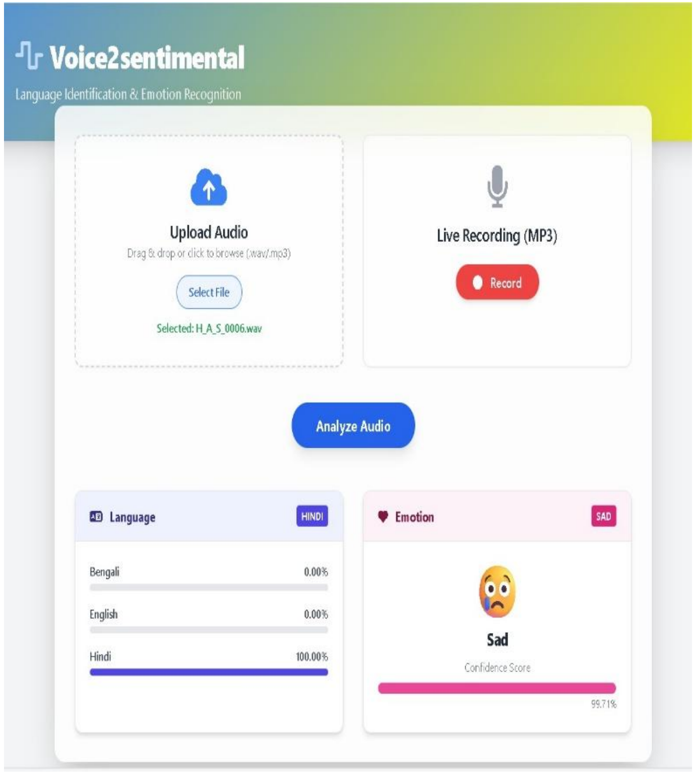


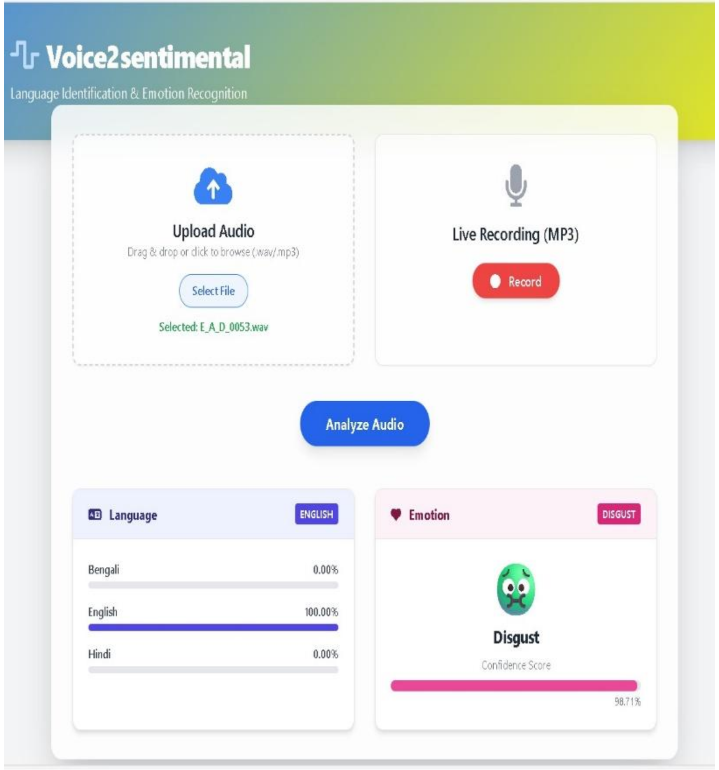
Detailed CNN Block Architecture (MFCC → Conv Blocks → Output)

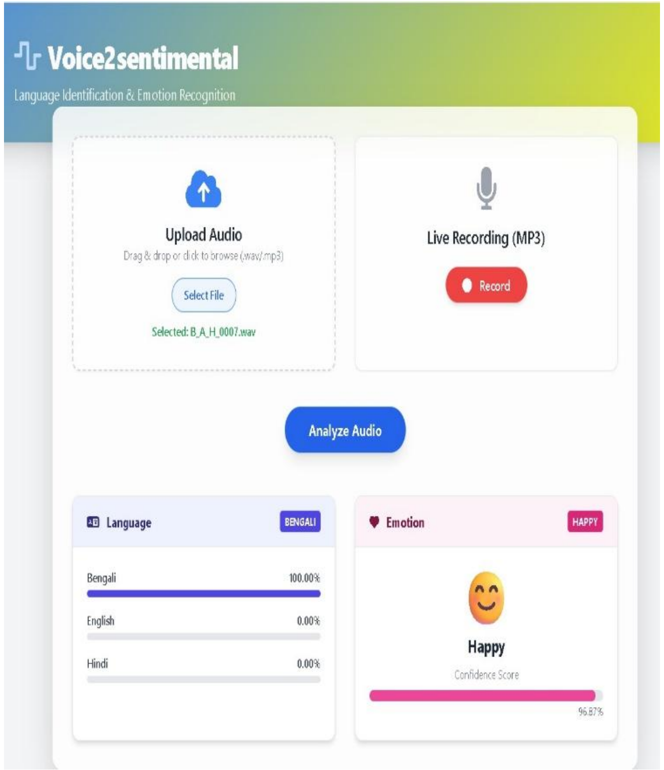


CNN Model Architecture (Audio Feature Learning)

UI Design-Application







Testing Results

- **Tables, charts, accuracy scores**

Emotion	Precision	Recall	F1-score
Angry	0.94	0.94	0.94
Disgust	0.87	0.91	0.89
Fear	0.94	0.82	0.87
Happy	0.92	0.88	0.90
Neutral	0.90	0.94	0.92
Sad	0.85	0.91	0.88

Test Case ID	Input Type	Expected Output	Result
TC1	Clean audio (single speaker)	Detect correct language & emotion	Passed
TC2	Noisy audio	Output with slightly lower confidence	Passed
TC3	Different accents	Stable predictions	Passed
TC4	Very short audio (<1 sec)	Low confidence or error message	Passed
TC5	Live microphone input	Real-time output	Passed
TC6	Mixed-language utterances	Dominant language predicted	Passed

Metric	Score
Total Accuracy	90.12%
Weighted Precision	90.28%
Macro Precision / Recall / F1	0.90 / 0.90 / 0.90

Performance comparisons

Model	Accuracy	Notes
Traditional MFCC + SVM	72–78%	Poor generalization
LSTM (baseline)	80–84%	Better, but slower & unstable
Proposed CNN Model	**90.12%**	Highest accuracy, best precision, fast inference

```
==================================================
OVERALL CLASSIFICATION REPORT
==================================================
Total Accuracy:  90.12%
Weighted Precision: 90.28%
--------------------------------------------------
              precision    recall  f1-score   support

       angry       0.94      0.94      0.94      2570
     disgust       0.87      0.91      0.89      2278
        fear       0.94      0.82      0.87      2443
       happy       0.92      0.88      0.90      2791
     neutral       0.90      0.94      0.92      3241
         sad       0.85      0.91      0.88      2781

    accuracy                           0.90     16104
   macro avg       0.90      0.90      0.90     16104
weighted avg       0.90      0.90      0.90     16104

Confusion Matrix:
[[2415   52   16   40   33   14]
 [  28 2062   18   37   58   75]
 [  38   95 2000   54   42  214]
 [  83   68   48 2449  101   42]
 [  12   28    6   40 3059   96]
 [   4   72   46   40   91 2528]]
```

Further implementation scope of your System

Why this system is better :

- It combines voice emotion and text sentiment, providing better results than systems that only use one method
- It performs well in real-time, even in noisy environments and supports multiple Indian languages
- It detects deeper signals like stress, sarcasm, and frustration that most systems miss

Benefits for users :

- It offers more natural and emotionally aware AI interactions
- Users can get a quicker sense of mood in customer care or online services
- It provides valuable insights into user emotions for teachers, doctors, and managers

Future market potential :

- There is high demand in call centers, telemedicine, mental health apps, and education technology
- The use of emotion-aware virtual assistants like Alexa, Siri, and chatbots is growing
- It can be useful in HR interviews, online tests, and security monitoring

References

SUPREME KNOWLEDGE FOUNDATION

Dataset link

- https://www.kaggle.com/datasets/evilspirit05/emotion
- Speech emotion recognition Hindi : https://share.google/bsH2mgPN9cJnL7mWQ
- Hindi Speech Recognition : https://share.google/RPWxU50WGvELVEc3E
- https://drive.google.com/drive/folders/1KNC9BxJ2bsKVmQJC6FZ55X9skic8X9-V

Research papers

[1] Z. Aldeneh and E. Mower Provost, "Detecting depression in speech using spectral harmonics," in *Proc. INTERSPEECH*, 2017.

[2] S. Anand and S. R. Patra, "Voice and text based sentiment analysis using natural language processing," in *Cognitive Informatics and Soft Computing*, Singapore: Springer, 2022, pp. 517–529.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

[4] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, 2008.

[5] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014.

[6] J. Deng, Z. Zhang, and B. Schuller, "Deep learning for emotion recognition in speech," *IEEE Trans. Affective Comput.*, 2019.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[8] H. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, 2017.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

SUPREME KNOWLEDGE FOUNDATION

Thank You