Dissertation Title:

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

Subject code: PROJ – AIML 781

Subject : Project – I

Dissertation Done by :

Student Name & Roll No: Debolina Samanta [25330822008],

Jamima Khatun[25330822009],

Uma Saha [25330822021],

Aditya Choudhary [25330822026]

Guided by : Dr. Dhrubasish Sarkar & Sonali Das

Degree Program : B.Tech in Computer Science Engineering

(Specialization in Artificial Intelligence and Machine Learning)

Research Area : *Multimodal Sentiment Analysis using Speech and Text in the field of Artificial Intelligence and Human-Computer Interaction*

Dissertation carried out at :

Supreme Knowledge Foundation Group of Institutions, Hooghly, West Bengal

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL, PIN- 741249

Contents

Broad Area of Work :

This project falls under the broader domain of Artificial Intelligence (AI), specifically intersecting the field of Speech Processing, Emotion Recognition, and Natural Language Processing (NLP). The main objective is to develop a system that can understand human emotions through both speech and textual input simultaneously.

Primarily, the work involves analyzing voice signals to recognize emotions based on features such as tone, pitch, and intensity. On the other side, spoken content is transformed into text and analyzed for its sentiment — whether it is positive, negative, or neutral.

A unique aspect of this system is detecting cases where the spoken tone and textual meaning contradict each other, such as sarcasm, disguised insults, or fake praise, providing a more realistic understanding of the speaker's intent.

By combining these two analyses, we create a system capable of providing a complete and context-aware understanding of the speaker's emotional state. This is useful in human–computer interaction, virtual assistants, mental health analysis, and customer service sentiment tracking.

1. Data mining or Speech Processing: Extraction of meaningful features from audio signals. Speech to text conversion using automatic speech recognition. Speech emotion recognition: feature extraction like chroma features, zero crossing rate, MFCC.
2. Machine learning: Classification used for map audio to emotion, decision trees, SVM, k-NN, Naïve Bayes, Random Forests, ANN and clustering used k-Means, Hierarchical clustering. Use PCA to remove redundant feature and LDA for enhance class separability.
3. Time- series Modeling: The sequential and temporal nature of speech, the system incorporates Hidden Markov Models (HMMs) to effectively model and predict emotion transitions over time. HMMs are particularly suited for capturing the dynamic nature of vocal expressions and have proven effective in speech-based applications.
4. Application sustainability: This system promotes sustainability by supporting mental health monitoring and emotionally aware human-computer interaction.

Background :

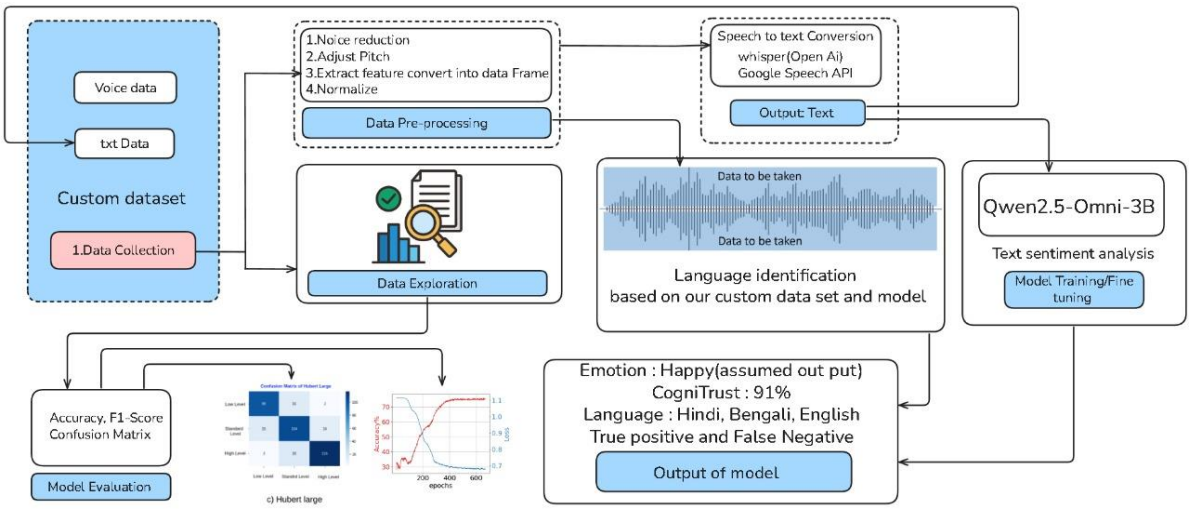
The growing demand for emotionally intelligent based on artificial intelligence systems has prompted researchers to explore sentiment and emotion recognition in human communication. Traditionally , Speech Emotion Recognition (SER) and Textual Sentiment Analysis have been treated as independent tasks, relying on either acoustic signals or textual content for interpreting user emotions. Sehgal et al.[1] highlight the necessity of emotional analysis in real-time communication by describing an interactive voice response (IVR) system that incorporates analysis into speech recognition.. Also in Pathak et al. [2] introduced an emotion-aware and text-to-speech system where it integrates sentiment recognizer to generate more speech with expression. Similarly, Anand and Patra [3] demonstrated how voice and text inputs can be separately processed using Natural Language Processing (NLP) techniques to derive sentiment, showing promise but still limited to unimodal analysis. Despite advances in these areas, real-world emotional expression is inherently multimodal, where voice tone and spoken content often convey complementary or even contrasting sentiments. Several recent studies have begun exploring cross-modal approaches. Another key limitation of past research is the lack of a unified, real-time pipeline that integrates automatic speech recognition (ASR), speech emotion detection, and textual sentiment classification in a synchronized manner. The Voice2Sentiment system is proposed a new end-to-end implementation that bridges this gap that unifies speech emotion recognition, speech-to-text conversion, and textual sentiment analysis into one integrated framework. Unlike prior fragmented models, this system captures prosodic features like pitch, tone, MFCC and linguistic sentiment using real-time automatic speech recognition(ASR), followed by dual-path emotion detection. Inspired by the project like Vāc [4], which questioned whether a Text-to-Speech (TTS) engine could replicate emotional intent , and by Kumari and Sowjanya [5] , who proposed a single framework for text, image, and voice sentiment mining, this project integrates multimodal input streams into a cohesive system capable of detecting emotion from audio and text in real time .

Additionally, Singh [6] and Rao et al.[8] emphasized the growing role of machine learning and deep learning in predicting voice -based sentiment, further supporting the need for unified systems like Voice2Sentiment. While Kim et al.[7] proposed a multimodal feature learning framework (MSDLF

-K) that combines speech and text for sentiment detection , showing superior performance over single modality models.

In this project we focus on the advancement of emotional aware systems has significantly shaped field of Artificial Intelligence. Traditionally Speech Emotion Recognition (SER) and Textual Sentiment Analysis have been explored separate research problems. While SER focuses on identifying emotions using features such as pitch , energy, and tone. Textual analysis evaluates the sentiment of written or transcribed speech content. However , as emotion are often conveyed jointly through both voice and language, analyzing them in isolation limits the systems interpretability. To overcome this problem , this Voice2Sentiment project introduces a new end-to-end multimodal implementation that combines speech emotion recognition , automatic speech recognition , and text-based sentiment analysis in a single intelligent framework. This system first analyses voice signals to detect emotional states based on acoustic features like chroma features, zero-crossing rate, and MFCC(Mel Frequency Cepstral Coefficients). The spoken content is converted into text using ASR, after which sentiment classification determines the polarity (positive, negative, or neutral) of the transcribed message. By integrating both modalities , the system provides a more comprehensive understanding of the speaker's emotional state.

The overview of the methodology of the project :



Objectives :

1. In this project we develop a voice input processing system to measure clarity, recognize emotions, extract and optimize text with sentiment and save outcomes.
2. The proposed system uses existing methods to improve the voice

quality, recognizing emotions, converting speech to text and analyze it with LLMs.

3. We try to achieve the accuracy rate of at least 70-80% for emotion detection from clear voice input.
4. We are focusing on a Word Error Rate (WER) for not more than 10% while extracting the text from the noise clear speech.
5. Process and analyze at least 50 voice inputs per hour with an average response time of under 10 seconds per input.
6. Save all processed text data and emotions securely in a database.
7. Utilize existing datasets to train and test system using machine learning prediction like classification and clustering as well.
8. The work progress focus on Natural Language Processing (NLP), speech processing, machine learning, creating a system that combines these features is realistic.
9. Implement a modular architecture that supports scalability and real-time deployment.
10. Machine learning evaluation metrics like F1-score, confusion matrix, AUC-ROC to assess system performance and fine-tune models iteratively.
11. The system supports that the user feedback loops to enabling continuous learning and improve the emotion detection accuracy over time.
12. Detect mismatched tone and meaning (e.g., sarcasm, irony, or hidden emotions) by combining speech tone features with textual sentiment for more accurate sentiment interpretation.
13. This system will recognize not only English but also the other languages like Bengali, Hindi as well.
14. We use the real-time voice dataset for better accuracy and model prediction.

Scope of Work

The “Voice2Sentiment” project focuses on developing and designing a comprehensive, real-time system that captures, processes, and interprets emotional content from human voice inputs. The system integrates advanced “speech processing”, “natural language processing” and “machine learning” techniques or methods to recognize emotions and analyze sentiments from spoken content.

The project includes :

1. ***Voice input Acquisition*** *:* Capturing clear voice data from user using various resource that will support real-time data input.
2. ***Noise Reduction and Clarity Enhancement*** : Preprocessing audio improve pitch, mid, travel etc.. . if required (Generally required).
3. ***Automatic Speech Recognition (ASR):*** Converting speech into dataframe using real-time transcription while maintaining low loss .
4. ***Feature Extraction*** *:* Extracting audio , tome , MFCC, pitch, zero-crossing rate, and chroma for emotion recognition.
5. ***Text(dataframe) Sentiment Analysis*** *:* Analysing transcribed text using LLMs or transformer-based NLP models to classify sentiment as positive, negative, or neutral.
6. ***Emotion Recognition*** *:* Classifying emotions from audio features using machine learning algorithms like SVM, Random Forest, ANN, and time-series models such as HMM.
7. ***Tone–Meaning Mismatch Detection:*** Detect sarcasm, irony, and hidden emotions by combining speech tone features with textual sentiment.
8. ***Business Applications***: Customer feedback analysis, call center emotion monitoring, and brand/social media sentiment tracking.
9. ***Mental Health Applications:*** Detect stress, anxiety, or depression, monitor emotional well-being in therapy, and provide real-time distress alerts.
10. ***Multi-Language Support***: Hindi, Bengali, English, with scope for more languages.

Literature References

[1] Sehgal RR, Agarwal S, Raj G. Interactive voice response using sentiment analysis in automatic speech recognition systems. In2018 International Conference on Advances in Computing and Communication Engineering (ICACCE) 2018 Jun 22 (pp. 213-218). IEEE.

[2] Pathak A, Majethia H, Singhall B, Bhor S, Venkatesan M. Emotion-Aware Text to Speech: Bridging Sentiment Analysis and Voice Synthesis. In2024 3rd International Conference for Innovation in Technology (INOCON) 2024 Mar 1 (pp. 1-7). IEEE.

[3] Anand S, Patra SR. Voice and text based sentiment analysis using natural language processing. InCognitive Informatics and Soft Computing: Proceeding of CISC 2021 2022 May 31 (pp. 517-529). Singapore: Springer Nature Singapore.

[4] Kulkarni S, Barbado L, Hosier J, Zhou Y, Rajagopalan S, Gurbani VK. Project Vāc: Can a Text-to-Speech Engine Generate Human Sentiments?. In2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) 2021 Oct 13 (pp. 103-108). IEEE.

[5] Kumari G, Sowjanya AM. An integrated single framework for text, image and voice for sentiment mining of social media posts. Revue d'Intelligence Artificielle. 2022 Jun 1;36(3):381.

[6] Singh AK. Prediction of voice sentiment using machine learning technique. In2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART) 2021 Dec 10 (pp. 162-166). IEEE.

[7] Kim TY, Yang J, Park E. Msdlf-k: A multimodal feature learning approach for sentiment analysis in korean incorporating text and speech. IEEE Transactions on Multimedia. 2024 Dec 24.

[8] Rao A, Ahuja A, Kansara S, Patel V. Sentiment analysis on user-generated video, audio and text. In2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) 2021 Feb 19 (pp. 24-28). IEEE.