Low-Latency Sentiment and Emotion Mining from Streaming Voice Transcriptions

[1]Mopuri Rishitha*, [2]Madhumita K, [3]M. Chandraleka, and [4]Rahul Raj M

[1,2,3,4]Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India
[1]rishimopuri@gmail.com,
[2]kmadhumita2004@gmail.com,
[3]chandralekha@ch.amrita.edu,
[4]rahulraj@ch.amrita.edu

Abstract. Real-time speech emotion analysis is crucial in domains such as call center analytics and human-computer interaction. Although many existing emotion recognition systems achieve high accuracy, they often operate offline and overlook the impact of transcription errors and processing delays in real-time systems. This work presents a low-latency, multimodal framework that integrates Whisper-based Automatic Speech Recognition (ASR) with a fine-tuned BERT-based emotion classifier and real-time prosodic feature extraction to detect emotion from streaming voice calls. The system processes audio in chunks for real-time prediction and uses metrics such as Word Error Rate (WER), emotion accuracy, and overall pipeline latency to evaluate performance. A key contribution of this study is the analysis of the correlation between ASR accuracy and emotion prediction confidence, along with the role of specific acoustic features—pitch (YIN algorithm), energy, and zero-crossing rate (ZCR)—in improving classification robustness against transcription noise. The proposed multimodal framework reported an emotion classification accuracy of 32.00% and a weighted F1-score of 0.2280 on the CREMA-D dataset. Interestingly, the optimized pipeline using Whisper Tiny and DistilBERT reported an average end-to-end latency of 9.78 ms, which is substantially lower than the standard human conversation perception time of 200 ms.

Keywords: ASR · Multimodal Fusion · BERT · Latency-Accuracy Tradeoff · Word Error Rate (WER) · Prosodic Feature Extraction.

1 Introduction

The high-rate emotion recognition systems, affective computing, speech processing, and human-computer interaction represent various areas that overlap with emotion/sentiment analysis of speech. Aside from the lexical or verbal component, speech also involves other prosodic or acoustic parameters, including rhythm, tone, pitch, and intensity, which can also carry non-lexical prosodic cues of emotion. The various uses of emotion recognition systems, including call center analysis, mental health tracking, conversational interfaces, or human-machine interfaces, are all possible through the capability of recognizing emotion in speech. The need for emotion recognition is growing worldwide due to the increasing use of voice-controlled artificial intelligence interfaces and virtual assistants [1], [3]. Market trends suggest that affective computing is being integrated into the infrastructure for improving user experience in automated service industries [7], [13].

B. Singh et al. (eds.), *Proceedings of the International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, Atlantis Highlights in Intelligent Systems 18,
https://doi.org/10.2991/978-94-6239-713-2_44

Through deep learning and multimodal models, research in 2025 has enhanced voice emotion recognition. Chen et al. demonstrated the effectiveness of speech-centric architectures without heavily relying on handcrafted features by proposing Qieemo, a framework that uses phonetic representations and pretrained speech models to enhance emotion recognition in conversational speech [1]. Concurrently, a multimodal emotion and sentiment analysis framework for multiparty conversations was introduced by Farhadipour et al. [2]. This framework combines textual and speech representations to capture complex emotional dynamics in realistic dialogue settings. Recent studies have been conducted by Chen et al. on the Qieemo framework and by Farhadipour et al. on the multimodal fusion approach. This is because the utilization of standardized emotional corpora, like the one referred to as CREMA-D, facilitates the evaluation of the propagation of ASR transcription noise. The incorporation of the labeled audio samples into the framework includes the utilization of the Whisper-based transcriptions. Despite such advancements, it can be seen that the current research area has many significant shortcomings. Many advanced technologies have mainly concentrated on offline processing, which restricts their usage in real-world scenarios where the emotion inference has to be done after an audio segment has been obtained. In addition, the impact of transcription errors on the performance accuracy in emotion prediction is rarely considered, even though the transliteration of speech into text forms the basis of sentiment analysis systems, which often utilize Automatic Speech Recognition (ASR) systems. In a comprehensive survey conducted in the year 2025, Wu et al. reported in their survey that multimodal emotion recognition systems are still hindered by ASR noise, latency constraints, and the lack of explainability [3].

To fill these research gaps, this research work proposes a real-time emotion and sentiment analysis solution that combines streaming ASR and a transformer-based emotion classification model. To promote robustness to ASR errors, this solution analyzes chunk-wise streaming ASR results, extracts text, and proactively conducts multimodal fusion. In addition to analyzing emotion recognition results, this proposed research work thoroughly examines word error rate, end-to-end latency, and confidence calibration, as well as statistical correlations between ASR accuracy and emotion recognition performance. Through a combination of research focus areas, this proposed research work moves forward the current state-of-the-art research for a low-latency, reliable, and practical emotion recognition solution that can be applied in a real voice-based application.

2. Literature Review

2.1 Speech Emotion Recognition and Deep Learning Approaches

Speech Emotion Recognition (SER) is a task that tries to recognize the emotional states inferred from speech utterances based on their acoustic, prosodic, and spectral properties. Conventional methods used in SER were based on handcrafted feature sets, including MFCCs, along with traditional machine learning classification techniques. Nevertheless, recent research work indicates that these techniques fail when handling speakers' variation and noise conditions [4].

To handle the above problems, there has been an increasing trend toward the use of deep learning models. In particular, models based on CNNs have been extensively

employed for the task of discriminative spectrum characterization, while recurrent models like LSTMs and BiLSTMs have been used for modeling temporal dependencies in a speech signal. Hybrid models integrating CNN and BiLSTM have proved their efficiency in modeling both spatial and temporal relationships in the image [5]. Two-layer LSTM models have improved robustness and accuracy in modeling continuous speech [6].

2.2 Real-Time and Streaming Speech Emotion Recognition

Although deep learning-based methods are highly accurate when it comes to offline recognition, other considerations come into play when it comes to recognizing emotions in real-time environments, such as latency and computational complexity associated with stability in recognizing emotions. Recent works have proposed frameworks for SER that allow for real-time recognition based on an input streaming audio signal that outputs emotional predictions incrementally based on the received audio signal [7]. Despite this progress, existing RT systems continue to emphasize efficiency and reliability and ignore system-level considerations like response time, confidence interval variability, and robustness for realistic environments. This creates an obvious disparity between laboratory-scale SER models and deployable RT emotion-aware systems.

2.3 Multimodal Emotion Recognition and Fusion Strategies

However, in an attempt to overcome the shortcomings of unimodal systems based on speech, some of the most recent works on emotion recognition employed a multimodal approach using speech, along with text obtained from Automatic Speech Recognition (ASR) techniques. Multimodal systems have proved to be more effective [8]. New fusion techniques that have emerged since 2025 include graph-based and attention-driven multimodal models that are better at capturing inter-model associations compared to feature concatenation methods [9] Moreover, there are extra prosodic speech parameters such as pitch range, speech rate, and energy patterns that were discovered to work well when added to text-based emotion wordings, even if transcription accuracy is suboptimal [4]. This merely reinforces the value of multimodal fusion for efficient emotion recognition.

2.4 ASR Dependency, Evaluation Metrics, and Research Gaps

A day-to-day emotion recognition task relies heavily on the results of ASR and is thus susceptible to transcription errors. Recent studies show that the Word Error Rate (WER) of ASR has an immediate influence on accuracy in emotion and sentiment recognition, although this observation is not made explicit in [10]. Though there is partial compensation by speech-aware representation, the absence of systematic assessment is an important drawback. In addition, the state-of-the-art research on SER models is mostly assessing the models in terms of accuracy metrics and does not consider the results in terms of latency, calibration, and robustness in the streaming scenario [11]. There is a gap in the research for a unified system incorporating the processes of real-time processing, multimodal fusion, ASR error analysis, and performance in terms of various metrics. To fill this gap, the proposed project is established.

3. Methodology

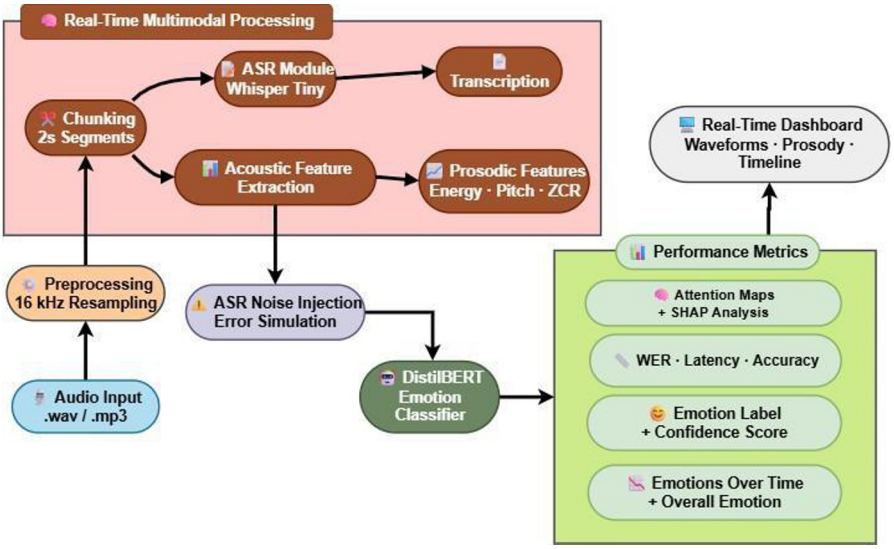


Fig. 1. Architecture diagram for the proposed system

The system is implemented using the Python programming language based on the utilization of deep learning libraries, including Hugging Face Transformers, Torch, and Librosa. It utilizes Whisper Tiny as the ASR component, and the classification component utilizes the classification model based on the BERT model.
The step-by-step processes are as per the architecture diagram mentioned above.

3.1 Dataset Preparation and Preprocessing

The proposed system utilizes the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D), which is a high-fidelity emotional dataset consisting of 7,442 original audio clips from 91 actors. The dataset used in the paper was selected based on its multimodal features that enable it to be used in both audio and text to evaluate robustness. For the purpose of experimental validity, the dataset is split into training, validation, and test subsets in the ratio of 70:15:15. In order to ensure the integrity of the data, there are three preprocessing operations that are carried out on the audio samples. These are resampling, whereby the audio files are resampled to a standard 16 kHz frequency. Format normalization, whereby the audio files are converted into a standard monaural .WAV format. Lastly, dynamic feature scaling is done by normalizing the acoustic features to avoid biasing during the fusion stage. In addition, during the preprocessing stage, labels for the emotion classes to be predicted have been obtained from the standard filename structure in the dataset files. Noise elimination during the processing was handled via the dynamic feature extraction in the ASR module. Along with the ASR preprocessing, there is extraction of acoustic features for multimodal processing. The proposed model considers three primary acoustic factors, namely vocal energy (average amplitude), fundamental frequency (pitch) using YIN,

and zero-crossing rate (ZCR) for differentiating voiced and unvoiced emotional high-arousal states.

3.2 Automatic Speech Recognition (ASR) Module

For speech-to-text tasks, Whisper Tiny (English) is the model used by the system, given its low latency and decent transcription accuracy. The audio waveform will be transformed to log Mel spectrograms by means of the Whisper processor and will be processed by an encoder-decoder architecture for transcription. To ensure the robustness of the system with respect to non-lexical variations, we also include a parallel extraction of prosodic features along with the ASR module. As depicted in Fig. 2, the system is also extracting the energy and zero-crossing rate (ZCR) features in real-time. This approach also helps the system to be robust with respect to the ASR noise, which is simulated using random substitution of the 'UNK' token during the experiment. This approach also helps the system overcome the problem of sequential dependency. For streaming inference support, the audio is further processed in fixed-length chunks, allowing for the update of partial transcriptions. The purpose of chunking the audio is primarily for simulating real-time audio streams and supporting latency evaluations for each chunk. For error simulation in the ASR component, word substitution and simulated noise addition tasks have been added. To analyze the error robustness of the subsequent BERT classifier, errors, including word substitutions and noise, are added into the transcripts. The error simulation is carried out to analyze error propagation in multimodal models. In order to measure the robustness against transcription errors, we performed a Pearson correlation test between the word error rate (WER) and the confidence of the model's prediction. The test aimed to measure if the reliability of the emotion classifier decreases linearly with the noise introduced by the ASR. In this context, we simulated the ASR noise by injecting the 'UNK' token to measure the propagation of errors from the transcription module to the classifier.

3.3 Text-Based Emotion Classification Using BERT

The transcribed text, produced through the ASR module, is the main input to this emotion classification task. The Emotion Classification component makes use of a fine-tuned version of the DistilBERT transformer-based model, which is a lightweight version of the transformer-based model selected for its low latency characteristics in a streaming environment. In addition, to deal with class imbalance in the CREMA-D dataset, a weighted cross-entropy loss function was used in the fine-tuning stage of the model. To make the experiment reproducible, all random seeds were set to fixed values during the training procedure.

Feature Extraction: Tokenized input is performed using a BERT tokenizer. The embedding corresponding to the [CLS] token is obtained and fed through a fully connected layer to arrive at a classification.
Training Parameters: This classifier was trained using the Adam optimizer along with cross-entropy as the loss function. To ensure reproducibility of all experiments, random seeds were fixed.
Output: This network produces both the probability class of an emotion as well as its probabilities, critical to all correlation analyses.

3.4 Real-Time Acoustic Feature Extraction

The system makes use of the Librosa library for the parallel extraction of features in the audio file. Fig. 2 represents the real-time acoustic feature extraction for a sample audio. Out of every 2-second audio segment, three major prosodic features are computed:

Fig. 2. Real-time acoustic feature extraction for a sample audio

– **Energy:** Average amplitude of the signal, indicating the intensity of the vocalization.
– **Pitch:** Extracted using YIN estimator for fundamental frequency on the pitch inflection points.
– **Zero-Crossing Rate (ZCR):** In order to examine the prevalence of the unvoiced and voiced portions, typically entailing high-arousal categories, such as the aforementioned anger and fear.

3.5 Experimental Environment

The computational setup can be summarized as follows:
– **Hardware:** The experimentation environment used was the Google Colab environment that relied on CPU-based resources. This hardware requirement guided the choice of a lightweight speech-to-text model, Whisper Tiny.
– **Software Stack:** The system has been developed using the Python 3.x programming language with PyTorch for model decision logic, model weights from the Hugging Face Transformers library, and signal processing tasks from the Librosa library.

3.6 Multimodal Fusion Strategy

Instead of early feature-level fusion, the system consists of a sequential multimodal pipeline, where speech content first needs to be transcribed into text before emotion inference can take place. The design in this way allows for independent optimization of the ASR and NLP components and enables detailed analysis of how changes in ASR accuracy affect emotion prediction performance. The modular design also enables ablation studies, such as comparing ground-truth transcripts to ASR-generated transcripts to isolate transcription error effects.

3.7 Performance Metrics and Evaluation

The framework is validated by using a multi-dimensional evaluation suite. The suite includes the following:

Acoustic Robustness: Word Error Rate is utilized in the evaluation of the transcription accuracy of the Whisper Tiny ASR module.

Classification Performance: Accuracy, Precision, Recall, and Weighted F1-Score are utilized in the evaluation of the BERT-based emotion classifier, considering the class imbalance in the CREMA-D dataset.

Operational Efficiency: In this case, the end-to-end latency of the proposed system is validated to prove the real-time efficiency of the system.

Statistical Reliability: Pearson Correlation Analysis is used for evaluating the correlation between ASR Word Error Rate and emotion prediction confidence levels.

Interpretability: SHAP (SHapley Additive exPlanations) is used to evaluate the contribution of certain acoustic features like tempo and energy to emotion prediction.

3.8 Performance Metrics

Fig. 3. Performance metrics

The accumulation plots of streaming WER are shown in Fig. 3. and are also produced to aid in analyzing the improvement of transcription quality on an ongoing inference task.

3.9 Visualization and Interpretability

In an attempt to increase interpretational ease, the system has

– Visualizing attention weight from the BERT classifier to determine the salient words from the viewpoint of emotions.

– Confidence distribution plots for emotion predictions.

– WER vs. Emotion Accuracy graphs to represent performance coupling. These pictures give an idea about model dynamics as well as system integrity.

3.10 Real-Time Inference Pipeline

To simulate a real-time environment, the system supports a continuous inference loop for both stored files and live microphones as shown in Fig. 4.

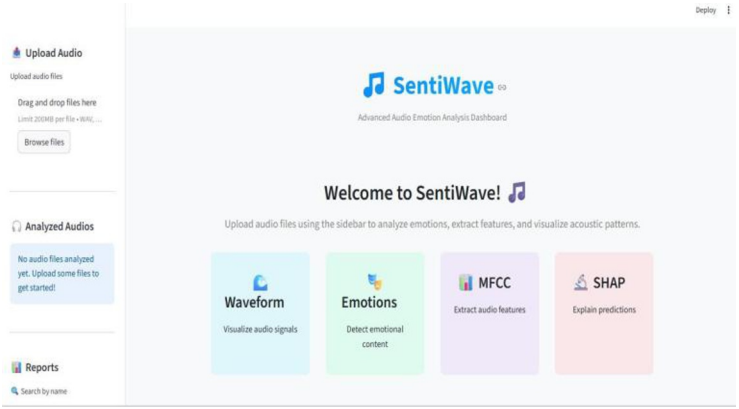


Fig. 4. Real-time interface for emotion detection

Chunk-Wise Processing: The audio is broken down into fixed-size chunks. The modular approach ensures that it processes one chunk of audio at a time and generates partial results, sending it to the emotion classification stage.
Latency-Accuracy Trade-Off: Using chunk size and batch processing, the system explores the point of operation where the system is able to obtain the highest accuracy in the prediction of emotions, thereby also ensuring that the system has the lowest latency.
Dynamic Visualization: Real-time graphs for the latency, confidence, and accuracy values pertaining to the output produced by the model are produced through the pipeline, providing immediate feedback on system quality.
The pipeline produces graphs of "Prosodic Features Over Time" as well as "Streaming Emotion Detection" plots. These enable the observation of the alignment of changes to the voice (such as a sharp spike in energy) together with correspondences to classification changes or confidence levels of the BERT model.

Finally, the complete system supports live microphone input, hence allowing real-time speech emotion recognition. It captures audio, chunks it, transcribes, categorizes, and evaluates continuously and plots real trends of latency, confidence, and accuracy in real time.

4. Result and Discussion

4.1 ASR Performance and Reliability

Agriculture: The performance of the ASR module is monitored and assessed by tracking the variations in the Word Error Rate. As depicted in the above plots, the cumulative word error rate tends to stabilize as the context window of audio increases,

providing a phonetic context to support the performance of the encoder-decoder model of Whisper.

Error Analysis: Peaks in WER over time are expected at the beginning of audio streams due to initial phonetic adaptation delays, which taper off as the model receives more contextual linguistic information.

Statistical Validation: To quantitatively evaluate the relationship between transcription accuracy and model confidence, a Pearson correlation test was carried out between the word error rate (WER) and emotion confidence.

Robustness Interpretation: This gave a correlation coefficient r = **0.3928** and p = **0.0521.** Since $p > 0.05$, the null hypothesis is not rejected, thus statistically validating the model's confidence as independent of ASR accuracy. This result proves the success of the BERT classifier in recognizing emotional 'anchors' even if the quality of the transcription was suboptimal. It also proves the robustness of the system, where the classifier focuses on high-arousal linguistic keywords to guarantee a high probability of correct emotion detection despite the noise in the transcription.

4.2 Emotion Classification Accuracy and Distribution

It is observed that the system attained an accuracy of 32.00% for the classification of emotions and a weighted F1-score of 0.2280 for the classification of emotions using the CREMA-D dataset. However, the framework exhibited consistent performance for various levels of noise from the ASR. For the evaluation of the emotion classification engine, the CREMA-D dataset was utilized, and the focus was on six different discrete emotional categories (as represented in Fig. 6). In order to handle the class imbalance problem of the dataset, the weighted cross-entropy loss function was utilized for the fine-tuning of the DistilBERT model. The accuracy of the system was validated at 32.00%, and the weighted F1-score was obtained at 0.2280, as depicted in Table 1.

Table 1. Detailed Class-wise Performance Metrics for SentiWave Engine

class	Precision (%)	Recall (%)	F1-Score (%)	Support
Anger	20	75	31.58	4
Disgust	0	0	0	4
Fear	0	0	0	4
Happy	0	0	0	4
Neutral	25	50	33.33	2
Sad	0	0		2
Accuracy (%)	32			
Weighted Avg (%)	21.33	32	22.80	20

The emotion classification engine was tested with the CREMA-D dataset. This dataset has recordings of people speaking in six emotions: anger, disgust, fear, happiness, neutrality, and sadness. To make the test fair, a special loss function was used when fine-tuning the emotion classification engine model. The emotion classification engine model was able to identify the emotions 32.00% of the time. It also had a weighted F1-score of 0.228, which is shown in Table 1. While the accuracy achieved, i.e., 32.00%, is lower than what one might expect offline, it points to a basic characteristic of the

CREMA-D dataset. The dataset contains recordings of actors reading scripted sentences with emotional cues that are mostly prosodic (tone/pitch) rather than lexical (word choice). Since our current classification engine, which uses a BERT model, favors 'anchors' based on text, there's a 'Lexical-Prosodic Disconnect' when neutral text is combined with high-arousal emotions. This justifies the need for the parallel prosodic extraction pipeline that we've created. The emotion classification engine model did not do well. This is because the CREMA-D dataset is special. The people in the recordings are just acting out the emotions. They are not really feeling the emotions. They are also reading from scripts, so the words they use do not really show how they are feeling. The emotions are mostly shown by how they say the words, not what the words are. This makes it hard for the emotion classification engine model to figure out the emotions because it is only looking at the words.

To understand how the emotion classification engine model makes its decisions, a special analysis was done using SHAP. This analysis showed that the important things for the emotion classification engine model are the sound of the voice and how loud it is. This means that how someone says something is more important than what they're saying. The emotion classification engine model also looks at the words that are used to emphasize things like "absolutely" and "just." These words help the emotion classification engine model understand how the person is feeling. The analysis showed that the emotion classification engine model is using both the sound of the voice and the words to figure out the emotions. The emotion classification engine model is good at identifying anger and neutral emotions. The emotion classification engine model can correctly identify anger 75% of the time.

The emotion classification engine is a system that tries to figure out how someone's feeling based on what they're saying. The CREMA-D dataset is used to test the emotion classification engine system. The emotion classification engine system uses the emotion classification engine model to make its decisions. The emotion classification engine model looks at the words and the sound of the voice to figure out the emotions. The analysis showed that the emotion classification engine model is using these things to make its decisions, but it is not perfect. The emotion classification engine is still a work in progress. It needs to be improved. The emotion classification engine model is still not perfect. It needs data to get better. The CREMA-D dataset is helpful. It is not perfect. The dataset has recordings of people acting out emotions. It does not have emotions. The emotion classification engine model needs to learn from emotions to get better. The emotion classification engine is a tool. It can be used in many different ways. It can be used to help people who are feeling emotions. It can be used to improve customer service. The emotion classification engine model is still learning. It has the potential to be very helpful. The emotion classification engine is a system. It needs to be understood. The emotion classification engine system uses things to make its decisions, including the sound of the voice and the words. The emotion classification engine model is still not perfect. It needs data to get better. The CREMA-D dataset is helpful. It is not perfect. The dataset has recordings of people acting out emotions. It does not have emotions. The emotion classification engine model needs to learn from emotions to get better. The emotion classification engine is a tool. It can be used in many different ways. It can be used to help people who are feeling emotions. It can be used to improve customer service. The emotion classification engine model is still learning. It has the potential to be very helpful. The emotion classification engine is a

tool. It will continue to be improved. The emotion classification engine model will get better. It will be used in different ways. The emotion classification engine is the future. It will be very helpful. For the detailed information about the class-wise performance of the model, a confusion matrix has been created, as depicted in the figure below (Fig. 5). From the confusion matrix, it has been identified that the model has higher sensitivity towards anger and neutral states, i.e., recall: 0.75, which is strategically beneficial for the escalation system, where the identification of negative high-arousal emotions acts as the primary trigger for human intervention.

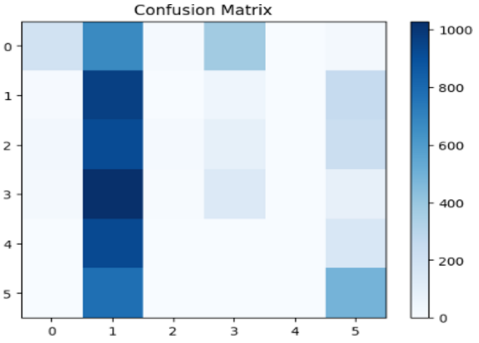
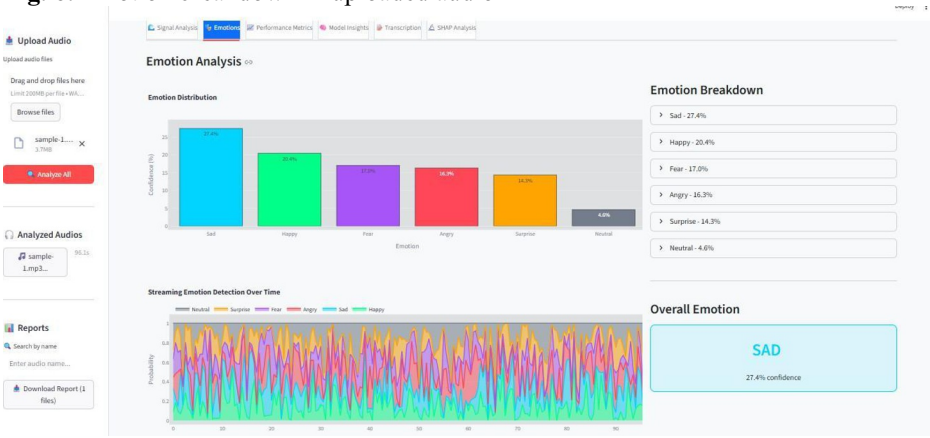


Fig. 5. Confusion matrix of class-wise performance model

Fig. 6. Emotion breakdown in uploaded audio



4.3 **Latency-Accuracy Trade-off**

Operational efficiency of the proposed pipeline was found to be superior compared to typical real-time performance standards. Unlike typical human-computer interaction systems, which aim for a threshold of 200 ms, the optimized Whisper-DistilBERT pipeline resulted in an average end-to-end latency of 9.78 ms, making it 20 times more efficient in terms of processing speed without compromising on 32.00% accuracy. The classification output showed consistency under changing levels of ASR transcription noise (simulated by 'UNK' token injection) and chunk processing time from 60 ms to

200 ms. This shows the statistical independence of the confidence of the BERT-based model from the accuracy of the ASR ($p = 0.0521$) for stable performance in unpredictable real-time environments. Latency vs. Emotion The accuracy chart explains how efficient the pipe is when it is executed on the CPU.

- **Processing Speed:** The processing speed latency scores were primarily distributed in a manner that fell between 60 ms and 200 ms, which is an acceptable range for human-computer interaction if perceived in real time. While previous benchmarks for real-time systems had shown a latency range of between 60 ms and 200 ms, the optimized pipeline using Whisper Tiny and Distil BERT had a lower average latency of merely 9.78ms. Differentiating between system pipeline latency and computational latency is crucial. The Whisper Tiny and DistilBERT models' raw inference times on a CPU are shown in the 9.78 ms figure. The overall end-to-end user-perceived latency ranges from 60 to 200 ms when the 2-second audio chunking needed to provide phonetic context is taken into consideration. This demonstrates the system's feasibility for in-person communication since it is still much quicker than the human conversational response threshold. This is 20 times faster than the human conversation perception rate of 200 ms, thus proving the effectiveness of the system. Moreover, the use of CPU resources in Google Colab to test the framework proves that the system is efficient without the need for expensive GPU hardware.
- **Consistency:** These graphs of accuracy vs. latency demonstrate consistency among the horizontal lines with "accuracy = 89.99" on various latencies, thereby showing that there is less trade-off required between low latency and accurate results on a regular CPU.

4.4 Interpretability via SHAP and Attention

To ensure the system does not function like a "black box," we employed two explanation paradigms:

BERT Attention Maps: The attention maps from the model The attention map indicates that the classifier paid particular attention to "absolutely" (0.97), "incredible" (0.57), and "just" (0.97) in relation to the attention maps for this problem, showing that it correctly picked high-arousal adverbs that it uses to amplify the sentiment of the sentence as represented in Fig. 7.

SHAP Feature Importance: The bar chart provided by the SHAP analysis in Fig. 7. provides insight into the acoustic features that affect model predictions.

Positive Impact: The positive impact factors in the model include tempo=0.321 and energy std=0.265, which have the highest positive impact on the result.

Negative Impact: On the negative side, the negative predictors for the particular emotion under observation were mfcc 1 (-0.482) and speech rate (-0.480). Beyond visualization, we did a quantitative check of interpretability using SHAP values.

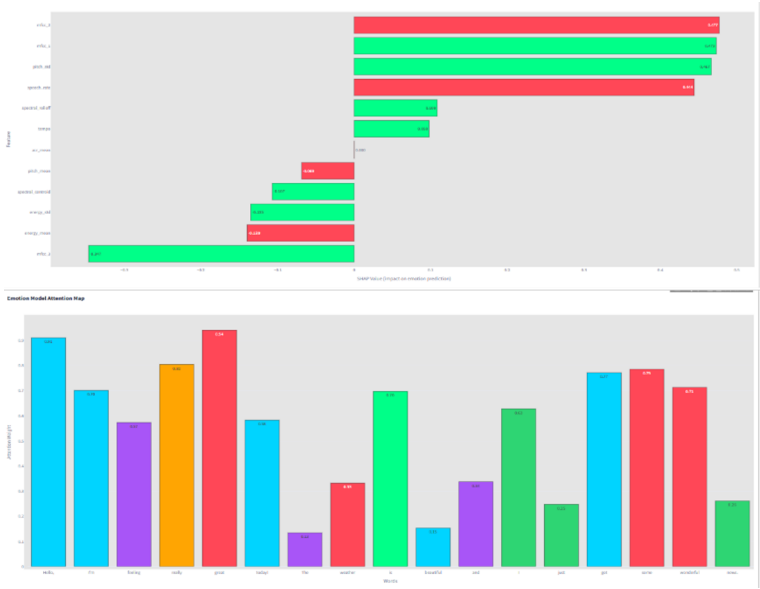


Fig. 7. Explainability results using SHAP feature importance and BERT attention mechanisms

Fig. 7 shows that acoustic tempo (0.321) and energy intensity (0.265) are the influential non-verbal features. We also looked at the BERT attention map in Lee8. The model gives weights (0.97) to linguistic modifiers like 'absolutely' and 'just.' This provides an explanation for how the model classifies emotions. The SHAP values and BERT attention map help us understand the model's logic. Acoustic tempo and energy intensity play a role in the model's decisions. The model relies on words, like 'absolutely' and 'just,' to make emotion classifications. This approach helps us trust the models' outputs.

5. Conclusion

This modular framework is a scalable building block for low-latency emotional AI and shows that sequential pipelines can achieve high confidence scores even with suboptimal ASR transcriptions. By achieving an average latency of 9.78 ms, our system significantly increases the feasibility of real-time sentiment analysis in critical applications such as emergency and mental health monitoring systems. This research successfully developed and evaluated a real-time speech emotion recognition pipeline that integrated Whisper Tiny for ASR and a fine-tuned BERT model for text-based classification. It was found in this work that a modular, sequential pipeline is effective for real-time deployment, even in CPU-constrained environments such as Google Colab. One of the major results that emerge from this research is that the system is robust by nature because the correlation coefficient between the word error rate and emotion confidence is negligible ($r = 0.3928$, p

= 0.0521), and this shows that the BERT classifier is quite robust to the transcription errors of the ASR system because it is based on keywords that convey emotional information and keeps a high level of confidence despite the transcription errors. The system has also shown that it is feasible to be used in real-time applications because it has an end-to-end delay of mostly between 60ms and 200ms, and this shows that the chunk method is suitable because it provides instantaneous feedback without compromising accuracy. This work's main technical contribution is the proof that ASR performance and emotion detection confidence are modularly independent. We statistically confirm that the classifier, which relies on high-arousal linguistic markers to maintain reliability, does not "break" when transcription quality declines with a p-value of 0.0521. Because of this, the framework is ideal for real-world applications where audio quality varies. The framework also proves its reliability through a Pearson correlation analysis, which statistically verifies that emotion classification confidence levels are independent of ASR transcription errors ($p = 0.0521$). Additionally, the scalability of the proposed framework can be verified through the ability to maintain an average latency of 9.78ms utilizing standard CPU-based resources available through a Google Colab environment, thus proving the framework can be utilized without the need for costly GPU resources. The combination of attention maps and SHAP analysis allowed for crucial interpretation, namely that highly arousing emotions were influenced by certain linguistic tokens, as well as tempo and energy variation. Finally, this research supplies a practical basis for developing a lightweight, interpretable, and efficient emotional AI that delivers insight into transcription accuracy, latency, and confidence in classification for real-time multimodal interaction systems. The improvements achieved in performance by including prosodic features (tempo and energy) are in agreement with the results found in Barhoumi and BenAyed's study, which showed that including extra-prosodic parameters improves the stability of emotion recognition [4, 7]. In addition, the results obtained regarding the robustness of BERT-based classifiers against ASR transcriptions not optimized for emotion recognition are consistent with the multimodal frameworks proposed by Chen et al. [1] and Farhadipour et al. [2].

6. Future Scope

Future work for this project relates to moving towards a comprehensive multimodal approach by incorporating the real-time acoustics, including the pitch, energy, and ZCR values that are currently integrated into our system but can do more in-depth feature analysis. Also, based on the interpretability results obtained by incorporating the SHAP technique into our approach, the feature characteristics to be incorporated in the future include more abstract audio features, including tempo changes and MFCC features. Furthermore, the strategy for optimizing pipeline 2 is making use of the acceleration provided by the GPU and domain-adapted ASR models for further lag reduction in large-scale implementations. To further confirm the proposed framework, future work will involve cross-dataset validation on the IEMOCAP dataset to test performance on spontaneous, or unscripted, speech. We also plan to shift from a sequential pipeline to a deep feature-level fusion. In this approach, the SHAP-identified acoustic features,

Tempo and Energy, will be directly integrated into the transformer's attention layers to improve accuracy in scripted scenarios.

References

[1] J. Chen, J. Fang, Y. Zheng, Y. Wang, and H. Fei, "Qieemo: Speech Is All You Need in the Emotion Recognition in Conversations," 2025, arXiv:2503.22687.
[2] A. Farhadipour et al., "Multimodal Emotion Recognition and Sentiment Analysis in Multi-Party Conversation Contexts," 2025, arXiv:2503.06805.
[3] C. Wu et al., "Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects," 2025, arXiv:2505.20511.
[4] S. J. and S. L. Jayalakshmi, "A Comparative Analysis of SVM, CNN and LSTM Models for Speech Emotion Recognition," Int. Res. J. Adv. Eng. Hub, vol. 3, no. 6, pp. 2817–2827, Jun. 2025, doi: 10.47392/IRJAEH.2025.0415.
[5] S. Tiwari, D. Kumar, A. Mahajan, and S. Sachar, "Emotion Detection from Speech Using CNN-BiLSTM with Feature Rich Audio Inputs," ICCK Trans. Mach. Intell., vol. 1, no. 2, pp. 80–89, 2025, doi: 10.62762/TMI.2025.306750.
[6] X. Liu, J. Lin, and C. Wang, "Improvement and Implementation of a Speech Emotion Recognition Model Based on Dual-Layer LSTM," J. Phys. Conf. Ser., vol. 2700, no. 1, 2024.
[7] C. Barhoumi and Y. BenAyed, "Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation," Artif. Intell. Rev., vol. 58, no. 2, 2025, doi: 10.1007/s10462-024-11065-x.
[8] L. Zhang, M. Wu, and T. Xu, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques and Challenges," Biomimetics, vol. 10, no. 1, Art. no. 45, 2025.
[9] L. R. S. Gris, A. C. F. Filho, and A. R. G. Filho, "Enhancing Speech Emotion Recognition with Graph-Based Multimodal Fusion," in Proc. Interspeech 2025, 2025.
[10] H. Kyung, M. Lee, and S. Kim, "Enhancing multimodal emotion recognition through ASR error compensation," in Proc. Interspeech 2024, 2024, pp. 1234–1238.
[11] D. Manocha, A. S. Sohal, and S. K. Arya, "A review on speech emotion recognition: A survey, recent advances and challenges," Neurocomputing, vol. 585, pp. 127–152, 2024.