© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715448

Multimodal Tone Sensitivity Analyzer for Customer Care Services

J. IMMANUEL KEVIN[1], DR.K. PONMOZHI[2]
[1]*PG Student, Department of Computer Application, SRM Valliammai Engineering College, Kattankulathur, Chennai, Tamil Nadu, India*
[2]*Associate Professor, Department of Computer Applications, SRM Valliammai Engineering college, Kattankulathur. Chennai, Tamil Nadu, India*

Abstract- ***Customer care services in many organizations lack real-time and objective evaluation of agent performance, often relying on manual auditing of recorded calls and basic text-based analysis. This paper presents a Multimodal Tone Sensitivity Analyzer, an AI-driven platform that addresses these limitations through integrated analysis of speech and textual data. The system consists of key components including audio input processing, speech-to-text transcription, natural language sentiment analysis, and acoustic emotion detection. The frontend interface enables users to upload or record calls, while the backend processes data using advanced AI models for transcription and evaluation. Audio signals are analyzed to detect emotional tones such as happiness, anger, sadness, and neutrality, while textual transcripts are evaluated for communication quality and agent effectiveness. A multimodal evaluation module combines these outputs to generate performance insights and structured reports. The system also supports scalable data storage and efficient retrieval of analysis results. This approach provides an automated, accurate, and cost-effective solution for improving customer service quality monitoring in modern call center environments.***

Index Terms- ***Multimodal Analysis, Speech Recognition, Sentiment Analysis, Voice Emotion Detection, Natural Language Processing, Customer Care Analytics, Call Center Monitoring, Acoustic Analysis, AI-Based Evaluation, Performance Analysis, Real-Time Processing, Cloud-Based System, Data Analytics, Automated Auditing***

I. INTRODUCTION

Customer care services in many organizations still depend on manual auditing of recorded calls, static evaluation criteria, and limited mechanisms to understand real-time agent performance. Supervisors often rely on reviewing a small sample of conversations, which results in inconsistent evaluation and lack of comprehensive insights. Additionally, traditional systems provide minimal support for analyzing emotional tone, leading to incomplete understanding of customer-agent interactions and reduced service quality.

A unified intelligent platform that can analyze both speech content and emotional tone, while providing real-time insights into customer interactions, can address these limitations effectively. The challenge lies in developing a system that integrates speech recognition, sentiment analysis, and voice emotion detection without requiring complex infrastructure or expensive hardware, making it suitable for academic and practical deployment.

The proposed Multimodal Tone Sensitivity Analyzer adopts a scalable and efficient approach. The system operates using standard computing environments and leverages cloud-based AI technologies for processing and analysis. Audio recordings are transcribed using speech recognition models, while natural language processing techniques evaluate sentiment and communication quality. In parallel, acoustic analysis detects emotional tones from speech signals, enabling a deeper understanding of interaction dynamics.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the system architecture. Section IV explains the proposed methodology. Section V presents the feasibility study. Section VI discusses the data flow diagram. Section VII covers database design. Section VIII outlines system implementation details. Section

IX presents results and discussion. Section X concludes the paper with future enhancements.

II. LITERATURE REVIEW

A. Speech Recognition in Customer Service Systems

Early customer service analysis systems relied on manual transcription of recorded conversations, which introduced delays and inconsistencies in data processing. With advancements in Automatic Speech Recognition (ASR), modern systems are capable of converting speech into text with high accuracy and minimal latency. Deep learning-based models, particularly those using transformer architectures, have significantly improved transcription quality across multiple languages. Cloud-based speech recognition platforms further enhance scalability by enabling real-time processing of large volumes of audio data without requiring complex local infrastructure.

B. Sentiment Analysis and Natural Language Processing

Sentiment analysis techniques have been widely used to evaluate customer interactions by analyzing textual data. Traditional approaches utilized machine learning algorithms such as Naïve Bayes and Support Vector Machines, while recent methods employ deep learning models for improved contextual understanding. Natural Language Processing (NLP) enables systems to identify sentiment polarity, detect key phrases, and evaluate communication effectiveness. However, text-based analysis alone often fails to capture emotional nuances present in spoken communication.

C. Voice Emotion Recognition Techniques

Speech emotion recognition systems analyze acoustic features such as pitch, tone, and intensity to detect emotional states in voice signals. Techniques based on Mel-Frequency Cepstral Coefficients (MFCC) and deep neural networks have shown significant improvements in identifying emotions such as happiness, anger, sadness, and neutrality. These systems provide valuable insights into the emotional dynamics of conversations, which are essential for evaluating customer service quality.

D. Multimodal Analysis in Communication Systems

Recent research emphasizes the importance of multimodal analysis, which combines textual and acoustic data for more accurate interpretation of interactions. By integrating speech transcription with emotion detection, multimodal systems provide a comprehensive understanding of both the content and tone of communication. Such approaches have been shown to improve performance in applications such as customer support monitoring and conversational AI systems.

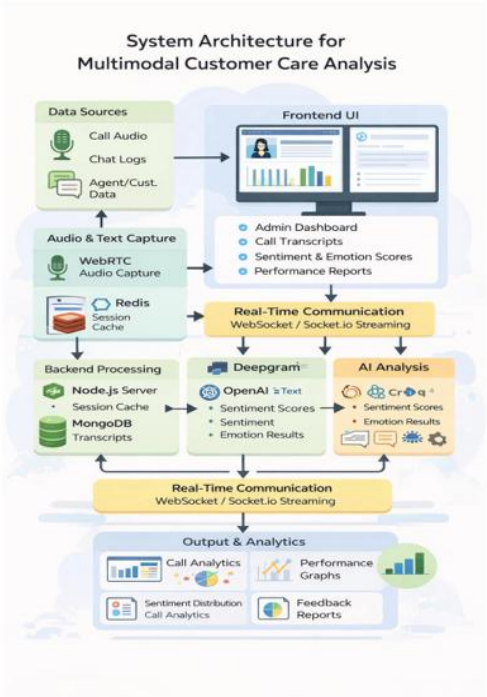
E. Cloud-Based Processing and Scalable Architectures

Cloud computing platforms have become widely adopted for processing and storing large volumes of interaction data. These platforms provide real-time data processing, scalability, and cost efficiency. The use of cloud-based architectures allows systems to handle multiple concurrent analyses and ensures seamless integration with existing customer service platforms, making them suitable for modern call center environments.

III. SYSTEM ARCHITECTURE

The proposed system is designed as a modular and scalable architecture that enables efficient processing and analysis of customer care interactions using artificial intelligence techniques. The system operates as a web-based application, where the presentation layer provides an interactive user

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715448



interface for uploading and analyzing customer service calls. The processing layer handles speech recognition, natural language processing, and voice emotion detection, while the data layer manages storage and retrieval of call records and analysis results. External tools and libraries are integrated for audio processing, visualization, and report generation.

The architecture is divided into three main tiers: (1) Client Tier — a user interface that allows supervisors and quality assurance teams to upload audio recordings, view transcripts, and analyze results through dashboards; (2) Processing Tier — responsible for converting speech to text, performing sentiment analysis, and detecting emotional tone from audio signals using machine learning models; (3) Data Tier — a centralized database that stores audio files, transcripts, sentiment scores, emotion labels, and performance reports for efficient retrieval and analysis.

The system achieves efficient processing by combining textual and acoustic analysis in a unified workflow.

A. Role-Based Access Control

The system defines multiple user roles such as Admin, Supervisor, and Quality Analyst to ensure controlled access to system functionalities. Each user role is stored in the database and validated during authentication. Role-based access control is implemented at both the application and database levels, ensuring that users can only access data and features permitted to their role. This prevents unauthorized data access and maintains system security. Frontend validation restricts access to specific dashboards, while backend controls enforce data-level permissions.

TABLE I Role Capabilities

Role	Key Capabilities
Admin	System configuration, user management, database monitoring, analytics dashboard
Supervisor	View analysis reports, monitor agent performance, access call history
Quality Analyst	Evaluate calls, review sentiment and emotion results, generate reports
User	Upload audio files, view transcripts, access analysis results

B. Frontend-Architecture

The frontend of the system is developed as a web-based interface that allows users to upload audio files, view transcripts, and analyze results through interactive dashboards. Efficient rendering techniques are used to ensure smooth performance during data visualization. The interface is designed to be user-friendly, with clear navigation and structured layouts for displaying sentiment scores, emotion detection results, and performance insights. Visualization components are optimized to handle frequent updates without affecting system responsiveness.

C. Backend-Architecture

The backend is responsible for processing audio data and performing analysis using artificial intelligence models. It handles speech-to-text conversion, sentiment analysis, and voice emotion detection. The backend also manages authentication, data storage, and communication between different system modules. Secure access mechanisms ensure that only authorized users can perform specific operations. Database functions are used to process and return

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715448

structured results efficiently, reducing system load and improving performance.

D. Notification-Layer

The system includes a notification mechanism that alerts users about analysis results and important updates. Notifications can be triggered when processing is completed or when specific conditions are detected, such as negative sentiment or high emotional intensity in conversations. These alerts help supervisors quickly identify critical interactions that require attention. The notification system operates efficiently and ensures timely delivery of updates without requiring constant user monitoring.

IV. IMPLEMENTATION METHODOLOGY

The system is implemented as a browser-native web application using a modern technology stack designed for efficient processing and scalability. This approach eliminates the need for specialized hardware in customer care environments, reducing deployment cost and complexity. The system operates using standard computing devices and leverages cloud-based technologies for speech processing, sentiment analysis, and voice emotion detection.

Layer	Technology	Purpose
Frontend	HTML, CSS, JavaScript / React	Web interface, dashboards, user interaction
Backend	Python (FastAPI) / Node.js	Processing logic, API handling, system control
AI Models	Speech Recognition, NLP Models	Transcription, sentiment analysis, text processing
Emotion AI	SpeechBrain / Acoustic Models	Voice emotion detection from audio signals
Database	JSON / PostgreSQL	Storage of transcripts, results, and records
Visualization	Chart.js / Dashboard Components	Display of sentiment, emotion, and analytics
Audio Tools	FFmpeg / Audio Processing Libraries	Audio conversion and preprocessing
Notifications	Web Alerts / System Notifications	Alerts for analysis results and critical events

A. Audio-Input-Preprocessing

The system interface allows users to upload or record customer care audio through a browser-based application. each audio input is stored with associated metadata such as timestamp and user identification. the input is validated for supported formats and passed through preprocessing steps including noise handling and audio normalization. the processed audio is then prepared for further analysis by downstream modules.

B. Transcription-Processing-Pipeline

After preprocessing, the audio is passed to the speech recognition module, which converts spoken content into textual transcripts. the transcription process is performed using advanced models capable of handling variations in speech patterns and accents. the generated text is stored in the database and made available for further analysis. this pipeline ensures consistent and accurate conversion of audio data into structured textual form.

C. Textual-Sentiment-Evaluation

The transcribed text is analyzed using natural language processing techniques to evaluate sentiment

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715448

and communication quality. the system identifies sentiment polarity such as positive, negative, or neutral and extracts relevant patterns from the conversation. these results help in assessing the effectiveness of communication and identifying potential issues in customer interactions.

D. Acoustic-Emotion-Detection

In parallel with textual analysis, the system processes the audio signal to detect emotional characteristics. acoustic features such as tone, pitch, and intensity are analyzed to classify emotions including happiness, anger, sadness, and neutrality. the results are combined with textual insights to provide a comprehensive evaluation of customer-agent interactions.

V. DATABASE DESIGN

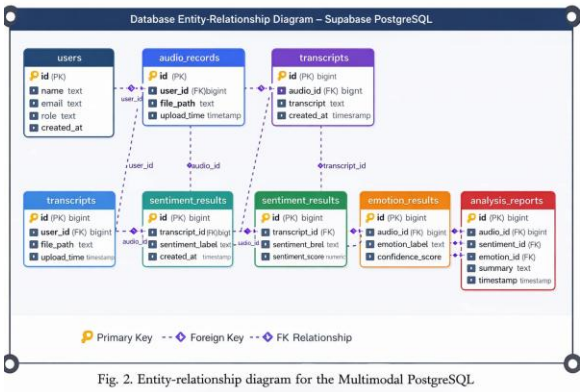


Fig. 2. Entity-relationship diagram for the Multimodal PostgreSQL

The database schema is implemented using a structured relational model designed to efficiently support the operational flow of the system. It comprises multiple interconnected tables whose relationships enforce data integrity and system logic at the database level, reducing dependence on application-layer validations. Each table stores specific information such as audio records, transcriptions, analysis results, and user details, ensuring organized data management. Foreign key constraints maintain consistency between related entities, allowing seamless data flow across different modules.

TABLE III Database Schema Summary

Table	Key Columns
users	id (PK), name, email, role, created_at
audio_records	id (PK), user_id (FK), file_path, upload_time
transcripts	id (PK), audio_id (FK), transcript_text, created_at
sentiment_results	id (PK), transcript_id (FK), sentiment_label, sentiment_score
emotion_results	id (PK), audio_id (FK), emotion_label, confidence_score
analysis_reports	id (PK), audio_id (FK), sentiment_id (FK), emotion_id (FK), summary, timestamp

Foreign key relationships enforce referential integrity throughout: buses reference both routes and users (driver); trips reference buses, routes, and the assigned driver; bookings reference passengers and trips; bus_positions reference trips.

VI. SIMULATION AND TRACKING ENGINE

Physical audio capture devices and real-time recording environments may not always be available in academic or testing scenarios, yet a customer care analysis system must remain fully functional without such dependencies. the system addresses this limitation by supporting both uploaded audio inputs and simulated processing workflows that replicate real-world interactions. this approach enables complete evaluation of the system's capabilities using standard computing environments without requiring live call center infrastructure.

A. Input-Mode

The system allows users to upload pre-recorded audio files through the web interface. each audio file is processed and stored with associated metadata such as timestamp and user details. this mode enables consistent testing and evaluation of the system by using sample datasets, allowing repeated analysis without dependency on live recordings.

B. Processing-Mode

In processing mode, the system performs speech-to-text transcription and emotion detection using

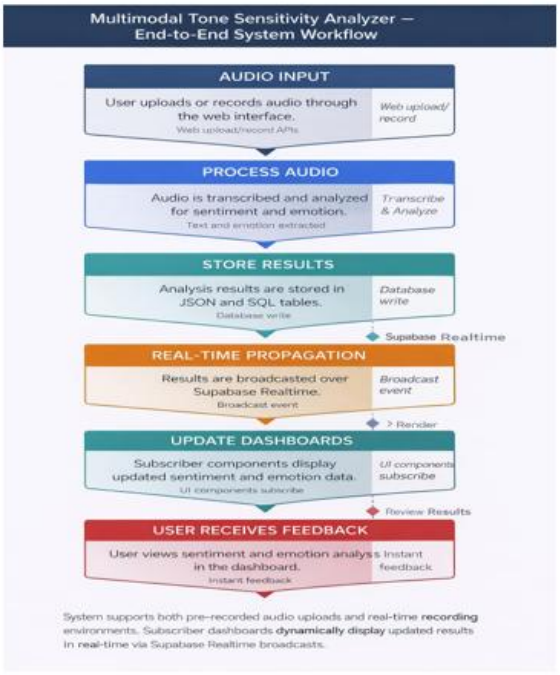
© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715448

predefined models. the transcription module converts audio into text, while the emotion detection module analyses acoustic features. both processes operate simultaneously and generate structured outputs that are stored in the database. this ensures consistent and efficient analysis of customer interactions.

C. Real-Time-Processing

The system enables near real-time analysis by processing uploaded audio immediately after submission. the results, including transcripts, sentiment scores, and detected emotions, are delivered to the user interface without delay. data is dynamically updated and displayed in dashboards, allowing users to view analysis results instantly. the system ensures proper handling of data flow and prevents duplication by managing processing states efficiently within the application.

VII. AUDIO PROCESSING AND VALIDATION



A. Audio-Input-Handling

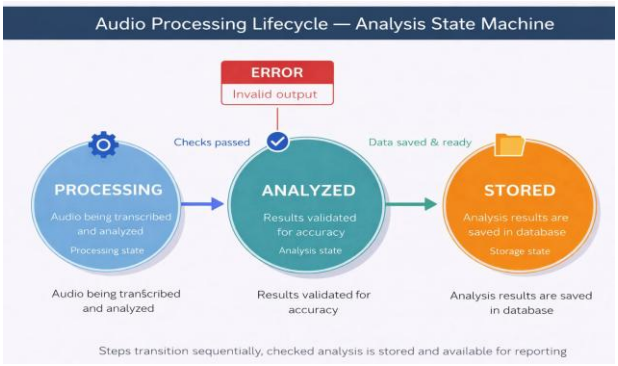
When a user uploads or records an audio file, the system assigns a unique identifier to the audio record and stores it along with relevant metadata such as user details and timestamp. The audio file is validated for format and quality before processing. This ensures that only compatible and usable audio inputs are passed into the analysis pipeline. The stored audio acts as the primary input for transcription and emotion detection modules.

B. Processing-Interface

The system processes the uploaded audio using speech recognition and emotion detection models. The transcription module converts speech into text, while the emotion analysis module evaluates acoustic features. The results are validated through system checks to ensure accuracy and completeness of the output. The processed data is then passed to the analysis module, where sentiment and performance evaluation are performed. All operations are handled within a structured workflow to maintain consistency and reliability.

C. Result-Lifecycle

The analysis results follow a defined lifecycle consisting of stages such as processing, analyzed, and stored. Initially, the audio is marked as processing while transcription and emotion detection are performed. Once analysis is completed, the results are marked as analyzed and stored in the database. These results can then be accessed for reporting and evaluation. The system ensures that each stage follows a proper sequence, preventing incomplete or inconsistent data from being used in performance analysis.



VIII. REAL-TIME ANALYSIS AND RESULT NOTIFICATIONS

A. Sentiment and Emotion Threshold Detection

On each processed audio event, the system evaluates sentiment scores and detected emotional tones using predefined thresholds. the textual transcript is

analyzed for sentiment polarity, while the audio signal is examined for emotional intensity such as anger, sadness, or neutrality. when the computed values fall beyond defined thresholds, the system identifies the interaction as critical. this evaluation is performed automatically within the analysis pipeline, adding minimal computational overhead and ensuring efficient detection of significant patterns in customer interactions.

the threshold values are designed to balance accuracy and reliability, ensuring that meaningful interactions are captured without generating excessive false alerts. these thresholds are configurable parameters and can be adjusted based on organizational requirements without modifying the system architecture or core processing logic.

B. Automated-Result-Handling

When threshold conditions are met, the system automatically updates the analysis status and stores the results in the database. the system records relevant details such as sentiment score, detected emotion, and processing timestamp. this process does not require manual intervention and ensures that all critical interactions are properly logged and available for further review.

to maintain data consistency, the system includes validation checks that prevent duplicate updates or incorrect state transitions. each analysis result follows a defined sequence, ensuring that processing, evaluation, and storage occur in a controlled and reliable manner.

C. Notification-Mechanism

The system generates notifications for significant analysis results to alert supervisors and quality assurance teams. alerts are triggered when specific conditions such as negative sentiment or high emotional intensity are detected. these notifications provide timely insights into customer interactions that may require immediate attention.

notifications are delivered through the system interface and can be extended to external communication channels if required. this ensures that users receive updates regardless of their activity status, enabling effective monitoring and faster response to critical customer service situations.

IX. ANALYTICS DASHBOARD

The analytics dashboard presents multiple categories of performance metrics related to customer care interactions. Interaction volume displays the total number of analyzed calls over a given time period, segmented by categories such as positive, negative, and neutral sentiment. This provides an overall view of customer interaction trends and helps identify patterns in service quality. Performance metrics aggregate evaluation scores across different agents and time intervals, enabling comparison and identification of areas requiring improvement. Emotion analysis reports highlight the distribution of detected emotions such as happiness, anger, sadness, and neutrality, offering deeper insights into customer-agent communication.

All metric categories are generated using backend processing functions that compute aggregated results directly from stored data. These functions return structured outputs that are efficiently delivered to the frontend interface for visualization. Performing aggregation at the database level reduces data transfer overhead and ensures consistency across multiple users accessing the dashboard simultaneously.

X. RESULTS AND DISCUSSION

The proposed Multimodal Tone Sensitivity Analyzer was evaluated through end-to-end functional testing of the system using the admin dashboard. The system successfully processed complete workflows including audio upload, transcription, sentiment analysis, emotion detection, and result visualization. The integrated modules worked efficiently to provide accurate and consistent analysis results without requiring complex infrastructure. The system demonstrated its capability to handle multiple audio inputs and generate real-time insights for monitoring customer care interactions. The following subsection presents the key output from the admin dashboard.

A. Admin Dashboard – Analysis and Monitoring
The Admin dashboard serves as the central interface for managing and monitoring customer interaction analysis. It allows users to upload or access recorded audio files and view detailed results including transcripts, sentiment scores, and detected emotional

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715448

tones. The dashboard presents summarized metrics such as sentiment distribution and emotion patterns across multiple interactions, helping in identifying trends in customer service quality. It also provides structured reports and visual insights that enable effective evaluation of agent performance and support decision-making for improving customer care services.

XI. COMPARATIVE ANALYSIS

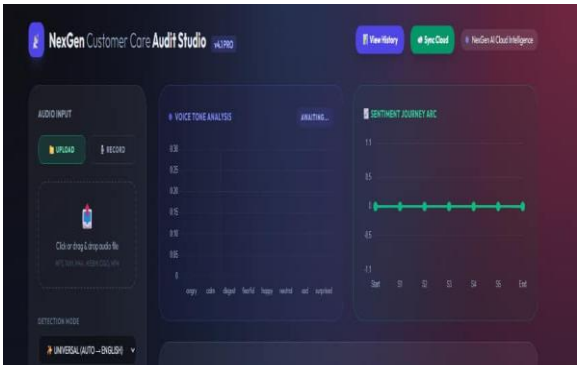
Table IV compares the proposed Multimodal Tone Sensitivity Analyzer with existing customer service evaluation approaches across key functional aspects addressed by the system. The most significant differentiators include the integration of speech recognition, sentiment analysis, and voice emotion detection within a single unified platform. Unlike traditional systems that rely only on manual auditing or text-based analysis, the proposed system combines both textual and acoustic features to provide a comprehensive evaluation of customer interactions. Additionally, the system supports automated processing and real-time analysis, eliminating the need for extensive manual effort. No existing approach effectively integrates multimodal analysis, automated evaluation, and scalable architecture within a single web-based application backed by a centralized data system.

TABLE IV Feature Comparison

Feature	Proposed System	Manual Auditing	Text-Based Analysis	Basic Call Logs
Automated call analysis	Yes	No	Yes	No
Speech-to-text transcription	Yes	No	Yes	No
Sentiment analysis	Yes	Partial	Yes	No
Voice emotion detection	Yes	No	No	No
Multimodal analysis	Yes	No	No	No
Real-time processing	Yes	No	Partial	No
Performance evaluation	Yes	Partial	Partial	No
reports				
Scalable architecture	Yes	No	Yes	No
Centralized data storage	Yes	No	Yes	Partial

XII. LIMITATIONS AND CHALLENGES

The system relies on speech recognition models to convert audio into text, which may be affected by background noise, accents, or unclear speech. In noisy environments or low-quality recordings, transcription accuracy may decrease, leading to



incorrect sentiment and emotion analysis results. Improving audio quality or using more advanced models can help reduce these limitations.

Voice emotion detection is based on acoustic features such as tone, pitch, and intensity, which may not always accurately represent the true emotional state of the speaker. Variations in speaking style or cultural differences can affect the interpretation of emotions. As a result, certain emotional classifications may not always be precise.

Although the system is designed for efficient processing, analyzing audio files using multiple AI models can require significant computational resources. Large volumes of audio data may increase processing time, especially when running on limited hardware or without optimized infrastructure.

While the system supports real-time analysis, handling a large number of simultaneous audio inputs may impact performance. Efficient scaling mechanisms and optimized data processing pipelines

are required to maintain responsiveness when the system is deployed at a larger scale.

XIII. CONCLUSION

This paper presented a Multimodal Tone Sensitivity Analyzer for Customer Care Services, a web-based system that integrates speech recognition, sentiment analysis, and voice emotion detection to evaluate customer-agent interactions. The system provides a comprehensive approach by combining textual and acoustic analysis, enabling accurate and objective assessment of communication quality. The centralized architecture allows efficient data processing, storage, and visualization through an interactive dashboard.

The proposed system eliminates the limitations of traditional manual auditing by automating the evaluation process and providing real-time insights into customer interactions. By analyzing both the content and emotional tone of speech, the system improves the reliability of performance monitoring and supports better decision-making for customer service management.

Future enhancements may include real-time live call monitoring, improved emotion detection models, multilingual support, and advanced analytics dashboards. These improvements can further enhance system performance, scalability, and usability, making it a powerful tool for modern customer care analysis and quality assurance.

XIV. ACKNOWLEDGMENT

The author expresses sincere gratitude to Dr. K. Ponmozhi, Associate Professor, Department of Computer Applications, SRM Valliammai Engineering College, for valuable guidance, continuous support, and encouragement throughout the development of this project. The author also acknowledges the contributions of open-source technologies and tools used in this work, including speech recognition and natural language processing frameworks, which played a significant role in the successful implementation of the system.

REFERENCES

[1] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in Proc. INTERSPEECH, 2009, pp. 312–315.

[2] D. Jurafsky and J. H. Martin, "Speech and Language Processing," 3rd ed., Pearson Education, 2021.

[3] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 1877–1901.

[4] Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[6] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," IEEE Intelligent Systems, vol. 28, no. 2, pp. 15–21, 2013.

[7] R. Cowie, E. Douglas-Cowie, and C. Cox, "Emotion Recognition in Human-Computer Interaction," IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32–80, 2001.

[8] S. Poria, E. Cambria, and A. Gelbukh, "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Multimodal Sentiment Analysis," in Proc. EMNLP, 2015, pp. 2539–2544.

[9] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Continuous Emotion Recognition in Speech Using Long Short-Term Memory Networks," in Proc. INTERSPEECH, 2010, pp. 597–600.

[10] Metallinou, S. Lee, and S. Narayanan, "Audio-Visual Emotion Recognition Using Gaussian Mixture Models," IEEE Trans. Multimedia, vol. 15, no. 7, pp. 1785–1797, 2013.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[12] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.