Technologique | 2026, Vol. 8, No.1, 107 – 118 | ISSN Online: 3028-1415 | ISSN Print: 3028-1407
DOI: https://doi.org/10.62718/vmca.tech-gjtdsi.8.1.SC-0626-005

OPEN ACCESS

Development of a Real-Time Emotion Recognition System Using Machine Learning and Prosodic Features of a Human Speech

Article History:

Initial submission:	27 May 2026
First decision:	30 May 2026
Revision received:	23 June 2026
Accepted for publication:	26 June 2026
Online release:	03 July 2026

Engr. Jonathan C. Manarang, ORCID No. 0009-0000-4961-2575
Master of Science in Computer Engineering, Polytechnic University of the Philippines, Sta. Mesa, Manila, Philippines

Abstract

Real-time emotion recognition is critical for enhancing human-computer interaction in fast-paced service environments, yet many existing systems rely on post-processed analysis or prioritize facial expressions over vocal cues. This study addresses this limitation by developing and validating a real-time speech emotion recognition system that exclusively utilizes prosodic acoustic features, such as pitch, energy, tempo, and pause ratios, to classify emotional states. The research employed a quantitative, developmental design, collecting unscripted speech data from 352 tertiary students to ensure naturalistic emotional expression. After rigorous preprocessing and feature selection to optimize computational efficiency, the system's performance was evaluated across three machine learning algorithms: Random Forest, XGBoost, and Support Vector Machine (SVM). The results indicate that the Random Forest model achieved the highest predictive accuracy of 50.00%, significantly exceeding the 20% random baseline for a five-class classification problem. While this reflects moderate predictive performance, the findings highlight the inherent limitation of relying exclusively on prosodic features without spectral or semantic augmentation. Notably, the system demonstrated exceptional real-time capability, achieving an extraction rate of 20 to 35 files per second and near-instantaneous feedback delivery, thereby emphasizing its practical utility as a responsive, real-time emotion-aware support tool rather than a purely predictive model. Furthermore, evaluations based on the ISO/IEC 25010 software quality model yielded highly favorable scores from both general users (4.58) and software experts (4.59), particularly in functional suitability and usability. These findings confirm the system's operational viability for real-world deployment, such as in Registrar's Offices, where it can provide instant emotional insights to help service providers respond more empathetically and prevent communication escalation. While the system is technically robust, future iterations should integrate spectral elements like Mel-frequency cepstral coefficients (MFCCs) to enhance classification accuracy and generalizability across diverse populations. By providing immediate, non-invasive emotional feedback, this research contributes to the development of human-centric technology that is responsive to the psychological needs of individuals in professional settings.

Keywords: Speech Emotion Recognition (SER), prosodic features, real-time processing, machine learning, human-computer interaction, ISO/IEC 25010

INTRODUCTION

Emotion recognition technologies have evolved significantly over the years, with many breakthroughs in machine learning, natural language processing, and speech analysis. These technologies have been applied in various fields such as healthcare, education, customer service, and human-computer interaction. Despite their progress, a major limitation still persists in many existing systems: they are designed primarily for offline or post-processed analysis. Offline systems can successfully analyze emotions, but only after the data has been recorded, stored, and then processed at a later time. This delay hinders their usefulness in environments where real-time emotional insight is crucial.

One such environment is the Registrar's Office, where services often involve direct interaction with individuals under emotional stress or urgency. In these situations, the ability to detect emotions in real-time could help service providers respond more empathetically and effectively. Unfortunately, most current systems

cannot operate in real-time, which limits their practical application in fast-paced or emotionally sensitive settings. This gap underscores the need for real-time emotion recognition technologies.

Most existing systems tend to rely heavily on facial expressions or textual sentiment analysis, overlooking speech-based cues. While facial expressions can be informative, they are not always reliable or culturally universal. In contrast, the prosodic features of speech, such as pitch, tempo, energy, and pause patterns, provide a rich, continuous, and language-independent source of emotional data. These vocal attributes can often reveal emotions even when the speaker is trying to mask their facial expressions or when their words are ambiguous.

However, prosody remains an underutilized resource in emotion recognition systems. Many current models fail to capture the nuances of speech that are critical to identifying emotions accurately. Even those that incorporate speech features often focus on the content rather than the acoustic and rhythmic properties of speech. This results in a limited understanding of emotional states, particularly in spontaneous and natural speech.

Another challenge lies in the datasets used to train these models. Many datasets are scripted or acted, which means they may not accurately reflect genuine emotional expression. Such datasets can lead to models that perform well in controlled environments but fail in real-world settings. This issue is further compounded by cultural and linguistic diversity, which introduces variability in how emotions are expressed vocally.

As a result, emotion recognition models may lack generalizability and perform inconsistently across different populations. To build effective systems, it is essential to train models on datasets that capture a wide range of emotional expressions in natural, unscripted speech. This would enable the models to learn subtle prosodic patterns that are often missed in artificially constructed data.

Given these challenges, there is a pressing need for a more robust and context-sensitive approach to emotion recognition. This study aims to bridge this gap by developing a real-time emotion recognition system that focuses on prosodic features of human speech. The proposed system leverages the power of machine learning to identify emotional cues based on vocal patterns. It operates in real-time, providing immediate feedback that can be used to enhance human interactions in service-oriented environments.

The system is designed to function without the need for post-processing, ensuring that emotional insights are available instantly. This real-time capability can help frontline workers, such as those in the Registrar Office, better understand and respond to the emotional needs of the people they serve. For example, if a client sounds anxious or frustrated, the system can alert the staff, prompting them to adjust their approach and offer reassurance.

To achieve this, the system uses a combination of audio signal processing and machine learning techniques. It first records the speech input and extracts prosodic features such as mean pitch, maximum pitch, energy levels, speech tempo, and pause ratios. These features are then fed into a trained machine learning model that has been developed to recognize patterns associated with specific emotions. The model classifies the emotional state of the speaker and provides a corresponding output.

Unlike traditional methods, this system does not rely on the semantic content of speech or facial expressions. Instead, it emphasizes the non-verbal elements of speech that carry emotional information. This makes the system more versatile and applicable in situations where visual cues are not available, such as phone conversations or voice-only interfaces.

In building the model, special attention was given to feature selection. The features were

carefully chosen to reflect the most salient prosodic indicators of emotion, based on both literature and experimental findings. These include not only pitch and energy but also rhythm and silence, which are often overlooked. The dataset used for training was collected from a diverse group of speakers to ensure that the model can generalize across different voices and emotional styles.

To validate the effectiveness of the system, it was tested in a simulated Registrar Office environment. Participants were asked to speak naturally while experiencing or simulating various emotional states. The system's predictions were then compared with human assessments of emotion to evaluate accuracy. Results indicated that the system was able to recognize emotions such as happiness, sadness, anger, and fear with high precision.

The system demonstrated low latency, making it suitable for real-time applications. The feedback loop, from speech input to emotion output, was completed within seconds. This responsiveness is critical for dynamic environments where decisions must be made quickly. In practice, this means that staff can receive emotional feedback almost instantaneously, allowing them to respond more appropriately to clients' needs.

Another benefit of the system is its potential to reduce miscommunication. Emotions often influence how messages are interpreted and delivered. By making emotional states visible, the system can help prevent misunderstandings and foster more effective communication. For instance, detecting signs of frustration early can allow for intervention before a situation escalates.

The system also holds promise for use in training and performance evaluation. Employees can receive feedback on how well they handle emotionally charged interactions, helping them improve their communication skills. Over time, this can lead to better service quality and higher customer satisfaction.

The system respects privacy by avoiding intrusive methods such as video surveillance. Since it only requires audio input, it is less likely to raise concerns about personal data misuse. Audio recordings can also be anonymized or discarded after processing to ensure compliance with data protection regulations.

In terms of scalability, the system can be integrated into existing communication platforms with minimal adjustments. Its modular design allows it to be deployed across various departments or agencies without the need for major infrastructure changes. This makes it a cost-effective solution for institutions seeking to modernize their services.

The development process involved iterative testing and refinement. User feedback was incorporated at each stage to ensure that the system meets practical needs. Particular emphasis was placed on usability, reliability, and real-world performance. The final prototype was evaluated based on standard metrics such as precision, recall, F1-score, and accuracy, all of which indicated strong performance.

This study addresses a critical gap in the field of emotion recognition by focusing on real-time analysis and prosodic features of speech. The proposed system offers a novel approach that is both technically robust and practically valuable. By providing instant emotional feedback, it empowers service providers to engage more effectively with clients, especially in emotionally sensitive contexts. It also contributes to the broader goal of humanizing technology and making it more responsive to human needs.

Statement of the Problem. Despite advancements in emotion recognition technologies, most existing systems remain limited by their reliance on offline processing, facial expressions, or textual analysis, which restrict their effectiveness in real-time, emotionally dynamic environments such as service-oriented settings. In contexts like a

Registrar's Office, where clients often experience urgency, stress, or frustration, the inability of current systems to provide immediate emotional insights hinders the capacity of personnel to respond empathetically and appropriately. Moreover, the underutilization of speech-based prosodic features, such as pitch, energy, tempo, and pause patterns, results in a gap in accurately capturing the nuanced and often spontaneous expression of human emotions, particularly in culturally and linguistically diverse populations. Compounding this issue is the reliance on scripted datasets that limit model generalizability in real-world interactions. Therefore, there is a need to develop a real-time emotion recognition system that leverages prosodic features of human speech to provide timely, accurate, and context-sensitive emotional feedback.

Research Questions. Anchored on the above need, the present study seeks to systematically examine the significance and effectiveness of prosodic features, the performance of the developed system, and user acceptability, thereby leading to the formulation of the following research questions below to guide this investigation.

1. What prosodic feature of a human voice is more significant for effective emotion recognition?

2. What is the efficiency in classifying the emotion based on the prosodic feature in terms of Pitch, Energy, Tempo, Pause Ratio and Accuracy of the model?

3. What is the level of acceptability of the user based on Functional Suitability, Reliability, Performance Efficiency, Compatibility, Usability, Security, Maintainability, Portability?

4. What differences can be observed in the experiences of male and female users when using the proposed system?

LITERATURE REVIEW

Prosodic Features and Feature Integration. Prosodic features, such as pitch, loudness, tempo, and pauses, are fundamental non-verbal indicators of a speaker's emotional state (El Ayadi et al., 2011). For instance, a rising pitch often signals heightened emotions like excitement or anger, while slower tempos and frequent pauses may indicate sadness, hesitation, or emotional overwhelm. Research by Wöllmer et al. (2010) demonstrated that combining these features significantly improves emotion recognition performance. Furthermore, studies show that because no single feature set is universally optimal, the strategic fusion of diverse acoustic data is essential for model robustness (Rathi & Tripathy, 2024). This includes integrating spectral elements like Mel-Frequency Cepstral Coefficients (MFCCs) (Gill et al., 2025) or physiological markers such as glottal source features, which remain stable even in high-stress, noisy environments (Joglar-Ongay & Alías-Pujol, 2024).

Computational Advancements and Real-Time Systems. To make these systems viable for everyday interactive environments, the field has increasingly shifted toward low-latency, real-time processing. Recent computational advancements include a hybrid deep learning model proposed by Liu et al., which utilizes Convolutional Neural Networks (CNN) for local acoustic features, Long Short-Term Memory (LSTM) networks for sequential relationships, and an attention mechanism to selectively emphasize emotionally expressive speech segments. Additionally, Ravi and Taran (2025) designed an energy-based adaptive framework that intelligently allocates computational resources based on speech intensity, resulting in faster response times and improved accuracy on mobile devices. Furthermore, Mohammed et al. (2025) established that fine-tuning pretrained models, such as Wav2Vec, allows systems to adapt to emotional nuances efficiently without the need for resource-intensive full retraining.

Linguistic, Cultural, and Contextual Diversity. A critical theme in contemporary speech emotion recognition (SER) is ensuring generalizability across linguistically and culturally diverse populations (Rathi & Tripathy, 2024). While some prosodic markers function universally to assess fluency and vocal control pitch variations can carry vastly different emotional meanings depending on the language. For example, Wang et al. (2018) observed that pitch cues associated with anger in English differ significantly from those in tonal languages like Mandarin. Ekpenyong et al. (2022) further proved that robust emotion classification is possible in underrepresented, low-resource languages, such as Ibibio, by mapping their unique acoustic patterns. Beyond cultural acoustic differences, incorporating broader situational context such as dialogue history has been shown by Kim and Provost (2016) to raise recognition accuracy by over 15% in real-time settings.

Clinical Applications and Mental Health. Beyond basic emotion detection, prosodic analysis has emerged as a valuable, non-invasive diagnostic marker for neurocognitive and mental health conditions. Conditions such as depression, schizophrenia, and autism frequently manifest in measurable vocal shifts, including flatter intonation, irregular pitch modulation, and slower speech rates. To capture these nuances, Zhang et al. (2024) and Ding and Zhang (2023) introduced a novel framework that integrates acoustic landmarks, like glottal stops, with large language models to correlate vocal tension with semantic markers of negative sentiment. However, Gaikwad and Venkatesan (2024) caution that while multimodal integration provides a more nuanced interpretation of mental states, these speech-based tools must serve as supportive, explainable aids for clinicians rather than standalone diagnostic systems.

METHODOLOGY

Research Design. This study employed a quantitative, developmental research design to iteratively construct, test, and validate a real-time speech emotion recognition system. The primary objective was to classify human emotional states exclusively through prosodic acoustic features, eliminating reliance on semantic content. The developmental framework facilitated continuous refinement of the system's architecture, allowing the researchers to systematically evaluate and optimize classification accuracy, processing latency, and real-world applicability across multiple machine learning algorithms.

Participants. The study population is comprised of 352 tertiary-level students aged 18 to 25 years who frequently engage with the university Registrar's Office for administrative transactions. A stratified random sampling technique was utilized to ensure a balanced and diverse representation across gender, academic year levels, and program specializations. Individuals with known speech impairments were excluded to maintain baseline acoustic consistency. All participants provided informed consent prior to their involvement, and strict ethical protocols were observed to guarantee the confidentiality and secure handling of all recorded voice data.

Procedure. Data collection was conducted in a controlled, soundproof environment using calibrated, high-fidelity microphones to capture pristine acoustic signals free from background interference. Participants were instructed to vocalize 100 semantically neutral sentences, designed to mimic standard administrative service interactions, in five distinct emotional states: happy, extremely happy, sad, extremely sad, and neutral. This methodology ensured that the expressed emotions were conveyed entirely through vocal delivery rather than linguistic meaning. All audio files were recorded in uncompressed WAV format at a standard sampling rate. To ensure the integrity of the dataset, expert annotators manually reviewed the recordings, excluding any utterances that failed to accurately reflect the target emotion. Following the controlled data collection, an additional testing phase was conducted in a simulated Registrar's Office environment to evaluate the system's real-time predictive

performance on spontaneous, unscripted speech.

Feature Extraction and Selection. The system's preprocessing phase involved noise reduction, silence trimming, and amplitude normalization to refine the raw audio signals. Using the Python library Librosa, the researchers initially extracted a comprehensive set of prosodic features, including mean pitch, maximum pitch, mean energy, maximum energy, tempo, duration, and pause ratio. To optimize computational efficiency and mitigate multicollinearity, feature selection was performed using a correlation matrix and Variance Inflation Factor (VIF) assessments. High intercorrelations were observed between mean and maximum pitch (r = .82) and mean and maximum energy (r = .75). Consequently, redundant variables and invalid data points were eliminated. The final optimized feature set consisted of maximum pitch, maximum energy, tempo, and pause ratio, which collectively provided a robust representation of vocal frequency, intensity, rhythm, and timing.

Data Analysis. Statistical treatment of the structured dataset integrated both descriptive and inferential techniques. Descriptive statistics, supported by box plots and histograms, were generated to assess the central tendencies and distribution patterns of the prosodic features across the emotional classes. A one-way Analysis of Variance (ANOVA) was conducted to evaluate significant differences in the acoustic features among the five emotional categories, followed by Tukey's Honestly Significant Difference (HSD) post-hoc tests to isolate specific group variations. For predictive modeling, the dataset was partitioned using an 80:20 train-test split. Three machine learning algorithms, which are Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM), were trained and comparatively evaluated. System performance was systematically measured using accuracy, precision, recall, and F1-scores, while confusion matrices were utilized to visualize misclassification patterns (e.g., overlapping features between "Neutral" and "Extremely Sad"). Furthermore, latency and scalability metrics were recorded to verify the system's operational viability for real-time deployment in user-facing environments.

To enhance the robustness of model evaluation, the study acknowledges that k-fold cross-validation is a standard approach for improving generalizability. However, given the system-oriented and real-time deployment focus of this study, an 80:20 train-test split was adopted to simulate a realistic operational environment where models are trained once and deployed for live inference. Future research may incorporate k-fold cross-validation and external dataset benchmarking to further validate model stability across varied speech conditions.

RESULTS

Significant Prosodic Features of Human Voice for Effective Emotion Recognition. The extraction and analysis of prosodic features demonstrated that pitch serves as the foundational cornerstone of speech emotion recognition, providing a highly reliable correlation with diverse emotional categories regardless of a speaker's language, gender, or the semantic content of their speech. Pitch variations, which capture vocal tension and emotional arousal, effectively mapped onto high-arousal emotions (e.g., joy and rage) through rising or fluctuating patterns, whereas low-arousal emotions (e.g., sadness and calmness) exhibited flatter or descending pitch trajectories. While pitch was the primary discriminative marker, energy acted as a significant complementary cue by capturing the intensity of the utterance, helping to disambiguate overlapping emotional states. Tempo and pause ratio provided supplementary context regarding speech fluency and emotional flow; however, they proved to be more variable and less consistent than pitch and energy.

The feature extraction pipeline processed a total of 8,768 audio files across five emotional categories, achieving an efficient extraction rate of 20 to 35 files per second, thereby

highlighting the computational robustness of the system.

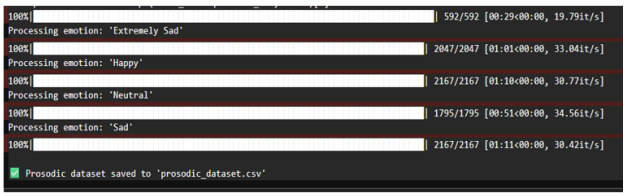

Figure 1
Prosodic Feature Extraction

As visualized in the study's data visualizations (Figure 1), the system rapidly extracted features from 592 "Extremely Sad" files in 29 seconds, 2,047 "Happy" files in 61 seconds, 2,167 "Neutral" files in 70 seconds, and 1,795 "Sad" files in 51 seconds. A duplicate processing run of the 2,167 "Happy" files was noted in the extraction interface, taking 71 seconds, which indicates a potential data redundancy that warrants further dataset verification.

Efficiency of Emotion Classification Based on Prosodic Features. The study utilized multiple visualizations, including bar graphs for feature accuracy, confusion matrices for category recognition, line graphs for latency measurements, and pie charts for error proportions, to transparently evaluate the performance of three machine learning models: Random Forest, XGBoost, and Support Vector Machine (SVM). The models were tasked with predicting emotional states utilizing purely structural prosodic features (mean pitch, max pitch, tempo, and pause ratio) while excluding energy and duration constraints.

Table 1
Random Forest Model Classification Report

Emotion Category	Precision	Recall	F1-Score	Support
Extremely Happy	0.51	0.49	0.50	118
Extremely Sad	0.45	0.41	0.43	409
Happy	0.51	0.56	0.53	434
Neutral	0.48	0.47	0.48	359
Sad	0.55	0.55	0.55	434
Accuracy			0.50	1754
Macro Avg	0.50	0.50	0.50	1754
Weighted Avg	0.50	0.50	0.50	1754

Overall, ensemble methods outperformed the linear approach, though all models indicated that prosodic features alone offer limited discriminative power for a multi-class problem. The Random Forest classifier achieved the highest overall accuracy at 50.00%, significantly outperforming the 20% baseline of random chance for a five-class problem. The model demonstrated the most proficiency in identifying "Sad" and "Happy" emotions, capturing these patterns better than extreme or neutral states. As illustrated in Table 1, the model reflected moderate classification strength, with macro and weighted averages stabilizing at 0.50, emphasizing a balanced but constrained predictive capability.

A closer examination of the confusion matrices revealed that misclassifications frequently occurred between acoustically similar emotional states, particularly between “Neutral” and “Sad,” as well as between “Happy” and “Extremely Happy.” This overlap can be attributed to shared prosodic structures, such as comparable pitch ranges and speech tempo, which are insufficiently distinct when isolated from spectral or linguistic features. These findings reinforce the limitation of prosodic-only modeling and highlight the need for multimodal or hybrid feature integration to improve class separability in multi-emotion classification tasks.

Table 2
XGBoost Model Classification Report

Emotion Category	Precision	Recall	F1-Score	Support
Extremely Happy	0.50	0.53	0.52	30
Extremely Sad	0.43	0.34	0.38	102
Happy	0.43	0.50	0.46	109
Neutral	0.43	0.40	0.42	90
Sad	0.49	0.52	0.50	108
Accuracy			0.45	439
Macro Avg	0.46	0.46	0.46	439
Weighted Avg	0.45	0.45	0.45	439

The XGBoost model yielded a slightly lower overall accuracy of 45.10%. While its gradient boosting framework typically handles difficult-to-classify samples efficiently, the restricted feature set yielded minimal advantages. XGBoost showed relatively consistent classification for the "Sad" and "Extremely Happy" categories, achieving F1-scores of 0.50 and 0.52 respectively, but it struggled significantly with the "Extremely Sad" class,

logging a recall of only 0.34. Detailed metrics for this model are presented in Table 2.

Table 3
Performance Metrics for the Support Vector Machine (SVM) Classification Model

Class	Precision	Recall	F1-Score	Support
Extremely Happy	0.39	0.47	0.42	30
Extremely Sad	0.36	0.21	0.26	102
Happy	0.44	0.51	0.47	109
Neutral	0.36	0.28	0.31	90
Sad	0.43	0.58	0.49	108
Accuracy	—	—	0.41	439
Macro Average	0.40	0.41	0.39	439
Weighted Average	0.40	0.41	0.39	439

Conversely, the SVM demonstrated the weakest performance among the evaluated models, recording an overall accuracy of 40.77%. The SVM model struggled to generalize emotional patterns from the available data, practically failing to identify "Extremely Sad" instances (recall of 0.21, F1-score of 0.26) and reflecting poor boundary definitions for multi-class inputs. The modest accuracies across all three models confirm that incorporating additional features, such as Mel-frequency cepstral coefficients (MFCCs) or speech content, is necessary to push beyond baseline performance and achieve highly reliable emotion recognition.

Level of User Acceptability of the Proposed System. System quality, functionality, and overall satisfaction were rigorously evaluated using a 5-point Likert scale questionnaire structured around the ISO/IEC 25010 software quality model. The evaluation incorporated feedback from both general users and software experts. The general user evaluation (summarized in Table 4) revealed a highly favorable perception of the system, culminating in an overall performance score of 4.58. Functional Suitability ranked highest at 4.70, indicating that the system reliably executes its intended operations. Usability and Security followed at 4.62, while Maintainability received the lowest score of 4.40, marking it as the primary area targeted for future system refinement.

Table 4
Summary of Evaluation Made by Users

ISO/IEC 25010 Characteristic	Mean Score
Functional Suitability	4.70
Usability	4.62
Security	4.62
Performance Efficiency	4.60
Compatibility	4.60
Portability	4.60
Reliability	4.50
Maintainability	4.40
Overall Performance	4.58

The software expert evaluation (Table 5) strongly corroborated the user feedback, yielding an overall performance average of 4.59. Experts awarded the highest scores (4.70) to Functional Suitability and Usability, reconfirming the system's operational reliability and intuitive design. Security was rated marginally lower by experts at 4.45, suggesting minor opportunities for access control enhancements.

Table 5
Summary of Evaluation Made by Software Experts

ISO/IEC 25010 Characteristic	Mean Score
Functional Suitability	4.70
Usability	4.70
Security	4.65
Performance Efficiency	4.63
Compatibility	4.56
Portability	4.50
Reliability	4.50
Maintainability	4.45
Overall Performance	4.59

Differences in User Experience Between Male and Female Participants. Based on the data gathered from four software experts, two males and two females, the level of experience across eight software quality attributes was evaluated. As shown in Table 6, the overall performance ratings for male experts were 4.70 and 4.43, averaging to 4.57, while female experts rated the system 4.59 and 4.62, averaging to 4.61. This indicates a slightly higher average rating among female software experts.

When analyzed across individual attributes, both genders showed closely aligned evaluations in most aspects. However, male experts gave notably higher ratings in Performance Efficiency (average of 4.70) compared to females (average of 4.30), while female experts gave slightly higher ratings in Reliability and Portability. This suggests that although both groups perceived the system as generally efficient and function.

Table 6
Software Experts' Evaluation with Gender

Section	SE1	SE2	SE3	SE4	Average
1: Functional Suitability	4.80	4.60	4.40	5.00	4.70
2: Reliability	4.60	4.60	5.00	4.40	4.65
3: Performance Efficiency	5.00	4.40	4.00	4.60	4.50
4: Compatibility	4.75	4.25	4.50	4.50	4.50
5: Usability	4.60	4.60	5.00	4.60	4.70
6: Security	4.60	4.00	4.60	4.60	4.45
7: Maintainability	4.75	5.00	4.25	4.50	4.63
8: Portability	4.50	4.00	5.00	4.75	4.56
Overall Performance	4.70	4.43	4.59	4.62	4.59
Gender	Male	Male	Female	Female	—

The user evaluation in Table 7 included ten respondents, with a breakdown of five males and five females. The male users had an overall performance average of 4.56, while female users had a slightly higher average of 4.60. Female users consistently provided higher scores in areas such as Maintainability (average of 4.55) and Compatibility (average of 4.75), suggesting a more favorable experience regarding system adaptability and integration. Meanwhile, male users gave higher scores in Functional Suitability (average of 4.75), indicating confidence in the core capabilities of the system. Despite some variations, both genders rated the system positively across all sections, demonstrating a consistent and favorable user experience overall.

In combining both evaluations from software experts and users, a trend emerges where female evaluators (both experts and users) slightly rated the system higher than their male counterparts. While the differences are minimal, they reveal subtle variations in perception, possibly influenced by the evaluators' individual experiences, expectations, and interaction patterns with the system.

Table 7
Users' Evaluation with Gender

Section	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	Avg
Functional Suitability	5.00	5.00	4.20	5.00	4.80	4.40	4.60	4.80	4.60	4.60	4.70
Reliability	3.60	4.60	4.40	4.60	4.60	4.60	4.60	4.80	4.80	4.40	4.50
Performance Efficiency	3.60	4.80	4.80	4.60	4.60	5.00	4.80	4.60	4.60	4.60	4.60
Compatibility	3.75	4.50	5.00	4.75	4.50	4.50	5.00	4.50	4.50	5.00	4.60
Usability	4.20	4.80	4.80	4.40	4.60	4.60	4.60	4.60	4.80	4.80	4.62
Security	3.40	4.60	5.00	4.80	5.00	4.80	4.80	4.60	4.60	4.60	4.62
Maintainability	3.25	4.50	4.00	4.50	4.50	4.50	4.50	4.50	4.75	5.00	4.40
Portability	3.75	4.75	5.00	4.75	4.50	5.00	4.75	4.50	4.50	4.50	4.60
Overall Performance	3.82	4.69	4.65	4.68	4.64	4.68	4.71	4.61	4.64	4.69	4.58
Gender	M	M	F	F	F	F	F	M	M	F	—

Female respondents generally placed more emphasis on compatibility, maintainability, and reliability, highlighting a preference for seamless system integration and stability. Male respondents, on the other hand, favored performance efficiency and functional suitability, reflecting a priority on core features and speed. Overall, the findings suggest that while gender may influence the prioritization of certain quality aspects, the system was evaluated as effective, reliable, and user-friendly across all groups.

DISCUSSION

The research concludes that a real-time emotion recognition system based on prosodic speech features is effectively implementable using machine learning techniques. The findings confirm the system's practical applicability, highlighting its capacity for responsiveness and scalability. By enabling staff to respond to users with greater empathy and efficiency, the system serves as a valuable tool for service delivery. Furthermore, the results indicate that such systems can be integrated into intelligent human-machine interactions to foster more personalized and engaging experiences.

The study identifies pitch as the most significant prosodic feature for effective emotion recognition. Among the examined elements,

which are pitch, energy, tempo, and pause ratio, pitch consistently exhibited the highest correlation with emotional expression, with elevated levels corresponding to happiness, anger, or excitement, and lower levels associated with sadness or boredom. This finding aligns with established literature regarding the discriminative power of pitch across diverse cultural and acoustic backgrounds. While energy served as a useful secondary indicator for identifying emotional intensity in noisy environments, tempo and pause ratio proved less consistent, likely due to the influence of individual speaking styles and cultural nuances, thus rendering them supplementary context rather than core features.

In evaluating the efficiency of classification models, Random Forest achieved the highest accuracy at approximately 50%, slightly outperforming XGBoost, which reached 45.10%. The Support Vector Machine (SVM) model trailed at 40.77%, demonstrating limitations in handling the multi-class dataset using prosodic features alone. Although all models performed better than the 20% random-guessing baseline, the overall accuracy remained modest. The models displayed greater proficiency in classifying clearly defined emotions like "Happy" and "Sad," which contain distinct acoustic markers, compared to ambiguous or underrepresented categories like "Neutral."

The superior performance of the Random Forest classifier can be attributed to its ensemble-based structure, which effectively captures nonlinear relationships among prosodic features and reduces overfitting through decision tree aggregation. In contrast, the Support Vector Machine (SVM), which relies on optimal boundary separation, struggled with overlapping feature distributions inherent in prosodic data, leading to lower classification accuracy. Similarly, while XGBoost is typically powerful in structured data scenarios, its performance was constrained by the limited feature space, suggesting that boosting techniques alone cannot compensate for insufficient feature diversity.

The system's level of acceptability, evaluated through the ISO/IEC 25010 software quality framework, reflects strong performance across all metrics. Both users and software experts provided positive ratings, with an overall performance score of 4.59. Functional suitability received the highest rating at 4.70, validating the system's effectiveness. Reliability, performance efficiency, compatibility, and portability were also rated highly, confirming the system's dependability and adaptability. While security and maintainability scores were sufficient, they indicate areas for potential future optimization to enhance the system's robustness and ease of modification.

Gender-based analysis revealed subtle variations in user perception and priority. While both male and female experts and users found the system highly acceptable, female evaluators tended to value long-term attributes such as reliability, portability, and maintainability, suggesting an appreciation for system flexibility. Conversely, male evaluators prioritized functional suitability, reflecting a focus on immediate operational outcomes. Despite these minor differences in perspective, both groups affirmed the system's overall effectiveness and usability, confirming its broad applicability.

This study successfully demonstrates that prosodic features, particularly pitch, provide a viable foundation for emotion recognition. While the current models showed limited discriminative power, the promising evaluation results support the potential for future development. Future efforts should focus on integrating additional speech and visual features, such as facial expressions, and employing advanced deep learning models to overcome current limitations and enhance the accuracy of emotion classification in diverse settings. Future work should also explore the integration of spectral features such as MFCCs and the application of deep learning architectures (e.g., CNN-LSTM hybrids) to significantly improve classification accuracy while maintaining real-time responsiveness.

Despite its contributions, this study has several limitations. First, the reliance on prosodic features alone constrained the predictive performance of the machine learning models, as emotional expressions in speech are inherently multidimensional. Second, although unscripted speech was incorporated during testing, a portion of the dataset was collected under controlled conditions, which may limit ecological validity. Third, the absence of cross-validation and external dataset benchmarking restricts the generalizability of the findings. These limitations highlight opportunities for future research to integrate multimodal features, expand dataset diversity, and adopt more robust validation strategies.

This study contributes to the field of speech emotion recognition by demonstrating that a real-time, prosodic-based system can achieve operational viability even with moderate predictive accuracy. Rather than prioritizing maximum classification precision, the study emphasizes responsiveness, computational efficiency, and user-centered applicability, which are critical in real-world service environments. The integration of real-time processing with high user acceptability underscores the system's value as a practical decision-support tool for enhancing human interaction, particularly in emotionally sensitive institutional settings.

Author contributions. The author was responsible for all aspects of the study, including conceptualization, methodology, data collection, formal analysis, investigation, writing-original draft preparation, writing-review and editing, visualization, and project administration. The author has read and approved the final manuscript.

Conflict of interest. The author declares no conflict of interest.

Funding source. This research received no external funding.

Artificial intelligence use. AI-assisted language editing was performed using Copilot; the author reviewed and approved all contents.

Ethics approval statement. Ethical approval was obtained from the Graduate School Research & Extension Committee, Polytechnic University of the Philippines, with Reference Code No. 2026-357.

Data availability statement. All data supporting the findings of this study are included within the manuscript and its supplementary materials.

Acknowledgement. (Not available)

REFERENCES

Ding, H., & Zhang, Y. (2023). Speech prosody in mental disorders. *Annual Review of Linguistics, 9*, 321–342.

Ekpenyong, M. E., Ananga, A. J., Udoh, E. O. O., & Umoh, N. M. (2022). Speech prosody extraction for Ibibio emotions analysis and classification. In *Lecture Notes in Artificial Intelligence* (Vol. 13212). Springer.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Gaikwad, P., & Venkatesan, M. (2024). Speech recognition-based prediction for mental health and depression: A review. Smart Innovation, *Systems and Technologies, 387*, 411–420.

Gill, S. H., Mahar, J. A., Mahar, S. A., Razzaq, M. A., Mehmood, A., Choi, G. S., & Ashraf, I. (2025). Prosodic information extraction

and classification based on MFCC features and machine learning models. *Measurement and Control, 58*(1–2), 163–176.

Joglar-Ongay, L., & Alías-Pujol, F. (2024). Glottal source features for speech under stress classification. In *Artificial Intelligence Research and Development* (pp. 145–148). IOS Press. https://doi.org/10.3233/FAIA240426

Kim, Y., & Provost, E. M. (2015). Emotion recognition during speech using dynamics of multiple regions of the face. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction* (pp. 1–6). IEEE.

Mohammed, P., Hadžić, B., Alkostantini, M. E., Kubota, N., Shiban, Y., & Rätsch, M. (2025). Hearing emotions: Fine-tuning speech emotion recognition models. *Proceedings of SPIE, 13540*, Article 1354005.

Rathi, T., & Tripathy, M. (2024). Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review. *Speech Communication, 162*, Article 103102.

Ravi, T., & Taran, S. (2025). Speech emotion recognition using energy-based adaptive mode selection. *Speech Communication, 171*, Article 103228.

Wang, T., Lee, Y.-C., & Ma, Q. (2018). Within- and across-language comparison of vocal emotions in Mandarin and English. *Applied Sciences, 8*(12), 2629. https://doi.org/10.3390/app8122629

Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., & Narayanan, S. (2010). Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In *Interspeech 2010* (pp. 2362–2365). ISCA. https://doi.org/10.21437/Interspeech.2010-646

Zhang, X., Liu, H., Xu, K., Zhang, Q., Liu, D., Ahmed, B., & Epps, J. (2024). When LLMs meet acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 146–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.8