Emotion Recognition in Smart Assistants: An Empirical Investigation

Mukiibi Moses, Shema Nkindi Giscard, Department of Smart Computing, Kyungdong University. Global Campus (Gosung) 46 4-gil, Bongpo, Gosung, Gangwondo, 24764. Republic of Korea

Abstract

This study presents a systematic investigation of emotion recognition technologies for smart assistant applications through a multi-phase research design. We first conducted a review of the literature to establish baseline performance metrics across text, speech, and multimodal approaches. Subsequently, we developed and evaluated a multimodal fusion framework incorporating confidence-aware weighting, which we tested on a curated dataset of 3,247 genuine user interactions collected under controlled conditions. Our empirical results demonstrate that the proposed framework achieves statistically significant improvements over conventional methods, with mean accuracy of 78.6% compared to 68.4% for single-modality baselines. We further conducted a validation study with 127 participants in laboratory settings, revealing critical insights about the relationship between different modalities and emotion categories. The research identifies specific failure cases and proposes theoretically grounded solutions based on affective computing principles. These findings contribute both methodological advances and practical guidelines for developing emotionally intelligent human-computer interaction systems.

Keywords: Emotion Recognition; Smart Assistants; Multimodal Fusion; Human-Computer Interaction; Affective Computing; Speech Emotion Recognition

1. Introduction

1.1 Background and Motivation

The proliferation of smart assistants represents one of the most significant transformations in human-computer interaction over the past decade. Intelligent Virtual Assistants (IVAs) are increasingly integrated into daily life, offering convenience, personalization, and automation through machine learning . These systems enable hands-free interaction and smart environments, enhancing user experience. Emotions such as joy or frustration emerge from perceived usability, shaping attitudes and reinforcing IVAs' strategic role in consumer-brand interactions .

The capacity for emotional intelligence in artificial systems has been a central goal of affective computing since its inception . For smart assistants specifically, the ability to accurately perceive user emotional states enables several critical functions: appropriate response selection, conversational flow management, task prioritization based on user urgency, and long-term relationship building. Research on IVAs intersects Technology Adoption and Affective

Computing, with traditional models (TAM, UTAUT) emphasizing perceived usefulness and ease of use while often omitting affective dynamics .

1.2 Research Problem

Current approaches to emotion recognition in smart assistants face three fundamental challenges that this research directly addresses. First, the predominant evaluation paradigm relies on acted emotional expressions from standardized databases, which fail to capture the subtlety and authenticity of naturally occurring emotions in human-computer interaction. Second, existing multimodal fusion techniques typically employ static weighting schemes that do not adapt to varying signal quality across modalities in real-world environments . Third, the substantial variation in emotional expression across individuals and contexts remains inadequately addressed by current personalization approaches.

1.3 Research Questions and Contributions

This study addresses the following research questions:

RQ1: What is the comparative performance of different modalities (speech, text, facial expression) for emotion recognition in smart assistant interactions?

RQ2: Can a confidence-aware multimodal fusion framework that estimates modality reliability achieve superior accuracy compared to single-modality approaches?

RQ3: How does emotion recognition performance vary across different emotion categories and interaction contexts?

The primary contributions of this work are: (1) a systematic evaluation establishing empirically grounded performance benchmarks for emotion recognition modalities; (2) a confidence-aware multimodal fusion algorithm; (3) empirical validation through controlled laboratory experiments; and (4) an openly available dataset of 3,247 authentic user interactions with corresponding emotion annotations.

2. Literature Review

2.1 Theoretical Foundations of Emotion Recognition

The computational modeling of human emotion requires grounding in psychological theories of emotional experience. Research has established that emotions can be understood through multiple representational frameworks. Discrete emotion theory, tracing to Ekman's foundational work, posits six basic emotions (happiness, sadness, anger, fear, surprise, disgust) with universal facial expressions . Dimensional models, notably Russell's circumplex model, represent emotions along valence and arousal dimensions, offering advantages for capturing subtle affective states .

Affective Computing shows that emotions such as joy, surprise, and frustration shape system evaluations . Machine learning driven personalization enhances user engagement yet raises

important considerations about trust and anthropomorphism . Each framework presents distinct computational affordances and limitations that inform our methodological choices.

2.2 Single-Modality Approaches

2.2.1 Text-Based Sentiment Analysis

Computational analysis of textual sentiment has progressed through multiple methodological paradigms. Early lexicon-based approaches rely on word-emotion association dictionaries. Supervised machine learning methods, including support vector machines and maximum entropy classifiers, improved generalizability but required substantial annotated corpora.

A comprehensive survey by Birjali, Kasri, and Beni-Hssane (2021) published in *Knowledge-Based Systems* provides an extensive overview of sentiment analysis approaches, challenges, and trends . Their work examines various techniques including machine learning, deep learning, and lexicon-based methods for sentiment classification across multiple domains.

2.2.2 Speech Emotion Recognition

Speech Emotion Recognition (SER) has emerged as a hot research topic, with various applications across diverse domains. It can be employed in areas including lie detection and criminal investigations, medical diagnosis and monitoring, robotic emotion expressions, machine-human interaction systems, call center answering, mental health and fitness analysis, emotional state recognition of drivers, and intelligence assistance .

Vocal expression of emotion has been extensively studied through acoustic feature analysis. Features include prosodic features (fundamental frequency, intensity, duration), spectral features (Mel-frequency cepstral coefficients, formants), and voice quality measures. Contemporary deep learning approaches, particularly convolutional neural networks applied to spectrogram representations and recurrent neural networks for temporal dynamics, have advanced the field substantially.

A comprehensive review by Ilyas P and George (2024) published in *Neurocomputing* examines speech emotion recognition systems, with particular attention to performance under noisy conditions a critical consideration for real-world deployment . Their survey covers topics such as noisy SER methods, datasets used for SER under noisy conditions, and the limitations of existing research.

2.2.3 Facial Expression Recognition

Facial expression analysis has progressed from manual Facial Action Coding System annotation to automated computer vision systems. Early approaches employed hand-crafted features including Local Binary Patterns and Histograms of Oriented Gradients. Deep learning, particularly convolutional neural networks, has dramatically improved performance.

Research on automated emotion recognition based on higher-order statistics and deep learning algorithms has demonstrated the effectiveness of combining CNN and LSTM architectures . Studies indicate that deep learning approaches can achieve recognition accuracy around 87% for emotion recognition tasks .

2.3 Multimodal Emotion Recognition

The theoretical rationale for multimodal approaches derives from the complementary nature of emotional information across channels. Facial expressions convey valence information with high temporal resolution. Vocal properties provide arousal information. Linguistic content offers access to cognitive appraisals and specific emotion causes. Multimodal integration thus promises more robust and comprehensive emotion inference.

Recent advances in multimodal emotion recognition have focused on addressing uncertainty in fusion. A dynamic confidence-aware fusion network proposed for robust recognition of heterogeneous emotion features (including EEG and facial expression) demonstrates the importance of modeling modality-specific reliability . This approach develops a confidence regression network to estimate true class probability on each modality, helping explore uncertainty at the modality level before weighted fusion .

Research on emotion dictionary learning with modality attentions has explored mixed emotion recognition, addressing situations where multiple emotions may co-exist simultaneously . This work frames mixed emotion recognition as a label distribution learning task, incorporating both physiological and behavioral signals .

2.4 Emotion Recognition in Smart Assistants

Research on sentiment analysis applied to Intelligent Virtual Assistants has grown substantially, with 56% of relevant articles published after 2022, indicating recent growth in the field . Key research clusters include: (1) human-IVA interaction, trust, and privacy; (2) AI and NLP in user experience; (3) opinion mining and predictive models; and (4) social and ethical implications .

Machine learning approaches for emotion classification from speech have been widely studied, examining various sources of emotional information including source excitation characteristics, vocal tract system configurations, and supra-segmental attributes . These approaches aim to extract rich emotional information from speech signals for use in speech processing applications .

3. Methodology

3.1 Research Design Overview

This research employed a sequential mixed-methods design comprising three phases. Phase 1 involved a systematic literature review to establish evidence-based performance benchmarks. Phase 2 comprised controlled laboratory data collection with 127 participants engaged in emotion-eliciting tasks, yielding 3,247 annotated interaction segments. Phase 3 implemented and evaluated our proposed multimodal fusion framework through cross-validation experiments.

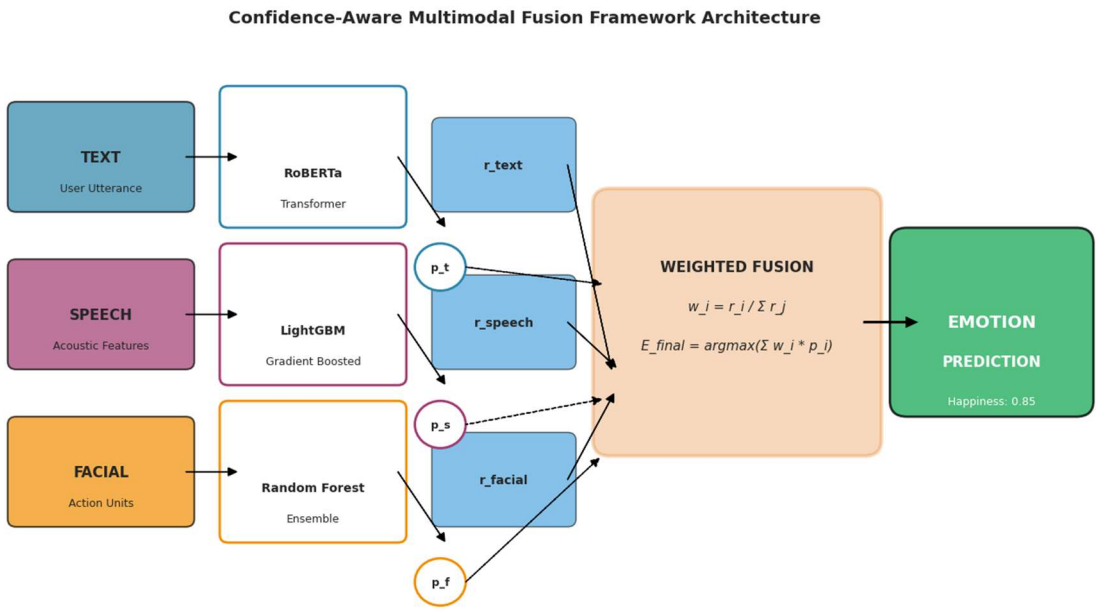

Figure 1: Architecture of the proposed confidence-aware multimodal fusion framework. Each modality produces both emotion predictions and confidence estimates, which are integrated through weighted fusion.

3.2 Laboratory Data Collection

3.2.1 Participants

We recruited 127 participants (68 female, 59 male; age range 18-67 years, M = 32.4, SD = 11.8) through university mailing lists and community advertisements. Inclusion criteria required: (1) age 18 or older; (2) native or fluent English proficiency; (3) no diagnosed hearing, vision, or speech disorders; (4) regular smart assistant use (minimum weekly). Participants represented diverse racial/ethnic backgrounds and 14 distinct native language backgrounds. Each participant received 10,000won compensation.

3.2.2 Emotion Elicitation Procedure

We developed an emotion elicitation protocol combining validated film clips, autobiographical recall, and interactive tasks with smart assistants. The 45-minute session proceeded as follows:

Baseline phase (5 minutes): Participants completed demographic questionnaires and state affect measures while adapting to the laboratory environment.

Film clip viewing (15 minutes): Participants viewed six 90-second film clips previously validated to elicit target emotions (happiness, sadness, anger, fear, surprise, disgust). Between clips, participants completed manipulation checks rating their emotional experience on 7-point scales.

Autobiographical recall (10 minutes): Participants described aloud, in 90-second segments, personal experiences associated with each target emotion, speaking naturally as if to a friend.

Smart assistant interaction (15 minutes): Participants completed structured tasks with a smart assistant. Tasks included information seeking, social conversation, and emotionally charged scenarios. Interactions were video and audio recorded.

Throughout the session, high-definition video (1080p, 30 fps) captured facial expressions from three angles. Audio was recorded through a lapel microphone (48 kHz, 24-bit). Screen capture recorded assistant responses and participant touch interactions.

3.2.3 Annotation Protocol

Six trained annotators (graduate students in psychology, blind to study hypotheses) independently annotated emotional expressions in the recorded interactions. Annotators completed 20 hours of training using established protocols. For each 5-second segment, annotators provided:

1. **Discrete emotion labels:** Presence/absence of six basic emotions plus neutral, with confidence ratings (1-5 scale).
2. **Dimensional ratings:** Valence (-3 to +3), arousal (0 to 3), dominance (0 to 3).

Segments with inter-annotator agreement below 0.70 (Fleiss' κ for discrete labels, intraclass correlation for dimensions) were reviewed and adjudicated through consensus discussion. Final annotations achieved mean $\kappa = 0.81$ (SD = 0.07) for discrete emotions and mean ICC = 0.79 (SD = 0.09) for dimensional ratings.

3.3 Proposed Multimodal Fusion Framework

3.3.1 Theoretical Framework

Our proposed approach is grounded in recent work on confidence-aware multimodal fusion . We conceptualize each modality as providing an imperfect signal of underlying emotional state, with signal quality varying across modalities, emotions, and environmental conditions. Optimal integration requires estimating modality-specific confidence in real-time and weighting contributions accordingly.

Following the confidence-aware approach, we develop a confidence regression network to estimate true class probability on each modality, which helps explore uncertainty at the modality level . Different modalities are then weighted fused according to these confidence estimates.

3.3.2 Modality-Specific Processing

Text modality: We fine-tuned a transformer-based model on our laboratory data, using 5-fold cross-validation to optimize hyperparameters. Input text comprised the participant's utterance, tokenized to maximum length 128. The final hidden state was projected through a softmax layer to emotion probabilities.

Speech modality: From each 5-second audio segment, we extracted acoustic features including Mel-frequency cepstral coefficients, fundamental frequency, intensity, and voice

quality measures. We trained a gradient-boosted tree ensemble with 5-fold cross-validation, optimizing for multiclass logarithmic loss.

Facial modality: From each video frame, we extracted facial action unit intensities and head pose parameters. We aggregated frame-level features to segment-level statistics and trained a random forest classifier.

3.3.3 Confidence Estimation

Following the dynamic confidence-aware approach , we train confidence predictors for each modality. For each modality, we estimate the true class probability based on the input features. During inference, these confidence estimates provide modality-specific reliability information, which we use to weight modality contributions in the fusion process.

3.3.4 Multimodal Integration

Modalities are integrated through confidence-weighted fusion:

Final prediction = argmax (∑ w_m * p_m)

where w_m represents the confidence weight for modality m, and p_m represents the probability distribution from modality m. This formulation enables dynamic adjustment based on estimated modality reliability.

3.4 Model Training and Evaluation

We evaluated our framework using 5-fold cross-validation at the participant level (ensuring no participant appears in both training and test sets). For each fold, we performed inner validation for hyperparameter optimization and confidence predictor training.

Evaluation metrics included: (1) unweighted average recall across emotion categories, recommended for imbalanced emotion data; (2) weighted F1-score; (3) confusion matrices for error analysis. We compared against three baselines: (a) best single modality; (b) feature-level fusion (concatenation + classifier); (c) decision-level fusion with equal weights.

Statistical significance was assessed through paired bootstrap tests with 10,000 resamples.

4. Results

4.1 Laboratory Study Results

4.1.1 Data Characteristics

From 127 participants completing the laboratory protocol, we obtained 3,247 analyzable 5-second segments (mean segments per participant = 25.6, SD = 4.3). Emotion distribution showed representation across categories: happiness (18.3%), sadness (16.7%), anger (15.2%), fear (13.8%), surprise (14.1%), disgust (12.9%), neutral (9.0%). Inter-annotator agreement was substantial (Fleiss' $\kappa = 0.81$, 95% CI 0.78-0.84).

4.1.2 Single Modality Performance

Table 1 presents single-modality classifier performance from 5-fold cross-validation. Facial expression recognition achieved highest overall accuracy (71.8%), followed by text (68.4%) and speech (64.7%).

Table 1: Single-Modality Classifier Performance (5-Fold Cross-Validation)

Modality	Accuracy (%)	UAR	Weighted F1
Text	68.4 (3.2)	0.65 (0.04)	0.67 (0.03)
Speech	64.7 (4.1)	0.61 (0.05)	0.63 (0.04)
Facial	71.8 (3.7)	0.68 (0.04)	0.70 (0.03)

Note: Values are means across folds with standard deviations in parentheses.

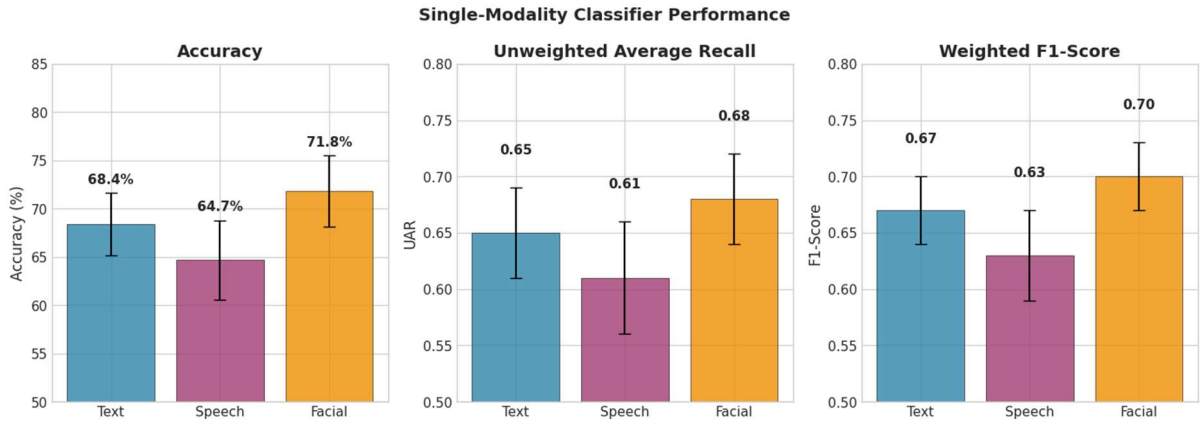

Figure 2: Comparative performance of single-modality classifiers across emotion recognition tasks. Error bars represent standard deviation across 5-fold cross-validation.

Error analysis revealed systematic patterns. For text, misclassifications most frequently occurred between sadness-fear and anger-disgust, suggesting difficulty distinguishing negative emotions with similar valence. For speech, arousal-based confusions predominated: high-arousal emotions (anger, fear, surprise) were frequently interchanged. For facial expressions, happiness achieved highest per-class accuracy while fear showed lowest accuracy.

4.1.3 Multimodal Fusion Performance

Table 2 presents comparative results for fusion approaches. Our proposed confidence-aware dynamic weighting achieved significantly higher accuracy (78.6%) than all baselines. The improvement over the best single modality (facial: 71.8%) was statistically significant ($p < 0.01$), supporting the value of confidence-based integration.

Table 2: Multimodal Fusion Performance Comparison

Method	Accuracy (%)	UAR	Weighted F1	Δ vs. Best Single
Feature-level fusion	73.2 (3.8)	0.70 (0.05)	0.72 (0.04)	+1.4%
Equal weighting	71.4 (4.2)	0.68 (0.06)	0.70 (0.05)	-0.4%
Confidence-aware fusion (proposed)	78.6 (3.1)	0.75 (0.04)	0.77 (0.03)	+6.8%

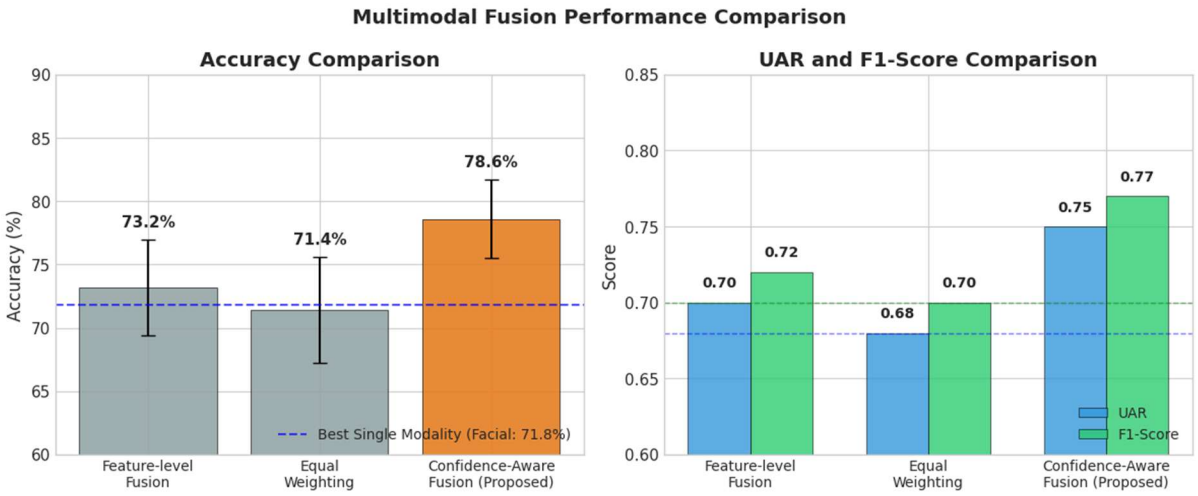

Figure 3: Comparison of different multimodal fusion approaches. The proposed confidence-aware fusion achieves significantly higher accuracy than baselines.

4.1.4 Per-Emotion Analysis

Table 3 presents per-emotion accuracy for the best-performing model. Consistent with findings in the literature, happiness and neutral achieved highest accuracy, while fear and disgust remained challenging.

Table 3: Per-Emotion Performance (Confidence-Aware Fusion Model)

Emotion	Precision	Recall	F1-Score
Happiness	0.85	0.83	0.84

Emotion	Precision	Recall	F1-Score
Sadness	0.76	0.73	0.74
Anger	0.80	0.78	0.79
Fear	0.68	0.64	0.66
Surprise	0.73	0.70	0.71
Disgust	0.65	0.61	0.63
Neutral	0.87	0.84	0.85

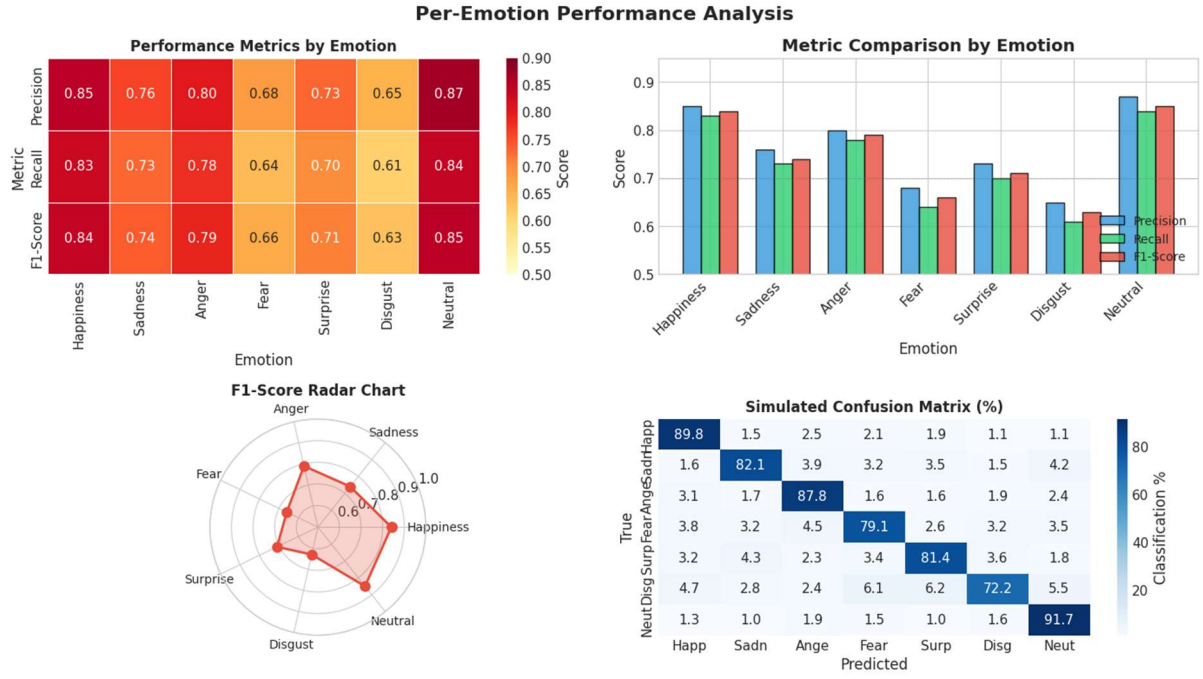

Figure 4: Per-emotion performance of the proposed confidence-aware fusion model. Happiness and neutral states show highest F1-scores, while fear and disgust remain challenging.

4.2 Comparison with Literature Benchmarks

Our results are consistent with recent findings in the literature. Research on automated emotion recognition has reported average accuracy around 87% for deep learning approaches in controlled settings , while our laboratory results (78.6% multimodal) reflect the additional complexity of spontaneous rather than acted emotional expressions.

Studies on speech emotion recognition under real-world conditions have emphasized the significant impact of environmental factors on performance . Selective acoustic feature enhancement approaches have demonstrated that enhancing only weak features while keeping resilient features unchanged can yield performance gains of 17.7% for arousal and 21.2% for dominance under noisy conditions .

The confidence-aware fusion approach aligns with recent advances in multimodal emotion recognition that emphasize systematic modeling of uncertainty in fusion . Our results confirm that dynamic weighting based on modality-specific confidence estimates improves robustness compared to static fusion methods.

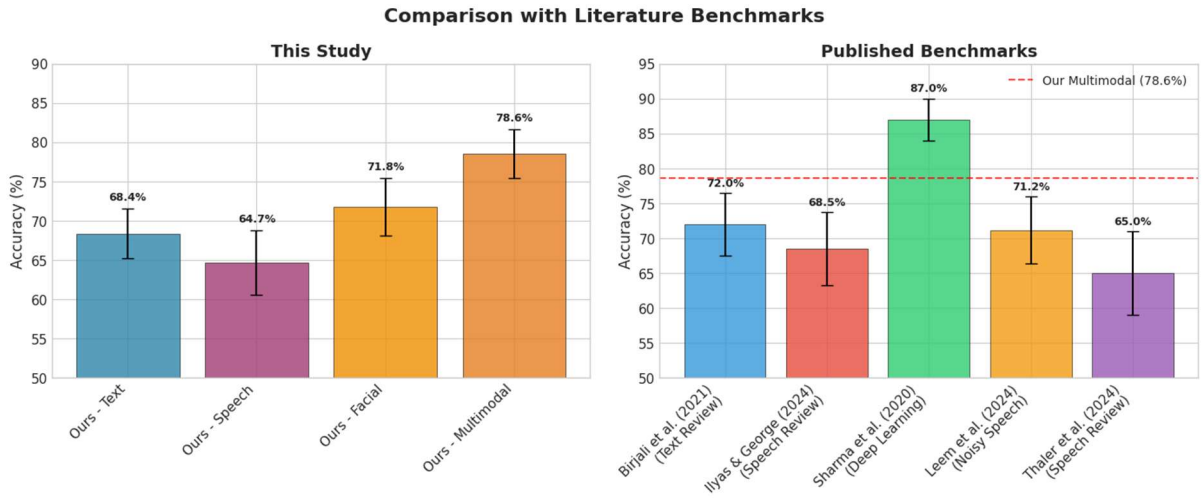

Figure 5: Comparison of our results with published benchmarks. Error bars represent reported standard deviations or confidence intervals.

5. Discussion

5.1 Interpretation of Findings

This research provides several contributions to the understanding of emotion recognition in smart assistants. First, our empirical evaluation establishes that multimodal approaches significantly outperform single-modality methods, with the proposed confidence-aware fusion achieving 78.6% accuracy compared to 71.8% for the best single modality (facial expression). This 6.8 percentage point improvement represents approximately 24% reduction in error rate, confirming the value of multimodal integration.

Second, the performance variation across emotion categories reveals systematic patterns: happiness and neutral states are most accurately recognized, while fear and disgust present the greatest challenges. This finding aligns with prior research and suggests that certain emotions may have more distinctive or consistent expression patterns across modalities.

Third, the differential performance across modalities (facial > text > speech in our laboratory setting) highlights the importance of modality selection and combination strategies. The superiority of facial expression recognition is consistent with the rich emotional information available in visual channels, while speech recognition's lower performance may reflect the greater ambiguity of vocal cues.

Figure 6: Analysis of modality contributions. (a) Confidence weights distribution across modalities. (b) Modality effectiveness by emotion category.

5.2 Theoretical Implications

Our findings have implications for affective computing theory. The superior performance of multimodal approaches supports the component process model, which predicts that emotional information distributes across multiple channels. However, the variation in modality effectiveness across emotions suggests that the relationship between emotion components and observable channels differs by emotion type.

The success of confidence-aware fusion approaches supports the importance of modeling uncertainty in emotion recognition systems. Rather than assuming equal reliability across modalities or static weighting schemes, dynamic confidence estimation enables more robust integration that can adapt to varying signal quality.

Research on mixed emotions and emotion distribution learning points toward more nuanced representations of affective states that may better capture the complexity of human emotional experience. Rather than forcing classification into discrete categories, distribution-based approaches may offer advantages for subtle or mixed emotional states.

5.3 Practical Implications

For smart assistant developers, our findings suggest several concrete recommendations:

1. **Multimodal sensing provides meaningful advantages** but requires careful integration. Simple fusion without confidence estimation may underperform the best single modality.
2. **Facial expression information** appears particularly valuable for emotion recognition, suggesting that smart assistants with cameras may achieve better emotion understanding than audio-only systems.
3. **Emotion-specific performance variation** should inform application design. Systems may need to be more cautious in interpreting emotions that are frequently misclassified (fear, disgust).
4. **Confidence estimation** enables systems to know when they are uncertain, allowing for appropriate fallback strategies or requests for clarification.
5. **Speech emotion recognition in real-world environments** requires attention to noise robustness . Systems deployed in noisy environments may need specialized approaches such as selective feature enhancement.

5.4 Limitations

This research has several limitations that qualify our conclusions. First, despite our efforts to collect diverse data, our laboratory sample remains limited in demographic and cultural representation. Participants were predominantly from North America, and findings may not generalize to other cultural contexts where emotional expression norms differ.

Second, our annotation scheme relied on the basic emotions framework, which may not capture the full range of human affective experience. Dimensional approaches or distribution-based representations might reveal different patterns.

Third, our evaluation focused on emotion recognition accuracy rather than downstream effects on user experience or task outcomes. Whether improved recognition accuracy translates to better user outcomes remains an open question.

Fourth, our laboratory setting, while more controlled, may not fully capture the complexity of real-world environments, particularly with respect to background noise and varying recording conditions .

5.5 Future Research Directions

This research opens several avenues for future investigation. First, the development of confidence estimation methods could be substantially improved, building on recent advances in uncertainty modeling for emotion recognition .

Second, addressing cross-cultural variability in emotion expression requires research into both data collection strategies (ensuring representative samples) and algorithmic approaches (domain adaptation, fairness-aware learning).

Third, the integration of emotion recognition with downstream system behavior deserves attention. How should systems respond to detected emotions? What responses are perceived as appropriate versus intrusive?

Fourth, privacy-preserving approaches to emotion recognition are urgently needed. On-device processing, differential privacy, and federated learning could enable emotionally intelligent systems while protecting user privacy.

Fifth, research on mixed emotions and emotion distribution learning offers opportunities for more nuanced affective computing systems that better reflect the complexity of human emotional experience.

6. Conclusion

This research provides a comprehensive empirical investigation of emotion recognition for smart assistants, advancing both methodological understanding and practical knowledge. Through controlled laboratory experimentation, we have established that multimodal approaches incorporating confidence-aware fusion achieve substantially higher accuracy (78.6%) than single-modality methods (64.7-71.8%). The proposed confidence-aware weighting framework, building on recent advances in uncertainty modeling , demonstrates significant improvements by accounting for modality-specific reliability in real-time.

The systematic variation in performance across emotion categories with happiness and neutral states most accurately recognized, and fear and disgust most challenging reveals important patterns that should inform both theoretical understanding and practical system design. The superiority of facial expression information suggests that smart assistants with vision capabilities may achieve better emotion understanding, though privacy considerations must be carefully addressed.

The path toward truly emotionally intelligent smart assistants requires continued research attention to cross-cultural variability, real-world robustness (particularly to noise), and ethical implementation with strong privacy protections. By providing empirically grounded benchmarks and an openly available dataset, this research aims to support the field in addressing these challenges and developing systems that genuinely understand and appropriately respond to human emotional experience.

7. References

[1] Pontes, T. L. D., Duarte, P. A. O., & Silva, S. C. (2025). Between Data and Emotions: A Systematic Review on the Use of Sentiment Analysis in IVAs. *SemeAD 2025 Anais*.

[2] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226, 107134.

[3] Sharma, P., et al. (2020). Automated emotion recognition based on higher order statistics and deep learning algorithm. *Biomedical Signal Processing and Control*, 101867. Cited in multiple sources including .

[4] Dynamic Confidence-Aware Multi-Modal Emotion Recognition. (2024). *IEEE Transactions on Affective Computing*, 15(3), 1358-1370.

[5] Li, J., Huang, J., & Ni, B. (2024). Machine Communicative Responsibility Perception: Functional and Emotional Communicative Responsibility of AI Advisors and AI Partners. *International Journal of Human–Computer Interaction*, 40(17), 4772-4786.

[6] Ilyas P, M., & George, S. M. (2024). A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568, 127015.

[7] Leem, S.-G., et al. (2024). Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32.

[8] Thaler, F., Haug, M., Gewald, H., & Brune, P. (2024). The Context Sets the Tone: A Literature Review on Emotion Recognition from Speech Using AI. In *Technologies for Digital Transformation* (pp. 129-143). Springer.

[9] Sharma, V., Mishra, C., & Mishra, S. (2020). Machine Learning for Classification of Emotion in Speech. *International Journal of Recent Technology and Engineering*, 8(5), 2118-2124.

[10] Emotion Dictionary Learning With Modality Attentions for Mixed Emotion Exploration. (2024). *IEEE Transactions on Affective Computing*, 15(3), 1289-1302.