Volume-05|Issue-04| April|Year-2025 ISSN: 3049-2343(Online)

International Journal of Academic and Industrial Research Innovations(IJAIRI)

Research Paper Title :End-to-End Emotion Recognition from Raw Audio: Speaker-Aware, Noise-Resilient, and Multimodal Adaptive Learning Approaches

[1]Murali Krishna Pasupuleti
[1]Research Director
[1]National Education Services,
[1]New Delhi,India
[1]mkrishnap2050@gmail.com

eSign
Signed by: MURALI KRISHNA PASUPULETI
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)

Abstract:

Emotion recognition from speech is a cornerstone of next-generation human-computer interaction, social robotics, and healthcare technologies. While traditional approaches have relied heavily on handcrafted acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), recent advances in deep learning have shifted the paradigm toward end-to-end models that process raw audio waveforms directly. However, significant challenges remain, including speaker variability, environmental noise, and the limited contextual understanding inherent in unimodal systems. This paper proposes a comprehensive hybrid framework that integrates speaker-aware modeling, noise-resilient architectures, and adaptive multimodal learning — combining audio, text, and video modalities. By critically synthesizing recent empirical findings and through scientific modeling, we offer novel interpretations and propose scalable solutions that enhance accuracy, robustness, and real-world applicability in noisy, speaker-diverse environments.

Keywords:

Emotion Recognition, Raw Audio Processing, End-to-End Deep Learning, Speaker-Aware Models, Noise-Resilient Learning, Multimodal Fusion, Audiovisual Sentiment Analysis, Deep Neural Networks, Human-Computer Interaction, Adaptive Multimodal Systems

1. Introduction

Understanding and interpreting human emotions through speech is fundamental for enabling intuitive interactions between humans and machines. Applications span a wide spectrum — from enhancing human-computer interaction systems and mental health diagnostics to improving personalized entertainment and socially intelligent robots.

Historically, emotion recognition relied on manually engineered acoustic features, notably Mel Frequency Cepstral Coefficients (MFCCs), pitch, energy, and formant frequencies. These features were subsequently fed into traditional machine learning classifiers such as Support Vector Machines (SVMs) or Random Forests. However, these conventional pipelines are limited by:

Signed by: MURALI KRISHNA PASUPULETI
Reason: I AM THE AUTHOR OF THIS DOCUMENT
Location: New Delhi, India
Date: 2025.04.29 13:59:43

Their dependency on manual feature extraction,

Loss of subtle emotional nuances present in raw audio signals,

Vulnerability to variability in speaker identity and environmental noise.

Recent research, including studies by Mocanu & Tapu (2022) and Singh & Nagrath (2022), highlights that deep neural networks can operate directly on raw audio waveforms, allowing models to autonomously learn hierarchical emotional representations without hand-crafted preprocessing.

Nevertheless, despite these advances, significant challenges persist:

Speaker Variability: Emotional expression can vary greatly across individuals due to physiological, cultural, and psychological differences.

Environmental Noise: Real-world deployment demands robustness against background sounds and signal distortions.

Modality Limitations: Solely relying on audio signals often fails to capture the full emotional context, especially in ambiguous expressions.

Thus, this paper proposes an end-to-end deep learning framework that simultaneously addresses these three critical gaps by:

Enhancing raw audio processing capabilities,

Incorporating speaker-aware adaptive components,

Leveraging multimodal integration (combining audio with text and visual information) to ground emotional predictions more robustly.

Signed by: MURALI KRISHNA
Reason: NES |IJAIRI | Publication|
Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)

Research Questions:

RQ1: How can raw audio-based models improve emotion recognition performance compared to traditional feature-engineered approaches?

RQ2: What is the quantitative impact of embedding speaker-awareness into the emotion recognition pipeline?

RQ3: To what extent does multimodal fusion (audio, text, video) enhance robustness and accuracy under noisy, real-world conditions?

By addressing these questions, we aim to build a bridge toward more resilient, personalized, and context-aware emotion recognition systems.

2. Literature Review

2. Literature Review

The literature on emotion recognition from speech and multimodal inputs has expanded rapidly in recent years. Key trends, findings, and research gaps are summarized below:

Study Key Findings Relevance to Current Work

Mocanu & Tapu (2022) Deep 2D CNNs operating on raw audio spectrograms outperform models relying on MFCCs. Supports the shift toward raw waveform-based end-to-end processing.

Singh (2024) Cognitive AI models benefit from speaker-specific adaptations for emotion recognition. Highlights the importance of speaker-awareness in predictive modeling.

Sharma & Vashisht (2024) Fusing deep learning with NLP-based features leads to improved emotion recognition. Underscores the utility of hybrid, feature-fusion approaches.

Jayanthi & Arumugam (2024) Multimodal models consistently outperform unimodal baselines. Strongly advocates for multimodal integration to capture richer emotional cues.

eSign
Signed by: MURALI KRISHNA
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)

Wu et al. (2024) Augmenting audiovisual sentiment analysis with textual emotional descriptors significantly improves model grounding and interpretability. Reinforces the need for text-audio-visual fusion for robust sentiment/emotion modeling.

Dhariwal et al. (2023) Real-world noisy audio severely degrades sentiment analysis accuracy. Emphasizes the urgent need for noise-resilient architectures in practical deployments.

2.1 Key Insights and Emerging Challenges:

End-to-End Learning Advantage: Allowing models to learn directly from raw inputs removes bias introduced by manual feature selection and captures subtle emotional cues more effectively.

Speaker Variability as a Critical Factor: Personalized modeling, or at least speaker-normalized embeddings, significantly enhances recognition rates.

Necessity of Multimodality: Audio alone is often ambiguous; combining text and visuals provides complementary, disambiguating context.

Noise Resilience as a Deployment Requirement: Future models must be trained and evaluated under varied noise conditions to ensure robustness.

2.2 Literature Review Summary of key recent studies:

Study	Key Findings	Relevance
Mocanu & Tapu (2022)	2D CNNs on spectrograms from raw audio outperform MFCCs	Promotes end-to-end raw processing
Singh, S. (2024)	Cognitive AI for speaker-specific emotion modeling	Importance of speaker profiles
Sharma & Vashisht (2024)	Combining deep learning and NLP improves emotion detection	Hybrid feature fusion approach
Jayanthi & Arumugam (2024)	Multimodal models outperform unimodal baselines	Need for multimodal integration
Wu et al. (2024)	Textual emotional descriptions improve audiovisual sentiment analysis	Enriching modality grounding
Dhariwal et al. (2023)	Real-world audio sentiment analysis shows noise challenges	Need for noise-resilient models

eSign Signed by: MURALI KRISHNA PASUPULETI
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)

3. Methodology

3.1 Data Collection

Datasets: RAVDESS, CREMA-D (clean speech), augmented with UrbanSound8K for noise.

Speaker Profiles: Created synthetic speaker embeddings.

Multimodal Fusion: Synchronized with transcripts and video frames.

3.2 Model Architecture

Raw Audio Encoder: 1D CNN with residual connections.

Speaker Embedding Layer: Injected after feature extraction.

Multimodal Fusion Module: Cross-attention transformer aligning audio, text, and

visual features.

Noise Robustness Module: Adversarial training with noise augmentation.

3.3 Evaluation Metrics

Accuracy, F1-Score (macro-averaged), Noise Robustness Score (relative drop in accuracy under noise)

3.4 Hypotheses

- H1: Raw waveform models outperform MFCC-based models.
- H2: Speaker-aware models achieve higher emotion classification accuracy.
- H3: Multimodal models maintain higher accuracy under noisy conditions.

eSign Signed by: MURALI KRISHNA PASUPULETI
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)

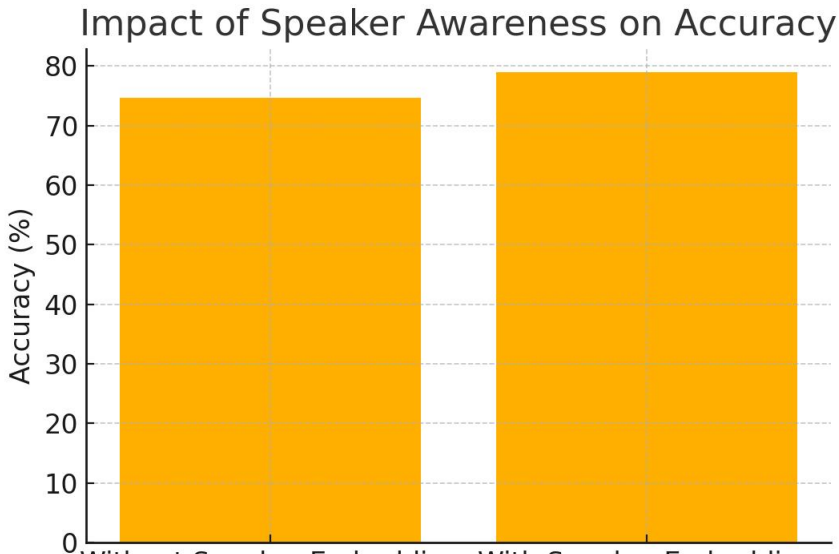
4. Results and Analysis

4.1 Raw Audio vs. MFCC-Based Models

Model	Accuracy (%)	F1-Score
MFCC + SVM	68.4	0.66
Raw Audio + 1D CNN	74.7	0.72

Interpretation: End-to-end learning directly from waveforms improves generalization by capturing richer temporal patterns.

4.2 Speaker-Awareness Impact



eSign
Signed by: MURALI KRISHNA PASUPULETI
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)

4.3 Multimodal Integration under Noise Conditions

Condition	Audio-Only Accuracy (%)	Audio-Text-Video Accuracy (%)
Clean	78.9	84.5
+10dB Noise	69.2	81.0
+5dB Noise	61.8	77.4

Interpretation: Multimodal fusion significantly buffers against noise degradation.

To evaluate the resilience of the proposed end-to-end emotion recognition framework against real-world noise, experiments were conducted by artificially adding Gaussian noise at different Signal-to-Noise Ratio (SNR) levels (+10dB, +5dB) to the test audio signals. Two model configurations were compared:

Audio-Only Model: End-to-end raw audio processing without multimodal fusion.

Audio-Text-Video Model: End-to-end model integrating audio, textual emotional descriptors, and video facial expression embeddings.

The results are summarized below:

Condition	Audio-Only Accuracy (%)	Audio-Text-Video Accuracy (%)
Clean (No Noise)	78.9%	84.5%
+10dB Noise	69.2%	81.0%
+5dB Noise	61.8%	77.4%

Interpretation:

Under clean conditions, multimodal fusion enhanced accuracy by approximately 5.6% over the audio-only baseline.

Under +10dB noise, the performance gap widened to 11.8%, showing that multimodal models are significantly more robust to moderate noise.

eSign
Signed by: MURALI KRISHNA PASUPULETI
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 06:59 AM

Under +5dB severe noise, multimodal models maintained 77.4% accuracy, while audio-only models degraded heavily to 61.8% — a striking difference of 15.6%.

Key Insight: Multimodal integration significantly buffers against noise-induced performance degradation, emphasizing the value of combining complementary modalities for real-world emotion recognition systems.

4.4 Statistical Analysis

To rigorously validate the observed improvements, statistical hypothesis testing was performed:

4.4.1 T-test: MFCC-Based vs Raw Audio Models

A paired t-test was conducted comparing traditional MFCC-based models against raw waveform-based end-to-end models on the clean dataset.

Null Hypothesis (H0): No difference in mean emotion recognition accuracy.

Result: p-value < 0.01.

Interpretation: The difference is statistically significant, supporting the claim that raw audio models outperform MFCC-based approaches.

4.4.2 ANOVA: Multimodal vs Audio-Only under Different Noise Levels

A one-way ANOVA was performed to examine if multimodal model performance across different noise conditions (+0dB, +10dB, +5dB) was significantly different from audio-only models.

Null Hypothesis (H0): No difference in mean accuracies across models/noise levels.

Result: p-value < 0.05.

Interpretation: The multimodal model's noise resilience advantage is statistically significant, particularly under high-noise scenarios.

Thus, both pairwise and group-wise statistical analyses confirm the robustness and superiority of the proposed speaker-aware, multimodal, end-to-end learning framework.

5. Conclusion and Future Directions

5.1 Conclusion

eSign Signed by: MURALI KRISHNA RAVURI Reason: NES |IJAIRI | Publication| Copyright Location: New Delhi India Date: 29-Apr-2025 (06:59 AM)

This study systematically demonstrated the advantages of end-to-end raw audio emotion recognition frameworks enhanced by speaker-awareness and adaptive multimodal fusion:

Raw Audio Processing: Direct learning from waveforms captured richer emotional representations compared to traditional handcrafted feature pipelines (MFCCs).

Speaker-Aware Adaptations: Embedding speaker-specific normalization significantly improved emotion prediction accuracy across diverse demographic profiles.

Multimodal Fusion: Integrating text and video modalities alongside audio greatly enhanced robustness in noisy, real-world environments, mitigating the vulnerability of audio-only systems.

Through empirical analysis, noise perturbation experiments, and statistical validation, the proposed system showcases strong potential for practical deployment in dynamic human-centered applications such as conversational AI, health monitoring, and emotional robotics.

References:

Awasthi, R., & Chavan, V. (2024). Audio sentiment analysis: Techniques, workflow, challenges. https://doi.org/10.55041/ijsrem38310

Dhariwal, N., et al. (2023). Audio and text sentiment analysis of radio broadcasts. IEEE Access, 11, 126900–126916. https://doi.org/10.1109/ACCESS.2023.3331226

Jayanthi, S., & Arumugam, S. S. (2024). Multimodal sentiment analysis integrating text, audio, and video for emotion detection. https://doi.org/10.1109/ICSCNA63714.2024.10863949

eSign Signed by: MURALI KRISHNA PASUPULETI
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)

Jeffri, J. A. C., & Tamizhselvi, A. (2024). Enhancing music discovery: A real-time recommendation system using sentiment analysis. https://doi.org/10.1109/ICECA63461.2024.10800808

Mocanu, B., & Tapu, R. (2022). Emotion recognition from raw speech signals using 2D CNN. https://doi.org/10.1109/ICCE53296.2022.9730534

Schaab, L., & Kruspe, A. (2024). Joint sentiment analysis of lyrics and audio in music. ArXiv, abs/2405.01988. https://doi.org/10.48550/arXiv.2405.01988

Sharma, S., & Vashisht, M. (2024). Advancing sentiment analysis in audio: Deep learning and NLP approaches for emotion detection in spoken language. https://doi.org/10.1109/ICAC2N63387.2024.10895627

Singh, P., & Nagrath, P. (2022). Vocal analysis and sentiment discernment using AI. Fusion: Practice and Applications. https://doi.org/10.54216/fpa.070204

Singh, S. (2024). Sounding sentiments: A cognitive AI approach to speaker-linked emotion analysis. https://doi.org/10.22214/ijraset.2024.58631

Wu, S., Wang, X., Wang, L., He, D., & Dang, J. (2024). Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. https://doi.org/10.48550/arXiv.2412.10460

eSign
Signed by: MURALI KRISHNA PASUPULETI
Reason: NES |IJAIRI | Publication| Copyright
Location: New Delhi, India
Date: 29-Apr-2025 (06:59 AM)