==========
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 25:7 July 2025
==========

A Review of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Mithilaj JS
Research Scholar
Department of Linguistics
University of Kerala
mithilaj@keralauniversity.ac.in

Dr. S.A. Shanavas
Professor (Rtd.)
Department of Linguistics
University of Kerala, Trivandrum

Dr.D.Muhammad Noorul Mubarak
Associate Professor
Department of Computer Science
University of Kerala, Trivandrum

Abstract

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a systematically designed and validated multimodal database for English language. It is focused to aid research in emotion recognition, neuroscience, psychology, affective computing and more. This includes 7,000 recordings from 24 professional actors, contains broad variety of emotional expressions captured through both speech and song modalities in audio-visual formats. This paper provides a detailed review of RAVDESS, discussing its design and development, unique features, validation, significance of its multimodal nature in research and highlighting its contributions to advancing emotion research. This paper analyses comparisons with other emotional databases to highlight RAVDESS advantages. The paper also explores its various applications in clinical research and machine learning. Finally, the review underscores potential future directions for enhancing RAVDESS, including expanding its cultural diversity and integrating advanced emotion detection algorithms.

Introduction

The study of emotion forms the foundation of both psychological and neuroscientific research, with important applications in fields such as human-computer interaction, artificial intelligence, and therapeutic practices. Emotions play a crucial role in shaping decision-making, behavior, and social communication, making the accurate detection and interpretation of emotional states vital across disciplines. Rather than being hardwired, emotions are now understood to be dynamically constructed by the brain through real-time predictive processes and interoceptive cues [1]. To advance this area of research, there is a growing need for rich, authentic datasets that reflect the complexity of emotional expressions in real-world scenarios. Multimodal corpora—combining elements like speech, gesture, and text—are particularly valuable for developing AI systems capable of human-like interaction [2].

In response to this need, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was created as a comprehensive multimodal resource that offers naturalistic representations of emotion. Unlike conventional datasets, RAVDESS uniquely incorporates both spoken and sung expressions, acknowledging the expressive power of vocal modulation and melody in emotional communication. This review seeks to offer an in-depth analysis of the RAVDESS dataset, highlighting its design, key features, practical uses, and potential future contributions to emotion research.

Importance of Multimodal Communication in Emotion Research

The human experience of emotion is inherently multimodal, involving a dynamic interplay between verbal and non-verbal signals. Elements such as facial expressions, vocal tone, and body language work together to shape how emotions are perceived and interpreted. However, traditional approaches to emotion research have often relied on unimodal stimuli—such as static facial images or isolated voice recordings—which fail to capture the fluid and integrated nature of emotional expression.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) overcomes this limitation by providing a **multimodal dataset** that includes synchronized audio and video recordings of emotional performances. This approach offers several important advantages:

1. Ecological Validity

In everyday communication, emotions are rarely conveyed through a single channel. For instance, a smile is typically accompanied by a warm, gentle tone, while anger may manifest through both a sharp voice and intense facial expressions. RAVDESS's multimodal structure enhances the ecological validity of emotional stimuli, offering a more realistic representation of how emotions are naturally expressed and experienced in real-life interactions.

2. Multisensory Integration

Empirical research has demonstrated that the brain processes emotional cues from multiple sensory inputs in a coordinated manner. The congruence between facial movements and vocal tone, for example, can strengthen emotional recognition, while incongruity may cause confusion or misjudgment. By including both audio and visual modalities, RAVDESS enables researchers to explore the neural and cognitive mechanisms underlying multisensory integration in emotional perception.

3. Broad Applicability

Thanks to its comprehensive design, RAVDESS is applicable across a wide spectrum of disciplines. It supports studies ranging from the neuroscience of emotional processing to the development of sophisticated emotion recognition systems for human-computer interaction. This versatility is one of the dataset's most valuable attributes, positioning it as a key resource for advancing multimodal emotion research in both scientific and applied domains.

Development of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

The development of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) involved a systematic and thoughtful approach to ensure its value and reliability as a resource for emotion research. Key considerations included the careful selection of actors, the design of emotional stimuli, and a rigorous validation process.

Actors and Stimuli Design

RAVDESS features performances by **24 professional actors**—equally split by gender (12 male and 12 female)—all of whom speak with a neutral **North American English accent**. The selection prioritized individuals with strong acting skills to ensure the consistent and expressive portrayal of emotions across modalities.

The dataset comprises **7,356 high-quality audio-visual recordings**, categorized into two primary types of stimuli:

- **Speech Stimuli**: Lexically matched sentences spoken in English, each performed to express one of **eight distinct emotions**: *neutral, calm, happy, sad, angry, fearful, surprised,* and *disgusted*.
- **Song Stimuli**: Musical phrases sung to represent **six emotions**: *neutral, calm, happy, sad, angry,* and *fearful*.

Each emotional expression is presented at **two intensity levels**—*normal* and *strong*—allowing researchers to investigate how variations in emotional intensity influence recognition accuracy and perception.

The inclusion of both speech and song reflects an understanding of the unique role of melodic expression in emotional communication. Song often evokes stronger and more nuanced emotional responses, making it a valuable complement to speech in studying affective processing.

To ensure emotional authenticity, all expressions in RAVDESS were induced rather than spontaneous. The actors employed established acting techniques, such as method acting and emotional memory recall, to convincingly portray each emotional state [3].

Emotional Stimulation and Intensity

The emotional categories used in RAVDESS were grounded in Ekman's theory of basic emotions, which posits that certain emotional expressions—such as happiness, anger, sadness, and fear—are universally recognized across cultures. Each selected emotion is presented at two intensity levels—*normal* and *strong*—enabling researchers to explore how emotional intensity influences perception and recognition. This is particularly significant in fields like clinical psychology, where the strength of an emotional response can offer critical insight into a person's emotional well-being or mental health condition.

Validation Process

To ensure the accuracy, consistency, and authenticity of the emotional expressions in RAVDESS, a rigorous validation process was undertaken. A total of 247 participants were involved in assessing each stimulus based on three criteria: emotional accuracy, intensity, and genuineness. These evaluations were collected using a standardized rating scale, and the feedback was employed to refine the dataset, ensuring that the expressions were clearly conveyed and uniformly interpreted across all recordings.

Furthermore, a subset of participants participated in a test-retest reliability assessment, which measured the consistency of emotional recognition over time. This step was essential to confirm that the stimuli elicited stable and reproducible interpretations across different observers. Such validation procedures reinforce RAVDESS's credibility and robustness as a tool for emotional research.

Comparison with Other Emotional Databases

To evaluate the performance and unique contributions of RAVDESS, it is essential to compare it with other widely used emotional databases such as **GEMEP**, **CREMA-D**, **MSP-IMPROV**, and **eNTERFACE'05**.

- **GEMEP (Geneva Multimodal Emotion Portrayals)** is a multimodal dataset featuring French-speaking actors. While it shares RAVDESS's multimodal design, its focus on the French language limits its suitability for English-language research. Additionally, although GEMEP includes a broader emotional range, it lacks standardized intensity levels, reducing its consistency for intensity-based analysis.

- **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)** is an English-language dataset that, like RAVDESS, combines audio and visual modalities. However, its reliance on crowd-sourced ratings introduces variability in emotional labeling. In contrast, RAVDESS employs a more controlled and rigorously validated participant group, ensuring greater consistency and reliability.
- **MSP-IMPROV** offers both scripted and natural emotional interactions in a multimodal format. While it provides valuable insights into everyday emotional communication, it does not include **musical stimuli**, a distinctive feature of RAVDESS. This makes RAVDESS particularly suited for research on **musical emotion recognition**.
- **eNTERFACE'05** contains audiovisual recordings of actors portraying six basic emotions. However, its smaller size, limited emotional diversity, and lack of a formal validation process reduce its applicability compared to RAVDESS. The thorough validation and inclusion of song in RAVDESS enhance its robustness as a comprehensive resource for emotion research.

Applications of RAVDESS

One of the most significant strengths of RAVDESS lies in its wide applicability across various research domains. Its multimodal and emotionally diverse dataset makes it a valuable tool in fields such as **neuroscience**, **psychology**, **clinical research**, and **affective computing**.

Neuroscience and Psychology

In neuroscience, RAVDESS can be used to study the neural correlates of emotional processing by examining how the brain responds to emotional stimuli across visual and auditory modalities. Its design supports investigations into how regions such as the **amygdala**, **prefrontal cortex**, and **superior temporal sulcus** respond to different emotional cues. For instance, **fMRI** studies can utilize RAVDESS stimuli to map brain activation patterns associated with specific emotions.

In psychology, RAVDESS serves as a powerful tool for examining **emotion recognition** and **empathy**. Behavioral studies can use RAVDESS to assess how individuals perceive and respond to emotional expressions, contributing to research on **emotional intelligence** and social cognition. Its controlled and validated emotional recordings help researchers explore subtle emotional cues and their interpretations across populations.

Clinical Research

RAVDESS also holds strong potential in clinical research, especially in the study of **mood disorders**, **autism spectrum disorders (ASD)**, and **social anxiety**. The inclusion of a song corpus makes it particularly suitable for **music therapy** research. Researchers can explore how patients with conditions like depression respond to emotional stimuli in both speech and song, and measure changes in perception following interventions such as **cognitive-behavioral therapy (CBT)** or medication.

For example, individuals with depression might exhibit different neural or behavioral responses to happy or sad expressions when compared to healthy controls. RAVDESS enables such comparative studies, offering insights into the emotional processing deficits linked to various disorders. Furthermore, the database can support the development and assessment of **intervention programs** aimed at enhancing **emotional recognition** and **social communication** skills.

Affective Computing and Machine Learning

RAVDESS plays a crucial role in the field of **affective computing**, where the goal is to create systems that can detect, interpret, and respond to human emotions. Its large, balanced dataset is ideal for training **machine learning models** in tasks such as facial emotion recognition and vocal tone analysis.

Researchers can develop algorithms capable of classifying emotions from speech, facial expressions, or combined audiovisual inputs. These models have applications in **virtual assistants**, **social robots**, **interactive games**, and **therapeutic technologies**. The inclusion of musical expressions makes RAVDESS especially valuable for exploring **creative, non-verbal emotional communication**, a domain that remains challenging for artificial systems to decode.

In **natural language processing (NLP)**, RAVDESS can enhance the emotional awareness of conversational agents. By training models to detect the emotional tone in spoken dialogue, developers can build AI systems that adapt their responses based on user sentiment, leading to more **empathetic and engaging interactions**.

In-Depth Analysis of Emotional Expression and Perception

Speech Modality

The speech modality in RAVDESS provides lexically matched statements expressed with a wide range of emotions. This design allows researchers to isolate emotional prosody (tone, pitch, intonation) from semantic content, making it ideal for studying how voice parameters convey emotion.

Key research applications include:

- Analyzing vocal cues such as pitch, tempo, and loudness to distinguish between emotions (e.g., anger vs. fear).
- Investigating how listeners interpret prosodic features in emotional speech.
- Exploring the neural and cognitive mechanisms involved in emotional processing of spoken language.

The database's controlled and validated speech samples enable precise and replicable analysis, making it a valuable tool in both behavioral and neuroimaging studies.

Song Modality

The song modality in RAVDESS offers a rare and important resource for exploring the emotional power of music, an area often underrepresented in emotion research. Music, unlike speech, evokes deep emotional responses through melody, harmony, and rhythm, interacting with vocal tone to shape emotional interpretation.

Research opportunities include:

- Studying the emotional impact of music in music therapy and mood regulation.
- Investigating how melodic structure and vocal performance convey emotion beyond verbal content.
- Exploring the role of music in social bonding and emotional communication.

Result Analysis

Validity Task

To assessed the accuracy of emotion identification in speech and song across modalities (audio-video, video-only, audio-only) and intensities (normal, strong). Key findings include:

Accuracy Measures:

Speech:

Mean propotion correct = 0.72

Song:

Mean propotion correct=0.71; Unbiased hit rate (Hu)=0.53 (moderate).

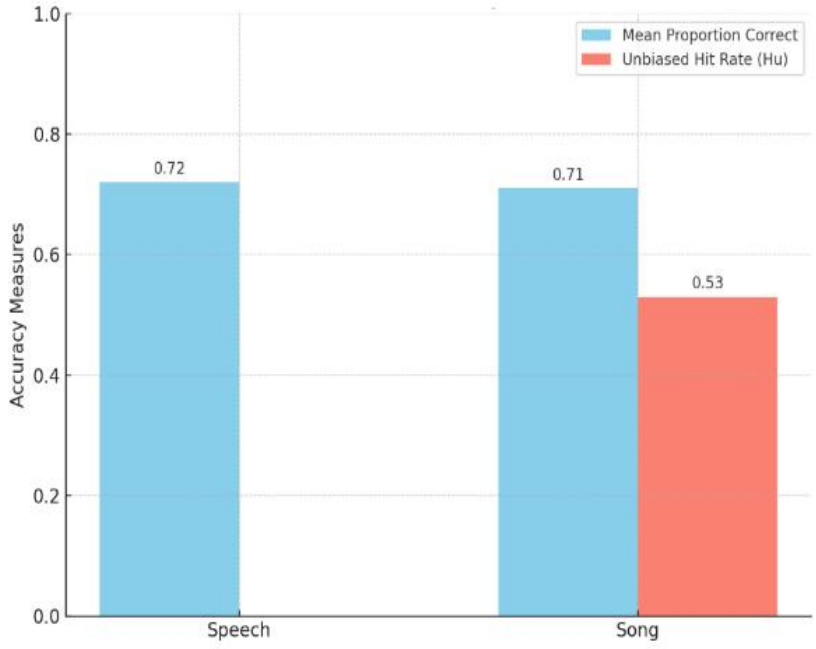

Figure 1:Emotional Identification accuracy in speech and song

The accuracy measures reveal that emotions were identified with comparable success in both speech and song modalities. The mean proportion correct was 0.72 for speech and 0.71 for song, indicating that participants could recognize emotions effectively in both forms of expression. However, the unbiased hit rate (Hu) for song was 0.53, reflecting moderate accuracy after accounting for potential guessing or response biases. This suggests that while emotional cues are present and generally understood in sung expressions, there may be greater variability or ambiguity in interpretation compared to speech. Overall, the findings highlight the robustness of emotional communication in both spoken and musical forms, with slightly higher clarity in speech.

Modality Effects:

Audio-video accuracy = 0.81; Video-only accuracy = 0.75;

Audio-only accuracy = 0.60;

followed by a significant main effect of modality; **F** = 941.68, **p**< 0.001

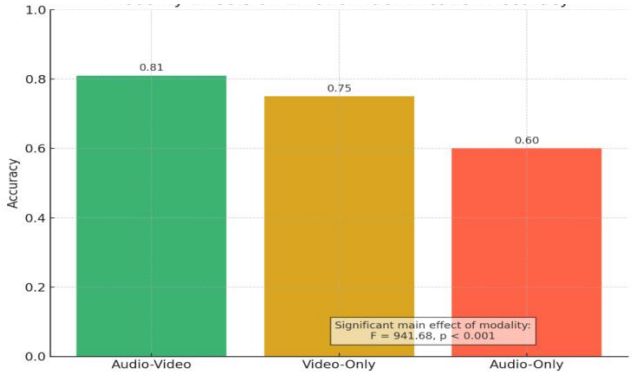

Figure 2: Modality effects of emotional identification accuracy

The analysis of modality effects revealed clear differences in emotion identification accuracy across sensory channels. Accuracy was highest when both audio and visual cues were available (audio-video: 0.81), followed by the video-only condition (0.75), with the lowest accuracy observed in the audio-only condition (0.60). This pattern indicates that the integration of both auditory and visual information significantly enhances emotional understanding. The differences were statistically significant, as confirmed by a main effect of modality, F = 941.68, p< 0.001. These findings underscore the importance of multimodal cues in accurately perceiving emotions, with visual input playing a particularly supportive role when auditory information is limited or ambiguous.

Intensity Effects:

Strong-intensity expressions were identified more accurately than normal-intensity.

M = 0.75 vs M = 0.68; **F** = 402.39, **p**< 0.001.

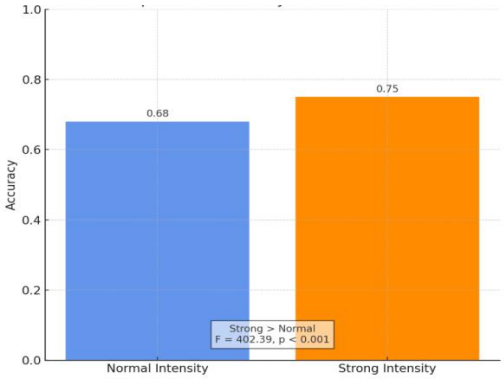

Figure 3: Effects on expression intensity on emotion identification

The analysis of intensity effects demonstrated that strong-intensity emotional expressions were identified significantly more accurately than those expressed with normal intensity. Specifically, the mean accuracy for strong-intensity expressions was 0.75, compared to 0.68 for normal-intensity expressions. This difference was statistically significant, as indicated by a main effect of intensity, $F = 402.39$, $p < 0.001$. These results suggest that heightened emotional intensity enhances the clarity of expressive cues, making emotions easier to recognize and interpret across modalities.

Emotion Specific Perfomance:

Speech: Neutral (M = 0.87) and Angry (M = 0.81) were most accurately identified, while Sadness (M = 0.61) was least accurate.

Song: Angry (M = 0.84) and Neutral (M = 0.78) scored highest, while Fearful (M = 0.65) and Calm (M = 0.63) were lowest.

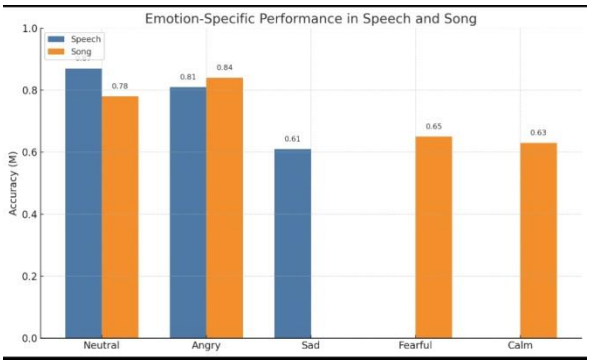

Figure 4: Emotion Specific Perfomance - Evaluation

Emotion-specific analysis revealed notable differences in identification accuracy across categories for both speech and song. In the speech modality, **Neutral** expressions were most accurately recognized (*M* = 0.87), followed closely by **Angry** expressions (*M* = 0.81), while **Sadness** was the least accurately identified (*M* = 0.61). In contrast, within the song modality, **Angry** expressions had the highest accuracy (*M* = 0.84), followed by **Neutral** (*M* = 0.78). The lowest accuracy rates in song were observed for **Fearful** (*M* = 0.65) and **Calm** (*M* = 0.63) expressions. These findings suggest that some emotions—particularly anger and neutrality—are more universally or robustly conveyed across modalities, while others like sadness, fear, and calm may be more ambiguous or context-dependent in how they are expressed and interpreted.

Interrater Agreement

Kappa values:

- Substantial agreement for strong-intensity expressions
  Speech: κ = 0.62; Song: κ = 0.61
- Moderate agreement for normal-intensity expressions
  Speech: κ = 0.53; Song: κ = 0.52

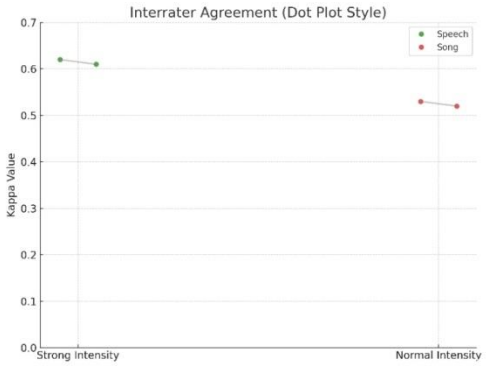

Figure 5: Kappa values

Confusion Patterns:

Common misidentifications included:
Calm confused with Happy (19%),
Sad with Neutral/Calm (17%) and
Happy with Neutral/Calm (14%).

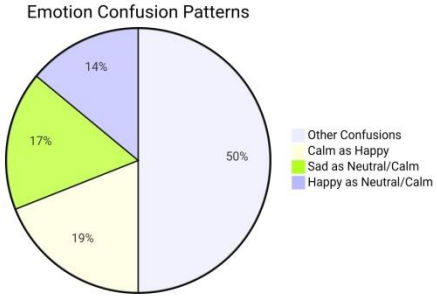

Figure 6: Confusion patterns

Intensity and Genuineness Ratings

Interrater Reliability (ICC):

Single-rater ICCs for intensity and genuineness were poor (0.07–0.22)
Multiple-rater ICCs improved to fair-good (0.42–0.74)

IntensityRatings:

ratings were higher for speech (M = 3.6) than song (M = 3.55).
Strong intensity stimuli (M = 3.83) were rated as more intense than normal intensity stimuli (M = 3.31).

Test-Retest Reliability

Intrarater reliability (Cohen's κ):

Substantial for strong-intensity expressions Speech: κ=0.76; Song: κ=0.77; &

Moderate for normal-intensity Speech: κ = 0.70; Song: κ = 0.68.

ICCs for test-retest reliability:

improved from fair to good
Single-rater: 0.42–0.46 Multiple-rater: 0.59–0.63

Broad Implications of RAVDESS in Contemporary Research

RAVDESS holds substantial value in modern research across domains where emotion, cognition, and technology intersect. Its validated, multimodal emotional expressions make it a reliable benchmark for interdisciplinary studies.

Enhancing Human-Computer Interaction (HCI)

In the field of Human-Computer Interaction, RAVDESS plays a pivotal role in improving the emotional intelligence of digital systems, including:

- Virtual assistants
- Gaming interfaces
- Social robots

These systems can be trained to detect and respond appropriately to users' emotional states. For instance, a virtual assistant utilizing RAVDESS could recognize signs of frustration or anger in a user's voice and respond with calming language or offer empathetic assistance.

The integration of RAVDESS into affective computing can support the development of systems that:

- Interpret emotional nuance in speech and facial expressions.
- Adapt their interaction strategies based on real-time emotional input.
- Improve user experience in areas like customer service, education, and mental health support.

As reinforcement learning continues to push the boundaries of machine intelligence (e.g., in complex environments like *StarCraft II*), databases like RAVDESS ensure that such systems also develop emotional awareness, making interactions more human-like and effective.

Contributions to Social and Behavioral Sciences

In social and behavioral sciences, RAVDESS offers a valuable foundation for examining how emotions are expressed, perceived, and influence interpersonal interactions.

Key applications include:

- Studying emotional communication in in-group vs. out-group settings or hierarchical relationships.
- Exploring how emotion shapes interpersonal dynamics, such as trust, empathy, or dominance.
- Investigating the role of emotional cues in moral judgment, decision-making, and leadership behavior.

For example, researchers can analyze how a speaker's emotional tone affects the choices of listeners during negotiations or persuasive communication. The controlled stimuli in RAVDESS make it easier to isolate the causal effects of specific emotions, enhancing the precision of experimental designs.

Further directions

While RAVDESS is already a valuable resource for emotion research, there are several promising directions for its future development. One key area is expanding its cultural representation, as the current dataset features only North American-accented English. Including voices from diverse linguistic and cultural backgrounds—especially Indian languages like Malayalam—would enhance its global applicability and support cross-cultural studies on emotional expression. This would help researchers understand how cultural norms shape emotional communication and influence recognition accuracy. Another important direction is the integration with advanced technologies such as deep learning and real-time emotion tracking systems. By incorporating richer annotations like facial landmark tracking, pitch contours, and arousal levels, RAVDESS could serve as a more robust training and testing tool for emotion recognition models. Furthermore, the database could support real-world applications in fields like healthcare, education, and entertainment. For example, it could help develop emotionally responsive systems for clinical

settings, tools for teaching empathy and emotional intelligence, or emotionally rich digital characters in games and virtual reality. These enhancements would position RAVDESS at the forefront of affective computing and emotional AI research.

Conclusion

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) represents a significant advancement in the field of emotion research. Its comprehensive and multimodal nature that combined with rigorous validation and broad applicability makes it a precious resource for researchers across various disciplines. RAVDESS not only provides a robust foundation for studying emotional expression and perception but also offers a wealth of opportunities for future research and technological innovation.

Current AI systems lack robustness because they fail to model causal relationships, limiting their ability to generalize beyond training data [9]. As the field of emotion research continues to evolve, RAVDESS will surely play a critical role in shaping our understanding of emotions and their impact on human behavior. By expanding its cultural representation, integrating with advanced technologies and by developing real-world applications, RAVDESS has the potential to become an even more powerful tool for advancing emotion research and improving emotional intelligence in the world.

References

1. Barrett, L. F. (2017). *The theory of constructed emotion: An active inference account of interoception and categorization.Social Cognitive and Affective Neuroscience.*
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?*
3. Cohn, J. F., & De la Torre, F. (2015). *Automated face analysis for affective computing.*
4. Kipp, M., et al. (2021). *The MULTISIMO multimodal corpus: A resource for studying dialogue behavior.*
5. Koch, C., & Buice, M. A. (2022). *The single neuron as a decision-maker.Neuron.*
6. Lin, H., & Li, S. (2020). *A multimodal emotion recognition model based on feature fusion and decision fusion.*
7. Liu, Z., Zhang, H., Pan, Z., & Xue, J. (2019). *A multimodal emotion recognition method using speech and facial expression.*
8. Livingstone, S. R., & Russo, F. A. (2018). *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.*
9. Marcus, G. (2020). *The next decade in AI: Four steps toward robust artificial intelligence.*

10. Pell, M. D., et al. (2023). *The role of prosody in emotional speech recognition.*
11. Priyasamy, R., & Ramaswamy, S. P. A. (2024). *Multimodal emotion recognition: Trends and challenges.WIREs Data Mining and Knowledge Discovery, 14*, e1563.ResearchGate+1MDPI+1
12. Schuller, B. W. (2021). *Speech emotion recognition: Two decades in a nutshell.*
13. Stevens, R. L., & Russo, F. A. (2018). *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)…*
14. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). *End-to-end multimodal emotion recognition using deep neural networks.IEEE Journal of Selected Topics in Signal Processing, 11*, 1301–1312.arXiv+2MDPI+2Wikipedia+2
15. Vinyals, O., et al. (2019). *Grandmaster level in StarCraft II using multi-agent reinforcement learning.Nature.*
16. Wu, Y., Mi, Q., & Gao, T. (2025). *A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions.Biomimetics, 10*(7), 418.Nature+6MDPI+6Wiley Online Library+6
17. Wu, Y., Zhang, S., & Li, P. (2025). *Multi-modal emotion recognition in conversation based on prompt learning with text–audio fusion features.Scientific Reports, 15*, 8855.Nature+2Nature+2MDPI+2
18. Wu, Y., et al. (2025). *Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects* [Preprint]. *arXiv.*arXiv+1MDPI+1
19. Survey on multimodal music emotion recognition. (2025). *arXiv:2504.18799.*arXiv+1arXiv+1
20. Khan, M., Tran, P.-N., Pham, N. T., El Saddik, A., & Othmani, A. (2025). *MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion.Scientific Reports, 15*, 5473.