Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 3 | Mar-2026
ISSN (O) 3107-6696

A Multimodal AI-Based Interview Assessment System Using Facial Emotion Recognition and Speech Confidence Analysis

[1] Prof. A. R. Bhuyar, [2]Aniruddha Saraf, [3]Charwak Bhonde, [4]Rohit Hage, [5]Shatayu Balapure, [6]Vedant Khairkar

[1,2,3,4,5,6]*Department of Information Technology SIPNA COET, Amravati*

Abstract – Interview performance depends not only on domain knowledge but also on non-verbal communication, emotional control, and vocal confidence—factors often overlooked by con- ventional mock interview platforms that primarily assess textual correctness. This paper presents an intelligent AI-powered mock interview system designed for holistic evaluation of both tech- nical knowledge and soft skills. The proposed system adopts a multimodal framework that integrates Natural Language Pro- cessing (NLP) for semantic answer evaluation, facial emotion recognition for analyzing non-verbal cues, and speech analysis for assessing vocal confidence. Facial emotions are predicted using a Convolutional Neural Network (CNN) trained on the FER-2013 dataset, while speech confidence is evaluated through acoustic feature extraction and classification. A hybrid semantic similarity and keyword-matching approach is used to assess answer relevance. By jointly analyzing these modalities, the system simulates a realistic interview environment and generates a comprehensive performance report. Experimental evaluation shows that the emotion recognition model achieved 53.98The proposed approach enables automated, objective, and holistic interview feedback, making it a practical tool for candidate preparation and behavioral improvement. This work contributes toward the development of intelligent interview training systems that bridge the gap between technical evaluation and soft-skill assessment..

Index Terms—AI mock interview, multimodal assessment, NLP, facial emotion recognition, speech confidence, interview evaluation.

I. INTRODUCTION

The interview process plays a crucial role in academic admissions, job recruitment, and professional evaluations. Be- yond assessing technical competence, interviews are designed to evaluate behavioral traits such as confidence, emotional control, and communication skills. Research in communication psychology shows that interviewer perceptions are strongly influenced by non-verbal cues, including facial expressions and vocal delivery [1], [7]. With the rapid adoption of virtual interviews, evaluation has increasingly shifted to online plat- forms; however, assessment methods remain largely subjective and dependent on human judgment.

Most existing mock interview platforms primarily evaluate the correctness of textual answers. While this is important for measuring knowledge, it does not fully reflect real interview conditions. In practice, candidates are judged not only by what they say but also by how they say it and how they present themselves emotionally. Nervousness, hesitation, and low vocal confidence can negatively influence interviewer perception even when answers are technically correct [1]. Consequently, systems that rely solely on textual scoring fail to capture key behavioral dimensions of interview performance. Recent advancements in Artificial Intelligence (AI) have opened new possibilities for automated and objective interview evaluation. Progress in computer vision enables facial emotion recognition systems capable of identifying stress, fear, and confidence through visual cues [4]. Similarly, speech process- ing techniques can analyze vocal attributes such as fluency, pauses, and pitch variation to estimate confidence levels [7]. Natural Language Processing (NLP) models further allow semantic evaluation of responses beyond simple keyword matching, improving the accuracy of answer assessment [6].

To address the limitations of existing tools, this research proposes a multimodal AI-based mock interview evaluation system that integrates facial emotion recognition, speech con- fidence analysis, and NLP-based answer evaluation within a unified framework. By jointly analyzing these modalities, the system provides a comprehensive and objective performance assessment that more closely resembles real interview eval- uation criteria. This multimodal approach aims to support candidates in improving both technical accuracy and soft-skill presentation, thereby enhancing overall interview prepared- ness.

II. LITERATURE REVIEW

The development of AI-based interview assessment systems has gained attention with the growth of virtual hiring plat- forms. Early systems focused on conversational agents and automated feedback for mock interviews, improving candidate preparedness but relying mostly on text-based evaluation [2], [8]. However, real-world interviews require assessment beyond textual correctness, including behavioral and communication cues.

A. Facial Emotion Recognition

Facial Emotion Recognition (FER) has been widely stud- ied using datasets such as FER-2013. CNN-based models effectively learn facial features like eyebrow movement and lip curvature [4]. Despite progress, FER remains challenging due to low-resolution inputs and subtle emotional variations, with typical accuracies between 50–70% on FER-2013 [4]. Most FER research emphasizes static emotion classification rather than behavioral trends over time, which are important in interviews.

B. Speech-Based Confidence Analysis

Speech analysis uses acoustic features such as pitch, MFCCs, and energy to detect emotional and psychological states. Vocal tone and fluency strongly influence perceived confidence [7]. Feature extraction tools like openSMILE sup- port large-scale speech analysis [9]. Machine learning models including SVM and Random Forest are commonly used, but many systems focus on emotion detection rather than interview-specific confidence assessment.

C. Automated Answer Evaluation

Automated grading has evolved from keyword matching to semantic evaluation. Keyword-based systems often fail with paraphrased responses [5]. Transformer-based models such as Sentence-BERT enable semantic similarity scoring aligned with human grading [6]. However, most answer evaluation systems operate independently of behavioral or vocal analysis.

D. Research Gap

Existing studies highlight several limitations:

- Most systems analyze only a single modality
- Behavioral and vocal cues are often overlooked
- Answer grading systems lack full semantic flexibility
- Few systems target holistic interview evaluation

To address these gaps, recent work emphasizes multimodal frameworks combining visual, speech, and textual cues [10]. The proposed system follows this direction by integrating facial emotion recognition, speech confidence analysis, and semantic answer evaluation into a unified mock interview platform, enabling more realistic and objective assessment [1], [7].

III. Proposed Methodology

A. System Architecture

The FER module continuously monitors a candidate's emo- tional state during the interview. Emotional expressions such as happiness, fear, nervousness, and neutrality serve as behav- ioral indicators reflecting confidence and psychological stabil- ity [1], [4]. Continuous emotion tracking enables behavioral evaluation beyond answer correctness and supports assessment of non-verbal communication patterns that influence inter- viewer perception [1].

The architecture consists of three primary components:

Frontend Interface

- Periodically captures webcam images at 3-second inter- vals
- Records candidate speech responses
- Transmits synchronized text, audio, and image data to backend modules

Backend AI Engine

- Facial Emotion Recognition (FER) Model [4]
- Speech Confidence Analysis Model [7], [9]
- NLP-based Answer Evaluation Model [5], [6]

Report Generation Module

- Aggregates outputs from all AI modules
- Computes composite performance metrics
- Produces a structured interview assessment report

This design enables simultaneous processing of heteroge- neous inputs, allowing holistic evaluation of both technical knowledge and behavioral traits [10].

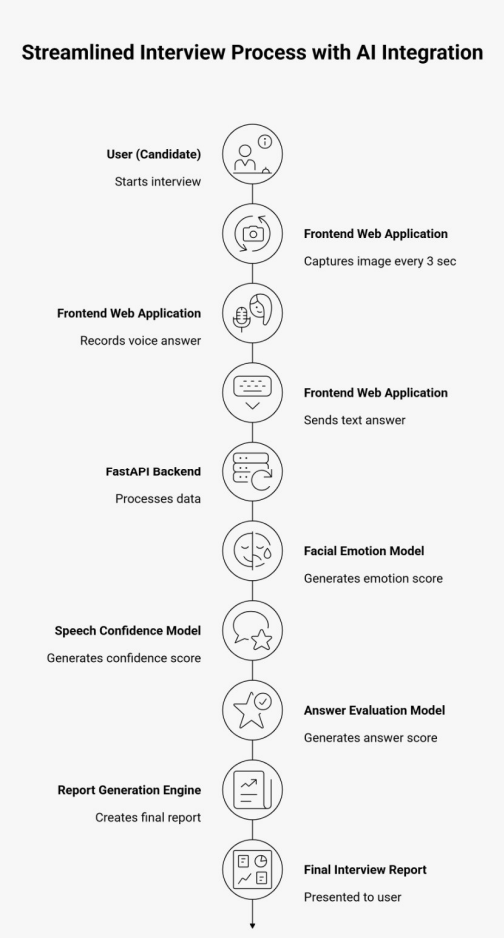

Fig. 1. Overall multimodal interview system architecture

A. System Workflow

The interview process follows a structured pipeline inspired by modern AI-based interview and assessment systems [8], [10]:

1) Activation of webcam and microphone
2) Periodic facial image capture
3) Question presentation to the candidate
4) Verbal response recording
5) Speech confidence analysis [7]
6) NLP-based answer evaluation [5], [6]
7) Iteration until interview completion

At the end of the session, a comprehensive performance re- port is generated to provide objective and data-driven feedback for candidate improvement [2].

B. Facial Emotion Recognition Model

1) Objective: The FER module continuously monitors a candidate's emotional state during the interview. Emotional expressions such as happiness, fear, nervousness, and neutrality serve as behavioral indicators reflecting confidence and psychological stability [1], [7]. Continuous emotion tracking enables behavioral evaluation beyond answer correctness and supports analysis of non-verbal communication cues that in- fluence interviewer perception [1].

2) Dataset Description: The model is trained on the FER- 2013 dataset, a benchmark dataset widely used in emotion recognition research [3], [4].

- 35,000+ grayscale facial images
- Original resolution: 48×48 pixels
- Seven emotion classes: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral

For improved feature learning, images are rescaled to 64×64 pixels before being input to the CNN, a common practice for enhancing spatial feature extraction in deep learning models [4].

Fig. 2. Sample images from the FER-2013 dataset showing variations in facial expressions

3) Preprocessing Pipeline:

1) Face detection using Haar Cascade classifiers
2) Grayscale conversion
3) Resizing to fixed dimensions (64×64)
4) Pixel normalization to [0,1]

These steps ensure consistent and noise-reduced inputs for model training.

4) CNN Architecture: A Convolutional Neural Network is employed due to its strong spatial feature learning capability [4]. The model consists of multiple layers that progressively learn hierarchical facial features from low-level edges to high- level emotional patterns, which is a standard approach in deep representation learning [3], [4]. Dropout and pooling layers help reduce overfitting and improve generalization in deep neural networks [3].

TABLE I

CNN ARCHITECTURE OVERVIEW

Layer	**Purpose**
Convolution	Extracts facial features
ReLU	Introduces non-linearity
Max Pooling	Reduces dimensionality
Fully Connected	Emotion classification
Softmax	Probability output

5) Feature Learning: The network learns discriminative cues such as facial muscle movements and expression patterns that correlate with emotional states [4]. Examples include:

- Eyebrow position → fear/anger
- Lip curvature → happiness
- Eye openness → surprise
- Facial tensio

6) Output Mapping:

Confusion Matrix-

The confusion matrix shows class-wise performance of the FER model across seven emotions. Major confusions are observed between *fear*, *sad*, and *neutral*, while *happy* has the highest number of correct predictions.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 3 | Mar-2026
ISSN (O) 3107-6696

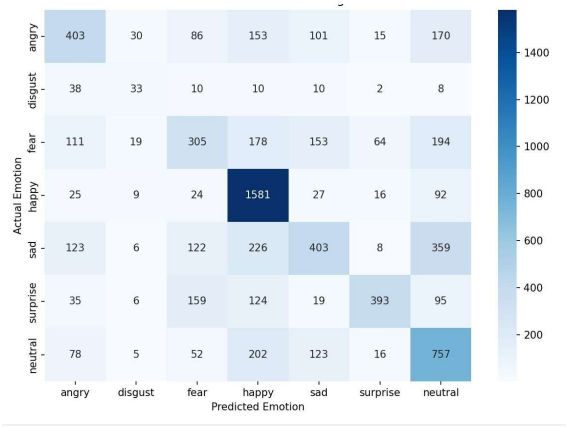

Fig. 3. Confusion matrix of the FER model

TABLE II

EMOTION-TO-INTERVIEW INTERPRETATION MAPPING

Emotion	Interview Interpretation
Happy	Confident
Neutral	Controlled
Fear	Nervous
Sad	Low confidence
Angry/Disgust/Surprise	Context-dependent

Emotion-wise Accuracy: Emotion-wise accuracy high- lights that *happy* has the strongest recognition rate. Lower accuracies in *fear* and *disgust* indicate difficulty in dis- tinguishing similar negative emotions.

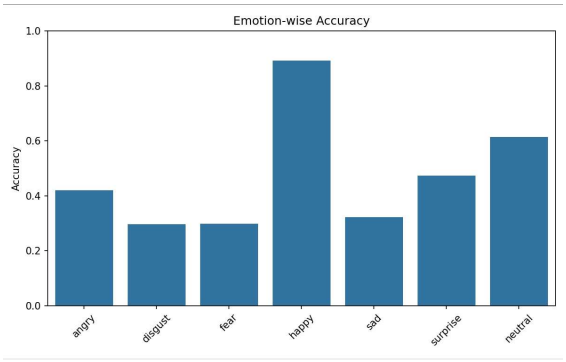

Fig. 4. Emotion-wise classification accuracy

7) *Performance:* Although moderate, this aligns with known FER challenges such as subtle expressions and illumination variation. Temporal emotion trends provide more reliable insights than single-frame predictions.

Metric	Value
Test Accuracy	53.98%

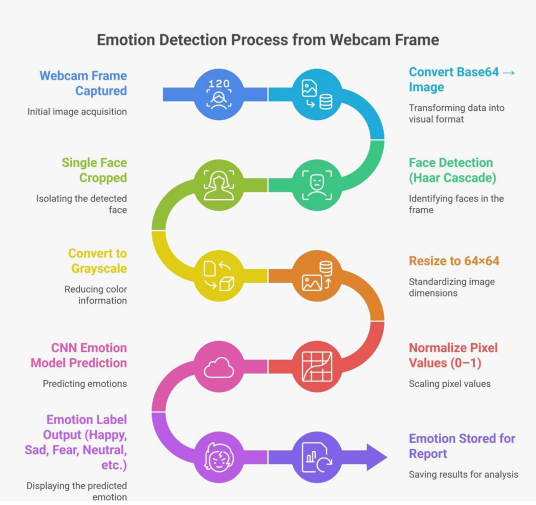

Fig. 5. Facial emotion detection pipeline

D. NLP-Based Answer Evaluation Model

1) Objective: The NLP module evaluates semantic correctness rather than relying solely on keyword matching. This allows accurate grading even when candidates para- phrase or use varied sentence structures, which is a known limitation of traditional keyword-based grading systems [5]. Semantic similarity methods improve agreement with human grading by capturing contextual meaning [6].

2) Core Technology: The system uses the SentenceTrans- formers framework with the *all-MiniLM-L6-v2* model, which is based on Sentence-BERT architecture designed for semantic similarity tasks [6].

- 384-dimensional embeddings
- Context-aware semantic representation [6]
- Optimized for similarity tasks [6]
- Lightweight and fast

3) Evaluation Pipeline: The evaluation pipeline follows established automated short-answer grading approaches [5]:

1) Phrase segmentation
2) Sentence embedding generation [6]
3) Cosine similarity matching
4) Keyword verification

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 3 | Mar-2026
ISSN (O) 3107-6696

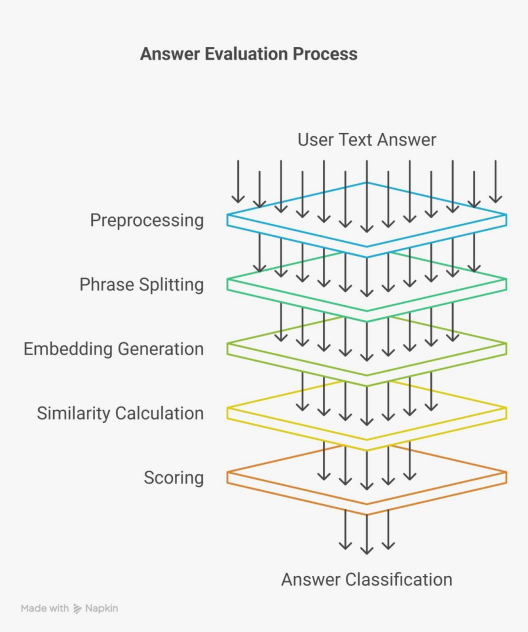

Fig. 6. Semantic answer evaluation pipeline

4) Similarity Computation:

$$Similarity = \frac{A \cdot B}{||A||\,||B||}$$

A threshold of 0.45 is used for concept matching.

5) Hybrid Scoring:

$$Final\ Score = (Semantic \times 0.7) + (Keyword \times 0.3)$$

TABLE III

HYBRID SCORING WEIGHTS

Component	Weight
Semantic Similarity	70%
Keyword Matching	30%

1) Classification Criteria: Agreement with human grading reached approximately 78%.

TABLE IV

ANSWER EVALUATION CRITERIA

Score Range	Evaluation
≥ 70%	Correct
40–69%	Partially Correct
< 40%	Incorrect

A. Speech Confidence Analysis Model

1) Objective: The primary objective of the Speech Con- fidence Analysis Model is to autonomously evaluate the paralinguistic attributes of a candidate's speech. Unlike the Natural Language Processing (NLP) module, which evaluates what is said, this module evaluates how it is said. The model quantifies psychological traits such as confidence, hesitation, and stress by analyzing acoustic biomarkers extracted from speech signals. Vocal commu- nication has been shown to strongly reflect emotional and psychological states [7].

2) Theoretical Framework and Motivation: In real inter- view scenarios, interviewers subconsciously assess vocal cues such as tone stability, speaking pace, and vocal en- ergy. Confident speakers tend to maintain steady pacing, dynamic intonation, and consistent loudness, while ner- vous speakers exhibit frequent pauses, monotone pitch, or irregular energy patterns. Research in vocal emotion communication confirms that pitch, intensity, and speech dynamics are key indicators of emotional and confidence states [7].

Soft skills and communication ability significantly in- fluence professional evaluation outcomes [2]. However, most existing automated interview systems focus solely on textual correctness and ignore these vocal indicators. This module addresses this gap by introducing speech- based confidence estimation to simulate human interview perception. Acoustic feature extraction frameworks such as openSMILE further support reliable analysis of speech characteristics [9].

3) *System Architecture:* The Speech Confidence Analysis Model follows an Input–Process–Output pipeline com- monly adopted in speech analysis systems [7].

Input

– Raw audio captured via microphone during interview responses

Processing

– Audio preprocessing (normalization, silence trim- ming)
– Acoustic feature extraction [9]
– Feature vector formation
– Machine learning–based classification

Output

– Numerical confidence score (0–100)
– Confidence category (High, Medium, Low)

4) Technical Implementation: **Technology Stack**

– **FastAPI**: Backend framework for handling audio uploads and API routing
– **Librosa**: Audio signal processing and feature extrac- tion

– **FFmpeg**: Audio format conversion to WAV
– **Scikit-Learn**: Random Forest classifier
– **NumPy**: Numerical computations

5) Dataset Description: The model is trained on a custom-curated dataset of recorded interview responses labeled by human evaluators. Human perception of con- fidence has been shown to correlate with vocal charac- teristics [7].

Class 0 – Nervous

– Frequent pauses
– Stuttering
– Low or unstable energy
– Flat pitch

Class 1 – Confident

– Clear articulation
– Steady speaking rate
– Consistent energy
– Dynamic pitch modulation

The dataset is split into training and testing sets to validate generalization.

6) Acoustic Feature Extraction: Acoustic feature extrac- tion is a well-established method for analyzing vocal emotion and confidence [7], [9]. The following features are extracted:

– MFCCs – capture vocal texture [9]
– Fundamental Frequency – tone representation [7]
– RMS Energy – loudness measure [7]
– Speaking Rate – tempo estimation [7]
– Silence Ratio – hesitation measurement [7]

7) Mathematical Formulation: **Speaking Rate**

$$SR = \frac{N_{syllables}}{T_{speech}}$$

Pitch Variance

$$\sigma_{f0} = \sqrt{\frac{1}{N}\sum (f0_i - \mu_{f0})^2}$$

RMS Energy

$$E_{rms} = \sqrt{\frac{1}{N}\sum x_i^2}$$

Silence Ratio

$$R_{silence} = \frac{T_{silence}}{T_{total}}$$

These acoustic metrics are widely used indicators of vocal confidence and emotion [7].

8) Machine Learning Classification: Extracted features are combined into a feature vector and fed into a Random Forest classifier. Random Forest models are effective for non-linear behavioral data modeling.

Reasons for using Random Forest

– Handles non-linear relationships
– Resistant to overfitting
– Provides feature importance analysis

9) Hybrid Scoring Method:

$$S = \alpha P_{ml} + \beta S_{pace} + \gamma S_{pitch}$$

$$\alpha + \beta + \gamma = 1$$

Hybrid scoring improves robustness by combining mul- tiple acoustic indicators [7].

TABLE V

SPEECH CONFIDENCE LEVELS

Score Range	Confidence Level
70–100	High
40–69	Medium
0–39	Low

10) Score Interpretation:

11) Experimental Results:

a. Classification Accuracy: 83.33%
b. Human Agreement: ∼72%

These results align with studies showing that vocal cues strongly influence perceived confidence [7].

12) *System Integration:* The Speech Confidence Score is combined with:

a. Facial Emotion Score
b. NLP Answer Evaluation Score

Multimodal fusion has been shown to improve behavioral analysis reliability [10]. Together, these metrics provide a holistic interview assessment. Soft-skill evaluation is increasingly recognized as important in professional set- tings [2].

IV. CONCLUSION

This paper presented a multimodal AI-based mock in- terview system for comprehensive evaluation of both technical knowledge and soft skills. Unlike conventional interview preparation tools that primarily assess answer correctness, the proposed framework integrates semantic answer evaluation,

facial emotion recognition, and speech confidence analysis to deliver a holistic assessment. By capturing both verbal and non-verbal cues, the system more accurately reflects real-world interview evaluation practices.

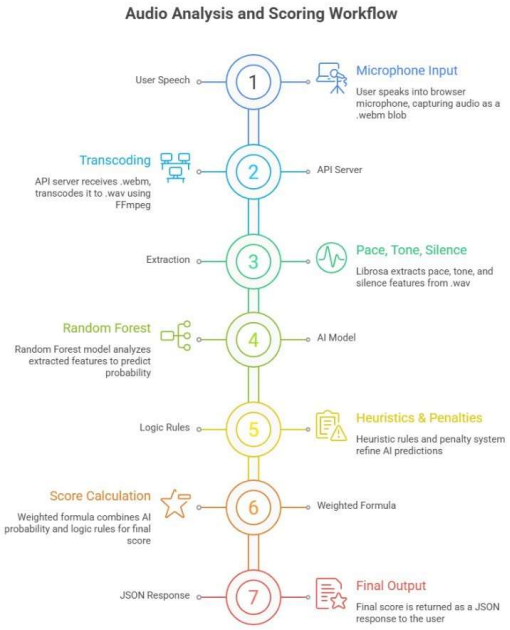

Fig. 7. Speech confidence analysis pipeline

Experimental results demonstrate that the system can effectively analyze candidate performance and provide meaningful, data-driven feedback. Participants were able to identify strengths and areas for improvement, par- ticularly in confidence, communication, and emotional stability. These findings confirm that multimodal AI can significantly enhance interview preparation by enabling realistic and objective performance assessment.

Overall, the proposed platform contributes toward intel- ligent interview training solutions that bridge the gap between knowledge-based evaluation and behavioral as- sessment. The study also highlights the broader potential of multimodal AI in recruitment and educational tech- nologies. Future research may focus on personalization, bias mitigation, and adaptive feedback mechanisms to further improve system reliability and user experience.

REFERENCES

[1] A. Mehrabian, *Nonverbal Communication*. Aldine-Atherton, 1972.

[2] J. J. Heckman and T. Kautz, "Hard evidence on soft skills,"*Labour Economics*, vol. 19, no. 4, pp. 451–464, 2012.

[3] I. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. ICONIP*, 2013.

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, no. 12, pp. 1–1, 2020.

[5]S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *International Journal of Artificial Intelligence in Education*, vol. 25, pp. 60–117, 2015.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embed- dings using Siamese BERT-networks," in *Proc. EMNLP*, 2019.

[7] F. Eyben, M. Wo¨llmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.

[8] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI*, 2020.

[9] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1–2, pp. 227–256, 2003.

[10] J. S. Black and P. van Esch, "AI-enabled recruiting: What is it and how should a manager use it?," *Business Horizons*, vol. 63, no. 2, pp. 215–226, 2020.