International Journal of General Engineering and Technology (IJGET)
ISSN (P): 2278–9928; ISSN (E): 2278–9936
Vol. 15, Issue 1, Jan – Jun 2026; 39–48

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

PROPOSED MULTIMODAL APPROACH FOR EMOTION AND BEHAVIOUR ANALYSIS IN INTERVIEW DIALOGUES

Dr. Shalini Gupta, Sana Khan, Priyanshi Yadav, Shruti Kumari & Sudeep Kaushal

Department of Computer Science and Engineering, Axis Institute of Technology and Management, Kanpur, Kanpur, Uttar Pradesh, India

ABSTRACT

This paper introduces a smart system that automatically analyses interview performance using different types of data. It improves on traditional methods, which often rely on subjective or general feedback, by providing objective and data-based insights into how a candidate communicates and behaves.

The system looks at multiple aspects of an interview. It uses video analysis to study facial behaviour such as eye contact and expressions, audio analysis to understand speech patterns like pitch and energy, and text analysis to examine what the candidate says, including the use of filler words. All these insights are combined using a scoring method to give a clear and understandable performance evaluation.

Overall, this approach helps in gaining deeper insights into interview performance and can be used for scalable applications such as interview coaching, education, and recruitment.

KEYWORDS: *Speech Emotion Recognition (SER), Natural Language Processing (NLP), Multimodal Analysis, Interview Performance Evaluation, Affective Computing, Emotion Shift Modelling, Prosodic Feature Extraction, Temporal Behaviour Analysis, Computer Vision, Speech Processing*

Article History

Received: *24 Apr 2026* | ***Revised:*** *25 Apr 2026* | ***Accepted:*** *27 Apr 2026*

INTRODUCTION

As modern communication systems become increasingly human-centric, evaluating only what a person says is no longer sufficient. In real-world scenarios such as interviews, communication effectiveness depends equally on how it is delivered. Non-verbal cues—including facial expressions, eye contact, vocal tone, and speech fluency—play a crucial role in conveying confidence and competence. However, most existing interview evaluation systems focus primarily on textual responses, neglecting these behavioural aspects, which often leads to incomplete assessment of candidate performance.

This limitation is significant, as many candidates fail interviews not due to lack of knowledge, but because of nervousness, hesitation, or poor delivery. To address this gap, this work presents **Face2Fate: an Emotion-Aware Interview Coach**, a multimodal system designed to evaluate interview performance by integrating video, audio, and text analysis. The system captures facial expressions, gestures, and eye contact through video processing; analyses pitch, energy, and speech patterns through audio signals; and evaluates fluency and filler word usage through text transcription. These features are combined to generate a **confidence score along with actionable feedback**, enabling a comprehensive

and practical assessment of candidate behaviour.

Emotion recognition forms a core component of Human–Computer Interaction (HCI), where systems aim to infer human emotions from behavioural signals. Among various modalities, Speech Emotion Recognition (SER) has been widely studied due to its ability to capture rich prosodic and acoustic features. With the advancement of deep learning, significant progress has been made in improving emotion recognition accuracy. Early work by Chang-hyun Park et al. [8] utilized recurrent neural networks (RNNs) for SER. Subsequently, Jianwei Niu et al. [16] applied deep neural networks (DNNs), while Qirong Mao et al. [6] employed convolutional neural networks (CNNs) to extract invariant emotional features. Lee and Tashev [17] further improved temporal modelling using long short-term memory (LSTM) networks.

More recent advancements include capsule neural networks proposed by MA Jalal et al. [14] and highway neural networks explored by R. Shankar et al. [15], which enhanced feature learning and model efficiency. Furthermore, transformer-based architectures introduced by Shamane Siriwardhana et al. [12] enabled effective **multimodal emotion recognition**, combining speech, visual, and textual information for improved performance. These developments demonstrate the growing capability of deep learning models to capture complex emotional patterns across modalities.

Despite these advancements, most existing systems focus on **static emotion classification**, where a single emotion label is assigned to an entire input. This approach is insufficient for applications like interview analysis, where **behavioural dynamics**—such as transitions from confidence to hesitation—provide deeper insights into performance.

In contrast, the proposed system focuses on **applied multimodal behavioural analysis**, integrating visual, acoustic, and linguistic cues to evaluate not only emotional states but also communication quality and confidence. By combining insights from multiple modalities into a unified framework, Face2Fate aims to simulate a realistic interview evaluation environment and provide meaningful, data-driven feedback.

The proposed approach is scalable and can be extended to real-time analysis, personalized coaching systems, and other domains requiring emotion-aware human behaviour understanding.

LITERATURE REVIEW

Research in Speech Emotion Recognition (SER) and interview-based affect analysis has evolved through three primary streams: audio-only machine learning approaches, multimodal frameworks, and dialogue- or NLP-based emotion modelling. Each approach offers valuable insights into human affect; however, they face limitations in capturing dynamic emotional transitions and behavioural cues essential for real-world scenarios such as interview evaluation systems like Face2Fate.

- **Audio-Only Machine Learning Approaches:** Early studies primarily relied on acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs), prosody, and spectral energy to classify categorical emotions. Liu et al. [5] introduced a brain-inspired multi-task deep learning model that mimics human perception mechanisms, showing improved accuracy on the IEMOCAP dataset but limited generalization to spontaneous speech. Similarly, Jain et al. [3] employed Support Vector Machines (SVM) with MFCC features to achieve high classification accuracy across emotions like happiness and anger, though the handcrafted nature of features reduced robustness across diverse speakers. Shaila et al. [11] compared classical and deep learning models (RF, SVM, CNN, MLP) on RAVDESS, finding CNNs superior in accuracy but restricted by acted data that lacked conversational context.

- **Multimodal and Behavioural Modelling Approaches:** To overcome the limitations of audio-only methods, several studies incorporated additional modalities like text, facial expression, and gesture to provide a more holistic affective understanding. Naim et al. [7] proposed a multimodal regression framework combining prosodic, lexical, and facial features to predict interview performance, achieving strong correlations with human ratings. However, it focused on overall performance prediction rather than emotional evolution across dialogue turns. Burdisso et al. [2] analysed depression detection systems trained on DAIC-WOZ interviews, revealing how such models can unintentionally exploit dataset bias—e.g., learning from therapist prompts instead of genuine emotional responses. These multimodal efforts improved prediction reliability but still lacked temporal continuity.
- **Dialogue-Based and NLP/Transformer Approaches**: A third direction in affective computing leverages dialogue modelling and natural language processing to capture emotional context. Ang et al. [1] used prosodic cues to detect frustration and annoyance in human–computer dialogues, while Ranganath et al. [10] identified nuanced social tones such as friendliness or awkwardness in speed-dating conversations through combined acoustic and lexical features. More recently, Poria et al.[9] reviewed conversational emotion recognition models, highlighting progress in context-aware modelling through dialogue-level architectures such as Memory Networks and Transformers. Yet, they noted a persistent gap in modelling continuous emotion evolution—a gap particularly relevant for interview dialogues where subtle shifts in tone can indicate confidence, anxiety, or relief.

Despite these developments, a key limitation remains: most models focus on static or context-aware classification but do not adequately capture continuous emotional transitions. This limitation is particularly critical in interview scenarios, where subtle shifts in tone, expression, and fluency directly reflect confidence, anxiety, or composure—factors explicitly targeted in the Face2Fate system.

PROPOSED FRAMEWORK

The proposed framework presents a **multimodal emotion-aware interview analysis system** that evaluates candidate performance by integrating video, audio, and text-based features. Unlike purely text-based approaches, the system captures both verbal and non-verbal cues by analysing facial expressions, vocal characteristics, and speech content. These multimodal signals are combined to generate a confidence score and behavioural feedback, enabling a comprehensive understanding of user performance during an interview.

The framework is designed to capture both instantaneous emotional states and temporal behavioural transitions, allowing deeper insights into confidence, nervousness, hesitation, and overall communication quality throughout the interaction.

Overview of the Framework

The system follows a structured pipeline consisting of input acquisition, multimodal feature extraction (video, audio, text), preprocessing, emotion and behaviour analysis, and confidence scoring.

Unlike traditional systems that rely on a single modality, this framework integrates:

- **Video analysis** for facial expressions, eye contact, and gestures
- **Audio analysis** for pitch, energy, and speech dynamics
- **Text analysis** for fluency, filler words, and linguistic patterns

The goal is to provide a holistic and realistic evaluation of interview performance, making the system scalable, practical, and suitable for real-world deployment.

Workflow of the Framework

- **Step 1:** Input Acquisition: The system captures video and audio recordings of candidate responses during an interview session. These inputs serve as the foundation for multimodal analysis.
- **Step 2:** Modality Separation: The recorded input is divided into three components:
    - **Video stream** → for visual behaviour analysis
    - **Audio stream** → for speech signal processing
    - **Text stream** → obtained via speech-to-text conversion.
- **Step 3:** Video Analysis (Non-Verbal Features): The video module extracts behavioural cues using computer vision techniques:
    - **Facial expressions** → Emotion detection using deep learning models (EfficientNet-based HSEmotion)
    - **Eye contact** → Head pose estimation and iris tracking (MediaPipe)
    - **Blink rate** → Eye Aspect Ratio (EAR)
    - **Hand movements and gestures** → Motion tracking (MediaPipe Hands)

  These features help evaluate **confidence, engagement, and body language**.
- **Step 4**: Audio Analysis (Vocal Features): The audio module processes speech signals to extract:
    - **Pitch variation** (YIN algorithm) → reflects tone and expressiveness
    - **Energy variation (RMS)** → indicates confidence and emphasis
    - **Speech duration and pauses** → capture hesitation and fluency

  These features provide insights into **vocal confidence and speaking style**.
- **Step 5**: Text Analysis (Linguistic Features): Speech is converted into text using an ASR model (Whisper). The resulting transcript undergoes:
    - **Text preprocessing** (cleaning, tokenization, normalization)
    - **Filler word detection** (e.g., *um, uh, like*)
    - **Fluency analysis** using filler ratio and sentence structure

  This module evaluates **communication clarity and verbal fluency**.
- **Step 6:** Multimodal Fusion and Scoring: Outputs from all three modules are combined using a **weighted scoring mechanism**:
    - Video features → behavioural confidence
    - Audio features → vocal confidence

- Text features → linguistic fluency

These are integrated to generate a **final confidence score**, representing overall interview performance.

- **Step 7:** Behavioural Transition Analysis: To capture performance dynamics:
  - Features are analysed over time
  - Behavioural shifts (e.g., confident → nervous, fluent → hesitant) are identified
  - A temporal profile of candidate performance is generated

This enables deeper understanding beyond static evaluation.

- **Step 8:** Output and Reporting: The system produces:
  - **Confidence score** (overall performance indicator)
  - **Module-wise feedback** (video, audio, text)
  - **Identified weaknesses** (e.g., low eye contact, monotone voice, high filler usage)
  - **Behavioural trend analysis**

These outputs can be visualized in a dashboard for user feedback and improvement.

METHODOLOGY (WITH MATHEMATICAL FORMULATION)

The proposed system follows a multimodal methodology that integrates visual, acoustic, and linguistic features to evaluate interview performance. The system extracts feature from video, audio, and text, and combines them using a scoring mechanism to generate a final confidence score.

Eye Blink Detection using Eye Aspect Ratio (EAR)

To analyse **blink rate and eye openness**, the Eye Aspect Ratio (EAR) is computed using facial landmarks detected via MediaPipe.

$$EAR = \frac{|p_2 - p_6| + |p_3 - p_5|}{2|p_1 - p_4|}$$

Where:

- p1,p2,,p6 represent eye landmark coordinates
- | | denotes Euclidean distance

A lower EAR indicates eye closure (blink), while a stable EAR indicates consistent eye contact. Frequent blinking or unstable EAR patterns may indicate **nervousness or lack of confidence**.

Audio Energy using Root Mean Square (RMS)

To evaluate **speech** intensity **and vocal confidence**, the Root Mean Square (RMS) energy of the audio signal is calculated:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$$

Where:

- x_i represents the amplitude of the audio signal
- N is the total number of samples

Higher RMS values indicate stronger and more confident speech, while low or fluctuating RMS values suggest hesitation or weak delivery.

Pitch Estimation (YIN Algorithm)

Pitch is estimated using the **YIN algorithm**, which computes the fundamental frequency of speech. Variations in pitch help identify monotony, expressiveness, and emotional tone. Stable pitch indicates controlled speech, while excessive variation may indicate nervousness.

Text-Based Fluency (Filler Ratio)

To evaluate **speech fluency**, a filler ratio is computed:

$$Filler\ Ratio = \frac{Number\ of\ Filler\ Words}{Total\ Words\ Spoken}$$

Multimodal Confidence Scoring

The final **confidence score** is computed by combining features from all three modalities using a weighted approach:

$$C = w_v V + w_a A + w_t T$$

Where:

- C = Final confidence score
- V = Video-based score (eye contact, expressions, gestures)
- A = Audio-based score (pitch, energy, pauses)
- T = Text-based score (fluency, fillers)
- w_v, w_a, w_t are weights such that:

$$w_v + w_a + w_t = 1$$

As a baseline configuration, the weights may be set as $w_v = 0.4$, $w_a = 0.3$, and $w_t = 0.3$, reflecting the relatively higher importance of visual cues in face-to-face interview contexts. These values can be adjusted based on domain-specific requirements or empirical tuning in future work. This fusion ensures a **balanced evaluation** of verbal and non-verbal communication.

DISCUSSION

The proposed **Face2Fate: Emotion-Aware Interview Coach** demonstrates the effectiveness of a multimodal approach in evaluating interview performance. By integrating video, audio, and text features, the system is able to capture both **verbal and non-verbal aspects of communication**, which are critical in real-world interview scenarios.

The inclusion of visual features such as facial expressions, eye contact, and gestures provides valuable insights into candidate confidence and engagement. At the same time, audio features such as pitch and energy help identify vocal stability and hesitation, while text-based analysis contributes to understanding fluency and linguistic clarity. The combination of these modalities enables a more **holistic evaluation** compared to traditional systems that rely solely on textual responses.

It is expected that **different modalities capture different dimensions of behaviour**. For instance, acoustic features are highly sensitive to micro-level variations such as pauses and pitch instability, which often indicate nervousness. In contrast, textual features are more effective in identifying patterns of hesitation and coherence through filler word usage and sentence structure. Visual cues further complement these by reflecting non-verbal confidence through eye contact and expressions.

The multimodal fusion strategy plays a crucial role in balancing these inputs. By assigning appropriate weights to each modality, the system ensures that no single feature dominates the evaluation, resulting in a more **robust** and **reliable confidence score**. Additionally, temporal analysis allows the system to track **behavioural changes over time**, enabling the detection of transitions such as increasing confidence or sudden drops in performance.

Despite its strengths, the system has certain limitations. The accuracy of emotion recognition models may vary depending on lighting conditions, background noise, and speech clarity. Furthermore, real-time processing of video and audio data can introduce computational overhead. The current system also relies on predefined scoring weights, which may not generalize perfectly across all users.

Overall, it is anticipated that **multimodal behavioural analysis significantly enhances interview evaluation**, making the system more aligned with real-world human judgment. This supports the idea that combining multiple modalities is expected to provide deeper insights into candidate performance than single-modality approaches.

LIMITATIONS

Human speech is never emotionally uniform. A systematic review of existing Speech Emotion Recognition (SER) and conversational modelling studies reveals several limitations that hinder the development of natural, context-aware emotion analysis systems. These challenges are discussed below.

- **The Dataset Dilemma: Acted vs. Authentic Emotions:** A substantial proportion of existing SER models rely on acted or semi-scripted datasets such as RAVDESS, IEMOCAP, and EMO-DB. While these corpora facilitate controlled experimentation, they tend to exaggerate emotional intensity and fail to capture micro-level cues—such as hesitation, sighs, or tension release—that are prevalent in real interviews. Few studies, such as Naim et al. (2015) [7], utilized spontaneous interview recordings, but their objective was limited to performance prediction rather than analysing temporal emotional evolution. This highlights the pressing need for authentic, context-rich emotional corpora that mirror natural human interactions.
- **Static vs. Dynamic Emotion Modelling and Shift Tracking:** Most prior works treat emotion as a static categorical label, classifying each utterance independently (e.g., happy, sad, angry, neutral). However, real-world interactions involve gradual emotional transitions—for instance, calmness evolving into tension or relief. Recent advances such as TIM-Net (2022), which models multi-scale temporal dependencies, and EmoTrans(2024), which predicts conversational emotion transitions, represent important steps toward dynamic emotion modelling. Yet,

these models remain constrained to predefined datasets and often overlook subtle tone variations that emerge in unstructured, real-life conversations.

- **Hearing Without Understanding: Feature and Contextual Limitations:** Traditional SER systems predominantly rely on low-level acoustic descriptors (e.g., MFCCs, prosodic statistics) or unimodal deep audio networks. While these features are effective for broad emotional categories, they fail to incorporate contextual elements such as question type, topic shift, or speaker's previous emotional state. Even advanced deep models—such as those proposed by Tzirakis et al. (2018) [13] and Liu et al. (2023) [4]—focus on spectral and temporal representations without integrating dialogue context, thereby limiting their ability to detect nuanced emotional tone shifts.

From the perspective of the proposed Face2Fate system, these limitations highlight key challenges in real-world deployment. The reliance on pre-existing datasets may not fully represent natural interview behaviour, and the absence of complete contextual understanding can affect the accuracy of confidence estimation. Additionally, real-time multimodal processing introduces computational complexity. Addressing these challenges remains an important direction for future improvements of the system.

CONCLUSION

This work presented **Face2Fate**, a multimodal emotion-aware interview analysis system that evaluates candidate performance by integrating video, audio, and text-based features. By capturing both verbal and non-verbal cues such as facial expressions, vocal characteristics, and speech fluency, the system provides a more comprehensive and realistic assessment compared to traditional methods.

The proposed approach demonstrates that multimodal analysis improves the understanding of confidence, hesitation, and communication quality**,** while temporal evaluation enables detection of behavioural transitions during responses.

Despite limitations related to dataset constraints and computational complexity, the system establishes a strong foundation for intelligent interview coaching applications. Future work can focus on real-time feedback and personalized evaluation to further enhance system effectiveness.

REFERENCES

1. *Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A., "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," Proc. ICSLP, 2002.*
2. *Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M., "A text classification framework for simple and effective early depression detection over social media streams," Expert Systems with Applications, 2020.*
3. *Jain, R., et al., "Speech emotion recognition using support vector machine," International Journal of Computer Applications, 2018.*
4. *Liu, Z., et al., "Speech emotion recognition using deep learning: A review," IEEE Access, 2023.*
5. *Liu, P., et al., "A brain-inspired multi-task learning model for speech emotion recognition," IEEE Transactions, 2020.*

6. *Mao, Q., Dong, M., Huang, Z., & Zhan, Y., "Learning salient features for speech emotion recognition using convolutional neural networks," IEEE Transactions on Multimedia, 2014.*
7. *Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E., "Automated analysis and prediction of job interview performance," IEEE Transactions on Affective Computing, 2015.*
8. *Park, C. H., Lee, D. W., & Sim, K. B., "Emotion recognition of speech based on RNN," Proc. ICMLC, 2002.*
9. *Poria, S., Cambria, E., Bajpai, R., & Hussain, A., "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, 2017.*
10. *Ranganath, R., et al., "It's not you, it's me: Detecting awkwardness in social interactions," Proc. EMNLP, 2013.*
11. *Shaila, S., et al., "Speech emotion recognition using machine learning and deep learning techniques," 2021.*
12. *Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., & Nanayakkara, S., "Multimodal emotion recognition with transformer-based self-supervised feature fusion," IEEE Access, 2020.*
13. *Tzirakis, P., et al., "End-to-end multimodal emotion recognition using deep neural networks," IEEE Journal, 2018.*
14. *Jalal, M. A., Loweimi, E., Moore, R. K., & Hain, T., "Learning temporal clusters using capsule routing for speech emotion recognition," Proc. Interspeech, 2019.*
15. *Shankar, R., et al., "Automated emotion morphing in speech using highway networks," Proc. Interspeech, 2019.*
16. *Niu, J., Qian, Y., & Yu, K., "Acoustic emotion recognition using deep neural networks," IEEE, 2014.*
17. *Lee, J., & Tashev, I., "High-level feature representation using recurrent neural networks for speech emotion recognition," Interspeech, 2015.*