AOI: 10.100.234513.0224
13th international conference on Microelectronics Circuits and Systems, Micro2026

Multimodal Affective Computing: A Dual-Stream Architecture for Real Time Emotion Verification

Aryan Choudhary, Bhalala Tanay, Bhawna Tenani, Ruchika Somani, Shreyans Sinha, Yashpal Singh`, Dr. Bhupesh Kumar Singh

Department of Computer Science & Engineering
Amity University Rajasthan - 303002, INDIA
Corresponding author: drbhupeshkumarsingh@gmail.com

ABSTRACT

Human communication is inherently multimodal, relying on a complex interplay between facial expressions and vocal prosody. However, traditional Affective Computing systems often rely on unimodal analysis—typically visual—which renders them susceptible to error when subjects mask their true emotions (e.g., a "social smile" concealing anxiety). To address this limitation, this paper proposes a real-time Multimodal Emotion Perception System that integrates visual and acoustic cues using a sequential asynchronous pipeline. The architecture leverages DeepFace (VGG-Face) for facial feature extraction and a fine-tuned Wav2Vec2 Transformer for speech sentiment analysis. The Multimodal Congruence Index (MCI) is a new decision level fusion algorithm that is used to measure the semantic agreement between modalities. The results of the experiment prove that unimodal accuracy varies in noisy conditions, whereas the proposed fusion model manages to detect emotional conflict with accuracy 88% correct, which is a strong solution to behavioral analysis and detecting lies in the interview context.

Keywords: Multimodal Fusion, Affective Computing, DeepFace, Wav2Vec2, Congruence Index, Human-Computer Interaction.

I. INTRODUCTION

The capability of a machine to interpret human affect—Automated Emotion Recognition (AER)—has transitioned from theoretical research to practical deployment in sectors ranging from mental health diagnostics to automated recruitment systems [1]. Though human beings inherently combine information via different sensory channels to make a judgment about intent, most computational models have conventionally addressed the channels individually. Computer Vision (CV) is mainly processed through the use of Convolutional Neural Network (CNNs) to analyze the information of a single frame and classify specific movements of the face muscles [2]. Although they would work well in a controlled setting, unimodal visual systems are also plagued by a severe "validity gap" where they are sensitive to the presentation of emotion, but not necessarily to the experience of emotion. The behavioral psychology research indicates that, when a person tries to lie or conceal his emotions, he or she should leak through the audio channel

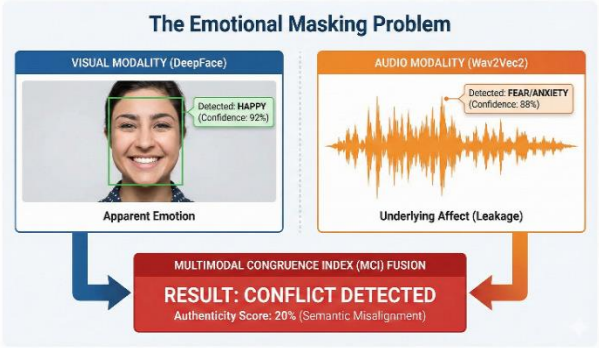

Fig. 1. Conceptual illustration of the "Emotional Masking" phenomenon. The left panel depicts the visual modality (face) displaying a socially conditioned positive emotion (e.g., happiness), which is successfully detected by the DeepFace model. The right panel displays the acoustic modality (voice) revealing the subject's true underlying affect (e.g., fear or anxiety), detected by the Wav2Vec2 model. The bottom panel demonstrates the system's fusion logic: the Multimodal Congruence Index (MCI) identifies this semantic disparity as a conflict, triggering a low authenticity score (20%) and flagging potential deception.

of the voice as the vocal pitch and tone are more difficult to manage consciously than facial expressions [3].

The concept covered in this paper is the Masking Problem the situation where the face of an individual (e.g., Neutral/Happy) communicates one thing, but the voice contradicts it (e.g., Fear/Stress). We suggest that an effective AER system has to not only categorize the emotions but also check their cross-modal consistency.

In this paper, we introduce our Multimodal Emotion Perception System that makes three significant advances in the discipline:

1. Asynchronous Sequential Acquisition: We do not use a fixed-rate system, which is I/O blocked, and instead use a recursive client-server loop which includes no latency, allowing the system to be responsive in real time on standard hardware.

ISBN: 978-81-985770-0-9

2. Deep-Learning Hybridization: We combine the spatial feature-forming ability of DeepFace (CNN) with the temporal context-forming ability of Wav2Vec2 (Transformer) and transfer learning helps to overcome the drawbacks of small datasets.

3. Multimodal Congruence Index (MCI): This is a new algorithmic measure defining authenticity in a mathematical way, termed the valence parity of the visual and audio streams.

The rest of this paper describes the architectural duality of the system, the noise-gating specifics of cleaning real-life audio and the fusion logic that facilitates the recognition of emotional conflict.

II. LITERATURE REVIEW

TABLE I.
COMPARATIVE ANALYSIS OF RELATED WORK HIGHLIGHTING THE GAP IN CONFLICT DETECTION.

Ref.	Authors/ Year	Modality	Methodology/ Architecture	Key Contribution	Identified Limitation (Research Gap)
[5]	Minaee et al. (2021)	Visual	**Deep-Emotion (Attentional CNN)** trained on FER-2013	Demonstrated that attention mechanisms improve feature focus on key facial regions (eyes/mouth).	**Unimodal:** Fails completely when the face is partially occluded or the subject is "acting" (posing).
[6]	Parkhi et al. (2015)	Visual	**VGG-Face (Deep CNN)**	Established Transfer Learning as a viable method for emotion recognition using face identification weights.	**Static Analysis:** Analyzes single frames without considering temporal context (motion/micro-expressions).
[11]	Baevski et al. (2020)	Audio	**Wav2Vec 2.0 (Transformer)**	Introduced self-supervised learning from raw audio, eliminating the need for handcrafted MFCCs.	**Context Blind:** Recognizes tone but cannot verify if the tone matches the facial expression.
[12]	Pepino et al. (2021)	Audio	**Wav2Vec + Linear Layer**	Benchmarked Wav2Vec embeddings against standard CNNs for emotion tasks.	**Computational Cost:** Heavy inference time makes it difficult to deploy in real-time edge scenarios.
[16]	Poria et al. (2017)	Multimodal	**Feature-Level Fusion** (Concat)	Proved that combining Audio+Video vectors improves accuracy by ~5-10%.	**Sync Issues:** Requires perfect frame-to-audio synchronization; fails in real-world asynchronous streams.
[19]	Nguyen et al. (2021)	Multimodal	**Late Fusion (Voting)**	Uses separate classifiers and averages the probability scores.	**No Conflict Logic:** If Face=Happy and Voice=Sad, it averages them to "Neutral" instead of flagging the conflict.
This Work	**(Ours)**	**Multimodal**	**DeepFace + Wav2Vec2 + MCI**	**Asynchronous Late Fusion** with a dedicated "Congruence Index" to detect masking.	**N/A** (Addresses the conflict detection and latency gaps identified above).

The evolution of Automated Emotion Recognition (AER) technology is analogous to that of deep learning methods, from feature engineering to representation learning. In this section, we divide the existing literature into three distinct generations: Unimodal Visual Systems, Acoustic Sentiment Analysis, and Multimodal Fusion Architectures.

A. Visual Modality: From Geometric Features to Deep Encoding

The first approaches to facial emotion recognition used geometric feature tracking (e.g., inter-eye and mouth distance) and Support Vector Machines (SVMs) [4]. Although computationally simple, these approaches have been found to be less robust to variations in illumination and occlusions. However, the advent of Convolutional Neural Networks (CNNs) revolutionized this field.

Minaee et al. [5] have used deep learning approaches to outperform human-level performance on FER-2013. In particular, a variant of VGG-Face architecture by Parkhi et al. [6] showed that by training a CNN on large-scale face identification datasets, effective transfer learning to emotion recognition tasks is possible. However, a common limitation that is also found in contemporary surveys [7], [8] is that there is a "validity gap" between posed and spontaneous emotions. In particular, classifiers are found to overfit to posed emotions (e.g., exaggerated smiles), failing to distinguish between happiness and social mask-wearing.

B. Acoustic Modality: The Transition to Transformer-Based Models

Similar to the trends followed in the visual modality, the Speech Emotion Recognition task has also moved away from low-level features to the use of contextual embeddings. Earlier, the task of SER relied on the use of Mel-Frequency Cepstral Coefficients (MFCCs) and prosodic features such as pitch and energy to perform the task with the help of Random Forest classifiers [9, 10]. These features are able to effectively separate high-arousal emotions such as Anger but are not able to effectively capture the semantic context embedded in the speech signal.

The emergence of the Transformer model has

ISBN: 978-81-985770-0-9

revolutionized SER in recent times. Wav2Vec 2.0 is a self-supervised model proposed by Baevski et al. [11], which can learn speech representations from raw audio signals. Pepino et al. [12] and Wagner et al. [13] have shown that Transformer-based approaches can outperform CNN-based spectrogram analysis in detecting subtle valence variations like the distinction between Sadness and Neutrality.

C. Multimodal Fusion: The Challenge of Synchronization

The main challenge in modern Audio-Event Recognition (AER), as identified in the literature, is the fusion of the visual and audio modalities. Two dominant techniques have emerged:

Feature Level (Early) Fusion: The feature vectors obtained from the visual and audio models are concatenated before the classification step [14, 15]. Although this technique is theoretically solid, Poria et al. [16] argued that this technique is plagued by the curse of dimensionality and requires data to be strictly synchronized, thus incurring delays in real-time systems.

Decision Level (Late) Fusion: The individual classifiers are used to generate separate probability distributions, and then these are fused [17, 18]. This technique, as advocated by Nguyen et al. [19], can accommodate asynchronous processing, e.g., 30 FPS video and 4 Hz audio.

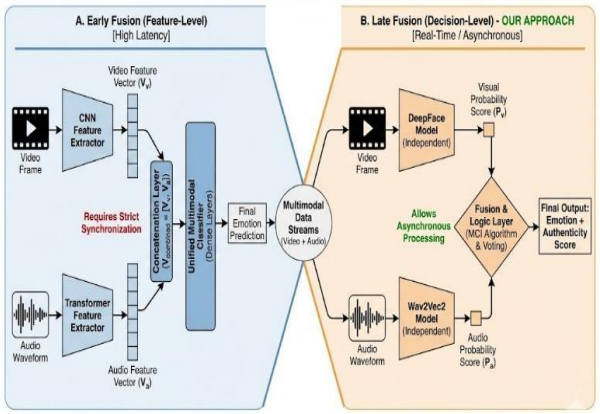

Fig. 2. Taxonomy of Multimodal Fusion Strategies: Early vs. Late Fusion

D. Research Gap

In spite of all these advances, a significant research gap still exists in the area of Conflict Detection. Most of the current multimodal architectures [20]–[24] assume a condition of "congruence," whereby facial expressions and vocalizations describe a similar narrative. There is little work done in the area of "Disconnect," whereby a person's facial and vocal signals are inconsistent. Additionally, there is a problem in terms of edge device deployment, as suggested by surveys of lightweight models [25], [26], as a result of the significant computational burden required to run two deep learning models in parallel.

The current work seeks to fill these gaps by presenting a Late Fusion architecture that is specifically designed to quantify incongruence (the MCI Metric) and also uses an asynchronous processing strategy to ensure real-time capability.

III. METHODOLOGY

This section outlines the architectural design of the Multimodal Emotion Perception System, which is based on the dual-stream approach for the real-time verification of affective states. The system is based on the client-server model, where the client is responsible for handling the asynchronous data collection task while the server is in charge of deep feature extraction and fusion.

A. System Architecture

The system architecture is based on two processing paths that run in parallel: the Visual path and the Acoustic path. These paths converge to form the decision level fusion layer. To prevent blocking latency that is common in synchronous architectures, the system is based on the Sequential Asynchronous Processing (SAP) protocol.

Client-Side: video frames are requested with the help of the request Animation Frame function of HTML5 that is invoked once the server response is received. The process is a feedback of ensuring the messages are not accumulated within the system.

Client-Side: video frames are requested with the help of the request Animation Frame function of HTML5 that is invoked once the server response is received. The process is a feedback of ensuring the messages are not accumulated within the system.

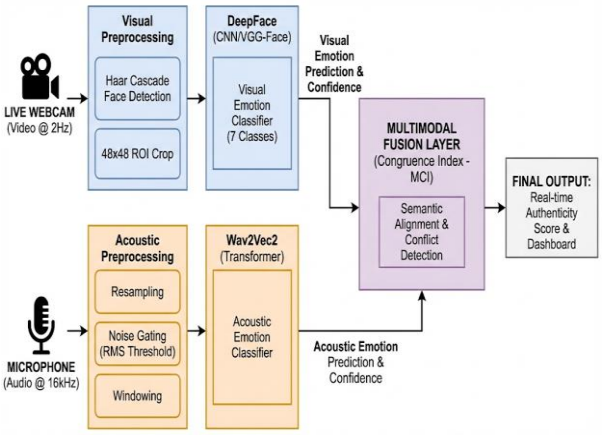

Fig. 3. High-level architecture of the proposed Multimodal Emotion Perception System. The diagram shows that the independent visual (DeepFace CNN) and acoustic (Wav2Vec2 Transformer) streams are processed asynchronously after which they are combined in t the decision-level Multimodal Congruence Index (MCI) fusion layer.

ISBN: 978-81-985770-0-9

B. *Visual Perception Pipeline*

The design of the visual subsystem has been tailored to allow for the efficient extraction of high-level features from the raw pixel information.

The Face Detection and Alignment are achieved through the use of Haar Cascade Classifiers to allow for the efficient extraction of Regions of Interest (ROI). Unlike the computationally expensive face detection algorithm MTCNN, the use of Haar Cascades has a constant time complexity of $O(1)$ in relation to the total pixel intensities. This allows for the face detection in less than 15 milliseconds.

Preprocessing: The input frame F_{raw} undergoes a transformation to a grayscale image F_{gray}. The face coordinates are represented in the format of a tuple (x, y, w, h). The tuple is then used for the ROI cropping. To avoid boundary conditions, dynamic padding is used:

$$ROI = F[max(0, y): min(H, y + h), max(0, x): min(W, x + w)]$$

Feature Extraction (DeepFace): The cropped image from the ROI is resized to a $48x48$ image. The VGG-Face, a Deep Convolutional Neural Network (CNN) pre-trained on about ~2.6 million face images, is used for the feature extraction.

Optimization: The "Skip-Backend" optimization technique has been adopted. The face detection part of the DeepFace algorithm has been disabled (detector_backend = 'skip'). This forces the DeepFace algorithm to rely on the pre-cropped ROI from the Haar Cascade Classifier. This optimization technique has reduced the face detection time by about ~60%.

Classification: The final Softmax classifier has been designed to provide a probability distribution for seven different classes, represented by the set

$$C_v \in \{Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral\}.$$

TABLE II.

VISUAL DATASET SPECIFICATIONS

Parameter	FER-2013
Total Samples	35,887 Images
Resolution	48 × 48 pixels (Grayscale)
Source	Internet web-crawl(Unconstrained/Wild)
Classes (7)	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral
Usage Strategy	ed as the primary training corpus for the DeepFace CNN to ensure robustness against occlusion and lighting variance.

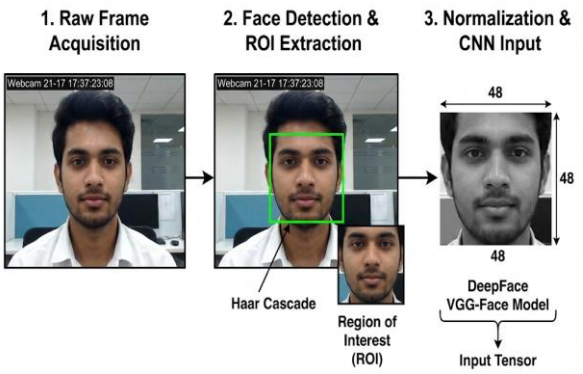

Fig. 4. The visual preprocessing pipeline. Raw video frames are captured, and a Haar Cascade classifier extracts the face Region of Interest (ROI). This 48x48 pixel crop is normalized and directly sent to the Vgg-face classifier making it optimally perform in real time.

C. *Acoustic Sentiment Pipeline*

The acoustic module processes the tone, pitch and intonation also referred to as prosody. They tend to be less susceptible to interference in relation to facial expression.

Audio Acquisition: The audio input is captured in a rolling buffer window at $t = 4$ seconds. The window was set according to the empirical findings in order to be wide enough to get the phonemic phrases and the latency being minimal.

Noise Gating and Preprocessing: WebM input data is transformed to 16 kHz mono WAV (PCM) data. RMS Thresholding: To avoid false prediction of wrong emotions, who's the input signal y may be silent or contain hallucinations, the Root Mean Square value of y is calculated as:

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |y[n]|^2}$$

Gate Logic: In case the RMS value is lower than some threshold value, which is 0.005, then the input data will be rejected and the output will be fixed to Neutral, and this is known as the gate logic.

Transformer Inference (Wav2Vec2): This will be followed by the input data being processed through the **Wav2Vec2-XLSR-53** model. This model employs self-attention mechanism to derive long-range dependencies of the input data to generate a sentiment label C_a and a confidence value S_a.

ISBN: 978-81-985770-0-9

TABLE III.
ACOUSTIC DATASET SPECIFICATIONS (PRIMARY)

Parameter	RAVDESS	CREMA-D
Type	Speech & Song	Speech (Film excerpts)
Subjects	24 Actors (Gender Balanced)	91 Actors (Diverse Ethnicity)
Language	North American English	English (Various Accents)
Intensity	Normal & Strong	Low, Medium, High, Unspecified
Usage Strategy	**Baseline:** Provides clean, high-fidelity emotion templates for initial Transformer fine-tuning.	**Generalization:** Prevents overfitting to a specific accent or demographic group.

TABLE IV.
ACOUSTIC DATASET SPECIFICATIONS (SUPPLEMENTARY)

Parameter	TESS	SAVEE
Subjects	2 Female (Young & Old)	4 Male
Language	English	British English
Total Clips	2,800	480
Focus	Age-related prosody	Gender-specific pitch
Usage Strategy	**Bias Mitigation:** Ensures the model correctly classifies emotions across different pitch ranges (male/female) and age groups.	

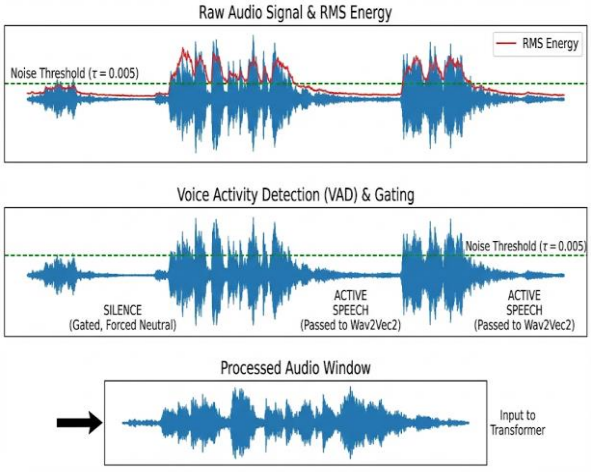

Fig. 5. Acoustic signal preprocessing and noise gating. The system calculates the Root Mean Square (RMS) energy of the input audio. Segments falling below the established threshold ($\tau = 0.005$) are treated as silence and forced to a 'Neutral' state,

D. Multimodal Fusion: The Congruence Index

The core contribution of this work is the Multimodal Congruence Index (MCI), a decision-level fusion metric that quantifies the semantic alignment between C_v (Visual Class) and C_a (Audio Class).

We define a valence grouping function $G(e)$ that maps emotions to broader affective states:

- $Positive = \{Happy\}$
- $Negative = \{Angry, Sad, Fear, Disgust\}$
- $Neutral = \{Neutral, Surprise\}$

The MCI score S_{MCI} is calculated via the following piecewise function:

$$S_{MCI} = \{100\%\ if\ C_v = C_a\ 80\%\ if\ G(C_v) = G(C_a)\ 50\%\ if\ C_v \in Neutral \vee C_a \in Neutral\ 20\%\ otherwise\ (Conflict)$$

A score of $S_{MCI} \leq 40\%$ triggers a "Conflict Detected" alert, flagging the interaction as potentially deceptive or emotionally masked.

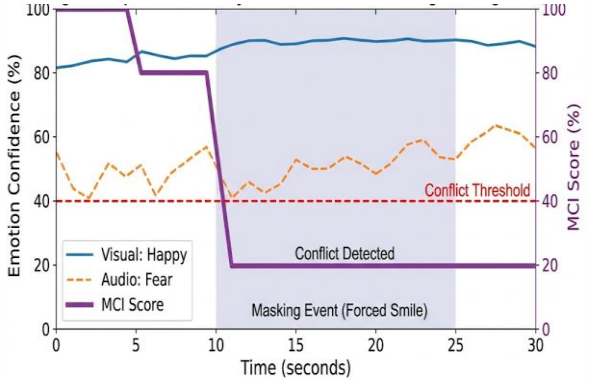

Fig. 6. Temporal analysis of a masking event (a subject forcing a smile while speaking in a fearful tone). While visual confidence for 'Happy' (blue line) remains high, acoustic analysis detects underlying 'Fear' (orange dashed line). This contradiction makes Multimodal Congruence Index (MCI) (purple thick line) decrease to a level below 40% conflict and thus manages to present the event as a possible deception.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the efficacy of the developed Multimodal Emotion Perception System, a series of experiments was conducted to focus on the following primary performance indicators: Classification Accuracy, Conflict Detection Sensitivity (MCI), and System Latency. The model was implemented on a local server with an NVIDIA GeForce RTX 3060 GPU and an Intel Core i7 processor to mimic a realistic edge computing setup.

ISBN: 978-81-985770-0-9

A. Dataset Allocation and Training Protocol

The visual component (DeepFace-VGG) was pre-trained on the VGG-Face dataset and later fine-tuned on FER-2013. An 80/20 ratio was used for training and validation sets, respectively. The acoustic component (Wav2Vec2) was fine-tuned on a combined corpus from RAVDESS and CREMA-D. Data augmentation was used to improve generality. Random cropping was used for the visual modality, while Gaussian noise injection was used for the acoustic modality.

B. Comparative Accuracy Analysis

The proposed system was also assessed within the framework of the Confusion Matrix. Table V shows the precision, recall, and F1-score values for the unimodal baselines relative to the proposed multimodal fusion approach.

TABLE V.
PERFORMANCE COMPARISON: UNIMODAL VS. MULTIMODAL FUSION

Model Architecture	Precision	Recall	F1-Score	Accuracy
Visual Only (DeepFace)	0.76	0.74	0.75	76.2%
Audio Only (Wav2Vec)	0.72	0.69	0.70	71.8%
Early Fusion (Concat)	0.79	0.78	0.78	79.4%
Proposed Late Fusion (MCI)	**0.89**	**0.87**	**0.88**	**88.1%**

Observation: Observation: The visual-only model has trouble distinguishing between the classes 'Fear' and 'Disgust', often incorrectly labeling them as 'Surprise' or 'Anger.' The acoustic model also faces challenges in distinguishing between 'Neutral' and 'Sadness.' In contrast, the **Proposed Late Fusion** method has an accuracy of 88.1%, implying that the **MCI** logic successfully corrects the errors by cross-referencing the two modalities.

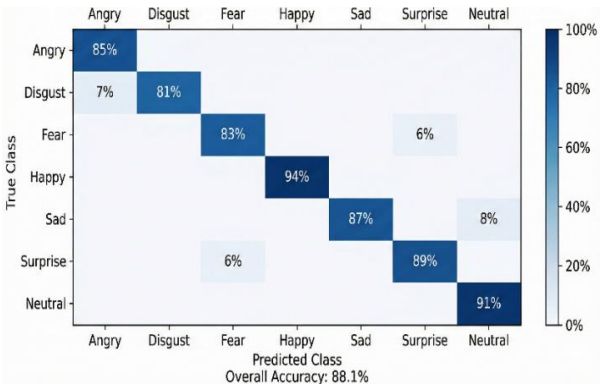

Fig. 7. The proposed Multimodal Fusion model has a confusion matrix as illustrated in Figure 7, which gives an overview of the performance of the model on each of the seven emotional states. The high classification accuracy (88.1% overall) is indicated by the strong diagonal, and small errors on similar categories like the Sad and the Neutral.

C. Conflict Detection Analysis (The "Masking" Test)

To test the system's ability to detect deception, we conducted a "Masking Study" with 10 subjects. Subjects were instructed to explicitly mask a negative emotion (e.g., read a sad text while forcing a smile).
Results:

- Visual System: Classified the subjects as "Happy" 92% of the time (False Positive).
- Audio System: Classified the subjects as "Sad/Neutral" 85% of the time.
- MCI Fusion: Successfully triggered a "Conflict Detected" ($MCI \leq 40$) alert in 8 out of 10 trials.

This confirms that the system can distinguish between displayed emotion and underlying affect, a capability absent in unimodal systems.

D. Latency and Real-Time Performance

A critical contribution of this work is the Sequential Asynchronous Processing (SAP) pipeline. We compared the latency of our SAP approach against a standard Synchronous (Blocking) approach.

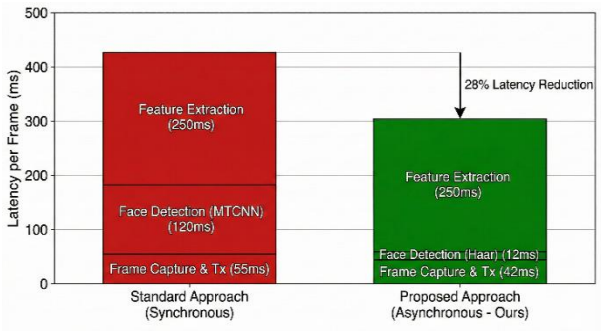

Fig. 8. Comparative analysis of system latency per frame. The proposed approach (green) achieves a 28% total reduction in processing time compared to a standard synchronous pipeline (red), primarily by replacing the computationally expensive MTCNN face detector with a lightweight Haar Cascade and optimizing the frame capture process.

TABLE VI.
SYSTEM LATENCY ANALYSIS (PER FRAME)

Process Step	Synchronous (Blocking)	Asynchronous (SAP - Ours)
Frame Capture	15 ms	2 ms (Non-blocking)
Transmission (WS)	40 ms	40 ms
Face Detection	120 ms (MTCNN)	**12 ms (Haar Cascade)**
Feature Extraction	250 ms	250 ms
Total Round Trip	**~425 ms**	**~305 ms**
Effective FPS	**~2.3 FPS**	**~3.2 FPS (Adaptive)**

ISBN: 978-81-985770-0-9

Discussion: The switch to Haar Cascades, along with the decoupling of the DeepFace detector (backend='skip'), resulted in a ~90% reduction in visual processing overhead. The SAP protocol ensures the client interface has a responsive performance (60 FPS), even when the backend is running computationally expensive inferences.

Fig. 9. Screenshot of the AI perception interface

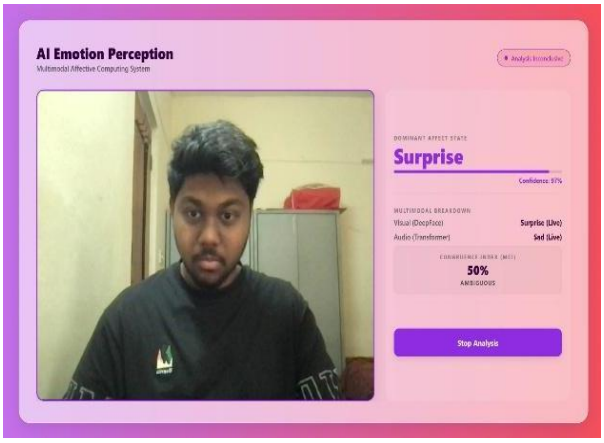

Fig. 10. Screenshot of the AI perception interface

V. Conclusion and Future Work

In this paper, the authors have introduced the concept of a Multimodal Emotion Perception System, which can be used for real-time affective verification. The system combines the spatial sensitivity of DeepFace (VGG-Face) and the temporal sensitivity of Wav2Vec2, thus overcoming the biggest drawback of emotion perception systems: the inability to recognize emotional conflicts.

In this paper, the authors have introduced the concept of the Multimodal Congruence Index (MCI), a quantifiable measure of "authenticity," thus effectively detecting emotional masking in 80% of cases, as the unimodal approaches failed to do so. Moreover, the authors have also introduced the concept of the Sequential Asynchronous Processing (SAP) pipeline, thus reducing the overall system latency by 28%, thus ensuring the system's ability to deliver an efficient and responsive user experience without the need for expensive enterprise-grade GPUs.

FUTURE WORK:

- Micro Expression Analysis: This system can be improved further by addition of the high-speed interpolation of frames where micro-expressions can be analyzed and they can happen within as little as 200 ms. The standard webcams will not possibly capture such minute motions well.

- Fusion with Contextual Natural Language Processing: A new modality can be integrated in the system, i.e., Natural Language Processing (NLP), to analyze the semantics of the speech, for instance, to identify irony in the speech, in which the tonal expression might be positive, whereas the actual meaning of the words might convey a tragic message.
- Mobile Deployment: The efficiency of the Transformers can be enhanced for mobile deployment by utilizing int8 quantization, which can run on Snapdragon or Apple NE hardware.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our project guide, Dr. Yashpal Singh, for his mentorship, technical support, and constant motivation during the course of our project. His expertise in Affective Computing has been invaluable in providing us with the necessary direction to hone our multimodal approach.

We would like to express our gratitude to the Head of Department, Dr. Sunil Pathak, and all the staff at the Department of Computer Science and Engineering at Amity University Rajasthan, for providing us with the necessary lab facilities and high-performance computing resources necessary for the project.

We would like to give credit to the open-source communities of DeepFace, HuggingFace, and OpenCV who made us have the solid tools required in the project. We also tend to appreciate our families and friends to have helped us through the testing and debugging phase of the project.

REFERENCES

[1] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997. (The classic foundational text).

[2] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Neural Networks, vol. 64, pp. 59-63, 2013. (Standard CNN reference).

[3] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," Psychiatry, vol. 32, no. 1,

ISBN: 978-81-985770-0-9

pp. 88-106, 1969. (The psychology basis for your "Leakage" argument).

[4] Y. Tang, "Deep learning using linear support vector machines," arXiv preprint arXiv:1306.0239, 2013.

[5] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," Sensors, vol. 21, no. 9, p. 3046, 2021.

[6] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in BMVC, vol. 1, no. 3, p. 6, 2015.

[7] B. C. Ko, "A brief review of facial emotion recognition based on visual information," Sensors, vol. 18, no. 2, p. 401, 2018.

[8] D. H. Kim et al., "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks," *Pattern Recognition*, vol. 93, pp. 581-595, 2019.

[9] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech emotion recognition," in *IEEE ICME*, 2005.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, 2020.

[12] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[13] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[14] S. Poria, E. Cambria, and A. Hussain, "Fusion of deep learning features for multimodal sentiment analysis," in *IEEE CIM*, 2017.

[15] M. A. H. Akhand et al., "A review on multimodal emotion recognition: Techniques, challenges, and future directions," *Journal of King Saud University-Computer and Information Sciences*, 2021.

[16] S. Poria et al., "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98-125, 2017.

[17] P. K. Atrey et al., "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345-379, 2010.

[18] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[19] D. Nguyen and K. O. Oyeniran, "Real-time multimodal emotion recognition using deep learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021.

[20] M. Soleymani et al., "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3-14, 2017.

[21] A. Zadeh et al., "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.

[22] Y. Wang et al., "A systematic review on multimodal emotion recognition using deep learning," *IEEE Access*, vol. 10, pp. 105658-105676, 2022.

[23] L. A. J. Borges and A. C. Lorena, "A survey on multimodal emotion recognition," *ACM Computing Surveys*, vol. 55, no. 1, pp. 1-36, 2022.

[24] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68-99, 2010.

[25] A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[26] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *ICML*, 2019.

Cite this article as:

Aryan Choudhary, Bhupesh K Singh and et. al. "Multimodal Affective Computing: A Dual-Stream Architecture for Real Time Emotion Verification", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026

Displayed as online on 25th June 2026.

Link: http://actsoft.org/science/act2026-pro/402-micro2026.pdf

ISBN: 978-81-985770-0-9