Multimodal Emotion Recognition in Speech: A Forensic Analysis of Facial Landmark Dynamics and Acoustic Features

*Vernika Mehta, Dr. Surbhi Mathur
National Forensic Sciences University, Gandhinagar, Gujarat, India

Abstract:
Emotion recognition is one of the core themes of affective computing, particularly in forensic science, where identification of subtle emotional cues can aid psychological evaluation, deception detection, and witness reliability assessment. The work outlined presents a multimodal approach combining speech acoustics with facial landmark analysis to investigate how emotional states affect communication. Focus was on five sustained vowels, /a:/, /i:/, /u:/, /ɔ:/ and /o:/ selected for their distinctive articulatory and acoustic properties, which vary with emotions and speaker groups. Examining vowels in emotionally varying speech and correspondence to facial landmark movements revealed statistically significant emotion-specific patterns across gender and emotional categories. The approach employs Dlib 68-point facial landmark model, OpenCV, and acoustic analysis toolkits like PRAAT, with data stored frame-by-frame for meticulous examination. A custom emotional speech dataset was collected from Indian speakers under controlled conditions, with a representative demographic sample and reproducible results. The dataset containing raw video recordings and frames labeled with facial landmark coordinates is available on reasonable request to the corresponding author, ensuring ethical utilization. To ensure transparency and reproducibility, protocols were devised for system installation, data acquisition, analysis, and interpretation. The method offers a robust, non-invasive technique with implications for forensic science and affective computing.

Keywords: Facial Landmarks analysis, Facial Expressions, Speech Acoustics, Multimodal Emotion Recognition, Multimedia Forensics, Vowel-Based Emotion Profiling

Declaration of Interest statement

- This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

1. Introduction

Emotion recognition is a key component of affective computing, which aims to bridge the gap in human-machine communication by enabling systems to detect and understand human emotions. Emotions, being multimodal, can be conveyed through various channels, including voice, gestures, facial expressions, and physiological indicators. Widespread are speech and facial expressions, which offer numerous additional clues about an individual's emotional state. For instance, a software programme could employ voice analysis to detect variations in tone and pitch to ascertain whether a person is feeling happy or sad during a conversation. Additionally, facial recognition technology could analyse micro-expressions to identify subtle changes in facial muscles that indicate emotions such as surprise or anger.

This study presents a novel approach in forensic science by integrating multimodal emotion recognition, specifically speech acoustics and facial landmark analysis, to gain a better understanding of how emotional states influence human communication. It can be beneficial to be able to decode emotions through vocal and facial patterns to evaluate witness credibility, detect deception, or profile suspects. In traditional forensic techniques, facial movement and

vocal modulation are largely ignored. This research offers measurable and statistically significant markers by examining sustained vowel articulations across different emotional states and linking them to facial landmark changes, which can enhance forensic interpretations. Furthermore, the identified gender- and emotion-specific patterns strengthen the granularity of these assessments. This study is particularly significant within culturally specific contexts, such as the Indian demographic presented here, where norms for emotional expression may differ. It fills a critical research gap by offering a scientifically validated, non-invasive method for emotional profiling that could improve the accuracy and reliability of investigative procedures in legal, psychological, and security domains.

1.1. Speech Acoustics and Emotional Expression

Our speech conveys a significant amount of information regarding our emotions. The intonation, mood, rhythm, and pace of our speech can all provide individuals with insight into our feelings when we communicate. For instance, people often raise their voices and speak at a quicker pace when they are angry or enthusiastic. However, their tone tends to lower, and their speech becomes more sluggish when they are depressed or fatigued. It is now significantly simpler to detect these emotional signals in speech with the aid of technologies such as openSMILE (Eyben et al., 2010). Computers have been able to recognise even the smallest patterns in human utterances as a result of deep learning. This facilitates their comprehension of our emotions (Chang et al., 2023).

Table 1- Summary of how basic emotions modulate voice via- pitch (F0), amplitude and duration (Adapted from Chang et al., 2023)

Emotion	Mean F0 (Hz)	F0 Range	Mean Amplitude	Duration
Anger	Higher	Wider or Similar	Higher	Shorter
Fear	Higher	Narrower, Wider, or Similar	Higher or Lower	Shorter
Happiness	Higher	Wider or Similar	Higher or Equal	Shorter or Longer
Sadness	Lower or Similar	Narrower, Wider, or Similar	Lower	Longer

Table 1 outlines the overarching patterns of prosodic features, including mean fundamental frequency (F0), F0 range, amplitude, and duration, corresponding to four primary emotions: anger, fear, happiness, and sadness. These differences highlight how emotional expression shapes the voice qualities of speech.

1.2. Facial Landmark Patterns and Emotion Recognition

Facial expressions are primary sources of emotional information, as specific muscle movements indicate various moods. In humans, these serve as a potent, non-verbal means of communication. They transform how we convey emotions and interact with others. This research investigates the impact of emotions on speech by analysing facial landmarks, key points on the face that help to decipher an individual's feelings and speech patterns. The findings offer intriguing insights for forensic science and affective computing.

The Facial Action Coding System (FACS) (Ekman & Friesen, 1978) provides a comprehensive framework for categorising facial movements into distinct Action Units (AUs), each representing a unique muscle contraction. To conduct a thorough analysis of these emotions, it is necessary to accurately identify the significant features of the face, i.e, facial landmarks, including the corners of the eyes, eyebrows, and lips. To increase the accuracy of facial landmark identification and, consequently, the accuracy of emotional recognition, geometric feature-based techniques have been used in recent studies. (Shanthi & Nickolas, 2022).

Related Work

Multimodal Emotion Recognition: Integrating Speech and Facial Cues

Unimodal systems function quite effectively, yet employing multiple types of input provides a clearer understanding of an individual's feelings. Multimodal emotional detection systems combine verbal and facial cues to compensate for the limitations of each. (Mittal et al., 2020) introduced the M3ER system in 2020. It improves performance on benchmark datasets by combining voice, text, and facial inputs using a multiplicative fusion method. (Yan et al.,2024) proposed a system that utilizes advanced fusion algorithms to combine voice, body movements, and facial expressions, thereby enhancing recognition accuracy. These methods demonstrate the importance of incorporating different types of data to gain a comprehensive understanding of people's feelings. (Singh et al., 2023) proposed a hybrid 3D-CNN + ConvLSTM model for video-based facial expression recognition, capturing spatial–temporal emotion patterns accurately with fewer parameters and low latency and hence well-fitted for real-time embedded deployment. Their work highlights the necessity of combined spatial and temporal modeling—a paradigm that can possibly be used to improve our landmark-based approach using temporal neural modules to capture smooth facial motion dynamics during speech in a more principled way. In addition, recent advances in facial recognition models and visual attention models have proven to be of great potential in enhancing landmark precision. A research paper published in ("Optimization of 2D and 3D Facial Recognition," 2025) introduced a hybrid model combining CBAM-AlexNet and ResNeXt models to enhance 2D and 3D facial recognition procedures. The process illustrated enhanced robustness and accuracy in facial feature extraction across varied conditions. Such visual feature boosting methods have the potential to enhance landmark-based emotion profiling used in our research, particularly in forensic research where reliability and accuracy are top priorities. Xiong et al. (2025) emphasizes the use of combining local–global attention mechanisms with facial landmark distributions to improve the accuracy of expression recognition. The model, as reported, uses contrastive learning on Dlib-based landmark datasets to encode detailed information and wide facial context effectively, hence achieving improved accuracy under various conditions. Use of such attention-based landmark refinement has the potential to improve the readability and accuracy of our forensic emotion profiling, particularly when combined with acoustic features. A notable recent work in this direction is by (Salas-Cáceres et al., 2024), who proposed an audio-visual deep learning architecture that combined audio and visual channels with temporal understanding by embracing Long Short-Term Memory (LSTM) networks. Their model tested different fusion techniques—concatenation, Principal Component Analysis (PCA), autoencoders, and EmbraceNet—and reported state-of-the-art performance on established benchmark datasets like RAVDESS and CREMA-D, with achieved accuracies of 88.11% and 80.27%, respectively. The work highlights the significance of multimodal integration of emotions and dynamic modeling, and their results complement our goal of fusing speech and facial signals. Although our contribution is parallel in a forensic context with focus on close facial landmark and acoustic examination, the work is a suitable point of reference for future advancement involving deep temporal architectures and adaptive fusion techniques. There has been

extensive research in the areas of emotion detection and facial recognition that utilizes artificial intelligence. There is a lack of research on how human facial expressions influence speech acoustics and how facial landmarks change with different expressions in men and women. Another notable direction is the application of reduced-feature facial emotion recognition in immersive environments. A recent study titled “Facial emotion recognition with a reduced feature set for video game and metaverse avatars” Bellenger, Chen, and Xu (2024) presents a real-time system using only 11 Ekman Action Units derived from Dlib’s 68-point model. This approach significantly reduces data and computational requirements, making it scalable for large virtual environments like online games or metaverse platforms. The system achieves high accuracy and shows promise in enhancing avatar-based social interaction, particularly for users with difficulties interpreting facial expressions. These innovations highlight how real-time, low-latency emotion detection can influence affective computing and educational applications, bridging the gap between virtual and emotional presence. The same principles could inform forensic avatar simulations or virtual reconstructions of emotional behavior.

Comprehending the connection between vocal patterns and facial expressions is central to the development of effective emotion recognition systems, particularly in high-risk applications such as forensics. Despite the advances made with technology, facial and vocal content conveying delicate emotions is unmapped. Identifying how emotions influence vocal tone, speech rhythm, and facial structure has great utilitarian significance in lie detection, witness evaluation, and emotionally intelligent systems. Usually, changes brought about by emotions are clearly visible in terms of frequency. For instance, whereas grief or anger often lowers frequencies, happiness or excitement may raise them (Kamiloğlu et al., 2020).

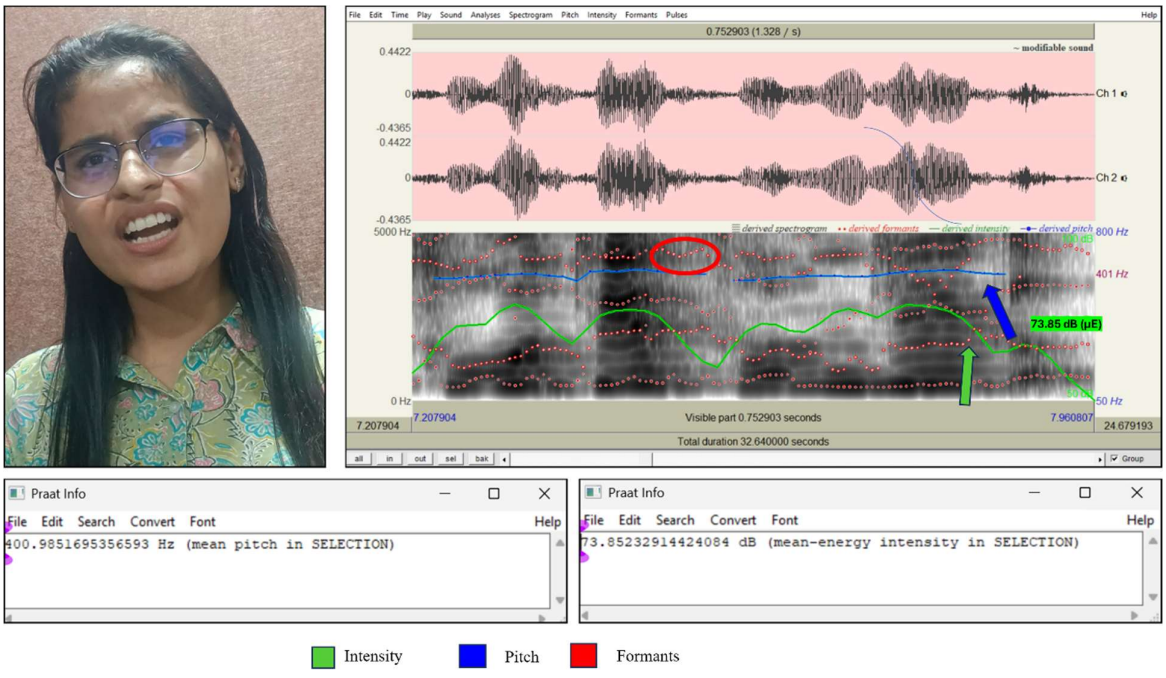

Figure 1: Acoustic and visual analysis of an angry emotional expression in a female speaker using Praat software.

The speaker in Figure 1 shows obvious signs of rage in the left panel, including a stiff jaw and strained facial muscles. Typical markers of high-arousal emotions, such as anger, are acoustic measures showing a high mean pitch of 400.98 Hz and a heightened mean intensity of 73.85 dB. This multimodal analysis correlates acoustic parameters with facial cues, enhancing

emotion detection accuracy. Because different emotions carry unique combinations of facial and vocal cues, the multimodal approach improves emotional interpretation. In forensic applications, where subtle emotional expressions can be used to verify testimonies, identify deception, or analyse recordings for authenticity, such insights are especially valuable. In conclusion, frequency, pitch, and intensity, combined with facial landmark analysis, provide a thorough framework for understanding how emotions impact speech and facial expressions. Therefore, this collaboration will not only deepen our understanding of emotional communication but also hold great promise for the advancement of affective computing and forensic science.

Table 2- Overview of selected multimodal emotion recognition studies (2018–2021). (Schoneveld, Othmani, & Abdelkawy, 2021)

Study Title	Authors	Publication Year	Key Findings
Emotion Recognition in Speech using Cross-Modal Transfer in the Wild	Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, Andrew Zisserman (Albanie, Nagrani, Vedaldi, & Zisserman, 2018)	2018	Developed a method to learn speech emotion embeddings without labeled audio by transferring annotations from facial expressions, achieving state-of-the-art results on benchmark datasets.
Emotion Recognition System from Speech and Visual Information based on Convolutional Neural Networks	Nicolae-Catalin Ristea, Liviu Cristian Dutu, Anamaria Radoi (Ristea, Duţu, & Radoi, 2019)	2020	Proposed a real-time emotion recognition system combining visual and audio data using deep CNNs, demonstrating improved accuracy over single-modality approaches.
Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition	Liam Schoneveld, Alice Othmani, Hazem Abdelkawy (Schoneveld et al., 2021)	2021	Introduced a deep learning approach utilizing knowledge distillation and model-level fusion of audio-visual data, outperforming previous methods in predicting valence on the RECOLA dataset.
M3ER: Multiplicative Multimodal Emotion Recognition Using	Trisha Mittal, Uttaran Bhattacharya, Rohan	2020	Presented a method combining facial, textual, and speech

Facial, Textual, and Speech Cues	Chandra, Aniket Bera, Dinesh Manocha (Mittal et al., 2020)		cues with a multiplicative fusion strategy, achieving mean accuracies of 82.7% on IEMOCAP and 89.0% on CMU-MOSEI datasets.

An overview of a few selected studies on emotion recognition using multimodal data sources, including speech, facial expressions, and text input, is provided in Table 2. Highlighting developments in deep learning methods, fusion strategies, and performance gains across multiple benchmark datasets, the list includes study titles, authors, publication years, and key findings.

2. Methodology

Emotion, speech acoustics, and facial expressions are entwined- a truth gaining traction in both psychology and artificial intelligence. Particularly useful for understanding human emotions entirely are multimodal techniques, which combine multiple signals. Given past studies that have typically focused on modalities in isolation, this study aims to address a significant gap. It will specifically examine how emotions influence both visual and auditory expressions taken together. Thus, the primary focus of research is on understanding the precise interactions between facial landmark patterns and voice acoustics as they convey emotions. The integration of these modalities, according to evidence, may result in a better understanding of artificial emotional intelligence (Wang et al., 2023) and interpersonal communication.

Finding acoustic features, making meaningful connections with these features, and analysing the acoustic patterns in light of the connections made are the three steps involved in decoding emotions in speech. This emphasises how crucial an integrated methodology is to precise emotional analysis. Ultimately, this methodological framework not only addresses the research issue at hand but also facilitates a deeper understanding of how emotional dynamics operate in interpersonal relationships, leading to significant breakthroughs in both academic research and practical technological applications. (Wang et al., 2023).

We built and utilized a large dataset of facial coordinates to investigate how facial landmarks evolve as people speak in various emotional tones. The target group consists of Indian men and women who speak Hindi fluently. They are between 18 and 40 years old. We selected participants to represent a diverse range of individuals in India, ensuring that their ages and genders were evenly distributed. Ethical approval for the study was granted by the University Ethical Committee, National Forensic Sciences University (Certificate No. NFSU/SDSR/IEC/Certificate/272/21, dated 17th September 2021). The study examines the pronunciation of five pure vowels in four different emotional states to provide a comprehensive understanding of how emotions influence speech output.

Figure 2- Sample set of five long vowels (/a:/, /i:/, /ɔ:/, /o:/, /u:/) across four emotional states used in the study.

Four distinct emotional expressions—Happy, Angry, Sad, and Normal—as well as five sustained vowel sounds—/a:/, /i:/, /ɔ:/, /o:/, /u:/—that were used in the study to investigate the facial and acoustic traits associated with the production of emotional speech are depicted in the Figure 2.

2.1. Video Recording Setup

A mobile device was employed to record the video data. Samples were obtained from the OnePlus smartphone (Model: ONEPLUS A6000; Software Version: V3.14.40). A resolution of 1080p and a frame rate of 30 frames per second were chosen. The room was soundproofed to ensure the clarity of the speech samples, thereby reducing external noise and visual distractions during the recordings. The camera was set up about 104.5 cm away from the participant and zoomed in 2x to make their facial features more straightforward for identifying landmarks. This arrangement enabled high-quality video capture suitable for precise facial landmark and movement analysis. The emotional states of the participants were divided into four categories. These emotions were chosen for their recognisability and ubiquity, establishing a solid basis for investigating the correlation between speech patterns and emotions.

Figure 3- Visual representation of facial expressions (Happy, Angry, Sad, Neutral) in a male speaker.

Figure 3 depicts facial cues associated with specific emotions- Happy, Angry, Sad, and Neutral, used in the study to examine the relationship between facial expressions and vocal traits during emotional speech production. These expressions were recorded under controlled settings to examine facial landmark dynamics during emotional speech. Both acoustic and visual metrics examine emotional discourse. Long vowels are collected from speech and analysed, with the script specifically tailored to ensure their frequent occurrence. Participants recite the identical script while conveying four distinct emotions- anger, happiness, sadness, and neutrality to ensure consistency in verbal content. The recordings capture various facial and articulatory motions, including lip and tongue movements, jaw extension, iris movement, eye blink frequency, and other facial landmarks. Facial landmarks are identified and utilised to examine the impact of emotional states on vocal acoustics and expressiveness.

2.2. Facial Landmark Detection: On the face, facial landmarks are specific sites including the corners of the mouth, the borders of the eyes, and the tip of the nose. These sites

were identified and followed using a devised code that precisely gauges the distances between them depending on the emotional state of the speaker and the vowel being voiced.

```
START
→ Input: video_file
→ Load face_detector and landmark_predictor
→ Initialize data structures for distances & coordinates
→ video_capture ← cv2.VideoCapture(video_file)
→ total_frames ← gets the total frame count
→ Initialize paused = False, reverse = False, frame_number = 0
WHILE True:
   IF not paused:
      IF not reverse:
         ret, frame ← read next frame
         frame_number++
         Reset frame_number if at the end of the video
      ELSE:
         frame_number--
         Reset frame_number if less than 0
   IF frame read failed:
      BREAK
   gray ← cvtColor(frame, GRAY)
   faces ← face_detector(gray)
   FOR each face in faces:
      Get landmarks using landmark_predictor
      DRAW landmarks & label points
      → Calculate all specified distances between landmarks:
         - dist(62, 66), dist(61, 67), dist(63, 65), etc.
         - Append each to its respective list
         - Append the current frame_number to the frame_list
      → Generate a coordinate frame with a distance text overlay
   → Display video & coordinate_frame windows
   → Save the current frame image to a file
   → Capture key input:
      - q: quit
      - p: pause/unpause
      - r: reverse/unreverse
      - f: play forward
   → Adjust frame_number if reversed or playing
END WHILE
→ Release video_capture
→ Destroy all windows
→ Create a DataFrame from all distances & coords
→ Save the DataFrame as an Excel file
PRINT "execution complete"
END
```

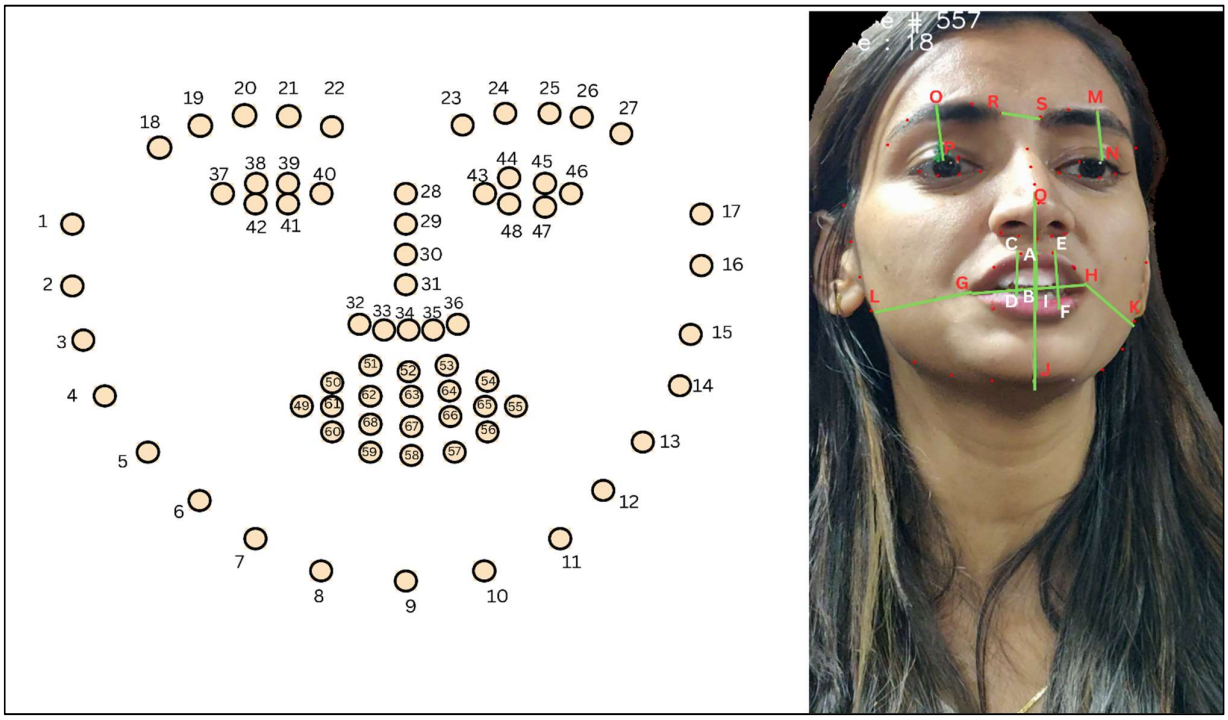

Figure 4- 68-point facial landmark model shown on a generic face and actual participant frame

Figure 4 depicts 68 essential facial features superimposed on a human face (left) alongside their real-time application on a subject's face (right), utilised to track the motions of vital areas including the eyes, eyebrows, nose, lips, and jaw. Designated landmark points (e.g., points 61–68) in the perioral region are crucial for examining lip and jaw movements during speaking and emotional expression, which helps in affective computing and forensic emotion recognition studies.

2.3. Landmark Detection and Distance Measurement

Custom Python scripts were created to extract and analyse facial landmarks from video sources automatically. First, a facial landmark detection library enabled the identification of face landmarks frame by frame. Once the landmarks were localised, particular frames matching the articulation of specific vowels were programmatically marked. Measures between predefined pairs of landmarks were computed at these critical times. These distances - both vertical and horizontal were chosen in line with articulatory traits including lip separation, jaw displacement, and eyebrow elevation. Facial motions during the utterance of every vowel sound across many emotional expressions were quantified using the measures.

The study involved analysing the speech of the subjects to understand how various emotions influence vowel articulation. The study aimed to identify patterns and notable variations associated with each emotional state by measuring the distances between facial landmarks during the pronunciation of various vowels.

Figure 5- Real time Facial landmark detection on participants during emotional speech articulation.

Captured using Python code, the image displays four people with face landmarks and measurements overlaid during a conversation. These landmarks track dynamic facial traits in real-time, including jaw alignment, lip curves, and eye location. Frame number, timestamp, and matching coordinate values provide the data necessary for analyzing face movement differences between individuals and emotional expressions.

Table 3- Description of 11 Measured Distances Between key Facial Landmarks During Emotional Speech Analysis

S.No	Distance Between Landmarks	Description
1	Distance between A and B	The vertical distance between the centre of the lips
2	Distance between C and D	Vertical distance between the lips
3	Distance between E and F	Vertical distance between the lips
4	Distance between G and H	Horizontal distance between the lips
5	Distance between I and J	Distance between the lower lip end and the chin
6	Distance between H and K	Distance between the jaw and the left corner of the lip
7	Distance between G and L	Distance between the jaw and the right corner of the lip
8	Distance between M and N	Distance between the eyebrows and eyelashes
9	Distance between O and P	Distance between the eyebrows and eyelashes
10	Distance between J and Q	Horizontal distance between the chin and the nose
11	Distance between R and S	Distance between the eyebrows

Table 3 presents a systematic summary of the principal face measurements derived from the landmarks depicted in the preceding image, facilitating the quantitative assessment of facial movement across different emotional states.

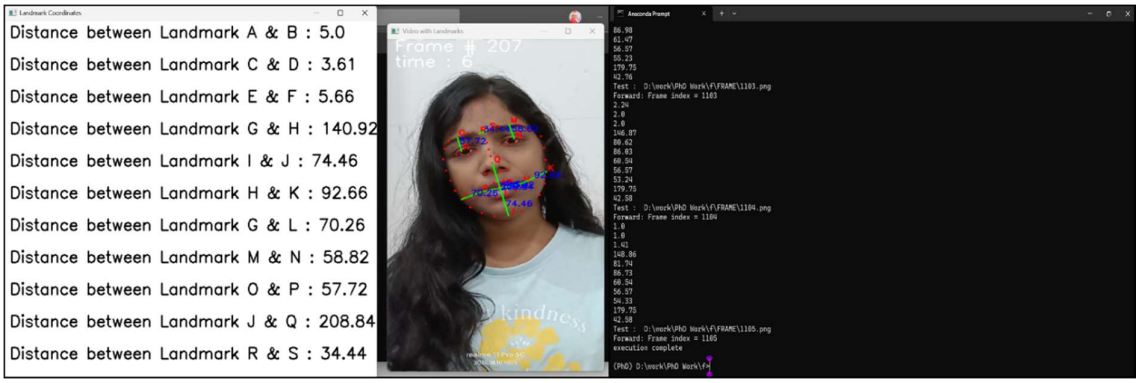

Figure 6- Facial landmark distance analysis in real-time during vowel articulation

Figure 6 shows the technique of facial landmark coordinate analysis in speech production. Computed distances between particular landmark pairs- such as lips, chin, eyebrows are displayed in real-time on the left. The Anaconda interface (right) records live coordinate data. Center: video frame; Left: distance overlay between key facial features. This arrangement enables the precise measurement of facial movements during the articulation of specific vowels. Rigorous statistical analysis was conducted once the data were gathered to determine the relevance of variations in facial landmark positions across different emotional states for different vowels. This meant matching the distances for every vowel and emotional state to spot regular trends and deviations.

3. Results

Our results indicate that distinct emotional states are associated with specific patterns of facial landmark movement. These differences can be consistently measured and examined, proving the promise of facial landmark analysis as a valuable tool for emotional identification.

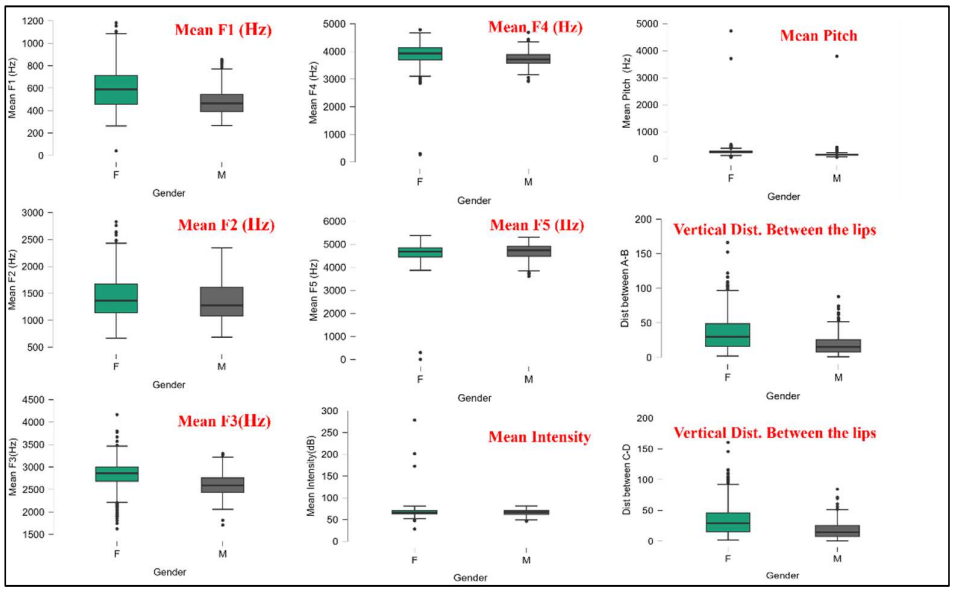

Figure 7- Box plots comparing acoustic features (F1–F5, Pitch, Intensity) and lip distances between male and female participants.

Alongside vertical distances (A–B and C–D) between the lips, Figure 7 shows boxplots illustrating the distribution of different acoustic parameters : Mean Formant Frequencies (F1 to F5), Pitch, and Intensity- alongside male and female participants. There are apparent gender-

based differences; Female participants generally show wider lip spacing and higher formant frequencies.

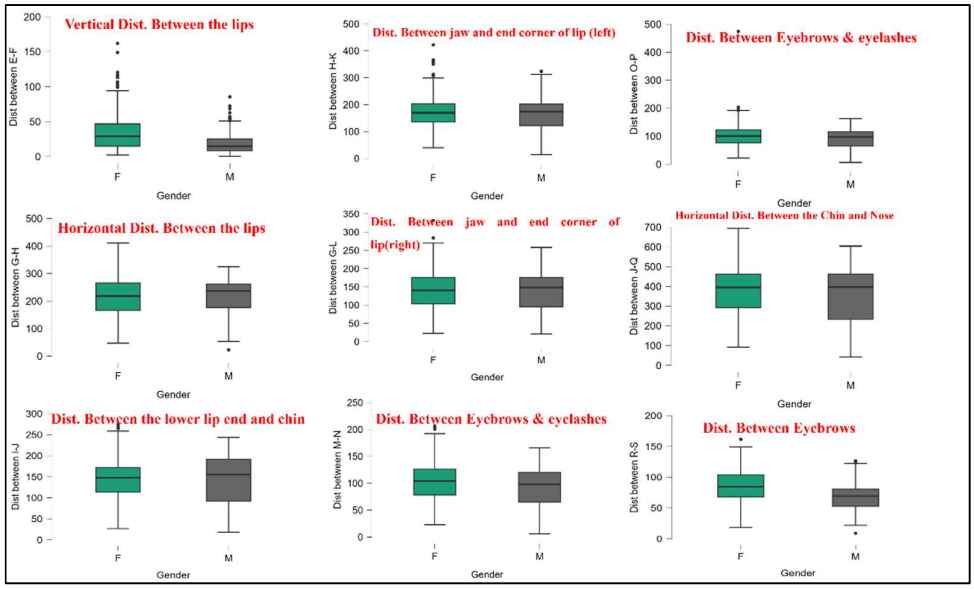

Figure 8: Box plot comparisons of facial landmark distances during vowel articulation across gender.

Box plots comparing male and female subjects across a range of facial landmark distances are shown in Figure 8. These include the distances between the lips and the jaw or chin, the eyebrow-to-eyelash/eyebrow spacing, and the vertical and horizontal lip spans highlighting structural and expressive variation.

Table 4-Statistical significance of acoustic and facial features across vowel, expression, gender, and their interactions.

	Vowel	Expression	Gender	Vowel * Gender	Gender * Expression	Vowel * Expression	Gender * Vowel * Expression
Mean F1	✓	✓	✓	✓	✗	✗	✗
Mean F2	✓	✗	✓	✓	✗	✗	✗
Mean F3	✓	✗	✓	✗	✗	✗	✗
Mean F4	✓	✓	✓	✗	✗	✗	✗
Mean F5	✓	✗	✓	✗	✗	✗	✗
Mean Intensity	✗	✓	✗	✗	✗	✗	✗
Mean Pitch	✓	✓	✓	✓	✗	✗	✗

Distance between A-B Center of the upper lip	✓	✓	✓	✓	✗	✗	✗
Distance between C-D Lower lip region	✓	✓	✓	✓	✗	✗	✗
Distance between E-F Horizontal distance near the mouth area	✓	✓	✓	✓	✗	✗	✗
Distance between G-H Cheek-to-cheek width	✗	✓	✗	✗	✗	✗	✗
Distance between I-J Chin to jawline	✗	✗	✗	✗	✗	✗	✗
Distance between H-K, Width between the sides of the jaw	✗	✗	✓	✗	✗	✗	✗
Distance between G-L Upper cheek points	✗	✗	✗	✗	✗	✗	✗

Distance between M-N Outer corners of the eyes	✗	✗	✓	✗	✗	✗	✗
Distance between O-P, Inner corners of the eyes	✗	✗	✓	✗	✗	✗	✗
Distance between J-Q Chin to the mouth	✗	✗	✓	✗	✗	✗	✗
Distance between R-S Horizontal distance between eyebrows	✗	✗	✓	✗	✗	✗	✗

✓- p< 0.05 ✗ - p> 0.05

About the main effects of vowel, expression, gender, and their interactions (vowel x gender, gender x expression, vowel x pitch, facial geometric features, distances between facial landmarks), this table 4 summarises the statistical relevance of various acoustic (mean formant frequencies, intensity, pitch) and facial geometric features. Whereas a cross (✗) indicates no significant effect (✗) - p> 0.05, a checkmark (✓) denotes a statistically significant effect (✓)- p< 0.05. With some interaction effects, acoustic elements, including mean F1 to F5 and mean pitch, show a firm reliance on vowel and gender. While distances involving the eyebrows and chin show less notable interactions, facial features—especially distances between lips and jaw points—are more sensitive to expression and vowel.
Providing insights on multimodal communication analysis, the table shows how both vocal and facial traits contribute differently to distinguish vowels, emotions, and gender.

4. Discussion

Beyond mere intellectual curiosity, this study has ramifications. Detecting emotions through facial expressions can aid forensic science in psychological profiling, the accuracy of witness testimony analysis, and the detection of deceit. Understanding the physiological expressions of emotions in speech can help individuals gain a deeper understanding of legal situations, particularly in cases where nonverbal cues are crucial.
The study reveals notable variations in the locations of facial landmarks associated with the pronunciation of certain vowels between male and female speakers. The ANOVA results, which show multiple comparisons across vowel groups produced statistically significant F-values and p-values ($p < 0.05$), therefore demonstrating the existence of

systematic variation. Especially, the mean variations in landmark placements across vowel sounds were somewhat significant. For instance, the highly significant ($p < 0.001$) difference between the articulation of /uː/ and /oː/ suggests that vowel articulation influences facial landmark movement patterns. These results highlight how facial kinematics during speech depend on the properties of vowels. Moreover, the results confirm the use of facial landmark analysis as a reliable technique for identifying and evaluating both emotional and phonetic effects on speech production. The method presents a potential way to advance multimodal speech analysis, as it can effectively capture notable alterations across several vowels, thereby reflecting its sensitivity to small changes in facial expressions and articulatory motions.

Multimodal emotion recognition has made significant progress, but several issues remain to be addressed before it can be effectively applied in real-world settings, particularly in fields that involve handling sensitive data, such as forensics. It's essential to consider cultural differences because people from different countries often express their emotions in distinct ways, which can lead to misinterpretation of meaning. Although this study only examines Indian individuals, future research could explore parallels between cultures. The effectiveness of detection systems can be influenced by factors such as gender, age, and attitude, which affect how people express their emotions. Additionally, usability challenges such as poor lighting or partially obscured faces can compromise the system's accuracy in tracking facial movements and emotions. At first glance, these tools appear simple, yet they often involve significant complexity. In addition to privacy concerns, ethical considerations are vital. Securing proper authorisation and handling sensitive emotional data with care is particularly crucial in forensic contexts. To develop emotion recognition systems that are both accurate and empathetic, it is essential to evaluate all relevant individual, social, ecological, and ethical considerations.

5. Conclusion

Together with their interactions, the ANOVA analysis results shed light on the relevance of several acoustic and facial data concerning the variables of vowel, expression, and gender.

For the primary impacts of vowel and gender, all formant frequencies (Mean F1 to F5) and Mean Pitch displayed statistically significant effects ($p < 0.05$), among the acoustic characteristics; expressiveness was substantial only for F1 and F5. Especially, Mean Intensity showed no appreciable changes in any one component or interaction. Using the three primary effects—vowel, expression, and gender—the mean pitch was modified and shown to be sensitive to both articulatory and speaker-related aspects. Only a few two-way interactions (e.g., Vowel x Gender) approached significance, though none of the acoustic characteristics exhibited relevance for the three-way interaction (Gender x Vowel x Expression). Vowel, expression, and gender all had a significant impact on the face distance parameters—that is, distances A–B, C–D, and E–F—as well as the vowel × gender interaction. More peripheral facial distances, such as those involving the chin-to-mouth area (J–Q) and the outer and inner eye corners (M–N, O–P), however, mainly failed to reach statistical relevance across most scenarios. Gender alone clearly affected the horizontal distance between the eyebrows (R–S). Furthermore, several two-way interactions—especially those involving expression as a component—e.g., gender × expression, vowel × expression, and the three-way interaction—suggested that expression had a less consistent influence across the measured variables compared to vowel and gender since many two-way interactions failed to show significant effects ($p > 0.05$). With emotion playing a more limited or context-dependent role, our findings generally highlight the primary influence of vowel and gender in altering both acoustic and facial traits during

speech production. These findings provide a more comprehensive understanding of how emotions are expressed and perceived through both vocal and visual modalities, with significant academic implications for various fields, including psychology, linguistics, and affective computing. This work highlights the considerable potential for integrating affective computing and facial landmark analysis into forensic investigations. Establishing a trustworthy approach for emotion detection will help strengthen the dependability and accuracy of forensic studies, thereby supporting more accurate and perceptive results in forensic science.

6. FUTURE SCOPE

Face landmark analysis could be linked with other biometric data, such as voice tone and pulse rate, in further stages of the research. This multimodal method could enhance the accuracy and robustness of emotional identification, thereby providing a more comprehensive understanding of how emotions influence speech. Future studies should focus on developing robust models that can adapt to such variability, potentially through domain adaptation methods and transfer learning. Furthermore, improving the generalizability of these systems will involve widening the datasets to include a broader spectrum of emotional displays and demographic diversity. The study also advises conducting longitudinal studies to investigate how emotional factors influence facial expressions and speech over time or under various circumstances, such as high-stress versus low-stress surroundings. This can offer a closer understanding of the consistency and stability of emotional reactions. Future studies might include more face landmarks to increase the study's level of precision. This would enable a closer analysis of facial expressions and speech patterns, potentially revealing more subtle emotional signals.

One of the high-potential approaches to expanding our work is the integration of emotional signals with scene-aware behavioral modeling. Recent research, as represented by the paper entitled "Scene-aware Human Pose Generation using Transformer" (Yao, Chen, Niu, & Sheng, 2023) introduces a transformer-based solution to generating human poses sensitive to contextual information in a scene. Through the use of attention mechanisms and representative pose template capabilities, the system successfully maps human behavior into environmental semantics. These developments have wide-ranging implications for forensic simulation enabling emotionally and spatially consistent reenactment of suspect and witness behavior. The integration of this work with our multimodal emotion recognition system might enable the development of holistic frameworks that not only account for the expression of emotions but also anticipate the way that individuals might physically react in particular situations. The integration of emotion detection and pose synthesis has the potential to contribute in significant ways to intelligent surveillance, forensic reenactment, human-robot interaction, and immersive content creation.

References

1. Albanie, S., Nagrani, A., Vedaldi, A., & Zisserman, A. (2018). Emotion recognition in speech using cross-modal transfer in the wild. *Proceedings of the 26th ACM International Conference on Multimedia*, 292–301. https://doi.org/10.1145/3240508.3240578
2. Bellenger D, Chen M, Xu Z. Facial emotion recognition with a reduced feature set for video game and metaverse avatars. *Comput Anim Virtual Worlds*. 2024; 35(2):e2230. https://doi.org/10.1002/cav.2230

3. Chang, H.-S., Lee, C.-Y., Wang, X., Young, S.-T., Li, C.-H., & Chu, W.-C. (2023). Emotional tones of voice affect the acoustics and perception of Mandarin tones. *PLOS ONE, 18*(4), e0283635. https://doi.org/10.1371/journal.pone.0283635
4. Ekman, P., & Friesen, W. V. (1978). *Facial action coding system*. Consulting Psychologists Press. https://books.google.co.in/books?id=08l6wgEACAAJ
5. Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. https://doi.org/10.1145/1873951.1874246
6. Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review, 27*(2). https://doi.org/10.3758/s13423-019-01701-x
7. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(2), 1359–1367. https://doi.org/10.1609/aaai.v34i02.5492
8. Optimization of 2D and 3D facial recognition through the fusion of CBAM AlexNet and ResNeXt models. (2025). *ResearchGate*. Retrieved June 30, 2025, from https://www.researchgate.net/publication/386507996_Optimization_of_2D_and_3D_facial_recognition_through_the_fusion_of_CBAM_AlexNet_and_ResNeXt_models
9. Ristea, N.-C., Duţu, L. C., & Radoi, A. (2019). Emotion recognition system from speech and visual information based on convolutional neural networks. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 1–6. https://doi.org/10.1109/SPED.2019.8906538
10. Salas-Cáceres, J., Lorenzo-Navarro, J., Freire-Obregón, D., & Castrillón-Santana, M. (2024). Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-20227-6
11. Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters, 146*, 1–7. https://doi.org/10.1016/j.patrec.2021.03.007
12. Shanthi, P., & Nickolas, S. (2022). Facial landmark detection and geometric feature-based emotion recognition. *International Journal of Biometrics*. https://www.inderscienceonline.com/doi/10.1504/IJBM.2022.121799
13. Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., & Singh, S. (2023). Facial expression recognition in videos using hybrid CNN & ConvLSTM. *International Journal of Information Technology, 15*(4), 1819–1830. https://doi.org/10.1007/s41870-023-01183-0
14. Wang, J. Z., et al. (2023). Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion. *Proceedings of the IEEE, 111*(10), 1236–1286. https://doi.org/10.1109/JPROC.2023.3273517
15. Wang, Y., et al. (2023). A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials, 25*(1), 319–352. https://doi.org/10.1109/COMST.2022.3202047
16. Xiong, K., Qing, L., Li, L. *et al.* Facial expression recognition based on local–global information reasoning and spatial distribution of landmark features. *Vis Comput* **41**, 535–548 (2025). https://doi.org/10.1007/s00371-024-03345-y

17. Yan, J., et al. (2024). Multimodal emotion recognition based on facial expressions, speech, and body gestures. *Electronics, 13*(18), Article 18. https://doi.org/10.3390/electronics13183756
18. Yao, J., Chen, J., Niu, L., & Sheng, B. (2023, August 4). Scene-aware human pose generation using Transformer. *arXiv*. https://doi.org/10.48550/arXiv.2308.02177