iJRASET
International Journal For Research in
Applied Science and Engineering Technology

Volume: 13 Issue: X Month of publication: October 2025

DOI: https://doi.org/10.22214/ijraset.2025.74760

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue X Oct 2025- Available at www.ijraset.com

AI-Powered Interview Stress and Confidence Analyzer

Vaibhav Thakare[1], Aditya Bhujade[2], Harsh Umredkar[3], Dr. Kapil Gupta[4]
Computer Engineering, St. Vincent Pallotti College of Engineering and Technology

Abstract: In the contemporary recruitment environment, the evaluation of a candidate's stress and confidence remains highly subjective. This paper proposes an intelligent system, the AI-Powered Interview Stress and Confidence Analyzer, designed to provide objective metrics by analyzing both facial expressions and vocal biomarkers. While the system architecture includes both modalities, the core contribution of this paper is a robust methodology for the vocal analysis component, which addresses the critical limitations of existing approaches that rely on acted emotional data. Our proposed solution for the Face module centers on creating a novel, context-specific dataset of authentic interview audio. We detail the complete system's architecture, including the system architecture employs a Convolutional Neural Network (CNN) to interpret facial emotions. For the vocal analysis component, the framework is designed to use Mel-Frequency Cepstral Coefficients (MFCCs), a feature set that has proven effective for classifying stress and emotion in speech. This paper outlines a clear path toward developing a comprehensive, non-intrusive tool to complement traditional interview procedures, providing holistic, data-driven insights for recruiters and candidates.

Keywords: Voice Stress Analysis, Confidence Detection, Deep Learning, Convolutional Neural Network (CNN), Mel-Frequency Cepstral Coefficients (MFCCs), Interview Analytics.

I. INTRODUCTION

The job interview is a critical juncture in an individual's career path, serving as the primary mechanism for employers to assess a candidate's suitability for a role. Beyond technical qualifications, interviewers evaluate soft skills, communication abilities, and the candidate's demeanor under pressure. However, the assessment of psychological states such as stress and confidence is often subjective, inconsistent, and susceptible to the inherent biases of the interviewer. The physiological and psychological stress experienced by a candidate during an interview can significantly impact their performance, yet there are few objective tools to measure it. This lack of objective measurement creates a significant challenge in ensuring fair and equitable evaluation. We divided our project into two parts as facial expression and vocal modulation. In first part we delt with the facial expression through 10 emotions (happy, sad, proud, anger, neutral, etc)[7][8]. In which we used CNN model of deep learning which have 3 layers. In the end, we calculate the confidence and stress based on emotions. The human voice is a rich, non-invasive source of information that reveals the internal psycho-physiological state of a speaker. Research has consistently shown that psychological stress induces measurable acoustic modifications in vocal patterns. The presence of psychological stress induces tangible changes in a person's speech, altering key vocal metrics such as fundamental frequency (F0), perturbations in frequency (jitter) and amplitude (shimmer), the overall speech rate, and other spectral qualities. Early studies in voice stress analysis (VSA) focused on detecting deception by analyzing micro-tremors in the vocal cords, establishing a direct relationship between emotional stress and the correctness of answers in testing scenarios. These foundational concepts have paved the way for more advanced, automated analysis using machine learning.[3][4]

Recent advancements in deep learning have revolutionized the field of speech analysis. Modern approaches often utilize Convolutional Neural Networks (CNNs) to classify emotions and stress with high accuracy.[8][9][11] Studies have demonstrated the efficacy of using spectral features like Mel-Frequency Cepstral Coefficients (MFCCs) as input for these models. For instance, Gupta et al. successfully employed a CNN with Mel Spectrograms from the RAVDESS dataset to recognize emotions, while Zainal et al. developed a specialized CNN architecture (SSNNA) that achieved 93.9% accuracy in classifying stress in female voices using MFCCs. These works underscore the potential of deep learning to create robust, speaker-independent models for stress detection.[1][2][5] While many systems focus broadly on stress or emotion, there is a clear opportunity to develop a specialized tool tailored to the unique context of job interviews, analyzing both stress and the often-overlooked dimension of confidence.

This paper proposes an AI-Powered Interview Stress and Confidence Analyzer that focuses on vocal inputs. Our primary objective is to develop and evaluate a deep learning model that can:

1) Extract salient spectral features, specifically MFCCs, from audio recordings of interview responses.
2) Utilize a Convolutional Neural Network (CNN) to classify the speaker's perceived level of stress and confidence.
3) Provide a framework for an objective, data-driven tool that can supplement the traditional interview process, offering valuable feedback for recruiters and aiding candidates in their professional development.

By focusing on the interview context, this work aims to bridge the gap between general-purpose emotion recognizers and the specific need for objective assessment tools in human resources.

II. METHODOLOGY

This section explains the process followed to detect facial emotions and map them to confidence levels, divided into subsections for clarity.

A. Dataset

The methodology for this research was conducted in two phases. The first phase involved developing and integrating a model for facial emotion analysis. The second phase of this work involved a critical analysis of existing approaches to vocal stress detection, which informed the proposal of a new methodology centered on a custom-collected, context-specific dataset.[7]

1) Facial Expression Analysis Component

This component focuses on detecting facial emotions from video and mapping them to confidence levels.[7]

Dataset and Preprocessing

The Facial detection model was developed using two datasets:

- FER-2013: A public dataset with ~25,000 grayscale images across 7 emotion labels, commonly used in emotion recognition research. [7][9][11]
- Custom Web-Scraped Dataset: An additional ~25,000 images were collected through web scarping , covering 10 emotions (Angry, Determined, Disgust, Excited, Fear, Happy, Neutral, Surprise, Sad, and Proud). These images were manually cleaned to remove irrelevant or mislabeled samples.

The following preprocessing steps were applied:

- Resizing: All images were resized to 48×48 pixels.
- Grayscale Conversion: Images were converted to grayscale to reduce computational load and focus on facial structures.
- Normalization: Pixel values were scaled from a 0–255 range to 0–1.
- Data Splitting: The dataset was divided into 80% for training and 20% for testing.

2) Model Architecture and Training

A Convolutional Neural Network (CNN) was implemented from scratch. The architecture included **three** convolutional layers (32-64-128) with ReLU activation, pooling layers, dropout layers for preventing overfitting, fully connected layers, and a Softmax output layer for classification.[8][9][11] The model was trained using the Adam optimizer and categorical cross-entropy as the loss function.

3) Application Integration

The trained model was integrated into a real-time video analysis pipeline that uses Haar cascades for face detection. Detected faces are preprocessed and passed to the model , and the predicted emotions are mapped to confidence levels (Confident, Underconfident, Neutral).[8][11] The system outputs annotated videos and CSV logs with timestamps.

B. Vocal Modulation Analysis Component

This component critiques existing methods for vocal stress analysis and proposes a more robust methodology.

1) Initial Exploratory Models and Their Limitations

Our initial investigation attempted to build a classifier using a combination of public emotional speech datasets (RAVDESS, CREMA-D, TESS, Emo-DB). We experimented with several models:

- LSTM Model: An LSTM was trained on features like Mel-Frequency Cepstral Coefficients (MFCCs) and their derivatives. While it performed well on training data, it overfitted and failed to generalize to new voices.
- DAIC-WOZ Dataset: An attempt to train an LSTM using PHQ-8 scores from the DAIC-WOZ dataset also failed to yield the desired classification performance.
- XGBoost with eGeMAPS: A powerful XGBoost model using the comprehensive eGeMAPS feature set also did not achieve the desired output.

2) Challenges with Staged Emotional Datasets

The experiments consistently showed that models trained on public datasets could not reliably classify stress in real-world cases. The primary reason is that these datasets consist of speech from actors portraying emotions. This "acted stress" is often exaggerated and lacks the subtlety of genuine psychological stress found in a high-stakes interview, preventing the model from generalizing effectively.

3) Proposed Methodology: Custom Interview Dataset

Based on these limitations, we propose a new methodology centered on a novel, context-specific dataset. **Data Acquisition**: We are collecting a custom dataset by conducting mock interviews with **student** participants in a controlled laboratory environment to capture authentic vocal stress and confidence markers.

4) Data Annotation

Mock interviews with student participants in a controlled laboratory environment to capture authentic vocal stress and confidence markers.

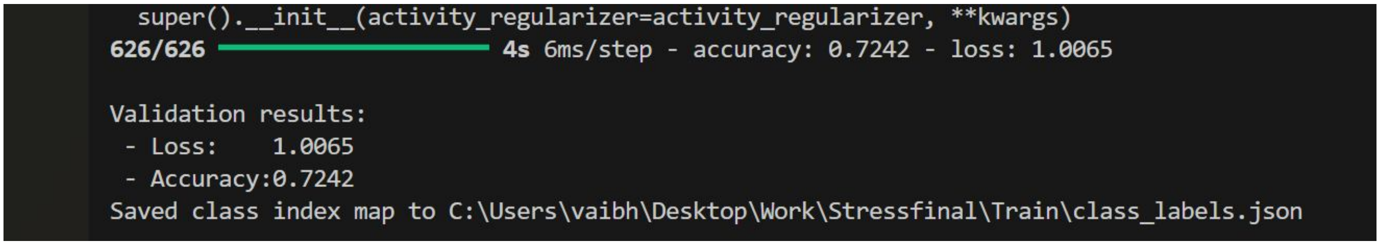
- Data Annotation: Each participant's interview recording will be segmented into a single question and answer format. Each question's recording will be manually annotated for 'Stressed' or 'Confident' levels by a panel of trained annotators using a predefined rubric
- Feature Extraction: We will use comprehensive feature sets, primarily the eGeMAPS (88 acoustic features) and ComParE (over 6,000 features), to create a detailed representation of the speech signal.
- Model Selection and Training: The extracted features will be used to train and evaluate several models, including XGBoost and LSTM, to find the best architecture for this task. Performance will be measured using Accuracy, Precision, Recall, and F1-Score on an 80/20 train-test split.

Architecture for this specific task. The data will be split into training (80%) and testing (20%) sets, and performance will be measured using Accuracy, Precision, Recall, and F1-Score.

III.RESULT

Our facial expression analysis model, a Convolutional Neural Network, was tested on a set containing 10,005 images. A review of the **40-**epoch training cycle revealed that the model overfitted the training data. On the test data, the model achieved an overall accuracy of **72%.** Performance varied significantly across categories(Fig 1); it was most effective at identifying '**Happy**' (0.55 F1-score) and **'Surprise'** (0.46 F1-score), but had considerable difficulty with classes like 'Determined' and 'Proud', which scored F1-scores of just 0.07 and 0.19, respectively.(Fig-1,2)

A detailed breakdown of the model's predictions is shown in the Fig. 2 confusion matrix . This highlights which classes were most frequently confused with one another; for instance, 'Sad' was often misclassified as 'Disgust', *'Neutral', or 'Angry'.*



```
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
626/626 ━━━━━━━━━━━━━━━━━━━━ 4s 6ms/step - accuracy: 0.7242 - loss: 1.0065

Validation results:
 - Loss:    1.0065
 - Accuracy:0.7242
Saved class index map to C:\Users\vaibh\Desktop\Work\Stressfinal\Train\class_labels.json
```

Fig .1.-Terminal Picture for Accuracy

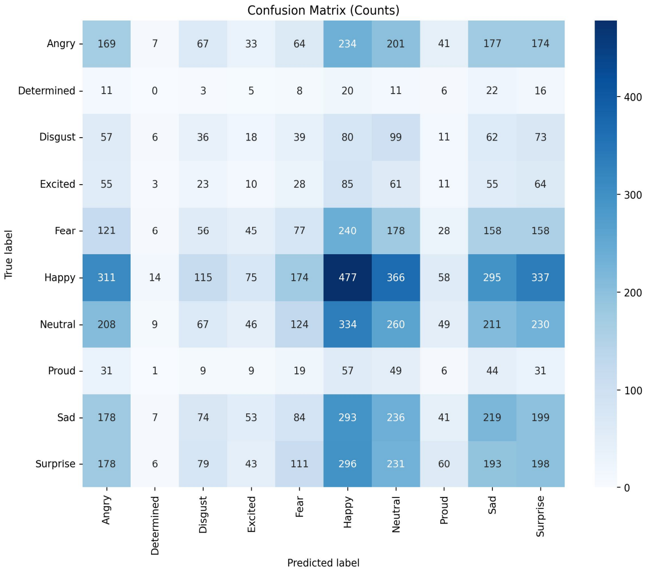


Fig .2.

Confusion matrix of model predictions on the test dataset(Counts)

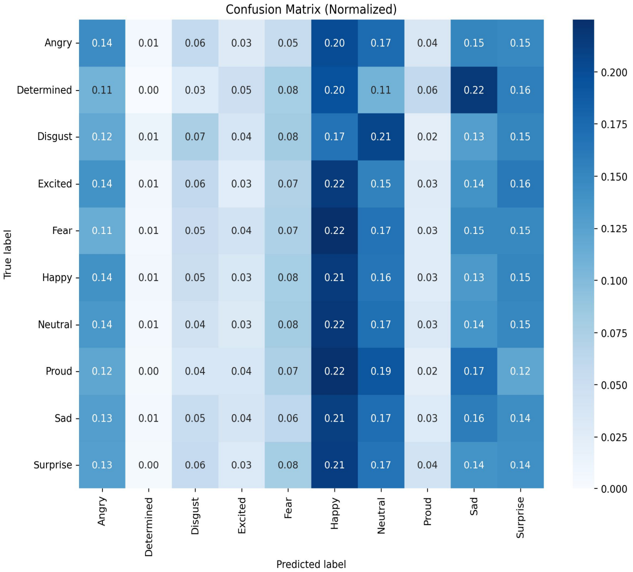


Fig .3.

Confusion matrix of model predictions on the test dataset(Normalized)

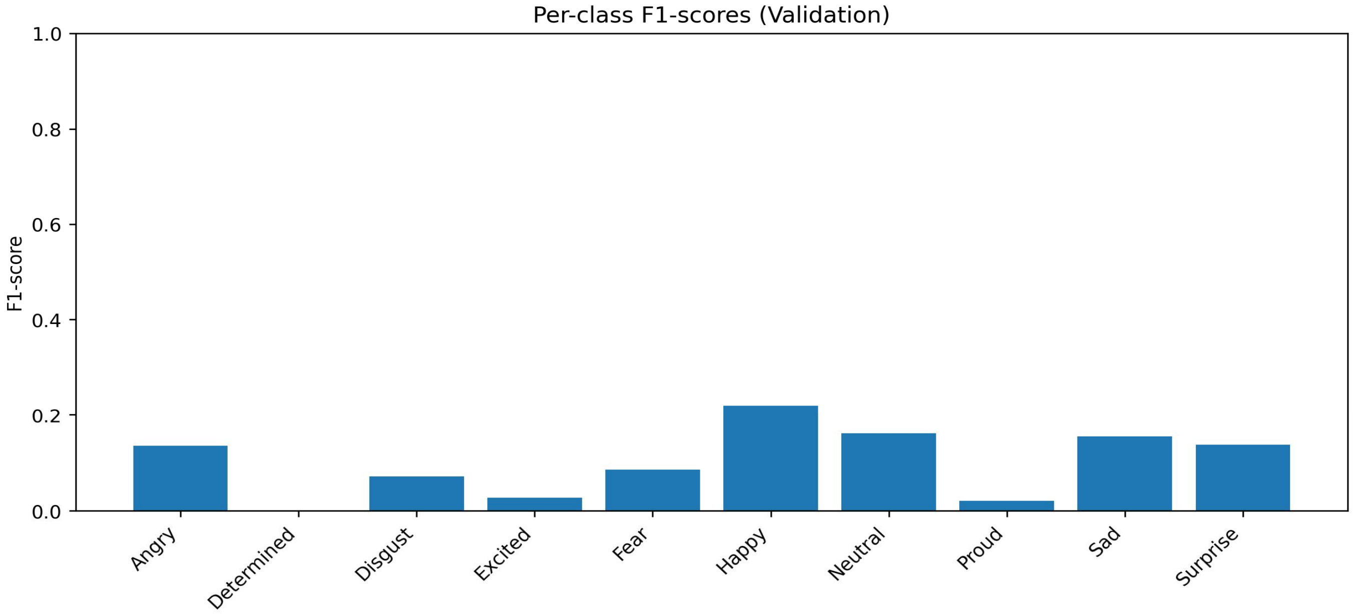


Fig.4.- F-1 Score of all ten emotions

IV.CONCLUSIONS

This paper detailed the development of an AI-powered analyzer for stress and confidence. As presented in the results, the facial expression recognition module achieved a modest accuracy of **72%.** The key finding from its evaluation was the significant overfitting observed during training, which limits the model's ability to generalize to new, unseen data. This establishes a functional but limited baseline for facial analysis.

The Reamaning contribution of this work remains the proposed methodology for the vocal analysis component. Our initial findings confirmed that standard emotional datasets are insufficient for this task, validating the need to create a novel, context-specific dataset from mock interviews.

Future work will focus on improving collect more dataset for the vocal model and improve the accuracy of vocal model. The long-term objective is to fuse both modalities, creating a robust and reliable tool to bring data-driven objectivity to the interview process.

FUTURE WORK AND EXPECTED OUTCOMES

The immediate next step is to complete the data acquisition phase by conducting the planned mock interviews. Following this, we will execute the manual annotation process to label the collected recordings for stress and confidence. Once the dataset is fully prepared and labeled, the feature extraction process will begin, utilizing the eGeMAPS and ComParE feature sets.

Subsequently, the machine learning models detailed in the methodology, including XGBoost and LSTM, will be trained and validated. Performance will be rigorously measured using standard metrics such as Accuracy, Precision, Recall, and F1-Score to identify the most effective model.

We expect that the models trained on our custom dataset will demonstrate significantly better generalization and performance on real-world interview audio compared to models trained on acted data. The primary expected outcome is a validated methodology and a robust classifier capable of providing an objective, data-driven assessment of stress and confidence. For the broader system, future work will also focus on enhancing the facial expression model. This includes exploring dataset balancing,

Furthermore, we recognize the inherent limitations of relying solely on classifying basic emotions to determine complex states like stress. Recent multi-modal research has shown that in controlled stress-inducing scenarios, facial expressions do not always group into the classic emotion categories, and that facial muscle movements alone may not be reliable predictors of physiological stress responses. This underscores the importance of our dual-modal approach. By eventually fusing vocal and facial data, we aim to create a more robust and nuanced classifier that addresses the limitations of a single-modality system, as advocated for by current research.[12] Data augmentation, and transfer learning to further improve its performance.

The ultimate long-term goal is to investigate fusion techniques. We plan to explore both feature-level and decision-level fusion of the vocal and facial model outputs. This will allow us to create a more comprehensive and accurate classifier that leverages the strengths of both modalities to provide a single, holistic assessment of a candidate's stress and confidence.

V. ACKNOWLEDGMENT

The authors would like to thank the student participants who volunteered their time for the mock interview sessions. We also extend our gratitude to the faculty and staff at ST. vincent Pallotti college of engineering and technology, Computer Engineering branch for their invaluable support, guidance, and for providing the necessary resources and facilities to conduct this research.

REFERENCES

[1] Gupta, S., Gambhir, S., Gambhir, M., Majumdar, R., Shrivastava, A.K., and Pham, H. 2025. A deep learning approach to analyse stress by using voice and body posture. Soft Computing. 29 (2025), 1719-1745.

[2] Zainal, N.A., Asnawi, A.L., Ibrahim, S.N., Azmin, N.F.M., Harum, N., and Zin, N.M. 2025. Utilizing MFCCS and TEO-MFCCS to classify stress in females using SSNNA. IIUM Engineering Journal. 26, 1 (2025), 324-335.

[3] Kaklauskas, A., Vlasenko, A., Seniut, M., and Krutinis, M. 2009. Voice Stress Analyser System for E-Testing. In Proceedings of the 2009 Ninth IEEE International Conference on Advanced Learning Technologies. 693-695.

[4] Sondhi, S., Vijay, R., Khan, M., and Salhan, A.K. 2016. Voice Analysis for Detection of Deception. In Proceedings of the 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS).

[5] Sandulescu, V., Andrews, S., Ellis, D., Dobrescu, R., and Martinez-Mozos, O. 2015. Mobile App for Stress Monitoring using Voice Features. In Proceedings of the 5th IEEE International Conference on E-Health and Bioengineering (EHB 2015).

[6] Chidaravalli, S., Jayadev, N., Divyashree, P., Yadav, G.A., and Prajwal, B. 2022. Stress and Anxiety Detection through Speech Recognition and Facial Cues using Deep Neural Network. International Journal of Innovative Research in Technology (IJIRT). 9, 2 (2022), 1040-1044.

[7] Sharmila Chidaravalli1, Namratha Jayadev2, Divyashree P3, Ghanavi Yadav A4, Prajwal B5 1,2,3,4,5 Stress and Anxiety Detection through Speech Recognition and Facial Cues using Deep Neural Network Dept. of Information Science & Engg., Global Academy of Technology, Bangalore, India

[8] Almeida, J. and Rodrigues, F. 2021. Facial Expression Recognition System for Stress Detection with Deep Learning. In Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021). 1 (2021), 256-263

[9] Bhagat, D., Vakil, A., Gupta, R.K., and Kumar, A. 2024. Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). Procedia Computer Science. 235 (2024), 2079-2089.

[10] Ismail, N. 2017. Analysing Qualitative Data Using Facial Expressions in an Educational Scenario. International Journal of Quantitative and Qualitative Research Methods. 5, 3 (2017), 37-50.

[11] Kumar, G.S., Cheriyan, J., Aparna, N., and Swathy, J. 2025. Unleashing Facial Expression Recognition for Stress Detection Using Deep CNN Model. Procedia Computer Science. 259 (2025), 306-315.

[12] Ringgold, V., Burkhardt, F., Abel, L., Kurz, M., Müller, V., Richer, R., Eskofier, B.M., Shields, G.S., and Rohleder, N. 2025. Multimodal stress assessment: Connecting task-related changes in self-reported stress, salivary biomarkers, heart rate, and facial expressions in the context of the stress response to the Trier Social Stress Test. Psychoneuroendocrinology. 180 (2025), 107560.