INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 04 | APRIL - 2023 IMPACT FACTOR: 8.176 ISSN: 2582-3930

AI Based Multimodal Emotion and Behavior Analysis of Interviewee

Aaditya Jadhav[1], Rushikesh Ghodake[2], Karthik Muralidharan[3] , G.Tarun Varma[4], Prof. Vijaya Bharathi J.[5]

[1-4]Department of Computer Engineering & Pillai College of Engineering, Navi Mumbai, India
[5]Department of Computer Engineering & Pillai College of Engineering, Navi Mumbai, India

---

Abstract - The COVID-19 epidemic has recently increased the popularity of virtual interviews, and globalization and technology have made them a popular option for hiring. Virtual interviews, however, provide more difficulties and problems for both interviewers and interviewees than conventional face-to-face interviews. The difficulty in comprehending the interviewee's behavioral features is one of the main issues with virtual interviews. There is a suggestion for a machine learning and deep learning-based method to detect and examine changes in the interviewee's behavior and personality features in order to address this problem. This strategy might remove interviewer bias and offer a more unbiased evaluation of interviewees. We offer a computational framework that counts the interviewee's communication-related performance and provides performance feedback based on the analysis of multimodal data like voice and facial expressions. Speech to text API is used to separate a video taken during the interview into audio and visual frames, as well as to extract text from the audio. The face is recognized from the visual frames, and emotions are assessed. Using machine learning and deep learning techniques, facial expressions are categorized as happy, fearful, sad, neutral, surprised, etc. Similar to this, a candidate's speech fluency is evaluated based on audio cues. The emotions of a candidate are discovered from the text. emotion, a candidate's speaking fluency, and the text's sentiment.

Key Words: Virtual Interview, Visual Frame, Facial Expression, Audio Cue, Speech fluency, Sentiment.

1.INTRODUCTION

Emotional recognition is the process of identifying and understanding the emotions that people feel in response to various circumstances that can vary from person to person. Accurately reading these emotions is critical to good communication, as emotions play an important role in determining human behavior. People can express their feelings verbally and non-verbally, including through voice, body language, facial expressions and many other ways. Recognizing these signals is critical to social engagement and bonding. To predict a person's emotional state, emotion perception involves analyzing emotional information from multiple sources. However, relying on one emotion recognition method can be difficult to accurately assess a person's emotions. For example, simply looking at an object or event is not enough to know a person's emotional state. Therefore, it is important to approach emotion detection as a multimodal challenge that takes into account various emotional indicators, including those present in body language, facial expressions and tone of voice. This allows for a deeper understanding of a person's emotional state and more accurate emotional predictions.

Facial gestures are important ways to express emotions in nonverbal communication. The ability to recognize facial expressions is crucial in numerous fields, such as healthcare and human-computer interaction. We observe that 7% of knowledge is exchanged between people in writing, 38% by voice, and 55% by facial expressions. There are six basic emotions: happiness, sadness, surprise, fear, disgust, and anger. There is evidence that people feel these emotions regardless of their culture. Emotions can be expressed in two orthogonal dimensions: valence and arousal. Speech expression recognition is one of the key elements of the human-machine interface system. They will convey their feelings by voice and face. Speech recognition systems are widely used for emotion detection. Early experiments on emotion recognition in speech contemplated manual extraction of features from speech for classification purposes. Sentiment analysis from the text is performed to determine the subjective information it contains and to understand the opinions, sentiments, and emotions emanating from the text. Typically, this process involves any textual data source to determine the essential information.

2.LITERATURE SURVEY

Here relevant techniques in literature are reviewed. It describes various techniques used in the work. The brief information about the referred research papers is explained in this section.

Literature Review

In paper [1], a simulated speech platform that can analyze visual, audio and text is proposed. More than one hundred

Chinese speakers from various professional backgrounds have been recognized. Automatic Dependency Detection (ARD) is used for video processing, linear regression for audio processing, and ARD for text analysis. The platform provides comprehensive analytics based on available features.

In paper [2], a web application is designed for recruiters and applicants to walk nicely through all the steps that need to be taken. A custom dataset of nearly one hundred images was collected for video analysis and Mozilla's deep speech-to-text query dataset. Deepface, faceAPI for image processing and short-term temporal (LSTM) for text. Experiments show that the proposed model produces the results in the case of two documents.

In paper [3], communication plans and facial expressions show good performance. IEMOCAP data was used. CNN is used for image processing and LSTM is used for speech processing. In addition, many small cores are used to replace large core convolutions, which reduces learning loss.

In paper [4], the job offer analyzes the candidates' performance in the interview. They store speech data and speech library data for audio, and FER2013, CK+ datasets for video. MFCC, logistic regression for audio processing, Haar cascade for image processing, Gabor filter. They provide a system by which respondents can be scored. Candidate's behaviors, facial expressions, eye contact, reactions can be analyzed in future studies.

In paper [5], the proposed function provides an automatic call evaluation from audio and video cues to evaluate the behavior of the interviewees. Libri Speech and student databases are used for audio, and CK and JAFFE datasets are used for video. It provides feedback on successful candidates by analyzing various indicators such as facial expressions and speech. Feedback offers candidates a variety of possibilities.

In paper [6], the proposed cnn model is very effective. During real-time detection, the model lacks robustness. FER 2013 dataset is used for videos. CNNs are used to process images. It can use other big data and use different models to increase the accuracy of the model.

In paper [7], the concept of neural network is effective. They manage the EmoDB dataset they use. MFCC, feedforward, neural network for voice feature extraction and speech processing. As mentioned the performance is good but it takes a while when it's done.

In paper [8], various speech perception classification models are used. A general speech database for audio. MFCC, Random Forest, SVM, MLP, KNN are used to remove noise and speech. In the research, it has been shown that the random forest is a good classifier.

In paper [9] provides a method for analyzing and separating the views of 3 rounds of discussion to 0.7189 (F1score) is higher than other samples. It might be good to continue the mixed method using emotional expressions and emoji gestures.

In paper [10], the proposed model is a combination of deep learning and machine learning methods. According to ML classifiers, SVM gives the highest accuracy with 78.97%. Among the DL methods, the Bi-GRU model reached the highest accuracy with 79.46%.

3.PROPOSED WORK

The proposed work consists of three main phases: emotion recognition from facial expressions, speech emotion recognition from audio cues, and sentiment analysis from text. It is important to analyze different modalities in order to get the best results and output. As per interview analysis is considered, the primary part is emotion recognition from the images frames followed by speech recognition from audio cues and sentiment analysis from the text.

Figure 3.1 represents the proposed architecture of the system. The recorded video of an interview is input for the proposed system. From the input video the image frames are extracted and passed towards the preprocessing step. Audio cues are extracted from the input video in order to analyze the emotion from the audio cues with the help of machine learning and deep learning techniques, Further the text is generated from the audio cues for the sentiment analysis purpose. Image frames, audio cues and generated text has been passed to a preprocessing followed by application of machine learning and deep learning model respectively. The performance score is provided by considering the results of all the models.

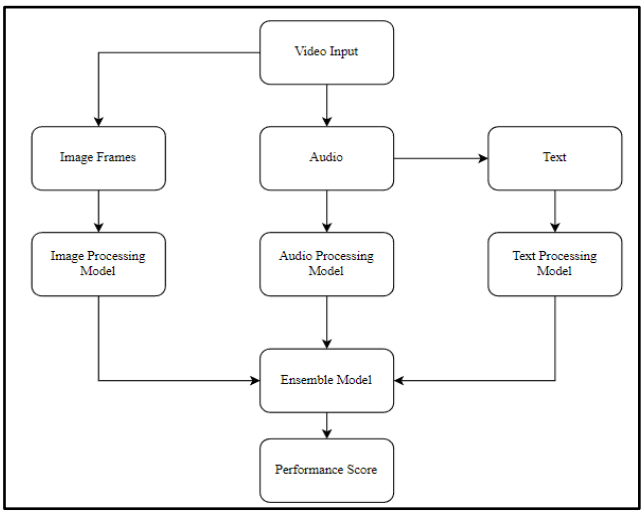

Fig. 3.1 Proposed Architecture

A. Emotion Recognition from Image Frames

The proposed work for emotion detection from image frames are preceded as follows:

a. Convert the intake videotape into a piece of frames. A crucial frame is uprooted from the sequence of images.
b. Describe the pollee face from the videotape frames.

c. Classify the pollee feelings by using machine literacy and deep literacy algorithms.
d. Finally, calculate the percentage of the emotion.

Image processing model shown in figure 3.2.

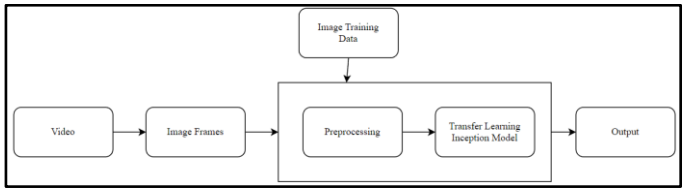

Fig. 3.2 Image Processing Model

B. Speech Recognition from Audio Tones

- **Audio Processing Model**

The proposed work of speech recognition from audio data as follows:

a. The speech of the candidate is input to the system.
b. The signals of speech are transformed into a number of frames.
c. The dynamic portions of MFCC features are regularized to form point vectors.
d. Later the point vectors are classified by using machine learning and deep learning algorithms to find fluency of the speech.

Audio processing model shown in figure 3.3.

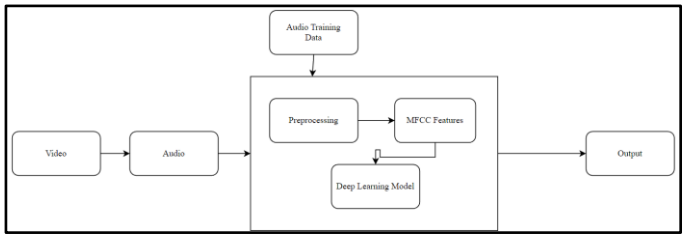

Fig. 3.3 Audio Processing Model

- **Speech Fluency:** The tone aspect has been used to calculate the fluency of the speech. Assume that the fluctuation of every respective person's tone is proportional to the maximum and minimum of its amplitude of sound, hence the normalization of the amplitude is carried out in order to calculate the speech fluency. For any person with the same fluency of speech but different in volume can have the same normalized values in terms of pitch and fluency.

C. Sentiment Analysis from Text

The proposed work for sentiment analysis from text is preceded as follows:

a. Extract the text from the audio cues.
b. Apply natural language techniques on the text data.
c. Analyze the sentiment of the text using a machine learning algorithm.

Text processing model shown in figure 3.4.

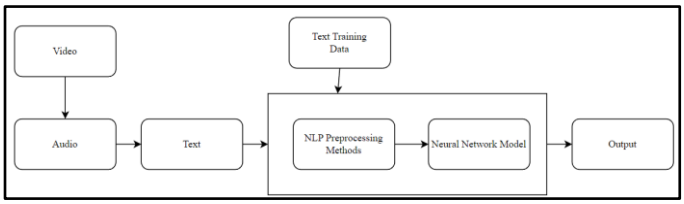

Fig. 3.4 Text Processing Model

D. Behavior Analysis

Behavior mapping based on emotional context

Behavior analysis involves various parameters that affect its class or type such as physical movement, emotional feelings, environment conditions etc. From all of them currently we had worked on analysis of emotions based on different types of aspects. Hence from all affection conditions the emotion values can be used and can be mapped to certain expected behavior values. This all worked on frequency of the emotions, all emotion frequencies get ensembled to calculate the final score of the behavior.

4.METHODOLOGY

In this section we discuss types of methods, which we followed for respective systems i.e., algorithms, techniques, to develop the proposed system.

A. Facial Emotion Recognition Techniques

- **HaarCascade Frontal Face Classifier:** The HaarCascade Frontal Face Classifier is a widely-used face detection algorithm that was developed by Paul Viola and Michael Jones [4]. Its purpose is to identify faces within an input image and provide the coordinates of each detected face. By obtaining these coordinates, it becomes possible to resize the image and subsequently analyze the emotions displayed on each face [4]. In the current project, we have utilized the HaarCascade classifier to detect faces.

- **Inception-V3 Model:** Inception-v3 is a pre-trained neural network in the ImageNet database with over one million images. The network comprises 48 layers and is capable of classifying images into 1000 different categories, ranging from common objects like keyboards and pencils to animals. Due to its extensive training on a diverse range of images, the network has developed an ability to recognize and learn rich feature representations for a variety of visual content. The model functions by extracting features from input images in the initial stage and then

classifying them based on those features in the subsequent stage.

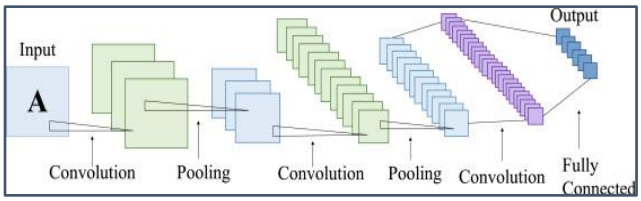

Fig. 4.1 Inception-V3 Model

B. Speech Emotion Recognition Technique

- **Mel-frequency Cepstral Coefficients (MFCC):** MFCC is a commonly used and highly effective technique for extracting features from audio signals. Cepstrum refers to the logarithmic spectrum of the signal spectrum in the time domain. Neither the frequency domain nor the time domain completely describes the generated spectrum. The Mel frequency cepstral coefficients (MFCC) are coefficients that make up the mel frequency cepstrum, which is a description of the short-term power spectrum of a sound. MFCC offers high frequency resolution while minimizing noise in the low frequency region compared to the high frequency region. It is often used as a feature for speech classification problems due to its ability to accurately represent the shape of audio signals [4]. To obtain MFCC functionality from audio signals, the PyAudio library in Python programming language is commonly used.

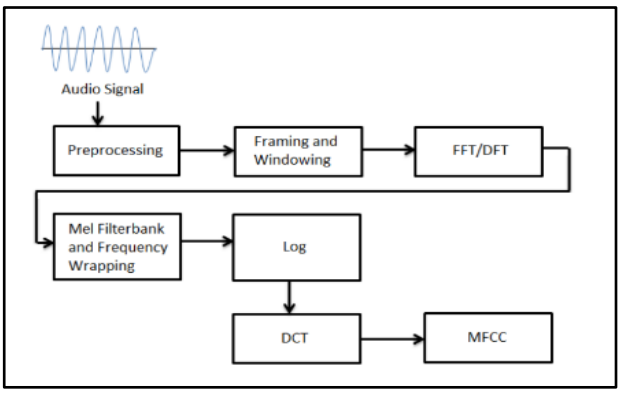

Fig. 4.2 MFCC Block Diagram

- **Long Short Term Memory (LSTM):** Long-term memory is a pivotal element of our capability to flash back and reuse information over extended ages of time. intermittent neural networks (RNNs) are a type of neural network that can effectively model successional data by using the affair of the former step as input to the current step. Still, traditional RNNs can struggle to maintain delicacy when prognosticating long- term dependencies in the data. This is where long short- term memory (LSTM) networks come by. LSTMs are a type of RNN that are specifically designed to handle time- series data like speech and textbook, and are suitable to learn and retain long-term dependencies in the data by introducing memory cells. Unlike traditional RNNs that have a single retired state passed over time, LSTMs are suitable to store information in these memory cells for longer ages of time, which helps to alleviate the problem of grade evaporation that can occur during the backpropagation process.

C. Sentiment Recognition form Text

- **Natural Languages Processing Concepts:** Text preprocessing mainly involves natural language processing concepts they are as follows:

1. **Tokenization:** Tokenization is basically unyoking an expression, judgment, paragraph, or an entire textbook document into lower units, similar to individual words or terms. Each of these lower units are called commemoratives. Tokenization can be done to either separate words or rulings. If the textbook is resolved into words using some separation fashion it's called word tokenization and the same separation done for rulings is called judgment tokenization. In the process of tokenization, some characters like punctuation marks may be discarded. Before recycling a natural language, we need to identify the words that constitute a string of characters.

2. **Filtration:** One of the key steps in processing linguistic data is removing noise so that machines can more easily detect patterns in the data. Text data contains a lot of noise, in the form of special characters such as labels, punctuation marks, and numbers. If these appear in the data, it is difficult for the computer to understand. We therefore have to process the data and delete these elements. Also, it is important to pay attention to the capitalization of words. If we include both uppercase and lowercase versions of the same word, the computer considers them different entities even though they may be the same.

3. **Stemming:** Normalization is a common technique used in natural language processing to standardize text data. One approach to normalization is stemming, which involves generating various morphological variants of a word's base or stem form to capture different tenses or grammatical variations. The aim is to reduce the number of unique forms a word can take without changing its meaning. This technique is often referred to as the "tribe" approach, and the programs or algorithms that use it are sometimes called "tribal" voters or algorithms.

- **Feature Extraction**

  In machine learning, feature extraction is an important step that involves transforming raw data into code that can be processed while preserving the important information in the original data. This is generally more effective than directly applying machine learning algorithms to raw data. One key characteristic of large datasets is the high number of variables or features that must be processed, often requiring significant computing resources to handle the complexity of the data.

- **Neural Networks**

  Neural networks are a family of algorithms designed to identify relationships in datasets through a process that mimics the way the human brain works. In this sense, a neural network refers to a system of neurons, whether organic or artificial. Neural networks can adapt to changing devices; so the network can produce good results without having to create a production model. The concept of neural networks started with artificial intelligence and quickly gained popularity in the development industry. Multilayer neural networks are called "deep" networks and are used in deep learning algorithms.

5.DATASET DESCRIPTION

Datasets are taken from the various platforms. Three types of datasets are used for this project the description of the datasets are described in the following section.

1. Image Dataset

Affectnet Emotion Images Dataset is used to train the machine literacy and deep literacy models. AffectNet is a large facial expression dataset with around0.4 million images manually labeled for the presence of eight (neutral, happy, angry, sad, fear, surprise, nausea, disdain) facial expressions along with the intensity of valence and thrill. AffectNet is by far the largest database of facial expression, valence, and thrill in the wild enabling exploration in automated facial expression recognition in two different emotion models.

2. Audio Dataset

The Toronto Emotion Speech set has been used. It contains stimulants. These stimulants were modeled on the Northwestern University Auditory TestNo. 6 (NU- 6; Tillman & Carhart, 1966). There are 2800 stimulants in aggregate. A set of 200 target words were spoken in the carrier expression by two actresses (aged 26 and 64 times) and recordings were made of the set portraying each of seven feelings (wrathfulness, nausea, fear, happiness, affable surprise, sadness, and neutral).

3. Text Dataset

The GoEmotions dataset is utilized for text investigation. GoEmotions, a human-annotated dataset of 58k Reddit comments extracted from popular English-language subreddits and labeled with 27 emotion categories, is referred to as "A Dataset of Fine-Grained Emotions."

6.RESULT AND DISCUSSION

The experimental results and the sample screenshots of the proposed system are provided in this given section.

A. Proposed System GUI Screenshots

1. Home Page

Fig. 5.1 Home Page

2. Login Page

Fig. 5.2 Login Page

3. Dashboard

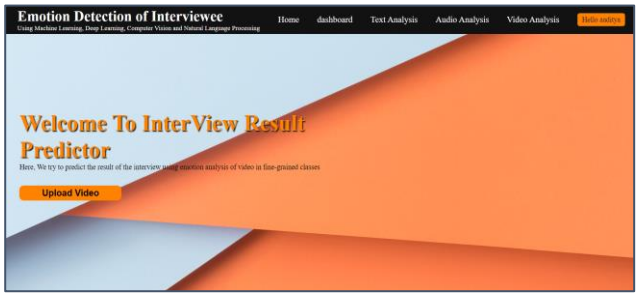

Fig. 5.3 Dashboard

4. Interview Performance Statistics

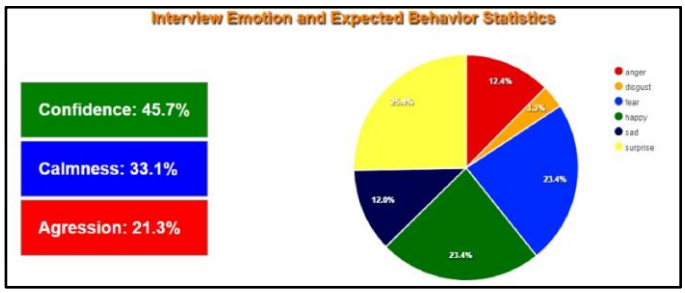

Fig. 5.4 Interview Performance Statistics

5. Performance Report

Fig. 5.5 Performance Report

B. Performance Evaluation:

The training and testing accuracies of the different Machine learning and Deep learning models are provided in this section. Inception-V3 and LSTM are used for image processing, while LSTM is used for audio processing and Neural Networks is used for text processing.

Table 5.1 Accuracies of models

Models	Inception-V3 + LSTM	LSTM	Neural Network
Training data Accuracy in (%)	91	97	87
Testing data Accuracy in (%)	65	97	53

7.CONCLUSION AND FUTURE SCOPE

The proposed system analyzes the visual, audio and textual features of the interview and evaluates the candidate's personality traits and the performance of the interview. The HaarCascade Frontal Face Classifier has been used in this project to detect faces, Mel frequency cepstral coefficient (MFCC) used for the audio feature extraction. Fluency of the speech has been calculated using amplitude and word frequency. Textual data is preprocessed with the help of natural language processing. Integration of InceptionV-3 model and LSTM is used to analyze the features of the image frames while Long Short Term Memory (LSTM) has been used for the analysis of audio features. As per as the textual features are concerned neural networks provide better performance. The system is able to calculate the behavior on the basis of emotional feelings. Statistical method is used to calculate behavior. The models used in this project obtained satisfactory results. In future the accuracy of models should be increased by considering the more optimized datasets. High-end software and hardware can be used for better performance.

REFERENCES

[1] Chou, Y. C., Wongso, F. R., Chao, C. Y., & Yu, H. Y. (2022, April). An AI Mock-interview Platform for Interview Performance Analysis. In 2022 10th International Conference on Information and Education Technology (ICIET) (pp. 37-41).

[2] Mhadgut, S., Koppikar, N., Chouhan, N., Dharadhar, P., & Mehta, P. (2022, February). vRecruit: An Automated Smart Recruitment Webapp using Machine Learning. In 2022 International Conference on Innovative Trends in Information Technology (ICITIIT) (pp. 1-6).

[3] Cai, L., Dong, J., & Wei, M. (2020, November). Multi-modal emotion recognition from speech and facial expression based on deep learning. In 2020 Chinese Automation Congress (CAC) (pp. 5726-5729).

[4] Adepu, Y., Boga, V. R., & Sairam, U. (2020, November). Interviewee performance analyzer using facial

emotion recognition and speech fluency recognition. In 2020 IEEE International Conference for Innovation in Technology (INOCON) (pp. 1-5).

[5] Parvathi, D. S. L., Leelavathi, N., Ravikumar, J. M. S. V., & Sujatha, B. (2020, July). Emotion Analysis Using Deep Learning. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 593-598).

[6] Jaymon, N., Nagdeote, S., Yadav, A., & Rodrigues, R. (2021, February). Real time emotion detection using deep learning. In 2021 International conference on advances in electrical, computing, communication and sustainable technologies (ICAECT) (pp. 1-7).

[7] Sham-E-Ansari, M., Disha, S. T., Chowdhury, A. I., & Hasan, M. K. (2020, June). A neural network based approach for recognition of basic emotions from speech. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 807-810).

[8] Rohan, M. A., Swaroop, K. S., Mounika, B., Renuka, K., & Nivas, S. (2020, October). Emotion Recognition Through Speech Signal Using Python. In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (pp. 342-346).

[9] Rashid, U., Iqbal, M. W., Skiandar, M. A., Raiz, M. Q., Naqvi, M. R., & Shahzad, S. K. (2020, October). Emotion Detection of Contextual Text using Deep learning. In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-5).

[10] Bharti, S. K., Varadhaganapathy, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K., & Mahmoud, A. (2022). Text-Based Emotion Recognition Using Deep Learning Approach. Computational Intelligence and Neuroscience, 2022.