International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 7, ISSN No. 2455-2143, Pages 176-180
Published Online November 2021 in IJEAST (http://www.ijeast.com)

EMOTIONAL RECOGNITION USING FACIAL EXPRESSIONS AND SPEECH ANALYSIS

Ms. Teena Verma (Assistant Professor),
Sahil Niranjan, Abhinav K.Gupta, Vinay Kumar, Yash Vashist
HMR Institute of Technology & Management, GGSIPU, New Delhi, India
Department of Computer Science & Engineering

***Abstract*- Emotional recognition can be made from Many sources including text, speech, hand, body language and facial expressions. Currently, most sensory systems use only one of these sources. People's feelings change every second and one method used to process emotional recognition may not reflect emotions in the right way. This research recommends the desire to understand and explore people's feelings in many similar ways speech and face. We have chosen to explore, sound and video inputs to develop an ensemble model that gathers the information from all these sources and displays it in a clear and interpretable way. By improving the emotion recognition accuracy, the proposed multi-sensory emotion recognition system can help to improve the naturalness of human computer interaction. Speech, hand, body language, and facial expressions are all examples of sources for emotional recognition. Most sensory systems currently use only one of these sources. People's feelings fluctuate by the second, therefore one method for processing emotional identification may not accurately reflect emotions. This study suggests that there is a need to comprehend and explore people's sentiments in many ways that voice and face do. Various emotional states were utilised in this case. Speech, facial expressions, and both can be used to detect emotions in the proposed framework. Audio, and video inputs and construct an ensemble model that collects data from all of these sources and presents it in a clear and understandable manner. The suggested multi-sensory emotion recognition system can help to increase the naturalness of human-computer interaction by boosting emotion recognition accuracy.**

Keywords—CNN, SVM, Facial Recognition, Speech Recognition, Python, Machine Learning, etc.

I. INTRODUCTION

Computers now play a critical part in a wide range of people's lives. Traditional computer integration, such as keyboard and mouse, is insufficient to meet people's demands. As a result, it is critical to contribute to the environment's cooperation between a computer and a person. Human communication via computers is now far less natural than communication among interconnected people. Complete computer communication does not yet allow for face-to-face contact with a person.

Computers should be able to support the way people communicate in order to improve natural delivery. They must be able to recognise and express emotions, which is a critical requirement. They should also be able to perceive and understand the feelings of their fellow humans in order to establish a natural connection such as communication.

Emotions are one of the most complicated aspects of the human body, both physically and mentally. Knowing our partners' emotional feelings is crucial and accurate in the community's daily life. When it comes to computers, however, getting the sensation of coworkers is quite tough.

Speech, gestures, eye contact, and facial expression are all used in human communication. Facial expressions, bodily movements, gestures, and eye contact are basic units for verbal communication in speech and basic units for verbal communication.

Geometry and appearance are two common strategies for automatic FER systems that have been used extensively in the literature.

Speech, gestures, eye contact, and facial expression are all used in human communication. Facial expressions, bodily and eye contact are basic units for verbal communication in speech and basic units for verbal communication.

Geometry and appearance are two common strategies for automatic FER systems that have been used extensively in the literature.

Pre-processing, face detection, feature extraction, and expression classification are usually the four stages of facial expression recognition. We used a variety of deep learning approaches (convolutional neural networks) to detect the key seven human emotions: anger, disgust, fear, happiness, sorrow, surprise, and neutrality in this study.

Words in speech are the basic units of verbal communication, while facial expression, bodily movements, gestures, and eye contact are the basic units of nonverbal communication. Take, for example, a phone call, where speech alone suffices for communication. Nonverbal communication units are extremely significant, especially when communicating face to face.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 7, ISSN No. 2455-2143, Pages 176-180
Published Online November 2021 in IJEAST (http://www.ijeast.com)

Furthermore, facial expressions reflect emotions that are significantly more intense than words. As a result, in order to have a more authentic human- computer interface, the computer should be able to discern emotion from speech, facial expressions, and body motions. We introduced a multi-sensory emotion recognition framework in this dissertation to recognise emotion from speech and facial expression. Ease of Use

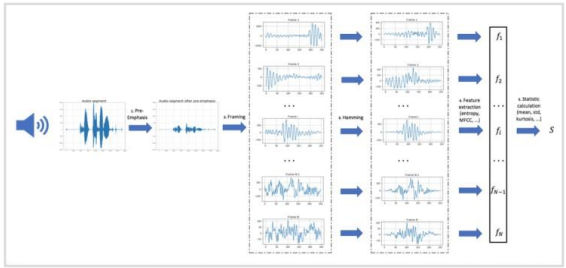
Dataset- In this paper, the RAVDESS database was used to assess our methods. It features male and female (gender balanced) performers who were requested to act out six various emotions (happy, sad, furious, disgust, fear, surprise, and neutral) at two levels of emotional intensity. The following table summarises the emotion distribution in RAVDESS:

RAVDESS

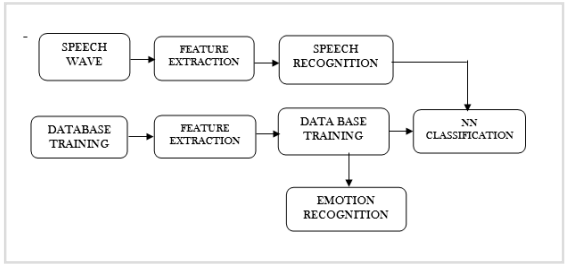
Emotions	Happy	Sad	Angry	Scared	Dis-gusted	Sur-prised	Neutral	**Total**
Man	*96*	*96*	*96*	*96*	*96*	*96*	*96*	**672**
Woman	*96*	*96*	*96*	*96*	*96*	*96*	*96*	**672**
Total	*192*	*192*	*192*	*192*	*192*	*192*	*192*	**1344**

II. METHODOLOGY

Every piece of soZware created is unique and necessitates a unique SDLC strategy based on internal and external considerations.
Gathering and analysing requirements: During this phase, all possible needs for the system to be developed are gathered and documented in a requirement specification document.
System Design: This phase examines the requirements specifications from the previous phase and prepares



the system design. System design aids in designing overall system architecture as well as describing hardware and system requirements.
Implementation: The system is first constructed in small programmes called units, which are then combined in the next phase, using inputs from the system design. Unit testing is the process of developing and testing each unit for its functioning.
AZer each unit has been tested, all of the units built during the implementation phase are merged into a system. The entire system is then tested for any flaws or failures aZer it has been integrated.
System deployment: Once the product has passed functional and nonfunctional testing, it is deployed in the customer environment or released to the market.
Maintenance: In the client environment, there are several issues that arise. Patches have been provided to address these issues. In order to improve the product, newer versions have been produced. Maintenance is carried out in order to bring about these modifications in the customer's environment. All of these phases are connected in a cascade, with progress appearing to flow smoothly downwards (like a waterfall) through them. The next phase begins only aZer the preceding phase's established set of goals has been met and it has been signed off, hence the term "Waterfall Model." Phases do not overlap in this model.)



The steps we followed while developing this project are-:

1. The problem statement is analysed.
2. Gathering the specifications for the requirements
3. An examination of the project's viability.
4. The creation of a general layout.
5. Based on past related efforts in this subject as documented inpublications.
6. Selecting an algorithm development approach.
7. Examining the many advantages and disadvantages.
8. Beginning the project's development
9. Making progress on analgorithm.
10. Algorithm analysis and improvement of accuracy
11. Coding in Python according to the described algorithm.

A- Speech Recognition
The goal of speech emotion recognition is to automatically detect a person's emotional or physical state based on his voice. The emotional state of a person's speech is an important aspect in human communication and engagement since it provides feedback in communication without changing the linguistic content.

Pre Emphasis filter- To begin, apply a pre- emphasis filter to the audio signal to boost all of the high frequencies before beginning feature extractions. A pre-emphasis filter has several advantages: it balances the frequency spectrum because high frequencies have smaller magnitudes than lower frequencies, and it avoids numerical issues when computing the Fourier Transform.

$$yt = xt - \alpha xt-1$$

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 7, ISSN No. 2455-2143, Pages 176-180
Published Online November 2021 in IJEAST (http://www.ijeast.com)

Typical values for the pre-emphasis filter coefficient α are 0.95 or 0.97.

Framing- AZer the pre-emphasis filter, we must divide the audio stream into frames, which are short-term windows. Window sizes for speech processing typically range from 20 to 50 milliseconds, with 40 to 50 percent overlap between two consecutive windows. The most common frame size is 25ms, with a 15ms overlap (10ms window step).
The main motivation behind this step is to avoid the loss of frequency contours of an audio signal over time because audio signals are non- stationary by nature. Indeed, frequency properties in a signal change over time, so it does not really make sense to apply the Discrete Fourier Transform across the entire sample. If we suppose that frequencies in a signal are constant over a very short period of time, we can apply Discrete Fourier Transform over those short time windows and obtain a good approximation of the frequency contours of the entire signal

Hamming- We multiply each frame by a Hamming window function aZer splitting the signal into numerous frames. It improves signal clarity by decreasing spectral leakage and any signal discontinuities. If the beginning and end of a frame do not match, the signal will appear to be discontinuous, and the Discrete Fourier Transform will display nonsense. The Hamming function ensures that the signal's beginning and end are in sync while smoothing itout.

Feature Extraction- We get a matrix of features per audio file aZer extracting the speech features from the preprocessed audio signal (described in the previous section). The first derivatives of each of those features are then computed to capture frame-to-frame changes in the signal. Finally, we compute the mean, median, standard deviation, kurtosis, skewness, 1 percent percentile, 99 percent percentile, min, max, and range between min and max for these attributes. For each audio signal, a vector of 200 potential characteristics is obtained. Before training and testing the classifier, some post-processing may be required.

B- Face Recognition
Dimension Reduction through Auto-Encoding**:**
The first thing that was done was to auto-encode the photos in order to shrink their size. By reducing the input image to a dimension of 12 × 12 pixels, the developed auto-encoder suggests a dimension reduction of more than 95%.
Signal processing, feature extraction, and classification are the three parts of the typical speech emotion recognition procedure. Signal processing separates raw audio streams into meaningful units using an acoustic filter. The delicate point in speech emotion recognition is feature extraction, because features must properly characterise the emotional content of a human speech while not relying on the lexical content or even the speaker. Finally, emotion classification will assign emotion labels to feature matrices.

Auto encoding- Since auto-encoding allows dimension reduction, we can again reduce the dimension to a dimension of 2 in order to be able to represent the different classes on a simple graph using T-Stochastic Neighbour Embedding (TSNE)techniques.

XCeption and Depthwise Separable convolutions- Xception is a Depthwise Separable Convolutions- based deep convolutional neural network architecture. Google researchers came up with the idea. Inception modules in convolutional neural networks are described by Google as an intermediate step between normal convolution and the depthwise separable convolution operation (a depthwise convolution followed by a point-wise convolution). In this sense, a depthwise separable convolution can be thought of as an Inception module with the most towers possible. This result leads them to propose a new deep convolutional neural network architecture based on Inception, but with depthwise separable convolutions in place of Inception modules.

$$Hn = \alpha - (\beta \cos(2\pi n)/(N-1))$$

where$\alpha = 0.54, \beta = 0.46and0 \le n \le N-1$withN the window length.

Illustration- A fundamental component of multimodal sentiment analysis is the ability to understand emotions using video data rather than just photos. The following is the procedure for deploying our trained model on a webcam stream:

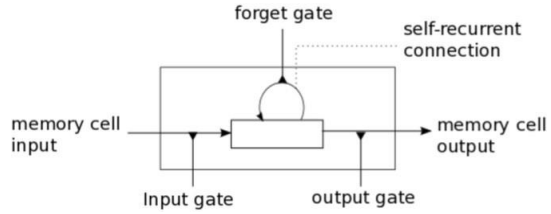
- add a grayscale filter to deal with fewer inputs
- recognise the face and zoom in on it
- manage several faces • examine the video picture by image
- lower pixel density to the same as the train set
- convert the input image to a model readable input
- anticipate the input's emotion

III. MODELING AND ANALYSIS

LSTM- Hochreiter Schmidhuber [1997] proposed Long Short Term Memory architectures, which have an advantage over traditional feed-forward neural networks and RNN. LSTMs do, in fact, have the ability to remember patterns selectively over long periods of time. A memory cell allows this to happen. An input gate, a neuron with a self- recurrent connection, a forget gate, and an output gate make up its one-of-a-kind structure. The self- recurrent connection assures that a memory cell's state does not change from one timestep

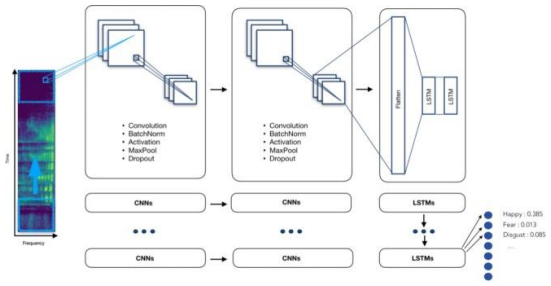
International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 7, ISSN No. 2455-2143, Pages 176-180
Published Online November 2021 in IJEAST (http://www.ijeast.com)

to the next. The gates use a sigmoid layer and a pointwise multiplication operation to fine-tune the interactions between the memory cell and its environment.

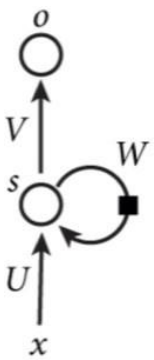


Time distributed convolutional neural network- Convolutional Neural Networks (CNNs) have outstanding recognition performance for computer vision tasks, but Recurrent Neural Networks (RNNs) have outstanding performance in many sequential data processing applications. Combining a deep hierarchical CNNs feature extraction architecture with a recurrent neural network model that can train to distinguish sequential dynamics in a speech signal is the concept of a temporally distributed convolutional neural network.

The primary idea behind a time distributed convolutional neural network is to use a rolling window (with a fixed size and time step) all the way through the log-mel-spectrogram. Each of these windows will be the entry point for a convolutional neural network made up of four Local Feature Learning Blocks (LFLBs), with the output of each of these convolutional networks being fed into a recurrent neural network made up of two LSTM (Long Short Term Memory) cells to learn long-term contextual dependencies. Finally, to forecast the emotion perceived in the voice, a fully linked layer with soZmax activation is used.



Recurrent Neural Networks- Recurrent neural networks take advantage of the sequential structure of data: unlike traditional neural networks, which presume inputs are independent of one another, these architectures gradually accumulate and record data through sequences



IV. RESULTS AND DISCUSSION

On the validation set, our model allows us to achieve a maximum score of 74%. We were able to save the weights associated with the best model using a Ker as feature called ModelCheckpoint, resulting in a 72 percent score on the test set.

Using deep learning and a time-distributed convolutional neural network, we were able to obtain a 10% improvement in performance over the typical SVM approach.

Our model yields findings that are quite satisfactory. For 7-way (happy, sad, angry, afraid, disgusted, surprised, neutral) emotions, our forecast recognition rate is around 65 percent, and for 6-way emotions, it's around 75 percent (surprised removed). In order to increase our performance, Hidden Markov Models (HMM) and Convolutional Neural Networks (CNN) appear to be strong choices for recognising speech emotion. Those classifiers, unlike SVM classifiers, are trained using short-term characteristics rather than global statistics features. For better capturing the temporal dynamic included in speech, HMM and CNN are deemed helpful.

To compare results, we tried different combinations of embeddings and classifiers. We explored a hybrid model employing averaged vector representations with TF-IDF weights, as mentioned at the end of the section on Word2Vec embeddings: there is a loss of accuracy compared to the whole Word2Vec embedding, but the results are better than the regular TF-IDF embedding. Let's go over the accuracy we got with each of the combinations we tried in our pipeline:

Model	EXT	NEU	AGR	CON	OPN
TF-IDF + MNB	45.34	45.11	45.24	45.31	45.12
TF-IDF + SVM	45.78	45.91	45.41	45.54	45.56
Word2Vec + MNB	45.02	46.01	46.34	46.38	45.97
Word2Vec + SVM	46.18	48.21	49.65	49.97	50.07
Word2Vec (TF-IDF averaging) + MNB	45.87	44.99	45.38	44.21	44.84
Word2Vec (TF-IDF averaging) + SVM	46.01	46.19	47.56	48.11	48.89
Word2Vec + NN (LSTM)	51.98	50.01	51.57	51.11	50.51
Word2Vec + NN (CONV + LSTM)	**55.07**	**50.17**	**54.57**	**53.23**	**53.84**

V. CONCLUSION

To summarise, for distinct input sources analysed individually, it is possible to design quite good classifiers for both personality characteristics and emotions recognition, each modality requiring its own collection of features and hyper-parameters. The next step in our study will be to create an ensemble model that can combine the results of both personality trait detection and emotion recognition to deliver a more comprehensive assessment of a user's interview. Our final model would incorporate a form of coherence

measure that expresses the resemblance between a user's emotional profile and the average characteristics of people in the same psychological category, as determined by our research. Our final model would incorporate a form of coherence score that expresses the resemblance between a user's emotional profile and the average characteristics of people in the same Big Five psychological category. Unsupervised clustering algorithms are oZen used in this case.

VI. REFERENCES

[1] B.Kratzwald, S.Ilie ´, M.Kraus, S.Feuerriegel, H.Prendinger. Deep learning for affective computing: text-based emotion recognition in decision support, Sep. 2018.https://arxiv.org/pdf/ 1803.06397.pdf

[2] https://www.youtube.com/watch?v=oP3MQyO-wwc

[3] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), https:// zenodo.org/record/1188976/?f=3.XAcEs5NKhQK

[4] https://en.wikipedia.org/wiki/ Facial_recognition_system

[5] T.Vogt, E.Andre´and J.Wagner. Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation. Affect and Emotion in Human- Computer Interaction, 2008.

[6] https://en.wikipedia.org/wiki/ Facial_identification

[7] T.Giannakopoulos. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. Dec. 2015https://doi.org/10.1371/ journal.pone.0144610

[8] https://en.wikipedia.org/wiki/ Convolutional_neural_network

[9] The Facial Emotion Recognition Challenge from Kaggle, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression- recognition-challenge/data

[10] End-to-End Multimodal Emotion Recognition using Deep Neural Networks,

https://arxiv.org/pdf/1704.08619.pdf