RESEARCH ARTICLE **SPECIAL ISSUE**

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S023

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Emotion Analysis Using Speech

M M Krupashree[1], Naseeba Begum[1], Nayana Priya [2], Nithya S [1], Rashmi Motkur[3]

[1]Department of Computer Science Engineering, Dayananda Sagar University, Bangalore, India
[2]Assistant Professor, Department of Computer Science Engineering, Dayananda Sagar University, Bangalore, India
[3]Professor, Department of Computer Science Engineering, Dayananda Sagar University, Bangalore, India

Email: mmkrupashree@gmail.com

Article History

Received: 28 February 2023
Accepted: 16 March 2023

Keywords:

deep learning;
CNN;
LSTM;
MFCCS;
mel-Spectrograms

Abstract

The main goal of our project is to identify the emotions a speaker evokes when speaking. For example, utterances uttered in states of fear, surprise, excitement, anger, or joy are loud and fast and have a large and wide pitch range, whereas utterances uttered in states of depression or fatigue are slow and deep. This is us We use deep learning techniques to build models that can identify human emotions through the analysis of speech and language patterns. The main reason for choosing this project is that speech sentiment analysis has become one of the largest commercialization strategies in which client moods and dispositions play a large role. Therefore, there is an increased demand for products or companies to recognize an individual's emotions and recommend appropriate products or assist him accordingly. It can also be used to monitor status. More recently, speech recognition and analysis have also been applied to medicine and forensics.

1. Introduction

Systems for recognizing speech emotions (SER) have developed from a specialized field to a crucial component of human-computer interaction (HCI). Instead of using conventional devices as input to understand rhetorical content and make it simple for human listeners to acknowledge, these HCI systems aim to speed up innate communication with machines through explicit speech interaction. In some applications, dialogue systems for lingual languages are used for call center consultations, music recommendation systems are made based on the user's mood, and emotion analysis from the speech is used in medical and forensic applications. (Senthilkumar et al.) However, there are many difficulties with HCIsystems, including noisy settings and different speaker accents that cause ambiguity that still needs to be properly resolved.

2. Literature Survey

Edward Jones et al (Jones) have presented a paper on Speech Emotion Recognition Using Deep Learning Techniques: A Review. These methods offer easy model training as well as the efficiency of shared weights. Limitations of deep learning techniques include their large layer-wise internal architecture, less efficiency for temporally varying input data, and overlearning during memorization of layer-wise information.

Ron Hoory et al (Hoory) have presented a paper on Speech Emotion Recognition Using Self-Supervised Features. They have clearly shown that well-designed combinations of carefully fine-tuned and averaged Upstream models and averaged Downstream models can significantly improve the performance of E2E SER models. This research paper aims to introduce a modular End To-End(E2E) SER

system based on an Upstream + Downstream architecture model paradigm

Mira Kartiwi et al (Kartiwi) have presented a paper on A Comprehensive Review of Speech Emotion Recognition Systems.This paper points out that deep learning techniques are considered best suited for the SER system over traditional techniques because of their advantages like scalability, all-purpose parameter fitting, and infinitely flexible function.

Srinivasa Parthasarathy et al (Parthasarathy) have presented a paper on Semi-Supervised Speech Emotion Recognition With LadderNetworks. The outcomes showed substantial increases when utilizing the suggested models, supporting the generalizability of the ladder networks.. The improvements were particularly high when using unlabeled data from the target domain, exploiting all the benefits of the proposed architecture.

3. Design

3.1. System Architecture

FIGURE 1 depicts the overall architecture of the project where input is taken in the form of speech or audio. Then we have to extract features from the input. Some of the main audio features extracted are Mel-frequency Cepstral Coefficient(MFCC), pitch, Mel-Spectrograms, chrome, zero-crossing rate, etc. Then the dataset is split into the training set and testing set. The very next step is training Deep-learning and Artificial Intelligence models using each feature extracted from the training dataset. A testing dataset is used to evaluate the developed model. Then we will save the model which gives the best accuracy. then we will be connecting the model to the interface and output the predicted emotion.

3.2. Data Flow Diagram

FIGURE 2 illustrates the data flow diagram of the project. Once the SER model is trained and tested, it is exported or embedded into an app. When you start the application it prompts the user to give input using Google Speech API. Input data is sent to the server via HTTP post request where it receives input and does feature extraction and tests extracted features using an already trained SER model/Then it predicts emotion and returns a JSON response.

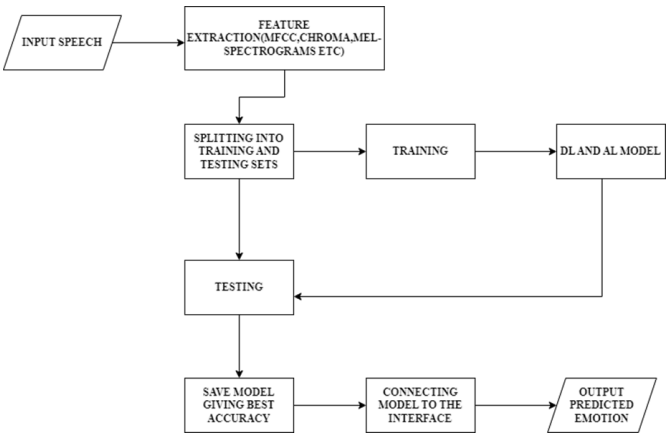


FIGURE 1. System

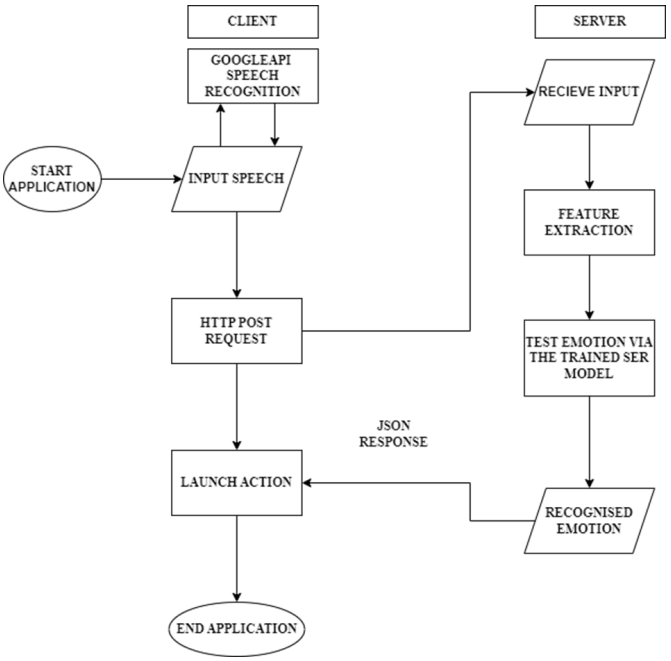


FIGURE 2. Data

3.3. Backend Architecture

FIGURE 3 depicts the backend architecture of our project where standard datasets are taken as input and divided into training samples and testing samples. Training samples undergo pre-processing such as converting audio waves to melspectogram and data augmentation which increases the diversity of the dataset by using standard augmentation techniques such as changing pitch, injecting noise, etc. The next step is feature extraction which extracts features such as MFCC, chroma, and Mel-frequency spectrograms. These extracted features are then sent to classifiers for predicting emotion. To evaluate the model, we will be using testing samples that undergo feature extraction and are sent to classifiers for predicting emotion. We can also test by pro-

viding live input via google speech recognition API which undergoes feature extraction and is sent to the model for predicting emotion.

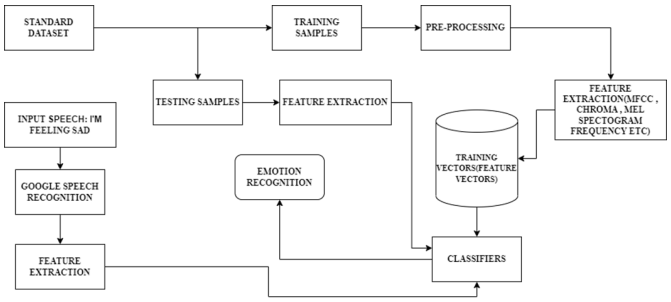


FIGURE 3. Backend Architecture of Speech Emotion Recognition System

4. Methodology

4.1. Existing system

The majority of the SER systems presently on the market employ traditional machine learning algorithms for emotion recognition, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gaussian Mixture Model (GMM), etc.. The accuracies of these models are low and have high computational complexity. However, there are other deep learning models such as Convolutional Neural Network(CNN), Quaternion Convolutional Neural Network(QCNN), and, Long-Short Term memory(LSTM), etc which give average accuracy of around 80 percent because of their computational complexity and many other reasons.

4.2. Proposed System

We propose an enhanced speech emotion recognition method that uses hybrid model of deep neural networks that is, CNN and LSTM to detect emotions elicited by the speaker.the method used Mel-frequency cepstral coefficients (MFCC), chromogram, Mel scale spectrogram in conjunction with spectral contrast to extract details about an audio file. These features are used to train our hybrid model which gives better accuracy compared to other existing models. The model classifies the speech audio in 8 different emotions such as neutral , calm , surprise , happy , anger , fearful , disgust, sad.

4.3. Proposed Methodology

The system's architecture makes it clear that we are using voice training. and it is then passed for pre-processing for the feature extraction of the sound which then gives the training arrays.These arrays are then used to form "classifiers "for making decisions about the emotion. So, a big data set of voices of different emotions is needed for the training sample. We searched on the web and found different sets of datasets some of which are mentioned below:

1.Crowd-sourced Emotional Multimodal Actors Dataset(Crema-D)

2. Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess)

3. Surrey Audio-Visual Expressed Emotion (Savee)

4. Toronto emotional speech set (Tess)

To begin with we created data frames and then the later step was data visualization and exploration wherein we have a wave plot and spectrogram of the audio input we have. Data augmentation is the process of creating new synthetic data samples by adding small perturbations to our initial training set.To generate syntactic data for audio, we can add noise, change the time, the pitch, and the pace. The objective is to make our model invariant to those perturbations and enhance its ability to generalize. In order for this to work adding the perturbations must conserve the same label as the original training sample.In images data augmentation can be performed by shifting the image, zooming, rotating.The next step being feature extraction where in we have extracted 5 features which are Zero crossing rate, chroma stft, RMS(root mean square)value, Mel Spectrogram to train our data Then on, the data preparation step where we have split the datasets into training and testing datasets.Further we have used the machine learning models such as Decision tree, KNN, MLP Classifier, LSTM and CNN Wherein, we have generated the confusion matrix and a classification report for each of the model.

5. Experimentation

5.1. Data Augmentation

FIGURE 4 shows wave plots of audio files after applying different data augmentation techniques such as injecting noise, stretching audio, and changing the pitch of audio in order to increase the diversity of the dataset and lessen the model's overfitting.

5.2. Feature Extraction

FIGURE 5 includes different features extracted from audio.

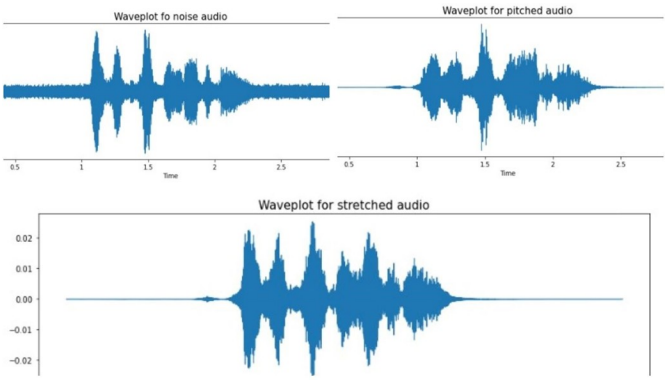


FIGURE 4. Waveplots of

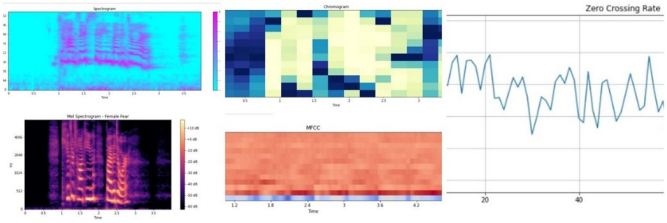


FIGURE 5. Different features extracted from audio

1. Mel-Spectrograms, which represent sound or audio on a mel scale. The mel scale is used because humans perceive sound differently from machines, which have a resolution that is the same across all frequencies as opposed to our higher resolution at lower frequencies. (Muppidi and Radfar) We convert our audio frequency to mel frequency because it has been found that simulating the human hearing characteristic during feature extraction improves the model's accuracy.

2. Chroma: A spectrogram is projected onto 12 bins to represent the 12 distinct semitones in the standard audio representation of audio. (Aftab et al.) On a typical chromatic scale, it displays the energy of each pitch that is present in the signal.

3. The zero-crossing rate is the speed at which a positive signal turns negative and vice versa. The frequency of the signal crossing the horizontal axis is another way to conceptualize it.

4. MFCC: It represents the short-time power spectrum envelope, which represents the vocal tract's shape.

5. RMS value: One of the most crucial parameters, it shows the signal's strength or power.

5.3. Data Pre-Processing

In this Data-preprocessing, we will be loading features into the X variable and emotions into the Y variable. Since detecting the emotions of the speaker is a multiclass classification problem, we will be using a one-hot encoding technique by which categorical data are converted into binary features of data (Prasomphan) . Then we will be splitting the dataset into the training set and testing set. In our project, 75 percent is training data, and the rest 25 percent is testing data. then we will be standardizing data using StandardScaler to make sure all variables contribute equally.

6. Result And Analysis

6.1. Decision Tree

```
In [67]:
#decision tree
from sklearn.tree import DecisionTreeClassifier
clf3 = DecisionTreeClassifier()

clf3 = clf3.fit(x_train,y_train)

y_pred = clf3.predict(x_test)

In [68]:
print("Training set score: {:.3f}".format(clf3.score(x_train, y_train)))
print("Test set score: {:.3f}".format(clf3.score(x_test, y_test)))

Training set score: 1.000
Test set score: 0.533
```

FIGURE 6. Training

FIGURE 6 depicts training and a test score of the decision tree model. It is observed that the training score is 100 percent which is unusual whereas the test score is around 40 percent which indicates the model is overfitting. (Singh and Goel) This is caused because the model is memorizing exact input and output pairs in training data instead of learning patterns. So, it underperforms when evaluated on test data.

6.2. KNN

FIGURE 7 illustrates training and test scores of the KNN model. It is observed that the training score is around 49 percent and the test score is around 37 percent. This accuracy is not good for deployment. (Vamshi and Krishna) The low test score is probably because the testset may contain new features which are not present in the training set and one more reason is the KNN is not good for large datasets because it is very costly to calculate the distance between existing points and new points which degrades model performance.

6.3. MLP Classifier

FIGURE 8 depicts the training and test scores of the MLP Classifier. It is observed that the training score is around 80 percent which is good but the test score is around 50 percent which makes the model not suitable for deployment.

```
[69]: #knn
      from sklearn.neighbors import KNeighborsClassifier
      clf1=KNeighborsClassifier(n_neighbors=4)
      clf1.fit(x_train,y_train)
[69]: KNeighborsClassifier(n_neighbors=4)
[70]: y_pred=clf1.predict(x_test)
[71]: print("Training set score: {:.3f}".format(clf1.score(x_train, y_trai
      print("Test set score: {:.3f}".format(clf1.score(x_test, y_test)))
Training set score: 0.492
Test set score: 0.372
```

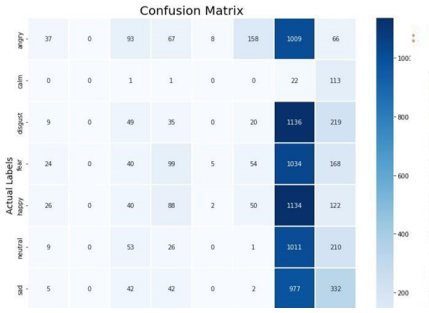
FIGURE 7. Training

```
#MLP Classifier
from sklearn.neural_network import MLPClassifier
clf2=MLPClassifier(alpha=0.01, batch_size=270, epsilon=1e-08, hidden_layer_sizes=(400,), learning_rate='adaptive', max_iter=
clf2.fit(x_train,y_train)
MLPClassifier(alpha=0.01, batch_size=270, hidden_layer_sizes=(400,),
              learning_rate='adaptive', max_iter=400)
print("Training set score: {:.3f}".format(clf2.score(x_train, y_train)))
print("Test set score: {:.3f}".format(clf2.score(x_test, y_test)))
Training set score: 0.804
Test set score: 0.535
```

FIGURE 8. Training

6.4. LSTM

FIGURE 9 illustrates the confusion matrix and the actual-predicted output of the LSTM model. It has been found that the model's accuracy is approximately 67%.Input data used in our project is sequential data and the LSTM model is predominantly used for this kind of data since it can remember long-term dependencies between time steps of data. (Aggarwal et al.) Training accuracy was good but it underperformed on testing data. This is because LSTMs are easily overfitted and implementing dropouts is quite difficult.

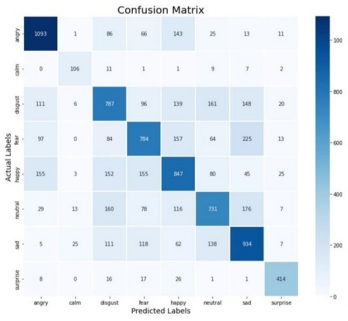


	Predicted Labels	Actual Labels
0	surprise	disgust
1	surprise	disgust
2	neutral	angry
3	sad	disgust
4	sad	fear
5	sad	fear
6	fear	happy
7	sad	happy

FIGURE 9. Confusion

6.5. CNN

FIGURE 10 illustrates the confusion matrix and the actual-predicted output of the CNN model. It is observed that the accuracy of the model is around 62 percent from the classification report mentioned in FIGURE 11 clearly. We can see our model is more accurate in predicting surprise, and angry emotions and it makes sense also because audio files of these emotions differ from other audio files in a lot of ways like pitch, speed, etc.



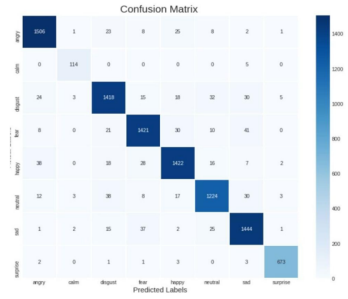
	Predicted Labels	Actual Labels
0	sad	disgust
1	disgust	disgust
2	angry	angry
3	happy	disgust
4	sad	fear
5	disgust	fear
6	angry	happy
7	happy	happy

FIGURE 10. Confusion

	precision	recall	f1-score	support
angry	0.73	0.76	0.74	1438
calm	0.69	0.77	0.73	137
disgust	0.56	0.54	0.55	1468
fear	0.60	0.55	0.57	1424
happy	0.57	0.58	0.57	1462
neutral	0.60	0.56	0.58	1310
sad	0.60	0.67	0.63	1400
surprise	0.83	0.86	0.84	483
accuracy			0.62	9122
macro avg	0.65	0.66	0.65	9122
weighted avg	0.62	0.62	0.62	9122

FIGURE 11. Classification

6.6. CNN-LSTM



	Predicted Labels	Actual Labels
0	female_fear	female_fear
1	male_angry	male_angry
2	female_fear	female_fear
3	female_neutral	female_neutral
4	female_angry	female_angry
5	female_surprise	female_surprise
6	male_sad	male_fear
7	male_happy	male_happy

FIGURE 12. Confusion Matrix And Output Table Of CNN-LSTMModel

FIGURE 12 illustrates the confusion matrix and the actual-predicted output of the CNN-LSTM model. It is observed that the accuracy of the model is around 94 percent from the classification report mentioned in FIGURE13 clearly. This has overall good accuracy because the CNN-LSTM hybrid model was used for speech emotion detection where CNN extracts features and LSTM will handle sequential learning.

```
print(classification_report(y_test_, y_pred))
```

	precision	recall	f1-score	support
angry	0.95	0.96	0.95	1574
calm	0.93	0.96	0.94	119
disgust	0.92	0.92	0.92	1545
fear	0.94	0.93	0.93	1531
happy	0.94	0.93	0.93	1531
neutral	0.93	0.92	0.92	1335
sad	0.92	0.95	0.93	1527
surprise	0.98	0.99	0.98	683
accuracy			0.94	9845
macro avg	0.94	0.94	0.94	9845
weighted avg	0.94	0.94	0.94	9845

FIGURE 13. Classification Report Of CNN-LSTMModel

MODELS	AVERAGE ACCURACY
GMM(Gausian mixture model)	72.61%
SVM(support vector machine)	78.16%
MLP(multilayer perceptron)	71.87%
Decision Tree	53.3%
Knn(k-nearest neighbours)	37%
MDT(Meta Decision Tree)	80%
Q-CNN(Quaternion Convolutional neural network)	77.97%
LSTM(Long short term memory)	67%
CNN(Convolutional Neural Network)	62%
CNN-LSTM	94%

FIGURE 14. Comparison Of Results

7. Comparison Of Results

FIGURE 14 illustrates the comparative study of the various classification models. The addition of behavioral features has improved the accuracy of the proposed model. After a thorough comparative study of different classification models, CNN-LSTM model has given the best accuracy.

8. Conclusion

In conclusion, because of its potential applications in a variety of domains, including human-computer interaction, healthcare, and psychology, the development of speech-emotion recognition (SER) systems has grown in importance as a research issue. The goal of SER is to automatically ascertain a speaker's emotional state from their speech signal.

In this article, we covered the speech preprocessing, feature extraction, and classification processes that make up a typical SER system. We covered a number of methods for each of these elements, including feature extraction methods like Mel frequency cepstral coefficients (MFCCs), signal processing methods like filtering, and classification algorithms like support vector machines (SVMs) and deep learning models.

Although SER has generally made tremendous progress in recent years, there is still much need for improvement.

Overall, there is still much opportunity for improvement even though SER has made significant progress in recent years. To increase the reliability and accuracy of SER systems and to make them viable for use in real-world circumstances, more research and development is required**.**

References

Aftab, Arya, et al. "LIGHT-SERNET: A Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition". *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022): 2022–2022. 10.1109/ICASSP43922.2022.9746679.

Aggarwal, Apeksha, et al. "Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning". *Sensors* 22.6 (2022): 2378–2378. 10.3390/s22062378.

Jones, Edward. "Speech Emotion Recognition Using Deep Learning Techniques : A Review". *Speech Emotion Recognition Using Deep Learning Techniques : A Review* (2019). 10 . 1109 / ACCESS.2019.2936124.

Kartiwi, Mira. "A Comprehensive Review of Speech Emotion Recognition Systems". *A Comprehensive Review of Speech Emotion Recognition Systems* (2021). 10 . 1109 / ACCESS . 2021 . 3068045.

Muppidi, Aneesh and Martin Radfar. "Speech Emotion Recognition Using Quaternion Convolutional Neural Networks". *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021). 10.1109/ICASSP39728.2021.9414248.

Parthasarathy, Srinivasa. "Semi-Supervised Speech Emotion Recognition with Ladder Networks". *Semi-Supervised Speech Emotion Recognition with Ladder Networks* (2019). 10 . 48550 / arXiv . 1905.02921;%20https://doi.org/10.48550/arXiv.1905.02921.

Prasomphan, Sathit. "Detecting human emotion via speech recognition by using speech spectrogram". *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2015). 10.1109/DSAA.2015.7344793.

Senthilkumar, N, et al. "Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks". *Materials Today: Proceedings* 57 (2022): 2180–2184. 10.1016/j.matpr.2021.12.246.

Singh, Youddha Beer and Shivani Goel. "A systematic literature review of speech emotion recognition approaches". *Neurocomputing* 492 (2022): 245–263. 10.1016/j.neucom.2022.04.028.

Vamshi, Kotikalapudi and Krishna. "Speech Emotion Recognition using Machine Learning". *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)* (2022): 2022–2022. 10 . 1109 / ICCMC53470 . 2022.9753976.

Embargo period: The article has no embargo period.

To cite this Article: Krupashree, M M, Naseeba Begum, Nayana Priya , Nithya S , and Rashmi Motkur. "Emotion Analysis Using Speech." International Research Journal on Advanced Science Hub 05.05S May (2023): 178–184. http://dx.doi.org/10.47392/irjash.2023.S023