JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Speech Emotion Recognition Using Machine Learning

Shaishav Tiwari
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India

Mohosmita Sengupta
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India

Nishtha Parlani
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India

Abhishek K
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India

Prajna K B
Assistant Professor, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India

ABSTRACT
Speech is one of the most innate ways for humans to communicate. We rely on it so much that we understand its significance while using other communication channels, such as emails and text messages, where we frequently utilize emoticons to convey our feelings. The detection and analysis of emotions are crucial in today's digital age of remote communication because they play a crucial role in communication. Because emotions are arbitrary, detecting them is a difficult process. How to quantify or classify them is a subject of debate. Speech emotion recognition is the process of accurately anticipating a human's emotion from their speech. It improves the way people and computers communicate.

Keywords
Speech, RAVDESS dataset, MFCC, Mel spectrogram, Chroma, Classification, Machine Learning, Deep Learning.

1. INTRODUCTION
One of the natural ways that people may communicate their feelings is through speech. Additionally, voice is simpler to acquire and analyse in real-world situations, which is why the majority of apps that rely on emotion identification use speech. A typical SER system operates by extracting information from speech, including spectral, pitch frequency, formant, and energy-related data. It then performs a classification job to forecast different classes of emotion. Speech style is crucial since it's necessary for conveying the speech's meaning. When we adopt particular speech styles for specific discussions or speeches, we are able to communicate our ideas in a way that the listener will understand better. Traditionally, feature extraction and classification have been the two fundamental components of the task of speech emotion recognition (SER). Using various front-end signal processing techniques, a speech signal is transformed into numerical values during the feature extraction stage. Extracted feature vectors are compact and, in theory, should include all of the crucial data from the signal. An appropriate classifier is chosen on the back end based on the task that needs to be accomplished.

The two essential elements of the speech emotion recognition (SER) challenge are feature extraction and classification. Various front-end signal processing techniques are used to transform a voice signal into numerical values during the feature extraction step. The ideal extracted feature vector should contain the most important data from the signal and have a compact shape. A suitable classifier is chosen in the backend based on the required job. An emotion identification system's performance only depends on characteristics and representations taken from the audio. They may be roughly divided into aspects that are time-based and frequency-based. The advantages and disadvantages of these traits have been thoroughly studied. There isn't a single audio feature that excels at all jobs involving processing of acoustic signals. Furthermore, Hand-crafted characteristics are used to meet the demands of the current issue. Our efforts have proven fruitful in separating presenting the speech in a hierarchical manner from these characteristics and determining underlying sentiment of the speech with the introduction of machine learning and deep learning techniques. Therefore, the model's performance in a specific voice recognition test depends significantly more on the selection more than the model architecture in terms of the feature. In this project, we focus on research projects addressing the analysis of acoustic cues from speech to recognize the speaker's emotions.

2. LITERATURE SURVEY
In order to address the issue of binary classification, Cao et al. presented a ranking SVM approach to gather data on emotion recognition. This ranking technique uses SVM algorithms to give instructions for certain emotions, processing the data from each speaker as a separate query before combining all rankers' predictions to use multi-class prediction. In two open datasets of performed emotional speech, Berlin and LDC, ranking techniques significantly outperform standard SVM in terms of accuracy. When it came to successfully recognising emotional utterances in both the acted data and the spontaneous data, which contains neutral strong emotional

utterances, ranking-based SVM beat classic SVM approaches. Balance accuracy (or unweight average, UA) was 44.4 percent. Rong et al. introduced an ensemble random forest to trees (ERF Trees) approach with a large number of attributes for emotion recognition, without referring to any language or linguistic information. This technique is used with tiny data sets that have a lot of characteristics. An experiment using a dataset of Chinese speakers' emotional speech was conducted to assess the suggested technique, and the findings show that it improved the rate of emotion identification. With 16 features in all, the female dataset had the highest maximum accuracy rate of 82.54 percent., while the natural data set with 84 characteristics had the lowest accuracy at just 16 percent.

Kyong Hee Lee, Do Hyun Kim: construction of a convolutional neural network for speech emotion recognition. This study used Mel spectrograms for a CNN and the speech features of MFCCs for an MLP network to create a deep neural network. With the help of this MLP, a test accuracy of around 75% was attained. However, with regard to CNN, the accuracy was almost 60%.

Haiyan Chen, Zheng Liu, Xin Kang: conducting research on vocal characteristics for speech emotion recognition based on four kinds of machine learning methods. With a total accuracy of 81.11 percent, the experimental findings demonstrate that the effect of speech recognition using an SVM model is the most notable. CNN model comes in second with a score of 80.56 percent. Accuracy rates for the KNN model and random forest model are 55.56 and 58.89 percent, respectively.

Harshini D, Pranjali B, Ranjitha M, Rushali J and Manikandan J: Design and Evaluation of Speech Emotion Recognition System using Support Vector Machines. Using a set of 26 characteristics, it can be seen that the suggested work can detect six and seven emotions with maximum recognition accuracy of 100 percent and 78.94 percent, respectively.

Zhou Qing, Wang Zhong, Wang Peng: Research on Speech Emotion Recognition Technology Based on Machine Learning. Through the findings, we can see that the work of speech emotion recognition has essentially been finished- and that the total effect of speech emotion recognition under the KNN algorithm has reached 78 percent.

Narayanan suggested using voice cues from a call-centre application to recognise domain-specific emotions. The main goal of this research is to identify both negative and positive emotions, such as anger and happiness. To work with various kinds of features, k-NN and linear discriminant classifiers are both employed. The outcome of the experiment supports the claim that combining acoustic and linguistic data yields the best outcomes. Results show that using three sources of information instead of just one improves classification accuracy by 40.7 percent for men and 36.4 percent for women. For men, the accuracy improvement ranges from 1.4 to 6.75 percent, whereas for women, it ranges from 0.75 to 3.96 percent.

3. OBJECTIVE AND SCOPE

A Speech Emotion Recognition system's main goal is to enhance the human-computer interface. It may also be utilized in lie detectors to track a person's psychophysical condition. Speech emotion recognition has recently found use in the fields of medicine and forensics. In this project, by utilising a range of machine learning algorithms and deep learning techniques, we will classify and analyse vocal signals in order to determine the underlying emotions. This model forecasts emotions based on the RAVDESS database, which contains recordings of two phrases in English with North American accents from each of the 24 voice actors, each speaking in one of eight different moods. A voice emotion recognition system could be helpful in a real-time setting to enhance industrial customer service. Customers' verbal comments on the service might be utilised to express their feelings, and the client could then be treated accordingly. Voice-based virtual assistants might gain from SER models like these in the marketplace and sector.

4. METHODOLOGY

4.1 Block Diagram of Speech Emotion Recognition System

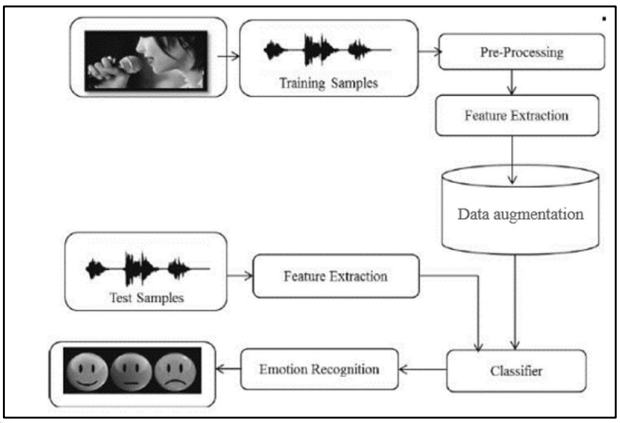

Figure 4.1: Basic Block diagram of an SER system.

4.2 Tools Used

4.2.1 Kaggle:

Data scientists and machine learning enthusiasts may connect online at Kaggle. Users of Kaggle may work together, access and share datasets, use notebooks with GPU integration, and compete with other data scientists to solve data science problems. In a web-based data-science environment, Kaggle enables users to explore, construct models, and locate and publish data sets.

4.2.2 Jupyter Notebook:

Software code, computational results, explanatory text, and multimedia files are all combined into one page by the interactive web application Jupyter. You can use the open-source and free Jupyter Notebook online application to create and share documents with live code, equations, visualisations, and text. is an interactive web tool known as a computational notebook, which is employed to mix software code, computational output, explanatory text, and multimedia elements in a single page.

4.2.3 Google Colab:

Colaboratory, sometimes known as "Colab," is a Google Research product Colab is especially well suited for teaching, data analysis, and machine learning. It enables anybody to create and execute arbitrary Python code through the browser. Technically speaking, Colab is a hosted Jupyter notebook service that offers free access to computer resources, including GPUs, and requires no setup to use.

4.3 Description of Dataset Used

Selecting an emotional speech database is one of the key elements of SER design. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is a recognised multi-modal database of emotional speech and song. The 24 professional actors in this gender-balanced database produce 104 different vocalisations representing various emotions, such as happy, sad, anger, fear, surprise, disgust, calm, and neutral.

4.4 Feature Extraction Techniques Used

A dimensionality reduction method called feature extraction splits a vast amount of raw data into smaller, convenient subgroups.

These huge data sets have the trait of having many variables that demand a lot of computational power to process. The term "feature extraction" refers to techniques for choosing and/or combining variables into features, which significantly reduces the quantity of data that has to be processed while properly and fully describing the initial data set.

In this project, we are focusing on three frequency domain features:

1. Mel-Frequency Cepstral Coefficients (MFCCs)
2. Mel Spectrogram
3. Chroma coefficients

4.4.1 Mel- Frequency Cepstral Coefficients (MFCCs):
An MFC is made up of a number of coefficients known as Mel-frequency cepstral coefficients (MFCCs). They come from a particular cepstral interpretation of the audio sample. The Mel-frequency cepstrum (MFC) differs from the cepstrum in that the frequency bands are evenly spaced on the M el scale, which more strongly resembles the response of the human auditory system than the linearly-spaced frequency bands used in the conventional spectrum.

4.4.2 Mel Spectrogram:
An alteration in a given quantity over time is referred to as a signal. When it comes to audio, air pressure is the variable quantity. The air pressure can be continuously measured by taking samples. We sample the data at a variety of frequencies, but most frequently at 44.1kHz, or 44,100 samples per second. A waveform of the signal has been recorded, and software can be used to interpret, alter, and analyse it. A spectrogram whose frequencies are scaled to the Mel scale is known as a Mel spectrogram.

4.4.3 Chroma coefficients:
The chroma feature is a description that summarises the tonal component of an audio musical stream. Chroma characteristics can be thought of as a necessary prerequisite for high-level semantic analysis, such as chord identification or determining harmonic similarity. Better outcomes on these complex tasks are made possible by an extracted chroma feature of higher quality. Chroma feature extraction uses Short Time Fourier Transforms and Constant Q Transforms.

4.5 Machine Learning Algorithms Used
A cutting-edge aspect of computer science called "machine learning," or "ML," aims to have computers do tasks without having to be explicitly trained to do so. Machine learning employs a variety of methods to develop algorithms that learn from data sets and generate
predictions. It is utilised in data mining, a method for identifying patterns and models in data sets when correlations were not previously recognized.

4.5.1 Random Forest (RF):
In the ensemble learning method known as random forests or random decision forests, which is used for classification, regression, and other tasks, a significant portion of decision trees are built during the training phase. The class that the preponderance of the trees choose is the output of the random forest for classification problems. A supervised machine learning technique called a random forest is built using decision tree algorithms.

4.5.2 Support Vector Machine (SVM):
Boser, Guyon, and Vapnik's COLT-92 paper that introduced the Support Vector Machine (SVM) was published in 1992. Support vector machines (SVMs) are a group of connected supervised learning techniques applied to regression and classification. They belong to the 31 family of generalised linear classifiers. Support Vector Machine (SVM), in other words, is a classification and regression prediction tool that automatically detects over-fitting to the data while enhancing predictive accuracy using machine learning theory.

4.6 Deep Learning Techniques Used
The term "deep learning" refers to how many levels of transformation the data goes through. Deep learning systems specifically have a significant credit assignment path (CAP) depth. The series of transformations leading from input to output makes up the CAP. CAPs define the relationships between input and output that could be causative. In the deep learning family of machine learning techniques, data models are developed in a way that makes them specific to a certain task. For a range of tasks, including pattern recognition, decision-making, classification, speech recognition, and picture and speech-to-text conversion, deep learning in neural networks is often used.

4.6.1 Convolutional Neural Network (CNN):
A neural network type called a convolutional neural network, or CNN or ConvNet, is particularly adept at processing input with a grid-like architecture, like an image. A binary representation of visual data is a digital picture. It is made up of a grid-like arrangement of pixels, each of which has a pixel value to indicate how bright and what colour it should be. In the feature extraction step, several filters and layers are used to extract information and features from the photos. Once this phase is complete, the images are sent on to the classification phase, where they are categorised according to the problem's target variable.

5. RESULTS & DISCUSSION

5.1 Accuracy using Support Vector Machine (SVM):

	precision	recall	f1-score	support
angry	0.70	0.77	0.73	158
disgust	0.67	0.77	0.72	150
fear	0.72	0.78	0.75	180
happy	0.68	0.65	0.66	150
neutral	0.80	0.81	0.81	204
sad	0.78	0.62	0.69	160
surprise	0.79	0.71	0.75	150
accuracy			0.73	1152
macro avg	0.73	0.73	0.73	1152
weighted avg	0.74	0.73	0.73	1152

----accuracy score 73.4375 ----

Fig 5.1. Confusion matrix of SVM classifier

5.2 Table of Actual vs Predicted emotion using SVM:

	Actual	Predicted
5542	neutral	neutral
883	fear	fear
5534	neutral	neutral
1779	sad	sad
4454	angry	angry
...	...	...
1968	happy	neutral
1836	fear	fear
4103	neutral	neutral
2193	neutral	neutral
2062	angry	angry

1152 rows × 2 columns

Fig 5.2. Actual vs Predicted table of SVM classifier

5.3 Accuracy using Random Forest (RF):

0.8324652777777778

	precision	recall	f1-score	support
angry	0.82	0.88	0.85	153
disgust	0.85	0.87	0.86	172
fear	0.90	0.74	0.81	141
happy	0.82	0.81	0.81	124
neutral	0.79	0.93	0.86	253
sad	0.92	0.72	0.81	155
surprise	0.80	0.80	0.80	154
accuracy			0.83	1152
macro avg	0.84	0.82	0.83	1152
weighted avg	0.84	0.83	0.83	1152

```
[[134   4   1   1   3   1   9]
 [  5 150   1   2  10   2   2]
 [ 11   4 105   3  10   4   4]
 [  3   0   7 100   8   1   5]
 [  0   7   0   7 235   1   3]
 [  2   9   1   6  17 112   8]
 [  9   3   2   3  13   1 123]]
```

Fig 5.3. Confusion matrix of RF classifier

5.4 Table of Actual vs Predicted emotion using RF:

	Actual	Predicted
5542	neutral	neutral
883	fear	fear
5534	neutral	neutral
1779	sad	sad
4454	angry	angry
...	...	...
1968	happy	happy
1836	fear	fear
4103	neutral	neutral
2193	neutral	disgust
2062	angry	angry

1152 rows × 2 columns

Fig 5.4. Actual vs Predicted table of RF classifier

5.5 Confusion Matrix and Heat Map using CNN:

```
Confusion matrix, without normalization
[[135   1   1   2   2   0   1]
 [  1 142   0   1   3   3   1]
 [  4   0 137   5   2   1   0]
 [  0   1   1 155   4   1   2]
 [  0   0   1   1 227   2   0]
 [  0   0   3   6   6 143   0]
 [  1   0   1   2   1   1 161]]
```

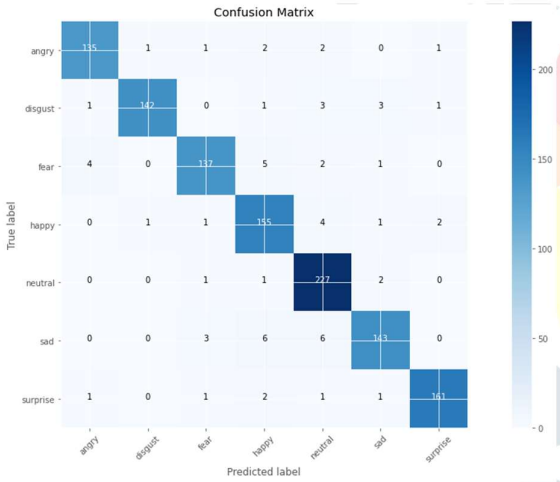

Fig 5.5. Confusion Matrix and heat map using CNN

Table 1: Accuracy for all 3 models

Algorithms	Accuracy
Support Vector Machine	73.4% without scaling; 78.9% with scaling
Random Forest	83.24%
Convolutional Neural Network	94.66%

6. CONCLUSION

The conclusion of human emotion is a complex task that can be used as a level for any model of feeling recognition. It employs a variety of feelings, including disgust, anger, fear, surprise, happy, sad, calm, and neutral. This project's implementation will allow us to apply machine learning and deep learning to detect the emotion in speech, which will enhance human-computer connection. As a result, three different types of features (MFCC, MS and Chroma) were taken from the RAVDESS database, and a mixture of these characteristics was then provided. These features were used to develop and test machine learning models that could identify emotional states. In order to categorise eight emotions, we presented a speech emotion recognition (SER) system that makes use of two machine learning algorithms and one deep learning technique:

- Random Forest
- Support Vector Machine
- Convolutional Neural Networks

Three types of features were extracted from the RAVDESS database:

- Mel Frequency Cepstral Coefficients
- Chroma
- Mel Spectrogram

These characteristics were combined and presented. We looked examined how classifiers and characteristics affect the precision with which speech emotions are recognised. Using the Random Forest classifier, we achieved an accuracy of 83.24 percent, using Support Vector Machine classifier, an accuracy of 73.4 percent, and using Convolutional Neural Networks, an accuracy of 94.66 percent.

7. FUTURE WORK

This technology can be used for marketing, enhancing customer service in contact centres, and creating virtual voice-based assistants who can recognise human emotion and respond appropriately. This model can also be used by multiple industries to other different services like marketing company suggesting consumers to buy products based on their emotion. Based on knowledge of the driver's mental state, SER is employed as an in-car board system. To ensure their safety and stop mishaps from happening, the device can receive speech input from them.

REFERENCES

[1] Ringeval, F.; Sonderegger, A.; Sauer, J.S.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013.

[2] Lotfian, R.; Busso, C. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. IEEE Trans. Affect. Comput. 2019.

[3] Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.

[4] Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access 2019.

[5] Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. IEEE Trans. Multimedia. 2018.

[6] New, T.L.; Foo, S.W.; Silva, L.C.D. Classification of stress in speech using linear and nonlinear features. In Proceedings of the 2003.

[7] IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 Proceedings (ICASSP '03), Hong Kong, China, 6–10 April 2003.

[8] Hossain, M.S.; Muhammad, G.; Song, B.; Hassan, M.M.; Alelaiwi, A.; Alamri, A. Audio– Visual motion-Aware Cloud Gaming Framework. IEEE Trans. Circuits Syst.Video Technol. 2015.

[9] Oh, K.; Lee, D.; Ko, B.; Choi, H. A Chatbot for Psychiatric Counselling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation. In Proceedings of the 2017 18th IEEE International Conference on Mobile Data Management (MDM), Daejeon, Korea, 29 May–1 June 2017 Speech Emotion Recognition using Machine Learning Department of Electronics and Communication Engineering, NMIT, Benagaluru-560064 18.

[10] N. Morgan, ―Deep and wide: Multiple layers in automatic speech recognition, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 7–13, 2012.
[11] J.M. Garcia-Haro, E.D. Oña, J. Hernandez-Vicen, S. Martinez and C. Balaguer, "Service Robots in Catering Applications: A Review and Future Challenges", Electronics, vol. 10, no. 1, pp. 47, 2021.