INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 04 | APRIL - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

Machine Learning Based Speech Emotion Recognition System

B Chakradhar[1] D Bhavana[2] G Rakesh[3] M Chandrashekar[4] A Sandeep[5]
Assistant Professor, Department of CSE, Raghu Engg. College, Visakhapatnam, AP, India.[1]
B.Tech Students, Department of CSE, Raghu Engg. College, Visakhapatnam, AP, India.[2,3,4,5]

***Abstract*- In the last decade, there has been significant research into Automatic Speech Emotion Recognition (SER). The primary goal of SER is to improve human-machine interfaces. It can also monitor someone's psychological state for lie detection applications. Recently, speech emotion recognition has found uses in medicine and forensics. This paper recognizes 7 emotions using pitch and prosody features. The majority of speech features used here are in the time domain. A Support Vector Machine (SVM) classifier categorizes the emotions. The Berlin emotional database was used for this task. A good recognition rate of 81% was achieved. The reference paper for this work recognized 4 emotions and obtained a 94.2% recognition rate. However, the reference paper used a more complex hybrid classifier, while this work focuses on recognizing more emotions with a simpler model.**

.

KeyWords: Blindness, Smart Stick, BlinDar, GPS, ESP8266, Internet of Things, RF Tx/Rx, MQ2.

I.INTRODUCTION

Human emotions are difficult to quantify. However, facial expressions and speech can provide clues into someone's emotional state. For example, speech conveys information about the message, speaker, language, and emotions. Speech recognition systems aim to identify emotions based on vocal properties, though speech can be an unreliable indicator of true feelings even for humans.

Speech emotion recognition using machine learning has a broad scope and numerous potential applications across various fields. Here's an overview of the scope of work involved:

1. **Data Collection and Preprocessing**: Gathering speech datasets containing recordings of individuals expressing different emotions. Preprocessing involves tasks like noise reduction, feature extraction (e.g., Mel-frequency cepstral coefficients), and normalization.
2. **Feature Extraction**: Extracting relevant features from the speech signals that capture the emotional content. These features may include pitch, intensity, formant frequencies, and spectral characteristics.
3. **Model Selection**: Choosing appropriate machine learning models for the task. Common choices include Support Vector Machines (SVMs), Random Forests, Gradient Boosting Machines (GBMs), deep learning architectures like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), and more recent architectures like Transformers.
4. **Model Training**: Training the chosen model(s) on the labeled dataset. This involves splitting the data into training, validation, and test sets, and tuning hyper parameters to optimize performance.
5. **Evaluation Metrics**: Defining appropriate metrics to evaluate the performance of the models. These may include accuracy, precision, recall, F1-score, and confusion matrices.
6. **Model Evaluation and Validation**: Evaluating the trained models on the test set to assess their performance in classifying emotions from unseen data. Cross-validation techniques may also be employed to ensure the robustness of the models.

7. **Model Interpretability**: Understanding how the models make predictions and interpreting their decisions. Techniques like feature importance analysis and model visualization can help in this regard.
8. **Deployment and Integration**: Integrating the trained model into real-world applications. This could involve deploying the model as part of a speech-enabled system, such as virtual assistants, customer service bots, or sentiment analysis tools.
9. **Continuous Improvement**: Continuously refining and updating the model based on feedback and new data to improve its performance and adaptability to changing contexts.

Ethical Considerations: Considering the ethical implications of speech emotion recognition systems, including privacy concerns, bias in the data or models, and potential misuse of the technology.

Overall, speech emotion recognition using machine learning is a multidisciplinary field that combines aspects of signal processing, machine learning, psychology, and human-computer interaction to develop systems that can accurately interpret and respond to human emotions conveyed through speech.

The major motivation behind this work comes from a desire to improve the naturalness and efficiency of human-machine interaction. The reference paper that was chosen has only been able to successfully recognize four emotions. The work presented here has classified seven emotions with an overall good recognition rate. In general, speech analysis systems use various techniques to extract characteristics from the raw signal. For emotions, the relevant information lies in pitch, prosody, and voice quality. The next step in this methodology is to identify the features that discriminate the labeled training speech data and discard non-discriminative features. This is achieved by calculating cross-validation between parameters, creating a grid of parameters, and selecting the one with the highest cross-validation.

II. PROPOSED TOPOLOGY

The Emotional profiles (EP) are constructed using SVM with Radial Basis Function (RBF). Emotion-specific SVMs are trained for each class as self-versus others classifiers. Each EP contains n-components, one for the output of each emotion-specific SVM. The profiles are created by weighting each of the n-outputs by the distance between the individual point and the hyper plane boundary. The final emotion is selected by classifying the generated profile. This is done by one vs one comparing of each emotion to the existing profile of the emotion.

Fig.1 comprehensively explains the methodology followed in this paper. Emotion recognition is done using two modules. The first module is the feature extraction module and the second is the classifier module. In the feature extraction module, we have used a feature set comprising pitch, prosody and voice quality features. Several classifiers exist for the task of emotion recognition.

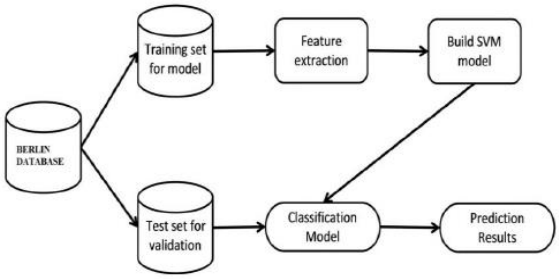


Fig. 1 Block Diagram of the proposed Speech Recognition System

The different classifiers are SVM, MLP (MultiLayer Perceptron), HMM(Hidden Markov Model), GMM(Gaussian Mixture Model), ANN(Artificial Neural Networks) etc. The SVM classifier yields good results even from small test samples and hence it is widely used for speech emotional recognition [3][4][5][6]. The SVM

classifier is therefore used for the proposed work. Because of the Structural Risk Minimization, SVM classifiers usually have better performance than others.

The Berlin Emo DB is the speech corpus used for training and testing[2] The Berlin emotional database consists of 10 speakers (5 male and 5 female). Each one of the speakers is asked to speak 10 different texts in German. The database consists of 535 speech files. The speech files are labeled into 7(Table I) emotion categories *anger*, *boredom, disgust, fear,*
happiness, sadness and neutral.

A. Pitch

The voiced regions looks like a near periodic signal in the time domain representation. In a short term, we may treat the voiced speech segments to be periodic for all practical analysis
and processing. The periodicity associated with such segments is defined is 'pitch period T◦' which gives 'Fundamental Frequency F◦'.

B. Entropy

It expresses the abrupt change in the signal.

C. Auto Correlation

It is the correlation of the signal with itself. It is the similarity between observations as a function of the time lag between them. It is a mathematical tool for finding repeating patterns, fundamental frequency or noise in signal.

D. Energy

The amplitude of the speech signal varies appreciably with time. Short Time energy provides a convenient representation that reflects these amplitude variations. The major significance
of this is that it provides a basis for distinguishing voiced speech from unvoiced speech.

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} \left(x[m] \times w[\hat{n}-m]\right)^2 = \sum_{m=-\infty}^{\infty} x^2[m] \times w^2[\hat{n}-m] \quad (1)$$

x[m] = Amplitude of Speech Signal
w[n] = Window Function.

E. Jitter and Shimmer

A frequent back and forth changes in amplitude (from soft to louder) in the voice is shimmer. Shimmer Percent provides an evaluation of the variability of the peak-to-peak amplitude within the analyzed voice sample. Jitter represents the relative period-to-period (very short-term) variability of the peak-to peak amplitude. It is defined as varying pitch in the voice, which causes a rough sound. Compared to shimmer, which describes varying loudness in the voice, Jitter is the undesired deviation from true periodicity of an assumed periodic signal.

Jitter Percent provides an evaluation of the variability of the pitch period within the analyzed voice sample. It represents the relative period-to-period (very short-term) variability.

$$Jitter = \frac{\left|(T_0)_i - (T_0)_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N}(T_0)_i} \quad (2)$$

$$Shimmer = \frac{|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (3)$$

F. HNR

It provides an indication of the overall periodicity of the voice signal by quantifying the ratio between the periodic (harmonic part) and a periodic (noise) components. It describes quality of speech hence a important parameter in emotion recognition.

$$HNR = 10\log_{10}\left\{\frac{\sum_{i}^{NFFT/2}|S_i|}{\sum_{i}^{NFFT/2}|N_i|}\right\} \quad (4)$$

G. ZCR

It is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. It is an important parameter to understand the variation of speech, it is also useful in differentiating between voiced and unvoiced speech.

$$Z_{\hat{n}} = \sum_{m=-\infty}^{\infty} 0.5\,|\mathrm{sgn}\{x[m]\} - \mathrm{sgn}\{x[m-1]\}| \times w[\hat{n} - m] \quad (5)$$

Z = Zero Crossing Rate

H. Statistics

•*Standard Deviation*: Standard deviation (represented by the symbol sigma) shows how much variation or dispersion exists from the average (mean), or expected value.

• *Spectral Centroid*: It is the weighted mean frequency. It indicates where the center of mass of the spectrum lies.

The spectral centroid is a good predictor of the brightness of a sound. Brightness here refers to the energy content of speech with time.

• *Spectral Flux*: It is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing

the power spectrum for one frame against power spectrum for the previous frame. More precisely, it is usually

calculated as the Euclidean distance between the two normalized spectra.

• *Spectral Roll off* : Spectral Roll off point is defined as the *nth* percentile of the power spectral distribution, where n is usually 85% or 95%.

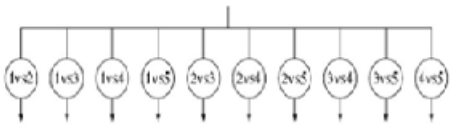


Fig. 3.1. One-vs-one max wins SVM voting scheme

III. TEST DIFFERENT AUDIO FILES

	speech	label
0	/content/drive/MyDrive/archive (21)/TESS Toron...	fear
1	/content/drive/MyDrive/archive (21)/TESS Toron...	fear
2	/content/drive/MyDrive/archive (21)/TESS Toron...	fear
3	/content/drive/MyDrive/archive (21)/TESS Toron...	fear
4	/content/drive/MyDrive/archive (21)/TESS Toron...	fear
...	...	...
2795	/content/drive/MyDrive/archive (21)/TESS Toron...	disgust
2796	/content/drive/MyDrive/archive (21)/TESS Toron...	disgust
2797	/content/drive/MyDrive/archive (21)/TESS Toron...	disgust
2798	/content/drive/MyDrive/archive (21)/TESS Toron...	disgust
2799	/content/drive/MyDrive/archive (21)/TESS Toron...	disgust

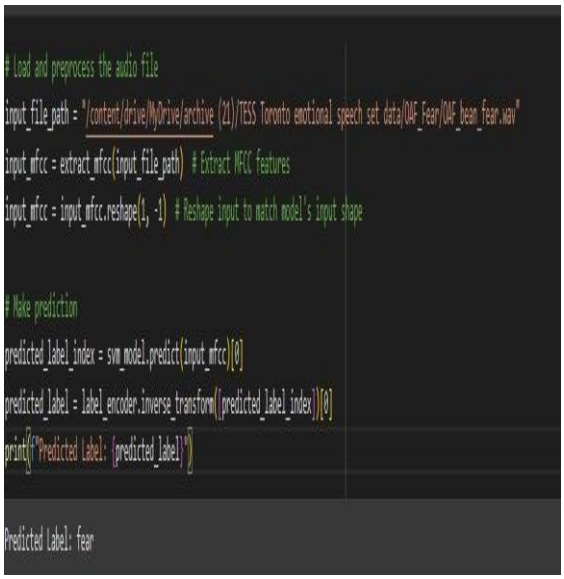
2800 rows × 2 columns

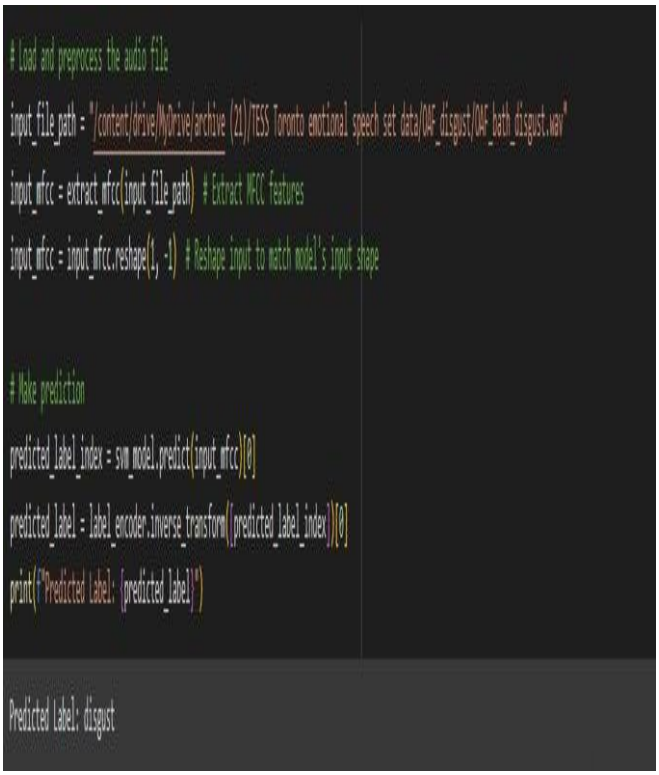
```
df["label"].unique()
```

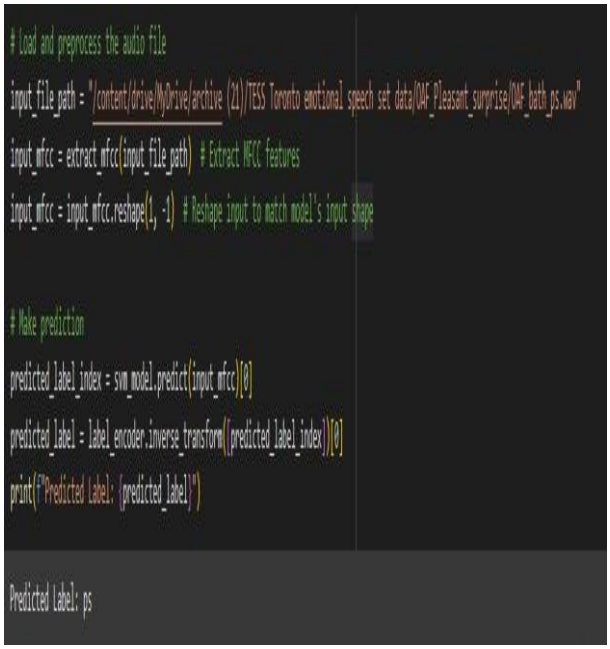
array(['fear', 'sad', 'happy', 'neutral', 'ps', 'angry', 'disgust']

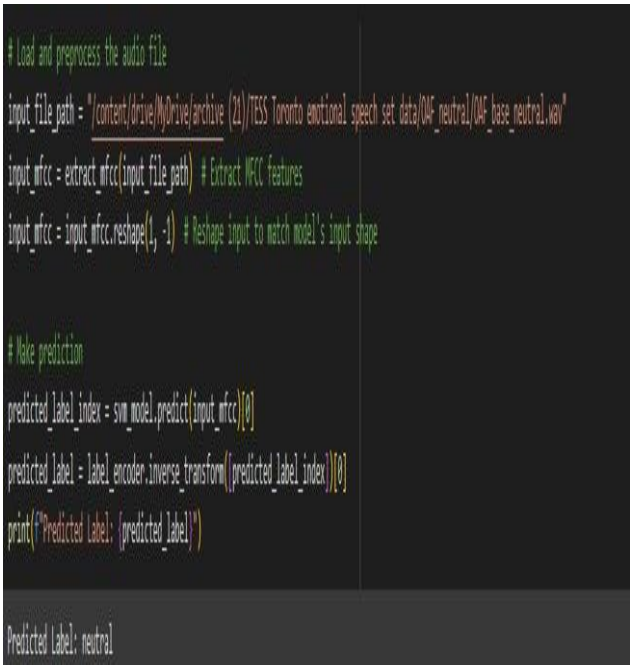
IV.RESULT

RESULT OUTPUT FOR DIFFERENT EMOTIONS









```
# Load and preprocess the audio file
input_file_path = "/content/drive/MyDrive/archive (21)/TESS Toronto emotional speech set data/YAF_angry/YAF_beg_angry.wav"
input_mfcc = extract_mfcc(input_file_path)  # Extract MFCC features
input_mfcc = input_mfcc.reshape(1, -1)  # Reshape input to match model's input shape

# Make prediction
predicted_label_index = svm_model.predict(input_mfcc)[0]
predicted_label = label_encoder.inverse_transform([predicted_label_index])[0]
print(f"Predicted Label: {predicted_label}")
```

```
Predicted Label: angry
```

```
from IPython.display import Audio

# Path to your recorded audio file
recorded_audio_file = "/content/recorded_audio2.wav"

# Load and preprocess the recorded audio file
input_mfcc = extract_mfcc(recorded_audio_file)  # Extract MFCC features
input_mfcc = input_mfcc.reshape(1, -1)  # Reshape input to match model's input shape

# Make prediction
predicted_label_index = svm_model.predict(input_mfcc)[0]
predicted_label = label_encoder.inverse_transform([predicted_label_index])[0]
print(f"Predicted Label: {predicted_label}")

# Play the recorded audio file
Audio(recorded_audio_file)
```

```
Predicted Label: disgust
```

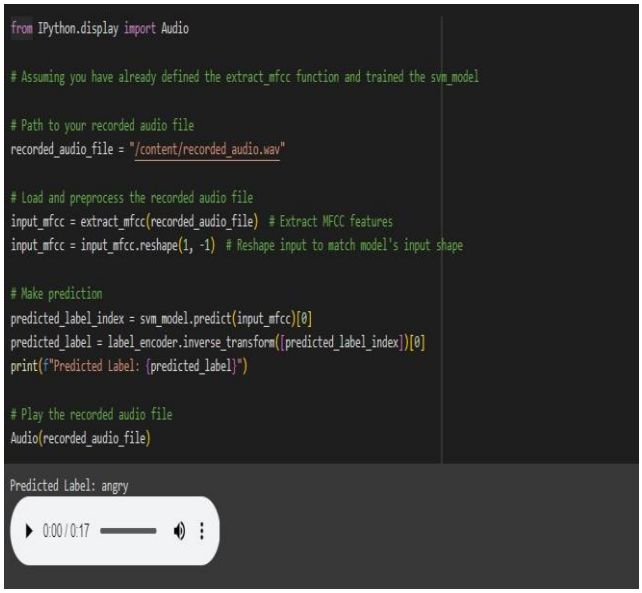
0:12 / 0:29

```
# Load and preprocess the audio file
input_file_path = "/content/drive/MyDrive/archive (21)/TESS Toronto emotional speech set data/YAF_sad/YAF_base_sad.wav"
input_mfcc = extract_mfcc(input_file_path)  # Extract MFCC features
input_mfcc = input_mfcc.reshape(1, -1)  # Reshape input to match model's input shape

# Make prediction
predicted_label_index = svm_model.predict(input_mfcc)[0]
predicted_label = label_encoder.inverse_transform([predicted_label_index])[0]
print(f"Predicted Label: {predicted_label}")
```

```
Predicted Label: sad
```

```
# Load and preprocess the audio file
input_file_path = "/content/drive/MyDrive/archive (21)/TESS Toronto emotional speech set data/YAF_happy/YAF_back_happy.wav"
input_mfcc = extract_mfcc(input_file_path)  # Extract MFCC features
input_mfcc = input_mfcc.reshape(1, -1)  # Reshape input to match model's input shape

# Make prediction
predicted_label_index = svm_model.predict(input_mfcc)[0]
predicted_label = label_encoder.inverse_transform([predicted_label_index])[0]
print(f"Predicted Label: {predicted_label}")
```

```
Predicted Label: happy
```



V CONCLUSION

This paper presents an approach to speech emotion recognition using a simple SVM classifier, achieving 81% accuracy. Disgust proved the most challenging emotion to recognize, even for humans, due to its complex nature. Compared to the reference paper, the proposed method obtains good overall accuracy across seven emotions using a compact feature set. Future work could improve recognition rates further by exploring hybrid classifiers [4]. Although the reference paper reports better accuracy, this work recognizes more emotions with high accuracy, without relying on complex cepstral features. This simplifies the system and reduces runtime significantly.

REFERENCES

[1] B. W. a. T. G. Lingli Yu, ”A hierarchical support vector machine based on feature-driven method for speech emotion recognition,” Artificial Immune Systems - ICARIS, pp. 901-907, 2013.

[2] http://pascal.kgw.tu-berlin.de/emodb/index-1280.html

[3] R. P. a. T. P. Alexander Schmitt, ”Advances in Speech Recognition,” Springer, pp. 191 200, 2010.

[4] A. Joshi, ”Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm,” International Journal of Advanced Research in Computer Science and Software Engineering, pp. 387-392, 2013.

[5] D. S. L. N. Akshay S. Utane, ”Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine,” International Journal of Scientific & Engineering Research, no. 5, pp. 1439-1443, 2013.

[6] K. S. R. S. G. Koolagudi, ”Emotion recognition from speech using global and local prosodic,” Springer, 2012.

[7] Soegaard, M. and Friis Dam, R. (2013). The Encyclopedia of Human-Computer Interaction. 2nd ed.

[8] Developer.amazon.com. (2018). Amazon Alexa. [online] Available at: https://developer.amazon.com/alexa

[9] Store.google.com. (2018). Google Home Tips & Tricks – Google Store. [online] Available at: https://store.google.com/product/google_home_learn

[10] Apple. (2018). iOS - Siri. [online] Available at: https://www.apple.com/ios/siri/

[11] The Official Samsung Galaxy Site. (2018). What is S Voice?. [online] Available at: http://www.samsung.com/global/galaxy/what-is/s-voice/ [Accessed 2 May 2018].

[12] Gartner.com. (2018). Gartner Says 8.4 Billion Connected. [online] Available at: https://www.gartner.com/newsroom/id/359891

[13] H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[14] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[15] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

[16] J. Rong, G. Li, and Y.-P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.