International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Computer Speech Recognition Using Machine Learning

A. Sri Laxmi Prasanna [a], B. Teja Goud [b], M. Priyanka [c]

[a] *Student Nalla Narasimha Reddy Education Society's Group of Institutions (Autonomous), Affiliated By Jntuh Hyderabad, Department Of Electronics and Communication Engineering , Hyderabad, Telangana, Pin Code 500088, India*
[b] *Student Nalla Narasimha Reddy Education Society's Group of Institutions (Autonomous), Affiliated By Jntuh Hyderabad, Department Of Electronics and Communication Engineering , Hyderabad, Telangana, Pin Code 500088, India*
[b] *Student Nalla Narasimha Reddy Education Society's Group of Institutions (Autonomous), Affiliated By Jntuh Hyderabad, Department Of Electronics and Communication Engineering , Hyderabad, Telangana, Pin Code 500088, India*
Under the Guidence of Swetha .T, ASSISTANT PROFESSOR ,Nalla Narasimha Reddy Education Society's Group of Institutions

ABSTRACT

Both the task of accurately predicting human emotion from speech and the accuracy of the prediction are involved in speech emotion recognition. It leads to improved human-computer interaction. Although emotions are arbitrary and difficult to annotate in audio, "Speech Emotion Recognition(SER)" makes it possible to predict someone's emotional state. Animals like dogs, elephants, horses, and others use the same idea that explains how they can understand human emotion. You can predict someone's emotional condition by listening to their tone, pitch, expression, behavior, and other characteristics. They are regarded to have limited verbal emotional expression skills. Using a limited number of examples, the classifiers are taught to identify speech emotion.

Keywords: Support Vector Machine, Machine Learning, RAVDESS, MFCC, Mel Spectogram, chroma.

1. Introduction

A voice signal is one of the quickest and most natural ways for people to communicate with one another. The quickest and most effective way for human-machine interaction is speech signals. All of a person's senses are utilized to the fullest extent possible to ensure awareness of the message they have received. While it comes naturally to humans, emotional recognition is a tremendously difficult assignment for machines. Therefore, an emotion recognition system makes use of emotion knowledge in a way that enhances communication between humans and machines. The feminine or make speakers' emotions are recognized in speaking through speech. Some of the examined speech features are the linear prediction cepstrum coefficient (LPCC), fundamental frequencies, and the Mel+ frequency cepstrum coefficient (MFCC).

These attributes form the foundation of speech processing. It is unclear which speech elements are more effective in differentiating between different emotions, which makes emotion detection from speakers' speech highly challenging. Due to the existence of and various forms of speech together with some of the existing datasets that have been the subject of previous studies. The fourth section briefly describes alternative feature extraction methods for identifying speech emotions before focusing on a review of the classification portion. We have discussed KNN, SVM, CNN, recurrent neural network, etc. in this part. The application of deep learning for voice emotion recognition is briefly discussed in the sixth part. Various speaking rates, styles, sentences, and speakers, there is an introduction of accosting variability that influences the aspects of speech

2. METHODOLOGY

A machine learning (ML) model is used to create the voice emotion detection system. Similar to every other ML project, the implementation process includes additional fine-tuning processes to improve the model's performance. A visual summary of the procedure is provided by the flowchart (see Figure 1). Data collection is the initial phase, which is crucial. The data that is being provided to the model-building process will allow it to learn, and the data will serve as the basis for all the decisions and output that the model will produce. The second step, referred to as feature engineering, consists of a number of machine learning operations that are carried out on the gathered data.

These steps take care of the various data representation and data quality problems. The third step, where an algorithmic-based model is constructed, is frequently regarded as the core of an ML project. To understand the data and prepare itself to react to any incoming data, this model employs a machine learning (ML) method. The built model's functionality must be assessed as the last stage.

The process of creating a model and analysing it is frequently repeated by developers in order to compare the effectiveness of various algorithms. Results of comparisons aid in selecting the ML method most appropriate to the issue.

2.1 Models for training and testing

The system receives training data, which includes expression labels and weight training for that network. The input is an audio file. The audio is then subjected to intensity normalization. To prevent the impact of the presenting sequence of the samples from affecting the training performance, the Convolutional Network is trained using a normalized audio. The weight collections that are produced as a result of this training procedure produce the best outcomes when used with the learning data. The dataset retrieves the system with pitch and energy during testing, and based on the final network weights learned, it provides the identified emotion. Each of the five expressions that make up the output is represented as a numerical value in the output. Based on the individual's bpm value, three emotions—Relaxed/Calm, Joy/Amusement, and Fear/Anger—are identified. Based on the theories of "color psychology" and "shape psychology," the colors and shapes used in the created art are similar to the emotions observed.

2.2 Speech Database:

In this study, various speech databases are used to validate the suggested techniques for identifying speech emotions. Berlin and AIBO are the two datasets that are most frequently utilized. German-language actors recorded Burkhardt et al. Technical University Berlin's Department of Technical Acoustics served as the location of record. Five male and five female German actors each read one of the selected sentences as part of the dataset contribution. Various documented emotions include disgust, indifference, fear, anger, happiness, and sadness. Aibo, a robot created by Sony that is controlled by a human operator, was used to play with and interact with 51 youngsters in order to gather data for another emotional database. The five collected emotions in AIBO are emphatic, angry, positive, neutral, and positive..

2.3 Extracting Features:

The next step is to extract the features from the audio files that will aid our model in differentiating between them. To extract features One of the libraries used for audio analysis in Python is the LibROSA library, which is what we utilize. Since its invention in the 1980s, MFCCs have been the cutting-edge feature while doing Speech Recognition jobs. What sound emanates depends on this geometry. We should be able to appropriately represent the phoneme being produced if we can precisely establish the form. The short time power spectrum envelope exhibits the form of the vocal tract, and it is the responsibility of MFCCs to faithfully represent this envelope.

2.4 Classification Methods:

A variety of classification techniques, including support vector machines (SVM), hidden Markov models (HMM), neural networks, Knearest neighbours, and Gaussian mixture models (GMM), are used to develop appropriate classifiers for modeling emotional states. On the other hand, a standard level classifier might not succeed on highly emotional statuses. As an illustration, the ranking SVM technique does not significantly increase emotion recognition when compared to the combination of SVM and RBF. Some hybrid/fusion-based techniques outperform standalone strategies in terms of recognition rate.

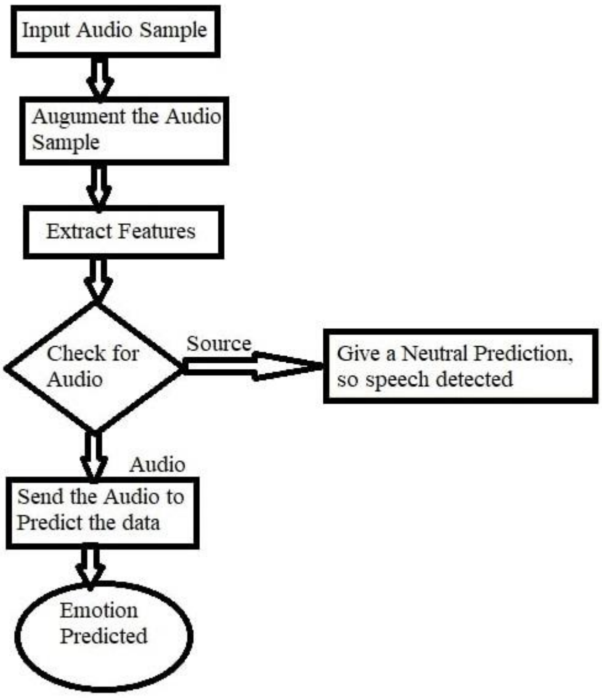
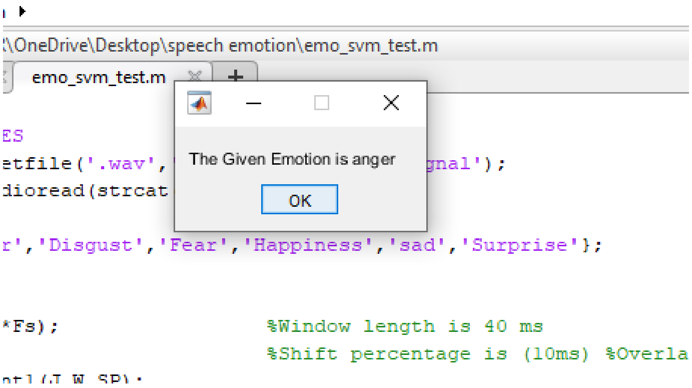

Input Audio Sample
Augument the Audio Sample
Extract Features
Check for Audio
Source
Give a Neutral Prediction, so speech detected
Audio
Send the Audio to Predict the data
Emotion Predicted

Figure 1: Training process workflow



```
:\OneDrive\Desktop\speech emotion\emo_svm_test.m
emo_svm_test.m

ES
etfile('.wav',                      gnal');
dioread(strcat

r','Disgust','Fear','Happiness','sad','Surprise'};


*Fs);                   %Window length is 40 ms
                        %Shift percentage is (10ms) %Overla
```

The Given Emotion is anger

3. DESCRIPTION OF BLOCK DIAGRAM

Numerous factors in the voice signal that convey emotional traits are present. What features should be employed is one of the challenging issues in emotion recognition. Energy, pitch, formant, as well as other spectrum features including linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC), and modulation spectral features, are only a few of the frequent aspects that have been retrieved in recent study. To extract the emotional features for this work, we have chosen MFCC and modulation spectral features.

The most popular way to characterizes the spectral characteristics of voice signals is via the mel-frequency cepstrum coefficient (MFCC). These are the finest for voice recognition because they take into account how sensitively humans perceive frequencies. The Fourier transform and the energy spectra were calculated for each frame and then transferred into the Mel-frequency scale. The first 12 DCT coefficients of the calculated discrete cosine transform (DCT) of the Mel log energies generated the MFCC values used in the classification procedure. The procedure for determining MFCC is often depicted. In our study, where speech signals are captured at 16 KHz, we extract the first 12 order of the MFCC coefficients. We compute the mean, variance, standard deviation, kurtosis, and skewness for each order coefficient, and we do this for all other frames in an utterance. A 60-dimensional feature vector makes up each MFCC.

A long-term spectro-temporal representation inspired by auditory signals is used to extract modulation spectral features (MSFs). These characteristics are generated by simulating the human auditory system's spectro-temporal (ST) processing, which takes into account both the regular acoustic frequency and the modulation frequency. Figure 3 shows the processes for computing the ST representation. The spoken input is initially divided by an auditory filter bank (19 filters in all) to produce the ST representation. The modulation signals are created by computing the Hilbert envelopes of the critical-band outputs. The Hilbert envelopes are then further processed via a modulation filter bank to perform frequency analysis. The proposed features are known as modulation spectral features (MSFs), which refer to the spectral components of the modulation signals. Recursive feature elimination (RFE) selects either the best- or worst-performing feature and then eliminates it using a model (such as linear regression or SVM). Recursive feature elimination (RFE) aims to pick features by repeatedly taking into account smaller and smaller sets of features since these estimators assign weights to features (for instance, the coefficients of a linear model). The prediction strength of each feature is then determined after the estimator has been trained on the first set of features. Then, the existing set of features is reduced in size by eliminating the least crucial features. Once the appropriate number of features to pick has been reached, the technique is recursively repeated on the pruned set. The recursive feature elimination method of feature selection was implemented in this study. Once the appropriate number of features to pick has been reached, the technique is recursively repeated on the pruned set. In this study, we used basic linear regression to create the recursive feature elimination method of feature ranking (LR-RFE). Another linear model, such as SVM-RFE, an SVM-based feature selection approach, is also used in research with RFE. Guyon et al. chose significant feature sets using SVM-RFE. It can decrease classification computing time while also increasing classification accuracy rate.

4. RESULTS

When the speech signal is given as input. Emotions are recognized as follows.

FIGURE 6: MAGNITUDE SPECTRUM OF SPEECH AND IT'S SPECTRAL ENVELOPE DERIVED FROM THE SYNTHESIS FILTER USING LPC COEFFICIENTS

6. CONCLUSION

In this paper, a new speech enhancement algorithm using Linear predictive coding techniques and adaptive synthesis filter has been presented. The proposed approach proves its reliability to improve the speech intelligibility without affecting the signal quality referring to the performance evaluation such as SNR. Finally, the real-time test of speech denoising has been successfully implemented in TMS320C6713 platform and reveals that the proposed algorithm has significantly improved the speech intelligibility**.**

References

1. ITU-T Recommendation (1996) G.723.1, Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 & 6.3 kbit/s, March.
2. CCITT Recommendation (1992) G.728, Coding of Speech at 16 kbit/s Using Low-delay Code Excited Linear Prediction.
3. ITU-T Recommendation (1995) G729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS- ACELP), December.
4. ITU-T Recommendation (2002) G722.2, Wideband Coding of Speech at Around 16 kbit/s Using Adaptive Multi-rate Wideband (AMR-WB), January.
5. 3G TS 26.190 V1.0.0 (2000-12) (2000) Mandatory Speech CODEC Speech Processing Functions; AMR Wideband Speech CODEC; Transcoding Functions (Release 4), December.
6. 3GPP TS 26.171 (2001) Universal Mobile Telecommunications System (UMTS); AMR Speech CODEC, Wideband; General Description (Release 5), March.
7. Bessette, B., Salami, R., Lefebvre, R. et al. (2002) The adaptive multi-rate wideband speech codec (AMR-WB). IEEE Trans. Speech Audio Process., 10 (8), 620–636.
8. Kondoz, A.M. (1995) Digital Speech Coding for Low Bit Rate Communications Systems, John Wiley & Sons, Inc., New York.
9. Kuo, S.M. and Morgan, D.R. (1996) Active Noise Control Systems – Algorithms and DSP Implementations, John Wiley & Sons, Inc., New York.
10. Tian, W., Wong, W.C., and Tsao, C. (1997) Low-delay subband CELP coding of wideband speech. IEE Proc. Vision Image Signal Process., 144, 313–316