VOCULAR: SPEECH EMOTION RECOGNITION

Akshita Jain[1], Vikas[2], Lakshay Singhal[3], Tanishqa[4], Ms. Anjani Gupta[5]
[1,2,3,4] Students, [5]Assistant Professor
, Department of Information Technology
[1, 2,3,4,5] Dr. Akhilesh Das Gupta Institute of Technology & Management

Abstract: ***Speech Emotion recognition is gaining more demand and need for it is increasing enormously. Many systems have been developed, to identify the emotions from the speech signal. In this paper speech emotion recognition we used Decision Tree classifier and Convolution Neural Network. The classifiers are used to differentiate emotions such as anger, happiness, sadness, surprise, neutral, calm, disgust state, etc. The dataset for the speech emotion recognition system is the emotional speech samples and the emotion recognized from these speech samples is recorded and according to the recognized emotion the application can suggest some books, videos, quotes, events happening around from the database and can redirect the user so that he/she can view the details of the same. The classification performance is based on extracted features. Various datasets are explored for training emotion recognition model. The dataset used for emotion recognition is RAVDESS dataset. Final accuracy of this emotion recognition model using Convolution Neural Network(CNN) is 85%(~).***

Index Terms: ***emotion recognition; speech emotion recognition system; speech processing module and feature extraction; classifier selection and training method; conclusion;***

I. INTRODUCTION

Communication can be done in many ways but the most common, easy and natural method of communication among the humans is by speech signals. That's why speech can also be considered as the most efficient and fastest method between the humans and the machines. From all the senses available in the human beings, people can understand the emotional state of the other people through communication very easily. But this task is not that much easy for the machines. Therefore, the main purpose of making this technology of speech emotion recognition system is to use emotion related knowledge in such a way that communication between humans and machines can be improved to some extent.

In speech emotion recognition, the emotions from the audios of male or female speakers are acknowledged. In the past century some speech features were studied which involved the fundamental frequencies, Mel frequency cepstrum coefficient (MFCC), Linear Prediction Coefficients (LPC), linear prediction cepstrum coefficient (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT) and Perceptual Linear Prediction (PLP), etc., which form the basis for speech processing even today. In one of the research the spectrograms of real and acted emotional speech were Studied and found similar recognition rate for both, which recommend that later one can be use for the speech emotion recognition system. In another research a correlation between emotion and speech features were present. Further humans and machine emotion recognition rate was Compared, in which same recognition rates were found for both.

II. EMOTION RECOGNITION

Emotion recognition from the voice of the speaker is extremely difficult thanks to the next reasons: In differentiating between various emotions in particular of speech features which are more useful isn't clear. Due to the existence of the numerous sentences, speakers, various accents, speaking styles, speaking rates, accosting variability, speech features get directly affected. An equivalent utterance may show different emotions. Each emotion may correspond to the varied portions of the spoken utterance. Therefore it's extremely difficult to differentiate these portions of utterance. Another problem is that emotion expression is relying on the speaker and his or her culture and environment. Because the culture and environment gets change the speaking style also gets change, which is another challenge before the speech emotion recognition system. There could even be two or more kinds of emotions, future emotion and transient one, so it isn't clear which type of emotion the recognizer will detect.

Emotion recognition from the speech information could also be the speaker dependent or speaker independent. The different classifiers available are Decision Tree Classifier, Artificial Neural Network (ANN), k-nearest neighbors (KNN), and Support Vector Machine (SVM). The application of the speech emotion recognition system include the psychiatric diagnosis, intelligent toys, lie detection, within the decision center conversations which is that the foremost vital application for the automated recognition of emotions from the speech, in car board system where information of the psychological state of the driving force may provide to the system to start out his/her safety.

III. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition is the emotional processing unit that employs advances in the field of affective computing and machine learning in order to estimate the emotional state of the user. Speech signals extracts emotional feature using computer and analyses the characteristic parameters and finally concludes the emotional states. At present, speech emotion recognition appears to be the foremost emerging topic within the field of AI, besides, it had been a hot

research topic of signal processing.
Speech emotion recognition is usually beneficial for applications, which require human-computer interaction like speech synthesis, customer service, education, forensics, medical analysis, and entertainment and security fields.
This speech emotion recognition and processing system mainly involves three parts, i.e., Speech Signal, Speech Processing Module and Emotion Recognizer.
The structure diagram of speech emotion recognition is shown in Figure 1.

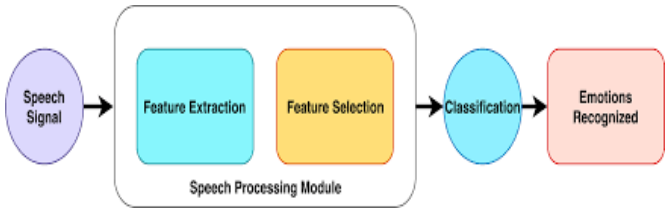

Figure 1: Structure of Speech Emotion Recognition System

In this system, the accuracy of speech emotion recognition directly depends on the quality of feature extraction of speech emotion. Within the method of feature extraction, it always took the entire emotion sentence as units for feature extracting, and extraction contents were four aspects of emotion speech, which were several acoustic characteristics of your time construction, amplitude construction, fundamental construction, and format construction.
The need to find out a datasets of the significant emotions by an automatic emotion recognizer system is a main trouble in speech emotion recognition system.
The evaluation of the speech emotion recognition system is based on the level of genuineness of the database which is employed as an input to the speech emotion recognition system. If the inferior database is employed as an input to the system then incorrect conclusion may be drawn. The datasets of the speech emotion recognition system may contain the important natural emotions or the acted ones.
The RAVDESS dataset is the Ryerson Audio Visual Database of Emotional Speech and Song dataset. This dataset has a total of 7356 files rated by 247 persons 10 times on the basis of emotional validity and genuineness. The entire dataset is of 24.8GB from 24 actors. Primary emotions are neutral, calm, happy, sad, angry, fearful, disgust, surprised.

IV. SPEECH PROCESSING MODULE AND FEATURE EXTRACTION

Before the extraction of the important attributes in the speech and identification, speech signals has to be processed to remove the noise.
The purpose of feature extraction is for instance a speech signal by a predetermined number of components of the signal. This is because all the information in the acoustic signal is too cumbersome to deal with, and some of the information is irrelevant in the identification task.
The speech waveform can be changed by accomplishing the feature extraction of speech emotion to a sort of parametric representation at a comparatively lesser rate for subsequent processing and analysis. This is usually called the front end signal-processing. It transforms the processed signal to a brief but logical representation that's more dependable. Front is the initial element within the sequence and the standard of the next features (pattern matching and speaker modeling) is suffering from the standard of the front end.
Therefore, acceptable classification springs from excellent and quality features. In present automatic speaker recognition (ASR) systems, the procedure for feature extraction has normally been to get a representation that's comparatively reliable for several conditions of an equivalent signal, even with the alterations within the environmental conditions, while retaining the portion that characterizes the knowledge.
Feature extraction approaches usually results in a multidimensional feature vector for each of the speech signal. A wide variety of options are available to represent the speech signal for the popularity process, like perceptual linear prediction (PLP) and Mel-frequency cepstrum coefficients (MFCC). MFCC is the best known and very popular.
Feature extraction is that the most relevant portion of speaker recognition. Features of speech have an important part within the segregation of a speaker from others. Feature extraction reduces the level of the unproductive part of the speech signal which can cause any damage to the facility of speech signal.
Before all the the features are extracted, there are many sequences of preprocessing phases that are carried out. The preprocessing step is pre-emphasis. This is achieved by passing the signal through a FIR filter which is usually a first-order finite impulse response (FIR) filter. This is succeeded by frame blocking, a way of partitioning the speech signal into frames. It removes the acoustic interface existing within the start and end of the speech signal.
The framed speech signal is then windowed. Band pass filter may be a suitable window that's applied to attenuate disjointedness at the beginning and finish of every frame. Hamming and Rectangular windows are the two famous categories of windows. It increases the sharpness of harmonics, removes the discontinuity of signal by tapering the beginnings and ending of the frame zero. The spectral distortion formed by the overlap is also reduced by this.
There are many emotional states. The feature of fear has a high average value. Therefore the statistics of pitch, energy and some spectrum features can be extracted to recognize and analyze the emotions from speech.
One of the main speech features which depict emotion is energy and the study of energy depends on the short term energy and short term mean amplitude. The pitch signal which is also known as glottal wave form is the main feature which indicates emotion in speech. The pitch signal depends on the tension of the vocal folds and the glottal air pressure. The pitch signal is characterized by two features that are pitch frequency, and the glottal air speed at the vocal opening time. Number of harmonics that are present in the spectrum gets affected by pitch frequency.
Speech signal may be a slow time varying signal, quasi-stationary, when observed over an adequately short period of your time between 5 and 100 milli sec, its behavior is

comparatively stationary. As results of this, short time spectral analysis which incorporates MFCC, LPCC and PLP are commonly used for the extraction of important information from speech signals. Noise may be a serious challenge encountered within the process of feature extraction, also as speaker recognition as an entire.

In feature extraction all the essential speech features that are extracted may not be helpful and important for speech emotion recognition. If all the extracted features gives as an input to the classifier this is able to not guarantee the simplest system performance which shows that there's a requirement to get rid of such a non-useful features from the base features. Therefore there's a requirement of systematic feature selection to scale back these features.

V. CLASSIFIER SELECTION AND TRAINING METHOD

In the speech emotion recognition system after calculation of the features, the simplest features are provided to the classifier. A classifier recognizes and analyzes the emotion within the speaker's speech. Various sorts of classifier are proposed for the task of speech emotion recognition. Bayesian Networks, Decision Trees, Gaussian Mixtures Model (GMM), K-nearest neighbors (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Linear Discriminate Analysis, Artificial Neural Network (ANN), etc. are the classifiers utilized in the speech emotion recognition system. Each classifier has some merits and demerits over the others.

The classifiers that we used in speech emotion recognition are Decision Trees and Convolution Neural Network (CNN). Decision Trees are a simplistic classifier which makes observations on data and maps these observations to decisions on class ownership. It functions by constantly querying a test instance to gain more information about which class it may belong through a combination of if-then rules. This study experimented with a number of data pre-processing approaches on prosodic features from speech samples achieving up the 68% accuracy for this classifier.

All voices labeled with respective emotions are prepared for training the model. The proposed CNN model was implemented using Tensor Flow. The RAVDESS is a validated database of emotional speech and songs. The database is balanced on the basis of gender consisting of 24 actors, vocalizing matching statements. The total of 7356 recordings were rated 10 times on carious features like emotional validity, intensity, loudness and genuineness.

The training took around 60 minutes and the best accuracy was achieved after 200 epochs. Training data model was performed sequentially. 67% of the data is trained and overall of 33% data is tested successfully. And finally the accuracy of 85.47 % was achieved. It is important to notice here that the overall accuracy is high.

VI. CONCLUSION

Speech emotion recognition systems supported the Decision Trees and Convolution Neural Network classifiers. The important issues in speech emotion recognition system are the signal processing unit during which appropriate features are extracted from available speech signal and another may be a classifier which recognizes emotions from the speech signal. The mean accuracy of most of the classifiers for SER for speaker independent system is less than that of the speaker dependent.

Automatic emotion recognitions from the human speech are increasing now each day because it leads to the higher interactions between human and machine. To improve the emotion recognition process, combinations of the given methods are often derived. Also by extracting simpler features of speech, accuracy of the speech emotion recognition system are often enhanced.

REFERENCES

1. M. E. Ayadi, M. S. Kamel, F. Karray, “Survey on Speech Emotion Recognition : Features, Classification Schemes, and Databases”, Pattern Recognition 44, PP.572-587, 2011.
2. Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.
4. C.-W. Huang and S. S. Narayanan, “Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition,” pp. 1–19, 2017. 14.
5. H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” Neural Networks, vol. 92, pp. 60–68, 2017. 15.
6. A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network,” 2017 Int. Conf. Platf. Technol. Serv., pp. 1–5, 2017.
7. C. Szegedy, V. Vanhoucke, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” 2014.
8. F. Noroozi, N. Akrami, and G. Anbarjafari, “Speech-based emotion recognition and next reaction prediction,” 2017 25th Signal Process. Commun. Appl. Conf. SIU 2017, no. 1, 2017.
9. A. Graves, A. Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.
10. Sabur Ajibola Alim and Nahrul Khair Alang Rashid (December 12th 2018). Some Commonly Used Speech

Feature Extraction Algorithms, From Natural to Artificial Intelligence - Algorithms and Applications, Ricardo Lopez-Ruiz, IntechOpen, DOI: 10.5772/intechopen.80419

