REAL TIME SPEECH EMOTION RECOGNITION

KrutikaThakare
BE Computer Engineering
Datta Meghe College of Engineering
Navi Mumbai,India

Sakshi Wadhe
BE Computer Engineering
Datta Meghe College of Engineering
Navi Mumbai,India

ShivakumarDasari
BE Computer Engineering
Datta Meghe College of Engineering
Navi Mumbai,India

Dr.ChhayaPawar
Assistant Professor
Department of Computer Engineering
Datta Meghe College of Engineering
Navi Mumbai,India

Abstract *: speech signal is one of the most natural and fastest methods of communication between humans. Many systems have been developed by various researchers to identify the emotions from the speech signal. In differentiating between various emotions particularly speech features are more useful and if not clear is the reason that makes emotion recognition from speaker's speech very difficult. There are a number of the dataset available for speech emotions, it's modeling, and types that helps in knowing the type of speech. After feature extraction, another important part is the classification of speech emotions so the paper has compared and reviewed the different classifiers that are used to differentiate emotions such as sadness, neutral, happiness, surprise, anger, etc. The research also shows the improvement in emotion recognition system by making automatic emotion recognition system adding a deep neural network. The analysis has also been performed using different ML techniques for Speech emotions recognition accuracy in different languages.*

1.Introduction

Emotion is the most important component of being human, and very essential for everyday activities, such as the interaction between people, decision making, and learning. It eases the communication between people and makes it representative. It is important to detect and recognize the emotion in computer systems which people interact with, to enhance the communication between users and machines. Moreover, we need to know the current state of the user to enhance the accuracy and throughput of the system. In order to make the computer understand and recognize emotion, One need to understand the sources of them in our body. Emotion could be expressed verbally like some known words or non-verbally like the tone of voice, facial expression and physiological changes in our nervous system. Voice and facial expression are not reliable indicators of emotion because they can either be fake by the user or may not be produced as a result of a specific emotion.

Speech Emotion Recognition (SER) is the task of recognizing emotional aspects of speech irrespective of the actual semantic contents. While humans even at early ages can easily perform this task as a natural part of speech communication, the ability to do it automatically using computer software is still an ongoing subject of research. Adding emotions to machines has been recognized as a key factor in making machines appear and act more like humans . If the next generation of human-machine communication systems is expected to have emotional capabilities, machines need to be able to understand not only what people say, but also what kind of emotions they convey. Only through this capability, can a fully meaningful and functional conversation based on mutual human-machine trust and understanding be achieved. Robots capable of understanding emotions could provide appropriate emotional responses and exhibit emotional personalities. In some circumstances, real people such as actors, teachers or social commentators could be replaced by computer-generated characters having the ability to conduct very natural and convincing conversations by appealing to human emotions.

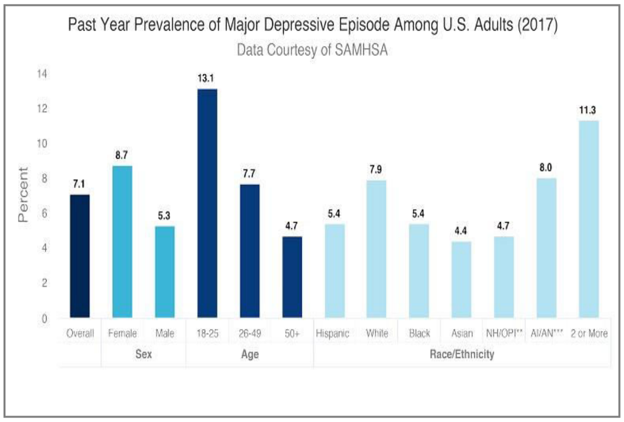
2. Overview

Emotion recognition in spoken dialogues has been gaining increasing interest all through current years. Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction (HCI). It has a potentially wide, such as the interface with robots, banking, call centers, car board systems, computer games etc. For class-room orchestration or E-learning, information about the emotional state of students can provide focus on enhancement of teaching quality. For example, teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why learners emotional state should be considered in the classroom. In general, the SER is a computational task consisting of two major parts: feature extraction and emotion applications, such as the interface with robots, banking, call centers, car board systems, computer games etc. For classroom orchestration or E-learning, information about the emotional state of students can provide focus on enhancement of teaching quality. For example, teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why learner's emotional state should be considered in the classroom.

3. Need Of Speech Emotion Recognition Emotion recognition in spoken dialogues has been gaining increasing interest all through current years. Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction (HCI). It has a potentially wide applications machine classification.

Major depression is one of the most common mental illnesses in the country. An estimated 17.3 million adults in the US reported having at least one major depressive episode over the course of a year, a 2017 report by the Substance Abuse and Mental Health Services Administration (SAMHSA) shows as follows.

That's 7.1% of all adults ages 18 and older. Women have a higher prevalence of experiencing a major depressive episode than men (8.7% compared to 5.3% for adult males). SER can be a great help for treatment of depression.



4. Previous works

Speech signals can be modeled using acoustic patterns that vary in frequency, time and intensity. One of the most common ways of observing these variations simultaneously is via the speech spectrogram, which represents speech energy as a function of both time and frequency. The spectrogram is one of the most simple and effective ways of visualizing the time frequency evolution of spectral, prosodic and energy features characterizing speech acoustics.

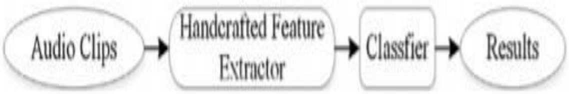
H. Cao, R. Verma, and A. Nenkova Proposed a ranking SVM method for synthesize information about emotion recognition to solve the problem of binary classification. Un weight average (UA) or Balance accuracy achieved 44.4%.[1] **Chen, X. Mao, Y. Xue, and L. L. Cheng** Aimed to improve speech emotion recognition in speakerindependent with three level speech emotion recognition method. This method classify different emotions from coarse to fine then select appropriate feature by using Fisher rate. The recognition rates for three level are 86.5%, 68.5% and 50.2%.[2] In conclusion, there are only limited studies that considered applying multiple classifier to speech emotion recognition.

5. Speech Emotion Recognition System

There is a pattern recognition system stage in speech emotion recognition system that makes them both same. Energy, MFCC, Pitch like derived speech features patterns are mapped using various classifiers.

It consists of five main modules are:

Speech input: Input to the system is speech taken with the help of microphone audio. Then equivalent digital representation of received audio is produced



through pc sound card.

Feature extraction and selection: There are 300 emotional states of emotion and emotion relevance is used to select the extracted speech features. For speech feature extraction to selection corresponding to emotions all procedure revolves around the speech signal.

Classification: Finding a set of significant emotions for classification is the main concern in speech emotion recognition system. There are 300 emotional states contains in a typical set of emotions that makes classification a complicated task. **Recognized emotional output**: Fear, surprise, anger, joy, disgust and sadness are primary emotions and naturalness of database level is the basis for speech emotion recognition system evaluation.

6. Dataset for Speech Emotions

In the field of affect detection, a very important role is played by suitable choice of speech database. The RAVDESS database is used for good emotion recognition system as given below.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files. The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

4.Types Of Speech

On the basis of ability they have to recognize a speech recognition systems can be separated in different classes. Following are the classification:

- Isolated words: In this type of recognizers sample window both sides contains low pitch utterance. At a time only single word or utterance is accepted by it and there is need to wait between utterances by speaker as these systems have listen/non-listen states. For this class isolated utterance is a better name.
- Connected words: In this separate utterance can run together with minimal pause between them otherwise it is similar to isolated words.
- Continuous words: It allows users to speak naturally and content are determined by computer. Creation of recognizers that have continuous speech capabilities are difficult due to determination of utterance boundaries by utilizing a special method.
- Spontaneous words: It can be thought of as speech at basic level that is natural sounding and not rehearsed. Variety of natural speech features are handle is the ability of spontaneous speech with ASR system.

7. Speech Features

Relevant emotional features extraction from speech is the second important step in emotions recognition. To classify features there is no unique way but preferably acoustic and linguistic features taxonomy is considered separately. Due to extreme difference concerning these extraction methods and database used is another distinction. An importance is gain by linguistic features in case of spontaneous or real life on other hand their features lose their value in vase of acted speech. Earlier only small set of features were used but now larger number of functional and acoustic features are in use for extraction of very large feature vectors. In this section explanations of acoustic, linguistic features and functional are discussed.

Acoustic features: Large statics measures of energy, duration and pitch is used to characterized acoustic features that are derived from speech processing. In order to mask particular items in speech of humans ainvoluntary and voluntary acoustic variation is basic used for emotion recognition using acoustic features. Measurement of energy, pitch or voiced and unvoiced segments is in seconds that can represent duration features by applying different types of normalization. Words, utterance, syllables or pauses like phonemes unit’s higher phonological parameter duration is exclusively represented.

Linguistic features: In reaction of our emotional state an important role is played to grammatical alternations or words chosen by us. Bag-of-Words and N-Grams are two prime methods from number of existing techniques used for analysis. To predict next given sequence a probabilistic base language model is used and N-grams is a numerical representation form of texts in automatic document categorization. Reduction of speech complexity by elimination of irrelevant words and stopping word that do not increase a general minimum frequency of occurrence is useful before applying this technique. Cries, laughs, sighs, etc non-linguistic vocalizations can be integrated into vocabulary.

Feature selection: To describe phenomenon from a larger set of redundant or irrelevant features is a subset of features selected by feature selection. Feature selection is done to improve the accuracy and performance of classifier. Wrapper based selection methods are generally used approaches that employ an accuracy of target classifier as optimization criterion in a closed loop fashion. In this features with poor performance are neglected. Hill climbing, sequential forward search is commonly chosen procedure with a sequentially adding and empty set. These features give performances improvement. Selected subset of features effects are ignored by use of filter methods which is a second general approach. Reduced features sets obtained from the acted and non-acted emotions difference is very less.

8. Feature Extraction

There are number of methods for feature extraction like Linear predictive cepstral coefficients (LPCC), Power spectral analysis (FFT), First order derivative (DELTA), Linear predictive analysis (LPC), Mel scale cepstral analysis (MEL), perceptual linear predictive coefficients (PLP) and Relative spectra filtering of log domain coefficients(RASTA).Here MFCC method is used for feature Extraction. **Mel frequency cepstral coefficients (MFCC):** It is considered as one of the standard method for feature extraction and in ASR most common is the use of 20 MFCC coefficients. Although for coding speech use of 10-12 coefficients are sufficient and it depend onthe spectral form due to which it is more sensitive to noise. This problem can be overcome by using more information in speech signals periodicity although a periodic content is also present in speech. Real cesptal of windowed short time fast Fourier transform (FFT) signal is represent by MFCC [5]. Non linear frequency is use. The parameters similar to humans used for hearing speech are used to extracts parameters using audio feature extraction MFCC technique. Other information is deemphasizes and arbitrary number of samples contain time frames are used to divide speech signals. Overlapping from frame to frame is used to smooth the transition in most systems and then hamming window is used to eliminate the discontinuities from each time frame.

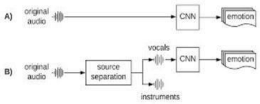
9. Classification

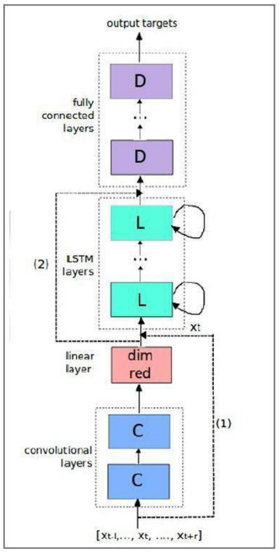
The best features come after features calculation is provided to the classifier. In expression of speaker's speech an emotion is recognizes by classifier and for speech emotion recognition number of classifiers have been proposed by various researchers. Here SVM classifier is used for classification purpose. **Support Vector Machine (SVM) classifier:** Human computer Interface (HCI) subset automatic emotion and speech recognition has become widely researched topic with the advent of digitization of every possible avenue. As we understand machines, machine has also understood us as menial jobs are taken with machines. From given sample amplitude, pitch and MFCC features are extracted and it run across growing and existing database of training samples. Ashwini Rajasekhar, et.al, (2018), have distinguished the given sample using SVM and speaker utterance is detected using MFCC [4]. In the end SVM classifier differentiates between fear, anger, sadness, happiness and updates the database accordingly. Amiya Kumar, et.al, (2015), have introduced a novel approach by combining MFCC, LPCC derived features, energy, ZCR, pitch prosody features, MEDC dynamic features for automatic recognition of speaker's emotion state [3]. Then happy, surprise, anger, sad, disgust, neutral and fear are seven discrete emotional states identified using multilevel SVM classifier in five native assamese languages. The proposed approach is evaluated for combination of features in terms of accuracy thatshows a good result for speaker independent cases as compared to individual features.

10. Jupyter Lab is an open-source, web-based UI for Project Jupyter and it has all basic functionalities of the Jupyter Notebook, like notebooks, terminals, text editors, file browsers, rich outputs, and more. However, it also provides improved support for third party extensions.Librosa is a Python library for analyzing audio and music. It has a flatter. package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code.

Along with this a webpage is created for recoring sounds and then recoginzing emotion for that sound.

11. CNN-LSTM Model foe SER Experiments have been performed on speech transcriptions and speech features independently as well as together in an attempt to achieve accuracies greater than existing state-of-the-art methods. Different combinations of inputs have been used in different DNN architectures. Recent research into speech processing has shown the successful application of deep learning methods and concepts such as CNNs and Long Short-Term Memory (LSTM) cells on speech features. CNNs have been shown, by extensive research, to be very useful in extracting information from raw signals in various applications such as speech recognition, image recognition.





CNN-LSTM Model

12. MLP Classifier

This is a Multi-layer Perceptron Classifier; it optimizes the log-loss function using LBFGS or stochastic gradient descent. Unlike SVM or Naive Bayes, the MLP Classifier has an internal neural network for the purpose of classification. This is a feed forward ANN model.

13. Results

In this section, we described the experiment environment and report the recognition accuracy of using MLP classifier on emotional speech database. We used RAVDES database for network training and validation. To evaluate the classification error 10cross validation test were used. We used 70% of data for training and 30 % for testing. The To further improve the efficiency of system combination of more effective features can be used that enhances the accuracy of speech emotion recognition system. Future works will investigate ways of implementing the spectrogram classification approach to SER on mobile phones, call centers, and

online communication facilities.

13. References

1. H. Cao, R. Verma, and A. Nenkova,

neural network structure used is a simple LSTM. It consists of two consecutive LSTM layers with hyperbolic tangent activations followed by two classification dense layers. These results in general have suggested that deep learning methodologies are appropriate for modeling affective states and, more importantly, indicated that feature extraction may not be necessary for as deep learning models are able to identify high level of data abstraction automatically.

This study is commenced by implementation of CNN-LSTM system. Subsequently, hyper-parameters were modified by changing the convolution kernel size etc. The model is constructed in Python and trained for 100 epochs. All the experiments presented here are run for data files of each individual participant and then the average of the resulting models prediction accuracy and other performance metrics are reported.

12. Conclusion and Future Scope

In the paper brief introduction about speech emotion recognition is given along with the speech emotion recognition system description. In the field of Speech recognition, a very important role is played by a suitable choice of speech database. For good emotion recognition system mainly RAVDES database is used. On the basis of ability, they have to recognize a speech recognition system can be separated in different classes are isolated, connected, spontaneous and continuous words. Relevant emotional features extraction from the speech is the second important step in emotions recognition. To classify features SVM is used here.

Through this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voicebased virtual assistants or chat bots, in linguistic research, etc

“Speaker- sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Compute. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2019.

2. 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. 2018.
3. Y. Kumar, N. Singh, —A First Step towardsan Automatic Spontaneous Speech Recognition System for Punjabi Language‖, International Journal of Statistics and Reliability Engineering, pp. 81-93, 2015.
4. A. Rajasekhar, M. K. Hota, —A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines.
5. A. Nogueiras, A. Moreno, A. Bonafonte, J. B. Marino, —Speech Emotion Recognition Using Hidden Markov Model‖, Eurospeech, 2001.
6. https://www.psycom.net/depression.central. html
7. https://www.researchgate.net/publication/322873355_Speech_Emotion_Recognition_Methods_and_Cases_Study