Human Emotion Recognition from speech in audio through physical features

Akshay Chatterjee[1], Ghazaala Yasmin[2]

[1] Department of Computer Science and Engineering, St Thomas' College of Engineering and Technology, Kolkata 700 023, India
akshay.chatterjee2015vit@gmail.com
[2] Department of Computer Science and Engineering, St Thomas' College of Engineering and Technology, Kolkata 700 023, India
me.ghazaalayasmin@gmail.com

Abstract. Emotion is the part of human life which highly influence decision making compare to other things. The human emotion is not only conveyed by the body language but also through speech. Speech delivers the maximum information regarding the feelings of human. Nowadays still there are many scenarios that these emotions are not being recognized properly. This is why this area has taken the attention of the researcher to explore. This fact has motivated us to propose a methodology to discriminate human emotions precisely. The proposed methodology has selected a set of features which carry more information about emotional feelings. The audio files database has been preprocessed .Physical features means both time domain and frequency domain features have been extracted from the preprocessed data. The files have been classified based on the feature. Some well-known classifier has been adopted for classification to achieve better outcome.

Keywords: Emotion Recognition. Feature extraction, classification, Speech

1 Introduction

Emotion recognition through speech has taken the center of interest among the researchers because of its ambiguous nature. As a result it is difficult to achieve efficient accuracy while implementing the techniques for the emotion recognition system. Emotion is the most natural activity of human which can be expressive gesture. Speech is most expressive way to carry out the human emotions. This fact has motivated researchers to choose speech for exploring the domain of emotion recognition. Emotion detection is required in many applications. It is an important factor of the human computer interaction field. Our proposed work is to detect the emotion from an audio speech. An efficient human emotion recognition system will lead to more natural and friendly interaction between the human and digital computer.

1.1 Related Work

Numerous approaches on emotion recognition has been proposed to achieve better result. Nanavare and jagtap[1] have been suggested recognition of human emotions from speech processing through hidden markov model by means of dynamic programming. Wanare Shankar and Dandare[2] have proposed human emotion recognition from speech through neural network and support vector machine. Karpagavalli and Chandra [3] has given a review of different speech recognition techniques. Zeng, Zhihong, et al.[4] have summerze a survey of affect recognition methods in Audio, visual, and spontaneous expressions. Ververidis et.al[5] have proposed emotional speech recognition through different audio feature. Neiberg et.al [6] has propounded emotion recognition in spontaneous speech using Gaussian mixture model. Song et.al [7] have suggested Audio-Visual based Emotion Recognitionn through hidden markov model. Ramakrishnan et.al [8]has highlighted algorithm and application of Enhancement, Modeling and Recognition for speech. Vogt et.al [9]have given a review on automatic recognition of emotions from speech. Pokorny et.al [10] proposed an efficient technique for detection of negative emotions in speech signals using bags-of-audio-words. Zhongzhe et.al [11] have analysed recognition of emotions in audio signals through different techniques. . Pawar et.al [12] has suggested recognition and Classification of human emotion from audio. Kamińska et.al [13] has proposed a technique for recognition of human emotion from a speech signal based on Plutchik's model.

.

2 Proposed Methodology

Emotion can be categorized into different groups based on the various features of speech. They are discriminative in nature. While observing the various characterictics of speech it has been observed that each kind of emotion has different features. The proposed work has aim to extract those features from the input speech . Then we have used some dimensionality reduction principle to reduce the size of features. We have classified the speech according to the features. From the past works it has been found that spectral flux, MFCC are one of the essential features for emotion classification. We have used four types of classifiers in our proposed work. And we have also depicted a clear comparison between them. Fig.1 gives a clear view of the proposed work.

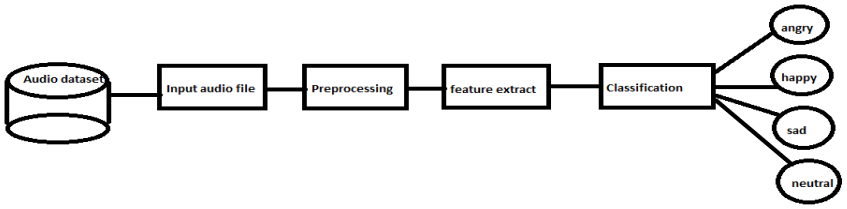

Fig. 1. Schematic block diagram for Emotion Recognition System

2.1 Feature Extraction

There are two kinds of features in audio. They are clip level and frame level features. The features are again divided into physical features and perceptual features. Physical features are of two types. Perceptual features contains Pitch, tempo, etc. Time domain features include Zero Crossing rate(ZCR), Energy, etc. Frequency domain features include Spectral flux, etc. In our proposed work, we have used ZCR, Energy, spectral flux, and MFCC.

2.1.1 Zero Crossing Rate (ZCR)

It is a time domain feature. The audio signal changes the sign from positive to negative and back. The rate of change of sign along a signal is called zero-crossing rate. It's always a key feature to classify sounds. ZCR is defined in eqn.1

$$z = \sum_{p=1}^{n-1} sign[x(p-1) * x(p)] \tag{1}$$

Here, n indicates the number of samples present in the r^{th} frame and

$$sign[x] = \begin{cases} 1, \text{ if } x > 0 \\ 0, \text{ otherwise} \end{cases} \tag{1.1}$$

It has been found that during different emotions the sign change is an important factor. So we have considered the rate of sign change feature for our classification of emotions. Mean and standard deviation has been calculated from ZCR.

2.1.2 Energy

Short-Time Energy (STE) is a simple but effective time-domain feature, which is widely it is calculated for finding the variation in the emotion of a person. The energy of the pth frame is defined in equ.2.

$$Ep = \sum_{n=1}^{r} [xp(n)]^2 \tag{2}$$

2.1.3 Mel-Frequency Cepstral Coefficients (MFCC)

It is Mel Frequency Cepstral Coefficient. In audio frequency bands are not linear and arranged by their mel-scale. In other words it is a scale of melody and calculates the difference in level. The first thirteen coefficients of MFCC are highly discriminative in nature in terms of classification. So we have used MFCC as a feature. In Fig. 2, x-axis denotes the audio dataset number and Y-axis denotes the MFCC value of the corresponding audio file. Fig.2 shows the 13 coefficient of MFCC in the plot.

$f_n = 2595 * \log_{10}(1+f)$; f is the frequency of the given speech signal.

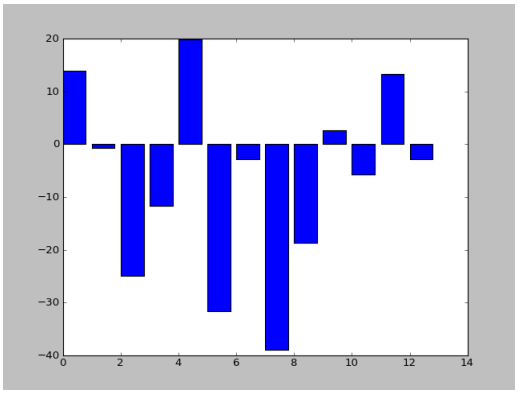

Fig. 2. MFCC plot for speech

2.1.4 Spectral Flux

It is also a frame level feature and a frequency domain feature. It is the squared difference between the normalized magnitudes of the spectra of the two successive frames. The mean and standard deviation has been calculated from the spectral flux with respect all frames. It has been noticed that the spectral flux changes precisely as the human emotion changes in their speech. Fige.3 shows the plot for spectral flux. The mathematical expression for spectral is given in eq.3 .N is the total number of frames. The square of the difference of sample value of frame t-1 and t for sample has been computed. Henceforth mean and standard deviation has been calculated.

$$\sum_{x=1}^{n} \{Nt(x) - Nt-1(x)\}^2 \tag{3}$$

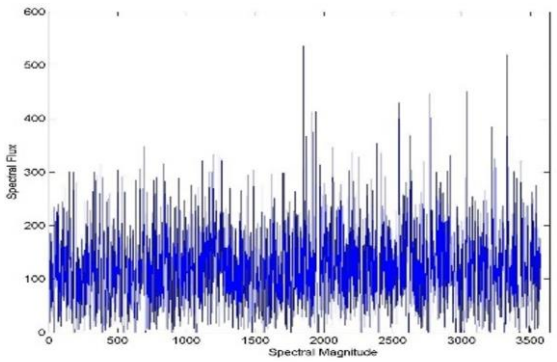

Fig. 3. Spectral Flux of sample speech file

2.2 Classification

Classification is defined as the process of grouping the given observation into a set of categories. This term involves two types of learning process; one is supervised learning where the set of class for identification is mentioned. Another is unsupervised learning also termed as clustering which categorize the data based on some inherent similarity. The classification algorithm is known as classifier. For classification weka tool [14] has been used to check the accuracy of the proposed work. Classifiers like naïve bayes, multilayer perceptron, have been taken into account. Furthermore, Random forest has been chosen as it has been noticed that this classifier gives the better result among all decision tree algorithm for the proposed system. For each classifier ten folds cross validation has been done.

2.2.1 Random-forest Classifier: It is a supervised learning technique. It creates a subset of decision trees from a set of training variables. Then it takes feedback from the trees and creates a voting system and finds the maximum efficient set. From the past outcome, it has been found that Random-forest is one of the powerful classifier.

2.2.2 Support Vector Machine: This is a supervised model that helps in classification and regression. Given a set of different data as an input it will just find out what are the various types are there and classify them in non probabilistic and binary form. It Is a path from the artificial neural network which was unable to give the same accuracy to the functions what svm performs. It follows obeys algorithms like support vector classification, Support vector clustering,etc. Classification algorithm is solved using some kernel functions like: Linear ,Polynomial and Radial Basis Function. It helps in classifying

2.2.3 Multilayer Perceptron: This also comes under supervised learning technique. It is a feedforward neural network which contains atleast three layers of nodes. Every layer of node except the input one uses non-linear activation function. It can classify non linear datasets. That's why it is appropriate for classification in case of our dataset.

2.2.4 Naïve bayes Classifier: This is scalable and class probabilistic classifiers and is based on Bayes theorem. It requires parameters linear in the number of variables. It is a supervised model that trains the features and classifies by labeling each class. This consist of a group of algorithms like GaussianNB, MultinomialNB,etc. Gaussian Naïve Bayes naïve bays has been used for the proposed methodology.

3. Experimental Result

The verification for the suggested method has been tested for wide range of heterogeneity in speech dataset. These data sets have been made up with 400 files having duration of each file is 120 seconds. These datasets have been collected from recording of different speakers, some files have been downloaded from the Internet. To determine the feature more precisely, each speech signal has been broken into multiple frames consisting 8 seconds each. The sound has been preprocessed through noise removal to achieve better accuracy. In Table.1 the accuracy of proposed work has been shown which has been classified by different types of classifiers. Here the accuracy has been summarized based on the classifier which has given better accuracy compare to other classifier. Total 18 features (13 MFCC+ 2 ZCR + 2 Spectral flux + 1 Energy) have been subjected for classification. Since the feature set small, feature selection algorithm has not applied on the extracted features.

Table 1. Classification Accuracy (in %) For proposed work

Classific Scheme	**Happy**	**Angry**	**Sad**	**Neutral**
Random-Forest	91.5	92.4	93.7	91.9
Support vector Machine	81	80.6	80.4	80.3
Multlaye perceptron	96	95.3	92.6	94.9
Naïve Bayes'	90.4	90.3	91.6	90.7

3.1 Comparative Analysis

The proposed methodology has been compared with the work of Nanavare, and Jagtap [1] on the sample data collected for the proposed work. The same has also been tested on the technique proposed by Song, Mingli, et al. [7] and Kamińska et.al [13]. Nanavare et.al has proposed the methodology based on hidden markov model. The same technique has been used by Song, Mingli, et al. [7]. Kamińska et.al [13] has adopted audio features from speech. Table.2 shows the result achieved by the proposed technique is coming up with better result compare to the other techniques..

Table. 2. Comparative study (in %) for the substantial work.

Precedent Approach	**Happy**	**Angry**	**Sad**	**Neutral**
Nanavare, and Jagtap [1]	88.5	86.7	85.1	86
Song, Mingli, et al. [7]	87	84.2	83	82.8
Kamińska et.al [13]	90.02	89.4	88.3	87.3
Proposed work	96	95.3	92.6	92.5

4. Conclusion

The proposed system has developed to discriminate different type of human emotion more precisely. The novelty of the system is achieving better result by extracting small set of feature. The proposed method can be explored more with new features in the future. New feature selection algorithm will be implemented. The proposed system will be tested with different indiscriminative emotions. Furthermore the system will be tested with some bench mark data. The method classifies the emotion into four groups it can be extended for more classes of emotion which can be used for worldwide human computer interaction fields.

References

1. Nanavare, V. V., and S. K. Jagtap. "Recognition of Human Emotions from Speech Processing." *Procedia Computer Science* 49 (2015): 24-32.
2. Wanare, Miss Aparna P., and Shankar N. Dandare. "Human Emotion Recognition From Speech." system 6 (2014).
3. Karpagavalli, S., and E. Chandra. "A Review on Automatic Speech Recognition Architecture and Approaches." *International Journal of Signal Processing, Image Processing and Pattern Recognition* 9.4 (2016): 393-404.
4. Zeng, Zhihong, et al. "A survey of affect recognition methods: Audio, visual, and spontaneous expressions." IEEE transactions on pattern analysis and machine intelligence 31.1 (2009): 39-58.
5. Ververidis, Dimitrios, and Constantine Kotropoulos. "Emotional speech recognition: Resources, features, and methods." *Speech communication* 48.9 (2006): 1162-1181.
6. Neiberg, Daniel, Kjell Elenius, and Kornel Laskowski. "Emotion recognition in spontaneous speech using GMMs." *Ninth International Conference on Spoken Language Processing*. 2006.
7. Song, Mingli, et al. "Audio-visual based emotion recognition-a new approach." *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. Vol. 2. IEEE, 2004.
8. Ramakrishnan, Srinivasan. "Recognition of emotion from speech: a review." *Speech Enhancement, Modeling and Recognition-Algorithms and Applications*. InTech, 2012.
9. Vogt, Thurid, Elisabeth André, and Johannes Wagner. "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation." *Affect and emotion in human-computer interaction*. Springer, Berlin, Heidelberg, 2008. 75-91.
10. Pokorny, Florian B., et al. "Detection of negative emotions in speech signals using bags-of-audio-words." *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015.
11. Zhongzhe, X. I. A. O. *Recognition of emotions in audio signals*. Diss. Thesis Google Scholar, 2008.
12. Pawar, Anuja, and M. E. Student. "Recognition and Classification of Human Emotion from Audio." *International Journal of Engineering Science* 13790 (2017).
13. Kamińska, Dorota, and Adam Pelikant. "Recognition of human emotion from a speech signal based on Plutchik's model." *International Journal of Electronics and Telecommunications*58.2 (2012): 165-170.
14. Srivastava, Shweta. "Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining." *International Journal of Computer Applications* 88.10 (2014).