Speech Emotion Recognition

S.Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh
Dept. of ECE, Amrita School of Engineering, Amrita Vishwa Vidyapeetam, Bangalore, India
Email: s_lalitha@blr.amrita.edu, abhishek_madhavan@yahoo.in, 36bharath@gmail.com, k.s.saketh@gmail.com

***Abstract*—In the past decade a lot of research has gone into Automatic Speech Emotion Recognition(SER). The primary objective of SER is to improve man-machine interface. It can also be used to monitor the psycho physiological state of a person in lie detectors. In recent time, speech emotion recognition also find its applications in medicine and forensics. In this paper 7 emotions are recognized using pitch and prosody features. Majority of the speech features used in this work are in time domain. Support Vector Machine (SVM) classifier has been used for classifying the emotions. Berlin emotional database is chosen for the task. A good recognition rate of 81% was obtained. The paper that was considered as the reference for our work recognized 4 emotions and obtained a recognition rate of 94.2%. The reference paper also used hybrid classifier thus increasing complexity but can only recognize 4 emotions.**

***Index Terms*—Emotion recognition, Pitch, Prosody, SVM, Berlin database**

I. Introduction

Human emotions are very difficult to comprehend from a quantitative perspective. Facial expressions are one of the best ways of guessing the emotional state of a person. Speech is another modality that can be used. Speech is a complex signal which contains information about the message, speaker, language and emotions. There are various kinds of emotions which can be articulated using speech. Emotional speech recognition is a system which basically identifies the emotional state of human being from his or her voice; speech is very misleading even for humans to judge the emotion of the speaker[1].

A major motivation comes from the desire to improve the naturalness and efficiency of human-machine interaction.The reference paper[4] that was chosen has been able to successfully recognize only 4 emotions. The work presented here has classified 7 emotions with a overall good recognition rate. In general, the systems for speech analysis uses various techniques for the extraction of characteristics from the raw signal. Concerning emotions, the relevant information is in the Pitch, Prosody and in the Voice quality. The next step in this strategy is to discover the features which discriminate the speech data (to the training labels) and to discard the non-discriminative features. This is achieved by calculating the cross validation between parameters after which grid of parameters is created; the one with highest cross validation is selected. 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 hyperplane 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.

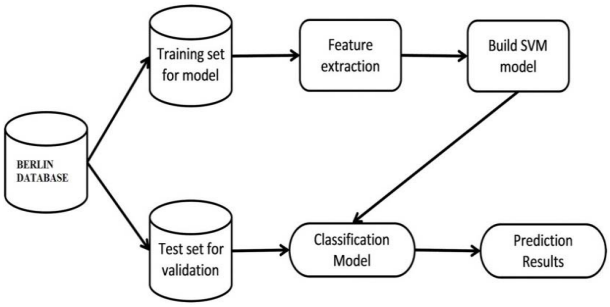

Fig. 1. Block Diagram of the proposed Speech Recognition System

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. 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.

II. Emotion Corpus

The BerlinEmo 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*.

978-1-4799-5496-4/14/$31.00 ©2014 IEEE

TABLE I
THE BERLIN DATABASE

EMOTIONS	NUMBER OF SAMPLES
Anger	127
Boredom	81
Disgust	46
Fear	69
Happiness	71
Sadness	62
Neutral	79

This is the first module and very important because speech features will determine the accuracy. Feature extraction was performed using MATLAB©. The features considered are [4]:

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 $T_\circ$' which gives 'Fundamental Frequency $F_\circ$'.

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.

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

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.

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

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

F. *HNR*

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

$$HNR = 10\log_{10}\{\frac{\sum_{i}^{NFFT/2}|S_i|}{\sum_{i}^{NFFT/2}|N_i|}\} \tag{4}$$

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.

$$Z_{\widehat{n}} = \sum_{m=-\infty}^{\infty} 0.5\left|\text{sgn}\{x[m]\} - \text{sgn}\{x[m-1]\}\right| \times w[\widehat{n} - m] \tag{5}$$

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 n^{th} percentile of the power spectral distribution, where n is usually 85% or 95%.

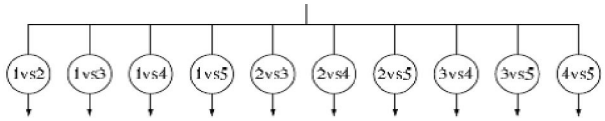

Fig. 2. One-vs-one max wins SVM voting scheme

III. Classifier

Support Vector Machine is a statistical classifier which classifies data into binary classes based on training data. Support vector machine constructs a hyperplane or set of hyperplanes in a high-dimensional space or infinite dimensional space, which can be used for classification, regression or other tasks. Speech emotion recognition is a multi-category classification problem. To classify using SVM we convert multi-classification into binary classification problem.

There are several methods to do this but we have opted for one-versus-one approach. The hyper planes of binary SVM are built from any two of all categories, so the number of binary SVM classifiers is k ×(k-1)/2. Here 'max-wins' voting method is used for One-Versus-One voting strategy, the k×(k-1)/2 binary SVM classifiers are trained in parallel.

For example, category *i* and category *j* trains with classifier C_{ij}, C_{ij} decides whether sample *x* belong to category *i* or category *j*. Therefore the number of *i* category votes increases by one, otherwise *j's* number of votes is incremented. When the process is over, the category with the most voters is the right category that the sample belongs to. The structure of One-Versus-One binary SVM is shown in Fig.2 Here 1-5 are set to represent 5 emotion categories of 2 speech corpuses, so 10 classifiers are trained. From the process of categorization, it is found that this method is less effective as the number of the classes increases, which will cause the decision to be slower.

LibSVM was used to perform the functions mentioned above.

IV. Results and Analysis

The data set was sub divided into a training set and testing set:

Training Set: 90.093%

Testing Set: 9.906%

Once the features of each speech signal were extracted, their statistical parameters were calculated. These were then compiled into the feature vector providing us with a training set and a testing set. Each statistical parameter was tested separately and the best results were found using Mean.

These labels were then appended to the training and testing matrices. The SVM used was the LibSVM. The SVM was trained using the Training data-set. Once training was complete a model was created and this was used for the testing phase.

A very good recognition rate was obtained for fear, anger and neutral emotions. For few emotions(happiness, boredom and sadness) had moderate recognition rate while disgust had a poor recognition accuracy. It can be postulated that with a bigger data set and a modified SVM better accuracy for these emotions can be obtained.

The overall emotion recognition efficiency is 81.132% which can be observed from the results of Table IV. The proposed work is definitely advantageous compared to the reference paper[4] which uses a feature set of 14, recognizing only 4 emotions although giving a better recognition rate.

TABLE II
The Labelling

Emotions	Label
Fear	1
Happiness	2
Boredom	3
Anger	4
Sadness	5
Neutral	6
Disgust	7

TABLE III
The Confusion Matrix

Emotion	Fear	Happiness	Boredom	Anger	Sadness	Neutral	Disgust
Fear	7	0	0	0	0	0	0
Happiness	0	5	0	2	0	0	0
Boredom	0	0	6	0	0	2	0
Anger	0	0	0	12	0	0	0
Sadness	0	0	2	0	4	0	0
Neutral	0	0	1	0	0	7	0
Disgust	1	1	1	0	0	0	2

TABLE IV
Accuracy

Emotions	Accuracy
Fear	100%
Happiness	71.43%
Boredom	75.00%
Anger	100.00%
Sadness	66.67%
Neutral	87.50%
Disgust	40%
Total	**81.132%**

V. Conclusion and Future Work

One approach to Speech emotion recognition has been presented in this paper and an accuracy of 81% was obtained using a simple SVM classifier. Disgust had a lower recognition rate as it is slightly complex in nature and difficult to be detected even by a human. The proposed work uses a compact feature set with a overall good recognition accuracy for seven emotions compared to the reference paper. The future scope

of this work would be in improving the rate of recognition by the use of hybrid classifiers[4]. Although the reference paper provides a better accuracy, the work presented in this paper has been able to incorporate more emotions and still provide a fairly high rate of recognition. Further as this work does not incorporate cepstral domain features or transforms for feature extraction, the complexity and thereby the run time is reduced significantly.

ACKNOWLEDGEMENT

Our work has been supported by Amrita School of Engineering, Bangalore.
We would like to thank Dr.Shikha Tripathi, Head of the Department, Electronics & Communication, Amrita School of Engineering for her support and guidance.
We would like to take this opportunity to wholeheartedly thank our teachers from the Department of Electronics & Communication, Amrita School of Engineering for their constant guidance.

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.