Speech emotion recognition using RBF kernel of LIBSVM

Article · June 2015

DOI: 10.1109/ECS.2015.7124760

Bhushan Yelure
Walchand College of Engineering, Sangli

Kishor Tayade
Government College of Engineering

SPEECH EMOTION RECOGNITION USING RBF KERNEL OF LIBSVM

Mr.Y.D. Chavhan
GCE, Karad
Karad, India
yashpal.chavhan@gmail.com

Mr.B.S.Yelure
GCE, Karad
Karad, India
bhushanyelure2008@gmail.com

Mr. K.N.Tayade
GCE, Karad
Karad, India
kntayade@rediffmail.com

***Abstract*— Automatic Speech Emotion Recognition (SER) is a current research topic in the field of Human Computer Interaction (HCI) with wide range of applications. The speech features such as, Mel Frequency cepstrum coefficients (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC) are extracted from speech utterance. The LIBSVM is used as classifier to identify different emotional states such as anger, happiness, sadness, neutral, fear, from Berlin emotional database. The results are taken by using RBF kernel of LIBSVM. It gives 93.75% recognition accuracy for RBF kernel.**

***Keywords*— Speech emotion, Emotion Recognition, LIBSVM, MFCC and MEDC, RBF**

I. INTRODUCTION

Emotion is a term for a mental and physiological state associated with a wide variety of feelings, thoughts, and behaviour. Emotions are subjective experiences, or experienced from an individual point of view. Emotion is often associated with mood, temperament and personality. But in general emotions are short-term whereas moods are long-term and temperaments or personalities are very long-term.

Human emotion can be of different types such as angry, happiness, sadness, neutral, fear, disgust, surprise, shy, bored etc.

Automatic Speech Emotion Recognition is a very recent research topic in the Human Computer Interaction (HCI) field. As computers have become an integral part of our lives, the need has risen for a more natural communication interface between humans and computers. To achieve this goal, a computer would have to be able to perceive its present situation and respond differently depending on that perception. Part of this process involves understanding a user's emotional state. To make the human-computer interaction more natural, it would be beneficial to give computers the ability to recognize emotional situations the same way as human does.

Automatic Emotion Recognition (AER) can be done by speech, facial expressions, brain signals and galvanic skin response etc. In the field of HCI, speech is primary to the objectives of an emotion recognition system, as are facial expressions and gestures. Speech is considered as a powerful mode to communicate with intentions and emotions.

In the recent years, a great deal of research has been done to recognize human emotion using speech information [1], [2]. Many researcher explored several classification methods including the Neural Network (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Maximum Likelihood Bayes classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN), Support Vector Machine (SVM) [3], [4].

The LIBSVM [5] supports multiclass problem and it is one of the mostly used SVM for the classification. It is simple and fast algorithm. It performs classification by constructing an N-dimensional hyperplanes that optimally separates the data into categories. The classification is achieved by a linear or nonlinear separating surface in the input feature space of the dataset. Its main idea is to transform the original input set to a high-dimensional feature space by using a kernel function, and then achieve optimum classification in this new feature space.

A Berlin Database of Emotional Speech [6] is used as a standard input speech emotion database for feature extraction and training SVM. The Berlin database of emotional speech was recorded at the Technical University, Berlin. The database contains speech with acted emotions in German language. It contains 493 utterances of 10 professional actors five males and five females who spoke 10 sentences with emotionally neutral content in 7 different emotions. The emotions were wut (anger), langeweile (boredom), ekel (disgust), angust (fear), freude (happiness), trauer (sadness) and neutral emotional state. While at the time of implementation only wut (anger), angust (fear), freude (happiness), trauer (sadness) and neutral these five emotions were taken.

Applications of Speech Emotion Recognition include psychiatric diagnosis, intelligent toys, lie detection, learning environment, educational software, and detection of the emotional state in telephone call center conversations to provide feedback to an operator or a supervisor for monitoring purposes.

II. SYSTEM IMPLEMENTATION

The importance of emotions in human-human interaction provides the basis for researchers in the engineering and computer science communities to develop automatic ways for computers to recognize emotions. As shown in figure 1, the input to the system is a .wav file from Berlin Emotion Database that contains emotional speech utterance from different emotional classes. After that features extraction process is carried out. In feature extraction process two features are extracted MFCC [7], [8] and MEDC [9]. After that the extracted features and their corresponding class labels are given as input to the LIBSVM classifier. The output of a classifier is a label of a particular emotion class. There are total five classes angry, sad, happy, neutral and fear. Each label represents corresponding emotion class.

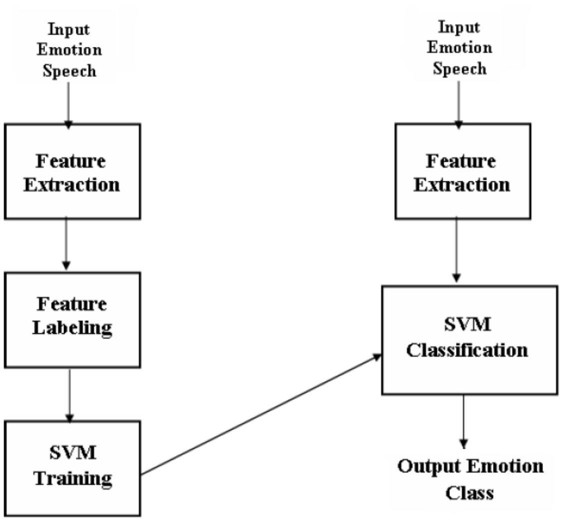

Figure 1. Speech Emotion Recognition System.

A. Feature Extraction

In previous works several features are extracted for classifying speech affect such as energy, pitch, formants frequencies, etc. all these are prosodic features. In general prosodic features are primary indicator of speaker's emotional state. Here in feature extraction process two features are extracted Mel Frequency Cepstral Coefficient (MFCC) and Mel Energy spectrum Dynamic coefficients (MEDC). Fig. 2 shows the MFCC feature extraction process. As shown in Fig. 2 feature extraction process contains following steps:

• Preprocessing: The continuous time signal (speech) is sampled at sampling frequency. At the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. This preemphasis is done by using a filter.

• Framing: it is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. Framing enables the non stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behavior within the short time period of 20-40 ms.

• Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame.

• FFT: Fast Fourier Transform (FFT) algorithm is ideally used for evaluating the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain.

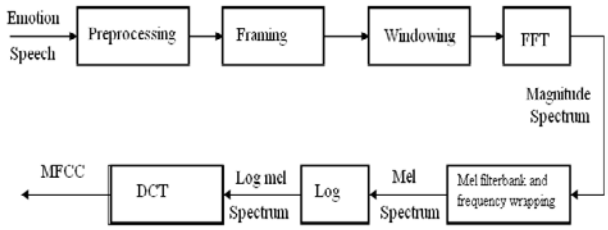

Figure 2. MFCC feature extraction

• Take Logarithm: The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition

• Take Discrete Cosine Transform: It is used to orthogonalise the filter energy vectors. Because of this orthogonalization step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components.

Another feature Mel Energy spectrum Dynamic coefficients (MEDC) is also extracted. It is extracted as follows: the magnitude spectrum of each speech utterance is estimated using FFT, then input to a bank of 12 filters equally spaced on the Mel frequency scale. The logarithm mean energies of the filter outputs are calculated $E_n(i)$, i= 1.....N. Then, the first and second differences of $E_n(i)$ are calculated. MEDC feature extraction process. The MEDC feature extraction process contains following steps shown in figure 3:

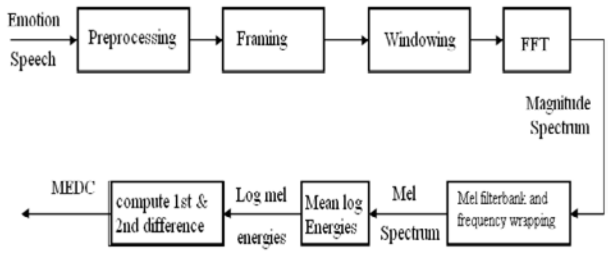

Figure 3. MEDC feature extraction

• Preprocessing, Framing, Windowing, FFT & Mel filterbank and Frequency wrapping processes of MEDC feature extraction are same as MFCC feature extraction.

• Take logarithmic mean of energies: In this process a mean log of every filter energies is calculated. This mean value represent energy of individual filter in a filterbank.

• Compute 1st and 2nd difference: The final Mel energy spectrum dynamics coefficients are then obtained by combining the first and second differences of filter energies**.**

B. Feature Labeling

In Feature labeling each extracted feature is stored in a database along with its class label. Though the SVM is binary classifier it can be also used for classifying multiple classes.

Each feature is associated with its class label e.g. angry, happy, sad, neutral, fear.

C. SVM Classification

In general SVM is a binary classifier, but it can also be used as a multiclass classifier. LIBSVM [5], [10] is a most widely used tool for SVM classification and regression developed by C. J. Lin. The LIBSVM provides linear, RBF, Polynomial and Sigmoid these four basic kernels. The system is train on Radial Basis Function (RBF) kernel. Advantage of using RBF kernel is that it restricts training data to lie in specified boundaries. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. The RBF kernel has less numerical difficulties than polynomial kernel. Polynomial kernels are less widely used than the RBF kernel. This might be because under similar training and testing cost, a polynomial kernel may not give higher accuracy.

III. EXPERIMENTATION AND RESULT

Berlin Emotion database contains 406 speech files for five emotion classes. Emotion classes Anger, sad, happy, neutral, fear are having 127, 62, 71, 79 and 67 speech utterance respectively. The LIBSVM is trained on MFCC and MEDC feature vectors using RBF kernel function. The LIBSVM is used to test these feature vectors. The experimentation is carried out by varying cost values for RBF kernel. Both gender independent and gender dependent experiments are performed. Using RBF kernel at cost value c=4, it gives recognition rate of 93.75% for gender independent case, 94.73% for male and 100% for female speeches.

The Confusion matrices using RBF kernel gender independent, male and female are shown in Table I, II, and III respectively.

Table IV shows the variation in percentage recognition with variable cost value ‘c’ of LIBSVMs RBF kernels for Gender independent, Male and Female cases. For Female case it gives maximum recognition rate ie. 100%. From these tables it is observed that after certain value of ‘c’ recognition percentage remains constant. Figure 4 show graphical representation of Table IV.

TABLE I. CONFUSION MATRIX OF RBF CLASSIFIER FOR GENDER INDEPENDENT

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	0	0	100	0	0
Neutral	0	6.25	0	93.75	0
Fear	0	0	30.76	0	69.24

TABLE II. CONFUSION MATRIX OF RBF CLASSIFIER FOR MALE

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	16.66	0	83.34	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	14.85	85.15

TABLE III. CONFUSION MATRIX OF RBF CLASSIFIER FOR FEMALE

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	0	0	100	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	0	100

TABLE IV. PERCENTAGE RECOGNITION USING VARIABLE COST VALUE ‘c’ FOR RBF KERNEL

	Cost Parameter									
	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8	c=9	c=10
Gender Independent	78.75	83.75	88.75	88.75	91.25	91.25	93.75	92.5	92.5	92.5
Male	86.84	92.1	92.1	94.73	94.73	94.73	94.73	94.73	94.73	94.73
Female	88.88	95.55	95.55	100	97.77	97.77	97.77	97.77	97.77	97.77

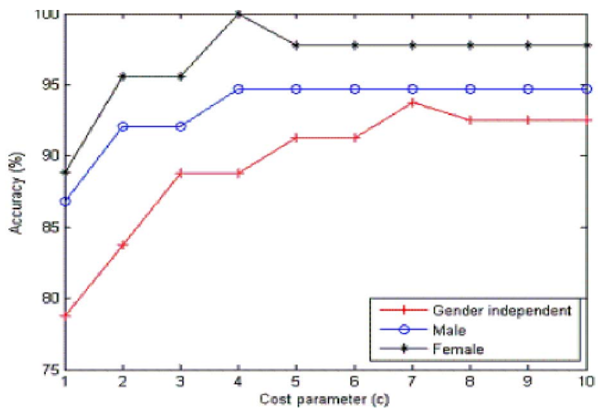

Figure 4. Accuracy Vs. Cost parameter Graph

CONCLUSION

In this paper Berlin emotion database of German language is used for feature extraction. MFCC and MEDC features are extracted from a speech files in .wav format. From experimentation and result it is observed that in case of gender dependent system gives slightly better result than gender independent case. It is speaker and text independent. It is also observed that results from LIBSVM by using RBF kernel function is 93.75%. Regarding LIBSVM using RBF kernels it is observed that by changing the parameters of a kernel functions better results can be obtain.

REFERENCES

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., Emotion recognition in human-computer interaction, IEEE Signal Processing magazine, Vol. 18, No. 1, 32-80, Jan. 2001.

[2] D. Ververidis, and C. Kotropoulos, Automatic speech classification to five emotional states based on gender information, Proceedings of the EUSIPCO2004 Conference, Austria, 341-344, Sept. 2004.

[3] Christopher. J. C. Burges, A tutorial on support vector machines for pattern recognition, DataMining and Knowledge Discovery, 2(2):955-974, Kluwer Academic Publishers, Boston, 1998.

[4] Tristan Fletcher, Support Vector Machines Explained, unpublished.

[5] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] Burkhardt, Felix; Paeschke, Astrid; Rolfes, Miriam; Sendlmeier, Walter F.; Weiss, Benjamin A Database of German Emotional Speech. Proceedings of Interspeech, Lissabon, Portugal. 2005.

[7] Fuhai Li, Jinwen Ma, and Dezhi Huang, MFCC and SVM based recognition of Chinese vowels, Lecture Notes in Artificial Intelligence, vol.3802, 812-819, 2005.

[8] M. D. Skowronski and J. G. Harris, Increased MFCC Filter Bandwidth for Noise-Robust Phoneme Recognition, *Proc. ICASSP-02*, Florida, May 2002.

[9] YL. Lin and G. Wei, Speech emotion recognition based on HMM and SVM, proceeding of fourth International conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.

[10] C.W Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification, *Technical Report*, Department of Comptuer Science & Information Engineering, National Taiwan University, Taiwan.