Comparison Between SVM & Other Classifiers For SER

Pranita N.Kulkarni
Student of M.E.
M.I.T. College, Aurangabad

Prof.D.L.Gadhe
Assistant Prof.
M.I.T. College, Aurangabad

Abstract

In this paper we approach to speech emotion recognition using support vector machine. Speech emotion recognition play important role in the field of Human Computer Interaction with wide range of application. To recognize the emotion from audio signal, many systems have been developed. The different classifier which is used for speech emotion recognition is reviewed in this paper. Emotion recognition is totally depends on speaker & utterance (Phrase). The support vector machine is used as classifier to classify different emotions such as anger, happiness, sadness, neutral & fear. The features extracting from these speech are the energy, pitch, linear prediction cepstrum coefficient (LPC), Mel frequency cepstrum coefficients (MFCC). This paper approaches towards the effectiveness, accuracy, simplicity of SVM classifier for designing the real time speech emotion recognition system.

1. Introduction

There are many ways of communication with one another such as speech, facial expression, eye contact, body language & so on [4]. But speech is one of the fastest & most efficient methods of communication between humans. And during communication, emotion plays an extremely important role in human life. Emotions [1][2] are nothing but the psychological description of one's feelings. There are some general emotions such as anger, happiness, sadness, neutral & fear. Speech emotion recognition helps to increase human-computer interaction. In our day-to-day life, computers have very important role & this method helps in communication between humans & computers. To make the human-computer interaction more efficient & natural, it would be beneficial to give computers.

Emotion recognition firstly classifies speech into categories, which are directly related to the psychological state of user. Emotional speech classification is not a easy task, it requires a set of successive operation such as voice activity detection, feature extraction, training & classification.

There are various methods used in field of speech emotion recognition for classification such as linear discriminateclassifier (LDC), K-nearest neighborhood (KNN), Gaussian mixture model (GMM), hidden Markov model (HMM), neural network (NN) & support vector machine (SVM).

The support vector machine is a learning algorithm addresses the general problem of learning to discriminate between positive & negative members of given n-dimensional vectors. The SVM is used for classification & regression purpose. The main idea of SVM classification is to a transform the original input set to a high dimensional feature space by using kernel function.

There is various application of speech emotion recognition like emotion recognition software for call center & helps in detection of emotional state provide feedback to an operator or a supervisor, lie detection, intelligent toys, psychiatric diagnosis, e-learning environment.

2. Literature Review

In speech emotion recognition, classifier recognizes the emotion in the speech. There are various types of classifier has been proposed which has advantages & disadvantages over the one another's. Generally there are two main types of information source can be used to identify the emotion of speakers, one is the word content of utterance & acoustic features. The linear discriminate classifier (LDC) is one of most simple well known & widely used classifier that finds a hyper plane that partitions the features into two decisions region. There is 70.5 % accuracy in LDC.

The K-nearest neighbor (K-NN) is amongst the simplest algorithm of all classifier. In this technique, firstly the object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its K nearest neighbors (k is a positive integer, typically small). Then classifier can classify all the utterance in the design set properly, if 'k' equal to 1, however its performance on the test set will reduced. K-NN classifier utilizes the information of pitch and energy and attains accuracy up to 64 % for four emotional set [11].

Gaussian Mixture Model (GMM) for speech emotion recognition is more suitable only when the global features are extracted from the training utterance. Also it provides a good approximation for the probability density functions by mixture of weighted Gaussian's. Expectation Maximization algorithm were used for computation of mixture coefficients. Each emotion is modelled in one GMM. [7] The decision is made for the maximum like hood model. The main advantage of utilizing GMM is of the fact that any signal by default will be in Gaussian shape & forms a bell shaped curve also due to its infinite range, it is also assumed that Gaussian distribution are well effective for speech recognition in presence of noise & speech spectra, because of its symmetry GMM attains 78.77 % accuracy.

Hidden Markov Model (HMM) has physical relation with speech signals production mechanism & due to this it is widely used in speech emotion recognition technique for isolation of word & emotion from speech. The HMM [9] is a doubly embedded stochastic process, which is not directly observable but has the capability of effectively modelling statistical variation in spectral features. HMM not only models the underlying speech sounds but also the temporal sequencing among the sounds. And this temporal modelling is advantageous for emotion recognition. The main limitation in building the HMM based recognition model is the features selection process. Because it is not enough that features carries information about the emotional states, but it must fit the HMM structure as well. HMM has better classification accuracy than other classifiers.

Neural Network (NN) has along history in classification pattern, due to their non-linear transfer function, their self-contained feature weighting capabilities and discriminative training. Neural Network classification technique is more suitable to classify to emotion anger & neutral. [5]

Support Vector Machine (SVM) plays an interesting role in the field of classification. Because it transfers the original features set to a high dimensional feature space by using the kernel function. Linear, polynomial, radial basis functions (RBF) are the kernel function which can be used in SVM model for large extent. SVM model show a high generalization capability due to their structural risk minimization oriented training. SVM classifier generally used in speech emotion recognition method due to their applications such as pattern recognition and classification problems. SVM classifier has correct classification rates of 89.4 %, 93.6 % & 88.9 % for male, female & gender independent cases resp. [6]

3. System Modeling

Dealing with the speaker's emotion is one of latest challenges in speech technologies. During speech emotion recognition, there are three different factors which want to recognize namely, speech recognition, then synthesis of emotional speech & finally emotion recognition. As we know that speech is a time varying signal, which represents the underlying patterns of emotions. To capture these time varying patterns of emotions, SVM can be effectively used [10]. The speech emotion recognition system mainly contains five modules such as follows:

- Speech Input (Audio n signal)
- Feature Extraction
- Feature Selection &Labeling
- SVM Classifier
- Recognized Emotional Output

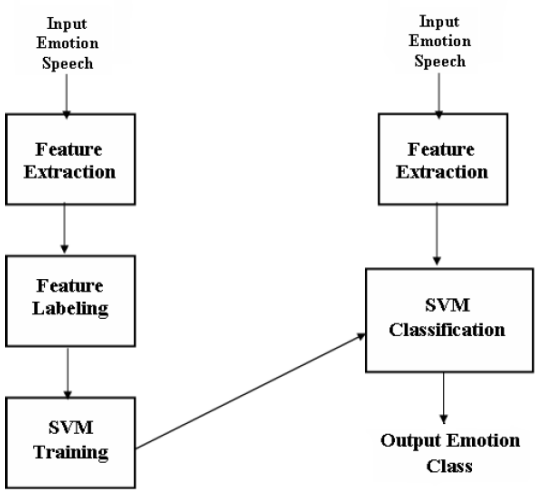

Figure 3.1 Implementation of Speech Emotion Recognition System

The main aim of speech emotion recognition is to automatically identify the emotional state of human being from his or her voice. And it is totally based on in-depth analysis of the generation mechanism of input speech signal. The level of naturalness of input speech signal is depend on data based used. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. Generally Berlin database can be used for emotion recognition. Berlin database is in German language, and consists of 535 emotional utterances recorded from 5 female & 5 male actors. Each speaker speaks at most 10 different sentences in 5 different emotion such as anger, happiness, sadness, neutral & fear.

3.1 Feature Extraction

Any emotion from the speaker's speech contains large numbers of parameters & the changes in these parameters will result in corresponding change in emotions. There are several features are extracted from speech such as energy, pitch, formant, frequencies, Mel Frequency Cepstrum Coefficients (MFCC), Mel Energy Spectrum Dynamic Coefficient (MEDC) etc all are called prosodic feature which are refers as primary indicator of the speakers emotional state. With the different emotional state, corresponding changes occur in the speak rate, pitch, energy & spectrum.

3.1.1 Energy Features

One of the most important speech features which indicates emotion is energy. To obtain the statistics of energy features, we use short term function which extracts the value of energy in each speech frame. Then we can calculate the statistics of energy in the whole speech sample by calculating the energy, such as mean value, max value, variation range & contour of energy.

3.1.2 Pitch Features

Pitch, often referred to as fundamental frequency (F0) is one of most important features for determination of emotion in speech. The pitch signal are also called glottal waveform which depends on the tension of the vocal folds & sub glottal air pressure & it is produced from the vibration rate of vocal cord. The pitch signal has two characteristics such as pitch frequency & glottal air velocity at the vocal fold opening time constant. Pitch frequency is directly get affected by numbers of harmonics present in the spectrum.

3.1.3 Formants Features

Spectral peaks of the sound spectrum |p(f)|' of the voice are called as formants. Formant is also used to mean an acoustics resonance of human vocal tract. It is often measured as an amplitude peak in the frequency spectrum of sound. Formants are measured by using a spectrogram or a spectrum analyzer [9]. The use of Linear Predictive Coding (LPC) to model formants is widely used in speech synthesis. The formant feature vector is 48 dimensional.

3.1.4 MelFrequency Cepstrum Coefficients (MFCC)

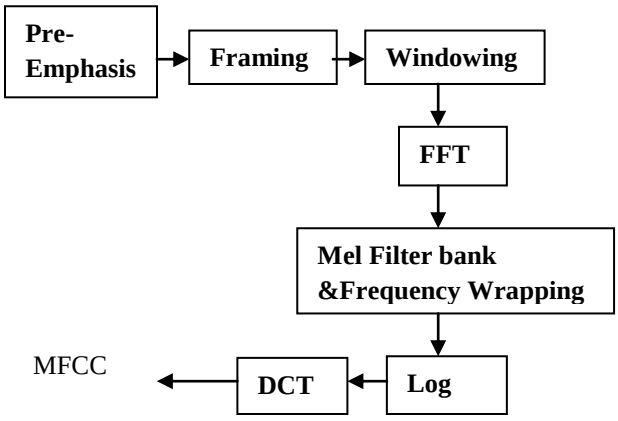

Figure 3.1.4.1 Structure of MFCC

- **Pre-Emphasis**

In this first step of MFCC, signal i.e. speech is passes through the filter which emphasizes the higher frequency. This process will increase the energy of signal at higher frequency.

Y [n] = X [n]- 0.95 X [n-1]

Let's consider a = 0.95, which make 95% of any one sample is presumedto originate from previous sample.

- **Framing**

Framing is the process of segmenting the speech signal into the small frames with the time length within the range of 20-40 ms. And speech signal are always obtained from analog to digital conversion (ADC).The voice signal is divided into framesof N samples. Adjacent frames are being separated by M (M<N).Typical values used are M = 100 and N= 256.Also framing enables the non stationary speech signal segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal.

- **Windowing**

Windowing is used as to give window shape to each individual frame, in order to minimize signal discontinuities at the beginning and the end of each frame.

If the window is defined as W (n), $0 \le n \le N-1$ where
N = number of samples in each frame
Y[n] = Output signal

X (n) = input signal
W (n) = Hamming window, then the result of windowing signal isshown below:
Y(n) = X (n) × W (n)
W(n)= 0.54-0.46 cos[2Πn/N-1]

- **Fast Fourier Transform**

Fast Fourier Transform is used for conversion of time domain to frequency domain of N sample speech signal. And this algorithm is used for evaluating the frequency spectrum of speech signal.The Fourier Transform is to convert the convolutionof the glottal pulse U[n] and the vocal tract impulse response[n] in the time domain.

- **Mel Filter Bank & Frequency Wrapping**

The Mel filter bank consists of overlapping triangular filters with the cut off frequencies determined by the centre frequencies of the two adjacent filters. The bank of filters according to Mel scale as shown in figure is then performed.

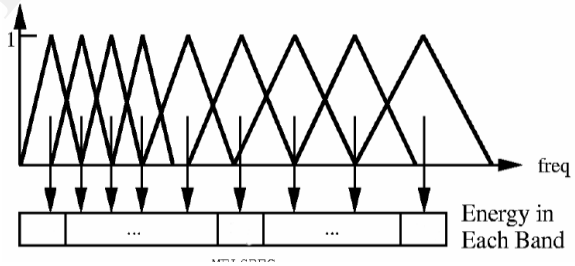

Figure 3.1.4.2 Mel scale filter bank.

This figure shows a set of triangular filters that are used to computea weighted sum of filter spectral components so that the output ofprocess approximates to a Mel scale. Each filter's magnitude frequencyresponse is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [7, 8]. Then, each filter output is the sum of its filtered spectral components.

- **Logarithm**

Logarithm is basically used for conversion of multiplication into addition. In this technique, it

simply converts the multiplication of the magnitude in the Fourier Transform into addition.

- **Discrete Cosine Transform**

Discrete Cosine Transform (DCT) converts the log Mel spectrum into time domain.The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance istransformed into a sequence of acoustic vector.

3.1.5 Mel Energy Spectrum Dynamic Coefficient (MEDC)

The MEDC feature extraction process contain following steps as shown in figure:

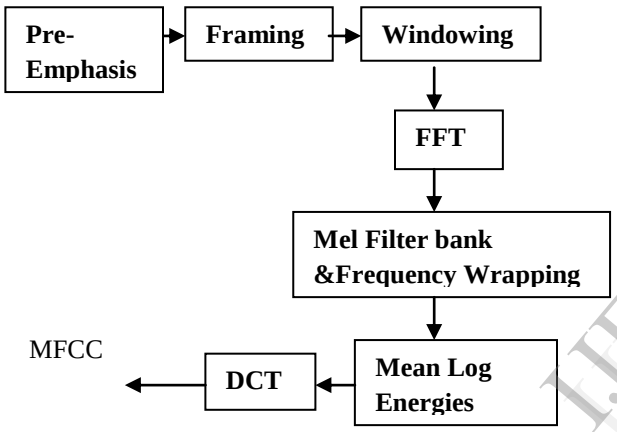

Figure 3.1.5 Structure of MEDC

- Preprocessing, Framing, Windowing, FFT & Mel filter bank & frequency wrapping process of MEDC feature extraction are same as MFCC feature extraction.The magnitude spectrum of each speech utterance is determined using FFT, then input to a bank of 12 filters equally spaced on the Mel frequency scale.

- **Mean Log Energies**
  In this process a mean log of every energies is calculated as $E_n(i)$, i=1…….N.Then, the first & second differences of $E_n(i)$ are calculated.

3.1.6 Linear Prediction Cepstrum Coefficients (LPCC)

Linear prediction coefficients which comes from Linear Predictive Coding (LPC) analysis is another alternative to filter bank analysis to represent the short-time spectrum.LPC analysis is an effective method to estimate the main parameters of speech signals.

3.2 Feature Labeling

In Feature labelling each extracted feature is stored in a database along with its class label. Here SVM is usedas a classifier to classify multiple classes. Each feature is associated with its class label e.g. angry, happy, sad, neutral, fear.

3.3 SVM Classification

SVM is a very efficient & simple classifier algorithm which is widely used for pattern recognition. Also it can have a very good classification performance than any other classifier. Thus we used this machine as classifier in this paper. LIBSVM is most widely used tool for SVM classification & regression. In SVM approach, the main aim of an SVM classifier is obtaining a function *f(x)*, which determines the decision boundary or hyper plane. This hyper-plane optimally separates two classes of input data points. SVM performs a non-linear mapping from an input space to a high-dimensional space through a kernel, which is an important component for SVM learning.

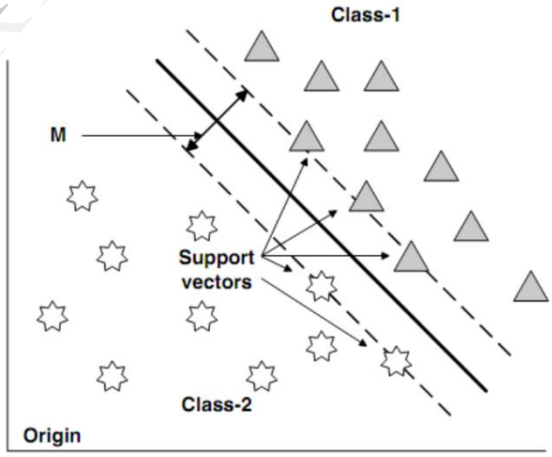

Figure 3.3 Linear Classification using SVM

In this work, we investigated three SVM kernels

1) Linear Kernel
2) Polynomial Kernel
3) Radial Basis Function (RBF) Kernel

But here we used Radial Basis function kernel in training phase. Advantage of using Radial Basis function kernel is that it restricts training data to lie in specified boundaries. The RBF kernel has less

numerical difficulties than polynomial & linear kernel.

4. Conclusion

Speech emotion recognition (SER) has become an area of active research interest in recent years. SER plays an important role in building more anthropomorphic human-computer interfaces. As a machine learning task, successful SER requires both reliable machine learning techniques and emotion discriminating features. Whilevarious learning algorithms have been proposed for SER, constructing powerful features, specifically spectral features, remains an open challenge. Emotion recognition is an important step toward implementingan emotional speech recognition system. The type and number of emotional states, extracted features, feature selection algorithm, and type of the classifier are importantfactors in the accuracy of emotion recognition systems.In this paper, support vector machine was studied for speech emotion recognition system. Speech feature were extracted from the emotional speech sample such as energy, pitch, formants, MFCC, MEDC. Automatic speech emotion recognitions are increasing now a day because it results in better interaction between human & computer.

References

[1] Christopher. J. C. Burges, A tutorial on support vectormachines for pattern recognition, Data Mining andKnowledge Discovery, 2(2):955-974, Kluwer AcademicPublishers, Boston, 1998.

[2] C.W Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide toSupport Vector Classification, *Technical Report*, Departmentof Computer Science & Information Engineering, National Taiwan University, Taiwan.

[3] S. Emerich, E. Lupu, A. Apatean, "Emotions Recognitions by Speech and Facial Expressions Analysis", 17th European Signal Processing Conference, 2009.

[4] L.R. Rabiner & B.H. Juang, *Fundamentals of speechrecognition* (Englewood Cliffs, NJ: Prentice-Hall, 1993).

[5] D. Morrison, Ruili Wang; & L.C. Silva, Spoken affectclassification using neural networks, *IEEE InternationalConference on Granular Computing, 2*, 2005, 583-586.

[6] Corinna Cortes and V. Vapnik, "Support-Vector Networks", MachineLearning, 20, 1995.

[7] Xuan-Hung et al., Speaker-dependent emotionrecognition for audio document indexing, *InternationalConference on Electronics, Information, andCommunications*, 2004.

[8]L.R.Rabiner and B.H.Juang. "Fundamentals of Speech Recognition", Upper Saddle River; NJ: Prentice-Hall, 1993

[9] Albornoz E. M., Crolla M. B. and Milone D. H."Recognition of Emotions in Speech". Proceedings of17th European Signal Processing Conference, 2009.

[10] 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] I. Chiriacescu, "Automatic Emotion Analysis

[11] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features and Methods", Elsevier Speech communication, vol. 48, no. 9, pp. 1162-1181, September, 2006.

[13] Ying Wang, Shoufu Du, Yongzhao Zhan, Adaptive and Optimal Classification of Speech Emotion RecognitionFourth International Conference on Natural Computation.

[14] O.Khalifa, S.Khan, M.R.Islam, M.Faizal and D.Dol, 2004."Text Independent Automatic Speaker Recognition".3rd International Conference on Electrical & Computer Engineering, Dhaka, Bangladesh, pp.561-564.

[15] 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.

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

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