Article can be accessed online at http://www.publishingindia.com

Real-time Speech Emotion Recognition Using Support Vector Machine

P. Vijayalakshmi*, A. Anny Leema**

Abstract

In this paper we present an approach for Real-time emotion recognition from speech using Support Vector Machine (SVM) as a classification technique. Automatic Speech Emotion Recognition (ASER) is an upcoming research area in the field of Human Computer Interaction Intelligence (HCII). Human emotions can be detected from their speech signals by extracting some of the speech acoustic and prosodic features like pitch, Mel frequency Cepstral Coefficient (MFCC)and Mel Energy Spectrum Dynamic Coefficient (MEDC). Here SVM classifier is used to classify the emotions as anger, fear, neutral, sad, disgust, happy and boredom. UGA and LDC datasets are used for offline analysis of emotions using LIBSVM kernel functions.With this analysis the machine is trained and designed for detecting emotions in real time speech.

Keyword: Support Vector Machine, Speech Signal, Experimentation, Emotion Analysis. Controller (PDC)

Introduction

Human Computer Interaction Intelligence is an emerging field which is used to improve the interactions between users and computers by making computers more respondable to the user's needs. Emotion Recognition is a recent research topic in the field of HCI which aims to achieve a more natural interaction between machine and humans. Today's HCI system has been developed to identify who is speaking or what he/she is speaking. If the computers are given an ability to detect human emotions then they can know how he/she is speaking and can respond accurately and naturally like humans do.

Automatic Emotion Recognition can be done in two ways; Speech signals and Image gestures. Since speech is considered as a most powerful tool for communication, dealing with speech emotions is one of the most difficult tasks. In recent years, many researches have been done to recognise emotions from human speech. Many methods and classification algorithms like Neural networks (NN), K-Nearest Neighbour (k-nn), Support vector machine (SVM), Hidden Markov model (HMM) have been developed to classify human emotions based on training datasets (Ayadi, Kamel & Karray, 2011).

However, many applications which are used today require affective information from real time data.So the consideration of real-time constraints is also important for emotion recognition from speech. Good results can be obtained by using standard classifier algorithms. Some applications where real time emotions can be used are in call centres, robots, gaming, in smartphones where songs can be played based on human emotion etc.

In this approach, basic features of speech signals like pitch, MFCC and MEDC are extracted from both offline and real time speech and they are classified into different emotional classes by using SVM classifier. We use SVM since it has better classification performance than other classifiers (Kulkarni & Gadhe, 2011). Support Vector Machine is a supervised learning algorithm which addresses general problem of learning to discriminate between positive an dnegative members of given n-dimensional vectors. The

* Master of Computer Applications Master of Computer Applications B.S.Abdur Rahman University, Chennai, Tamil Nadu, India. Email: viji181090@gmail.com

** B.S. Abdur Rahman University, Chennai, Tamil Nadu, India. Email: annyleema@bsauniv.ac.in

SVM can be used for both classification and regression purposes. The classification can be done linearly or non-linearly. Here the kernel functions of SVM are used to recognise emotions with more accuracy.

This paper is organized as follows: the second section tells on related work, the third section focuses on the implementation of the system. The next section discusses the experimentation and results obtained, and the final section summarizes main conclusions.

Related Work

Under this point the focus is on the literature that is available for speech and emotions.

Speech

Speech is the primary means of communication between humans. It is a complex signal containing information about message, speaker, language, emotional states and so on.

Speech is referred in terms of the speech production and speech perception of the sounds used in vocal language. Speech production defines how the spoken words are processed and how their phonetics are formulated whereas speech perception refers to process by which humans can interpret and understand the sound present in human language.

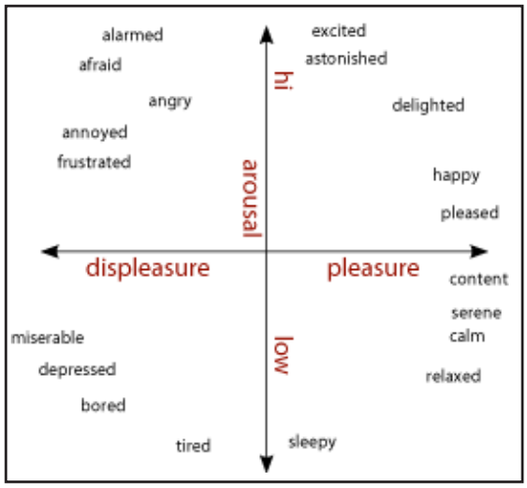
Emotions

Emotions are defined as changes in physical and psychological feeling that influences behaviour and thought of humans.It is associated with temperament, personality, mood, motivation, energy etc. The emotions can be categorised by grouping them with dimensions.

Dimensional Model of Emotions

Dimensional model usually conceptualize emotions by defining them in 2 or 3 dimensions. All dimensional models incorporate arousal and valence. In 1980, James Rusell introduced Circumflex of effect (Fig. 1).He mapped emotions in two dimensional axes as pleasure/displeasure in horizontal axis and high/low arousal in vertical axis.

Fig. 1: Dimensional Model of Emotions



Existing System

We survey existing emotional classification techniques and methods used for feature extraction and selection. As discussed in research by Lin & Wei (2005), speech prosodic and acoustic features are extracted in time domain and frequency domain. Here inorder to reduce complexity we choose only basic features like pitch, energy and formants.

Commonly many classifiers are used for emotion recognition like Guassian Mixture Models (GMM), Hidden Markov Models (HMM), Support Vector Machines (SVM), K-Nearest Neighbour (K-NN) etc. As specified by Kulkarni & Gadhe (2011), since the performance of SVM is better, it is taken in our project and the emotional states are classified in real time.

Some of the important research concerns of emotion recognition system are discussed below:

- Emotion recognition systems are influenced by speakers and language dependent information. But these systemshaveto be speaker and language independent (Koolagudi & Rao, 2010).

- There is no standard speech corpus to recognise emotions. Some speech corpus are recorded using experienced artists whereas some with inexperienced one. And some speech corpus had limited Emotional utterances. That is 2-3 emotional states.They do not contain different classes of emotions (Ververidis & Kotropoulos, 2006)
- Feature extraction is one of the most important factors in emotion recognition systems.Identification and selection of particular features from a list of features is a complex task. Along with features, computational models and techniques must also be identified which could make the performance better (Ayadi et al., 2011).
- Emotion Recognition system should recognise emotions in real time. It has to be robust and noise should be eliminated from the speech utterances which is a complex task.

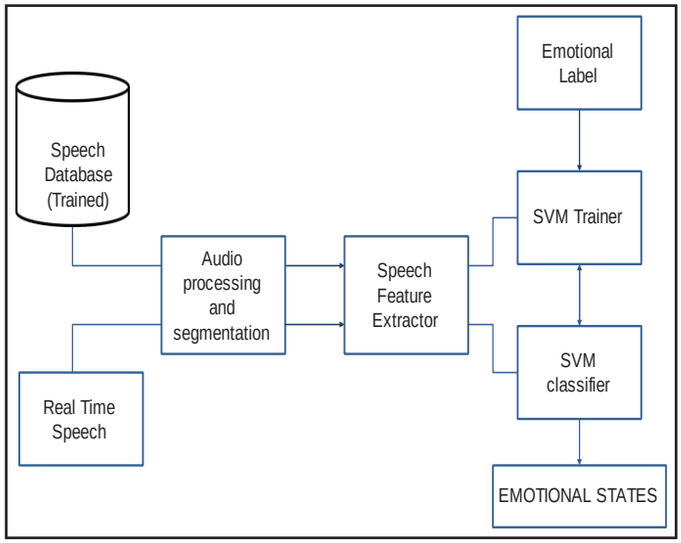
Since these are some of the concerns,a proper measure has been taken to avoid these problems and to make the system more accurate.

System Implementation

Dealing with speaker's emotion is one of the main challenging tasks in speech technology. The design of emotion recognition system basically contains different modules which are depicted in Fig. 2. They are Speech acquisition, Feature extraction from signals, Training the machine with feature sets and classifying emotions through SVM.

The input to the system will be given as a .wav file from any emotional database. Here we use LDC and UGC databases which contain emotional speech utterances with different emotional states. From the given input,the speech features like Pitch, MFCC and MEDC values are extracted and they are fed into LIBSVM classifier along with their class labels. The machine is trained in such a way to classify all possible emotions. Once the system is trained well with acted speakers, whenever a real time speech is given as an input the SVM classifier will automatically predict the emotional states by referring to the training sets. If not, again the machine will be trained iteratively until it gives accurate results.

Fig. 2: Emotion Recognition System



Speech Acquisition

The first module of emotion recognition system involves preprocessing of the given speech utterance. Since the input is given as continuous speech signal,it has to be segmented to individual frames for classification purpose. This step generally involves the digitization of the given input speech.

Feature Extraction

Feature extraction is the process by which the measurements of the given input can be taken to differentiate among emotional classes. In previous works,several features have been extracted from the input speech signals such as pitch, energy, formants, DTW etc. In order to reduce the complexity of our system and also to predict emotions in real time we extract the features like Pitch, MFCC (Ma, Huang & Li, 2005) and MEDC from utterances.To extract these features many algorithms and techniques are used.

Pitch Features

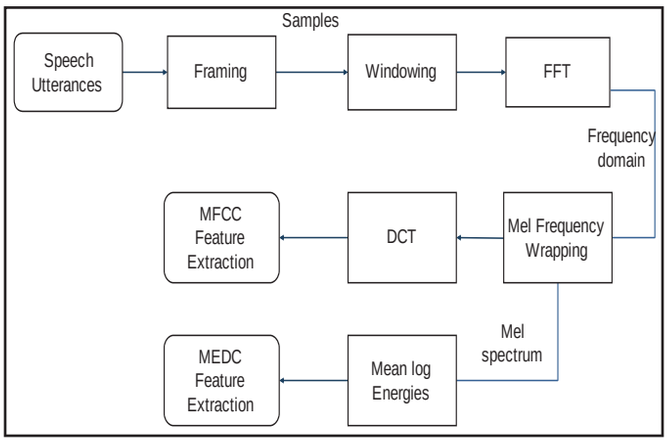
Pitch is a fundamental frequency F0 of a speech signal. It generally represents the vibration produced by vocal cords during sound production.Usually the pitch signals have two characteristics pitch frequency and glottal air

velocity. The pitch frequency is given by the number of harmonics present in spectrum. The estimation of pitch is one of the complex tasks.Here we use YIN Pitch Estimation algorithm to detect pitch values.

MFCC

Mel Frequency Cepstral Coefficients are most widely used feature in speech recognition. It was introduced by David and Mermelstein.The main purpose of the MFCC processor is to mimic the behaviour of the human ears.The main steps of MFCC are identified (Khalifa et al., 2004) and the block diagram for MFCC is shown in Fig. 3.

Fig. 3: MFCC and MEDC Feature Extraction



Usually the speech input is recorded with a sample rate of 16000 Hz through microphone.The steps for calculating MFCC are described below:

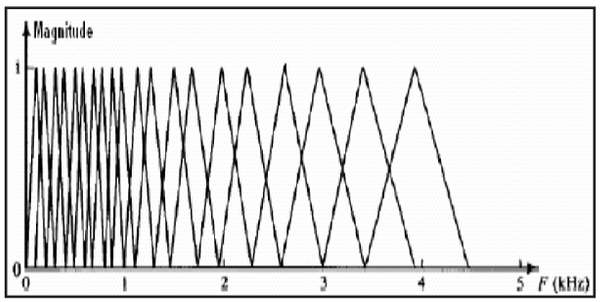
- Framing: In this step, the continuous input speech is segmented into N sample frames.The first frame consists of N samples, second frame consists of M samples after N, third frame contains 2M and so on. Here we frame the signal with time length of 20-40ms. Therefore the frame length of 16Khz signal will have 0.025*16000 = 400 samples.
- Windowing: This step in processing is used to window each individual frame inorder to remove the start and end discontinuities.Hamming window is mostly used to remove distortion.
- Fast Fourier Transform: Here FFT algorithm is used for converting the N samples from time domain to frequency domain. It is used for evaluation frequency spectrum of speech. In this project we use Cooley-Tukey algorithm of FFT.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \quad k = 0, \ldots, N-1.$$

- Mel Filter Bank: This step maps each frequency from frequency spectrum to Mel scale.The Mel filter bank will usually consist of overlapping triangular filters with cut off frequencies which is determined by center frequency of two filters.

The Mel filters are graphically shown in Fig. 4.

Fig. 4: Mel Filter Bank with Overlapping Filters



- Cepstrum: Here the obtained Mel spectrum is converted back to time domain with the help of DCT algorithm.

The result we obtain is Mel Frequency Cepstral Coefficient vector values.

MEDC

Mel Energy Spectrum Dynamic Coefficient is an another important spectral feature that is obtained from speech utterances (Khalifa et al., 2004).The process of MEDC techniques is same as MFCC (Fig. 3). Here the magnitude spectrum of each speech utterances is calculated from FFT algorithm and they are equally placed in mel frequency scales. The mean log energies of output obtained from filters are calculated.

$$\text{En (i) where i} = 1, 2 \ldots N$$

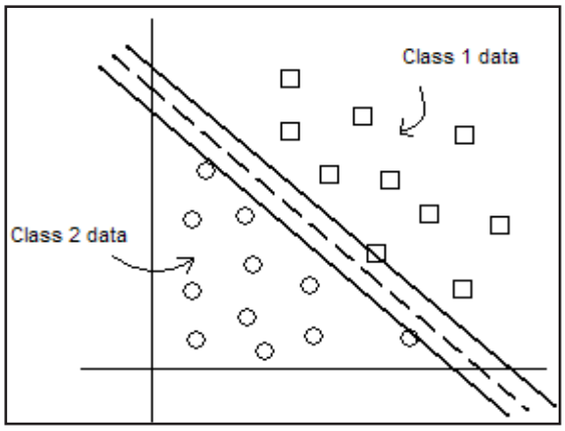
The obtained values represent Mel Energy Spectrum Dynamic Coefficient vector values.

SVM Training

During the training phase of emotion recognition system, the extracted feature values are labelled into different emotional classes. Here the feature values are stored in database with their corresponding class labels. The SVM kernel functions are used for training process.

SVM is a very simple and efficient classifying algorithm which is used for pattern recognition (Fig. 5) (Pao *et al.*, 2006). Support Vector Machines were introduced by Vladimir Vapnik in 1995.The main aim of this algorithm is to obtain a function $f(x)$, which is used to determine Hyper planes or boundaries. These hyper planes are used to separate different classes of data input points.

Fig. 5: SVM Classifier

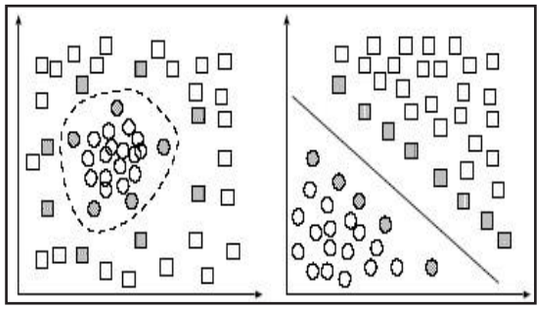


Though SVM follows a linear classification, it can also perform non-linear mapping of data points by using its kernel functions (Hsu, Chang & Lin, 2010). The SVM has three kernel functions namely,

1. Polynomial kernel
2. Linera kernel
3. Radial Basis kernel

In training phase we use Radial Basis kernel function because it restricts the training data to lie within the specified boundaries as depicted in Fig. 6. The RBF kernel has less numerical difficulties when compared with other kernel functions.

Fig. 6: SVM Kernel Fuctions



(a) Radial Basis Function (b) RBF mapping

LIBSVM is the most widely used tool for SVM classification and regression purpose. During training phase, feature values of speech signals that are obtained from emotional databases speech utterances are fed into LIBSVM along with their class labels. SVM model is developed for each emotional state by using their feature values. Here more than 80 utterances of speech from LDC datasets are used for developing SVM training model.

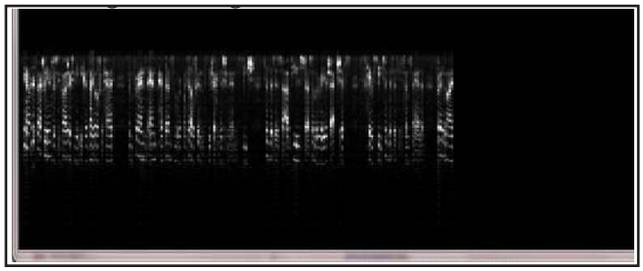
SVM Classification for Real Time Speech

Once the training model has been developed with more training datasets, it is easy to predict the emotional states in real time. The real time input speech is taken through microphone since noise distortion will be less. Speech features are extracted from signals and with the help of LIBSVM, these features values are compared with predicted values of SVM training models and the emotions are classified automatically.

Experimentation and Result

The performance of speech emotion recognition depends upon many factors such as the quality of speech that is obtained without noise, features extracted from speech signals and the classification method used. The spectrogram specifying frequency for each frame is given in Fig. 7.

Fig. 7: Spectrogram of Each Frame with Corresponding Frequency



Emotional Database

Here in this project we make use of UGA and LDC datasets (www.ldc.upenn.edu). The UGA datasets have emotions acted by students of university of Georgia.It contains nearly 800 samples. In LDC dataset each speech utterance is recorded for one to two seconds with 60ms

segments.For each emotional class there are about 25 utterances spoken and acted by speakers.

Emotion Recognition with LIBSVM

During training phase of Emotion Recognition system,we train the classifier with emotions namely Anger,Happy, Sad,Neutral,Disgust,Surprised.The LIBSVM is trained with MFCC and MEDC feature vectors using RBF kernel functions.Gender Independent files are used during experimentation.The result obtained by SVM is given in Table 1 with confusion matrix which represents classification of given class.

Table 1: Confusion Matrix of SVM for Training Models with RBF Kernel Function c = 9

Emotion Label	*Emotion Recognition(%)*					
	H	*S*	*F*	*S*	*N*	*A*
Happy	100	0	0	0	0	0
Sad	0	100	0	0	0	0
Fear	0	4.7	95.3	0	0	0
Surprise	2.9	0	18.6	78.5	0	0
Neutral	0	0	0	0	100	0
Anger	0	0	13.6	0	0	86.4

During real time recognition, gender independent speech and gender dependent speech are taken from speakers and their feature values are fed into LIBSVM. By changing the cost value c = 9 and d = 10.05 of RBF kernel functions (Onen & Alpaydin, 2011),we could predict the classification of emotional states. Their average values are specified in Table 2.

Table 2: Recognition(%) for Speaker Independent and Speaker Dependent Utterances

Emotions	*Speaker independent*	*Speaker Dependent*
Happy	56.2	79.6
Sad	59.8	82.5
Fear	53.6	69.2
Surprise	55.2	68.9
Neutral	60.3	85.1
Anger	51.3	74.4

Conclusion and Future Works

Today, the Speech Emotion Recognition has become one of the most important research areas. It plays an important role in Human Computer Interaction. Recognising emotions in real time is a complex task. The type and the number of emotional classes, feature selection, classification algorithm are the important factors of this system. Here we have used SVM for obtaining higher classification accuracy. MFCC and MEDC feature values are extracted from speech utterances and emotional states are classified using SVM RBF kernel function.For training set with acted speakers we obtained recognition rate by 81% with LDC datasets. During Real time recognition of speech the accuracy is nearly 60% which is obtained by changing the values of RBF cost value c = 9 and degree d = 10.05.Since emotion recognition is done in real time, distortion of noise makes the system with less accuracy.

In our future works, we work on experimenting the system by combining some other feature values with MFCC and MEDC inorder to make the performance of system better.This emotion recognition system is implemented in voice mail application which queues up the calls based on emotions.

References

Ayadi, M. E., Kamel, M. S. & Karray, F (2011).Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572-587.

Emotional Prosody Speech and Transcripts from the Linguistic Data Consortium. (2002). Retrieved from http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId = LDC2002S28.

Hsu, C.W., Chang, C. C. & Lin, C. J. (2010). *A Practical Guide to Support Vector Classification*. Department of Computer Science & Information Engineering, National Taiwan University, Taiwan.

Khalifa, O., Khan, S., Islam, M. R., Faizal, M. & Dol, D. (2004). Text Independent Automatic Speaker

Recognition. 3rd International Conference on Electrical& Computer Engineering, Dhaka, Bangladesh.

Kulkarni, P. N. & Gadhe, D. L. (2011). Comparison between SVM & Other Classifiers for SER. *International Journal of Research and Technology*, January, 2(1), 1-6.

Koolagudi, S. G. & Rao, K. S (2010). *Real Life Emotion Classification using VOP and Pitch Based Spectral Features*. India: Jadavpur University.

Lin, Y. L. & Wei, G. (2005). *Speech Emotion Recognition based on HMM and SVM*. Paper Presented on 2005

at Fourth International Conference on Machine Learning.

Ma, J., Huang, D. & Li, F. (2005). SVM based recognition of chinese vowels. *Artificial Intelligence*, 3802, 812-819.

Onen, M. G. & Alpaydin, E. (2011).Multiple kernel learning algorithms. *Journal of Machine Learning Research*, July, 12, 2211-2268.

Pao, T., Chen, Y., Yeh, J. & Li, P. (2006). *Mandarin Emotional Speech Recognition based on SVM and NN*. Paper presented on 2006 at 18th International Conference on Pattern Recognition (ICPR'06), (1, pp. 1096-1100).

Ververidis, D. & Kotropoulos, C. (2006). *A State of the Art Review on Emotional Speech Databases*. Presented at 11th Australian International Conference on Speech Science and Technology, Auckland, New Zealand.