Speech Emotion Recognition System

Ranjith N [1]
MTech, 4th SEM, Industrial Electronics,
RNSIT, Bangalore
ranjithharsha91@gmail.com

Srikanth G N [2]
Associate Prof, Dept of E & I,
RNSIT, Bangalore
srikanthgn@yahoo.com

Abstract— Speech Emotion Recognition (SER) is a research topic which has a wide range of applications. The features of speech such as, Mel Frequency cepstrum coefficients (MFCC) are extracted which are uttered in the speech. To classify different emotional states such as boredom, happiness, sadness, neutral, anger, from various emotional sound tracks from a database containing emotional speech SVM is used as classifier. SVM is used for classification of emotions. The accuracy obtained in SVM is very much higher.

Keywords— *Speech emotion, Emotion Recognition, SVM, MFCC*

*****

I. INTRODUCTION

In the field of Human computer interaction Speech Emotion Recognition is a recent research topic. Computers have become an integral part of our lives, hence the need for a more natural communication interface between computers and humans are necessary. A computer should be trained such that it should be capable of perceiving its present situation and respond accordingly. The main step is to understand the users emotional state. If the computers have the ability to recognize emotions as human beings it would be very beneficial in future.

Speech is the an essential to the objective of an emotion recognition system in the field of HCI . An effective mode to communicate with intentions and emotions is speech. From the Recent times, using speech information a great deal of research is carried out to identify human emotions [1], [2]. Many researches explored several classification methods including the Gaussian Mixture Model (GMM), Neural Network (NN), Hidden Markov Model (HMM), Maximum Likelihood Bayes classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN), Support Vector Machine (SVM) [3], [4]. Support Vector Machine is used as a classifier. It can classify by constructing an N-dimensional hyper planes that optimally separates the data points. The classification can take place either by a linear or nonlinear separating surface in the input feature space of a dataset. The main intention is to transform the original input set to a high dimensional feature space by using a kernel function, and then achieving appropriate classification in the new feature space.

In this paper section II Gives Requirement of Speech Emotion recognition, section III Gives information about emotion Database, section IV Gives SER Feature Extraction Technique & section V gives details about SVM classification and section VI gives Experimental Results & VII Is conclusion and future scope.

II. SER Requirement

Human-Robotic Interfaces: robots are taught to interact with humans and recognize emotions. So that robots, should be able to understand not only spoken words, but also other information, such as the emotional and other status and should act accordingly.

In call-centers: SER renders help to detect the potential problems that arise due to an unsatisfactory course of interaction. A frustrated client can be offered the assistance of human operators or from a reconciliation strategy.

In intelligent spoken tutoring systems: detection & adaption to a student emotions is considered to be an important strategy for enhancing the performance gap among the human and computer tutors as emotions can have considerable impact on the performance and learning.

III. EMOTION DATABASE

The important issue considered in the evaluation of a emotional speech recognizer is the degree of naturalness of the database to be used to assess its performance. Improper conclusions may be established if a poor-quality database is used. The design of a database is critical, & more important to carry out the classification task. The classification task can be defined based on the number and type of emotions in the database. Databases can be divided into three type:

- Acted emotions database, is obtained by asking an actor to speak with a predefined emotion.
- Real-life systems databases.
- Elicited emotions databases.

In the work carried out here the database of type One is created. Have recorded few person's voice in different emotions which are pre-defined. Later we calculate the MFCC and then create a database.

 ISSN: 2321-8169
2839 - 2842

IV. FEATURE EXTRACTION

Feature extraction stage is known as speech processing front end. Feature Extraction main aim is to simplify recognition of speech data without losing the acoustic properties from which a speech can be defined.
The schematic diagram for Feature extraction steps is depicted in Figure 1.

Processing: In processing the input signal is de-noised. The silent parts of the input speech signal does not contain any information, so they are eliminated by thresholding.
.

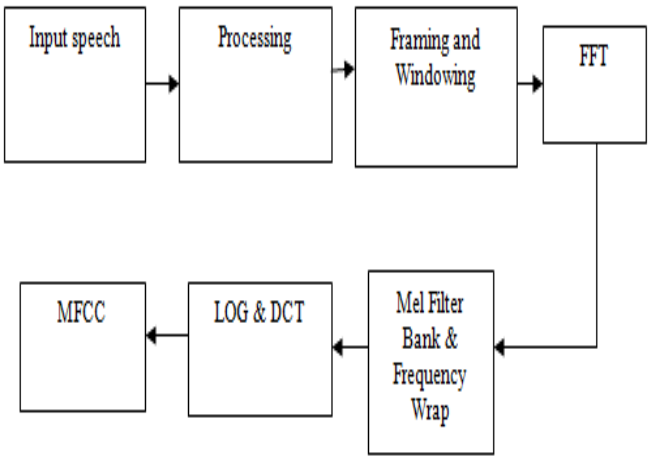


Fig 1 Feature Extraction process

Frame Blocking: The speech signal characteristics stays stationary for a short period of time interval (It is called quasi-stationary). Because of this, the speech signals can be and they are processed in short time intervals. The signal is divided into small frames. The overlapping of frames with its previous other frame are predefined. By overlapping the transition from frame to frame will be easy.
Windowing: Windowing step is carried out to eliminate the discontinuities at the edges of the frames. Hamming window is generally used.
FFT: FFT is the next step which is carried out to transform each frames. This algorithm is used to evaluate the speech spectrum. FFT is carried out in order to transform the domain from time to frequency.
Mel filter bank and Frequency warping: This filter bank is applied in the frequency domain; therefore it takes triangle-shape windows on the spectrum. An efficient way of implementing this is to consider the triangular filters in the Mel scale which are linearly spaced and have centre frequencies.
Taking Log: The real values defining real cepstrum uses the logarithm function. Real cepstrum uses only the information in the magnitude of the spectrum, which allows the reconstruction of the signal. Log is generally taken to reduce the computational complexities. It just converts the multiplication of magnitude in Fourier transform to addition.
DCT: By using DCT the number of vectors to components would be reduced and the information of filter energy vectors which are orthogonalized is compacted.
MFCC: After following the above steps Mel Frequency cepstral coffiecients can be obtained.

V. SVM CLASSIFICATION

Support vector machine (SVM)[11] is one of a effective approach that can be employed for pattern recognition. SVM basics is briefly is explained. SVM basics can be found in references Cristianini and Shawe-Taylor (2000), Burges (1998), and Scho lkopf and Smola (2000), Dmitriy Fradkin and Ilya Muchnik[11]. In SVM approach, the SVM classifier aim is to obtain a function *f(x)*, which determines the decision hyper-plane or boundary. The hyper-plane is used to separate the two classes of data points. The hyper-plane is as shown in Fig. 2.Where M is used to indicate the margin, which shows the distance from the hyper-plane to the closest point for both classes of data points (Ferna ´ndez Pierna et al., 2004; Gunn et al., 1998). In SVM, the data points can be separated into two types: linearly separable and non-linearly separable (Ferna ´ndez Pier-na et al., 2004).
In a linearly separable data points, a training set of instance-label pairs (x_n,y_n), $n = 1,2,3, \ldots, t$ where $x_n \in R^d$ and $y_n \in \{+1, -1\}$, the data points are termed as:

$$\langle w, x_n\rangle + h_0 \geq 1, \text{ if } y_n = 1,$$
$$\langle w, x_n\rangle + h_0 \leq 1, \text{ if } y_n = -1, \text{ ----(1)}$$

where $\langle w, x_n\rangle$ shows the inner product of w and x_n. The inequalities in Eq. (1) can be combined as in

$$y_n [\langle w, x_n\rangle + h] - 1 \geq 0, \text{ if } n = 1.....t, \text{ ------(2)}$$

The SVM classifier is used to place a decision boundary by using maximal margin among all possible hyper planes.
There are no problems with local minima in SVM. Back-propagation in artificial neural network (ANN), some local minima problems exist. This is termed as an advantage of support vector machines ANN.

For solving the dual optimization problem, most appropriate w and h parameters of optimal hyper-plane are estimated with respect to a_{k1}. The acquired optimal hyper-plane f(x) can be expressed as follows:

$$f(x) = \sum_{n=1}^{t} yn \cdot an < xn \cdot x > + h \text{ -----(3)}$$

The x_n is called a Support vector, If x_n input data point has a non zero lagrange multiplier a_n. The data points outside support vector is not necessary for calculating the f(x).
If data points are non-separable and non-linear SVM is used, Eq. (1) should be changed as:

$$\langle w, x_n\rangle + h \geq +1 - X_n, \text{ for } y_n = 1,$$
$$\langle w, x_n\rangle + h \leq -1 + X_n, \text{ for } y_n = -1, \text{------(4)}$$

where x_n is a non-negative slack variables and $x_n >= 0$, $n = 1, \ldots, t$. This x_n variable keeps the constraint violation as small as and provides the minimum training error (Huang & Wang, 2006). The optimal decision function (f(x)) is same as in Eq. (2).

 ISSN: 2321-8169
2839 - 2842

If non-linear SVM is used, a Kernel function *(Ø)* is used for training data points in input space transformed to a higher dimensional feature space.
The optimal hyper-plane f(x) of a non-linear SVM can be given as below:

$$f(x) = \sum_{n=1}^{t} yn \cdot an\ kernel\ < xn \cdot x > + h$$

$$f(x) = \sum_{n=1}^{t} yn \cdot an\ kernel\ < xn \cdot x > \text{-------(5)}$$

SV is a support vector number. Kernel functions which are linear, polynomial, radial basis function (RBF), Fourier, and exponential radial basis function (ERBF) are used for SVM. The kernel functions are given in Eqs. (6)–(10), respectively. To increase the classification accuracy of SVM the kernel parameters of these kernel functions should be properly set.

Linear Kernel Function:

$$\textbf{kernel } (x_n, x_m) = (x_n \cdot x_m) \text{ ------(6)}$$

Polynomial Kernel Function:

$$\textbf{kernel } (x_n, x_m) = (x_n \cdot x_{m+1})^d \text{-----(7)}$$

Radial Basis Function(RBF):

$$\textbf{kernel } (x_n, x_m) = \exp(-\|xn - xm\|^2 / 2\sigma^2) \text{----(8)}$$

Fourier:

$$\textbf{kernel } (x_n, x_m) = (\sin(d + 1/2)(x_n - x_m)) / (\sin(1/2(x_n - x_m))) \text{---(9)}$$

Exponential Radial Basis Function:

$$\textbf{kernel } (x_n, x_m) = \exp(-\|xn - xm\| / 2\sigma^2) \text{ ----(10)}$$

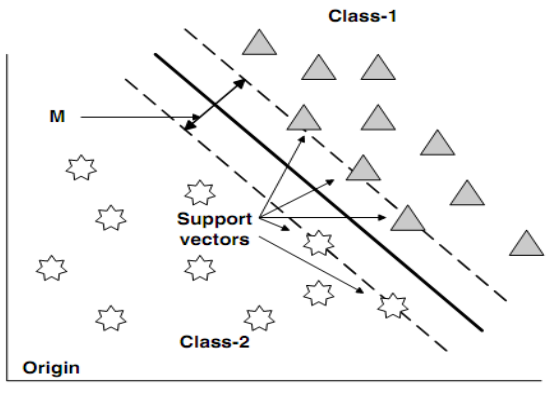


Fig 2 Separation of Two classes in SVM

d is degree of kernel function for polynomial and Fourier kernel functions. σ (sigma) is used to determine width of RBF and ERBF kernel functions.

Feature label:

Here each feature are extracted and then stored in a database and class label is assigned to them. The SVM is binary classifier. Features extracted will have a class label associated such as angry, happy, sad, neutral, boredom.

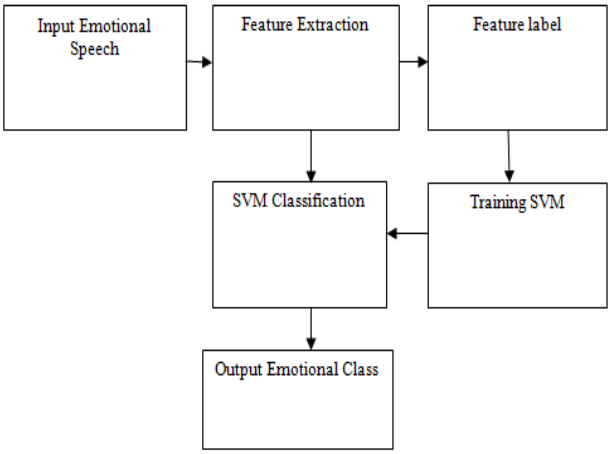


Fig 3 Block diagram of Speech recognition System

Feature Selection :

The ability of pattern recognition is relied on the differentiating ability of the features and identifying the most related subset from the original features. increasing the performance. The computational complexities can be minimized. Fig 3. illustrates the ideology.

Classification :

In recognition of human emotions, pattern recognition is an essential problem. SVM classifier is used to classify different output emotional states.

VI. EXPERIMENTAL RESULTS

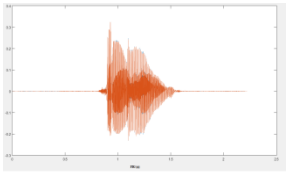
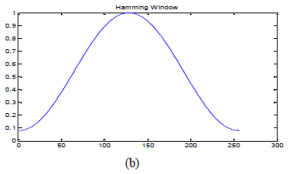
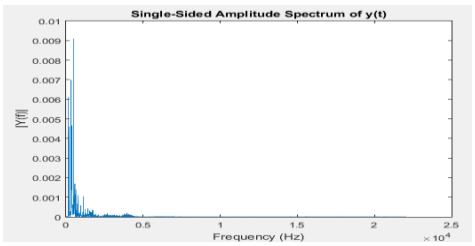


Fig (a)



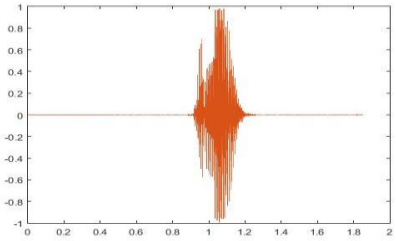
(b)



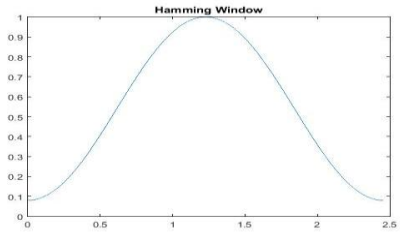
(c)

Fig (a) Hello signal of happy speech
Fig (b) Hamming window
Fig (c) Single sided amplitude spectrum.

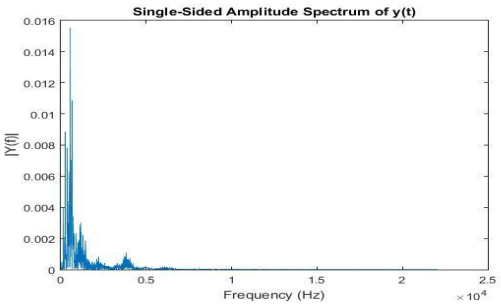
 ISSN: 2321-8169
2839 - 2842



(d)



(e)



(f)

Fig (d) Frame blocking of hello signal angry speech, Fig (e) Hamming window, Fig (f) Single sided amplitude spectrum of angry speech.

VII CONCLUSION

Whenever emotion changes, it's difficult to get accurate results. In the work carried out features such as frequency, and other features can be extracted whenever emotion changes. Few steps of MFCC algorithm are carried out on same words and sentences in four different emotional state and a database is created, which we use for classification. Here SVM is used to classify different emotions. The MFCC numbers are extracted from the speech data and the emotional state output of the recorded data is obtained in MATLAB command window.

ACKNOWLEDGEMENT

The Book *Digital Processing* of *Speech Signals.* by L.R. Rabiner and R.W. Schafer was very much helpful to start research in this field. I am very thankful to all the authors & researchers of the manuscript from where I got valuable information for completing this work.

REFERENCES

[1] Rabiner, L.R, Schafer, R.W, Digital Processing of Speech Signals, Pearson education, 1st Edition, 2004.

[2] Ferna ´ndez Pierna, J. A., Baeten, V.,Michotte Renier, A., Cogdill, R. P., & Dardenne, P. (2004). Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. Journal of Chemometrics, 18(7–8), 341–349.

[3] Gang, H., Jiandong, L., & Donghua, L. (2004). Study of modulation recognition based on HOCs and SVM. In Proceedings of the 59th Vehicular Technology Conference, VTC 2004-Spring. 17–19 May 2004 (Vol. 2, pp. 898–902).

[4] Guyon, I.,Weston, J., Barnhill, S., & Bapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422.

[5] S.Kim, P.Georgiou, S.Lee, S.Narayanan. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece, 2007.

[6] T.Bänziger, K.R.Scherer, ―The role of intonation in emotional expression‖, *Speech Communication*, Vol.46, 252-267, 2005.

[7] F.Yu, E.Chang, Y.Xu, H.Shum, ―Emotion detection from speech to enrich multimedia content‖, *Lecture Notes In Computer Science*, Vol.2195, 550-557, 2001

[8] Barra R., Montero J.M., Macias-Guarasa, DHaro, L.F., San-Segundo R., Cordoba R. Prosodic and segmental rubrics in emotion identification. Proc. ICASSP 2005, Philadelphia, PA, March 2005.

[9] Boersma P. Praat, a system for doing phonetics by computer. Glot International,vol. 5, no 9/10, pp. 341- 345, 2001.

[10] V. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.

[11] Dmitriy Fradkin and Ilya Muchnik, DIMACS Series in Discrete Mathematics and Theoretical Computer Science.