Speech Emotion Recognition using SVM algorithm

Vivek Kumar Dwivedi[1], Mr. Paurush Bhulania (Assistant Professor) [2]

Amity School of Engineering & Technology, Amity University, Uttar Pradesh, Noida Campus

Abstract- Speech Emotion Recognition is a recent research topic in the field of Human Computer Interaction. The objective of this paper is to use Support Vector Machine (SVM) classifier to classify seven different emotions happiness, anger, sadness, boredom, disgust, neutral, fear.The explored features include: pitch, mel-frequency spectrum coefficients (MFCC), Spectro-temporal features, formants and energy measurements. In this paper we use two different kernel:Linear (Homogeneous), Gaussian radial basis function kernel to get higher accuracy for emotion recognition in speech .Performance analysis is done by using the confusion matrix and the accuracy obtained is 95% on the basis of data set. Finally results for different combination of the features and on different databases are compared and we get SVM recognition accuracy is more than other algorithm**.**

Keyword- SVM, Speech emotion recognition, linear kernel, Gaussian radial basis function, MFCC

I. INTRODUCTION

Interpersonal communication is an interaction which involves the exchange of reciprocal ideas and emotions. [1]Gestures and sound are a way of conveying information in a human-to-human interaction. Speech, a special form of sound, is one of the fundamental ways of conveying information between people. . Words are not sole component of speech. Acoustic properties of speech also carry important affective features. Emotions exist in every part of the speech. Emotions in speech are transmitted from one communicator to another during an interaction.[2] As a result of exchange of emotions during an ongoing conversation, emotional state of a speaker may easily trigger an interlocutor emotional state resulting in a change in the speech style or tone. A lot of research works have been performed in identification of emotion through speech processing. Performance of speech recognition systems is usually evaluated in terms of accuracy and speed. Tin Lay New et. al proposed a text independent method of speech emotion classification, which makes use of short time log frequency power coefficients (LFPC) to represent the speech signals and a Hidden Markov Model (HMM) is used as a classifier. Average accuracy of 78% achieved & analyzed performance using LPCC and MFCC. Kamran Soltani et al studied the importance of the psychology and linguistics in spoken language man-machine interfaces. Average accuracy of 77% is achieved. Jana Tuckova et. al performed experimental analysis using parameters like fundamental frequency, formant frequency and statistical analysis was conducted for multilayer neural network (MLNN). [3]The average accuracy obtained using this technique is 75.93% for multiword sentences while that for one word sentences is 81.67%.In this paper we use Support Vector Machine (SVM) classifier to classify seven different emotion and accuracy obtained from basis of performance analysis is 95%.

In section 2 we explained speech emotion recognition system, acoustic feature used in speech emotion recognition, database used in this. In section 5 we describe the speech emotion algorithm SVM and section 6 we describe the simulation results of SVM and section 7 includes the conclusion and future scope of our work.[4]

Application of speech emotion recognition: Applications of emotion classification based on speech have already been used to facilitate interactions in our daily lives.

For example

1. In call centres apply emotion classification to prioritize impatient customers.

2. A warning system has been developed to detect if a driver exhibits anger or aggressive emotions.

3.For distance learning, to indentifying students emotion timely and taking appropriate action can improve the quality of teaching.

4. Emotion sensing has also been used in behaviour studies sound features have been widely explored in both the time domain and the frequency domain.

A. Speech emotion recognition system

Speech emotion recognition is nothing but an application of the pattern recognition system in which patterns of derived speech features such as Pitch, Energy, MFCC are mapped using classifier like ANN, SVM, HMM etc.

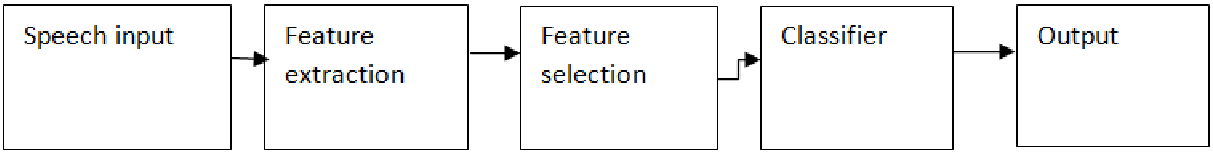


Fig.1: Block diagram of speech emotion recognition system

The system contains five major modules: speech input database, feature extraction, feature selection, classifier & recognized output as illustrated in figure – 1 above. Overall, the system is based on deep study of the production mechanism of speech signal, extracting some of features which contain information about speaker’s emotion & taking proper pattern recognition model to recognize states of emotion. Typically, a set of emotion having 300 emotional states [6]. Whenever, signal is passed to the feature extraction & selection process, the extracted speech features are selected in terms of emotion relevance. All over procedure revolves around the speech signal for extraction to the selection of speech features corresponding to emotions. Forward step is generation of database for training as well as testing of extracted speech features[7]. At the end, detection of emotions has been done using classifier with the usage of pattern recognition algorithm. The Speech emotion recognition is similar to the speaker recognition system but different types of approach to detect emotions make it secure & intelligent. The evaluation of the system is depending on naturalness of the input database.

B. Speech feature used in speech emotion recognition

In speech emotion recognition feature extraction is very important. Speech features can be divided into several categories. Speech features are divided into 4 categories; continuous, qualitative, spectral and TEO-based. Continuous features are pitch, energy and formants. Quantitative features are described as voice quality features which are harsh, tense and breathy voices[8]. Most popular acoustic features used in emotion recognition process are outlined below:

1) Pitch feature: Pitch is the fundamental frequency of the glottal excitation. Pitch depends on the tension of the vocal folds and subglottal air pressure. Pitch frequency is one of the widely used features in emotion from speech applications.

2) Spectral features: Mel-frequency cepstrum coefficients, linear predictive coding and log frequency power coefficients are the most popular spectral features. Mean and standard deviation of 13 Mel frequency cepstral coefficients (MFCC) are set as discriminating features in many studies.

3) Duration features :Mean and standard deviation of the duration of voiced and unvoiced segments, ratio between the duration of unvoiced and voiced segments are [4] duration features.

4) Energy features: Energy mean, standard deviation, maximum, 25% and 75% quantiles, and the inter quantile distance are the popular energy based features used in speech emotion recognition task.

5) Vocal tract features: Formants are a vocal tract feature. Each formant has its own bandwidth and center frequency. Slackened speech can be distinguished from an articulated speech using formant features. Other widely used feature is the energy of a certain frequency which corresponds to the critical bands of the human ear.

C. Database used in speech emotion recognition

Using a proper database is crucial in order to obtain a good accuracy. Low quality databases result in incorrect conclusions [5]. Generally, it is extremely difficult to produce a database representing the natural speech of a man or a woman in completely natural conversation. Many examples of humans talking exist, but very few of them illustrate speech in a natural environment. . In the latter case, some databases use corpora (i.e. large collections) of spontaneous speech, usually consisting of clips from live television, radio programs or call centers, with natural speech recorded in real-world situations. Capturing a faithful, detailed record of human emotion, as it appears in real life, is an incredibly challenging task. The assembly of databases (or datasets) has not traditionally been considered a high-profile or intellectually challenging area. Focus is explicitly placed in good quality recording and large samples that usually contain high arousal emotions (e.g. anger, sadness), while real human emotions are left relatively off-focus.

D. Emotion recognition algorithm

Support vector machine (SVM): The SVM is a high dimensional vector supervised learning method that is based on emotion assumptions. It predicts that the presence or absence of a specified feature of a class is not related to the presence or absence of all other features. It is very simple to program and execute it, its parameters are simple to assume, even on very large databases learning or training is very fast and effective and its accuracy is comparatively better in comparison to the other techniques[15].

SVM are also called as maximum margin classifiers. Firstly, the SVM theory is used to the solveibinary classification problems , rendering SVM ideal for the case under consideration, hence we .apply a psychologically-inspired binary cascade classification schema for identifying or to make a classifier . In this we use two different kernels.[16]

Let gi be the ith training vector.

1. Gaussian radial basis function kernel:

TSVM (gi , gj) = exp_((−_vi − vj) ^2)/σ2

where σ is a scaling factor; and tvsn=m.k;kbll;

2. Linear (Homogeneous):

TSVM (vi ., vj) = v Si vj

for various S values, for male subjects, female subjects, and both genders.

SVM with Gaussian radial basis function kernel are tested for various σ values with σ ∈ (0, 10], The best performance is obtained for σ = 1. For the case of SVM with Gaussian radial basis function kernel the two genders exhibit the same pattern: emotion recognition accuracy reaches at a fast rate its maximum for σ = 1,

Classification	Happiness	Neutral	Boredom	Sadness	Anger
Happiness	99.7	0	0	0	1.7
Neutral	2.9	90.5	1.5	0	0
Boredom	0	9.5	91.0	11.4	0
Sadness	0	0	6.0	88.6	0
Anger	4.5	0	0	0	90.1

Table 1. Confusion matrix (%)for the SVM with Gaussian radial basis function .ikernel (σ =1)

Speaker-independent emotion recognition accuracy of SVM with Gaussian radial basis function kernel for various σ values, for male subjects, female subjects, and both genders whereas it decreases strictly at a slower rate for greater σ values. The confusion matrix for σ = 1 is exhibited in Table 3 and the related accuracy equals 92.4%. Male emotion recognition accuracy is consistently greater than female emotion recognition, with exception of extreme low and high σ values. ii3The lower bound accuracy presented by SVM with Gaussian radial basis function kernel is 50.7% and can be attributed to poor parametrization. Linear SVM has the advantage of no need for parametrization.

Classification	Happiness	Neutral	Boredom	Sadness	Anger
Happiness	92.6	6	0	0	1.6
Neutral	0	98.9	1.3	0	0
Boredom	0	1.1	88.5	11.5	0
Sadness	0	0	8.9	88.5	0.8
Anger	1.6	0	0	0	87.7

Table 2 iiConfusion matrix (%) for the linear SVM

The corresponding confusion matrix is sketched in Table 2. It achievs an emotion recognition accuracy equal to 955%.

II. SIMULATION RESULTS AND DISCUSSION

Emotion Classification: In this project we recognise the mood of a user through their voice on the basis of their mood and classify into classifier:

6.1. Anger

6.2. Fear

6.3. Disgust

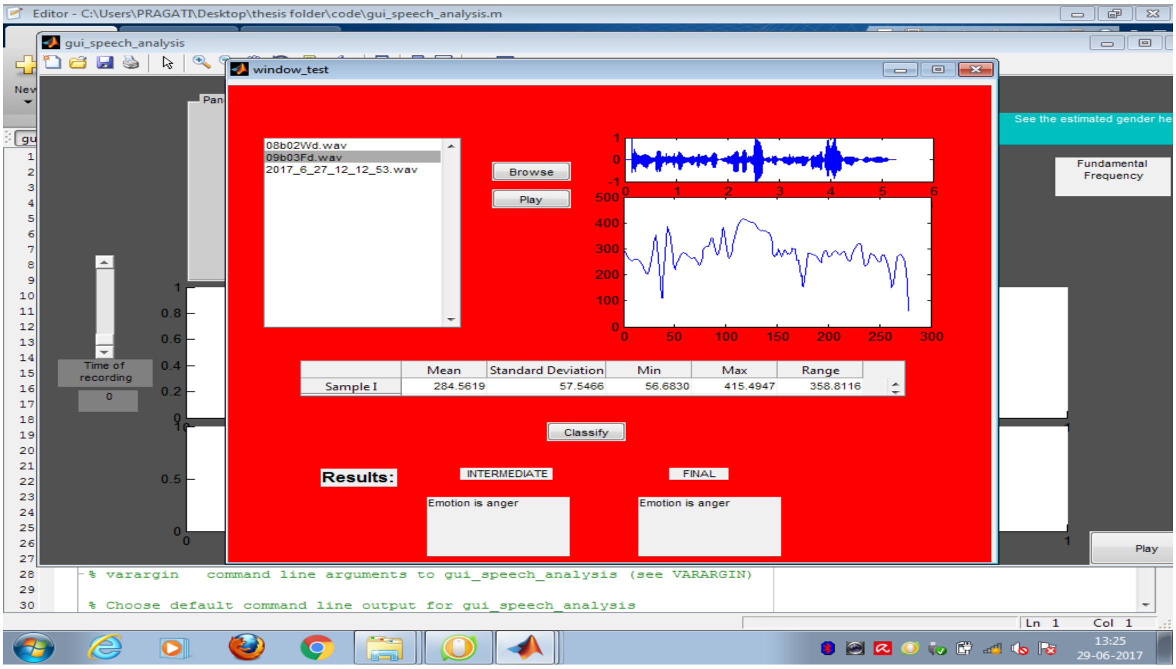
6.4.Happiness

6.5.Sadness

6.6. Neutral

6.7.Boredam

6.1.Anger emotion : Anger requires high energy to be expressed. Definition meaning of the anger is simple extreme displeasure [1]. In case of anger, aggression increases in which control parameter weakens. Anger is stated to have the highest energy and pitch level when compared with the emotions disgust, fear, joy and sadness. The widest observed pitch range and highest observed rate of pitch change are other findings about the emotion label anger when compared with other emotions[1]. Besides a faster speech rate is observed in angry speeches.



Editor - C:\Users\PRAGATI\Desktop\thesis folder\code\gui_speech_analysis.m

gui_speech_analysis

window_test

08b02Wd.wav
09b03Fd.wav
2017_6_27_12_12_53.wav

	Mean	Standard Deviation	Min	Max	Range
Sample I	284.5619	57.5466	56.6830	415.4947	358.8116

Results:

INTERMEDIATE: Emotion is anger

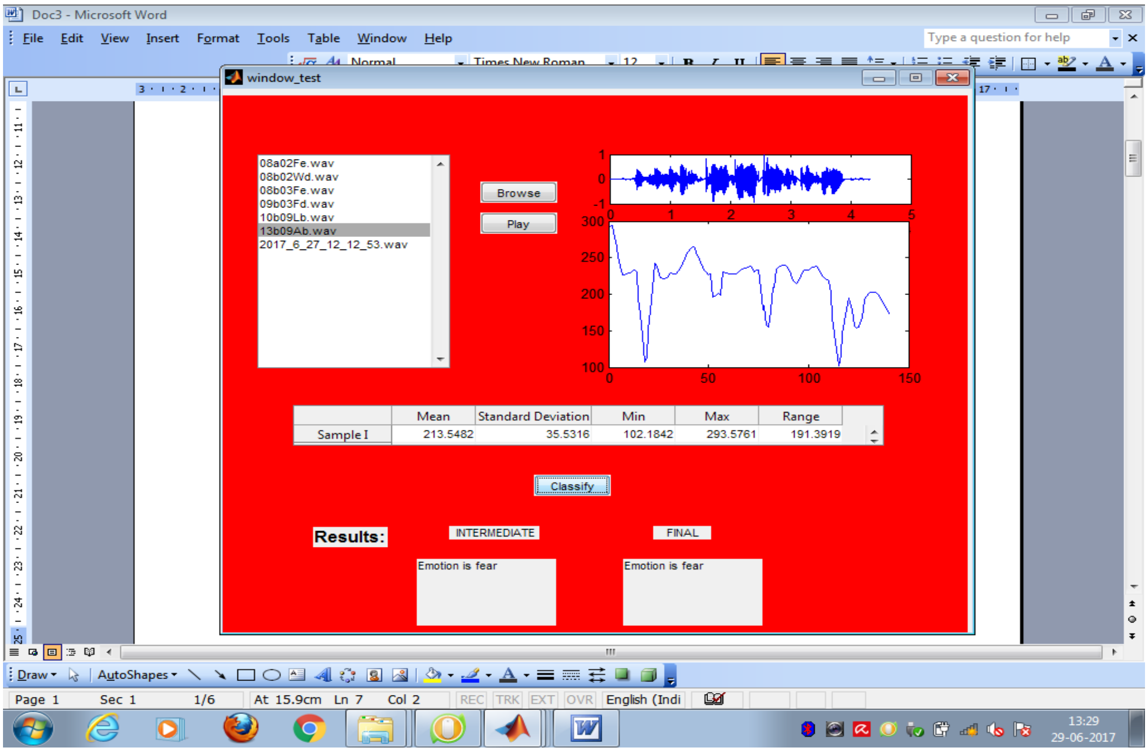
FINAL: Emotion is anger

```
% varargin   command line arguments to gui_speech_analysis (see VARARGIN)

% Choose default command line output for gui_speech_analysis
```

Fig.2: anger emotion

6.2 Fear emotion: In emotional dimension, fear has similar features to anger. High pitch level and raised intensity level are correlated with fear [2]. It is stated that fear has a wide pitch range. Highest speech rate is observed in fear speeches [1]. The pitch contour trend separates fear from joy. Although the pitch contour of fear resembles the sadness having an almost downwards slope, emotion of joy have a rising slope [2].



Doc3 - Microsoft Word

window_test

08a02Fe.wav
08b02Wd.wav
08b03Fe.wav
09b03Fd.wav
10b09Lb.wav
13b09Ab.wav
2017_6_27_12_12_53.wav

	Mean	Standard Deviation	Min	Max	Range
Sample I	213.5482	35.5316	102.1842	293.5761	191.3919

Results:

INTERMEDIATE: Emotion is fear

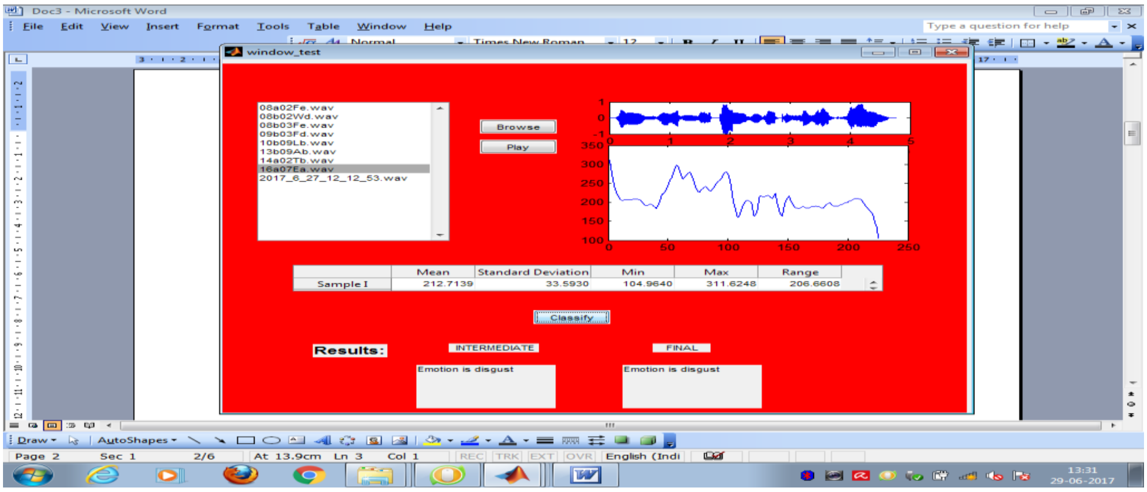
FINAL: Emotion is fear

File Edit View Insert Format Tools Table Window Help
Type a question for help
Browse
Play
Classify
Draw AutoShapes
Page 1 Sec 1 1/6 At 15.9cm Ln 7 Col 2 REC TRK EXT OVR English (Indi
13:29
29-06-2017

Fig.3: Fear emotion

6.3 Disgust emotion:

In [2], low mean pitch level, a low intensity level, and a slower speech rate is observed when disgust is compared with the neutral state. Disgust is stated the lowest observed speech rate and increased pause length [1].



08a02Fe.wav
08b02Wd.wav
08b03Fe.wav
09b03Fd.wav
10b09Lb.wav
13b09Ab.wav
14a02Tb.wav
16a07Ea.wav
2017_6_27_12_12_53.wav

	Mean	Standard Deviation	Min	Max	Range
Sample I	212.7139	33.5930	104.9640	311.6248	206.6608

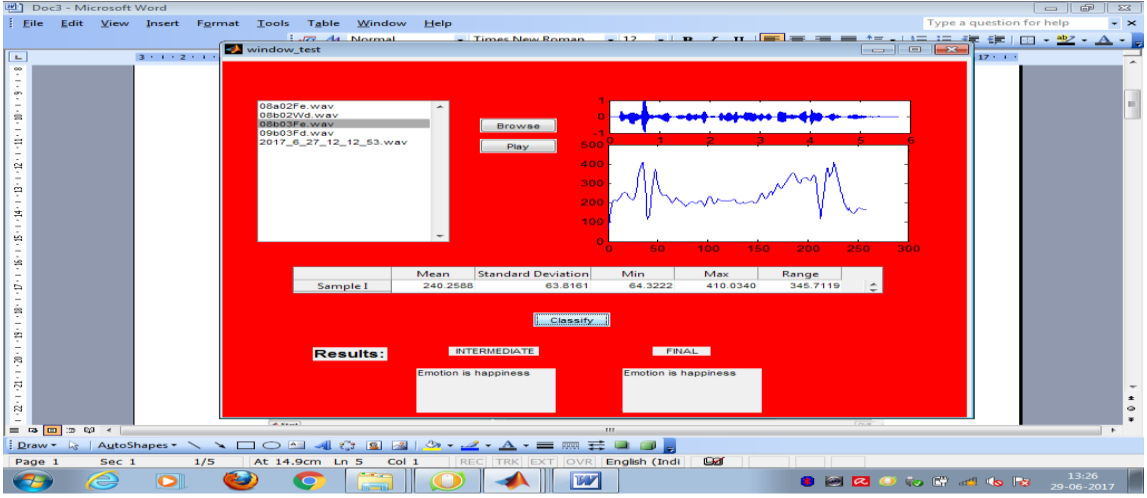
Results: INTERMEDIATE FINAL

Emotion is disgust

Emotion is disgust

Fig.4: Disgust Emotion

6.4 Happiness emotion: Happiness exhibit a pattern with a high activation energy, and positive valence. Strength of the happiness emotion may vary. In the emotional state happiness or joy, pitch mean, range and variance increases . In it is stated that fundamental and formant frequencies increases in case of smile. Moreover, amplitude and duration also increase for some speakers.



08a02Fe.wav
08b02Wd.wav
08b03Fe.wav
09b03Fd.wav
2017_6_27_12_12_53.wav

	Mean	Standard Deviation	Min	Max	Range
Sample I	240.2588	63.8161	64.3222	410.0340	345.7119

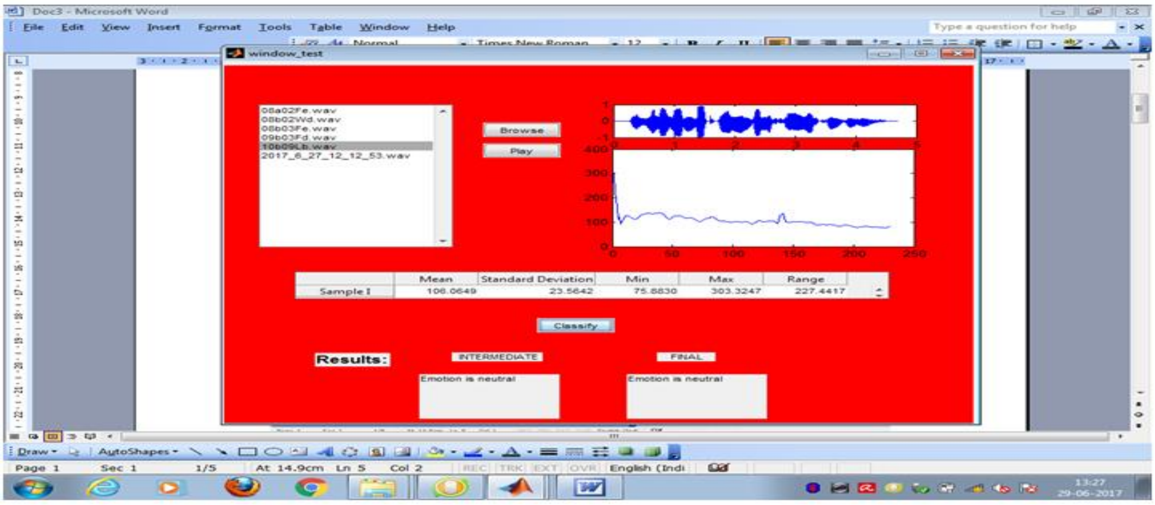
Results: INTERMEDIATE FINAL

Emotion is happiness

Emotion is happiness

Fig.5: Happiness Emotion

6.5 Neutral emotion:



08a02Fe.wav
08b02Wd.wav
08b03Fe.wav
09b03Fd.wav
10b09Lb.wav
2017_6_27_12_12_53.wav

	Mean	Standard Deviation	Min	Max	Range
Sample I	106.0649	23.5642	75.8830	303.3247	227.4417

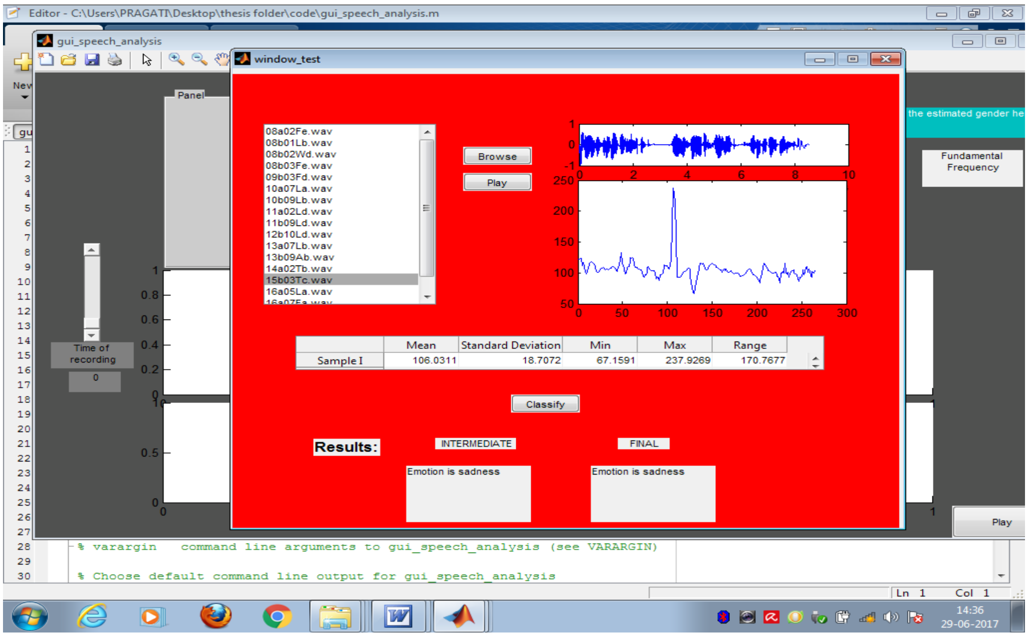
Results: INTERMEDIATE FINAL

Emotion is neutral

Emotion is neutral

Fig.6: Neutral Emotion

6.6 Sadness emotion: In emotional dimension, sadness requires very low energy. In addition, valence degree is negative. Sadness exhibits a pattern that is normal or lower than normal average pitch, a narrow pitch range and slow tempo [1]. Speech rate of a sad person is lower than the neutral one [2].



	Mean	Standard Deviation	Min	Max	Range
Sample I	106.0311	18.7072	67.1591	237.9269	170.7677

Results: INTERMEDIATE FINAL

Emotion is sadness

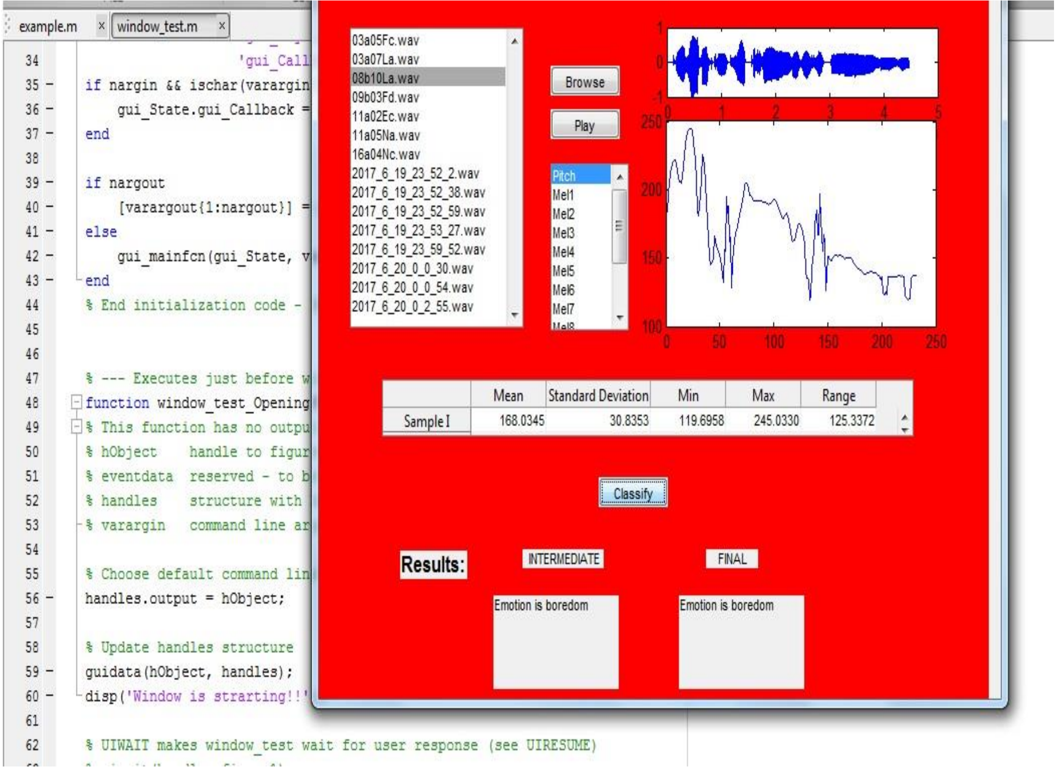
Emotion is sadness

```
% varargin   command line arguments to gui_speech_analysis (see VARARGIN)

% Choose default command line output for gui_speech_analysis
```

Fig.7: Sadness .Emotion

6.7 Boredom emotion: Boredom is a negative emotion with negative valence and low activation level same as sad. A lowered mean pitch and a narrow pitch range with a slow speech rate are defined as the properties of a bored expression [3].



```
                     'gui_Call
if nargin && ischar(varargin
    gui_State.gui_Callback =
end

if nargout
    [varargout{1:nargout}] =
else
    gui_mainfcn(gui_State, v
end
% End initialization code -


% --- Executes just before w
function window_test_Opening
% This function has no outpu
% hObject    handle to figur
% eventdata  reserved - to b
% handles    structure with
% varargin   command line ar

% Choose default command lin
handles.output = hObject;

% Update handles structure
guidata(hObject, handles);
disp('Window is strarting!!'

% UIWAIT makes window_test wait for user response (see UIRESUME)
```

	Mean	Standard Deviation	Min	Max	Range
Sample I	168.0345	30.8353	119.6958	245.0330	125.3372

Results: INTERMEDIATE FINAL

Emotion is boredom

Emotion is boredom

Fig.8: Boredom emotion

Emotion classificationof speechsignal

Sample	**Mean**	**Standard .ii3deviation**	**Min**	**Max**	**Range**	**Emotion**	**Accuracy**
08a02fe	204.5379	44.9043	77.7242	291.9683	293.5781	Happiness	Yes
08b01lb	187.5497	42.8591	78.7224	290.5677	215.4478	Happiness	Yes
02b09ab	213.5482	35.5316	102.4282	293.5781	191.3919	Fear	Yes
100a51d	106.4724	18.9341	66.2311	165.9720	165.9720	Disgust	No
16a07ea	212.7139	33.5930	14.9649	311.6248	206.6608	Disgust	Yes
03b01lb	113.5912	35.5358	66.5358	470.5882	403.779	Boredam	Yes
13a027a	132.6433	31.4061	80.8352	214.8494	134.0142	sadness	Yes
11a05Na	109.0524	14.5833	52.4239	179.4007	126.9168	Neutral	Yes
03a04wc	227.9200	48.6458	77.5875	300.1133	222.5259	Anger	Yes

Table-3On the basis of table-3 we can easily see that the recognition ratio is 95% on the basis of the dataset and ii on the basis of classifier Hence, recognition accuracy is more as comparison to other algorithm.

III. CONCLUSION AND FUTURE WORK

a. CONCLUSION

As technology evolves, interest in human like machines increases. Technological devices are spreading and user satisfaction increases importance. A natural interface which responds according to user needs has become possible with affective computing. The key issue of affective computing is emotions. Any research which is related with detection, recognition or generating an emotion is affective computing. User satisfaction or un-satisfaction could be detected with any emotion recognition system. Besides detection of user satisfaction, such systems could be used to detect anger or frustration. In such cases, user could be restrained like driving a car. In emotion detection tasks, speech or face emotion detections are the most popular ones. Easy access to face or speech data made them very popular. Speech carries a rich set of data. In human to human communication, via speech information is conveyed. Acoustic part of speech carries important info about emotions.MFCC are used for the feature extraction.Algorithm with the SVM's overall performance is tested. Finally results for different combination of the features and on different databases are compared and we get SVM recognition accuracy is more than other algorithm. Accuracy obtained from SVM is 95% basis of data set.

b. Future work

Due to very less knowledge about this field there are very few researches going on in the area of speech processing. But a large amount of work can be done by processing the spectral features effectively to recognize. Higher accuracy can be obtained using the combination of more features. Increasing the sigma value from the default value one, substantial results may be obtained.

IV. REFERENCE

[1]. Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. Journal of the Acoustical Society of America, 93, 1097–1108.

[2]. Ververidis, D., & Koropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication, 48, 1162–1181.

[3]. Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. ISCA Workshop on Speech and Emotion, 4, 151–156.

[4]. Wu, D., Parsons, T. D., & Narayanan, S. S. (2010). Acoustic feature analysis in speech emotion primitives estimation. Interspeech.

[5]. Ayadi, E., Moataz, Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition features, classification schemes and databases. Pattern Recognition, 44, 572–587

[6]. B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", IEEE international conference on acoustics, speech, and signal processing, vol.1, pp. I-577-80, 2004.

[7]. Sirko Molau, Michael Pitz, Ralf Schl¨uter, and Hermann Ney. Computing mel-frequency cepstral coefficients on the power spectrum. IEEE Transaction, 2011.

[8]. Mandar Gilke , Pramod Kachare , Rohit Kothalikar , Varun Pius Rodrigues and Madhavi Pednekar. MFCCbased vocal emotion recognition using ANN, International Conference on Electronics Engineering and Informatics, 150-154, 2012.

[9]. Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese and Andrea Sciarrone. Gender driven emotion recognition through speech signals for ambient intelligence applications. IEEE Transactions on Emerging Topics in Computing, vol. 1, no. 2, 244-257, December 2013.

[10]. Kamran Soltani and Raja Noor Ainon. Speech emotion detection based on neural networks. In 9th International Symposium on

Signal Processing and its Applications, 1 4244-0779-6/07, IEEE, 2007.8

[11]. Thapanee Seehapoch & Sartra wongthanavasu(2013), “Speech Emotion Recognition using Support Vector Machine”

[12]. Xiao, Z., E. Dellandrea, Dou W.,Chen L., "Features extraction and selection for emotional speech classification". 2005 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp.411- 416, Sept 2005.\

[13]. T.-L. Pao, Y.-T. Chen, J.-H. Yeh, P.-J. Li, "Mandarin emotional speech recognition based on SVM and NN", Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), vol. 1, pp. 1096-1100,September 2006.

[14]. Lin Y, Wei G, "Speech emotion recognition based on HMM and SVM". Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol.8, pp. 4898-4901. Agu 2005.

[15]. Jana Tuckova and Martin Sramka. Emotional speech analysis using Artificial Neural Networks. Proceedings of the International Multiconference on Computer Science and Information Technology, 141-147, 2010

[16]. H. Fletcher, Speech and Hearing in Communication. The Bell Telephone Laboratories Series, D. Van Nostrand Company, Inc.