Mood Estimation using Timbre Analysis and Pitch Detection

Akhilesh Rao, Hiteshi Acharya, Rohan Dotihal
Department of Electrical Engineering, New York University, New York.

***Abstract*— Human speech contains lot of information. In the case of strings of words spoken without considering the fashion in which they are spoken, one might miss out some of important aspects of utterance. Using human computer interaction one can exploit the additional information encoded in the speech to determine the emotional state of the speaker and also his personality traits. This paper explores the estimation of mood using different features extraction from the language and acoustic information contained in the speech.**

Keywords—mood estimation, feature extraction, human computer interaction

I. INTRODUCTION

The mood detection from the human speech is not a relatively new field of research, but has plethora of applications in our day to day lives. The human-computer or human-human interaction systems could provide with an improvised response by adapting to the mood. The various research conducted in the field of psychology and linguistics for understanding and modelling human mood from the speech has already attracted lot of attention in engineering fraternity. There has increased a need to not only for knowing what has been spoken but also the way in which it is spoken. Researchers in psychology and neuroscience have proved that the human mood plays a vital role in decision making and that it plays major role in the rational actions of human being. The emotional states conveyed in one's speech do not alter any linguistic content but carry important information about that speaker about his/her desires, intent, and reaction to the outer world. The identical utterances can be expressed in different emotion and conveys distinct meanings.

Automatic mood detection from human speech has lot of applications in various fields such as learning, security, entertainment and medicine. Considering the immense use of internet and rapid development of virtual learning applications, automatic tutoring systems can adjust the content and way of teaching based on the response using mood detection. The audio surveillance, mood detection will help detecting abnormal gestures for example stress, fear, nervousness, etc. and help identify suspicious human subjects. Also detecting moods of the patient would help in diagnosis and psychosis monitoring. In commercial applications such as customer services and call centers, mood detection could play a great role in adjusting the automatic responses based on the estimated mood by distinguishing between satisfaction and dissatisfaction conveyed through the speech. Mood detection could bring on more fun features and improve the adaptive interaction between human and machine.

There are variety of temporal and spectral features that can be extracted from a human speech. We have used statistics related to pitch, Mel Frequency Cepstral Coefficients (MFCCs) as inputs to classification algorithm. The mood detection accuracy helps to understand which features carry the most emotional information and the reason for it. It further helps us develop criteria to distinguish emotions together into classes. With the help of these techniques we can achieve high mood detection accuracy.

II. DATABASE

First step is creating a vast database by recording the various emotions. In this project the Surrey Audio-Visual Expressed Emotion (SAVEE) is been used. This database was recorded as a prerequisite to develop an automatic mood detection system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from standard TIMIT corpus and phonetically balanced for each emotion. The data was recorded in visual media laboratory with high quality audio visual equipment processed and labeled.

III. THEORETICAL BACKGROUND

The next step which is the most important step for each mood detection is the step where two major processes take place. First the feature extraction and second the extracted features that determine the mood the database used to create training sets and training sets. In this project, features were extracted from each utterance. The features like Energy Entropy block, Short Time Energy, Zero Crossing Rate, Spectral Roll Off, Spectral Centroid, and Spectral Flux have been extracted in MATLAB using functions. After all the features have been extracted the training set procedure takes place. Finally the last step of this mood detection is the classification, which is the detection of the emotion we have given as an input.

A. Zero Crossing Rate

In the case of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero. Speech signals are broadband signals and interpretation of average zero-crossing rate is therefore much less precise. However, rough estimates of spectral properties can be obtained using a representation based on the short time average zero-crossing rate. Since high frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero-crossing rate and energy distribution with frequency. A reasonable generalization is that if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced.

B. Short Time Energy

The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segments is much lower than the amplitude of voiced segments. The energy of the speech signal provides a representation that reflects these amplitude variations. The choice of the window determines the nature of the short-time energy representation. In our model, we have used Hamming window. The hamming window gives much greater attenuation outside the band pass than the comparable rectangular window. The attenuation of this window is independent of the window duration. Increasing the length, decreases the bandwidth. If N is too small, Energy will fluctuate very rapidly depending on the exact details of the waveform. If length is too large, Energy will change very slowly and thus will not adequately reflect the changing properties of the speech signal.

C. Spectral Roll Off

Spectral roll-off (SR) is a feature which is defined as a frequency under which 95% of signal's spectral energy is accumulated. It characterizes the inclination of the signal's spectrum. The SR feature has higher values for studio speech than for telephone speech, because telephone speech is filtered above 3400 Hz and has therefore lower values of SR. But the telephone speech signal is also filtered in the frequency band between 0-300 Hz. In this region speech normally has a lot of energy present. The result is that the values of the SR feature are therefore higher and the original SR feature is not a good discriminator between wide-band and narrow-band speech.

D. Spectral Centroid

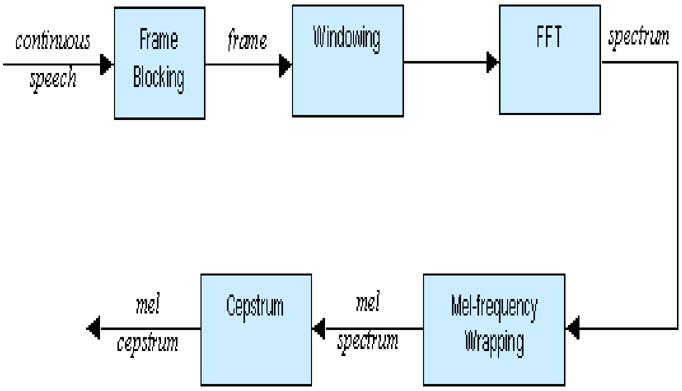
The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of mass" of the spectrum is. It has a robust connection with the impression of brightness of a sound. It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights. Spectral centroid also refer to the median of the spectrum. This is a different statistic, the difference being essentially the same as the difference between the unweighted median and mean statistics. Since both are measures of central tendency, in some situations they will exhibit some similarity of behavior. But since typical audio spectra are not normally distributed, the two measures will often give strongly different values. It is widely used in digital audio and music processing as an automatic measure of musical timbre.

E. Spectral Flux

Spectral Flux is computed as the magnitude of the difference between adjacent spectra. I t is generally higher for music signal than for speech signal as the speech signal changes more rapidly than music signal.

F. MFCC

Known as the Mel-Frequency Cepstrum, it emulates the human ear in terms of hearing. The log power spectrum of here is decorrelated using the DCT matrix to give the cepstral coefficients.



IV. Approach

A. Problems and constraints

The various problems and constraints affecting the accurate mood detection are given below. Smoking & Drinking Habits speed up the process of developing "vocal nodes" that usually has the effect of lowering ones speech fundamental frequency range. The effects of autocorrelation function which we have used for mood detection can be corrected by considering jitter and shimmer after taking into consideration to an extent.

During puberty [8], in case of males, the lower limit is very sharp and can be attributed to the change of pitch frequency during puberty. But in case of females, the lower limit is gradual and caused by the slowly changing dimensions of the vocal tract length only. Generally for both genders, the upper limit is very gradual and can possibly be attributed to changes in the glottis area and the internal control loops of the human articulatory system.

We have used clean signal. But this might not be the case. The signal will be noisy up to an extent depending on the environment. We have to perform morphological component analysis like Basis Pursuit Denoising, Iterative Soft Thresholding Algorithm to remove noise. But this can cause the error of over smoothing or distorting, if the weighted regularization isn't chosen correctly and could further misclassify the emotion for the signal. Additionally, we have assumed all vocal inputs are provided by healthy adult males.

An important factor we've constrained is sarcasm. Although sarcasm is shown subtly via speech, it includes certain baseline functions that makes it highly detectable. [10] effective way of modeling and applying prosodic contours to the task of automatic sarcasm detection. This approach focuses on using k-means clustering to determine common patterns of pitch and intensity contours. Uses Legendre polynomial expansions to represent prosodic contours. [11] to [13] give a good clue about this topic

B. Method

For the purpose of our analysis, we have mainly used the MFCC's along with other timbre detection features. There have been other methods devised that use Support Vector Machines, Neural Networks, decision tree, HMM and GMM etc for mood estimation, but all have provide less accurate estimation results using prosodic speech inputs. The HMM and GMM systems make use of a generative model for all speaker models. In addition GMM models require a large number of training samples to train the model and hence isn't feasible in real time applications as compared to MFCC. This will result in over-fitting and can misclassify the emotion of the speaker. More data based on SVM vs HMM and NN models can be referenced by [14] and [15]. Implementing the other features for the signal such as the zero crossing rate, spectral flux, spectral roll off, standard deviation, short time energy and the pitch detection using auto correlation help with better tonal classification of the sample allowing for more accurate mood estimation.

V. IMPLEMENTATION

We have assumed all averaged aged male speakers as the training and testing samples for this experiment. Assuming female and children of different ages introduces a change in the fundamental frequency, and other subtleties in the audio sample. Hence, while training data for the same emotion, we may introduce changes in the zero crossing rate, autocorrelation etc, which results in varied data sets specific to that emotion. In order to include these varied training data and still maintain accuracy for mood estimation, we can introduce jitter and shimmer as additional feature constraints to improve mood estimation for test data.

We train using 424 training samples overall. We train the system to detect 7 emotions, namely sadness, disgust, anger, fear, happy, neutral, surprise. For each signal in the training set, we compute autocorrelation and perform the timbre analysis to obtain all its tonal qualities for that particular emotion associated with it. After the training sets have been computed, we run the test samples to obtain its autocorrelation and tonal data through timbre analysis. We compare this data with all the test sample data. The output which gives the maximum autocorrelation match with the test matrix and the minimum difference for tonal qualities is matched to that particular emotion.

Now we cannot be certain the audio sample we are testing completely represents one emotion, hence we compute the probabilities of the emotions it represents.

VI. EXPERIMENTAL PART

We pass the test sample to the function compare_train_test. We then calculate the spectral flux, spectral centroid, spectral roll off, short time energy, zero crossing , standard deviation and auto correlation denoted by C,R,Z,E,F,SD,AT in MATLAB.

For every speaker and for every set of emotions, we find the overall accuracy compared to the test sample. We have 7 sets of emotions for 4 speakers

Anger set has 13 samples
Disgust set has 13 samples
Fear set has 13 samples
Happiness set has 13 samples
Neutral set has 13 samples
Sadness set has 13 samples
Surprise set has 13 samples

Consider 13 samples of anger for the first speaker. Then take the average of spectral flux, spectral centroid, spectral roll off, short time energy, zero crossing, standard deviation and auto correlation , individually. Also, perform the same for all emotions sets.

Example:
anger
C1(1,:)=array of all spectral centroid of anger samples
R1(1,:)=array of all spectral roll of anger samples
Z1(1,:)=array of all zero crossings of anger sample
E1(1,:)=array of all short time energy of anger samples
F1(1,:)=array of all spectral flux of anger samples
SD1(1,:)=array of all standard deviations of anger samples
AT1(1,:)=array of all autocorrelation of anger samples
overall accuracy(1)

Similarly,
disgust
C1(2,:)=array of all spectral centroid of anger samples
R1(2,:)=array of all spectral roll of anger samples

Z1(2,:)=array of all zero crossings of anger sample
E1(2,:)=array of all short time energy of anger samples
F1(2,:)=array of all spectral flux of anger samples
SD1(2,:)=array of all standard deviations of anger samples
AT1(2,:)=array of all autocorrelation of anger samples
overall accuracy(2)

fear
C1(3,:)=array of all spectral centroid of anger samples
R1(3,:)=array of all spectral roll of anger samples
Z1(3,:)=array of all zero crossings of anger sample
E1(3,:)=array of all short time energy of anger samples
F1(3,:)=array of all spectral flux of anger samples
SD1(3,:)=array of all standard deviations of anger samples
AT1(3,:)=array of all autocorrelation of anger samples
overall accuracy(3)

happiness
C1(4,:)=array of all spectral centroid of anger samples
R1(4,:)=array of all spectral roll of anger samples
Z1(4,:)=array of all zero crossings of anger sample
E1(4,:)=array of all short time energy of anger samples
F1(4,:)=array of all spectral flux of anger samples
SD1(4,:)=array of all standard deviations of anger samples
AT1(4,:)=array of all autocorrelation of anger samples
overall accuracy(4)

neutral
C1(5,:)=array of all spectral centroid of anger samples
R1(5,:)=array of all spectral roll of anger samples
Z1(5,:)=array of all zero crossings of anger sample
E1(5,:)=array of all short time energy of anger samples
F1(5,:)=array of all spectral flux of anger samples
SD1(5,:)=array of all standard deviations of anger samples
AT1(5,:)=array of all autocorrelation of anger samples
overall accuracy(5)

sadness
C1(6,:)=array of all spectral centroid of anger samples
R1(6,:)=array of all spectral roll of anger samples
Z1(6,:)=array of all zero crossings of anger sample
E1(6,:)=array of all short time energy of anger samples
F1(6,:)=array of all spectral flux of anger samples
SD1(6,:)=array of all standard deviations of anger samples
AT1(6,:)=array of all autocorrelation of anger samples
overall accuracy(6)

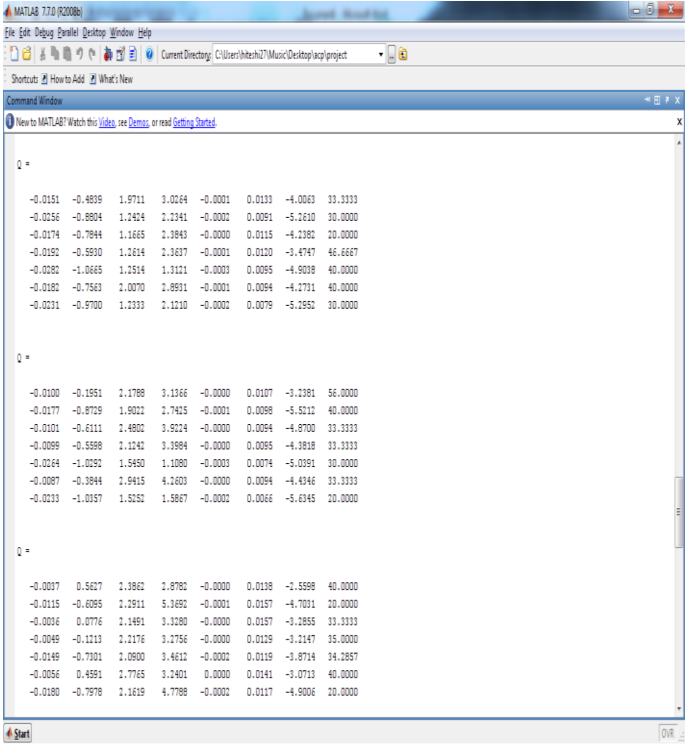
surprise
C1(7,:)=array of all spectral centroid of anger samples
R1(7,:)=array of all spectral roll of anger samples
Z1(7,:)=array of all zero crossings of anger sample
E1(7,:)=array of all short time energy of anger samples
F1(7,:)=array of all spectral flux of anger samples
SD1(7,:)=array of all standard deviations of anger samples
AT1(7,:)=array of all autocorrelation of anger samples
overall accuracy(7)
Now we calculate the Nearest Neighbor for every emotion

```
for 1:7
Q (1)=avgC1-C
Q2(2)=avgR1-R1
Q(3)=avgZ1-Z
Q(4)=avgE1-E
Q(5)=avgF1-F
Q(6)=avgSD1-SD
Q(7)=maxAT1-AT
Q(8)=overallaccuracy
end
```

We perform the above computation for every speaker. We then sum up the Q matrices of all speakers using:

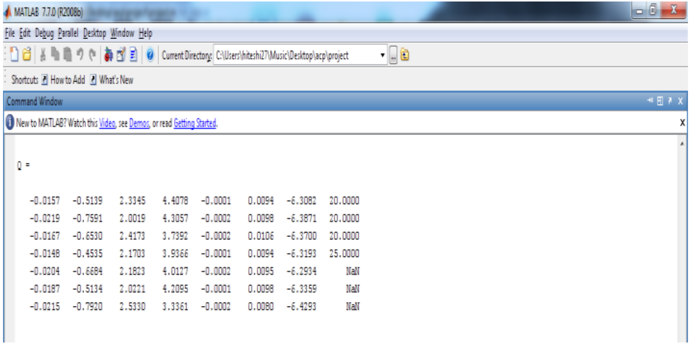
QQ=QQ+Q

For a 'fear' sample


MATLAB 7.7.0 (R2008b)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\hiteshi27\Music\Desktop\acp\project
Shortcuts How to Add What's New
Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Start
OVR

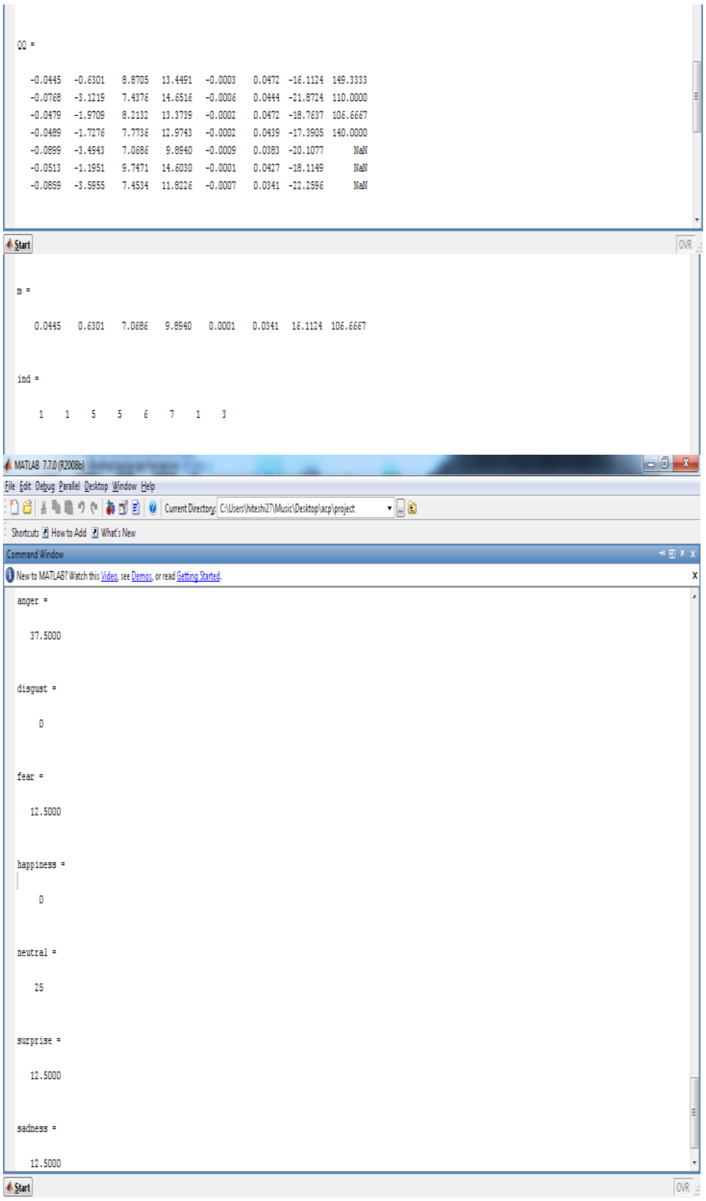
```
Q =

   -0.0151   -0.4839    1.9711    3.0264   -0.0001    0.0133   -4.0063   33.3333
   -0.0256   -0.8804    1.2424    2.2341   -0.0002    0.0091   -5.2610   30.0000
   -0.0174   -0.7844    1.1665    2.3843   -0.0000    0.0115   -4.2382   20.0000
   -0.0192   -0.5930    1.2614    2.3637   -0.0001    0.0120   -3.4747   46.6667
   -0.0282   -1.0665    1.2514    1.3121   -0.0003    0.0095   -4.9038   40.0000
   -0.0182   -0.7563    2.0070    2.8931   -0.0001    0.0094   -4.2731   40.0000
   -0.0231   -0.9700    1.2333    2.1210   -0.0002    0.0079   -5.2952   30.0000


Q =

   -0.0100   -0.1951    2.1788    3.1366   -0.0000    0.0107   -3.2381   56.0000
   -0.0177   -0.8729    1.9022    2.7425   -0.0001    0.0098   -5.5212   40.0000
   -0.0101   -0.6111    2.4802    3.9224   -0.0000    0.0094   -4.8700   33.3333
   -0.0099   -0.5598    2.1242    3.3984   -0.0000    0.0095   -4.3818   33.3333
   -0.0264   -1.0292    1.5450    1.1080   -0.0003    0.0074   -5.0391   30.0000
   -0.0087   -0.3844    2.9415    4.2603   -0.0000    0.0094   -4.4346   33.3333
   -0.0233   -1.0357    1.5252    1.5867   -0.0002    0.0066   -5.6345   20.0000


Q =

   -0.0037    0.5627    2.3862    2.8782   -0.0000    0.0138   -2.5598   40.0000
   -0.0115   -0.6095    2.2911    5.3692   -0.0001    0.0157   -4.7031   20.0000
   -0.0036    0.0776    2.1491    3.3280   -0.0000    0.0157   -3.2855   33.3333
   -0.0049   -0.1213    2.2176    3.2756   -0.0000    0.0129   -3.2147   35.0000
   -0.0149   -0.7301    2.0900    3.4612   -0.0002    0.0119   -3.8714   34.2857
   -0.0056    0.4591    2.7765    3.2401    0.0000    0.0141   -3.0713   40.0000
   -0.0180   -0.7978    2.1619    4.7788   -0.0002    0.0117   -4.9006   20.0000
```

Assign the emotion corresponding to maximum value of each feature in QQ


MATLAB 7.7.0 (R2008b)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\hiteshi27\Music\Desktop\acp\project
Shortcuts How to Add What's New
Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
Q =

   -0.0157   -0.5139    2.3345    4.4078   -0.0001    0.0094   -6.3082   20.0000
   -0.0219   -0.7591    2.0019    4.3057   -0.0002    0.0098   -6.3871   20.0000
   -0.0167   -0.6530    2.4173    3.7392   -0.0002    0.0106   -6.3700   20.0000
   -0.0148   -0.4535    2.1703    3.9366   -0.0001    0.0094   -6.3193   25.0000
   -0.0204   -0.6684    2.1823    4.0127   -0.0002    0.0095   -6.2934       NaN
   -0.0187   -0.5134    2.0221    4.2095   -0.0001    0.0098   -6.3359       NaN
   -0.0215   -0.7920    2.5330    3.3361   -0.0002    0.0080   -6.4293       NaN
```

Some example sample outputs



```
QQ =

  -0.0445   -0.6301    8.8705   13.4491   -0.0003    0.0472  -16.1124  149.3333
  -0.0768   -3.1219    7.4376   14.6516   -0.0006    0.0444  -21.8724  110.0000
  -0.0479   -1.9709    8.2132   13.3739   -0.0002    0.0472  -18.7637  106.6667
  -0.0489   -1.7276    7.7736   12.9743   -0.0002    0.0439  -17.3905  140.0000
  -0.0899   -3.4943    7.0686    9.8940   -0.0009    0.0383  -20.1077       NaN
  -0.0513   -1.1951    9.7471   14.6030   -0.0001    0.0427  -18.1149       NaN
  -0.0859   -3.5955    7.4534   11.8226   -0.0007    0.0341  -22.2596       NaN
```

```
m =

    0.0445    0.6301    7.0686    9.8940    0.0001    0.0341   16.1124  106.6667


ind =

     1     1     5     5     6     7     1     3
```

MATLAB 7.7.0 (R2008b)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\hiteshi27\Music\Desktop\acp\project
Shortcuts How to Add What's New
Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
anger =

   37.5000


disgust =

     0


fear =

   12.5000


happiness =

     0


neutral =

    25


surprise =

   12.5000


sadness =

   12.5000
```

Audio file	Spectral centroid	Spectral flux	Zero crossing	Short time energy	Spectral roll-off	Auto correlatin	Zero crossing	MFCC overall accuracy
Sa15,speaker1 (fear)	5	7	5	5	5	7	7	7
A15,speaker3 (anger)	1	1	5	5	6	7	1	1
F15,speaker1 (fear)	2	2	5	5	1	7	1	2

Emotion%	SA15	A15	F15
1 anger	0	50	37.5
2 disgust	0	0	0
3 fear	0	0	12.5
4 happiness	0	0	0
5 neutral	50	25	25
6 surprise	0	12.5	12.5
7 sadness	50	12.5	12.5

VII. Conclusion

The mood detection still remains a challenging problem primarily due to the inherent ambiguities in the human emotions traits. The research in this area is not as mature, but rapid progress is being made. The performance of automated systems for mood recognition using a wide range of annotated and content-based features and multimodal feature combinations have advanced significantly. We have developed a novel mood recognition system based on the processing of physiological signals. This system shows a recognition ratio much higher than chance probability, for three and four mood categories, respectively, when applied to the signal databases obtained. The advantages of our system include the reduction of required monitoring time, further extension of applicability to multiple users and the minimum amount of user inconvenience. . The system consists of characteristic feature extraction and pattern classification stages.

VIII. Future work

Efficiently fusing recognition systems of each sub group (audio, visual and psychological) in an online and offline condition and extending it further to developing an application. In the case of Sarcasm which is subtly shown via speech, it needs certain baseline functions that makes it highly detectable. Combining the mood detection mechanism along with the facial expressions would give more efficient and accurate results relevant to human emotion than just mood estimation results.

Acknowledgment

The authors wish to express gratitude towards Juan P Bello and Rachel Bittner for their valuable help. We are also thankful to authors Philip Jackson and Sanaul Haq for letting us use SAVEE database. Any opinions, findings and conclusion or recommendations used expressed in this features are those of the authors.

References

[1] C. Peter, Affect and Emotion in Human–Computer Interaction: From Theory to Applications, vol.4868, Springer-Verlag, New York, 2008.

[2] W. Yoon, K. Park, "A study of emotion recognition and its applications," Modeling Decisions for Artificial Intelligence, vol. 6417, pp. 455–462, 2007.

[3] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, "Being bored? Recognizing natural interest by extensive audiovisual integration for real-life application," Image Vis. Comput. vol. 27, no. 12, pp. 1760-1774, 2009.

[4] K. Van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, S. Baumann, "Fully generated scripted dialogue for embodied agents," Artificial Intelligence, vol. 172, no. 10, pp. 1219–1244, 2008.

[5] E. Lorini, F. Schwarzentruber, "A logic for reasoning about counterfactual emotions," Artificial Intelligence, vol. 175, no. 3, pp. 814–847, 2011.

[6] Amplitude Modulation Features for Emotion Recognition from Speech Md Jahangir Alam1, 2, Yazid Attabi2, 3, Pierre Dumouchel3 , Patrick Kenny2 , D. O'Shaughnessy1 1 INRS-EMT, University of Quebec, Montreal (QC), Canada 2 CRIM, Montreal (QC), Canada 3 École de technologie supérieure, Montreal, Canada

[7] Amplitude Modulation Features for Emotion Recognition from Speech Md Jahangir Alam1, 2, Yazid Attabi2, 3, Pierre Dumouchel3 , Patrick Kenny2 , D. O'Shaughnessy1 1 INRS-EMT, University of Quebec, Montreal (QC), Canada 2 CRIM, Montreal (QC), Canada 3 École de technologie supérieure, Montreal, Canada

[8] Approaching Multi-Lingual Emotion Recognition from Speech - On Language Dependency of Acoustic/Prosodic Features for Anger Detection Tim Polzehl1 , Alexander Schmitt2 , Florian Metze3 1Deutsche Telekom Laboratories / Quality and Usability Lab, Technische Universitat Berlin ¨ 2Dialogue Systems Group Institute, Information Technology University of Ulm 3Language Technologies Institute, Carnegie Mellon University, Pittsburgh

[9] A Multi-Modal Recognition System Using Face and Speech, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011 ISSN (Online): 1694-0814

[10] “Sure, I Did The Right Thing”: A System for Sarcasm Detection in Speech Rachel Rakov 1 , Andrew Rosenberg2 1 Linguistics Department, The Graduate Center CUNY, New York, USA 2 Computer Science Department, Queens College CUNY, New York, USA

[11] Kreuz, R. J. and Glucksberg, S. “How to be sarcastic: the echoic reminder theory of verbal irony”, Journal of Experimental Psychology: General, 118(4), 374-386, 1989.

[12] Clark, H.H., and Gerrig, R.J. “On the pretense theory of irony”, Journal of Experimental Psychology: General, 113(1), 121-126.

[13] Mueke, D.C. “The Compass of Irony”, Methuen, London, 1969.

[14] [24] Speaker Verification Using MFCC and Support Vector Machine, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong

[15] [25] C.C. Lin, S.H. Chen, T. K. Truong, and Yukon Chang, “Audio Classification and Categorization Based on Wavelets and Support Vector Machine,” IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 5, pp. 644-651, Sept. 2005.