INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Emotion Recognition from Speech Using Embedded Board OMAP 3530

Sravani Nellore[1], A.Ramesh Kumar[2], V.Naveen Kumar[3]

[1]PG student, Department of Electronics and Communication Engineering,
VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
sravaninellore@gmail.com
[2]Associate Professor, Department of Electronics and Communication Engineering,
VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
rameshkumar_aytha@yahoo.com
[3]Project Engineer, Research and Consultancy Center,
VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
naveenkumar_v@vnrvjiet.in

Abstract: *Speech emotion recognition (SER) is a current research topic in the field of human computer interaction (HCI) with wide range of applications. Speech is considered as a powerful means to communicate with intentions and emotions. Emotions add expressiveness to the natural language speech. In this paper the emotional state of a human being can be identified from his or her speech signal. Speech emotion recognition system can be used by disabled people for communication. The emotions from speech signal are recognized by considering the features of speech signal. The speech features such as Energy, Pitch, Mel-Frequency Cepstral Coefficients (MFCC) are extracted from speech utterance. In this paper Support Vector Machine (SVM) is used as a classifier to classify different emotional states such as anger, sadness, fear, boredom. SVM is simple and efficient algorithm which has a very good classification performance compared to other classifiers. The identification of emotion-related speech features is extremely challenging task. This paper describes design of speech emotion recognition system on beagle board OMAP 3530(ARM Cortex-A8 core) in Linux platform.*

Keywords**:** *Speech; Emotion Recognition; Feature Extraction; SVM; Database.*

1. INTRODUCTION

Current studies on emotion recognition concentrate on facial expressions and gestures etc. Automatic Emotion Recognition (AER) can be done in two ways, either by speech or by facial expressions. In the field of HCI, speech is primary objective of an emotion recognition system, as are facial expressions and gestures. In speech-based communications, emotion plays an important role. Emotion is an important cue to express one's feelings. It is also becoming more and more important in computer application fields such as healthcare, children education, etc. Its most important application is in intelligent human-machine interaction. Today there are many different algorithms which were used for various signal processing applications. In the recent years, a great research has been done to recognize human emotion using speech information [1], [2]. Many researchers explored several classification methods including the Neural Network (NN), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM), and Support Vector Machine (SVM) [7], [9]. In this paper the emotions from the speech signal are recognized by considering the features of speech signal by using SVM classifier. The basic emotions taken are classified as anger, happiness, sadness, boredom and fear. In this paper SVM is chosen as classifier which performs classification by constructing N-dimensional hyper planes that optimally separates the data into categories.

2. FEATURE EXTRACTION METHOD

Feature extraction stage is the most important one in the entire process, since it is responsible for extracting relevant information from the speech frames, as feature vectors. The speech signal contains a large number of information which reflects the emotional characteristics [3]. So in the research of speech emotion recognition, the most important thing is that how to extract and select better speech features with which most emotions could be recognized. The basic features extracted from speech signal are Pitch, MFCC, Energy, Voice quality and Intensity. We focus on providing an effective and robust feature extraction front-end for emotion recognition system, which only utilizes few robust features that have the potential to perform well in online and real life task.

a. Pitch:

Pitch is defined as the relative highness or lowness of a tone as perceived by the ear. The quality of sound depends on the number of vibrations per second produced by the vocal cords. The pitch signal has information about emotion [3]. Pitch is usually approximated by the fundamental frequency F0.Range of pitch for male=50-200HZ and for female=200-450HZ in voiced regions. The signal is periodic with a fundamental frequency between 80 Hz and 350 Hz. To calculate pitch the speech signal is broken into overlapping frames and a voiced speech value is taken from each

segment by autocorrelation method by taking time difference from one peak to next peak. Robust pitch detection is one of the most important technologies in speech signal processing.

b. MFCC:

The most popular set of feature vectors used in recognition systems is the Mel Frequency Cepstral Coefficients (MFCC). It is extracted from speech signal of spoken words used for joining two speech segments as shown in Figure.1. In this paper low frequency=0 and high frequency=8000.In addition to features that are commonly used for emotion the Mel-frequency Cepstral coefficients (MFCC) feature have proven to be one of the most effective feature sets first for speech processes, especially automatic speech recognition (ASR) and has less complexity.

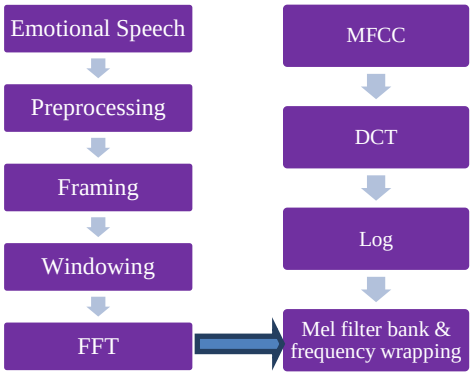

Figure.1: MFCC feature extraction

At the same time, the number of frames depends on the length of speech signal, sampling frequency, frame step, frame length. In this paper the length of speech signal is taken as 25ms at a frame rate of 10ms and a sampling frequency of 16 kHz, the frame step is 160 samples, and the frame length is 400 samples. MFCC is based on the characteristics of the human ear's hearing. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages. When the input speech emotion is given as input, processing steps are as follows:

Pre-processing and Framing: From whole speech utterance a short signal is taken removing silent parts as they do not carry any information. Signals are estimated for their energy to make it normalize. It is used to override noise frequency with Pre-emphasis value k=0.97. Speech signals are divided into frames of length 25ms and analysed every 10ms.
Windowing: A hamming window is applied to each frame to remove discontinuities in signal and ensure continuity between first and last data points.
FFT: It converts each frame from time domain signals into frequency domain and obtain frequency response of each frame.
Mel filter bank & frequency wrapping: Here triangular band pass filters are used to reduce dimensionality of features occupied.
Log: It computes the energy from raw signals before windowing and pre-processing is used for creating new frequencies and transforms multiplication into addition.
Discrete-cosine transform: It converts into time domain from frequency domain. It is used to compress the information of feature vector.

c. Energy:

Energy represents the loudness of the speech [3]. We calculate the energy for each segment by taking the summation of all the squared values of the samples amplitude. The amount of energy in a speech signal is taken in frame by frame process. Both MFCC and energy features are taken at same time. The amplitude of speech signal varies with time. More fluctuations may indicate active emotions, such as happiness and anger.

d. Voice quality:

The sound of someone's voice reaches us after traveling through the speaker's vocal tract. Therefore, the sound wave has certain characteristics, depending on the size and shape of the vocal tract. "Voice quality" refers to the properties of speech. Some voice quality features make things sound higher or lower. Some natural voice qualities said to map to affect and therefore assist in characterizing emotion in speech.

3. DATABASE

In this paper popular Berlin emotional speech Database (EMO-DB) [8] recorded in German language by male and female actors with different emotions is taken by asking an actor to speak with a predefined emotion. On the basis of Berlin database a new database is created in configuration file by name emovnr.conf with emotions sadness, happy, fear, boredom and anger are spoken by 5 male and female participants. Although, there are number of available databases well developed, it is usually a challenge to find a suitable one for a specific purpose. Therefore, researchers may prefer to design their own databases.
Choosing database: There are many aspects to be considered while trying to choose a database, such as the language, the scope (emotion analysis or recognition), subjects and observers (adults or children), naturalness (acted or natural), and the duration of phrases or dialogs etc.

4. SVM

LIBSVM is a library for Support Vector Machines (SVMs) [10], [11] which is most widely used tool for SVM classification and regression developed by C. J. Lin. Radial Basis Function (RBF) kernel is used in training phase. Support Vector Machines (SVMs) [6], [7] are a popular machine learning method for classification, regression and other learning tasks. In this paper SVC: support vector classification (two-class) method is used. A typical use of SVM involves two steps: Training a data set to obtain a model and using the model to predict information of a testing data set. Parameter selection is important for obtaining good SVM models. The Support Vector Machine is used as a classifier for emotion recognition. It performs classification by constructing an N-dimensional hyper planes

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

that optimally separates the data into categories [2]. SVM is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems. Under the conditions of limited training data, it can have a very good classification performance compared to other classifiers [3].

We choose SVM as our basic classifier due to its ease of training and its ability to work with any number of attributes. An optimal hyper plane is constructed to separate positive examples from negative ones. The examples with $y_i = +1$ are called positive examples, and those with $y_i = -1$ are negative ones. The main aim of SVM is to obtain a function f(x) that determines hyper-plane [2]. This hyper-plane optimally separates two classes of input data points. The separating hyper plane (margin) is chosen in such a way as to maximize its distance from the closest training examples of different classes. The below shown Figure.2 illustrates the geometric construction of hyper plane for two dimensional input space. Where M is margin, which is the distance from the hyper plane to the closest point for both classes of data points. As SVM is a binary classifier it scales each attribute of class labels with upper and lower boundaries [+1, −1].

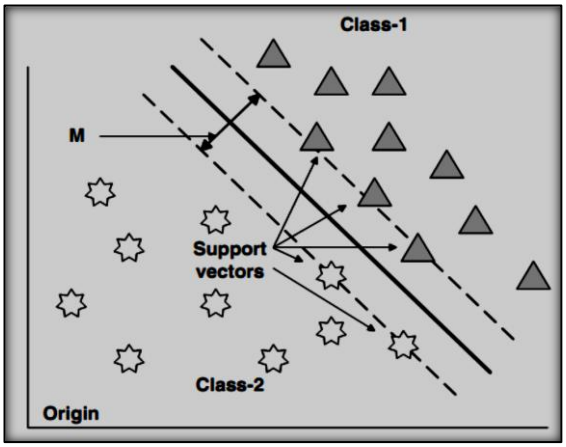

Figure.2: Hyper plane for two dimensional input space

a. Dimensional Approach:

The most commonly assumed dimensions are valence and arousal. The arousal dimension describes how excited one is, or how much energy is required to express the feeling. Feelings with high arousal induce some physical changes in the body such as increased heart rate, higher blood pressure and greater sub-glottal pressure resulting in change in speech as well such as making it louder, faster and have higher average pitch etc. The valence dimension describes to which extent the feeling is positive or negative, from unpleasant to pleasant. Emotional classification based on dimensions arousal and valence is shown in Figure.3.

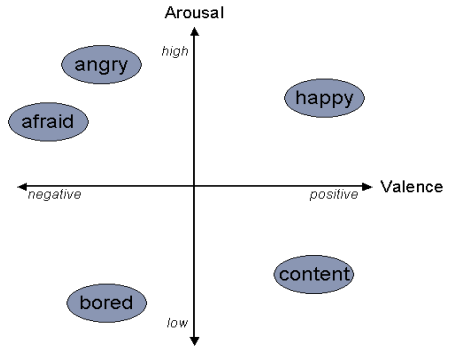

Figure.3: Classification of Emotions based on dimensions

5. SYSTEM DESIGN

a. Hardware Design

The hardware system consists of the Beagle board which is connected to the 5V DC Supply. The maximum supply voltage limitation of the Beagle board is 5V. The output of the embedded environment can be viewed in Touch Screen monitor. Secure Digital card (SD card) is inserted into the beagle board for mounting of u-Image file and root file system (RFS) which is used for driving the audio. MICOUT and MICIN are also a part of beagle board at which the audio jack is connected. Self-powered MIC is connected to MICIN for recording the speech. The below Figure.4 shows hardware design of embedded board consisting of OMAP3530 processor with SVideo, Stereo In & Out, USB Host, SD MMC, JTAG, LCD, Expansion pins, Reset & User buttons.

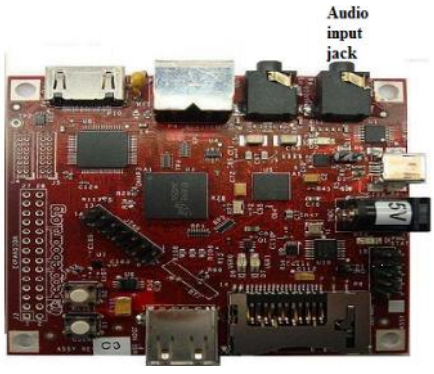

Figure.4: OMAP3530 Processor

i. OMAP 3530 Processor:

There are many features on this board which are useful for Open Embedded Developers. However, this project uses only few of the features. The Beagle Board is an OMAP 3530 platform. It has been equipped with a minimum set of features to allow the user to experience the power of the OMAP3530 [4]. By utilizing standard interfaces, the Beagle Board [5] is highly extensible to add many features and interfaces.

ii. Block Diagram of the System

The system block diagram is shown in Figure.5. Input is

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

given through MIC to the OMAP3530 Processor and OMAP3530 Processor is running with code written, extracts the features and recognizes the emotional state from the speech and displays the emotion on LCD.

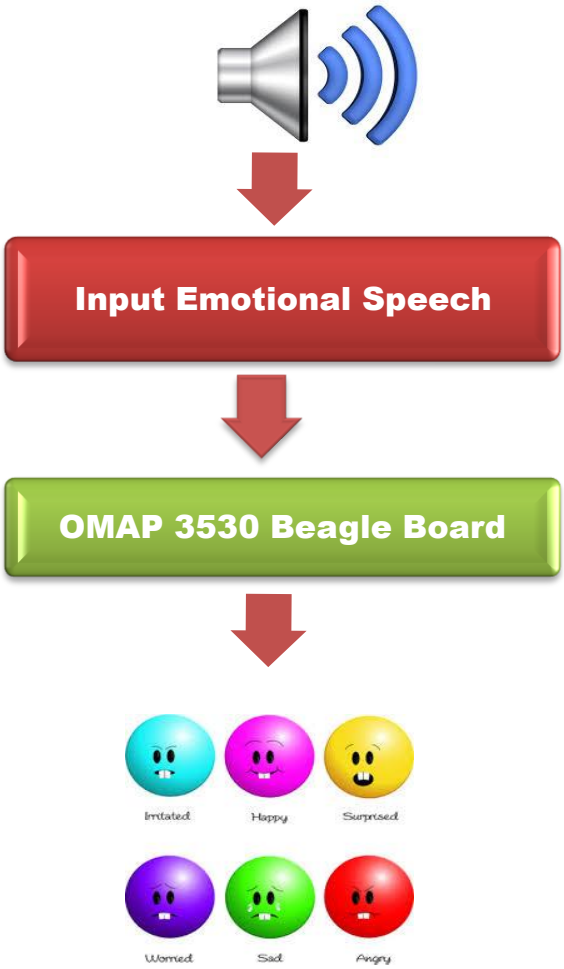

Figure 5: Block diagram of speech emotion system

Port Audio software is installed on OMAP3530 Processor and it is used to process the input speech. The port audio is used for taking input speech from microphone for live recording of speech emotions from sound card it has input/output library for cross platform works on Linux, Windows. It records the emotional speech then saves the sound into WAV files and forwards it depending on which function is invoked.

6. IMPLEMENTATION

We developed speech emotion recognition system on Beagle Board by porting Angstrom (mobile) operating system and OMAP3530 processor for ARM core in Linux platform which can be integrated into mobile devices. The programmable environment supports implementation in C++ assembly language; the entire code was compiled after several unsupported constructs in the code were removed.

Software:

•Linux on the Beagle Board:
• Angstrom (mobile operating system
Which is Linux Distribution) Porting Angstrom OS
Make two partitions on the SD/MMC card
•FAT partition (MLO, u-boot, u-Image)
•Ext2 partition
The five (5) boot phases
•ROM loads x-load (MLO)
• X-load loads u-boot
• U-boot reads commands
•Commands load kernel (u-Image)
•Kernel reads root file system.

7. PROCESS FLOW

The process flow consists of the main functions of the entire system. The main functions are Pre-Processing, feature extraction, SVM training, SVM classification. Flow chart of emotion recognition system is shown in Figure.6.

Steps:

a. In the first Step mode selection is done: Either we can choose Training or Testing.
b. Training is the most important step in the emotion recognition, once we have a set of features available.
c. First we train the classifier with some input speech (16 kHz .wav mono sample file) data of different emotional states. While training each feature along with a class label is stored in a corpus (database) is given for LIBSVM classifier.
d. The SVM is trained according to this labelled feature. After training the classifier, we can use it for recognizing the new given input.
e. In the testing part, the input of this system is .wav file, which come from the real-time speech.
f. Initially emotional speech sentence is given as input from the microphone the speech signal will be in analogue form which will be converted to digital form and saved in .wav format at a sampling frequency of 16khz is taken.
g. In pre-processing from the whole speech utterance taken the noise parts and silence parts are removed from the speech signal.
h. Then the signal is segmented into small frames of length 25ms using hamming window at a frame rate of 10ms with overlapping frames to compute the next set of speech parameters. Thus, overlapping segments are used for speech analysis.
i. In feature extraction process various features from speech signal are extracted like pitch, energy are calculated from each frame. By selecting the best features which are well recognized increases the performance of classifier.
j. In classification LIBSVM uses 2-dimensional space to separate two emotional classes and identifies class which is closest to the hyper plane is given as emotional output.
k. When the real-time speech is given as input if it matches with the pre-trained database the SVM checks for the probability of the emotion detection to be sad or happy and displays the output.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

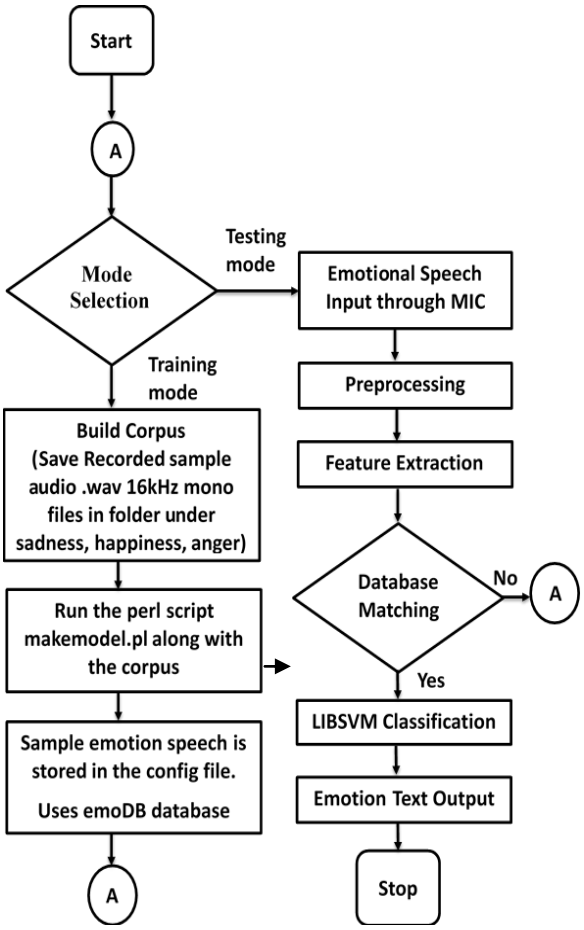

Figure 6: Flow chart of emotion recognition system

8. EXPERIMENTAL RESULT

In this paper the experiment is conducted by recording emotions from 5 male and 5 female participants each acting with different emotions. Each participant speaks 5 different sentences in 5 different emotions angry, happy, sad, boredom and fear. Files are chosen to be at 16-bit PCM, mono channel; sampled at a frequency of 16 KHz. In the experiment conducted we got the results as shown in the Table 1 and Table 2. In Table 1, female samples are compared with the corpus (trained database) and the detection percentage is classified in the Table1. Similarly for male samples the test is performed and the detection percentage of the male samples is classified in Table2

Table1: Experiment for Female samples

S.No	Emotion	Detection Percentage %
1.	Anger	70
2.	Boredom	80
3.	Fear	73
4.	Happiness	68
5.	Sadness	95

Table2: Experiment for Male samples

S.No	Emotion	Detection Percentage %
1.	Anger	74
2.	Boredom	86
3.	Fear	77
4.	Happiness	66
5.	Sadness	93

Figure7: Command run in a terminal to start the process

The above shown Figure.7 shows the initialization of all instances and starts recording. When the command config/emovnr.conf is run in terminal it starts processing of live emotion recognition from microphone registering all components from the component manager. Then reads config file which initializes list of all components specified. These components are fed to LIBSVM models which consist of loading instances scale and classes. Port audio device starts recording when all the components are loaded successfully. The below terminal windows shows that the emotion recognition from speech works in real time successfully. If the given emotional speech sample matches with the pre-trained ones it displays the output as fear or anger emotion depending on the probability. When the path of the command is given in the terminal it displays the output.

In the below shown Figure.8, fear emotion is recognized in beagle board when a fear emotional speech sample sentence is given as input from microphone. It saves the speech generating audio segments and ARFF file is created for each process. For one instance containing a feature vector it extracts features like Pitch, MFCC, Energy, Voice quality are identified from that speech signal and checks with the pre-trained database from LIBSVM models for the probability of that speech emotion to be fear or sad from the arousal and valence values which depends on the Pitch of speech signal.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

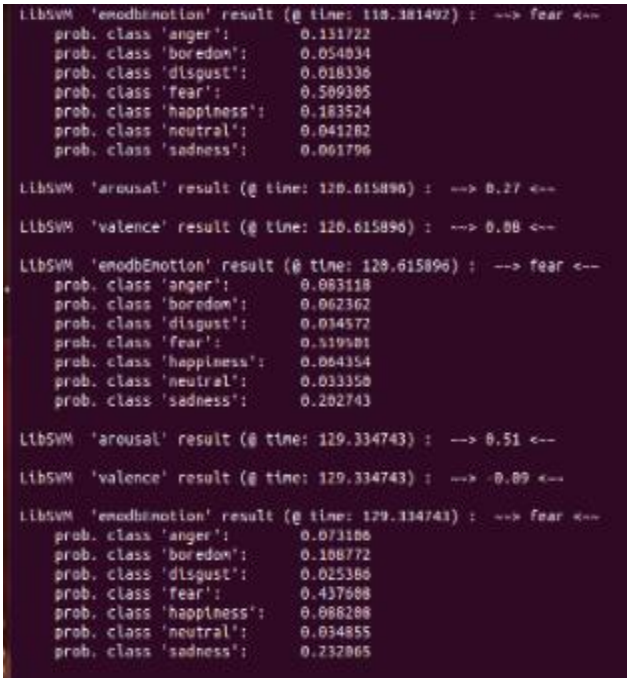

Figure8: Dimension and Emotion probabilities for fear emotion

In the above Figure. 8 the arousal and valence values are less for fear as the Pitch and Energy will be low for fear (low voices) checks probability for speech emotion to be fear. Classifies emotion as fear based on Pitch, MFCC, Energy and Voice quality. In the same way anger values will be high as Pitch and Energy will be high compared to fear emotion.
In the below shown Figure.9, anger emotion is recognized in beagle board when an anger emotional speech sample sentence is given as input from microphone.

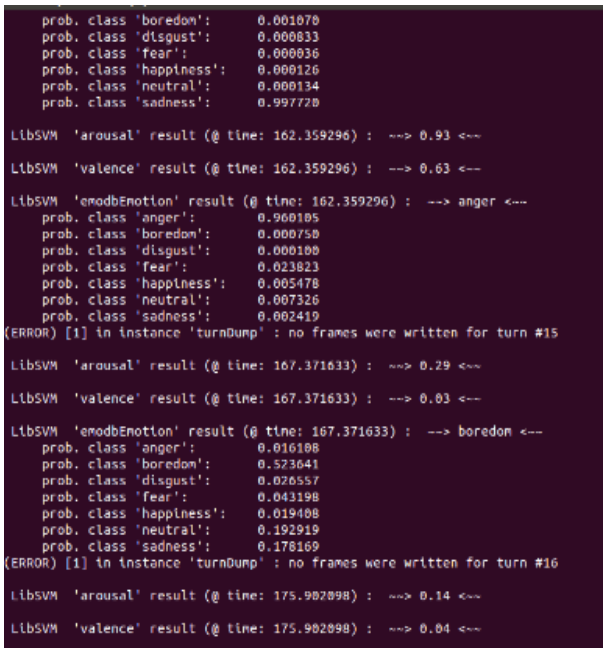

Figure9: Dimension and Emotion probabilities for anger emotion

9. CONCLUSION AND FUTURE SCOPE

In this paper, emotion recognition from speech system using embedded board OMAP 3530 is developed and tested successfully. The integration of speech emotions into human-machine interface has great challenges since emotions have to be recognized in real-time. Although it is difficult to get accurate result, we can show the variations that occur when emotion changes.The experimental results show that the speech emotions are identified efficiently. In future we can extend the implementation to detect new emotions. Emotion classification may also be useful for future consumer electronics or services. Making this a pluggable module which can be integrated into existing virtual assistants to give them the capability to recognize emotions. For example, since smartphones interface heavily with voice, they may be customized to automatically choose songs based on the user's current emotion.

REFERENCES

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

[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. (journal style)**

[3] **Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification", in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004. (conference style)**

[4] **http://linux.OMAP.com/mailman/listinfo/ Linux- OMAP-open-source. [Accessed: July. 13, 2013]. (General Internet site)**

[5] **http://elinux.org/Beagleboardbeginners [Accessed: June. 23, 2013]. (General Internet site)**

[6] **Christopher. J. C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, 2(2):955-974, Kluwer Academic Publishers, Boston, 1998. (journal style)**

[7] **Tristan Fletcher, Support Vector Machine Explained. (journal style)**

[8] **Burkhardt,Felix;Paeschke,Astrid;Rolfes, Miriam; Sendlmeier, Walter F.; Weiss, Benjamin A Database of German Emotional Speech. Proceedings of Interspeech, Lissabon, Portugal. 2005. (journal style)**

[9] **Christopher. J. C. Burges, A tutorial on support vector machines for pattern recognition, DataMining and Knowledge Discovery, 2(2):955-974, Kluwer Academic Publishers, Boston, 1998.9 (journal style)**

[10] **Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at**

http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[Accessed: Sept. 7, 2013]. (General Internet site)

[11] C.W Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification, Technical Report, Department of Comptuer Science & Information Engineering, National Taiwan University, Taiwan. (technical report style).