Development of a Real-time Embedded System for Speech Emotion Recognition

Amiya Kumar Samantaray

Department of Electronics and Communication Engineering

National Institute of Technology, Rourkela-769008

Development of a Real-time Embedded System for Speech Emotion Recognition

A Thesis submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

In
Electronics and Instrumentation Engineering

By

Amiya Kumar Samantaray

Roll No.: 110EI0255

Under the Guidance of

Prof. Kamala Kanta Mahapatra

Department of Electronics and Communication Engineering

National Institute of Technology

Rourkela-769008 (ODISHA)

May 2014

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGG.

NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA- 769 008

ODISHA, INDIA

CERTIFICATE

This is to certify that the thesis entitled "**Development of a Real-time Embedded system for speech emotion recognition**", submitted to the National Institute of Technology, Rourkela by **Amiya Kumar Samantaray, Roll No**. **110EI0255** for the award of the degree of **Bachelor of Technology** in Department of Electronics and Instrumentation Engineering, is a bonafide record of research work carried out by them under my supervision and guidance.

The candidate has fulfilled all the prescribed requirements. The thesis is based on candidate's own work, is not submitted elsewhere for the award of degree/diploma.

In my opinion, the thesis is in standard fulfilling all the requirements for the award of the degree of **Bachelor of Technology** in Electronics and Instrumentation Engineering.

Prof. Kamala Kanta Mahapatra

Supervisor

Department of Electronics and Communication Engineering

National Institute of Technology-Rourkela,

Odisha– 769008 (INDIA)

Dedicated to

NIT Rourkela

ACKNOWLEDGEMENT

I would like to convey our deepest gratitude towards our supervisor, Professor Kamala Kanta Mahapatra for his support and supervision, and for the valuable knowledge that he shared with us.

I would like to thank Professor A K Swain and Professor S K Das who have helped me to complete the thesis work successfully.

I would like to convey appreciation to all CYBORG, The robotics club members, for their encouragement and support.

I thank God for being on my side.

Amiya Kumar Samantaray

ABSTRACT

Speech emotion recognition is one of the latest challenges in speech processing and Human Computer Interaction (HCI) in order to address the operational needs in real world applications. Besides human facial expressions, speech has proven to be one of the most promising modalities for automatic human emotion recognition. Speech is a spontaneous medium of perceiving emotions which provides in-depth information related to different cognitive states of a human being. In this context, we introduce a novel approach using a combination of prosody features (i.e. pitch, energy, Zero crossing rate), quality features (i.e. Formant Frequencies, Spectral features etc.), derived features ((i.e.) Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding Coefficients (LPCC)) and dynamic feature (Mel-Energy spectrum dynamic Coefficients (MEDC)) for robust automatic recognition of speaker's emotional states. Multilevel SVM classifier is used for identification of seven discrete emotional states namely angry, disgust, fear, happy, neutral, sad and surprise in 'Five native Assamese Languages'. The overall experimental results using MATLAB simulation revealed that the approach using combination of features achieved an average accuracy rate of 82.26% for speaker independent cases. Real time implementation of this algorithm is prepared on ARM CORTEX M3 board.

Keywords—Linear Predictive Coding Coefficients, Mel Frequency Cepstral Coefficients, Prosody features, Quality features, Speech Emotion Recognition, Support Vector Machine.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATION

- HCI stands for Human-Computer Interaction
- ASER stands for Automatic Speech Emotion Recognition
- MFCC stands for Mel Frequency Cepstral Coefficients
- LPCC stands for Linear Predictive coding Coefficients
- MEDC stands for Mel Enegry Spectrum Dynamic Coefficents
- MATLAB stands for Matrix Laboratory
- HMM stands for Hidden Markov Model
- SVM stands for Support Vector Machine

Chapter 1: Introduction

1.1 Introduction

In the past decade, we have seen intensive progress of speech technology in the field of robotics, automation and human computer interface applications. It has helped to gain easy access to information retrieval (e.g. voice-automated call centers and voice search) and to access huge volumes of speech information (e.g. spoken document retrieval, speech understanding, and speech translation). In such frameworks, Automatic Speech Emotion Recognition (ASER) plays a major role, as speech is the fundamental mode of communication which tells about mental and psychological states of humans, associated with feelings, thoughts and behavior. ASER basically aims at automatic identification of different human emotions or physical states through a human's voice. Emotion recognition system has various applications in the fields of security, learning, medicine, entertainment, etc. It can act as a feedback system for real life applications in the field of robotics, where robot will follow human commands by understanding the emotional state of human. The successful recognition of emotions will open up new possibilities for development of an e-learning system with enhanced facilities in terms of student's interaction with machines. The idea can be incorporated in entertainment with the development of natural and interesting games with virtual reality experiences. It can also be used in the field of medicine for analysis and diagnosis of cognitive state of a human being. With the advancement of the human-machine interaction technology, a user-friendly interface is becoming even more important for speech-oriented applications. The emotion in speech may be considered as similar kind of stress on all sound events across the speech. Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. With the advance of the human-machine interaction technology, a user-

friendly interface is becoming more and more important for speech-oriented applications.

1.2 Literature Review

In recent years, a great deal of research has been done to recognize human emotions using speech information. Researchers have combined new speech processing technologies with different machine learning algorithms [1], [2] in order to achieve better results. In machine learning platform, speech emotion recognition belongs to supervised learning, following the generalized system model of data collection, feature exaction and classification. The extremely complex nature of human emotional states makes this problem more complicated in terms of feature selection and classification. Many Researchers have proposed important speech features which contain emotion information, such as prosody features [3] (pitch [4], energy, and intensity) and quality features [5], [6] like formant frequencies and spectra temporal features [7]. Along with these features, many state of- the-art derived features like Mel-Frequency Cepstral Coefficients (MFCC) [8], [9], Linear Predictive Coding have been suggested as very relevant features for emotion recognition. We have also considered some dynamic features like Mel-energy spectrum dynamic coefficients (MEDC) and have combined all the features to get a better result in emotion recognition. Many researches provide an in-depth insight into the wide range of classification algorithms available, such as: Neural Networks (NN), Gaussian Mixture Model (GMM) [10], Hidden Markov Model (HMM) [11], Maximum Likelihood Bayesian Classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support Vector Machine (SVM) [12], [13], [14], [15]. We have chosen Support vector machine for our research work as it gives better results in emotion recognition domain of various databases like BDES (Berlin Database of Emotional Speech) and MESC (Mandarin Emotional Speech Corpora).

1.3 Motivation

Humans have been endowed by nature with the voice capability that allows them to interact and communicate with each other. Hence, the spoken language becomes one of the main attributes of humanity. Intensive progress of speech technology in the field of robotics, automation and Human Computer Interface (HCI) applications which is future of the world. Emotion recognition system has various applications in the fields of security, learning, medicine, entertainment, etc. A feedback system for real life applications can be developed in the field of robotics, where robot will follow human commands by understanding the emotional state of humans. This research will open up new possibilities for development of an e-learning system with enhanced facilities in terms of student's interaction with machines. If incorporated in entertainment world with the development of natural and interesting games the virtual world will become the real world for us. It can be used in the field of medicine for analysis and diagnosis of cognitive state of a human being. Microsoft and Google has been trying to implement speech interactive system but till today they are not completely successful with flaws in real time integration with an online database.

1.4 Objective

- To develop a robust algorithm for emotion recognition for an Indian Language
- A real time embedded system implementing the same algorithm and a completely dedicated hardware for speech emotion recognition.

1.5 Work done

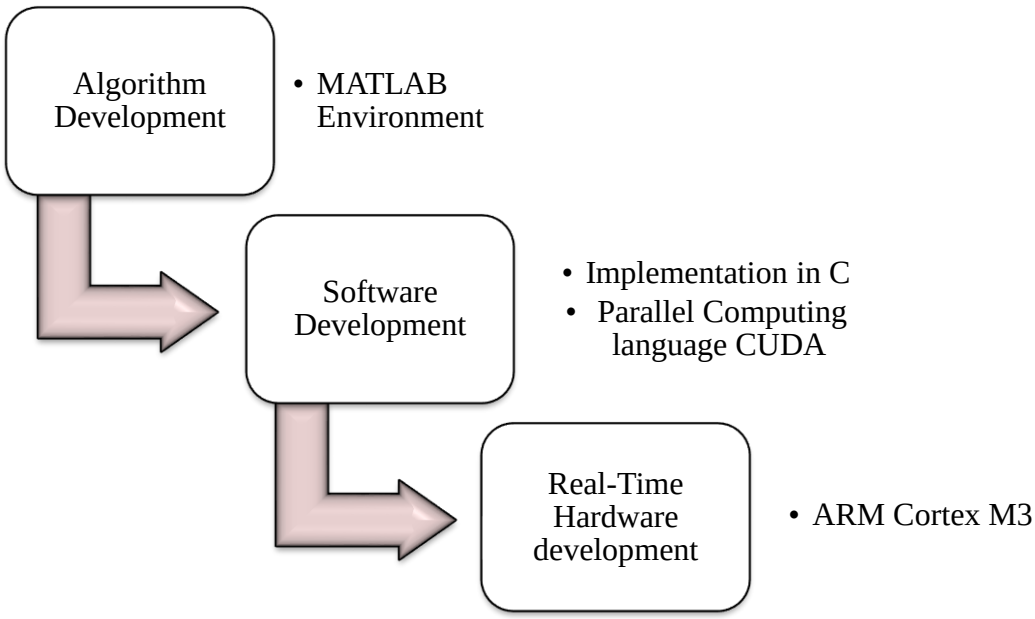

Fig. 1: Flow diagram of work done

1.6 Database Description

In our experiment, we have used utterances of "Multilingual Emotional Speech Database of North East India" (MESDNEI) [8], [10] which have been collected from different places of Assam. Thirty subjects randomly selected from a group of non-professional and first-time

trained volunteers were requested to record emotional speeches in 5 native languages of Assam (3 males and 3 females per language), to build the database. This database includes utterances belonging to seven basic emotional states anger, disgust, fear, happy, neutral, sad and surprise. Each person recorded 140 short sentences (20 per emotion) of different lengths in his or her first language. This makes the database, a combination 4200 utterances, enrich in various modalities in terms of gender and languages. The speech samples were recorded with 16 bit depth and 44.1 kHz sampling frequency. A listening test of emotional utterances was carried out for validation of the MESDNEI database.

1.7 Thesis Organization

Chapter 1 includes Introduction, Literature Review, Motivation, Work done , Database description and Thesis organisation.

Chapter 2 includes overall system description and various features related to emotion are explained in details. This also includes the formation of feature vector by applying various statistics and the use of SVM classifier.

Chapter 3 includes the hardware implementation in ARM Cortex M3 board.

Chapter 4 includes the conducted experiments and its results.

Chapter 5 describes the Future Scope and Conclusion.

Chapter 2: System Description

2.1 Introduction

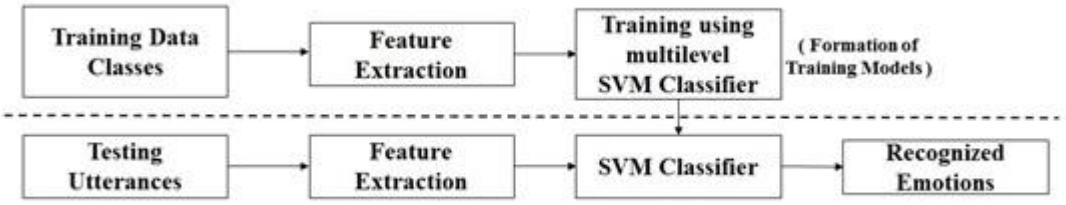

Fig. 2 Generalized system model for emotion recognition

Machine learning which concerns the development of algorithms, which allows machine to learn via inductive inference based on observation data that represent incomplete information about statistical phenomenon. Classification, also referred to as pattern recognition, is an important task in Machine Learning, by which machines "learn" to automatically recognize complex patterns, to distinguish between exemplars based on their different patterns, and to make intelligent decisions. A pattern classification task generally consists of three modules, i.e. data representation (feature extraction) module, feature selection or reduction module, and classification module. The first module aims to find invariant features that are able to best describe the differences in classes. The second module of feature selection and feature reduction is to reduce the dimensionality of the feature vectors for classification. The classification module finds the actual mapping between patterns and labels based on features. The objective of our work is to investigate the machine learning methods in the application of automatic recognition of emotional states from human speech.

Different Machine Learning Algorithms based on the input available at the time of training:

- Supervised learning algorithms are trained on labelled examples, i.e., input where the desired output is known. The supervised algorithm attempts to generalize a function or

mapping from input to outputs which can then be used to speculatively generate an output for previous unseen inputs.

- Unsupervised learning algorithms operate on unlabeled examples, i.e. input where the desired output is unknown. Here the objective is to discover structure in the data (e.g. through a cluster analysis), not to generalize a mapping from input to output.
- Semi-Supervised learning combines both labeled and unlabeled examples to generate an appropriate function or classifier.

This problem belongs to the class of supervised learning of pattern recognition as we have to train the machine for particular classes with labelled data.

2.2 Feature Extraction

Different emotional states can be recognized using certain speech features which can be either prosody features or quality features. Some Prosody features which can be extracted directly; includes pitch, intensity and energy are the most widely used features in the emotion recognition domain. Though it is possible to distinguish some emotional states using only these features, but it becomes very inconvenient when it comes to emotional states with same level of stimulation [5]. The difficulty in distinguishing between joy and anger can be lowered by reflecting some quality features. Formants and Spectral energy distributions are the most important quality features to solve the classical problem of emotion recognition using speech. While prosody features are preferred on the arousal axis, quality features are favored on valence axis. Some other features which are derived from the basic acoustic features like MFCC and

coefficients from Linear Predictive Coding are considered good for emotion recognition. Some dynamic features can be obtained from the variation of speech utterances in time domain by taking first order derivative of MFCC; named as MFDC. A method which combines all the above mentioned features is more promising than a method that uses only one type of features for the classification. Fig 3. Shows the overall model for feature extraction that has been used for both training of classifier and testing the unknown speech samples.

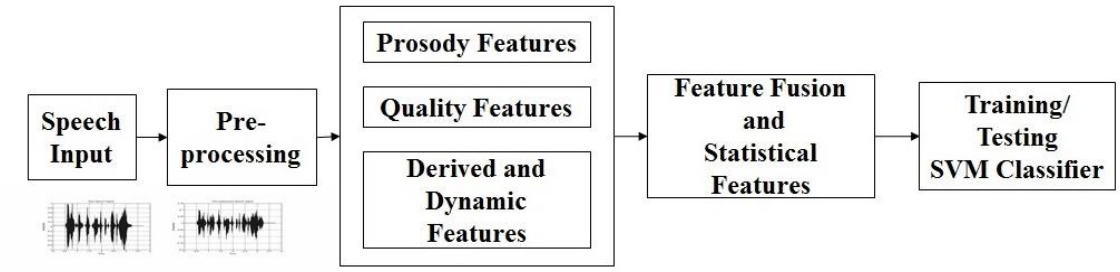

Fig. 3 Steps for feature extraction

2.2.1 Pre-processing

The speech samples which are going to be processed for emotion recognition should go through a pre-processing step that removes the noise and other irrelevant components of speech corpus for better perception of speech data. The preprocessing step involves three major steps such as pre -emphasis, framing and windowing. The pre-emphasis step is carried out on the speech signal using a Finite Impulse response (FIR) filter called pre-emphasis. The filter impulse response is given by

$$H(z) = 1 + a\, z^{-1} \text{ , where a = -0.937} \qquad (1)$$

The filtered speech signal is then divided into frames of 25ms with an overlap of 10ms. A hamming window is applied to each signal frame to reduce signal discontinuity and thus avoid

spectral leakage. Then speech emotion related features are extracted from the pre-processed speech data.

2.2.2 Prosody Features

The fundamental frequency, often referred to as pitch, is one of the main acoustic correlate of tone and intonation, and depends upon the number of vibrations per second by vocal cord and represents the highness or lowness of a tone as perceived by the ear. Pitch provides ample amount of information related to emotions as it is a perceptual property which is related to the tension of vocal folds and sub-glottal air pressure. Zero-crossing rate is a key feature for identification of percussive sounds and information retrieval, and gives information related to change in frequency components of speech. The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. The varying nature of speech signals insists for the use of energy related features which will show the variation of energy in the speech corpse associated with a short-term region. We have extracted short term energy and entropy from the speech signals for the formation of feature vector.

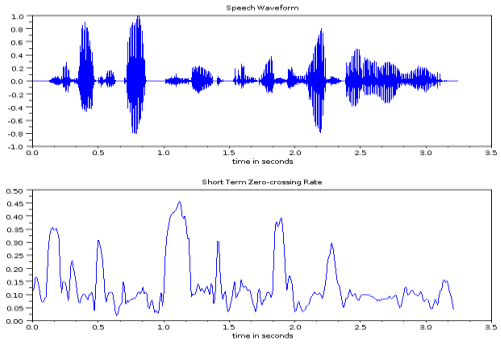

Fig 4. Zero Crossing Rate

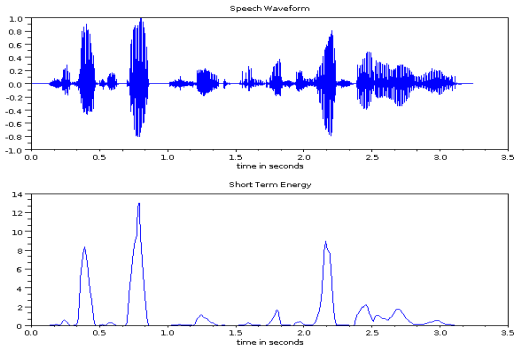

Fig 5. Short Term Energy

2.2.3 Quality Features

Spectral properties, which include spectral roll-off, spectral centroid, and spectral flux, can be extracted using Hilbert envelope. Spectral roll-off point can be defined as 85% percentile of the power spectral distribution. The roll off point is the frequency below which the 85% of the magnitude distribution is concentrated. This measure is useful in distinguishing voiced speech from unvoiced. Similarly, a feature extractor extracts the spectral centroid by measuring the center of mass of power spectrum by calculating the mean bin. The spectral flux can be found by calculating the difference between the current values of the magnitude spectrum bin in the current window and the corresponding value of the magnitude spectrum of the previous one. This provides a good measure of spectral change of the signal. Formants play an important role as feature and can be termed as the spectral peaks of the sound spectrum of voice, they are often measured as the amplitude peaks in the frequency spectrum of the sound. The first three formant frequencies can be taken as relevant features in the complete feature vector.

2.2.4 Derived and Dynamic Features

Along with the prosody and the quality features, MFCC and LPCC features are also extracted from the speech utterances for better classification. In speech processing, the basic human speech production model is described by source filter model. Source is related to the air expelled from the lungs. Filter is responsible for giving a shape to the spectrum of the signal in order to produce different sounds. As convolution of source and filter represents speech, two convoluted signals can’t be separated by linear filtering. First the non-linear combination is converted to linear combination. So we need a different scale like ‘Mel Scale’ for subjective measurement, which can be described as follows:

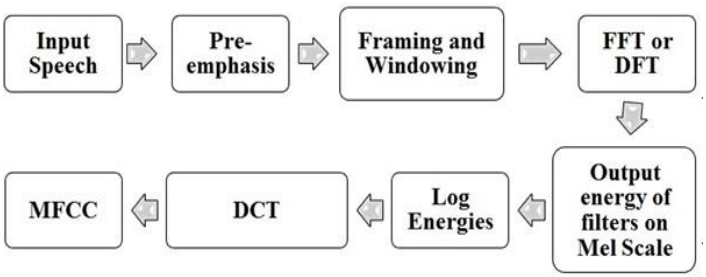

Fig 6. MFCC Feature Extraction

$$m = 2595 * \log\left(1 + \frac{f}{700}\right) \quad (2)$$

The process of finding MFCC can be described as shown in fig.6.

$$e(n) * h(n) = x(n). \quad (3)$$

Taking Z-Transform on both sides: $E(z)\,H(z) = X(z)$. (4)

Now taking log of both sides: $C(z) = \log X(z) = \log E(z) + \log H(z)$. (5)

We assume that H(z) is mainly composed of low frequencies and that E(z) has most of its energy in higher frequencies, in a way that a simple low-pass filter can separate H(z) from E(z). Hence, after filtering out H (z) and taking inverse Z-transform, the time domain signal e (n) can be obtained. The Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency, i.e., a nonlinear "spectrum-of-a-spectrum".

Linear Predictive Coding is a method of separating out the effects of source and filter from a speech signal. This is a way of encoding the information in the speech signal into a smaller space for transmission over a restricted channel. It encodes the signal by finding a set of weights on earlier signal values that can predict the next signal value.

$$Y[n] = a[1]\,Y[n-1] + a[2]\,Y[n-2] + \ldots + a[n] \quad (6)$$

The above mentioned equation is a very close match to our source filter model of speech production where we excite a vocal tract filter with either a voiced signal or noise source. It embodies the characteristics of particular channel of each person, and the same person with different emotional speech will have different channel characteristics, so we can extract these feature coefficients to identify the emotions contained in speech. This is an efficient method which can better describe the vowels; while the disadvantage is that the description of the consonants are less capable and less noise immunity. The filter coefficients derived by the LPC analysis contain information about the glottal source filter, the lip radiation and vocal tract itself. It is based on the source-filter model, where the vocal tract transfer function is modeled by an all-pole filter with a transfer function given by

$$H(z) = \frac{1}{1 - \sum a_i z^{-i}} \qquad \text{Where } a_i \text{ is the filter coefficients} \qquad (7)$$

The speech signal S is assumed to be stationary over the analysis frame and approximated as a linear combination of the past p samples.

MEDC extraction process is similar with MFCC. The only one difference in extraction process is that the MEDC is taking logarithmic mean of energies after Mel Filter bank and Frequency wrapping, while the MFCC is taking logarithmic after Mel Filter bank and Frequency wrapping.

2.3 Classification using SVM

The input audio signal was divided into frames and all the features were calculated for each frame. Now, In order to draw one conclusion from all the features of several frames of the input signal, we need to consider some kind of statistics. Statistical features [16] like Mean, Standard Deviation, Max and Range were considered for each feature over all the frames, and

a single feature vector was formed including all the statistical parameters, representing the input signal. Then, the normalized statistical feature vector was provided to the Support Vector Machine (SVM) classifier for training or testing.

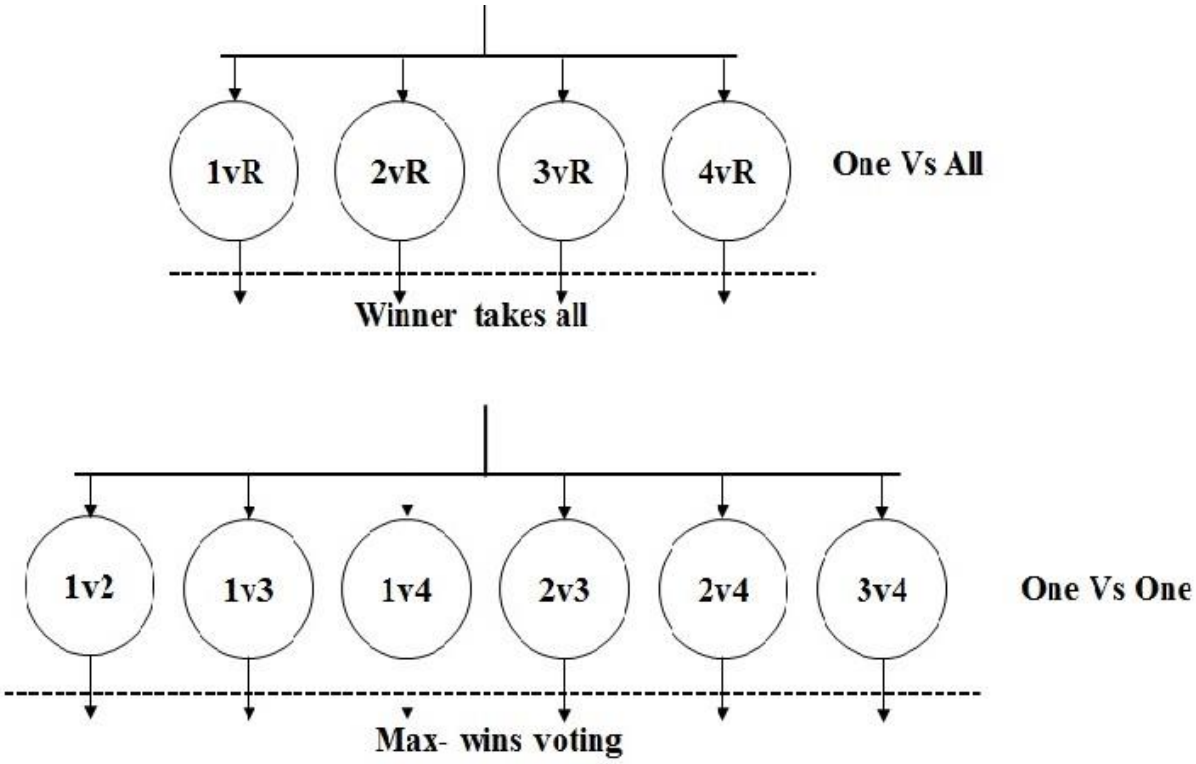

Fig 7. Example of multi-level SVM classification for four different classes

A single SVM is a binary classifier which can classify 2- category data set. For this, first the classifier is manually trained with the pre-defined categories, and the equation for the hyper-plane is derived from the training data set. When the testing data comes to the classifier it uses the training module for the classification of the unknown data. But, automatic emotion recognition deals with multiple classes. Two common methods used to solve multiple classification problems like emotion recognition are (i) one-versus-all [17], and (ii) one-versus-one [18]. Fig.7 demonstrates these two methods of multilevel SVM [19], [20] classification for four different classes. In the former, one SVM is built for each category, which distinguishes this category from the rest. In the latter, one SVM is built to distinguish between every pair of

categories. The final classification decision is made according to the results of all the SVMs with the majority rule. In the one-versus-all method, the category of the testing data is determined by the classifier based on the winner-takes-all strategy. In the one-versus-one method, every classifier assigns the utterance to one of the two emotion categories, then the vote for the assigned category is increased by one vote, and the emotion class is the one with most votes based on a max-wins voting strategy. We have used one versus all SVM classification method to recognize the emotional states in our experiment on MESDNEI.

Chapter 3: Hardware Implementation in ARM Cortex Board

3.1 Introduction

A dedicated single purpose standalone system which combines different mechanical, electrical and chemical components can be termed as embedded systems. These systems are abundant components of our day to day life. We interact with these kind of devices every day and it makes our life better. One of the sole purpose of this project is to develop an embedded system for an emotion recognition device which can take input in the form of audio signal. Developing a real time embedded system for speech emotion recognition is comparatively very tough task as the processors used in embedded applications are less powerful in terms of computational power and clock speed. We have chosen ARM Cortex series boards as they support a large range of functionalities for real time system development including DSP libraries. The algorithm described above is implemented in the same way as implemented in MATLAB environment. The development of C code, which is optimized for a low end machine with low resources, is an essential part of this embedded system.

3.2 ARM Architecture

ARM Cortex series is an example of Harvard Architecture which means it has separate data lines and instruction buses. Its instruction set combines the high performance typical of a 32 bit processor with high code density. In our project we have used ARM Cortex M3 board for implementation. The board has wide range of features having clock frequency of 72 MHz. Along with this the STM32F107VC ARM Cortex M3 processor includes 256 KB Flash and 64 KB RAM memory with a Color QVGA TFT LCD with touch screen.

The inbuilt micro SD interface in this board is used to store the training data sets for the experiment. This board also includes microphone and speaker in its peripheral. Fig. 8 describes the complete development board with all its peripherals and inbuilt hardware set up.

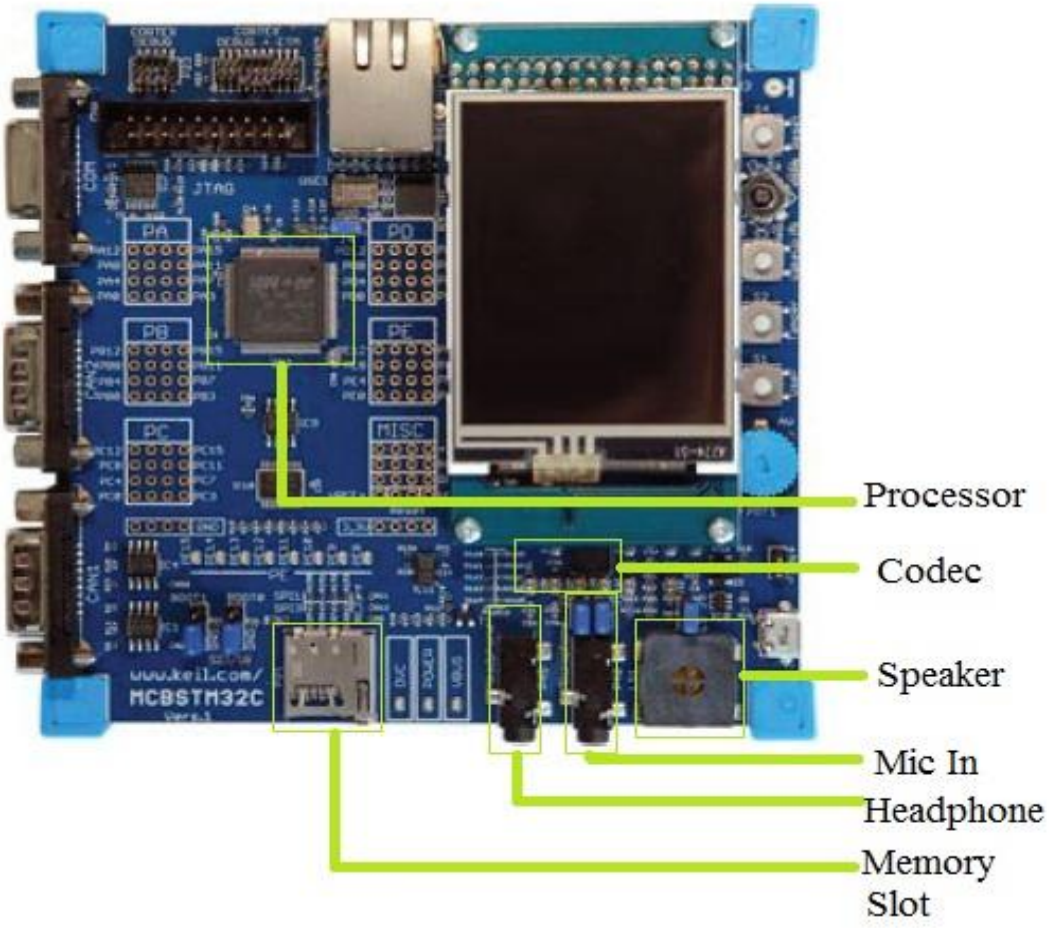

Fig 8. Development board of STM32F107VC ARM Cortex M3 processor

3.2.1 Hardware Set up

The hardware set up is much simple comparative to the algorithm development. The MATLAB code is written in C for the hardware implementation. This application reads the

pre-recorded speech samples stored in the memory card for the training and offline processes all the files using the in-built DSP library functions of ARM board. At the time of real time testing it uses the microphone and speaker for interacting with the environment. It takes the continuous speech input from the user and uses the SVM classifier for the classification of the current speech frame and inform the current emotional state of the user by giving a speech output through speaker. Fig. 9 describes the complete hardware block diagram of the setup.

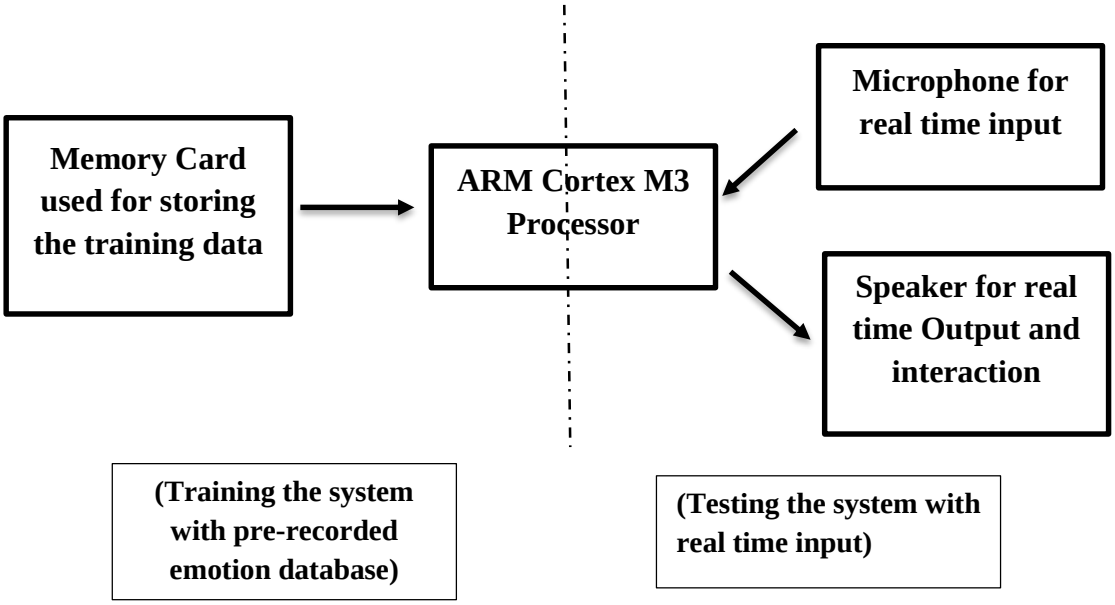

Fig 9. Generalized Hardware Set up

3.2.2 Hardware implementation

For implementing the code in ARM board first we need to read the standard .wav file from a microSD Card. The WAVE file format is a file format specification for the storage of multimedia files which is a subset of Microsoft's RIFF specification.

A RIFF file starts out with a specific file header followed by a pattern of data chucks. The "WAVE" file is represented by two sub-chunks: "fmt" and "data". The "fmt" sub-chuck describes the format of the sound information in the data sub-chunk. The "data" sub-chunk indicates the size of the sound information and contains the raw sound data. The header for the specific .wav file can be written as follows:

```
typedef struct
{
        char Chunk_ID[4];
        uint32_t ChunkSize;
        char Format[4];
        char FormatChunkID[4];
        uint32_t FormatChunkSize;
        uint16_t AudioFormat;
        uint16_t NumOfChannels;
        uint32_t SampleRate;
        uint32_t ByteRate;
        uint16_t BlockAlign;
        uint16_t BitsPerSecond;
        char OutputChunkID[4];
        uint32_t OutputChunkSize;
}
WavHeader;
```

This format is specified for the structure of wave file. All the .wav files meant for the training data are stored in the micorSD card and read one by one thorough the ARM processor. For reading the FAT32 file system like the microSD card we need SPI communication protocol which is required for reading the data from the memory card. Then the complete algorithm is developed using ARM cortex board and Keil Software. The output of the wave files and processing are displayed in the TFT touch screen based display set up.

Chapter 4: Experiment and Results of Simulation

4.1 Experimental Setup

In our experiment, we have taken speaker-independent training models for SVM. For each of the Assamese languages, we have taken utterances of five speakers as training set and the other speaker's speech samples for testing purpose. The feature vector includes four prosody features (Pitch, ZCR, Short-term Energy, Log-entropy), six quality features (first three formant frequencies, Spectral Roll-off, Spectral flux, Spectral centroid), 14 Mel Frequency Cepstral Coefficients, 12 Linear Predictive Coding Coefficients, 13 Mel-Energy spectrum Dynamic Coefficients. We have taken four statistics (mean, standard deviation, max and range) for each feature class in order to form a single feature vector for each utterance. This makes a feature vector of 196 features for each sample. A total of 600 speech samples for each language are used for training and 120 speech samples are used for testing purpose.

4.2 Result of simulation

Our experiment shows 79.3%, 78.57%, 82.8%, 89.23%, and 81.43% accuracy for emotion recognition in Assamie, Dimasa, Bodo, Karbi and Mishing language respectively which is shown in bar graphs in the mentioned order.

	Angry	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**	**Percentage**
Angry	**15**			**4**		**1**		**75**
Disgust	**1**	**16**					**3**	**80**
Fear			**18**		**1**	**1**		**90**
Happy	**5**			**15**				**75**
Neutral		**1**	**1**		**13**		**4**	**65**
Sad				**1**		**19**		**95**
Surprise				**3**		**2**	**14**	**70**
								78.57

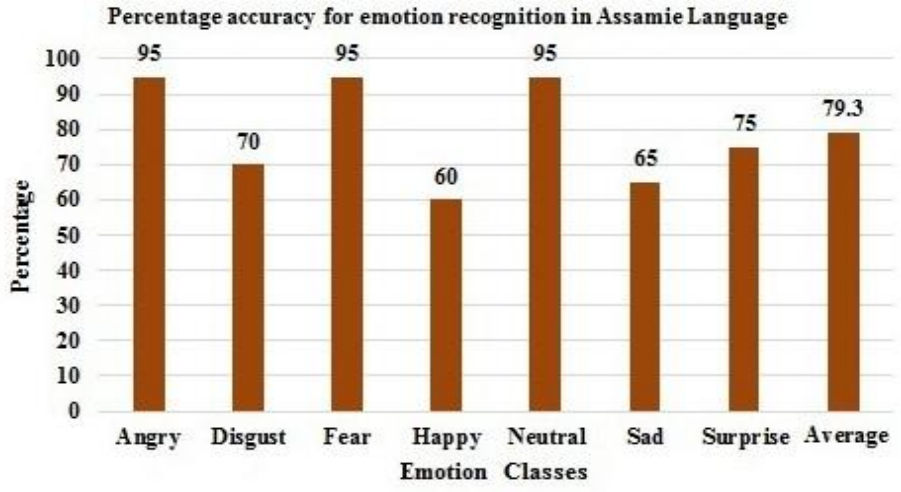

Fig 10. Emotion recognition accuracy in Assamie language

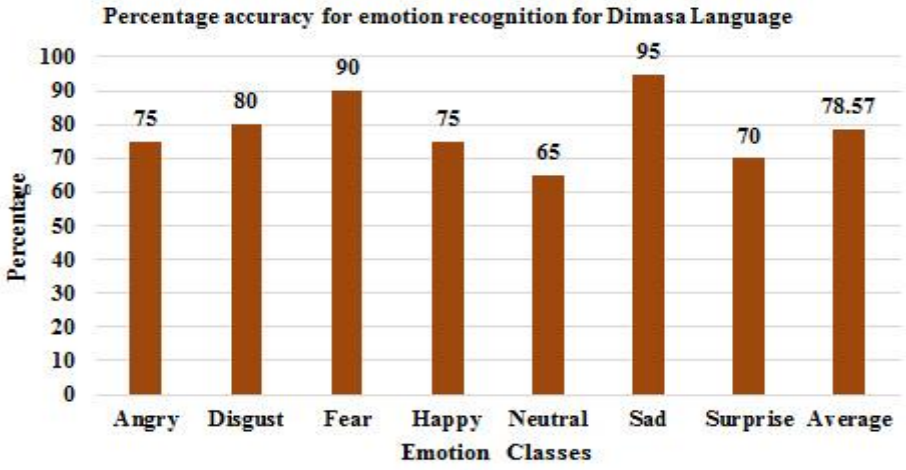

Fig 11. Emotion recognition accuracy in Dimasa language

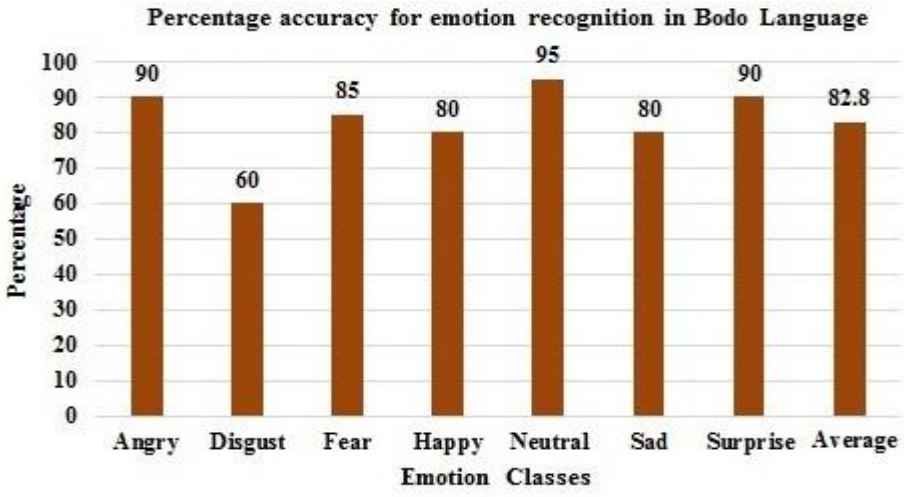

Fig 12. Emotion recognition accuracy in Bodo language

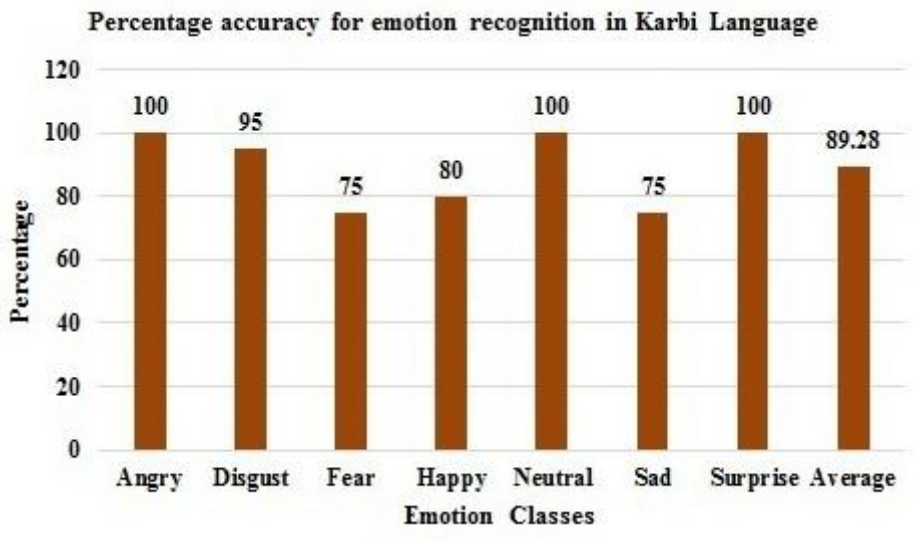

Fig 13. Emotion recognition accuracy in Karbi language

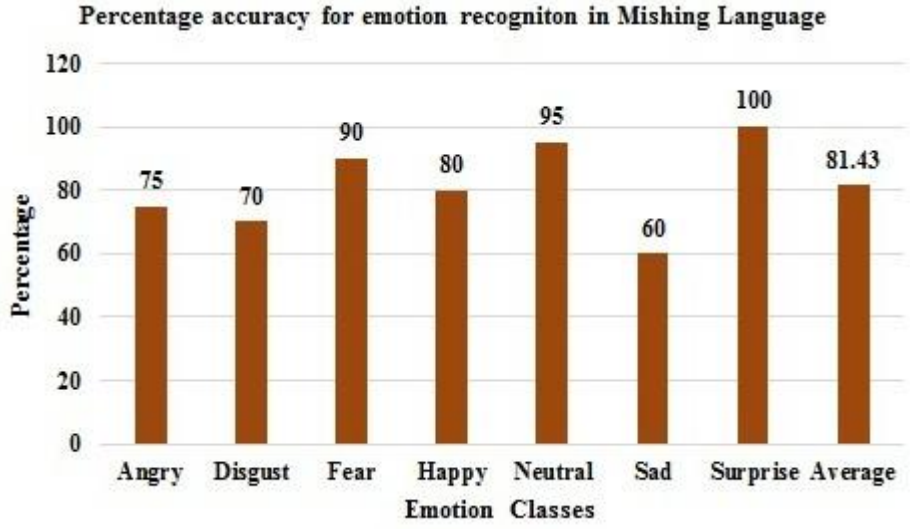

Fig 14. Emotion recognition accuracy in Mishing language

Chapter 5: Future Scope and Conclusion

Though a lot of work has been done in this project, but are ample amount of scope remaining to be done in future. An implementable and robust real time model for these applications can be a scope for future work. This will require an improvement in feature selection and Classification strategies of emotion recognition algorithm. Implementation of RTOS for real time embedded platforms can be a novel work as the future of technologies are concerned. Implementation in Multicore CPU or General Purpose Graphics Processing Unit like NVIDIA or AMD GPUs for cloud based platforms can also be done in future. With a lot of future scope, a development work to can be done for a user interface using android or java or C++ for portable devices. Development of this emotion recognition engine for rapidly increasing low power hand-held devices can be a novel work for future.

To sum up, this paper discusses a speech based emotion recognition engine using the combination of prosody, quality, derived and dynamic features with the help of SVM classifier. The classification methods reported in this work and those in different literatures are not comparable as they have used different databases and incomparable experimental protocols. Despite of several researches in the field of emotion recognition a real time model for this application has not been developed yet.

PUBLICATIONS

[1] B. Kabi, **A. K. Samantaray**, P. Patnaik, A. Routray, " Voice Cues, Keyboard Entry and Mouse Click for detection of affective and cognitive states: A case for use in technology-based pedagogy", *Fifth IEEE International Conference/ T4E,* Dec. 18-20, 2013.

[2] **A. K. Samanataray**, K. K. Mahapatra, B. Kabi, A. Kandali, A. Routray, " A Novel approach of speech emotion recognition with prosody, quality, derived features using SVM classifier". *IEEE/ ICRAIE, 2014. (Paper accepted)*

REFERENCES

[1] C. M. Lee, S. Member, S. S. Narayanan, and S. Member, "Toward Detecting Emotions in Spoken Dialogs," vol. 13, no. 2, pp. 293–303, 2005.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[3] I. Luengo and E. Navas, "Automatic Emotion Recognition using Prosodic Parameters" pp. 493–496, 2005.

[4] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 143–160, Aug. 2012.

[5] M. Borchert and a. Dusterhoft, “Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments,” *2005 Int. Conf. Nat. Lang. Process. Knowl. Eng.*, vol. 00, pp. 147–151, 2005.

[6] Y. Z. Y. Zhou, Y. S. Y. Sun, J. Z. J. Zhang, and Y. Y. Y. Yan, “Speech Emotion Recognition Using Both Spectral and Prosodic Features,” *2009 Int. Conf. Inf. Eng. Comput. Sci.*, pp. 0–3, 2009.

[7] S. Wu, H. Tiago“Automatic Recognition Of Speech Emotion Using Long-Term Spectro-Temporal Features,” 2009.

[8] A. B. Kandali, S. Member, A. Routray, and T. K. Basu, “Emotion recognition from Assamese speeches using MFCC features and GMM classifier.”

[9] M. Learning, M. In, T. Application, O. Speech, and E. Recognition, “ Machine Learning Methods In The Application Of Speech,” pp. 1–21.

[10] A. B. Kandali, A. Routray, and T. K. Basu, “Vocal emotion recognition in five languages of Assam using features based on MFCCs and Eigen Values of Autocorrelation Matrix in presence of babble noise,” *Commun. (NCC), 2010 Natl. Conf.*, 2010.

[11] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” *2003 Int. Conf. Multimed. Expo. ICME ’03. Proc. (Cat. No.03TH8698)*, vol. 1, pp. 1–4, 2003.

[12] Y. Pan, P. Shen, and L. Shen, “Speech Emotion Recognition Using Support Vector Machine,” vol. 6, no. 2, pp. 101–108, 2012.

[13] M. Dumas, “Emotional Expression Recognition using Support Vector Machines.”

[14] P. Shen and X. Chen, “Automatic Speech Emotion Recognition Using Support Vector Machine,” pp. 621–625, 2011.

[15] B. Schuller, G. Rigoll, and M. Lang, “Machine - Belief Network Architecture,” in *IEEE/ICASSP*, 2004, pp. 577–580.

[16] T. Iliou and C. Anagnostopoulos, “Statistical Evaluation of Speech Features for Emotion Recognition,” 2009.

[17] R. Rifkin, “In Defense of One-Vs-All Classification,” vol. 5, pp. 101–141, 2004.

[18] D. Fradkin and I. Muchnik, “Support Vector Machines for Classification,” vol. 0000, pp. 1–9, 2006.

[19] A. Hassan and R. I. Damper, “Multi-class and hierarchical SVMs for emotion recognition.”

[20] N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, and W. Heinzelman, “Speech-based Emotion Classification using Multiclass SVM with Hybrid Kernel and Thresholding Fusion" pp. 455–460, 2012.