ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

Emotion Recognition from Acted Assamese Speech

Nilim Jyoti Gogoi[1], Jeemoni Kalita[2]

Assistant Professor, Dept of Computer Science, University of Science and Technology, Meghalaya, India[1]

Assistant Professor, Dept of Electronics University of Science and Technology, Meghalaya, India[2]

ABSTRACT: Emotions play important roles in expressing feelings as it tend to make people acts differently. Determining emotions of the speaker is less complicated if we are facing him/her rather than from voice independently such as conversation in telephone. However it would be a great achievement if we able to detect with what emotion a speaker is speaking just by listening to the voice. This project is a small step towards it and we basically are focusing on determining emotions through the recorded speech and developing the prototype system. The ability to detect human emotion from their speech is going to be a great addition in the field of human-robot interaction. The aim of the work is to build an emotion recognition system using Mel-frequency cepstral coefficients (MFCC) and Gaussian mixture model (GMM). In this work four emotional states happy, sad, angry and neutral are taken for classification of emotions. Here we have considered only 10 speakers, 7 male and 3 female, all belonging to upper Assam region and all speak in the same accent. The experiments are performed for only speaker dependent and text independent case.

KEYWORDS: Energy, GMM, MFCC

I. INTRODUCTION

Human beings are highly gifted as they can express their emotions through speech and facial expressions and can recognize others emotional status by listening to the voice or looking at the face. Most of the times we human beings detect the emotional status of other persons by listening his/her voice and looking at his/her facial expression at the same time. However the recognition of emotion from listening the voice is a very complex process which comes naturally to human and they don't need to practice it. Based on different emotions human beings utter in different ways and the characteristics of speech changes accordingly. Also ways of expressing emotions by people in different parts of the world are different from others.

In this paper we have taken the Assamese language for consideration. Assamese language is the mother tongue of the state Assam, however, in Assam a lot of different tribes of people live (Missing, Deuri, Lalung, Kachari, Ahom, Chutia etc.) and all speak in different accent. Also the spoken language differs from upper Assam to lower Assam. This work is aimed at recognizing emotion from Assamese language spoken at the northern part of the Assam.

Considerable past work has been done in this field and many researchers have published their results stating that they have got a good recognition rate by using different features and different classification methods. However, selection of proper features and use of suitable classifier is the most important issue to be considered for a better accuracy.

In this work a system will be developed using Gaussian Mixture Model. Initially, when people speak, a microphone converts the analog signal of our voice into digital chunks of data that the system must analyse. It is from this data that the computer must extract enough information. Thus, the data that being extracted will process as an input to the classifier. As a conclusion, this system will determine the speaker emotion through speech. This kind of nonverbal information could open the way to new paradigms in human-computer interactions.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

II. RELATED WORK

Researchers have proposed various approaches for emotion recognition in the literature. To list a few of them, Mixed modelling of human's emotional states from speech and visual information has been used as well.A.B. Kandali and his team in their work on Assamese language found that the choice of delta-Log-energy and delta-delta Log-energy features do not improve the recognition score. So, the feature set they choose for further study consisted of 1 Log-energy, 14 MFCC, 14delta-MFCC, and 14delta-delta-MFCC. After exhaustive trials of experiment with GMM classifier for M[6, 12], the highest mean classification score rose to 74.4%. They also found that the surprise emotion is the most difficult one to disambiguate from other emotions, since surprise may be expressed along with any other emotion such as angry-surprise, fear-surprise, happy-surprise, etc [2]. Perusing used eight features chosen by a feature selection algorithm developed a real time emotion recognizer for call centre applications. He achieved 77% classification accuracy in two emotions ("agitation" and "calm") using neural networks. [3] Joy Nicholson and his team gained an accuracy rate of approximately 50% while developing a speaker and context independent system for recognition of emotion in speech using neural networks. They have designed and implemented a one-class-in-one network for emotion recognition [7]. Y. Pan found that the system that uses both spectral and prosodic features for recognition of emotion achieves better recognition rate than that only uses spectrum or prosodic features. More distinctly they have showed that, the recognition rate of the system which uses the combination of features MFCC, LPCC, MEDC, energy and pitch is a bit less than the systems that only use energy, pitch MFCC and MEDC features [1]

III. CORPUS BUILDING

There are a few speech databases available from the recent works. There are six available databases which we can refer: two of them are publicly available, namely the Danish Emotional Speech corpus (DES) and Berlin Emotional Database (EMO-DB), and the remaining four databases are from the Interface project with Spanish, Slovenian, French and English emotional speech. Work on Japanese, English and Assamese database is also been done in recent past. The speech database is of two types. Acted and real life data. As the name suggests itself, in an acted emotional speech corpus, a professional actor is asked to speak in a certain emotion, whereas in real databases, speech databases for each emotion are obtained by recording conversations in real-life situations such as call centres and talk shows. But it has been observed that there is a difference in the features of acted and real emotional speeches. This is because acted emotions are not felt while speaking and thus come out more strongly. However collecting real life data for emotion is not easy and it's also not helpful in training and testing due to noise and other disturbances [1]. Although several works have been done in recognition of emotion, not much work has been done in Assamese language. So no existing corpus was found for Assamese language. Hence we had to build a new corpus. For that purpose data collection was carried out at Assam Don Bosco University, Guwahati, Assam. A headphone-mic, a Laptop, the Virtual DJ software were used for single channel recording from 7 male and 3 female speakers of Assamese language in closed-room and noise-free environment, the speakers were actors with experience in different dramas. Each speaker was asked to utter 3 different sentences in the same emotion, and was repeated for all the four emotions happy, sad, angry and neutral. The recorded speech signals were sampled at 16 Khz and in the format of 1 channel 16 bit mono. Total 120 samples were collected for training and testing purpose.

IV. EXPERIMENTS

We have performed one experiment for the speaker dependent and text independent case. A total of 40 samples have been trained of 10 speakers in 4 different emotions using the [mu,sigma,c] function that is described in MATLAB. 11 samples of different text have been tested with each single trained sample. For the training purpose 8 components have been taken for the GMM. The trained GMM is tested with the features extracted from the test utterances. The total log likelihood of test vectors of one test utterance with respect to the trained GMM corresponding to each emotion-class is computed. The test utterance is considered to belong to that emotion-class with respect to which the total log-likelihood becomes the largest. Finally average of all the speakers' accuracy is calculated.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

V. RESEARCH APPROACH

Detecting emotion from speech can be viewed as a classification task. It consists of assigning, out of a fixed set, an emotion category e.g., sad, anger, sadness, & neutral, to a speech utterance. The system consists of four major parts as shown in fig: 1. The proposed system is described as

I. Speech Acquisition from speech recorded in Assamese language.
II. Feature Extraction using Mel-Frequency Cepstrum Coefficients (MFCC) will be used for the baseline system
III. Development of Gaussian mixture models (GMMs) based baseline emotion detection system for speaker independent and context dependent database.
IV. Study on different speech features will be conducted to find the best set of features that captures the maximum emotion related information of the speech signal.
V. Develop an enhanced system based on the above evaluation.

VI. FEATURE EXTRACTION USING MFCC

From the different past works it is learned that different features are extracted from speech signal and different results are obtained. Feature selection is always an important task and in this work we have opted for extracting Mel Frequency Cepstrum Co-efficient (MFCC) feature for training. Each speech signal is segmented in frames of 20ms, and the window analysis is shifted by 10ms. Each frame is converted to 12 MFCCs plus a normalized energy parameter. The first and second derivatives (D's and DD's) of MFCCs and energy are estimated. A total of 39-dimensional features have been calculated. For the purpose of extraction the MATLAB v-7.8 has been used along with VOICEBOX. For extracting the features, steps shown in fig: 2 are being followed.

VII. CLASSIFICATION USING GMM

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. At first the GMM is trained with all the speech signals individually for each speaker, then the GMM is tested with features extracted from the speech signals. The total log likelihood of each test speech is computed with respect to the trained GMM. The tested speech is considered to belong to that emotional state in which the total log-likelihood is the largest.

VIII. RESULTS AND CONCLUSION

In this work, MFCC features are extracted from the speech signal to perform emotion recognition. A total of 39 features are extracted and the GMMs are constructed using 8 components. Various emotion recognition systems (ERS) are developed for different speakers and tested with the speaker dependent data. For each speaker one model is built in each of the four emotions (angry, happy, sad and neutral) and tested with all the sentences uttered by that particular speaker in all the emotions. The recognition rate is calculated for each emotion and finally the recognition rate of the complete system is evaluated. Evaluated results are shown in Table1, Table2 and Table3.

From the experiments we learned that recognition rate is better for male speakers as compared with female speakers. From the above experiments it is also been noticed that differentiating between the sad and neutral emotion is bit difficult. It may be because of the tones in which sad emotion is expressed is somewhat similar with emotion of type neutral, as they mostly have the same types of features. The recognition rate achieved in the work is not fixed for Assamese language. In this work the database taken is small consisting of only 120 sentences from 10 different speakers. Recognition rate can be higher than what is achieved in this work if we work with a larger database.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

Table1: Recognition Rate for Female Speakers

Female	Recognition Rate
Angry	44.44
Happy	44.44
Sad	44.44
Neutral	33.33

Recognition rate for female speaker=41.66%

Table2: Recognition Rate for Male Speakers

Female	Recognition Rate
Angry	38.09
Happy	52.37
Sad	47.61
Neutral	38.09

Recognition rate for male speaker = 44.04%

Table3: Rate of recognition of the system

Female	Recognition Rate
Angry	40.02
Happy	50
Sad	46.67
Neutral	36.67

Recognition rate of the system = 43.385%

REFERENCES

[1] Gogoi N.J., Das P. "EMOTION RECOGNITION FROM SPEECH SIGNAL: REALIZATION AND AVAILABLE TECHNIQUES" International Journal of Engineering Science and Technology (IJEST) Vol. 6 No.5 May 2014, 188-191.
[2] Kandali A.B., Routray A., and Basu T.K. "Emotion recognition from Assamese speeches using MFCC features and GMM classifier."TENCON 2008 - 2008 IEEE Region 10 Conference, 2008 , Page(s): 1 - 5

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

[3] Meshram A.P., Shirbahadurkar S.D., Kohok A., Jadhav S. "An Overview and Preparation for Recognition of Emotion from Speech Signal with Multi Modal Fusion" The 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010 (Volume:5), Page(s): 446 – 452

[4] Rao K. S., Kumar T. P., Anusha K., Leela B., Bhavana I. and Gowtham S.V.S.K. "Emotion Recognition from Speech" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012, pages: 3603-3607.

[5] Wang J., Han Z., Lun S., "Speech emotion recognition system based on genetic algorithm and neural network" in Image Analysis and Signal Processing (IASP), 2011 International Conference on 21-23 Oct. 2011, pages 578-582.

[6] Dellaert, F. ,Polzin, T. Waibel, A. "Recognizing emotion in speech in Spoken Language", 1996. ICSLP 96. Proceedings, Fourth International Conference on 3-6 Oct 1996.vol.3 Page(s):1970 – 1973

[7] Nicholson, J., Takahashi, K., Nakatsu, R., "Emotion recognition in speech using neural networks" in Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on 1999.vol.2 Page(s): 495 – 501

[8] Nicholson, J., Takahashi, K., Nakatsu, R., "Emotion recognition and its application to computer agents with spontaneous interactive capabilities" in Multimedia Signal Processing, 1999 IEEE 3rd Workshop on 1999, Page(s): 439 – 444.

[9] Xiao Z., Dellandrea E., 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. Page(s): 411 – 416

[10] Joshi D.D., Zalte M.B. "Speech Emotion Recognition: A Review" IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) ISSN: 2278-2834, ISBN: 2278-8735. Volume 4, Issue 4 (Jan. - Feb. 2013), PP 34-37

[11] Park J.S., Kim J. and Oh Y.H. "Feature Vector Classification based Speech Emotion Recognition for Service Robots" IEEE Transactions on Consumer Electronics, Vol. 55, No. 3, AUGUST 2009

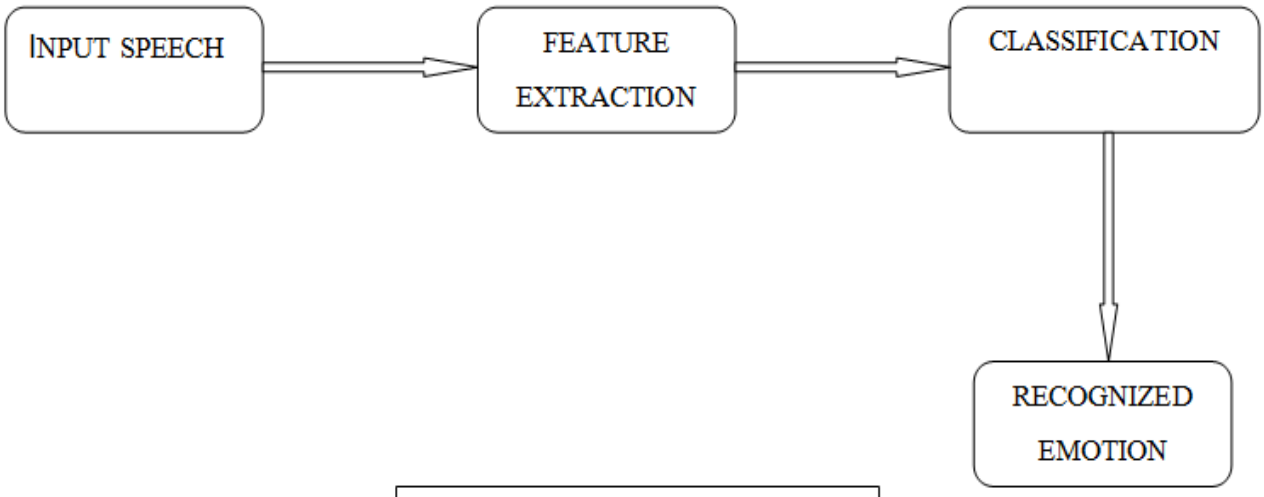

Figure1: System workflow

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

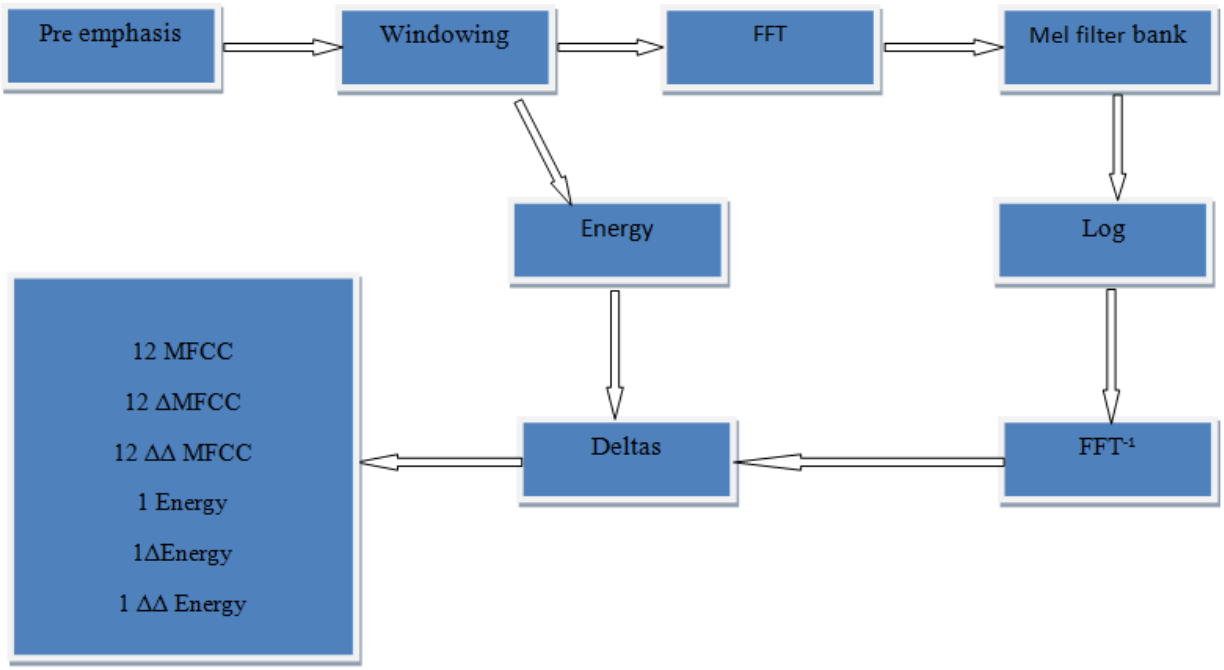

Figure2: MFCC blocks