Emotion Recognition from Isolated Marathi Speech using Energy and Formants

Rani Prakash Gadhe
Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004 (MS), India.

Ratndeep R. Deshmukh
Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004 (MS), India.

ABSTRACT
Speech emotion recognition is one of the interested research topics in today's time. Presently there are many attempts have been made for emotions recognition. The speech features such as energy and formants are extracted from speech. Angry, stress, admiration, teasing and shocking, these emotional states have been recognized on the basis of speech feature using K-Nearest Neighbor (K-NN) as a classification technique.

General Terms
Speech signals, Emotion Recognition, Speech Emotion Database.

Keywords
Human-Computer interaction, Speech, Emotion, Speech Emotion Recognition, Speech Database.

1. INTRODUCTION
Emotion plays an extremely important role in human's life for expressing mind state. Humans express their emotion via written and speech using different language. Enabling systems to interpret user utterances for human machine interaction therefore suggests also understanding transmitted emotional aspects. The user defined emotions can help the system to track the accurate behavior of user and also the mind state. Generally emotions recognition from speech is in the scope of research for the human-machine-interaction. In today's world attempting emotions from speech in detecting emotional speech analyze in general signal characteristics. In literature review many attempts done on speech emotion recognition system for identifying the proper and accurate emotions. Researchers have been using different techniques and this technique includes some of the feature selection and extraction methods. These systems used various features like Prosodic and spectral features where prosodic features included Pitch, Speech intensity glottal parameters and Spectral features included Mel-frequency cepstral coefficients (MFCC) and Linear Predictive cepstral coefficients LPCC [1].

Different classifiers have been used for speech emotion recognition some of this include Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Artificial Neural Network(ANN). A study using the features Energy and formants and classifier KNN performed the accuracy rate based on both results.

The paper is as follows: Section II describes database where Marathi dataset is developed and some of the samples are added from database. Section III defined feature extraction for speech features energy and formants. Section IV defined KNN classifier and the experimental results are evaluated which concludes the paper.

2. DATABASE
Dataset used in this paper is Marathi developed database. This database is developed by our self. All the samples are simulated by 100 speakers and each word was recorded thrice. The database consists of five emotions each having 8 words totally about 40 words. The database was recorded in noisy environment using PRAAT software. The frequency was set 16000HZ.

Table 1(a). Some of the samples of Angry emotion

Angry Emotion	English
समजताय काय स्वत:ला	What do you think of yourself
चल निघ	Get out
येड लागल का	Are you mad

Table 1(b). Some of the samples of Stress emotion

Stress Emotion	English
आता काय	Now what
नशिबच् फुटक्	Bad luck
मिच् का?	Why me?

Table 1(c). Some of the samples of Admiration emotion

Admiration Emotion	English
वा वा	wow
किती सुंदर	How beautiful
जब्बरदस्त	Fantastic

3. FEATURE EXTRACTION
3.1 Energy
Energy of the Discrete Speech Signal the amplitude of unvoiced segments is noticeably lower than that of the voiced segments. The short time energy of speech signals reflects the amplitude variation. In a typical speech signal we can see that its certain properties considerably changes with time. For example, we can observe a significant variation in the peak amplitude of the signal and a considerable variation of fundamental frequency within voiced regions in a speech signal. These facts suggest that simple time domain processing techniques should be capable of providing useful information of signal features, such as intensity, excitation mode, pitch, and possibly even vocal tract parameters, such as formant frequencies. Most of the short time processing techniques that give time domain features (Qn), can be mathematically represented as

$$Qn= \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m) \quad (1)$$

m=-∞ where T[] is the transformation matrix which may be either linear or nonlinear, X(m) represents the data sequence and W(n-m) represents a limited time window sequence. The energy of the discrete time signal is defined as

$$E = \sum_{m=-\infty}^{\infty} X^2(m) \quad (2)$$

Such a quantity has little meaning or utility for speech since it gives little information about time dependent properties of speech signal. In particular, the amplitude range of the unvoiced signal is generally much lower than amplitude of the voiced signal. The short time energy of the speech signal provides the amplitude variation which can be defined as

$$En= \sum_{m=-\infty}^{\infty} [x(m)W(n-m)]^2 \quad (3)$$

The major significance of En is that it provides a basis for distinguishing voiced speech signal from unvoiced speech signal. The energy function can also be used to locate approximately the time at which voiced speech become unvoiced speech and vice versa, and for high quality speech (high signal to noise ratio) the energy can be used to distinguish speech from silence. In voiced speech the short-time energy values are much higher than in unvoiced speech, which has a higher zero crossing rate [2, 3].

3.2 Formant

Formant is one of the important parameter which reflects the sound track features. Firstly linear prediction is applied on the signal to calculate the prediction coefficients, and then, those coefficients are used for estimating the sound track of frequency response curve, and the peak method is adopted to calculate the frequency of every formant. Both average formant frequency changing rate of the first formant [4].

The Linear predictive coding technique (LPC) has been used for the estimation of the formant frequencies. The analog signal is converted in .wav file digital format. The signal is transformed to frequency domain using FFT and the power spectrum is further calculated. Then the signal is passed through a Linear Predictive Filter (LPC) with 11 coefficients and the absolute values are considered. The roots of the polynomial are obtained which contain both real and imaginary parts. The phase spectrum is further displayed which clearly shows the formant frequencies [5].

4. KNN CLASSIFIER

In most of the pattern recognition system, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. Testing sample is classified on the basis of nearest distance which is calculated with training dataset. In both cases, the input consists of the k closest training examples in the future space. The output depends on whether k-NN is used for classification or regression [6].

At the classification phase, k is a user-defined constant value, and unlabeled vector is created for classifying the label which is most frequently present among the k training samples which is nearest to that query point. A commonly used distance metric for continuous variables is Euclidean distance. In our research, we use the system to extract the speech's feature like, pitch energy and formant. After the completion of speech feature extraction, we provide each speech sample to the corresponding emotion class label. After that the output of feature classification is provide as input to the K-NN classifier and result class is displayed which classifies the input sample in one of the five emotions.

Accuracy of the speech features is calculated on the basis of following formula.

Accuracy = (Correctly classified samples/Total number of samples) X 100.

The following table 5.1 displays the approximately achieved accuracy for Angry, stress, admiration, teasing and shocking using pitch and formants as discriminate factor.

Table 2. Accuracy result

Emotion	Energy	Formant
Angry	90%	100%
Stress	70%	90%
Admiration	100%	100%
Teasing	80%	90%
Shocking	90%	100%

5. CONCLUSION

In this paper we studied speech features for finding out the emotional state of speaker. On the basis of accuracy result we can defined that using energy and formant angry can be 90% and 100% recognized where as stress gives the lowest accuracy that is of a 70% accuracy with energy. Admiration, teasing and shocking gives 100% 80% and 90 % of accuracy with energy. Whereas using formant we get 100%, 90% and 100% for the same emotions respectively. Thus formant plays a vital role in recognizing emotions and we can conclude that using formant we can more accurately recognize emotions as compare to that of energy.

6. ACKNOWLEDGMENTS

This work is supported by University Grants Commission as Major Research Project Emotion recognition from Isolated Marathi Speech using Energy and Formants. The authors would like to thank the University Authorities for providing the infrastructure to carry out the research.

7. REFERENCES

[1] Shaikh Nilofer R. A.,Rani P. Gadhe, R. R. Deshmukh, V. B. Waghmare, P. P. Shrishrimal ," Automatic emotion recognition from speech signals: A Review," ISSN 2229-5518,International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015

[2] Vishal B. Waghmare, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, Ganesh B. Janvale,"Emotion Recognition System from Artificial Marathi Speech using MFCC and LDA Techniques ," Proc. of Int. Conf. on Advances in

Communication, Network, and Computing, CNC, Elsevier, 2014.

[3] D.S.Shete , Prof. S.B. Patil ,Prof. S.B. Patil," Zero crossing rate and energy of the Speech Signal of Devanagari Script," IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 4, Issue 1, Ver. I, PP 01-05, Jan. 2014.

[4] Xin min Cheng ,Pei ying Cheng, Li Zhao," A Study on Emotional Feature Analysis and Recognition in Speech Signal," International Conference on Measuring Technology and Mechatronics Automation, 978-0-7695-3583-8/09 © 2009 IEEE DOI 10.1109/ICMTMA.2009.

[5] Bageshree V. Sathe-Pathak, Ashish R. Panat" Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person," ISSN (Online): 1694-0814, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.

[6] Anuja Bombatkar, Gayatri Bhoyar, Khushbu Morjani, Shalaka Gautam,Vikas Gupta," Emotion recognition using Speech Processing Using k-nearest neighbor algorithm," International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, April 2014.