Comparative Analysis of Pitch and Formant for Recognizing Emotions of Isolated Marathi Speech

Shaikh Nilofer R.A.
Mtech Student,
Department of Computer Science & IT, Aurangabad (MS), India

R.R. Deshmukh
Professor,
Department of Computer Science & IT, Aurangabad (MS), India

V.B. Waghmare
Research Student,
Department of Computer Science & IT, Aurangabad (MS), India

ABSTRACT
Recognizing emotions from speech is a tuff task as we are not aware of the features which will accurately classify the emotions. This paper is an approach to show which speech feature classifies the emotions more accurately. The features compared here are Pitch and Formant while the classifier used is Linear Discriminant Analysis (LDA). The database used in this experiment was developed using 50 male and 50 female Marathi speaking native speakers. The emotions used here are Neutral, Happy, Sad, Surprise and Boredom. At the end of the experiment it was observed that formant recognized the emotions very efficiently and accurately with respect to that of energy.

General Terms
Emotion, Marathi database, formant, Pitch, Speech.

Keywords
Linear Discriminant Analysis (LDA), Emotion Recognition, Human Computer Interaction (HCI), Feature extraction.

1. INTRODUCTION
Though Emotion Recognition from speech now is not a new field for the researchers, many layers of it are yet to be open. At present vigorous research is going on detecting emotions by various means, one such mean is speech. For detecting the emotional state of a person machine requires some intelligence level. Emotion is not only associated with ones feeling but also is associated with behavior and state of mind. Emotion detection is a process through which the real emotion can be recognized During human to human interaction it becomes very easy to recognize the emotion but when the same is to be done by a machine problem arises.

Thus a lot of research had been done to solve this problem and many of them got succeed also, still one problem exist and that is of recognizing boundaries between emotion having same feature values. Emotion can be detected by extracting features from the samples of the speech corpus. Speech corpus contains some audio recording of sentences by the human that represents different kind of basic emotion like sad, boredom, happy, surprise, neutral .There are two type of information that can be conceived from emotional speech. First is related with linguist part which refers to the acceptance of all rules of pronunciation whereas second information concludes emotional state of speaker. Extracted features from emotional speech represent emotional information. Features can be classified as - spectral and prosodic features. Prosodic features are energy, pitch, intensity etc whereas spectral features are Mel frequency cepstrum coefficient, Linear prediction coefficient etc Speech emotion detection has plenty of applications in Call centers, E-commerce as customer satisfaction is to be done at customer care centers in both the cases, E – learning to detect the emotional state of learner so as to improve their presenting skills, Medical Field as a psychiatric diagnosis.[1]

The paper is as follows: Section II describes the features to be extracted briefly. Section III explains the classification method and finally the experimental results are evaluated which concludes the paper.

2. DATABASE AND DATA COLLECTION
Dataset used in this paper is a Marathi language based simulated database. The Headset used for recording was Snnehiser PC 360 as it has Noise Cancellation Facility. The database was recorded in noisy environment using PRAAT software. The frequency was set 16000HZ. All the samples are simulated by 100 speakers 50 male and 50 female native Marathi speaking speaker and each word was recorded thrice thus in total 12000 samples were considered. The database consists of five emotions each having 8 words totally about 40 words.

Table 1(a) some of the samples of Neutral emotion

Neutral Emotion	English
मग काय	What now
बोला आता	Say now
बरं बरं	Ok ok

Table 1(b) some of the samples of Happy emotion

Happy Emotion	English
किती छान	How nice
लय भारी	Awesome
अभिनंदन	Congrats

Table 1(c) some of the samples of Sad emotion

Sad Emotion	English
अरे देवा	Oh my god
काय करु	What to do
नकोये मला	I don't want

3. Feature Extraction

The features are extracted from a database of 40 words that is eight words of each emotion. We have extracted pitch and formant of each speech sample. Pitch is a prosodic feature whereas Formant is a spectral feature.

3.1 Pitch

The Pitch is the fundamental frequency of the vocal cords vibration (also called F 0) followed by 4-5 Formants (F 1 - F5) at higher frequencies. Male ~ 85-155 Hz; female ~ 165-255 Hz;Normally, pitch detection algorithms use short-term analysis techniques. For every frame xm, f (T|xm) that is am function of the candidate pitch periods is obtained

T. Algorithm determine the optimal pitch by maximizing (1)

$$T_m = \underset{T}{\operatorname{argmax}} f(T \mid x_m) \qquad (1)$$

A commonly used method to estimate pitch is based on detecting the highest value of the autocorrelation function in the region of interest. Given a discrete time signal x(n), defined for all n, the auto-correlation function is generally defined in (2):

$$R_x(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m) \qquad (2)$$

The autocorrelation function of a signal is basically a (non-invertible) transformation of the signal that is useful for displaying structure in the waveform. Thus, for pitch detection, if x(n) is exactly periodic with period P, i.e., x(n) = x(n + P) for all n, then it is easily shown that:

$$R_x(m) = R_x(m + P), \qquad (3)$$

i.e., the autocorrelation is also periodic with the same period. Conversely, periodicity in the autocorrelation function indicates periodicity in the signal. For a non stationary signal, such as speech, the concept of a long-time autocorrelation measurement as given by (2) is not really meaningful. Thus, it is reasonable to define a short-time autocorrelation function, which operates on short segments of the signal as:

$$R_x(m) = \frac{1}{N} \sum_{n=0}^{N'-1} [x(n+l)w(n)][x(n+l+m)w(n+m)],$$
$$o \le m \le M_0 \qquad (4)$$

where w(n) is an appropriate window for analysis, N is the section length being analyzed, N' is the number of signal samples used in the computation of R(m), Mo is the number of autocorrelation points to be computed, and l is the index of the starting sample of the frame. For pitch detection applications N' is generally set to the value in (5):

$$N' = N - m \qquad (5)$$

So that only the N samples in the analysis frame (i.e., x(l), x(l+1), . . . , x(l + N - 1)) are used in the autocorrelation computation. Values of 200 and 300 have generally been used for Mo and N, respectively, it is corresponding to a maximum pitch period of 20 ms (200 samples at a 10 kHz sampling rate) and a 30 ms analysis frame size.[2][3][4]

3.2 Formant

Formants are vocal tract features also known as spectral features. From the literature it is observed that vocal tract features are fairly reflected in frequency domain. Other vocal tract features include Mel Frequency Cepstral Coefficient (MFCC), Linear Perceptual Coding (LPC), Perceptual Linear Prediction coefficient (PLPC). All these features are extracted from a speech segment of length 20-30ms. Out of which formants are the widely extracted one. [6] In our work we have used PRAAT software for extracting the formants from the speech samples. Using PRAAT we can extract formants ranging from F1 to F5 as per our requirement.

4. Classifier – LDA

The linear discriminate analysis (LDA) is a statistical technique to classify data into mutually exclusive and exhaustive groups based on a set of measurable data's features. It is a well-known scheme for feature extraction and dimension reduction and has been used widely in many applications involving high dimensional data, such as pattern recognition, supervised learning and data classification. In LDA, the directions that will give a good separation to the different classes of the data need to be located. This is attained by projecting the data onto a lower-dimensional vector space so that the ratios of the between class distance to the within class distance is maximized, in order to achieve maximum separation. Here, if the data contains only two features, the separators between data groups will become lines. If there are three features, the separator will become a plane and if the number of features is more than three, the separator will become a hyper-plane.

The process of LDA is summarized below:

1. Calculating the within class scatter matrix. The amount of scatter between training data in the same class is first calculated using Equation (1) where Si is scatter matrix, mi is the mean of the training data xi within the class i and Xi is the covariance matrix of n the data. The within class scatter matrix Sw is then calculated as the sum of all the scatter matrices as shown in Equation (2) where C is the number of classes. A

$$S_i = \sum_{x\in X_i} (x - m_i)\,(x - m_i)^{-1} \qquad (1)$$

$$S_w = \sum_{i=1}^{C} S_i \qquad (2)$$

2. Calculating the between class scatter matrix. The amount of scatter between classes Sb is measured using Equation (3) where ni is the number of data in the ith class, m is the total mean of all training data, mi is the mean value of each class and C is the number of classes.

$$S_b = \sum_{i=1}^{C} n_i (m_i - m)(m_i - m)^{-1} \qquad (3)$$

3. Calculating the generalized eigenvectors and eigen values. The generalized eigenvectors and Eigen values of the within class and between class scatter matrices are computed.

4. Sorting the order of eigenvectors. The eigenvectors are sorted in a descending order depending on the magnitude of the Eigen values. Here, the sorted eigenvectors form the Fisher basis vector.

5. Projection of the training data onto the Fisher basis vectors. The training data are projected onto the Fisher basis vector by calculating the dot product of the training data with each of the Fisher basis vectors. Important feature is energy of speech signal. Speech energy is having more information about emotion in speech

5. EXPERIMENTAL RESULTS AND ANALYSIS

Detecting emotions of, neutral, happy, sad surprise and boredom from the speech signal was the main motive of this work. Of all the system modules the database played a vital role [6].

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

The following table 1 displays the approximately achieved accuracy for neutral, happy, sad, surprise and boredom using pitch and formants as discriminate factor.

Table 2 Accuracy result

Emotion	Pitch	Formant
Neutral	40%	90%
Happy	100%	90%
Sad	70%	80%
Surprise	90%	90%
Boredom	70%	70%

6. CONCLUSION

During this study we found that using pitch and formant happy can be 100 and 90% recognized where as neutral gives the lowest accuracy that is of 40% accuracy with pitch. Surprise, sad and boredom gives 90 70 and 70 % of accuracy with pitch. Whereas using formant we get 90, 80 and 70 for the same. Thus formant plays a very important role in recognizing emotions and we can say that using formant we can recognize emotions effectively compare to that of pitch. This study can be useful for various purposes lie detection tests, Call Centers working in Marathi language. Also it will give direction to the researchers willing to work in this field. Further number of samples can be increased, more features can be extracted, and multiple classification techniques can be applied on it in order to check the robustness of the system.

7. ACKNOWLEDGEMENT

This work is supported by University Grants Commission as Major Research Project as Isolated Marathi speech emotion recognition based on prosodic features. The authors would like to thank the University Authorities for providing the infrastructure to carry out the research.

8. REFERENCES

[1] Reshma, Maninder, Amarbir Singh, "Speech emotion recognition by Gaussian mixture model", International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2969-2971

[2] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal. "A comparative performance study of several pitch detection algorithms". IEEE Transactions on Audio, Signal, and Speech Processing 24, 399-417 1976.

[3] M. M. Sondhi, "New methods of pitch extraction," IEEE Trans.Audio Electroacoust., vol. AU-16, pp. 262-266, June 1968.

[4] Yi Kechu, Tian Fu, Fu Qiang, "YU YIN XIN HAO CHU LI", China Machine Press, Beijing, 2000

[5] Shashidhar G. Koolagudi · K. Sreenivasa Rao," Emotion recognition from speech: a review", Int J Speech Technol (2012) 15:99–117 DOI 10.1007/s10772-011-9125-1

[6] Chih-Hsien Huang, Cheng-Hung Tsai and Bo-Yi Li"The Corpus Preparation and Effective Feature Representation of Emotional Speech," Fourth International Conference on Innovative Computing, Information and Control 2009.