International Journal of Computer and Communication Technology

Volume 2 | Issue 1 Article 6

January 2011

Emotion Recognition using Fuzzy K-Means from Oriya Speech

Sanghamitra Mohanty
Dept of Computer Sc. & Application Utkal University Bhubaneswar, Orissa, India,
sangham1@rediffmail.com

Basanta Kumar Swain
Dept of Computer Sc. & Application Utkal University Bhubaneswar, Orissa, India, technobks@yahoo.com

Emotion Recognition using Fuzzy K-Means from Oriya Speech

Sanghamitra Mohanty [1], Basanta Kumar Swain [2]
Dept of Computer Sc. & Application
Utkal University
Bhubaneswar, Orissa, India
sangham1@rediffmail.com[1] , technobks@yahoo.com [2]

***Abstract*—Communication will be intelligible when conveyed message is interpreted in right-minded. Unfortunately, the right-minded interpretation of communicated message is possible for human-human communication but it's laborious for human-machine communication. It is due to the inherently blending of non-verbal contents such as emotion in vocal communication which leads to difficulty in human-machine interaction. In this research paper we have performed experiment to recognize emotions like anger, sadness, astonish, fear, happiness and neutral using fuzzy K-Means algorithm from Oriya elicited speech collected from 35 Oriya speaking people aged between 22-58 years belonging to different provinces of Orissa. We have achieved the accuracy of 65.16% in recognizing above six mentioned emotions by incorporating mean pitch, first two formants, jitter, shimmer and energy as feature vectors for this research work. Emotion recognition has many vivid applications in different domains like call centers, spoken tutoring systems, spoken dialogue research, human-robotic interfaces etc.**

Keywords- Emotion recognition, pitch, formant, jitter, shimmer, energy, Fuzzy K-Means.

1. INTRODUCTION

Human beings are replaced by machines (computers) in real life working domain in order to accomplish the desired task in faster as well as in lower cost. As a result of which human interacts with machine through speech which is natural modus of communications of human in order to avail required service. In other hand machine should be enriched with the ability to perceive, interpret, express and regulate emotions present in conveyed speech message for user friendly interactions in between human and machine [1,2,3]. Therefore our ongoing scientific research is towards development of realistic interactions between human beings and computers so that computers are able to perceive and respond to human non-verbal communication such as emotions. Hence, emotion recognition from speech objective is to automatically identify the emotional state of human being from his or her voice. It is seen that speech signal gets distorted from emotion to emotion as speaker utters speech in different physical state of mind. This paves the way for getting one of solutions of emotional speech recognition by measuring acoustic parameters from distorted speech wave in different emotions in the form of feature vectors and applying pattern recognition technique to yield the desired solution [4,5].

The purpose of this research work is to recognize emotion in speech for Oriya language. We have studied six emotions categories namely anger, sadness, astonish, fear, happiness and neutral. For training the speech emotion recognition system we have used speech corpus consisting of more than 900 emotional speech texts incorporating all six mentioned emotions. People are allowed to utter emotional texts collected from various drama scripts of Oriya language as professional actors in order to create emotional Oriya speech corpus. Finally emotional state of speaker is identified by applying Fuzzy K-Means algorithm as pattern recognition technique by utilizing mean pitch, first two formants, jitter, shimmer and energy as features.
In this paper, experiments will be presented with detail. It is organized as follows: Section 2 deals with design of Oriya emotional speech database, section 3 describes feature extraction, section 4 demonstrates about Fuzzy K-Means algorithm, section 5 shows experimental results, section 6 presents conclusion and discussion.

2. DESIGN OF ORIYA EMOTIONAL SPEECH DATABASE

The Oriya emotional speech corpus used in this study was designed by joint effort of the faculty, PhD students, research scholars and Master students from department of Computer Science and Application at Utkal University of Orissa. The speech language is Oriya and spoken by 35 speakers aged between 22-58 years old having sound health and normal heights belong to different Oriya speaking areas of state Orissa. Speakers were allowed to read text fragments in great naturalness of speech with respect to different emotions. Rerecording was allowed for speaker to read emotional sentences if desired emotion is not delivered at the time of recording by monitoring recorded speech [6,7]. Text fragments were taken from various Oriya drama scripts like "MO PEHENKALI BAJAIDE"(/*mo pehenkAli bajAide/),* "E DUNIA CHIDIA KHANA"(/e duniA chidiA khAnA/),

"BIDESINI"(/bidesinI/), "MUN POCHHIDELI MO APA SINDURA"(/mun pochhideli mo apA sindura/), "PHULA BAULA BENI"(/phula baula beni/). The corpus was digitized at 16000Hz with A/D conversion precision of 16 bits with mono channel under laboratory environment which may contain noticeable noise. 90% of our speech corpus was used as training data set and remaining 10% was used for test sample for this research work. Detail of emotional Oriya speech corpus is given in Table 1.

TABLE1. DETAILS OF EMOTIONAL ORIYA SPEECH DATABASE

Items	Details
Emotion Classes	Anger, Sadness, Astonish, Fear, Happiness and Neutral
Total Phrases Average words	900 10
Speakers	Male- 23 Female-12 Age-22 to 58
Environment	Laboratory Environment
Speech Style	Reading text fragments
Transducer	Noise cancelation microphone
Sampling Rate	16000Hz,16bit,1-channel

3. FEATURE EXTRACTION

Feature extraction is a vital stage for emotion recognition as selection of appropriate features influence the recognition rate greatly. Feature extraction is a process to convert an observed speech signal (speech waveform) to some type of parametric representation for further analysis and processing. For emotion recognition by machine, the acoustic waveform must be transformed into an electrical signal by a microphone or a transducer. Then the feature extraction stage is carried out by digitizing the speech waveform at a rate of 16 kHz to produce a sampled waveform. We have used hamming window of 25ms with 10ms shifting for each consecutive window [8,9,10].

A set of features that we have incorporated in this work namely mean pitch, first and second formants, jitter, shimmer and energy.

3.1 Pitch

The time duration one glottal cycle is referred as pitch period and the reciprocal of the pitch period is the corresponding pitch, also called as the fundamental frequency. We have used autocorrelation function to calculate pitch as autocorrelation method is more accurate, nose-resistant and robust than other methods [11].

Consider a discrete-time short-time sequence given by

$$S_n(m)=S(m)w(n-m) \quad (1)$$

Where w(n) is an analysis window of duration N_W. The short-time autocorrelation function $R_n(k)$ is defined as

$$R_n(k)=\sum S_n(m)\, S(m+k) \quad (2)$$

If S(m) is periodic with period P then $R_n(k)$ contains peak at or near the pitch period P. The location of a peak in the pitch period range provides a pitch value.

3.2 Formant

Linear predictive coding is used to extract formant frequencies in speech. Linear prediction is an adequate all-pole model to voiced speech signals. Parameters of all pole models are representative of formant positions. A difference equation on solution of linear prediction expresses each sample of the original signal as a linear combination of preceding samples. This difference equation is the linear predictor and the coefficients of equation are the linear predictive coding (LPC) coefficients. The formant frequencies are estimated from peaks of the linear prediction spectra of speech signals. In this paper, the first and the second formant frequencies (F1andF2) are extracted [2,12].

3.3 Jitter Measurement

Fixed pitch period is never maintained in time but can randomly vary over successive periods, this characteristic is referred to as jitter [8]. In other words we can define jitter (absolute) is the cycle-to-cycle variation of fundamental frequency i.e. the average absolute difference between consecutive periods, expressed as:

$$Jitter(absolute)=\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i-T_{i+1}\right| \quad (3)$$

Where T_i is the extracted F0 period length and N is the number of extracted F0 periods.

3.4 Shimmer Measurement

The amplitude of the air flow velocity within a glottal cycle may differ across consecutive pitch periods, this characteristics is called shimmer [8]. Alternately we can define shimmer (dB) is expressed as the variability of the peak-to- peak amplitude in decibels, i.e. the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20:

$$Shimmer\,(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(A_{i+1}/A_i\right)\right| \quad (4)$$

3.5 Energy

It is experimentally found that amplitude of speech signal varies significantly with time and most cases amplitude of unvoiced segments is much lower than amplitude voiced segments. The short time energy of speech signal represents these amplitude variations [2,8]. Energy can be defined as

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 \quad (5)$$

Energy at sample n is the sum of squares of the N samples n-N+1 through n.

4. Fuzzy K-Means Algorithm

We have followed fuzzy k-means algorithm as pattern recognition technique for emotion recognition in this research study. Fuzzy k-means technique is based on fuzzy set theory [4,14,15] . Thus the fuzzy k-means problem is treated as the optimization technique for obtaining the membership values along with the centroids of the individual clusters such that

$$\sum_{i}^{n}\sum_{j}^{m} M_{ik}^{2} * (X_i - C_k)^2 \text{ is minimized} \quad (6)$$

M_{ik} is membership value of the vector X_i belongs in the cluster k. X_i is the i^{th} vector. C_k is the k^{th} centroid.

Fuzzy K-means method for emotion pattern recognition is described in following steps

Step1: Select the number of clusters (let this number be k).

Step2: Initialize the membership values (M_{ik}) randomly.

Step3: Compute the centroids of the k clusters.

C1= [Vector1* (the member ship value of the vector 1 belongs to the cluster 1)^2 (i.e.) M_{11}^2 + Vector2* (the member ship value of the vector 2 belongs to the cluster 1)^2 (i.e.) M_{12}^2 +… Vector n* (the member ship value of the vector n belongs to the cluster 1)2 (i.e.) M_{1n}^2] / Sum of the squared values of the membership values belonging to the cluster 1.
Similarly the centroids C2, C3, C4, …. Ck are obtained.

Step4: Update the membership values M using the current values of the k centroids as given below.

```
for k= 1 to m
    for i=1 to n
        for  r= 1 to m
            Mik=[((vector(i) – cent(k))^2/((vector(i)– cent(r))^2]^2
        endfor
    endfor
endfor
```

Step5: Compute the sum of the squared difference between the previous membership value and the current membership value.

If the computed value is not less than the threshold value go to step3 to compute the next set of centroid and followed by next set of membership values. If the threshold value is less than the threshold value, stop the iteration. Thus the centroids are obtained using fuzzy k-means algorithm.

Step6: Compute the Euclidean distance of each object in the dataset from each of the centroids.

$$d(x,y) = \sqrt{(\sum(x_i - y_i)^2)} \quad (7)$$

Step7: Allocate each object to the cluster it is nearest to based on the distances computed in the previous step.

Step8: Compute the centroids of the clusters by computing the means of the attribute values of the objects in each cluster.

Step9: Check if the stopping criterion has been meet. If yes, go to step Step10 else go to Step6.

Step10: Stop

We experimented fuzzy k-means pattern recognition technique by setting k value as 6. Block diagram of emotion recognition using fuzzy K-means from Oriya speech is shown in the Figure 1.

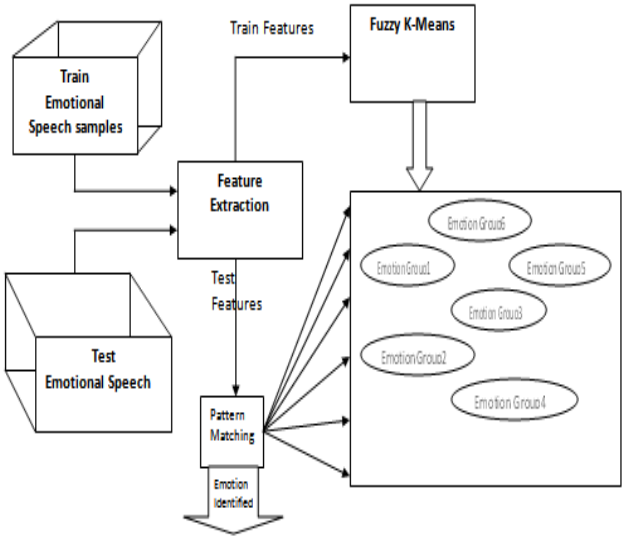

Figure 1. Block diagram of emotion recognition using Fuzzy K-Means from Oriya speech.

5. EXPERIMENTAL RESULTS

Human listener some time even fails to recognize emotion of speakers but when provided the recorded emotion speech they can easily recognize the emotion as listeners are well habituated with sentences during recording session in order to avoid that situation we jumbled together the emotional speech samples so that the listener will not get some prior knowledge while recognizing emotion. We have also used some new human listeners those are not selected for Oriya emotion speech corpus design stage to identify emotion of some emotional speech. We have used ten persons for identification of emotions. Confusion matrix for emotion recognition by human is shown in Table 2.

TABLE 2. CONFUSION MATRIX IN % BY HUMAN LISTENERS

Emotion	Anger	Astonish	Fear	Happiness	Sadness	Neutral
Anger	85.6%	0%	4.4%	0%	10%	0%
Astonish	0%	70%	5%	17%	5%	3%
Fear	3%	7%	77%	0%	10%	3%
Happiness	0%	16%	0%	84%	0%	0%
Sadness	0%	3%	11%	0%	66%	20%
Neutral	2%	3%	4%	6%	15%	70%

From the above experiment we found that human listeners can more accurately identify anger, happiness and astonish emotions than any other remaining emotions. The average emotion recognition accuracy is 75.43%.

TABLE 3. CONFUSION MATRIX IN % BY FUZZY K-MEANS

Emotion	Anger	Astonish	Fear	Happiness	Sadness	Neutral
Anger	72%	2.2%	4%	8%	13%	0.8%
Astonish	3%	80%	2%	13%	1%	1%
Fear	10.6%	1.2%	62%	2.3%	19.4%	4.5%
Happiness	8%	22%	0.5%	67%	0.5%	2%
Sadness	1%	2%	9%	1%	52%	35%
Neutral	0.5%	1%	0.5%	2%	38%	58%

From above confusion matrix drawn from Fuzzy K-Means technique we found that emotion astonish is well recognized as compared to anger followed by happiness. This pattern recognition technique recognizes the emotion sadness and neutral less in percentage in acoustic domain as compared to other emotions. The average emotion recognition accuracy is 65.16% even with lower percentage recognition rate of sadness and neutral emotion.

6. Conclusion and Discussion

We investigated this research study using both human points of view as well as using Fuzzy K-Means pattern recognition technique. The accuracy level for emotion recognition by human was 75.43% in average where as machine recognition based on Fuzzy K-Means technique was 65.16% in average. The reduction of recognition by machine is due to the following reasons:

- Human beings are trained with natural emotions from childhood which outperforms than machine which is trained using elicited emotions.
- Size of emotion database for training human neurons is much larger as compared to corpus used for training machine because human beings are getting free as well as natural emotions dataset from day to day life. Size of such a corpus for training is having greater effect on accuracy.
- Human beings are using linguistic knowledge as well as acoustic parameters while recognizing emotions from speech where as our technique is based on only acoustic parameters.

In this research we also found that sadness and neutral emotions are misclassified because the features like pitch and shimmer values are very close to each other for both emotions.

7. Acknowledgment

Financial support of DIT, MCIT Govt. of India for this work is highly acknowledged. All the people involved in speech corpus building work, for their kindness, their voice messages and time dedication.

References

[1] Mohanty, S., Bhattacharya, S., Bose, S., Swain, S., An Approach To Parametric based Mood Analysis In Oriya Speech Processing ,Proceedings of the International Symposium Frontiers of Research on Speech and Music(FRSM-2005),January 6- 7,2005,pp.105-109.

[2] Rabiner, L.R, Schafer, R.W, Digital Processing of Speech Signals, Pearson education, 1st Edition, 2004.

[3] S. Huang, L.S. Chen, and H. Tao, "Bimodal emotion recognition by man and machine," *ATR Workshop on Virtual Communication Environments,* Kyoto, Japan, Apr. 1998.

[4] Han.J, Kamber.M. Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, second edition, 2006.

[5] C. Espy-Wilson, "Articulatory strategies, speech acoustics and variability", From Sound to Sense, June, 2004.

[6] Becchetti,C, Ricotti,L,P, Speech Recognition Theory and C++ Implementation, Jhon Wiley & Sons,2009

[7] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. Speech Communication, 40, 33–60 (2003).

[8] Quatieri T.F., Discrete-Time Speech Signal Processing, Pearson education ,Third Impression-2007

[9] Shaughnessy, D, Speech Communications Human and Machine, Universities Press, 2nd Edition, 2001.

[10] Proakis, J.G, Manolakis, D.G. Digital Signal Processing, Pearson Education, 4th Edition, 2007.

[11] A. Protopapas and P. Lieberman, "Fundamental frequency of phonation and perceived emotional stress", *Journal of Acoustical Society of America*, v. 101, n. 4, pp. 2267-77, 1997

[12] Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combing acoustic features and linguistic information in hybrid support vector machine-belief network architecture," *Proc. ICASSP,* Montreal, Canada, pp. 577-580, 2004.

[13] G. Castellano, L. Kessous, "Multimodal emotion recognition from expressive faces, body gestures and speech", 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, September 2007N.

[14] Jackson. P, Introduction to Expert Systems, third edition, 2001.

[15] Sebe, N., Cohen, I., Huang, T.S.: Multimodal Emotion Recognition, Handbook of Pattern Recognition and Computer Vision, World Scientific, ISBN 981-256-105-6, January 2005.