AUTOMATIC EMOTION RECOGNITION FROM SPEECH SIGNAL

Surayya Ado Bala[1], KhushbooTaneja[2], Yahya Musa Shuaib[3], Rajiv Kumar[4]

[1,2,3,4]*Department of Computer Science and Engineering, Sharda University,Greater Noida, (India)*

ABSTRACT

This paper is an effort to recognize the emotion, from speech signals. The emotion recognition system uses vocal structural features of speech as features for emotion recognition. For finding emotions vocal expression of eight emotions (Happy, Angry, Sad, Depressed, Bored, Anxious, Fear and Nervous) were considered in Hausa language. In order to recognize emotions, features from these emotions are computed from acoustic pattern of the signals. For collecting features from emotions rhythm (duration, pitch and energy), enunciation (formant) and diction (ZCR) were considered. These features were finally combined and classified by using KNN classifier. The algorithm is tested on newly developed dataset for Hausa language based on audio signals. The recognition system is tested with 8000 (eight thousand) emotions collected from 100 (one hundred) different speakers with different age groups and genders.

Keywords: Emotion Recognition; Audio signal; Hausa language; Classification.

I. INTRODUCTION

Human beings expressed their emotions through speech. Audio Emotion Recognition from speech is a current research topic with wide range. Audio emotion recognition is a way to identify the emotional state of human from speech signals. Speech signals can rapidly deliver information or messages by human. Automatic speech emotion recognition is a challenging task in signal processing. Many paralinguistic properties exist like: gender, age, emotion, voice quality, stress and nervousness, dialect, pathological state, alcohol or drug consumption, charisma, just to mention a few. Among these properties, the emotion plays a key role in many applications like in call centers to detect angry customers [1-4], in entertainment electronics to gather emotional user feedbacks [5], in ASR to resolve linguistic ambiguities [6–8], and in text-to-speech systems to synthesize emotionally more natural speech [4-5]. Human being expressed their emotions through speech. Speech can act as an index for representing the emotions. In general the emotion is a reflection of uncertainties in real that’s why it is hard to define prototype for all real life emotions under one figure in psychological literature [1]. So knowledge discovery from speech signal need more attention. Rest of paper organized as section II represents related work, detailed methodology is discussed in section III, experimental results are discussed in section IV and section V concludes the paper.

II. RELATED WORK

A research is done on “Transformation of emotion based on intonation patterns for Hindi speech” by Agrawal et. al [3] in which they worked to transform neutral sentences into emotion rich sentences with an idea that changing an intonation pattern or structure of sentence can change the emotion associated with the sentence or phrase. And, for this fundamental frequency(f_0), energy contour were used as parameters to convert intonation emotion. The emotions under consideration during the experiment were: surprise, happiness, anger, sadness.

In 2013 , Bahugama and Raiwani [4] proposed a work over emo-voice model for speaker’s emotion identification using MFCC and vector quantization techniques in Hindi database for four basic emotions: Happy, sad , anger, and neutral. Recent trends in research of audio emotion recognition emphasized the use of combination of different features to achieve improvement in the recognition performance. System and prosodic features represent mostly mutually exclusive information of the speech signal. Therefore, these features are complementary in nature to each other. Combination of complementary features is expected to improve the intended performance of the system. In this work various features and techniques which were not been examined in previous researches for audio emotion recognition using developed Hausa data set that is not available in the public domain , been covered here in order to check their effects over recognition rate, for feature extraction like duration, zero crossing rate , pitch detection , formant estimation ,techniques: cepstral, LPC (linear prediction coefficient) for efficient and accurate estimation of results used in recognizing emotions.

The work done on Hindi speech signal in 2011 by Shashidhar et al. [7], emotions like anger, disgust, fear, happy, neutral, sad, sarcastic, and surprise are considered. These emotions are classified by using prosodic (energy, pitch, duration) and formant and Zero Crossing Rate (zcr). Further Chauhan et. al [8] proposed a text independent emotion recognition using spectral features over Hindi speech database achieved through IITKGP-SEHSC [7], in this the techniques used are mel frequency cepstral coefficient (MFCC) and Gaussian mixture model (GMM) for recognizing emotion: anger, happy, disgust, fear ,neutral, sarcastic and surprise,respectively.

In 2011, Warkhade et. al [9] proposed a speech emotion recognition system : using SVM and LIBS , where emotion database in Hindi and Berlin were used for analyzing the emotions.Prosodic features were extracted using MFCC and MEDC technique for emotional classification.

In 2012, Ahmad [10] proposed Transformation of emotions using pitch as parameter for Hindi speech”, for analysis of different emotion from neutral emotion based on pitch contour and proposed an algorithm for emotion conversion based on pitch factor involved simple rule for converting pitch points.

III. METHODOLOGY

The methodology used in this work is discussed below;

3.1 Proposed Model

The ultimate goal of system design should be its simplicity and efficiency. The Figure 3.2 shows architecture of a speech emotion recognition system. Generally the speech files are in .wav format. For proposed system .wav format files are used. The Hausa Emotion database contains files in .wav format.

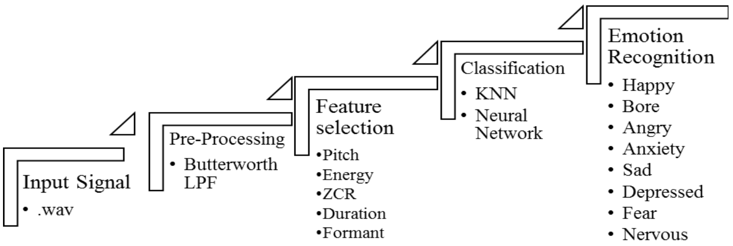

Fig. 1. Proposed model for Audio Emotion Recognition System

3.2 Data Collection

A suitable emotional speech database is an important requirement for any emotional recognition model. The quality of database determines the efficiency of the system. The emotional database may contain collection of acted speech or real data world.

Source data is collected from 100 non-professional speakers (include both genders), each person uttered the same sentence eight times in eight different emotions, which made a total of 8000 samples in Hausa language. The data was recorded through microphone using recording software. In order to reduce faulty collection, (same distance kept between microphone and source).

3.3 Preprocessing

For analyzing achieved speech signals, preprocessing of signal is required to extract errorless information regarding:

i. Duration of speech signals, zero crossing rate etc.

ii. Fundamental frequency variation, pitch variation using two different techniques (cepstral) in order to analyze difference.

iii. Formant frequency for detecting syllables which reflect changes in voice feature due to change in vocal tract shape on utterance of phrase with different emotions.

iv. Energy or amplitude variation within same phrase but with different emotions.

Here the signal is segmented in frames of 30ms and overlapping of 10ms between frames was made for interpolation to reduce the loss of information while processing of data. Then signals were filtered using butter worth low pass filter and after that framed signal was further divided into voiced and unvoiced signal as voiced frames differentiate themselves from unvoiced in signals by means of energy, zero crossing rate and classifying it favors accurate measurements.

TABLE I. **Literature Review Of Audio Emotion Recognition System**

S.No.	REF.	DATABASE	EMOTIONS	FEATURES EXTRACTED	CLASSIFIER	ACCURACY
1	Chieng et al. (2014)[1]	RML and Mandarin database, English	Anger, Fear,Happy, Sad,Surprise and Neutral	Pitch, ZCR ,MFCC ,Log energy and Teager Energy Operator	Hidden Marko Model and Support Vector Machine	84.2%
2	Milton, Roy and Selvi (2013)[2]	Berlin database	Anger, Boredom Happiness, Fear Sadness, Disgust Neutral	Spectral and prosodic	3 stage Support Vector Machine	68%
3	Agrawal, Prakash, and Jain (2010) [3]	Hindu Language	Surprise,Happiness,Anger and Sad	Fundamental frequency	Lineaer Modification Model	90.7%
4	Bahuguna et al (2013) [4]	English language	Angry, Disgust, Fear, Happy,Neutral, sarcastic and surprise	MFCC	KNN and SVM	89.2%
4	Panda et. al. [5]	Hindi language	Anger, Hapiness, Sadness,neutral and Fear	MFCC	Support Vector Mamchine(SVM)	93.75%
5	Shashidhar et al. (2011) [7]	Hindi Language	anger, disgust, fear, happy, neutral, sad, sarcastic, and surprise	Mel frequency cepstral coefficients (MFCCs), Energy, pitch and duration.	Support Vector Machine	77% male 81% female
6	Chauhan et. al [8]	Hindi Language	anger, happy, disgust, fear ,neutral, sarcastic and surprise,	mel frequency cepstral coefficient (MFCC) and Gaussian mixture model (GMM)		72% and 82% for text independent and dependent cases

3.4 Feature Extraction

As in previous researches related to emotion recognition in hindi speech didn’t include all acoustical features such as formants, zcr etc. which are also involved here in this dissertation work. The features extracted are shown in Fig. 2.

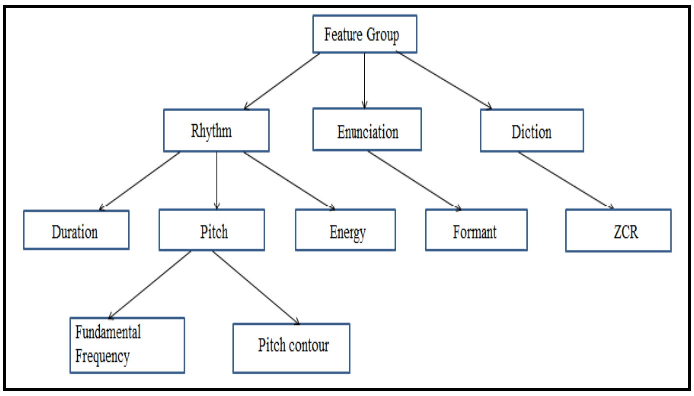

Fig. 2. Feature Extraction

3.4.1 Duration

Duration of signal gives time taken in speaking the phrase, which varies with number of pauses in the signal depend upon emotions associated with uttered phrase. This can be calculated as:

$$T = (N - \sum_{p=0}^{p=n} (P))/dt$$

Where,

T= duration of sample (sec)

N= length of sample

P=length of pause

dt= time rate

3.4.2 Zero Crosing Rate

This defines number of zero crossing per unit time, and can be measured using:

$$Z = n_c.(f/n)$$

where,

n_c = number of zero crossing per frame

Z=zero crossing rate per sample

f=sampling frequency (44100 hz)

n=length of frame (30ms)

3.4.3 Energy

Energy defines how much force been posed over a phrase while reciting it, it basically used to identify the intensity of signal, for maintaining accuracy, energy is calculated for short framed signal in order to avoid unvoiced part of signal. Energy can be calculated using:

$$E(x) = \sqrt[2]{\sum s_f^2}/_n$$

Where,

E = energy of sample

s= sample value of fth frame

n= frame length.

3.4.4 Pitch Detection

One of the essential acoustic feature for emotion recognition is pitch , defines the rate of vibration of human vocal chord while uttering a phrase. It has its various sub features as fundamental frequency, harmony, pitch contour. In this work, fundamental frequency contour ,f_0, pitch contour are taken into consideration. Pitch contour gives variation of pitch along the signal on utterance of a phrase with particular emotion .The fundamental frequency(f_0), that is first harmonic define as greatest common divisor of all harmonics, plays vital role for determination of pitch as it gives highest peak in a period of signal, therefore it often called pitch .Various algorithms have been proposed for reliable estimation of pitch but none of them are able to give accurate estimation of pitch. Commonly autocorrelation pitch detection algorithm is used for pitch estimation ,but in this work pitch been estimated through cepstral pitch determination algorithm, as this algorithm has advantages over autocorrelation based PDA,because in autocorrelation method effect of vocal tract and vocal source are convolved with each other but in cepstral approach they are independent and easily identifiable. For better results, voice signals are first filtered with butter worth low pass filter, and then separated in to voiced and unvoiced signals. Further,signals were segmented into short frames of 30ms with an overlapping of 10ms. It can be estimated as shown in Equation (1).

$$C(t) = F^{-1}(|\log(F(x(t))|^2) \quad \text{------}(1)$$

3.4.5 Formant

Formant frequencies has its significance in determining the phonetic content of speech signals as phonetic content cause peaks of vocal tract resonance in spectral envelop. Creaky and whispered vowels are rarely used. Several algorithms are in existence for finding formant frequencies such as, analysis by analysis with fourier spectra, peak picking on cepstrally smoothed spectra, linear prediction. However, linear prediction is best approach for determining formant frequencies. The first three formants (f1),f2, f3 are usually considered to favor clear and effective analysis.

3.5 Emotion Classification

For emotion recognition, classification of emotion is obvious and inevitable requirement to train and test the data. In order to classify each parameter feature such as standard deviation of fundamental frequency, standard deviation of energy and pitch contour, formant frequencies, duration, zero crossing result, KNN and Neural Network would be used. In pattern recognition, the k-Nearest Neighbors algorithm (or KNN for short) is a non-parametric method used for classification and regression In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether KNN is used for classification or regression. In KNN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In KNN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

3.6 Emotion Recognition

The audio emotion recognition system is implemented using MATLAB 2015a and the output are:

a) ANGRY

Angry is one of emotions that play an essential role in decision making, perception, learning and more. Anger or wrath is an intense emotional response. Often it indicates when one's basic boundaries are violated. Angry is one of emotions that play an essential role in decision making, perception, learning and more.

b) ANXIETY

Anxiety is an unpleasant state of inner turmoil, often accompanied by nervous behavior, such as pacing back and forth, somatic complaints and rumination.

c) BORED

Bore is an emotional state experienced when an individual is left without anything in particular to do, and not interested in their surroundings.

d) DEPRESSION

Depressed is a state of low mood and aversion to activity that can affect a person's thoughts, behavior, feelings and sense of well-being .

e) FEAR

Fear is an emotion induced by a threat perceived by living entities, which causes a change in brain and organ function and ultimately a change in behavior, such as running away, hiding or freezing from traumatic events.

f) HAPPY

Happiness is a mental or emotional state of well-being defined by positive or pleasant emotions ranging from contentment to intense joy. A variety of biological, psychological, religious, and philosophical approaches have striven to define happiness and identify its sources confusion matrix of 170 feature vector of each input and output sample.

g) NERVOUS

Nervous is irregularities in heart beat, parasympathetic co-activation and slightly altered behavior. Every sign originates from brain states.

h) SADNESS

Sad (also called heavy-heartedness) is emotional pain associated with, or characterized by feelings of disadvantage, loss, despair, helplessness, disappointment and sorrow.

IV. RESULT ANALYSIS

The performance of the proposed audio emotion recognition System is evaluated on Hausa databases. The figure below show the result analysis of the experiment. The experiments were conducted on 124 datasets from eight different emotions. The various feature extraction graphs are shown in Fig. 3. From figure it is clear that duration, ZCR, and Ferment gives better feature differentiation compared to others. Finally these feature vectors are combined together and further classified by KNN classifiers. The final confusion matrix is shown in Table II. From this table it is clear that Happy emotions has given better results whereas Bored and fear emotions performed poor results.

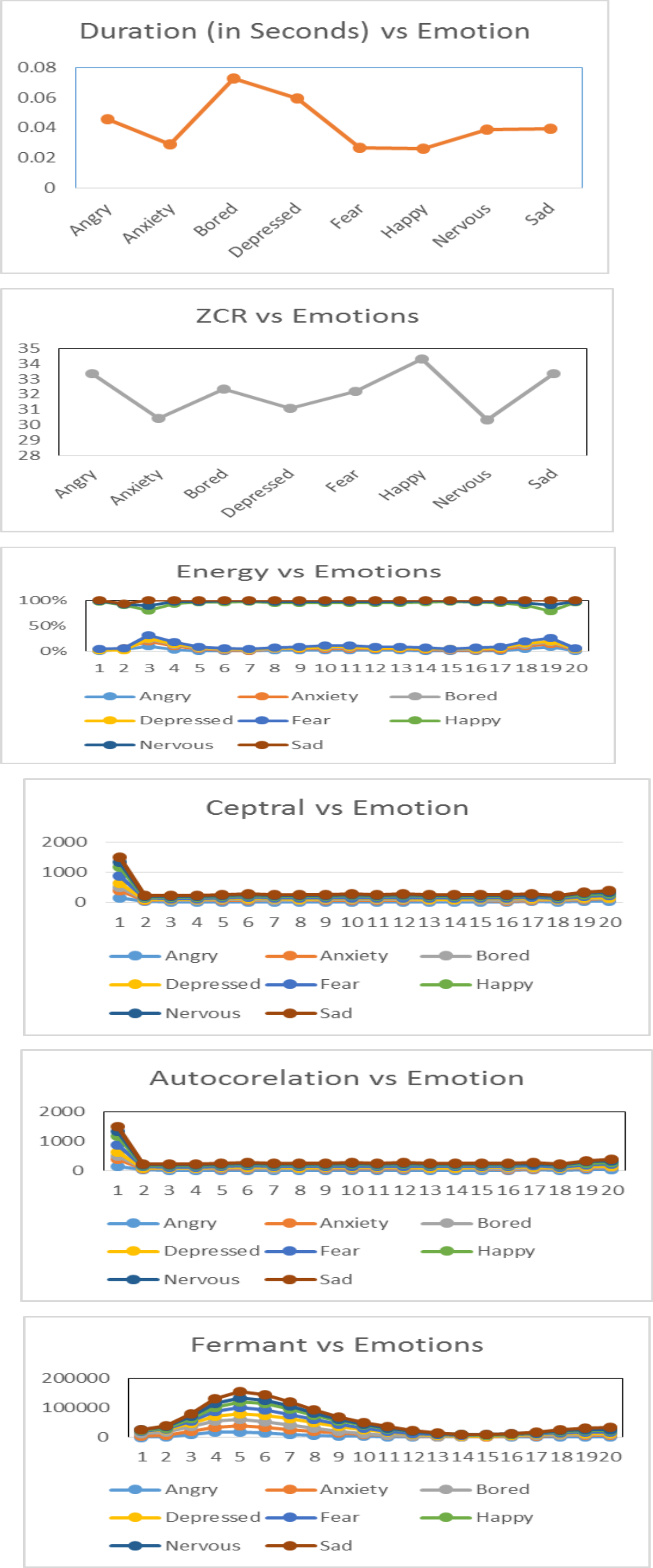

Fig. 3. Feature extracted from different methodology for different emotions.

TABLE II. **Confusion Matrix**

	Angry	Anxiety	Bored	Depressed	Fear	Happy	Nervous	Sad
Angry	75.00	12.50	8.33	0.00	0.00	0.00	4.17	0.00
Anxiety	8.33	70.83	0.00	8.33	0.00	0.00	12.50	0.00
Bored	4.17	8.33	62.50	12.50	8.33	4.17	0.00	0.00
Depressed	4.17	0.00	8.33	66.67	8.33	0.00	12.50	0.00
Fear	0.00	12.50	4.17	0.00	62.50	0.00	12.50	8.33
Happy	0.00	4.17	8.33	0.00	0.00	79.17	8.33	0.00
Nervous	0.00	0.00	0.00	8.33	4.17	0.00	75.00	12.50
Sad	0.00	0.00	12.50	0.00	8.33	0.00	8.33	70.83

V. CONCLUSION

Recognition of emotional states from speech is a current research topic with wide range. Emotion recognition through speech is particularly useful for applications in the field of human machine interaction to make better human machine interface. It is gaining a lot of importance due to its wide range of application in day to day life. In this paper, the features that are extracted are Duration, Zero Crossing Rate (ZCR), Pitch, Formant and Energy from Hausa Emotion Database. The emotions Angry, Anxiety, Bored, Depressed, Fear, Happy, Nervous and sad where recognizer using KNN classifier with 73.31% accuracy. These results may be improved by using various neural network classifiers.

REFERENCES

[1] Qi chieng shing ooi, kah phooi seng, Li-minn Ang, Li wern chew, “A new approach for audio emotion recognition (2014). Expert systems with applications science direct volume 41, issue 13, (October 2014) pp. 5858 5869.

[2] A. Milton, S. Sharmy Roy, S. Tamil Selvi, “SVM Scheme for Speech Emotion Recognition using MFCC Feature”, International Journal of Computer Applications, Vol. 69 – No. 9, pp. 34-40, May 2013.

[3] Agrawal, S. S., Prakash, N., & Iain, A, “Transformation of emotion based on acoustic features of intonation patterns for Hindi speech”, 2010, IJCSNS International Journal of Computer Science and Network Security, 10(9), 198-205.

[4] Bahuguna, Sushma, and Y. P. Raiwani. "Study of Speaker’s Emotion Identification for Hindi speech." International Journal on Computer Science and Engineering 5, no. 7 (2013): 629-634..

[5] Bhoomika Panda, DebanandaPadhi, Kshamamayee Dash, Prof. SanghamitraMohanty, “Use of SVM Classifier & MFCC in Speech Emotion Recognition System”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2, Issue 3, pp.226-230, ISSN:2277-128X, March 2012.

[6] Ekman, Paul. "An argument for basic emotions." *Cognition & emotion* 6, no. 3-4 (1992): 169-200.

[7] Shashidhar G. Koolagudi, Ramu reddy,Jainath Yadav , K.Sreenivasa Rao. "IITKGP-SEHSC:Hindi speech corpus for emotion analysis." IEEE (2011).

[8] Chauhan, Rahul, et al. "Text independent emotion recognition using spectral features." Contemporary Computing. Springer Berlin Heidelberg, 2011. 359-370.

[9] Wankhade, Sujata B., and YashpalsingChavhan PritishTijare. "Speech Emotion Recognition System Using SVM AND LIBSVM." International Journal Of Computer Science And Applications 4, no. 2 (2011): 0974-1003.

[10] Ahmad, Peerzada Hamid. "Transformation of emotions using pitch as a parameter for hindi speech." *TRANSFORMATION* 2, no. 1 (2012).