Detecting Emotions through Pitch and Formants for South Kamrupi Dialect of Assamese Language

Ranjan Das, Uzzal Sharma

Abstract*: *Speech is a form of voice signal that propagates through atmosphere and human beings communicate primarily through the exchange of speech. Along with their thoughts, human also expresses their emotions while communicating through the speech. Spoken dialects are the most native forms of speech communication and emotion is expressed very effectively through it. Speech is digitized to use it for speech processing. The speech processing task mostly comprises of feature extraction and pattern recognition. Pitch and Formants are two most fundamental speech features through which emotions can be modeled. Modeling emotion is a pattern matching task. This paper describes a successful empirical study of Pitch and Formants based emotion recognition system on the widely spoken, colloquial South Kamrupi dialect of Assamese language.

Index Terms*: *Automatic Speech Recognition(ASR), Emotion Detection, Pitch, Cepstrum, Formants.

I. INTRODUCTION

Speech is the expression of thoughts and feelings by articulating voice sound[1][2]. It is most naturally and effectively used as a means of communication by human. Human uses the speech communication by framing it into some language. Languages are spoken in the form of colloquial dialects. For every language, there are usually many dialectic forms. Different dialects of the same language are mostly characterised by variation in pronunciation as well as grammar, vocabulary, prosody, sentence structure, region of the inhabitants etc. Moreover, usually, the dialects do not have any standard, written form. People mostly use the spoken dialects to express their emotions effectively. Emotion is the sturdy feelings attributed by the circumstance, mood or attachment of a speaker[3]. On the other hand, speech can be comprehended as a digital signal. Various temporal and spectral features of digital speech signals can be extracted by speech processing. These features can be employed for emotion detection. This paper is an attempt to explore the pitch and formant frequency features for emotion detection on South Kamrupi dialect of Assamese language[4]. The following portion of this paper can be organized as follows. Section II explores the South Kamrupi dialect of Assamese language. Section III describes the basic ideas of acoustic feature extractions, especially on pitch and formants. Section IV discusses the various empirical issues of emotion detection on some sample speech data files recorded as a primary data source spoken in this dialect and attributed by emotion. Section V concludes this paper with its future perspectives.

Revised Manuscript Received on July 05, 2019.

Ranjan Das, Department of Computer Science & Engineering, School of Engineering, Assam Don Bosco Univeristy, Guwahati, Assam, India
Uzzal Sharma, Department of Computer Science & Engineering, School of Engineering,Assam Don Bosco Univeristy, Guwahati, Assam, India

II. THE SOUTH KAMRUPI DIALECT OF ASSAMESE LANGUAGE

Assamese is the first, among the 22 official languages enlisted in the eighth schedule of the constitution of Republic of India. Assamese is mostly used as the mother tongue in the Brahmaputra valley of the state of Assam. Moreover, it is used as a lingua-franca in all over North-Eastern India, particularly in the state of Nagaland, Arunachal Pradesh and Meghalaya. Besides, a handful of people from Bangladesh, Bhutan and Nepal also speak Assamese.The modern day Assamese language is evolved from Kamrupi dialect, formerly known as Kamrupi Prakrit or old Kamrupi language[5]. The word Kamrup is derived from the ancient Kmarupa kingdom of (4^{th} to 12^{th} century AD) and the distribution of this kingdom was there in Assam, parts from North Bengal, Arunachal Pradesh and Bhutan[6]. During medieval periods, scholars used this language for literary purpose, parallel to Sanskrit. The discovery of rock and the copper plate inscription of 5^{th} to 9^{th} century confirmed that this language was present all over the Brahmaputa Valley in it's ancient form for centuries.

In its present day form, Kamrupi dialect is spoken in Kamrup (Metro), Kamrup, Nalbari, Barpeta, Darang, Kokrajhar, Udalguri and Bongaingaon districts of Assam. However, Kamrupi is actually a colloquial, heterogeneous dialect with four distinct major spoken forms[7]. All these forms exhibit typical phonological and morphological characteristics. The major forms are

1) Central Kamrupi(Nalbari)
2) West Kamrupi (Barpeta)
3) North Kamrupi(Pathsala)
4) South Kamrupi (Palasbari)

The people of southern part of Kamrup district along with the western part of Guwahati city with an estimated population of around 19 lakhs do speak South Kamrupi dialect. This folk are purely cosmopolitan in the sense that they belong to various caste, tribe, creed and ethnicity. A sizable number of linguistic minority groups such as Marowari, Bihari, Bengali etc. are also present in this region. All these people use the South Kamrupi dialect as lingua-franca. There is no record of comprehensive computational linguistic study on this dialect so far[13].

The following section discusses some basics of acoustic feature extraction, especially pitch and formants which are used here as parameters for experimenting and modeling the human emotion.

III. ACOUSTIC FEATURE EXTRACTION

Human speech is the sound caused by the exhalation from the lungs by either the vibration of the vocal cords or by turbulence at some constriction points in the vocal tracts. Automatic Speech Recognition System (ASR) is the process of converting a recorded speech into an acoustic signal and to use them as pattern classifier[8]. For ASR, speech sound has to be perceived as digital signal first. Then, two operations which have to be carried out on that digital signal are : signal modeling and pattern recognition[9]. Signal modeling is all about converting the digital speech signal into a set of parameters. Pattern recognition is the study to ascertain the match between these parameters and previously known parameters. There are basically four major operations involved in the signal modeling : spectral shaping, feature extraction, parametric transformation and statistical modeling. Spectral shaping deals with the conversion of the speech signal from sound pressure wave into its digital form. Feature extraction is obtaining different features such as power, pitch, vocal tract configuration, formants etc. The features, thus obtained are converted into transformation parameters through differentiation and concatenation. Statistical modeling involves the analysis of these parameters in the form of signal observation vector.

For feature extraction, the main emphasis is on representation of the speech signal by a sequence of feature vectors[9]. In digital speech processing, there is ample number of available parameters which specifies the individualistic features of the speaker such as intensity, pitch, LPC, formants, energy level etc. It is learned that pitch and formants are the most crucial speaker dependent features which can be employed for human emotion detection[4].

A. Pitch

Pitch is generally accepted as the fundamental frequency of the speech signal[8]. It is a perceptual parameter in comparison to the best match with a pure sinusoid[9][10]. It mostly represents the quality of sound. It can only be extracted from clear, stable and noise free sound. It is a powerful psychoacoustic attribute. There are many factors which influence the pitch such as frequency of excitation of the vocal cords of the speaker, the size of the larynx, the length of the vocal cord etc. Pitch may also vary from different syllables within the same word. There are many different algorithms for pitch extraction[10]. Cepstral method of pitch extraction has been chosen for implementation because of its simplicity and accuracy.

B. Cepstral Method of Pitch Estimation

Cepstral method[9] assumes that the speech signal is composed of an excitation e(t) applied to the vocal tract filter and it has an impulse response v(t). The speech signal s(t) in the time domain t is given by the equation (1) and it is the convolution of e(t) and v(t).

$$s(t)=e(t)*v(t) \quad \text{..... (1)}$$

On expressing the equation (1) in the frequency domain, equation (2) is obtained.

$$S(w)=E(w)*V(w) \quad \text{..... (2)}$$

Here, in equation (2), S(w), E(w) and V(w) are the Fourier transformation of the continuous time functions s(t), e(t) and v(t). The excitation E(w) and the vocal tract filter V(w) are combined multiplicatively, so they are inseparable. On taking Logarithm of S(w), the excitation and vocal tract function becomes additive and thus, they become separable. It has been represented by the equation (3).

$$\log[S(w)]=\log[E(w)]+\log[V(w)] \quad \text{……(3)}$$

When, further Fourier Transformation is applied to the log spectrum, the resultant is called Cepstrum.

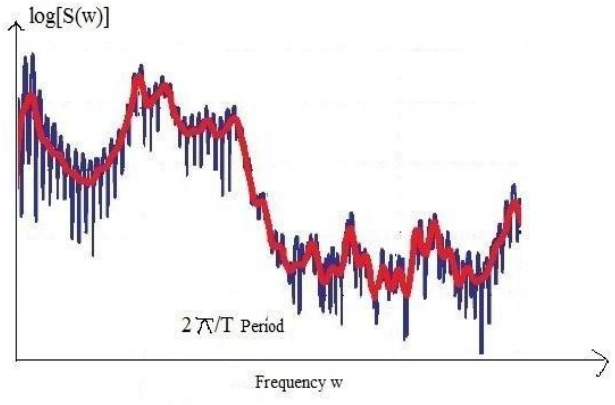

Fig. 1. Pitch component added to the vocal tract Response

The periodic component in the logarithmic spectrum at a frequency interval equivalent to the reciprocal of the pitch period T is given by the figure 1.These periodic components are becoming sharp peaks in the cepstrum. In the cepstral domain, the pitch can be estimated by picking the very first peak frequency of the resulting signal within a certain range. The magnitude of the pitch frequency exhibits the emotional level of the speaker.

C. Formants

Formant frequency is another principal analytical feature of the speech spectrum[9][11]. It is expressed as the spectral peaks of the voice sound of the speaker. It represents the acoustic resonance for the human vocal tract. At those spectral peaks, the acoustic energy around a particular frequency of the speech wave form is concentrated. Formants are usually measured as peaks of amplitude in the frequency spectrum of the sound wave. Usually, the first 3 formants, namely f1, f2 and f3 are considered for emotion detection[12]. For various vowel sounds, the range of the first formant, f1, typically lies between 270 to 730 Hz while the range of the second formant, f2 and the third formant, f3 lie between 840 to 2290 and 1690 to 3010 Hz respectively[4]. There are multiple, well known techniques available for formant frequency estimation[11]. Linear Predictive Code (LPC) based technique is chosen for this implementation. The following section discusses the dataset on which all the experiments are carried out and it demonstrates the

basic results which are obtained. A discussion is also made in parallel with the results.

IV. THE DATASET, EXPERIMENTS AND RESULTS

Dictionary defines emotion as the strong feelings derived from one's circumstances, mood, or relationships with others. It is a very complicated psychological state and it involves mostly three distinct components: a subjective experience, a physiological response and a behavioral or expressive response[12]. For the sake of convenience of our implementation, only three basic emotional states of the speakers are considered for identification; namely: happiness, neutral and anger.

An emotional primary dataset(mini corpora) is created which comprises of these 3 basic emotional states. Ten native South Kamrupi speakers from different cast, tribe and creed were volunteering the imitations of these emotions[13][14]. One sentence comprising of 7 words which actually signify specifically the day to day communication in South Kamrupi dialect of Assamese language has been chosen carefully to utter in all these three emotional states. It means, a total of 30 different data samples are recorded and each of these samples is in between 3.159 Seconds to 4.426 Seconds. The bit rate of all the samples is assigned to be 22050, resolution is chosen to be 16 bit and channel to be mono. A recording software CoolEditPro is used for all the recordings. Special care is taken to maintain a noise free recording environment. Matlab 15 is used exclusively to carry out all the experiments on these recorded speech dataset.

The following section describes various parameters evaluated while experimenting with pitch and formants for a recorded sample speech file of length 4.15 seconds. The graph in the Figure 2 is representing the speech file in its digitized form in time domain. It is followed by the graph in the Figure 3 which is the pitch frequency estimated via cepstrum based method on the sample speech file [10].

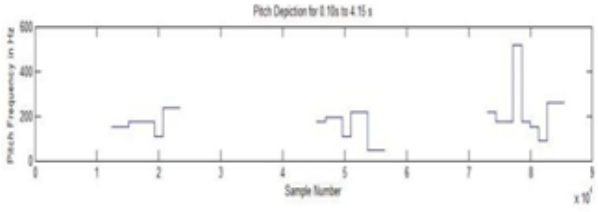

Fig.2. Sample Wave in Time domain

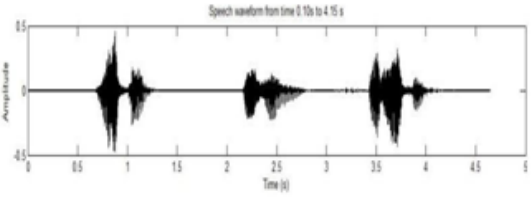

Fig. 3. Cepstrum based Pitch Frequency of the Sample

The power associated with the sample speech file is depicted by the periodogram as represented by the Figure 4.

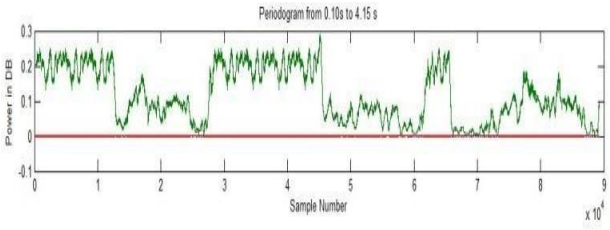

Fig.4. Power of the sample in Decibel

The vertical spectrographic representation of the frequency of the speech file is given by figure 5.

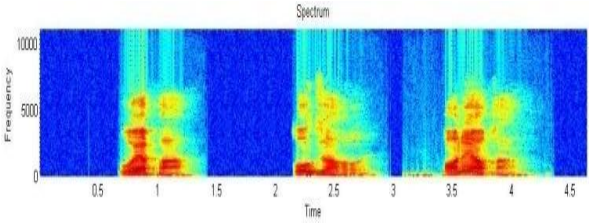

Fig.5. Vertical Frequency Spectrogram of the sample

Now finally, the formants associated with the sample files are depicted by figure 6. Linear Predictive Code(LPC) based formant frequency estimation technique is used in the experiment[11].

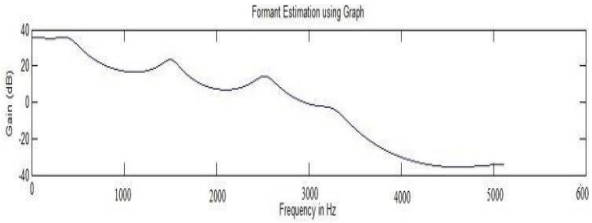

Fig.6. LPC based Formant Estimation

Thus, by using pitch and formants, a detailed empirical study has been carried out for identifying three basic emotional states namely : happiness, neutral and anger. It has been observed that pitch can be effectively used for detecting the emotional states neutral and anger while formants can be used for detecting the emotion, happiness. The results obtained demonstrate the facts that pitch frequency for neutral emotion are significantly low as compared to anger. Figure 7 depicts the histogram which signifies the empirical study for all the 10 speakers.

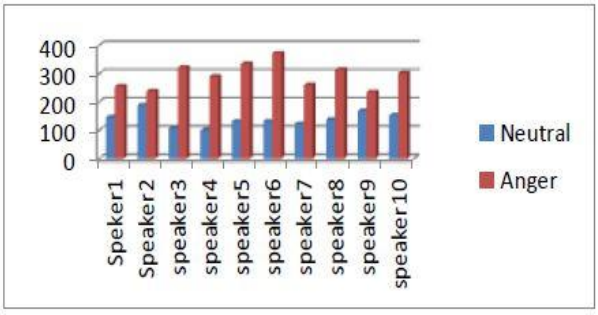

Fig.7. Pitch values for neutral and anger

It is found that the mean of pitch for neutral emotional state is 138.15 and the mean for the anger emotional state is 292.2.

On the other hand, the emotion happiness is well identified by the first three formant frequencies namely : f1, f2 and f3. The results of formant frequencies obtained are demonstrated by the histogram in the Figure 8.

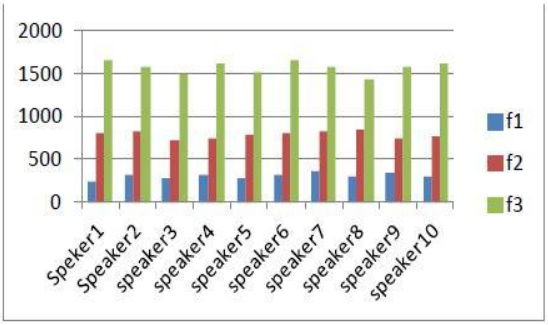

Fig.8. Formants f1,f2 and f3 for happiness

It depicts all the 10 speakers who are volunteering in the empirical study. It is obtained that the mean of f1, f2 and f3 for happiness are 303.1, 784.3 and 1571.9 in order.

V. CONCLUSION

It is potentially observed that different emotional states do exhibit different spectral behaviours in the speech communication of South Kamrupi dialect of Assamese language too, like every other verbal communication. Cepstrum based pitch estimation and Linear Predictive Code(LPC) based formant evaluation techniques are used experimentally to recognize three fundamental human emotional states in this study namely : happiness, neutral and anger. Pitch is the spectral feature which is successfully used to recognize the neutral and anger emotional state of the speaker whereas formant frequencies are used to recognize the happiness emotional state. There is ample scope to extend this study by enlarging the corpora and by incorporating other emotional states also like enthusiasm, disgust, sorrow, grief etc. Other acoustic features of speech can also be explored in parallel for better emotion detection. This research endeavour is the first attempt of emotion detection through speech processing in South Kamrupi dialect of Assamese language. It is surely enriching the dialect in particular, and the Assamese language as a hole.

REFERENCES

1. L.R.Rabiner and B. H.Juang, *Fundamentals of Speech Recognition.* London, United Kinddom : PTR Prentice Hall, 1993.
2. D.Jurafsky and J.H.Martin, *Speech & language processing : An introduction to natural language processing, computational linguistics, and speech recognition.* 2nd ed., Pearson Education India, 2014.
3. U. Sharma, "Identification of emotion from speech signal," in *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 2016, pp. 2805-2807,IEEE,2016.
4. Sathe-Pathak, Bageshree V., and A. R. Panat. "Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person." *IJCSI International Journal of Computer Science Issues* 9, no. 4 ,2012.
5. B.K.Kakati, *Assamese, its formation and Development,* Guwahati, Lawyers Book Stall. pp. 56-57, 1941.
6. U.N.Goswami, *A study on Kamrupi : A dialect of Assamese,* Lawyers Book Stall, Guwahati, Assam, 1970.
7. K. Medhi, "Assamese grammar and origin of the Assamese language" in Publication Board, 1988.
8. C.Rowden, "Speech Processing", Department of Electronics Systems Engineering, University of Essex, McGra-Hill Book Company, 1992.
9. V.Anderson and F.-T.Hsiao, "Speech Coding and recognition", IT University of Copenhagen, November, 2005.
10. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, no. 5, pp. 399–417, Oct. 1976.
11. Gargouri, Dorra, M. A. Kammoun, and A. B. Hamida. "A comparative study of formant frequencies estimation techniques." In *Proceedings of the 5th WSEAS international conference on Signal processing, Istanbul, Turkey*. 2006.
12. Irtiza, Naveen, and H. Farooq. "The study of fear-induced power modulations for Cognitive Man-Machine Communication." In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on, pp. 346-351. IEEE, 2015.
13. R. Das and U. Sharma, "Extracting acoustic feature vectors of South Kamrupi dialect through MFCC," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi,India,2016,pp.2808-2811,IEEE,2016.
14. J. Borah and U. Sharma, "Automatic identification of the dialects of Assamese Language in the District of Nagaon," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 2827-2831, IEEE,2016.

AUTHORS PROFILE

Ranjan Das is presently a research scholar in the department of Computer Science & Engineering, School of Engineering, Assam Don Bosco University, Guwahati, India. He has obtained an MTech degree in Information Technology from Tezpur University, India in 2011. His researches are in the fields of soft computing and speech processing. Recently, dialect modeling application development on colloquial dialect has been undertaken.

Uzzal Sharmahas obtained his MCA from IGNOU and completed PhD from Gauhati University. He has over 15 years of experience in the field of academics and in industry. His research area includes Speech Signal Processing and Software Engineering and Data Engineering. He has produced one PhD scholar and currently guiding 4 research scholars for their PhD degree. He has also guided many MTech, BTech and MCA students for their projects in different areas. He has published more than 25 research papers in journals (International and National) and conference proceedings (International and National). He also has 11 book chapters to his credit in edited book. He has also published five books as a sole author. Currently he is an Assistant Professor - Stage 2 at Assam Don Bosco University, Guwahati, India.