International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-1, Issue-6, January 2013

Recognition of the Tonal Words of BODO Language

Utpal Bhattacharjee

Abstract: The performance of a state-of-art speech recognition system degrades considerably when the recognizers are used to recognize the tonal words. This is due to the fact that at the time of developing those recognizers, the tonal property has not been considered.Bodo is a tonal language like other Sino-Tibetan languages. In this paper we consider how current models can be modified to recognize the tonal words. Two approaches have been investigated in this paper. In the first approach attempt has been made to develop a feature level solution to the problem of tonal word recognition. In the second approach, a model level solution has been suggested. Experiments were carried out to find the relative merits and demerits of both the methods.

I. INTRODUCTION

Most of the automatic speech recognition theory and systems are developed in the Indo-European context [1,2,3]. However, for global acceptability of the automatic speech recognition system, it must give a consistent performance for any language it operates. It has been observed that the state-of-art speech recognizer system suffer serious performance setback when it operates in Sino-Tibetan languages. One of the major reasons for such performance setback is due to the ignorance of tonal nature of those languages. Most of the languages in Sub-Saharan Africa, East Asia and South-East Asia are tonal. Thus, a major part of the world population speaks tonal language. Therefore, the capability of the automatic speech recognition system to process tonal language is a basic requirement for universal acceptability of these systems.

The paper is organized as follows: Section II is dedicated to an introduction to the tonal language in context of Bodo language. In Section III describes the baseline speech recognition system. In Section IV we present two alternative solutions for the tonal word recognition. Section V we describes the experiments carried out and present the results. The paper concludes in section VI.

II. AN INTRODUCTION TO TONAL LANGUAGE

The different pitch levels produce different types of tone in a language. Pitch is the acoustic result of the speed of the vibration of the vocal cord in the utterance of the voiced part of the sound. The vocal cord rapid vibration produces high-pitched sound and slow vibration produces low-pitch sound. Due to pitch contour movement, the tones may fluctuate and thus raising and falling tones are produced [4]. Pitch variation is found in all languages; however, its function is different from language to language. In some language, specially the Sino-Tibetan family of languages, the pitch difference distinguishing the meaning of one word from the other though they have the same phonetic structure. The pitch difference used in this way is called tones. Tones refer to the distinctive pitch level of a syllable. In many languages the tone carried by the word is very essential for the meaning of the word. Such languages are called tonal languages.

Tone may be on a single level of pitch, called level tone or may fluctuate and thus produce contour type of tones. As a result of the fluctuation, the level of tone may change and produce different categories of tones. If the pitch level rises during the articulation of the sound it is called rising tone. If the pitch level falls, the tone is called falling tone. There may be fluctuation in the middle to produce the tones rising-falling and falling-rising. Based on the pitch movement from the starting position, the tones may also be classified as mid-level, high-level and low-level due to their level-wise movement or they may be mid-rising, mid-falling, high-rising, high-falling, low-rising and low-falling due to their fluctuation from the starting position.

Bodo is a tonal language. It has two contrastive tones of contour type – rising, which rises still higher than its original pitch registered at the beginning of the syllable and falling, which falls still lower than its original pitch registered at the beginning of the syllable. Any of the two tones must co-occur with every syllable in the language. The falling and the rising tones may be marked with numeral 1 and 2. Some of the words in Bodo language where the basic syllable is same but meaning is changed due to tone is given below[4].

BodoTonal Words	Meaning
/^{1}si/	Cloth
/^{2}si/	To be wet
/^{1}su/	To wash
/^{2}su/	To measure
/^{1}huu/	To drive
/^{2}huu/	To give
/^{1}er/	To draw (a picture)
/^{2}er/	To increase
/^{1}sum/	To soak
/^{2}sum/	To be black
/^{1}ran/	To become dry

Revised Manuscript Received on 30 January 2013.
* Correspondence Author
Utpal Bhattacharjee*, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India.

/[2]ran/	To divide
/[1]gaɔ/	To feel thirsty
/[1]gaɔ/	Wing

III. BASELINE SPEECH RECOGNITION SYSTEM

A baseline speech recognition system has been developed using Mel Frequency Cepstral Coefficient as feature vector and Recurrent Neural Network (RNN) as recognizer. The theoretical detail of the system is given below:

A. *Recurrent Neural Network based Phoneme Recognizer*

The speech model has been constructed using a fully connected recurrent neural network. This network architecture was described by Williams and Zipser[5] and also known as Williams and Zipser's model. Let the network has N neurons and out of them k are used as output neurons. The output neurons are labelled from 1 to k and the hidden neurons are labelled from k+1 to N. Let P_{mn} be the feed-forward connection weight from m^{th}input component to the n^{th} neuron and w_{nl} be the recurrent connection weight from the l^{th} neuron to the n^{th} neuron. At time t, when an M-dimensional feature vector $U(t)$ is presented to the network, the total input to the n^{th} neuron is given by

$$(t) = \sum_{l=1}^{N} w_{nl} x_l(t-1) + \sum_{m=1}^{M} P_{nm} U_m(t) \quad \text{--- (1)}$$

where$x_l(t\text{-}1)$ is the activation level of the l^{th} neuron at time t-1 and $U_m(t)$ is the m^{th} component of $U(t)$. The resultant activation level $X_n(t)$ is calculated as

$$(t) = f_n(Z_n(t)) = \frac{1}{1+e^{-Z_n(t),}}, 1 \le n \le N \quad \text{---(2)}$$

To describe the entire network response at time t, the output vector $Y(t)$ is formed by the activation level of all output neuron, i.e.

$$Y(t) = [x_1(t) x_2(t) \dots\dots\dots\dots x_k(t)]^T \quad \text{----(3)}$$

Following the conventional winner-take-all representations, one and only one neuron is allowed to be activated each time. Thus, k discrete output states are formed. In state k, the k^{th} output neuron is most activated over the others. Let $s(t)$ denote the output state at time t, which can be derived from $Y(t)$ as

$$S(t) = \arg\max_{j=1}^{k} \{x_j(t)\} \quad \text{--- (4)}$$

The RNN has been described so far only for a single time-step. When a sequence of input vector $\{U(t)\}$ is presented to the network, the output sequence $\{Y(t)\}$ is generated by eq. (2) – (4). By eq. (5), $\{Y(t)\}$ can be further converted into an output scalar sequence $\{s(t)\}$, and both of them have the same length as $\{U(t)\}$. $\{s(t)\}$ is a scalar sequence with integer value between 1 to n. It can be regarded as a quantized temporal representation of the RNN output.

The fully connected RNN described above performs time aligned mapping from a given input sequence to an output state sequence of the RNN. Each element in the state sequence is determined not only by the current input vector but also by the previous state of the RNN. Such state dependency is very important if the sequential order of input vector is considered as an indispensable feature in the sequence mapping.

In the present study, the recurrent neural network has been used to construct a recognizer to recognize the isolated words of Bodolanguage

The Real Time Recurrent Learning (RTRL) algorithm [5] with sufficiently small learning rate has been used to train both the phoneme recognizer.

B. *Mel Frequency Cepstral Coefficients (MFCC)*

Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction method in speech recognition. The technique is called FFT based which means that feature vectors are extracted from the frequency spectra of the windowed speech frames.

The Mel frequency filter bank is a series of triangular bandpass filters. The filter bank is based on a non-linear frequency scale called the mel-scale. According to Stevens et al[6], a 1000 Hz tone is defined as having a pitch of 1000 mel. Below 1000 Hz, the Mel scale is approximately linear to the linear frequency scale. Above the 1000 Hz reference point, the relationship between Mel scale and the linear frequency scale is non-linear and approximately logarithmic. The following equation describes the mathematical relationship between the Mel scale and the linear frequency scale

$$f_{Mel} = 1127.01 \ln\left(\frac{f}{700} + 1\right) \quad \text{---} \quad (5)$$

The Mel frequency filter bank consist of triangular bandpass filters in such a way that lower boundary of one filter is situated at the center frequency of the previous filter and the upper boundary situated in the center frequency of the next filter. A fixed frequency resolution in the Mel scale is computed, corresponding to a logarithmic scaling of the repetition frequency, using $\Delta f_{Mel} = (f_{H(mel)} - f_{L(mel)})/(M+1)$ where $f_{H(mel)}$is the highest frequency of the filter bank on the Mel scale, computed from f_{max} using equation (5), $f_{L(mel)}$is the lowest frequency in Mel scale, having a corresponding f_{min} and M is the number of filter bank. The values considered for the parameters in the present study are: f_{max}=8 KHz, f_{min}=0 Hz and M=20. The center frequencies on the Mel scale are given by

$$f_{cm(Mel)} = f_{L(Mel)} + \frac{m(f_{H(Mel)} + f_{L(Mel)})}{M+1}, 1 \le m \le M$$

--- (6)

The center frequencies in Hertz, is given by

$$f_{cm} = 700\left(e^{\frac{f_{cm(Mel)}}{1127.01}} - 1\right) \quad \text{--- (7)}$$

Equation (7) is inserted into equation (5) to give the Mel filter bank. Finally, the MFCCs are obtained by computing the discrete cosine transform of $X'(m)$ using

$$c(l) = \sum_{m=1}^{M} X'(m) \cos(l \frac{\pi}{M}(m - \frac{1}{2})) \quad \text{--- (8)}$$

for$l$ = 1, 2, 3, ….., M where $c(l)$ is the l^{th} MFCC.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The time derivative is approximated by a linear regression coefficient over a finite window, which is defined as

$$\Delta c_t(l) = \left[\sum_{K=2}^{2} k \; c_{t-k}(m)\right].G, \; 1 \le l \le M \quad \text{--- (9)}$$

where $c_t(l)$ is the l^{th} cepstral coefficient at time t and G is a constant used to make the variances of the derivative terms equal to those with the original cepstral coefficients. In the present study we use first 12 coefficients excluding. The 0^{th} coefficient was not considered as it contains energy of the whole frame. To add the dynamic property of the speech signal the 1^{st} order derivatives is also added to the feature vector.

IV. ENHANCEMENT OF THE BASELINE SYSTEM

In the present study two alternative approaches have been taken for the recognition of the tonal words of Bodo language and their performances have been evaluated. In the first approach MFCC features has been combined with Prosodic features. In the second approach, two separate recognizers have been used for recognizing the base-syllable and tone respectively. In the following subsection we describe the algorithm used for detecting prosodic features and in the next subsection we describe the structures of the enhanced speech recognizers.

A. Algorithm for Prosodic Feature Extraction

Prosodic features are the rhythmic and intonational properties in speech, examples are voice fundamental frequency (F0), F0 gradient, intensity and duration. They are relatively simple in structures, and are believed to be effective in some speech recognition tasks. Prosody refers to non-segmental aspects of speech, including, for instance, syllable stress, intonation patterns, speaking rate and rhythm. One important aspect of prosody is that, unlike the traditional short-term spectral features, it spans over long segments like syllables, words, and utterances and reflects differences in speaking style, language background, sentence type, and emotions to mention a few. A challenge in text-independent speaker recognition is modeling the different levels of prosodic information (instantaneous, long term) to capture speaker differences; at the same time, the features should be free of effects that the speaker can voluntarily control [7].

The most important prosodic parameter for the recognition of tone is the fundamental frequency (or F0). Other prosodic features for tone recognition includes duration, speaking rate, formants, pitch and energy distribution/modulations among others. In has been observed that for tone recognition F0-related features yielded the best accuracy, followed by energy and duration features in this order [8,9,10].

Through a pitch detector algorithm [11], the pitch related acoustic features are extracted - including frame energy, the probability of voicing and pitch period. The same window size and frame rates are used to make the extracted pitch features more consistent with the original cepstral coefficients based features.

Thus, the speech signal s(n), is first divided into frames. For each frame, decisions are made for: (a) speech vs. non-speech and (b) the pitch period. The basic features of the algorithm are as described below.

First to discriminate between speech and non-speech, the signal energy level is computed using autocorrelation and it is then compared with fixed threshold. Cepstral coefficients are computed. In cepstral domain, first peak (R_0) is 0^{th} cepstral coefficient, which is partly depends on the frame energy. In voiced speech the second peak (R_1) is present showing the energy of F0. For unvoiced frame, no predominate 2^{nd} peak is present. Therefore, the ratio of R_1 against R_0 denoted by R_c is compare with a fixed threshold t. If R_c is longer than t, the frame is classified as voiced and the position of R_1 is the pitch period.

For the features to be useful for speech recognition, it is better to make soft decision instead of hard decision for speech silence differentiation. By using autocorrelation value e as a feature, we can estimate the conditional distribution Pr (e | non-speech) and Pr (e | speech) empirically using non-parametric estimation techniques (such as histogram). By using Bayes rule and empirical estimation of Pr (speech) and Pr (non-speech), we can estimate the probability, Pr (speech | e), for each frame.

The algorithm stated above generates two pitch related features for each frame, namely, the transfer energy En(t) and the pitch period. For using these features in real speech recognition application we are to normalize these parameters as described in the following paragraphs:

The energy of the voicing region is higher than that in unvoiced region and so it is intuitively a useful feature. However, the energy can be affected by loudness which is irrelevant to phonetic identity. In the present study, we use the transformed energy En(t), which is given by:

$$E_n(t) = \frac{E(t) - E_{channel}}{E_{\max} - E_{channel}} \quad \text{--- (10)}$$

where E(t), $E_{channel}$ and E_{max} are energy at frame t, average

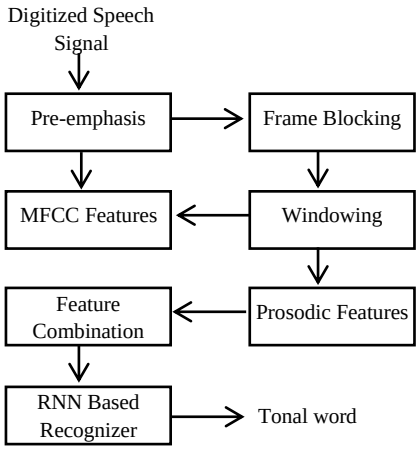

Fig.1: Tonal word recognizer using combined feature Vector

energy in the silence period and maximum energy across the whole utterance respectively. In our study, we consider two type of transformation of $E_n(t)$ which are given by log ($E_n(t)$) and $\Delta\log(E_n(t))$.

Pitch period or F0 is the most important feature because it directly related to tone. However, as the pitch period is only defined in the voiced region, depending on the pitch extraction algorithm,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

it is sometimes set to 0 during unvoiced and silence region.

This problem is similar to the problem of probability of voicing that can have zero variance if a hard 0/1 decision is made during feature extraction. Different solutions have been proposed to deal with this problem [12]. In the present investigation, it has been observed that pitch period of unvoiced frame are self-sustainable by itself and no special treatment is required. Therefore, the pitch period is normalized using average pitch of a sentence as described in the equation given below:

$$F_n(t) = \frac{F_0}{\overline{F_0}} \quad \text{--- (11)}$$

Since tone is actually a segmental feature, modelling the pitch per frame may not be sufficient in determining the tone pattern and as derivatives are the normal approaches for modeling frame dependency, therefore, the first order and the second order derivatives of the normalized pitch period, i.e., ΔF_n (t) and $\Delta^2 F_n$(t) has been considered. Therefore, the pitch related feature vector for frame *t* is given by:

$$U_p(t) = \{\log(E_n(t)), \Delta\log(E_n(t)), F_n(t), \Delta F_n(t), \Delta^2 F_n(t)\} \quad \text{---(12)}$$

B. *Modification in the baseline system*

In the first approach, we enhance the baseline system by adding prosodic feature to the feature vector of the baseline system. The digitized speech signal at 8 KHz, 16 bit mono resolution has been pre-emphasized by a pre-emphasized filter $1\text{-}0.96z^{-1}$ and then block into frame of duration 30 microseconds which conation 240 samples. To make the frame size multiple of 2, the fame size is adjusted to 256 samples. The frame rate is kept at 100 Hz. Each frame is multiplied by a Hamming window and the windowed signal is passes through two parallel process for the calculation of MFCC as well as Prosodic features. Once the features are calculated, they are concatenated and as a result we get a 29-dimensional feature vector. The feature vector is now used as the input to the RNN based speech recognizer for the recognition of tonal word recognition.

In the second approach, the job of recognizing the tone and the base-syllable has been distributed into two parallel system and the final results are combined to recognize the tonal word. The baseline configuration has been used for the recognition of the base-word. However, short-time cepstral mean and variance normalizationand has been used to the MFCC feature vector to compensate for the pitch related features. The detail of the method applied is given below:

In short-time mean and variance normalization (STMVN), *m* number of frame with *k* feature vector each has been normalized. That is, the space used for normalization is C(m,k). The normalization operation is given below:

$$C_{STMVN}(m,k) = \frac{C(m,k) - \mu_{st}(m,k)}{\sigma_{st}(m,k)} \quad \text{--- (13)}$$

Where m and k is the frame index and cepstral coefficient index respectively.$\mu_{st}(m,k)$and$\sigma_{st}(m,k)$ are the short-time mean and standard deviation respectively, defined as:

$$\mu_{st}(m,k) = \frac{1}{L}\sum_{j=m-L/2}^{m+L/2} C(j,k) \quad \text{---(14)}$$

$$\sigma_{st}(m,k) = \frac{1}{L}\sum_{j=m-L/2}^{m+L/2} \left(C(j,k) - \mu_{st}(m,k)\right)^2 \quad \text{---(15)}$$

Where L is the sliding window length in terms of frame.

The RNN based recognizer has been used for the recognition of the base-syllable.

To recognize the tone associated with the utterance of the word, we extract prosodic features from the windowed speech signal and a RNN-based recognizer has been used for the recognition of the tone. Once the base-syllable and the tone have been recognized, a tonal word recognizer has been used to recognize the tonal word.

V. EXPERIMENTAL SETUP

A. *Database Used for the Experiments*

All the experiments reported in this paper are carried out using a database of 3500 isolated Bodotonal words uttered by 25 speakers (13 male and 12 female). Each speaker utters 14 tonal words 10 times each. The recording has been done in a controlled environmental condition in a noise-free booth at 8 KHz with 16 bit mono format. The data is stored in WAV PCM format.

B. *Experiments and Results*

A baseline speech recognition system has been developed using MFCC feature vector and RNN.The digitized speech signal is first pre-emphasized using a pre-emphasized filter $1\text{-}0.96z^{-1}$ and blocked into frame of 256 samples each with frame frequency 100 Hz. The frames are multiplied by Hamming window and 12 MFCC coefficients are extracted from each frame along with its 1st order derivatives using the method explain in section III. Thus we get a 24 dimensional feature vector for each frame. These features are used as input to the RNN based speech recognizer. A RNN based speech recognizer has been developed consisting of 24 input units, 14 output units and 20 hidden units. The number of hidden units has been experimentally fixed.The sequentially arranged input vector has been given to the input of the RNN based speech recognizer and RTRL algorithm has been used to train the

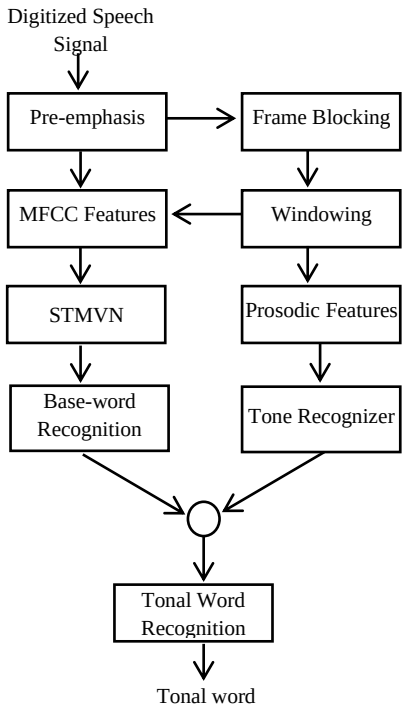

Fig.2: Tonal word recognition using parallel model for Base-word and Tone

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

recognizer. Single RNN has been used in the present study to recognize all the 14 Bodo tonal words considered in the present study. Twenty occurrences of each word has been considered for training, collected from 5 male and 5 female speakers. The system has been tested using remaining utterances and the performance has been evaluated.

Now the system is modified using the 1st approach as described in section IV. Prosodic features have been added with the cepstral features the combined features have been used for training and testing the system. The same dataset has been used for training and testing the system as above experiment. Now the RNN is modified to accommodate the increased dimension of the feature vector. The number of input nodes has been increased to 29, the output nodes which correspond to 14 test word remain same and the number of hidden nodes has been increased to 22, which is found to be suitable for this input/output ratio. The performance of the system has been evaluated.

Finally, the system has been enhanced using approach 2, described in section IV. The task of recognizing base syllable and tone has been separated. After STMVN to the cepstral features, the feature vector has been used as input to the RNN based base-syllable recognizer. Since there is only 7 base-syllables in the dataset considered in this study, the output unit is now limited to 7. Thus the recognizer consist of 24 input units, 7 output units and 15 hidden units, which is found to be suitable for this structure of the RNN. Further, another RNN based recognizer consisting of 5 input units, 2 output units and 3 hidden units has been used for recognizing two tones associated with the base-syllables. The results of the experiments have been present in Table-1.

Table-1: Results of the experiments for the recognition of tonal words

Recognition System	Feature Vector	Recognition Accuracy
Single RNN based Recognizer	MFCC	66.86
Single RNN based Recognizer	MFCC+ Prosodic	74.29
Separate Recognizer for Recognizing Base-word and Tone	MFCC for Base-word and Prosodic for Tone	83.57

VI. CONCLUSION

From the above experiments it has been observed the performance of a speech recognizer system degrades considerably when it is used for recognizing tonal words compared to the performance reported in our earlier work[13]. It is basically due to the fact that the feature extraction techniques remove the pitch related information of the speech signal. In the present study, when prosodic features, which basically pitch related information added to the feature vector, there is a sharp improvement of nearly 8% has been reported. However, this performance is still far behind. The poor recognition accuracy even after adding prosodic features may be due to the recognizer itself. Due to the more weight of the cepstral features, the recognizer may suppress tone related information. To overcome this problem, two separate recognizers have been used for recognizing the base-syllable and tone. It has been observed that as a result of using separate tone recognizer, the performance of the system improves considerably.

REFERENCES

1. Stephenson, T.A.; Doss, M.M.; Bourlard, H.; , "Speech recognition with auxiliary information," Speech and Audio Processing, IEEE Transactions on , vol.12, no.3, pp. 189- 203, May 2004
2. Venayagamoorthy, G.K.; Moonasar, V.; Sandrasegaran, K.; , "Voice recognition using neural networks," Communications and Signal Processing, 1998. COMSIG '98. Proceedings of the 1998 South African Symposium on , vol., no., pp.29-32, 7-8 Sep 1998
3. Abushariah, A.A.M.; Gunawan, T.S.; Khalifa, O.O.; Abushariah, M.A.M.; , "English digits speech recognition system based on Hidden Markov Models," Computer and Communication Engineering (ICCCE), 2010 International Conference on , vol., no., pp.1-5, 11-12 May 2010
4. Baro, M.R.; "The Boro Structure – A Phonological and Grammatical Analysis", Priyadini Printing Press, 2001.
5. Williams, R.J., Zipser, D: A learning algorithm for continually running fully recurrent neural networks. Neural Computation 1, 270--280 (1989).
6. Stevens, S., Volkmann, J., and Newman, E., "A Scale for the Measurement of the Psychological Magnitude Pitch." Journal of the Acoustical Society of America 8: 185–190, 1937.
7. Ng, Raymond WM, et al, "Analysis and Selection of Prosodic Features for Asian Language Recognition", International Journal of Asian Language Processing, 19(4):139-152, 2009.
8. Adami, A., Mihaescu, R., Reynolds, D., and Godfrey, J., "Modeling prosodic dynamics for speaker recognition", In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003), pp. 788–791, 2003.
9. Bartkova, K., D.L.Gac, Charlet, D., and Jouvet, D, "Prosodic parameter for speaker identification", In Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002), pp. 1197–1200, 2002.
10. Reynolds, D. et al, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition", In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003), pp. 784–787, 2003.
11. Li Tan and MontriKarnjanadecha, "Pitch Detection Algorithm: Autocorrelation Method and AMDF", Proceedings of the 3rd International Symposium on Communications and Information Technology, vol. 2, pp. 541-546, September 2003.
12. Wong, P.F. and Siu, M.H.; "Integration of Tone Related Features for Chinese Speech Recognition", Proceedings of ICSP' 02, PP 476-479, 2002.
13. Bhattacharjee, U.; "Environment and Sensor Robustness in Automatic Speech Recognition", International Journal of Innovation Science and Modern Engineering, Vol.1. No.2, pp 31-37, 2013.

AUTHOR PROFILE

Utpal Bhattacharjee received his Master of Computer Application (MCA) from Dibrugarh University, India and Ph.D. from Gauhati University, India in the year 1999 and 2008 respectively. Currently he is working as an Associate Professor in the department of Computer Science and Engineering of Rajiv Gandhi University, India. His research interest is in the field of Speech Processing and Robust Speech/Speaker Recognition.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)