World Journal of Engineering Research and Technology

WJERT

SJIF Impact Factor: 3.419

SPEAKER AUTHENTICATION USING ZERO CROSSING RATE WITH RESPECT TO BODO VOWEL PHONEME: A CLASSICAL EXPERIMENT

[1]Bimal Kumar Kalita* and [2]Dr. (Professor) Pran Hari Talukdar

[1]Research Scolar, Deptt. Of Usic & Instrumentation, Gauhati University, Assam.

[2]Ex- Hod, Deptt. Of Usic & Instrumentation, Gauhati University, Assam.

Article Received on 08/10/2016 Article Revised on 28/10/2016 Article Accepted on 17/11/2016

***Corresponding Author**
Bimal Kumar Kalita
Research Scolar, Deptt. of Usic & Instrumentation, Gauhati University, Assam.

ABSTRACT

The researcher fraternity in the field of Natural Language Processing have been always enthusiastic about the possibility of machine human interaction. In this area authentication of a certain person based on speaker's specificity in his/her voice is always attracting a section of scientists. Speaker authentication is a process of recognition and verification of the identity of a claimant speaker through the properties of his/her utterances i.e. analyzing the voice characteristics and then come to a conclusion whether the claim of the identity is correct or he/she is an imposter matching the pre existing data in the database. The Bodo language belongs to the Tibeto-Burman language family, which is a sub-group of the Sino-Tibetan language group. It is one of the popular Indian tribal languages, primarily spoken in Assam, the north-eastern state of India. Speakers are also found in adjoining areas of Assam and border areas of West Bengal and Bangladesh. In our paper experiments are performed with the method of speaker dependent modelling. System used consists of three phases.

i. Training ii. Testing and iii. Recognition/Authentication.

Five female and male speakers each are taken for speech database. Phoneme utterances are recorded with ten repetitions for feature extraction. Zero Crossing Rate is a feature vector which can be useful as an acoustic metric. In our study ZCR feature extraction of Bodo vowel phonemes are performed frame by frame. ZCR analysis will be helpful in designing a BODO

Speaker Recognition, Identification and Authentication System (SPERIA-B i.e. SPEaker Recognition, Identification and Authentication for Bodo).

KEYWORDS: Zero Crossing Rate, Speech / Speaker recognition / Authentication, feature vector, imposter.

Section A: INTRODUCTION

Speech is the tool through which human transfer their thoughts, feelings and knowledge, which gives them the edge over the other living fraternity in the universe. This is the prime reason why homo-sapience are well ahead in the race of knowledge accumulation and implementation. That makes us to dominate the universe. Speech is the most predominantly accepted, efficient and natural way for human beings to communicate. It is, therefore, sensible to investigate, develop and deploy technologies that facilitate speech-enabled human computer interaction, in environments where users may experience efficiency and convenience (Gaikwad, Gawali & Yannawar, 2010:24-28). With the various developments in the field of automatic speech recognition, dream to design a machine which can mimic the capability of a human to speak has enthralled engineers and scientists for centuries.[1] In forensic speaker authentication, two voice samples are compared to verify whether they are uttered by the same speaker or not. In (Rose, 2002), the author refers to scenarios in which voice samples are more incriminating than DNA samples. Therefore, developing a speaker recognition and authentication system can be useful in forensic applications as well.

Our focus in the present paper is on speech-enabled human computer interaction in access control, for the ultimate purpose of developing an efficient and effective speaker authentication system (SAS) with reference to Bodo language. Basically there are two approaches in speech recognition and verification/authentication:

a. Text dependent and
b. Text independent.

A *speaker dependent* system is intended for use by a single speaker, but a *speaker independent* system is intended for use by any or unanimous speaker. Developing Speaker independent system is critical to achieve for many reasons, mainly because parameters of the system becomes tuned to the speaker(s) that it was trained on, and these parameters tend to be highly speaker-specific. Error rates are typically 3 to 5 times higher for speaker independent systems than for speaker dependent ones (Lee 1988). Speaker recognition can be classified

into speaker identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance (Rabiner et.el. 1993).In speech authentication, *isolated speech* means single words; *discontinuous speech* means full sentences in which words are artificially separated by silence; and *continuous speech* means naturally uttered sentences. Isolated and discontinuous speech recognition is relatively easy because word boundaries are detectable and the words tend to be cleanly pronounced. Continuous speech is more difficult to deal with, because,

1. Word boundaries are unclear and
2. Their pronunciations are more corrupt due to the effects of *co-articulation*. In a typical evaluation, the word error rates for isolated and continuous speech were 3% and 9%, respectively (Bahl et al 1981).

Speaker Identification and authentication system consists of two essential modules, namely:
a. Feature extraction and b. Feature matching.

Speech Biometric authentication is the process of verifying a user's claim of identity by performing a one-to-one (1:1) comparison on the claimed username stored in the database. The voice biometric sample will be captured, modelled, and then compared with the biometric data in the database assigned to the claimed username. If the captured biometric data does not match the stored data then the user is rejected. However, if the captured data matches the stored value within a certain predefined threshold, then the user is successfully authenticated (Woodward, Orlans & Higgins, 2003:8). A human generated signal or attribute for authenticating a person's identity is known as a biometric. Biometrics is the process of identifying and confirming an individual's identity through measuring and comparing certain characteristics of their physiology or behaviour (Riley, Buckner, Johnson & Benyon, 2009:295-306).

Voice is a popular biometric for the following reasons

- natural signal to produce
- does not require a specialized input device
- Devices are easily available: telephones and microphone equipped PC

Speaker Authentication has become an attractive field for scientific investigation and exploration presently. The inspiration behind is the application of this biometric in several critically important sectors. For controlling access to protected resources Authentication of

personal identity is an essential requirement. Though robust speaker authentication remains a challenge, steady progress has been made in this sector till date. Bodo speech synthesis research is in early stage, recently a section of researcher have shown immense interest in this tribal language of the North-East India, which is recognized by the sixth schedule of Indian constitution and one of the instructional medium in the schools in Assam.

Section B: INTRODUCTION TO BODO LANGUAGE

The Bodo language belongs to the Tibeto-Burman language family, which is a sub-group of the Sino-Tibetan language group. It is one of the popular Indian tribal languages. Bodo language uses the **Devanagari script**. But, there is a huge difference in the usage of the letters in Bodo language from the Devanagari script. In bodo phonology there are six (6) pure vowels (monophthongs) and nine vowel glides or diphthongs along with sixteen (16) consonants. In this paper investigation will be limited to the following pure vowels.

Tables 1(A) and 1(B) gives a clear view of the vowels used in bodo language along with some very common words where the use of vowels are found in different positions.

	FRONT	CENTRAL	BACK
HIGH / CLOSE MID	इ /i/		उ /u/, औ /a^{w}/
MID / CLOSE MID	ए /e/		अ /o/
LOW / OPEN	आ /a/		

Table 1(A): BODO vowel phonemes and utterance positions.

Phonetic	I.P.A.	Manner & place of articulation	Bodo Examples with their meaning in English.		
			Initial	Middle	Final
इ	/i/ /इ/	High, front, unrounded	इनाइ/inai Evil	गिबि/gibi First	बोराइ/B*u*rai Old
उ	/u/ /उ/	High, back, rounded	उन्दु/undu Sleep	गुसु/gusu Cold	दुखु/dukhu Sorrow
ए	/e/ /ए/	Mid, front, unrounded	एनजर/enzor Rabit	बै/be this, it	बेसे/base how much
अ	/o/ /अ/	Mid, back, rounded	अमा/oma Pig	____	____
ओ	/*ɯ*/ /ओ/	High, back, unrounded	ओंखाम/*ɯ*ŋk^{h}am Rice	जोंनि/z*ɯ*ŋni Our	वोलो/b*ɯ*lu(force) गोसो/g*ɯ*su(mind)
आ	/a/ /आ/	Low, central unrounded	आं/aŋ I	जा/zat^{h}ai incident	आदा/ada elder brother

Bodo pure vowels with their I.P.A representation and some example of their use: Table 1(B):

Section C: Zero Crossing Rate

Zero Crossing Rate determines the information about the number of zero crossings present in a given signal. The concept behind zero crossing is to calculate how many times the signal waveform crosses the zero amplitude line by transition from a positive to negative or vice versa in a specific time.[25] In mathematical terms, a '***Zero Crossing***' is a point where the sign of a function changes (e.g. from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function. Spontaneously if the numbers of zero crossings are more in a given signal, the signal will be changed rapidly which implies that the signal contains the high frequency information. Like the similar way, if the numbers of zero crossings are less, the signal will be changed slowly denoting that the signal contains low frequency information. In the context of discrete-time signals, a zero crossing is said to occur if the successive samples have different algebraic signs. A zero crossing is said to occur when there is a sign difference between two successive samples. The rate at which zero crossings used to happen is a simple measure of the frequency content of a signal.

Since high frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero crossing rate and energy distribution with frequency.[2] The significance of ZCR are as follows:

1. Tool for end point detection. Detection process of when a speech utterance begins and ends in very important.[1] For the utterances that are from noisy environment, end point detection is quite difficult. For the silence section of the speech wave, ZCR is zero.
2. ZCR is proposed for sex determination and approximately 97% result for gender classification is obtained. ZCR is proposed for musical instrument identification and result reflects more effectively the difference in musical instrument.[3]
3. In general if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced. An algorithm of ZCR for voiced or unvoiced speech has been given below:

- Read the given speech signal S.
- Store the above signal into signal vector V.
- Read the specified window size Z.
- For i=(1:Z) 1 to Z
- Read and Store the first value of signal vector V into temporary variable T1.
- Read and Store the next value of signal vector V into another temporary variable T2.

- Compare the sign between T1 and T2
- If any sign change is detected from T1 to T2 then increment zero crossing counter as follows
- c = c + 1.
- T1 = T2.
- Next i (Increment i to get the next iteration for i and continue until all data in window size is reached or until all data in signal vector are processed).

Calculation of Zero Crossing Rate

The rate at which zero crossings used to happen is a simple measure of the frequency content of a signal. In case of a narrow band (NB) signal, the average zero crossing rate gives a reasonable way to estimate the frequency content of the signal. But in case of a broad band signal such as speech, it is much less accurate. Zero-crossing rate is measure of number of times the amplitude of the speech signals passes through a value of zero in a given time interval/frame.

$$ZCR = \frac{1}{N}\sum_{i=0}^{N-1} |sgn(x(i)) - sgn(x_{-1}(i-1))|$$

Where the values of function sgn() is defined as:

$$\text{sgn} = \begin{cases} 1 & \text{for } x(i) > 0 \\ 0 & \text{for } x(i) = 0 \\ -1 & \text{for } x(i) < 0 \end{cases}$$

As getting a completely noise free speech is improbable, because there will be some amount of background noise will always be there which interfere with the speech signal, meaning that the silent regions actually have quite high ZCR rate as the signal changes just one side of the zero amplitude to the other side and back again.[10] To nullify this possible effect or error in the ZCR value, there is a need of applying a tolerance threshold. In our experiments threshold value has been kept as ±0.001. Therefore, any Zero Crossing that is in the range -0.001 and +0.001will be rejected.

Section D: Experiments and Findings

For segmentation of the target speech signal sample, the free software Audacity is used and the segmented samples are stored in the database with .wav format. Due to the slowly varying nature of the speech signal, it is common to process speech signal into frames over which

then properties of the speech waveform can be assumed to remain relatively constant.[1]

Parameters used in our work are as follows:

1. Input vowel sound wave signals are recorded at the sampling frequency fs=16 kilo Hertz.
2. Hamming window specifications are kept as:

Window size=256 samples, window overlap= 100 samples.

Frame time =[(0:frameNumber – 1) * (frameSize –overlap) +0.5*frameSize] /fs

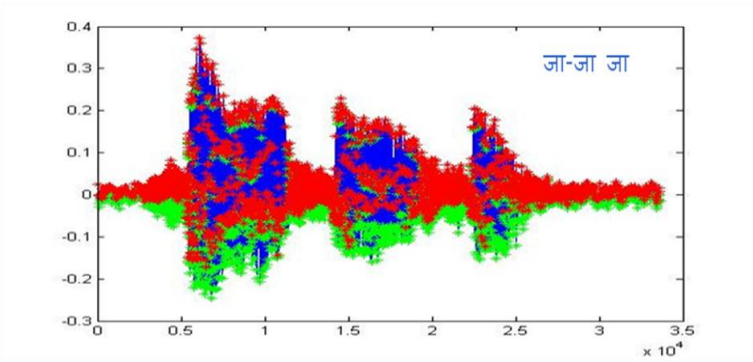
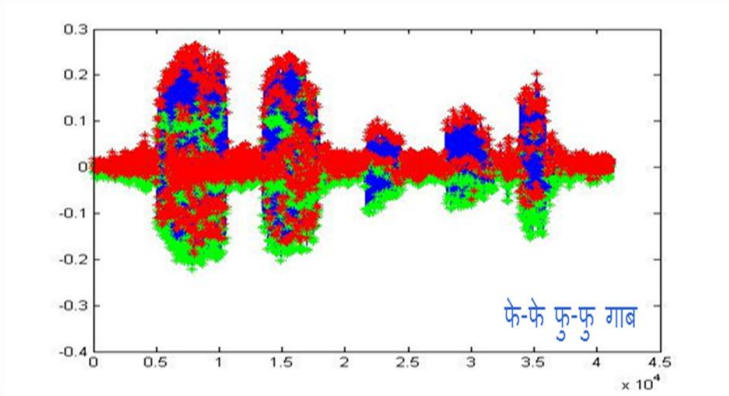


Figure 1: ZCR of BODO word जा-जा जा



फे-फे फु-फु गाब

Figure 2: ZCR of BODO words

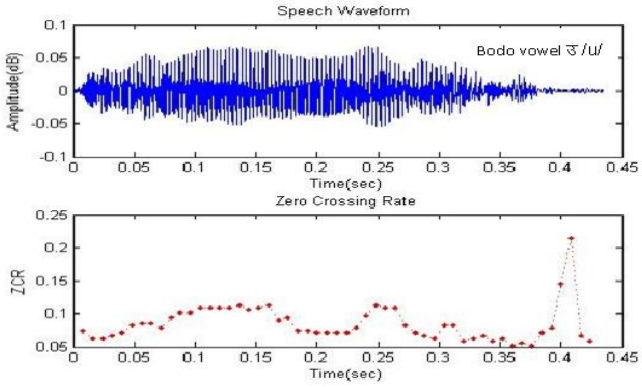


Figure 3: ZCR of BODO vowel उ /u/

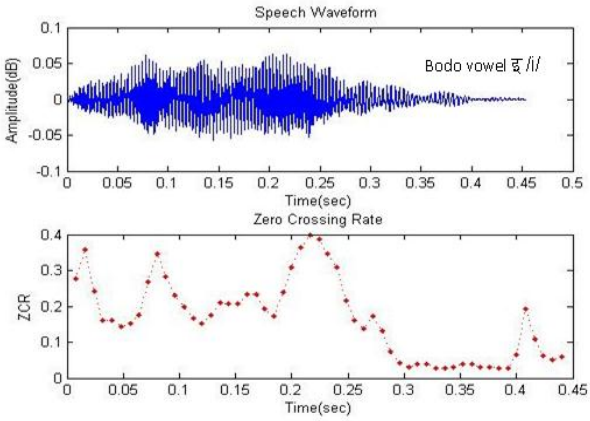


Figure 4: ZCR of BODO vowel इ /i/

Number of BODO male and female informants are taken 5 (five) each i.e. total ten speakers are considered in our experiments. Through Matlab program ZCR values are computed using step by step method applying the algorithm mentioned above for six (6) vowels of Bodo language, which are recorded frame by frame. Following tables depicts the behaviour of the recorded ZCR values:

Table 2: ZCR OF SIX BODO VOWELS UTTERED BY FEMALE INFORMANTS

Informants	Frames	Bodo vowels					
		/i/ /इ/	/u/ /उ/	/e/ /ए/	/o/ /अ/	/a/ /आ/	/*ɯ*/ /ओ/
Female1	0.0	46	47	47	49	50	50
	0.5	47	48	48	44	51	52
	1.0	43	50	49	47	53	55
	1.5	46	49	51	48	49	56
	2.0	47	50	50	49	48	58
Female2	0.0	48	50	50	48	50	54
	0.5	46	48	45	50	53	55
	1.0	50	52	50	52	48	54
	1.5	47	49	51	53	54	56
	2.0	48	50	49	49	56	55
Female3	0.0	52	49	48	52	50	53
	0.5	48	48	52	52	52	54
	1.0	47	50	50	51	51	55
	1.5	49	52	52	51	49	55
	2.0	52	51	49	50	49	56
Female4	0.0	47	52	50	54	53	58
	0.5	52	45	50	49	52	54
	1.0	54	49	51	50	51	55
	1.5	50	52	48	50	51	58
	2.0	51	50	53	52	50	53

Female5	0.0	45	48	50	50	50	54
	0.5	53	51	51	51	51	52
	1.0	50	52	50	51	50	56
	1.5	46	50	54	49	45	54
	2.0	45	50	49	53	53	53

Table 3: ZCR OF SIX BODO VOWELS UTTERED BY MALE INFORMANTS

Informants	**Frames**	**Bodo vowels**					
		/i/ /इ/	**/u/ /उ/**	**/e/ /ए/**	**/o/ /अ/**	**/a/ /आ/**	**/ɯ/ /ओ/**
Male1	0.0	52	52	50	52	48	46
	0.5	53	51	51	52	51	48
	1.0	52	55	50	50	46	52
	1.5	49	52	49	50	50	53
	2.0	49	52	50	51	50	55
Male2	0.0	51	52	49	54	50	51
	0.5	50	45	51	52	50	52
	1.0	53	51	49	52	48	51
	1.5	53	52	50	51	47	53
	2.0	52	49	50	50	52	52
Male3	0.0	53	55	49	50	48	51
	0.5	52	53	49	49	48	50
	1.0	53	52	50	49	49	50
	1.5	49	50	49	51	46	49
	2.0	53	52	51	48	50	50
Male4	0.0	52	53	50	49	48	55
	0.5	53	56	50	48	49	51
	1.0	53	55	49	52	49	52
	1.5	50	55	50	49	49	51
	2.0	51	53	49	48	52	50
Male5	0.0	52	50	49	49	49	53
	0.5	56	48	50	47	46	53
	1.0	53	52	49	48	51	49
	1.5	51	49	51	51	52	53
	2.0	54	49	51	48	50	49

In a classical way, to test authentication of speaker based on only the ZCR feature one single speaker is asked to utter the same vowel sound five times and the experiment is repeated for four more speakers (total five informants). These are performed in noise free studio environment. With repeated tests to avoid false rejection and false acceptance a threshold of ±5 is accepted which provides best result in our study. After the study of frame by frame values for a specific utterance by a particular informant for five times the following (table 4) trends were observed:

Table 4: Frame wise ZCR comparison for male and female speakers. Values beyond threshold are marked with Red.

Speaker (speaker number_ gender_ age)	Vowel uttered	ZCR for Bodo vowel utterances for five instances					
		Frame wise value Recorded in table 1 and 2	Utterance1	Utterance2	Utterance3	Utterance4	Utterance5
s1_f_20	/i/ /इ/	46 47 43 46 47	46 54 45 48 41	47 41 40 46 47	48 47 43 46 45	49 45 43 49 46	49 47 44 46 49
s2_m_20	/u/ /उ/	52 45 51 52 49	52 45 55 54 50	49 47 51 52 49	44 51 51 52 49	52 45 51 52 42	58 45 59 54 50

Utterance1 and utterance5 of speaker s1_f_20 are rejected. Similarly for speaker s2_m_20 utterance3 and utterance5 are rejected. In the same manner total five male speakers were tested for total six vowels for five repetition of the same vowel (5x6x5=150 utterances) and the same experiment is conducted for the female speakers. All together 150x2=300 utterances are tested.

Table 4 records the attempted authentication trials and the outcome. In our experiment which follows classical method of feature matching, authentication rate achieved is 79.67%. Though the final result is not satisfactory, if other features such as Short Time Energy, Mel Frequency cepstral coefficient, Linear Predictive coefficient are used along with ZCR are extracted and features are mixed to define the feature vectors, the authentication is expected to reach a very efficient and effective level.

Total number of speakers= 10 (5 male + 5 female) Total utterances by each speaker= 30 (6 vowels x 5 times)				
Grand total utterances= 300 (10 speakers x 30 utterances)				
	Successful	**Failed**	**Average**	**PERFORMANCE**
Speaker1	26	4	86.67	**79.67**
Speaker2	27	3	90.00	
Speaker3	21	9	70.00	
Speaker4	22	8	73.33	
Speaker5	22	8	73.33	
Speaker6	24	6	80.00	
Speaker7	25	5	83.33	

Speaker8	24	6	80.00	
Speaker9	25	5	83.33	
Speaker10	23	7	76.67	
				=80% (approximately)

Section E: result analysis and discussion

In our paper based on only single feature ZCR, a classical method of speaker authentication is represented for Bodo isolated speech words. It can be beneficial in developing a Speaker Identification and Authentication System for Bodo language. Due to the direct and indirect effects of various factors, speaker authentication possesses many inherent complexities making the process immensely difficult. The basic and natural rule is that each and every human being exhibits unique parameters in their speech waves. ZCR values can be invaluable contributor in this regard. Thus ZCR for all six vowel phonemes are studied through this paper and their behavior is monitored frame by frame in time domain. Experiment results achieves 80% authentication result which is moderate for a system but also proves worthy as it is based only on one feature. With a quality database and addition of more feature(s) will increase the authentication rate.

REFERENCES

1. Juang B.H., Rabiner Lawrence R. Rabiner (2004), Automatic Speech Recognition – A Brief History of the Technology Development'.
2. A U khan, L. P Baishya, S.K. Banchhor, "Hindi Speaking person identification using ZCR", IJSCE, july 2012; 2(3).
3. *S. K. Banchhor, A. Khan, "musical instrument recognition using Zero crossing Rate and Short Time Energy" , IJAIS, 2012; 1: 3.*
4. Baro Madhu Ram (1990),'Structure of Bodo Language', *N. L. Publications.*
5. Baro Madhu Ram, (2008), 'Structure of Bodo Language', *N.L Publications.*
6. Basumatary Phukan (2005), 'An Introduction to the Bodo Language', *Mittal Publication*
7. Bodo M.R. (1991), 'Assamese and Bodo, A Comparative Study', *Priyadini Publications.*
8. Borz. Porat, "A course in digital Signal Processing", John Willy & sons. 1997.
9. CATFORD, J. C. (1994), 'A Practical Introduction to Phonetics', *Oxford University Press.*
10. Chong F. L. (2004), 'Objective Speech Quality Measurement for Chinese Speech', *Master of Science Thesis, University of Canterbury.*

11. D. Talkin, (1987), 'Speech Formant Frequency estimation using dynamic programming with modulated transition cost', *AT&T Bell Lab, McGraw Hill, NJ.*
12. Allen J., Hunnicutt S., Klatt D. (1987). From Text to Speech: The MITalk System. Cambridge University Press, Inc.
13. Atal, B. S. and Hanauer, S.L. (1971), 'Speech analysis and Synthesis by Linear Prediction of Speech Wave', *J. Acoust Soc. Am.*, 50: pp 637-655.
14. Juang B.H., Rabiner Lawrence R. Rabiner (2004), Automatic Speech Recognition – A Brief History of the Technology Development'.
15. Rabiner L. R. and Schafer R. W., (2009), 'Theory and Application of Digital Speech Processing', pp 425-452.
16. Rabiner L.R. and Schafer R. W. (1978), 'Digital Processing of Speech Signals', *Englewood Clifts, N.J. Prentice Hill.*
17. Rabiner, L. R and B. H. Juang, (1993), 'Fundamental of Speech Recognition', *Prentice-Hall, Englewood Cliff, New Jersy.*
18. Talukdar P. H. (2010), 'Speech Production, Analysis and Coding', Lambert Academic Publishing.
19. Welling L. and Ney H. (1998), 'Formant Estimation of Speech Recognition', *IEEE trans, Speech and Audio processing Sept 1985; pp-134.*
20. *Costas panagiotakis and G. Tzigitas, "A speech/music discriminator based on RMS and Zero Crossing", IEEE transaction on multimedia. feb' 2005; 7(1).*
21. D.A.Reynolds, Speaker Identification and verification using Gaussian mixture speaker models, Speech Communication, 1995; 17(1): 91-108.
22. D. A. Reynolds, Experimental evaluation of features for robust speaker identification, IEEE Trans. Speech Audio Processing, 1992; 2(4): 639-643.
23. Zhang Wanfeng, Yang Yingchun, Wu Zhaohui, Sang Lifeng, Experimental Evaluation of a New Speaker Identification Frame work using PCA, IEEE International Conference on Systems, Man and Cybernetics, Oct 2003; 5: 4147- 4152.
24. G. Saha, Sandipan Chakroborty, Suman Senapat, A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications, Proceedings of the NCC 2005, Jan 2005.
25. M. Ito, R. Donaldson, "Zero-crossing measurements for analysis and recognition of speech sounds, IEEE Transactions on Audio and Electro acoustics", Sep 1971; 19, 3: 235 -242.

26. L. R. Rabiner , B. Atal, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, Jun 1976; 24(3): 201-212.
27. L. R. Rabiner, M. R. Sambur , " An algorithm for determining the end points of isolated utterances, Bell System Technical Journal (BSTJ), Feb 1975; 54: 297-315,
28. Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, Kong-Pang Pun, "An efficient MFCC extraction method in speech recognition, "IEEE ISCAS, 2006.
29. D. O'Shaughnessy, "Linear Predictive Coding", IEEE Potentials, 1988; 7(1): 29-32.
30. S. Furui, "Digital speech processing, Synthesis and Recognition", New York, Marcel Dekker, 2001.