Int. J. Open Problems Compt. Math., Vol. 3, No. 5, December 2010
ISSN 1998-6262; Copyright © ICSRS Publication, 2010
www.i-csrs.org

A Comparative Study for Identification of Sex of the Speaker With Reference to Bodo Vowels

[1]M.K. Deka, [2]S.K. Kalita and [3]S.K. Sarma

[1]Dept. of Instrumentation,
Gauhati University, Guwahati-14, Assam, India.
e-mail: manoj2007in@indiatimes.com

[2]Dept. of Computer Science,
Gauhati University: Kokrajhar Campus, Kokrajhar, Assam, India
e-mail: sanjib959@rediffmail.com

[3]Dept. of Computer Science,
Gauhati University, Guwahati-14, Assam, India.

Abstract

This work presents an application of Fundamental Frequency (Pitch), Linear Predictive Cepstral Coefficient (LPCC) and Mel Frequency Cepstral Coefficient (MFCC) in identification of sex of the speaker in speech recognition research. The aim of this article is to compare the performance of these three methods for identification of sex of the speakers. A successful speech recognition system can help in non critical operations such as presenting the driving route to the driver, dialing a phone number, light switch turn on/off, the coffee machine on/off etc. apart from speaker verification-caste wise, community wise and locality wise including identification of sex. Here an attempt has been made to identify the sex of Bodo speakers through vowel utterance by following Pitch value, LPCC and MFCC techniques. It is found here that the feature vector organization of LPCC coefficients provides a more promising way of speech-speaker recognition in case of Bodo Language than that of Pitch and MFCC.

Keywords: Fundamental Frequency or Pitch, Linear Predictive Cepstral Coefficient (LPCC), Mel Frequency Cepstral Coefficient (MFCC), sex of speakers, Bodo Language.

1 Introduction

The language used in the present study is Bodo language. Linguistically, the Bodo belongs to Sino-Tibeton group of languages [1]. The Bodo language has 22 phonemes which include 6 vowels, 14 consonants and 2 semi-vowels.

The Pitch, LPCC (Linear Predictive Cepstral Co-efficient), and MFCC as speech recognizer have been used by several workers [2] in the last two decades. This paper focuses on the identification of the speaker with reference to the sex of the speaker. How we can identify whether the speaker is male or female is studied in this paper through three techniques and gives a comparison among these techniques to find out which one technique is more reliable. In the present study, to reduce the volume of input data, clustering techniques [3].

2 Feature Extraction

In the present study, LPC-based Cepstral Coefficients, Mel Frequency-based Cepstral Coefficients and phonetically important parameters are used as feature vectors.

Speech signal is first subjected to low-pass filter to prevent the aliasing effect. The vowel speech waveform is then sampled at 8 KHz and quantized by a 16 bit resolution. The use of end point detection algorithm removes the unnecessary silence period at the beginning and at the end of the speech signal. To bring the signal in the spectral domain the signal is subjected to pre-emphasis procedure through a first order digital filter whose transfer function has been given by ($1-0.95z^{-1}$).Consecutive speech signal is blocked after every 31.25 milliseconds. 250 samples have been allotted to each of the 32 frames. Frames are overlapped by 20 milliseconds to produce a frame rate of 10 milliseconds. To reduce the undesired effect of Gibbs phenomenon[4], the frames are multiplied by a windows function (Hamming window), which is given by,

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \qquad (1)$$

where N is the number of sample in a block

3 Database of BODO vowels

The database used in the present study consists of isolated utterances of Bodo vowels recorded from 20 Bodo speakers comprised of 10 males and 10 females speakers. The testing data base has been created from 24 speakers from each group.

4 Fundamental Frequency Estimation

The opening and closing of the vocal folds that occur during speaking break the air steam into chains of pulses. The rate of repetition of these pulses is the pitch and it defines the fundamental frequency of the speech signal [4]. In other words, the rate of vibrations of the vocal folds is the fundamental frequency of the voice. The frequency increases when the vocal folds are made taut. Relative differences in the fundamental frequency of the voice are utilized in all languages to study the various aspects of linguistic information [5] conveyed by it. The general problem of fundamental frequency estimation is to take a portion of signal and to find the dominant frequency of repetition. Thus, the difficulties that arises in the estimation of fundamental frequency are (i) all signals are not periodic, (ii) those are periodic may be changing in fundamental frequency over the time of interest, (iii) signals may be contaminated with noise, even with periodic signals of other fundamental frequencies, (iv) signals which are periodic with interval T are also periodic with interval 2T, 3T etc., so we need to find the smallest periodic interval or the highest fundamental frequency, and (v) even signals of constant fundamental frequency may be changing in other ways over the interval of interest. In general, the fundamental frequency of the speech wave is estimated using autocorrelation. The mathematical model used for estimating the fundamental frequency is given below [6]:

A discrete short-time sequence is given by

$$sn[m]=s[m]w[n-m] \tag{2}$$

where w[n] is an analysis window of duration N_w. The short-time autocorrelation function $r_n[\tau]$ is defined by

$$s_n[m]=s[m]w[n-m]$$

$$=\sum_{m=-\alpha}^{\alpha} S_n[m]S_n[m+\tau] \tag{3}$$

where s[m] is periodic with period p, $r_n[\tau]$ contains peak at or near the pitch period p. For unvoiced sound no clear peak occurs near an expected pitch period. Location of the peak in the pitch period range provides a measure of pitch estimation and voicing decision. The above correlation pitch estimator can be obtained, more formally, by minimizing over possible pitch periods (p>0), and the error criterion is given by

$$E[p]=\sum_{m=-\alpha}^{\alpha} (S_n[m]-S_n[m+p])^2 \tag{4}$$

Minimizing E[p] with respect to p yields

$$p=max\left(\sum_{m=-\alpha}^{\alpha} (S_n[m]S_n[m+p]\right) \tag{5}$$

where $p > \epsilon$ i.e., p is sufficiently far from zero.

This alternative view of autocorrelation pitch estimation is used for detecting the pitch of Bodo vowels. The speech waveform and corresponding pitch spectra of six Bodo vowels have been depicted in Fig.-1 and Fig.-2 for both male and female informants. The estimated values of the fundamental frequency or pitch have been given in Table-1 for Bodo vowels.

Table 1: Pitch of six Bodo vowels

SEX OF THE SPECIMEN	FUNDAMENTAL FREQUENCY (Hz) FOR VOWEL					
	/a/	/e/	/i/	/o/	/u/	/w/
MALE	129.03	380.95	148.15	153.85	125.00	145.46
FEAMLE	228.57	242.42	266.67	250.00	117.65	250.00

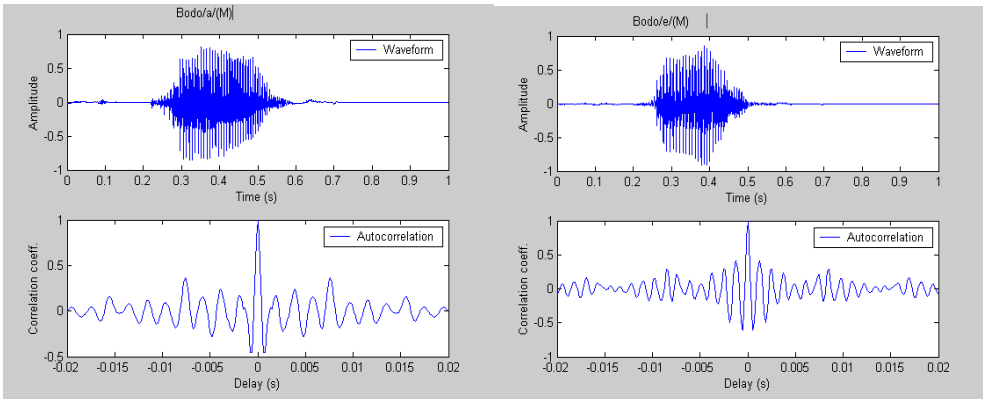

Fig 1: Estimation of typical pitch of Bodo vowel corresponding to male informants (time domain)

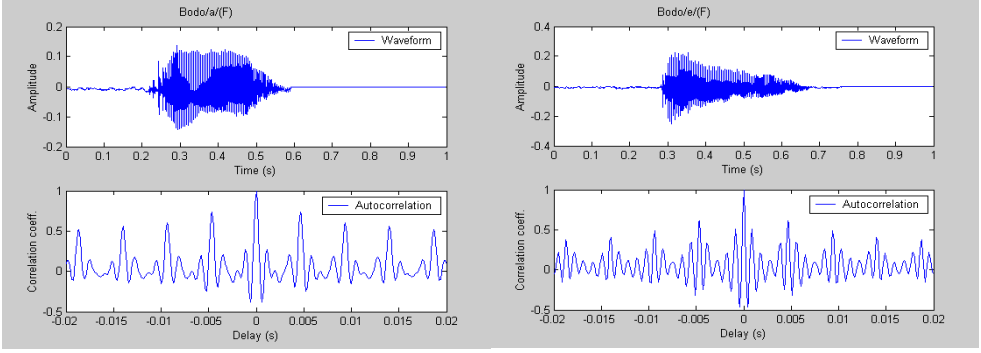

Fig 2: Estimation of typical pitch of Bodo vowel corresponding to female informants (time domain)

4.1 Results and Discussion

Typically, the pitch or fundamental frequency ranges from 80Hz to 160Hz for male speakers and from 140Hz to 400Hz for female speakers [3].The estimation of pitch finds extensive use in speech encoding, synthesis and recognition. In adult, generally the length of vocal folds in male is more than that of female counterpart. The more is the vocal fold length, less is the pitch frequency. Thus the pitch differs in male and female informants. In our present study, as given in Table-1, we observe that the values of pitch or fundamental frequency for female informant are higher than that of male informants as proposed by Pinto-et-al [7]. Thus, the pitch or fundamental frequency can be affectively used for sex verification of Bodo speakers.

5 Linear Predictive Cepstral Coefficients (LPCC)

Speech compression/Speech coding is a method for reducing the amount of information needed to represent a speech signal. LPC is one of the methods of compression that models the process of speech production. A digital method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. At a particular time, *t*, the speech sample s(t) is represented as a linear sum of the *p* previous samples [7]. The general algorithm for linear predictive coding involves an analysis or encoding part and a synthesis or decoding part. In the encoding, LPC takes the speech signal in blocks or frames of speech and determines the input signal and the coefficients of the filter that will be capable of reproducing the current block of speech. The LPC based cepstral coefficients are described as:

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, \qquad 1 \le k \le p \qquad (6)$$

$$c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, \qquad m > p \qquad (7)$$

where, c_k is an LPC based cepstral coefficient. It has been established that LPC based cepstral coefficients produce better recognition results when they are appropriately weighted. The weighting function is given as:

$$w(m) = 1 + \frac{Q}{2} \sin\left(\frac{\pi m}{Q}\right), \qquad 1 \le m \le Q \qquad (8)$$

where Q is the order of the cepstral coefficients. In addition to the cepstral coefficients, their time derivative approximations are used as feature vectors to account for the dynamic characteristic of speech signal. The time derivative is approximated by a linear regression coefficient over a finite window, which is defined as [1] –

$$\Delta\hat{c}_l(m) = \left[\sum_{k=-k}^{k} k\hat{c}_{l-k}(m)\right].G, \quad l \le m \le Q \qquad (9)$$

where $\hat{c}_l(m)$ the mth is weighted cepstral coefficients at time and is a constant, used to make the variances of the derivative terms, equal to those with the original cepstral coefficients.
In the present investigation, the following typical values are used –
n=240, p=10, Q=12, k=2 and G=0.316 [4], where p is predictor order.

5.1 Results and Discussion

In the present study, it has been observed that by careful selection of the feature set (Frame no 11, Frame no 12, Frame no 15) could increase the efficiency of the recognizer. It provides a basis for Bodo vowels recognizer. Fig 3, Fig 4, Fig 5 have shown the prominent distinction between male and female utterances for the Bodo vowel /a/.From our analysis for other vowels also, distinguishable feature vectors are found for male and female utterances for the same set of frames. The linear cepstral coefficients characteristics of six Bodo vowels, have been studied and depicted in Fig 3, Fig 4, Fig 5 for male & female informants respectively. It is clearly observed, that, the Bodo phonemes(vowels), show a clear distinction between the spectral characteristics of Male & Female while using LPCC. It is thus concluded in the present study that LPCC measure may be a better technique for speaker verification with respect to sex.

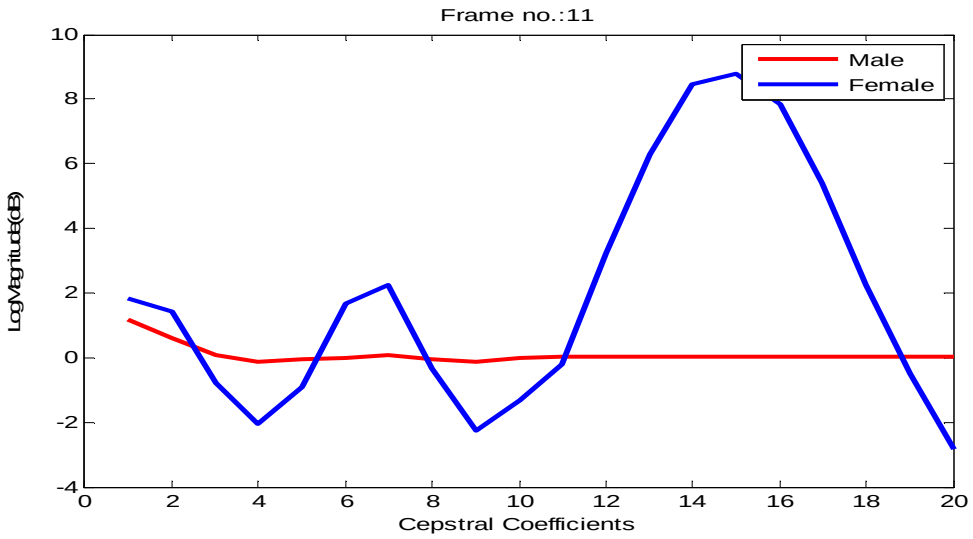

Fig 3: Cepstral Coefficients of the Bodo Vowel /a/ for Frame 11

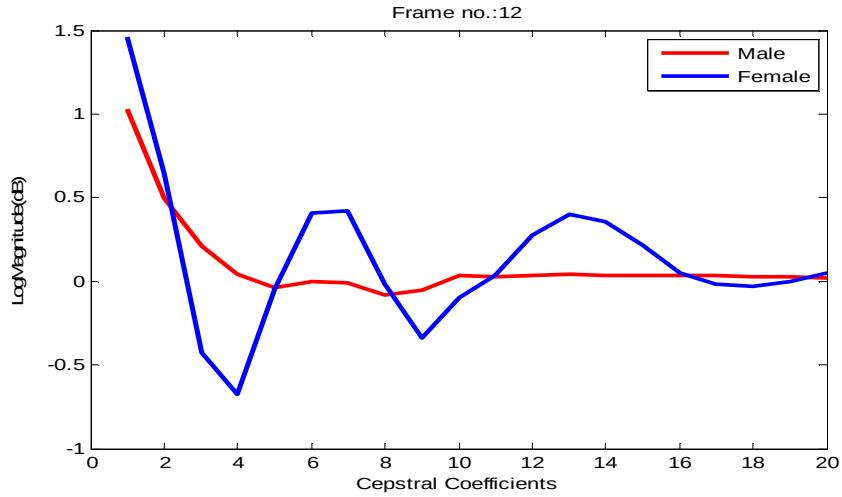

Fig 4: Cepstral Coefficients of the Bodo Vowel /a/ for Frame 12

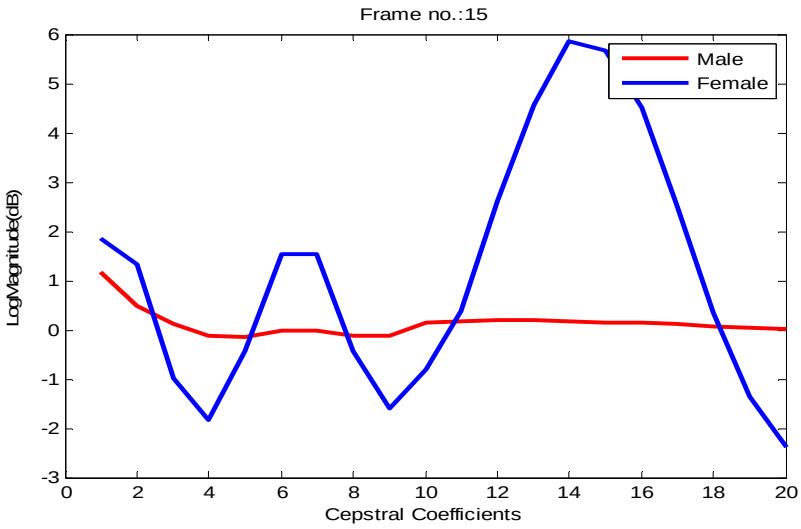

Fig 5: Cepstral Coefficients of the Bodo Vowel /a/ for Frame 15

6 Mel-Cepstrum

Instead of using linear prediction, another method is often used in speech recognition, namely the ***Mel-cepstrum***. This method consists of two parts: the cepstrum calculation and a method called Mel scaling. The cepstrum method is a way of finding the vocal tract filter H (z) with "homomorphic processing". Homomorphic signal processing is generally concerned with the transformation to linear domain of signals, combined in a nonlinear way. In this case, the two

signals are not combined linearly, as convolution can't be described as a simple linear combination. The speech signal s(n), can be seen as the result of a convolution between u(n) and h(n):

$$s(n)=b_0.u(n) * h(n) \quad (10)$$

where u(n) is a normalized excitation signal, b_0 is the gain of the excitation signal, and h(n) is a vocal tract transfer characteristic.

In frequency domain, in equation (10) is described as –

$$s(z)=b_0.U(z)H(z) \quad (11)$$

Since the term, representing excitation, U(z), and the vocal tract function, H(z), are combined multiplicative, it is difficult to separate them explicitly. But, if the log operation is applied to, the task will become additive:

$$Logs(z)=log(b_0.U(z)H(z))=log(b_0.U(z))+log(H(z)) \quad (12)$$

The additive property of the log spectrum also applies in an inverse transformation. The result of this operation is called a ***cepstrum***.

6.1 Measure of Mel Frequency Cepstral Coefficients (MFCC) for BODO Phonemes

In the earlier method of feature extraction, preferably, used the Linear Predictive Coding (LPC) [3, 4, 7]. Mel Frequency Cepstral Coefficient (MFCC) analysis has been widely used in signal processing in general and speech processing in particular [8]. This coefficient has a great success in speech recognition application.Fig.6 represents the algorithms involved while calculating Mel Frequency Cepstral Coefficients (MFCC).

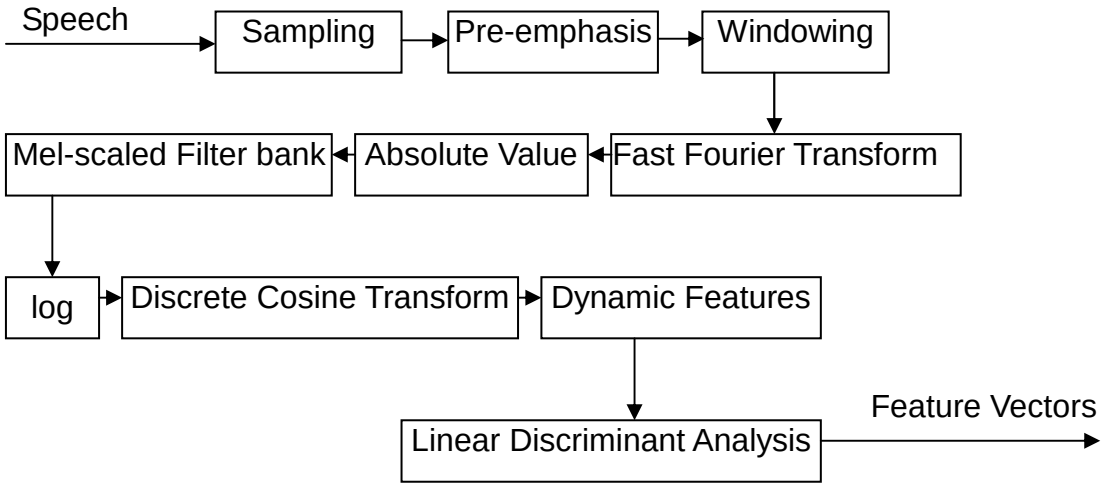

Fig 6: Algorithm for calculating Mel Frequency Cepstral Coefficients

6.2 Results and Discussion

The typical Mel frequency cepstral coefficients six Bodo vowels, corresponding to male and female informants, have been depicted in Fig 7, Fig 8 . As depicts in Fig.(7,8), it is clearly observed, that, the Bodo phonemes unable to show a clear distinction between the spectral characteristics of Male & Female while using MFCC i.e. MFCC technique does not provide sufficient clues to distinguish the male and female utterances.

Again, it is seen from the present study that range of the fundamental frequency for male is 80Hz to 160Hz whereas that of for female is from 140Hz to 400Hz [3].It is thus difficult to determine whether the sound is from male or female in the range from 140Hz to 160Hz.In this particular range the utterance may be from man or female, we cannot say definitely whether the sound is from man or female. In case of LPC, such type of problem is absent. In the frame no. 11, 12, 15 there are distinct differences between the male and female utterances as shown in Figures 3, 4, and 5.By observing these particular -numbers of frames we can easily and correctly differentiate the male or female sound. So, it can be concluded that LPCC method is more reliable than the Pitch based and MFCC based method for the identification of the sex of the Bodo informants.

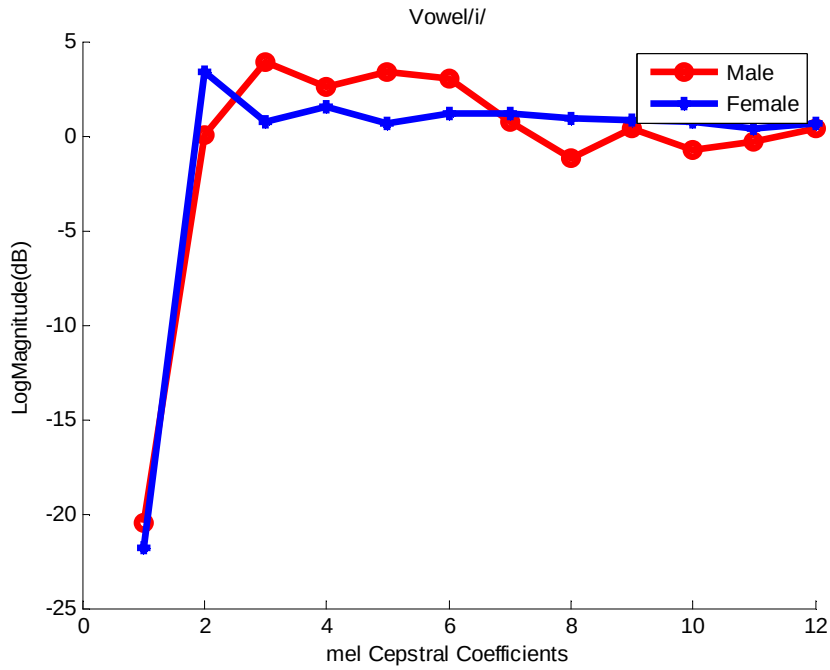

Fig 7: Mel Frequency Cepstral Coefficients of the Bodo Vowel /i/

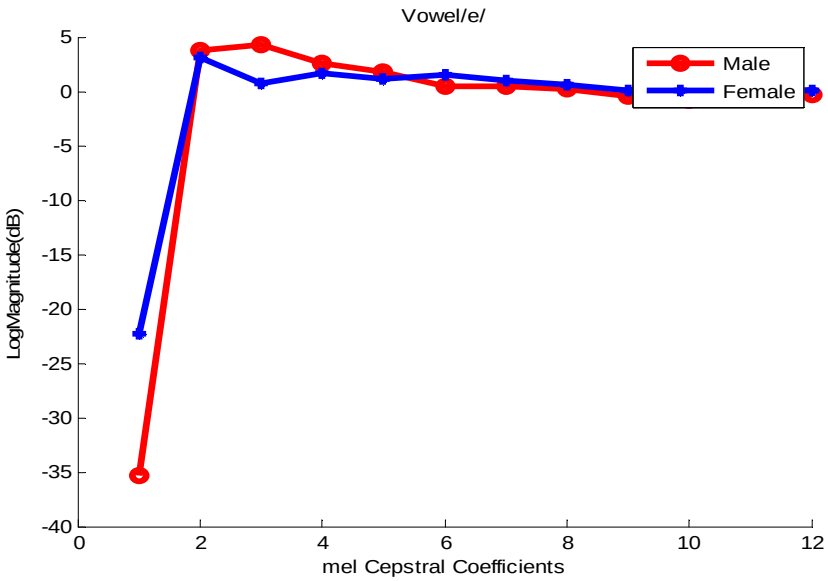

Fig 8: Mel Frequency Cepstral Coefficients of the Bodo Vowel /e/

7 Open Problem

Many other techniques like SOM, MSOM, SVM, HMM, ICA, PCA may be used with IPA standard sound.

References

[1] S.N. Goswami, Studies in Sino-Tibetan Languages, Published by M. Goswami, 1988.

[2] S. Furui, *IEEE Trans. Acoustic Speech, Signal Processing*, 34, 1986.

[3] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg and J.G. Wilpon, *IEEE Trans Acoustics, Speech, Signal Proc* ASSP-27,1979,336.

[4] J. Jantzen, Neurofuzzy Modeling, Technical Report 98-H-874, Technical University of Denmark: Oersted-DTU, 1998.

[5] J. Barman, S. Kalita, P.H. Talukdar, Feature Extraction of Bodo VowelsThrough LPC-Analysis, Proceedings of Frontiers of Research on Speech and Music (FRSM), 2004

[6] R. Lawrence, J. Biing-Hwang, *Fundamentals of Speech Recognitions*, Prentice Hall, New Jersey, 1993.

[7] L.R. Rabiner, K.C. Pan and F.K. Soong, On the performance of isolated word speech recognition using vector quantization and temporal energy contours, *AT and T Bell Lab. Tech. J.*, 63(1984), 1245– 1260.

[8] B.H. Juang, L.R. Rabiner and J.G. Wilpon, On the use of band pass filter in speech recognition, *IEEE Trans. On Acoustics, Speech and Signal processing*, (1987), 947 – 954.