Analysis and Recognition of Vowels in SHAI`YÂNG MIRI Language using Formants

Rizwan Rehman
Assistant Professor
Centre for Computer Studies
Dibrugarh University

Gopal Chandra Hazarika, Ph.D
Professor
Dept. of Mathematics
Dibrugarh University

ABSTRACT

Speech is generated by the co-ordination of various anatomical articulations known collectively as the vocal organs. The waveforms thus produced are either voiced or unvoiced speech. All the vowel sounds are voiced and have the sound source, the glottis. An amplification caused by the vocal tract filter is called resonance, and in speech these resonance are known as formants. Since each vowel has a different vocal tract shape, it will have a different format pattern. This paper is an attempt to analyze the vowels in SHAI`YÂNG MIRI language using the formant analysis. Analysis of the eight cardinal vowel speech samples in SHAI`YÂNG MIRI language is performed using formants, spectrogram, 19-channel auditory filter bank and through normalization of vowels.

General Terms

Speech Processing and Analysis, Speech Synthesis

Keywords

Formant, SHAI`YÂNG MIRI, MISHING, ASR, Vowel Analysis

1. INTRODUCTION

The nature of the speech signal and its acoustic properties can be studied by the analysis and presentation of speech signal in frequency domain [1]. In order to maintain the naturalness of oral communication between human and machines all aspect of speech must be involved. An automatic speech recognition system uses speech analysis as their first stage. All ASR systems are essentially pattern recognizers. Therefore, how well the formant frequencies are determined is essential aspect of most of the system for speech recognition and speech identification. Human listeners, to interpret a signal, might benefit from the pattern of the formant.

The vowels can be categorized by the temporal development of the formants [2]. The analysis and recognition of vowels can be used in the identification of vowels in continuous speech. The written text in most of the languages differs from the pronunciation therefore the correct pronunciation can be described by a set of symbolic representation [3]. In automatic speech processing two aspects are considered, first is the speech recognition and second is the speech synthesis [4]. Speech synthesis involves speech generation of voice waveform commonly generated from a written or stored text [5].

2. THE SHAI`YÂNG MIRI LANGUAGE

SHAI`YÂNG MIRI also known as MISHING language is spoken by the Mishing people residing mainly on the banks of Brahmaputra river. The origin of SHAI`YÂNG MIRI language is Tibeto-Burman language spoken by more than 500,000 people residing in Lakhimpur , Dhemaji, Sivasagar, Jorhat, Golaghat and Tinsukia district of Assam. SHAI`YÂNG MIRI language, in absence of its own script uses the Roman script for its lexicographical determinants. Therefore, there is a difference between the spoken form and the written form.

Based on the type of sound produced SHAI`YÂNG MIRI language consists of 41 phonemes of which 16 are consonants and 15 vowels including 10 allophones [6]. The vowels include /a/, /e/, /i/, /o/, /ô/, /û/, /ee/, /ea/. These eight vowels are the cardinal vowels with their seven long variances which are /ah/, /eh/, /ih/, /oh/, /uh/, /eyh/, /iuh/.

Table 1. Cardinal Vowels

Phoneme	Place of Articulation	Example
/a/	Front open	Esar
/e/	Central half open	Ta-peta
/i/	Front close	Moi`ya
/o/	Half open	Ko`pak
/ ô /	Back half close	Tat-p ôa
/u/	Back close	Bur
/ea/	Half close	Geelean
/ee/	Central close	Leegang

3. THE CHARACTERISTICS OF VOWEL

All the vowel sounds are voiced. The shape of the oral cavity and to some extend the shape of lips and duration for which sound is spoken, distinguishes one vowel sound from another. In case of vowels, speech is produced by the glottal source waveform travelling through the pharynx and as the nasal cavity is shut off, the waveform progresses through the oral cavity and is radiated into the open air via lips.

The system equation for vowel can be written as:

$$X(z) = G(z).A(z).B(z).C(z) \qquad (3.1)$$

Where G(z) is glottal source with A(z), B(z) and C(z) representing the transfer function of the pharynx, the oral cavity and lips respectively.

A(z) and B(z) are linearly combined so a single vocal tract transfer function:

$T(z) = A(z).B(z)$ can be defined such that equation (3.1) becomes:

$$X(z) = G(z).T(z).C(z) \qquad (3.2)$$

This can be represented in the time domain as:

$$x(n) = g(n) \times t[n] \times c[n] \qquad (3.3)$$

4. FORMANTS

Speakers with a lower fundamental frequency will have spaced harmonics whereas speakers with higher fundamental frequency will have widely spaced harmonics.

The amplification caused by the filters is called resonance and in speech these resonance are known as formants. Since each vowel has a different vocal tract shape, it will have different formant pattern and it is these that the listener uses as a main cue to vowel identity.

By convention formants are named F1, F2 and F3. The fundamental frequency is often called F0. The identity of vowel is nearly completely governed by the frequency of first two formants F1 and F2. The position of the harmonics depends on the fundamental frequency and the glottis whereas the spectral envelope is controlled by the vocal tract and hence it contains the information required for vowel and consonant identity.

4.1 Formant Structure of vowels in Shai`Yâng Miri Speech

The analysis of the 8 cardinal formants of Shai`Yâng Miri language were performed using the spectrum obtained by 19 filter bank as proposed by Holms' filters bank [7]. The Speech Filling System (SFS) software package is used for the analysis. SFS package is developed at University College London and is freeware used extensively in phonetics and linguistic. SFS package can be used for displaying both waveform and spectrogram as well as filter bank analysis among many other functions available in the package.

The short time spectrum can be used to study the acoustic properties of vowels [8]. We have examined the vowel speech using the 19-channel auditory filter bank and found that there can be many formants present in the spectrum of phoneme, but for the recognition purpose first three can be used.

The results of the analysis of vowels based on formants give better results. The analysis of the formant pattern of the vowel pronunciation shows that the first formant (F1) can be found in the range from 250 Hz to 1000 Hz while the range of the second formant (F2) is found in the range of 550 Hz to 2800 Hz. In figure 1 and figure 2, we have shown the SFS output of speech waveform, wideband spectrograms and output of 19-channel auditory filter bank for two vowels /a/ and /u/ in Shai`Yâng Miri language. The output clearly shows that there is a harmonic structure present in all the spectra. The lowest frequency peaks in the output of the filters represent the speakers' fundamental frequency while the other peaks are the resonant frequency of the vocal tract or we can say the formant.

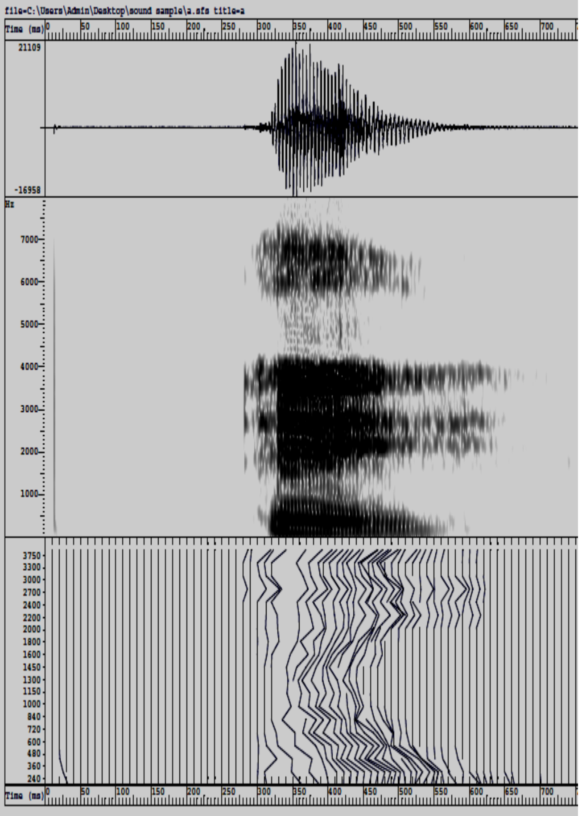

Fig 1: SFS output showing speech waveform, wideband spectrograms and 19-channel auditory filter bank for vowel /a/ in Shai`Yâng Miri language uttered by male speaker

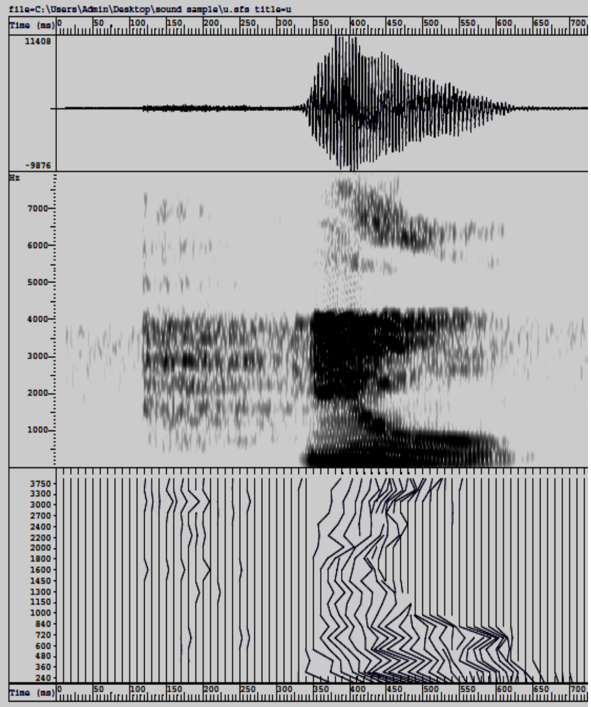

Fig 2: SFS output showing speech waveform, wideband spectrograms and 19-channel auditory filter bank for vowel /u/ in Shai`Yâng Miri language uttered by female speaker

5. LPC METHOD OF ANALYSIS

Linear Predictive Coding method is a popular speech analysis and synthesis technique used for encoding sound quality speech at a low bit rate. In LPC a simple system of speech production is used where we have an input source a[n] which passes through a linear time invariant filter b[n] to give the output speech signal c[n]. In the time domain this is

$$c[n] = a[n] \times b[n] \qquad (5.1)$$

The convolution makes it difficult to separate the source and filter in time domain it is achieved by transforming the system to the z-domain

$$H(z) = \frac{X(z)}{Y(z)} \qquad (5.2)$$

Assuming H(z) can be represented by all pole filters then

$$Y(z) = X(z).H(z) \qquad (5.3)$$

$$= \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}\ \mathrm{X(z)} \qquad (5.4)$$

Where $\{a_k\}$ are the filter coefficients.

By taking the inverse z-transform[2,] in time domain it becomes

$$\mathrm{c[n]} = \sum_{k=1}^{p} a_k y[n-k] + a[n] \qquad (5.5)$$

The coefficient of the filters are determined by minimizing the squared error between the real sample c[n] and the predicted samples at the output of the filters č[n]. The estimated sample č[n] depends on previous p samples according to

$$\check{c}[n] = \sum_{k=1}^{p} a_k c[n-k] \qquad (5.6)$$

where a_k are coefficient of LPC filter.

The error or the LPC residual at any time is given by

$$e_n = c[n] - \check{c}\,[n]$$

$$= c[n] - \sum_{k=1}^{p} a_k c[n-k]$$

The summed squared error E over the finite window length N is given by

$$E = \sum_n [e(n)]^2$$

Consider in one case that the predicted sample signal S_n is equal to the actual signal C[n] then,

$$C[n] = \sum_{k=1}^{p} a_k c[n-k] \qquad (5.7)$$

Taking z-transform of equation (5.7) we get

$$C(z) = \sum_{k=1}^{p} a_k s(z) z^{-k} \qquad (5.8)$$

Thus, the linear filter has the transfer function

$$C(z) = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \qquad (5.9)$$

The LP spectrum is obtained by plotting the magnitude response. Formant location can be obtained from the LP spectrum by considering the peaks.

6. EXPERIMENTAL PROCEDURE

The study was performed by recording the 8 cardinal vowels in Shai`Yâng Miri language spoken by 10 male and 10 female native speakers. The recording was done using Speech Filling System (SFS) package and then vowels were analyzed by measuring the formant of each vowel by using fmanal application of SFS package. This analysis provides a proper match between the tracks and the spectrogram. Formant values thus obtained are converted to BARK scale.

6.1 Procedure for Calculations

The transform used is that from Traunmüller (1990):

$$Bark = \left[\frac{26.81\,f}{1960+f}\right] - 0.53 \qquad (6.1.1)$$

where f is the frequency in Hz.

By converting all Hz measurement using the above equation i.e. 6.1.1, we can apply the same transform to all formant frequency obtained for any number of speaker.

6.2 Results

Table 2. Mean formant frequencies of phonetic vowel as produced by 10 male and 10 female native speakers

VOWELS	FORMANTS	
	F1 / Hz	F2 / Hz
/i/	368	2315
/e/	403	2119
/ea/	460	1720
/ee/	581	1551
/ a /	751	1255
/o/	558	1102
/ ô /	380	1080
/u/	316	1065

In order to compare the vowel uttered by different speakers of Shai`Yâng Miri language, vowel normalization is an important step. The normalization of vowels is done in our study for analyzing the vowels spoken by native Miri speakers. The result is shown in Table 3 below.

Table 3. Normalized Vowels of Shai`Yâng Miri language

VOWELS	FORMANTS	
	F1 (Bark)	F2 (Bark)
/i/	3.708	13.988
/e/	4.042	13.398
/ea/	4.566	12.000
/ee/	5.600	11.313
/ a /	6.897	9.935
/o/	5.411	9.119
/ ô /	3.824	8.995
/u/	3.192	8.909

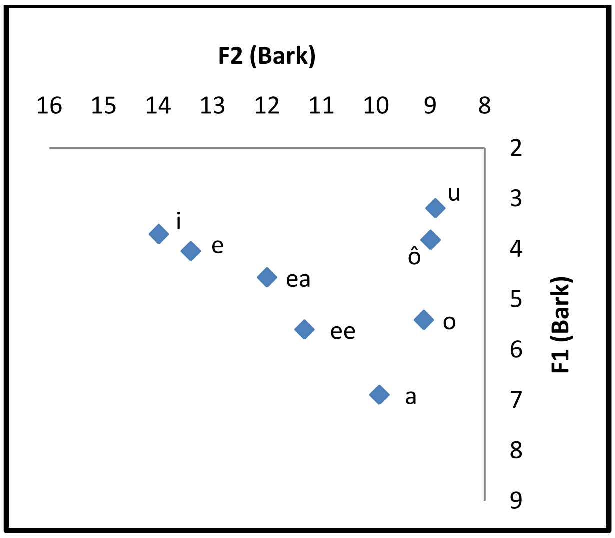

Fig 3: Data on Shai`Yâng Miri language in Table 3 plotted as a "vowel quadrilateral", with formant frequencies converted into Bark Scale

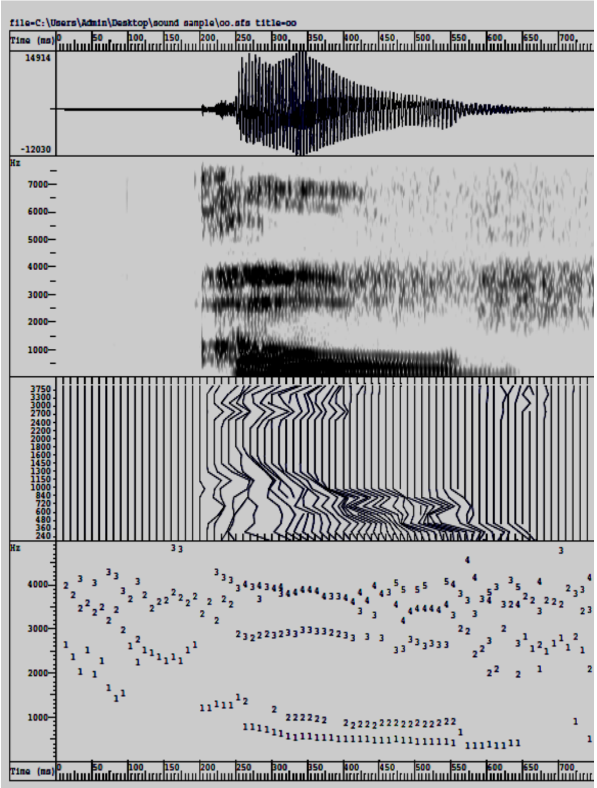

Fig 4: Example of SFS output showing speech waveform, wideband spectrograms and 19-channel auditory filter bank and fmanal output for vowel / ô /

7. CONCLUSION

In this paper, the authors have presented a way of recognizing the vowels of Shai`Yâng Miri language. The investigation is completely based on formant analysis, spectrogram analysis, analysis of 19-channel auditory filter bank output and normalization of vowels with respect to the formants F1 and F2 i.e. the first and the second formant. The LPC method is used for determining the frequencies and amplitudes of formants in speech.

The normalization points for each of the 8 cardinal vowels in Shai`Yâng Miri language have been found. The results were examined with the help of the SRS software for the recognition of vowels in continuous speech with a success rate of around 88-95%.

However, further investigation are planned to overcome the overlapping space in F1 and F2 plane, which would further enhance the success rate of recognition, and to utilize the result of this investigation in recognition of vowels more accurately in continuous speech.

8. ACKNOWLEDGMENTS

The authors would like to thank and acknowledge all the contributors and informants who helped in collecting the data and relevant information for the investigation.

9. REFERENCES

[1] Biljana Prica and Sinisallic. 2010 Recognition of Vowels in Continuous Speech by Using Formants.

[2] Y. A. Alotaibi and A. Hussain, "Comparative analysis of arabic vowels using formants and an automatic speech recognition system," International Journal of Signal Processing, Image Processing and Pattern Recognition vol. 3, 2010.

[3] Evandro Frozen and Dante Augusto Court Baronc, "Automatic Discovery of Brazilian Portugese Latter to Phoneme Conversion Rule Through Genetic Programming", Springer-2003.

[4] Dutoit and Thierry, " An Introduction to Text to Speech Synthesis", Dordrecht: Klower Academic, 1996.

[5] Lemmetty and Sami , "Review of Speech Synthesis Technology",Disponivel em:Acesso emzodez, 2000.

[6] J.F. NEEDHAM , "Outline Grammar of the Shai`Yâng Miri language ", Mising Agom Kcbang, 2003.

[7] J.Holmes, "The JSRU Channel Vocoder ", Proceedings IEE, vol. 127, 1980.

[8] L. Rabinder and B. Juang, "Fundamental of Speech Recognition", Prentice Hall, 1993.