Normalization and Automatic Recognition of Assamese Vowels

Bhaskarjyoti Sarma, Ph.D
Asst. Professor
Dept. of Assamese
Dibrugarh University

Gunadeep Chetia
Asst. Professor
Centre for Computer Studies
Dibrugarh University

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

ABSTRACT

Vowel normalization is a procedure of finding out the intermediate parameter from varied forms of utterance produced by different speakers of a particular language or same vowel token uttered by same speaker at different times. Every speaker has their own speech parameters and the variation of these parameters is the primary cause of speaker to speaker variation of speech sounds. The significant parameters that distinguish one vowel from the other are its formant values. In this paper an automatic vowel recognition procedure is described where the eight standard Assamese vowels are analyzed through Formant detection and Normalization. Here a standard approach is adopted for Normalization of vowels based on two Formants F1 and F2. A Vowel Recognition System based on Formant Estimation is developed and a preliminary recognition test for isolated vowels shows an expected recognition rate of 84-96% for arbitrarily selected speakers.

General Terms

Speech processing, Phonetics

Keywords

Vowel Normalization, Vowel Recognition, ASR, Formant Estimation, Assamese

1. INTRODUCTION

1.1 The Assamese language

Assamese is a major East Indian Language of Indo-European origin, spoken by almost 15.3 million people in Assam as their first language, and by more than 5 million people belonging to different linguistic communities of Assam as their second language. The Assamese language is also recognized by the Indian constitution as an official language of Assam. Assamese rather than being a language of a particular community, is more of a lingua franca for the most of the people of North-Eastern states. Assam being the host of a number of tribes (most of whom have their own distinctive language), Assamese served as a medium of communication between them. The differences of Assamese from other Indo-Aryan languages of India are mostly found in Phonology, Morphology, Syntax and in Vocabulary.

In Modern Standard Assamese (MSA) there are eight vowels, twenty one consonants, two semivowels, forty one two-vowel clusters where five two-vowel clusters are distinguished as substitute form and one hundred and forty one two-consonant clusters [1]. The distinctive vowels of MSA are /i, e, ɛ, a, ɔ̝, ɔ, o and u/, where /ɛ/ and /ɔ̝/ are two weak vowels. These two weak vowels shift their position according to the following vowels. For instance, if /ɛ/ and /ɔ̝/ is followed by high front /i/ and high back /u/, it is shifted to /e/ and /ɔ/, and if it is followed by low central vowel /a/, lower mid back /ɔ/ than /e/ and /ɔ/ are shifted to /ɛ/ and /ɔ̝/.

1.2 Dialects of Assamese

Dialects of Assamese also take an important role in the vowel perception by the people. The dialect of Assamese was broadly divided into two divisions, viz. Eastern Assamese Dialect (EAD) and Western Assamese Dialect (WAD) [2]. Later on the dialect variations of Assamese are further divided into three broad divisions, i.e. Eastern Assamese, Central Assamese and Western Assamese [3]. It has been observed that there are variations in the vowel perceptions in these dialects [4]. Considering all this dialectal variation, a Standard Assamese Dialect (SAD) is formed which consists of eight vowels in total.

1.3 Formants of Vowels

Vowels are voiced sounds which produce quasi-periodic pulses of air which are acoustically filtered as they propagate through the vocal tract. Formants are exactly the resonant frequencies of a vocal tract when pronouncing a vowel. [5] The formant with the lowest frequency is called F1, the second F2, and the third F3. Most often the first two formants, F1and F2, are enough to distinguish the vowel. These two formants determine the quality of vowels in terms of the open/close and front/back dimensions. Vowels generally have more than four distinguishable formants. However, the first two formants are the most significant in distinguishing the vowels and the vowel quality is often displayed in terms of a plot of the first formant against the second formant.

1.4 Objective of the Study

The objective of this study is to evaluate the variations of the vowel inventory of the Assamese language and also to compare the effectiveness of perception, to distinguish phonetic information and the dialectal or sociolectal differences distinct in phonemic level. Firstly, a Formant based analysis of the eight Assamese vowels is carried out and normalized to find out some standard formant ranges. Secondly, an Automatic Vowel Recognition System based on Formant Estimation is developed and vowels uttered by any arbitrary speaker are detected by comparing with the standard formant ranges.

2. NORMALIZATION OF VOWELS

2.1 Database Used

Required data are recorded in Speech Lab by using CSL-4500, in the Department of Assamese, Dibrugarh University. Ten male, ten female voices (between the age of 20-30) and 5 children voice (between the age of 8 to 12) are collected. All informants are from different regions of Eastern Assam. They are given to select carrier word where the vowel sounds are in initial, medial and final positions within the VC, CV and CVC syllabic structure. The male voices are recorded in 44.1 KHz, female and child voices are recorded in 16 KHz. Total recorded vowel tokens are ((10+10+5)*3)*3)*8= 1800. In close syllable the vowel tokens are taken in 15ms and in open syllable these are taken in 20 ms duration.

2.2 Formant Analysis

Ladefoged and Broadbent (1957) showed that three types of information are expressed when a speaker utters vowel sounds[6]. They are phonetic values, vocal tract shape and sociolinguistic variations between the speakers. From vocal tract shape the listeners can be able to distinguish speakers' gender. In terms of sociolinguistic variations, vowel utterance can give the information about the speakers' sociolinguistic group-characteristics. The first information is based on linguistic parameters and rest two provides speaker related information. Formant frequencies scientifically affect all three types of information [7,8, 9].

In day to day conversation the anatomical or physiological differences are generally ignored in perception level, where the listener can recognize the discourse of what they intend to express. It is due to sentence level perception capability of the listener. But in micro level research on vowel perception, especially when the research will have a goal to extract the data for computational perception, these differences are treated as redundant variation [10]

2.2.1 Formant plotting of Vowels

Using the CSL-4500 (KEYPENTEX) speech processing software, the beginnings and end points of the target vowels are placed in spectrographic displays. Formant tracks for the two formants, F1 and F2 are analyzed by using the Burg LPC algorithm implemented in CSL-4500, and visually checked by superimposing the tracks on a wideband spectrogram. When a mismatch is found between the tracks and the formant band in the spectrogram, the model order of the LPC-analysis is changed until a proper match is obtained between tracks and spectrogram. Once a satisfactory match is detected, the values for F1 and F2 are extracted at 15 to 20ms duration of the target vowel and stored for off-line statistical processing. Formant values are then converted to Bark and averaged over the ten male, ten female and five children speakers in each speaker group separately. These mean F1 and F2 values are plotted in acoustical vowel. In general various experiments show that the relationship between formant frequency of the vowels and identified vowel quality is not linear. Therefore, any two vowels produced from two different vocal tracts are measured by using the measurement of vowels normalizing method proposed by Lobanov [11]. The formula, which is followed in this study, is-

$F_n[v]N = (F_n[v]\text{-}MEAN_n)/S_n$

Where $F_n[v]N$ is the normalizing value for Fn[v] (i.e. for Formant of vowel. $MEAN_n$ is the mean value for formants for the speaker in question and Sn is the standard deviation of formants.) The common view amongst experimental phoneticians is that vowel quality can be quantified with adequate precision and validity by measuring the centre frequencies of the lower resonances in the acoustic signal [12]. The relationship between the formant frequencies and the corresponding perceived vowel quality is not linear. However, there is an empirical formula that adequately maps the differences in hertz-values onto the perceptual vowel quality domain, using the Bark transformation. Here the Bark formula advocated by Traunmüller is used to compute the perceptual distance between the vowel qualities from acoustic measurements [13].

Bark = $[(26.81 \times F) / (1960 + F)] – 0.53$
where, F represents the measured formant frequency in Hertz.

2.2.2 Normalization Results

Table 1. Normalized Vowels of MSA

	F1 (Hz)	**F2 (Hz)**	**F1 (Bark)**	**F2 (Bark)**
i	260.329	2268.791	2.543	13.908
e	347.104	2006.365	3.360	13.125
ɛ	406.421	1757.419	3.905	12.258
a	732.486	1445.971	6.637	10.953
ɔ	582.634	1121.783	5.440	9.256
ɔ̝	456.566	1075.097	4.346	8.978
o	378.354	1078.445	3.649	8.998
u	318.072	1067.818	3.090	8.927

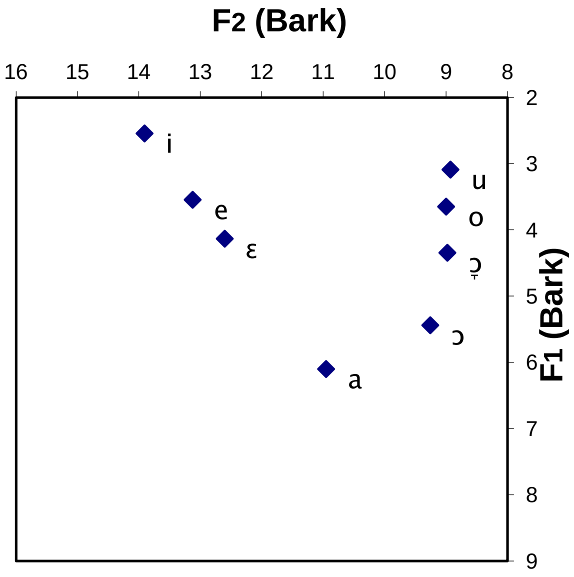

Fig 1: Formant plotting of Normalized vowels

2.2.3 *Range Estimation*

Table 2. Range of F1 and F2

	F1 (Hz)		F2(Hz)	
i	260	320	2240	2290
e	330	410	2030	1950
ɛ	390	460	1770	1720
a	690	740	1460	1430
ɔ	440	470	1140	1080
ɔ̣	370	450	1110	1050
o	340	400	1120	1060
u	290	360	1090	1030

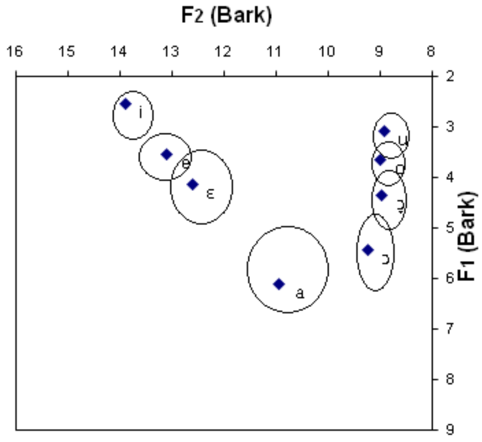

Fig 2: Range Plotting

3. AUTOMATIC RECOGNITION OF VOWELS

An automatic vowel recognition system was developed to recognize the vowels uttered by any native speaker. The Recognition system broadly consists of two stages, Formant Estimation from the input speech and comparison of estimated formants with the normalized range and detection of vowels accordingly.

3.1 Formant Estimation

For formant estimation, methods based on linear prediction analysis (LPC) have received considerable attention[14]. Root finding algorithms are employed to find the zeros of the LPC polynomial, or local maxima of the LPC envelope are searched using peak-picking techniques. The steps described below are followed to estimate formants from the input speech.

3.1.1 *Estimating Linear Prediction (LP) coefficients from the speech*

The first step is to perform the autocorrelation analysis of speech frames having length of 20 ms after multiplying it with a hamming window. After computing the autocorrelation sequence of this voiced frame of speech, the Toeplitz auto correlation matrix is generated as explained below.[15]

A given speech sample at time n, $\hat{s}(n)$, can be approximated as a linear combination of the past p samples such that

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k . s(n-k) \qquad (1).$$

where a_ks are the linear prediction coefficients and s(n) is the windowed speech sequence obtained by multiplying short time speech frame with a hamming or similar type of window which is given by,

$$s(n) = x(n).w(n) \qquad (2)$$

where ω(n) is the windowing sequence

We now form the prediction error e(n), defined as

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k . s(n-k) \qquad (3)$$

The total prediction error can be represented as follows

$$E = \sum_{n=-\infty}^{\infty} e^2 (n) \qquad (4)$$

which, using the definition of e(n) in terms of s(n) can be written as

$$E = \sum_{n=-\infty}^{\infty} [s(n) + \sum_{k=1}^{p} a_k . s(n-k)]^2 \qquad (5)$$

To solve equation (5) for the predictor coefficients differentiation of E is done with respect to each a_k and set the result to zero for k = 0,1,2,...p.

$$\frac{\partial E}{\partial a_k} = 0$$

giving

$$\sum_{n=-\infty}^{\infty} s(n-i).s(n) = \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s(n-i\_.s(n - k) \qquad (6)$$

where i=1, 2, 3...p. The equation (6) can be written in terms of autocorrelation sequence R(i) as follows,

$$\sum_{k=1}^{p} a_k R(i-k) = R(i) \qquad (7)$$

for i=1,2,3...p.

where the autocorrelation sequence used in equation (7) can be written as follows,

$$R(i) = \sum_{n=i}^{N-1} s(n)s(n-i) \qquad (8)$$

This can be represented in the matrix form as follows,

$$R.A = -r \qquad (9)$$

where R is the p x p symmetric matrix of elements R(i, k) = R(|i-k|), (1<=i, k<=p), r is a column vector with elements (R(1),R(2), ...R(P)) and finally A is the column vector of LPC coefficients (a(1), a(2), ...a(p)). It can be shown that R is Toeplitz matrix and can be represented as

$$R = \begin{bmatrix} R(1) & R(2) & R(3) & \dots & R(P) \\ R(2) & R(1) & R(2) & \dots & R(P-1) \\ R(3) & R(2) & R(1) & \dots & R(P-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R(P) & R(P-1) & R(p-2) & \dots & R(1) \end{bmatrix} \qquad (10)$$

Now, the LP coefficients can be computed as given in the following equation

$$A = -R^{-1}.r \qquad (11)$$

where R^{-1} is the inverse of the matrix R.

3.1.2 Plotting the LP Spectrum

In the frequency domain, the equation (3) can be represented as,

$$E(z) = S(z) + \sum_{k=1}^{p} a_k \, .S(z)z^{-k} \qquad (12)$$

From equation (12) we can construct the transfer function H(z) as follows

$$H(z) = \frac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}} = \frac{1}{A(z)} \qquad (13)$$

The frequency response of the LP filter is obtained by using z=e^{jw} in H(z) and the magnitude response is plotted to get what is known as LP spectrum. From the LP spectrum formant locations can be found by picking the peaks in the LP spectrum. A peak-picking algorithm is used to pick the peaks and estimate the formants. [16]

3.1.3 Comparison and detection of Vowels

After estimating the formants from the input speech signal, they are compared against the normalized formant range and vowels can be detected accordingly.

4. CONCLUSIONS

In this paper, an approach of Automatic recognition of Assamese vowels is presented. The method used is based on Formant Analysis and Normalization of Vowels in terms of first two formant values. After finding the normalization points, a standard range of formant values are estimated for each of the eight vowels in MSA. The automatic recognizer compares the formant values of the input vowels uttered by arbitrary speaker and recognize the vowels. The recognition rate is 84-96%. It has been found that the low recognition rate occurs for those vowels whose F1-F2 space overlaps, thus creating confusion in the recognizer. Further analysis will help to draw the exact boundaries in the overlapping region by examining the transitional features of these vowels and improve the recognizer.

5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the valuable contribution of the informants towards collection of the data.

6. REFERENCES

[1] Goswami, G.C,1990 Varna Prakash, Bani Mandir, Gauhati.

[2] Grierson, G.A, 1951. Linguistic Survey of India. Vol. III, Part-II & III, New Delhi: Motilal Banarshidas

[3] Goswami.G.C, Structure of Assamese , Gauhati: Gauhati U P, 1982. Print.Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Sarma, B. 2012, A Study of the Spoken Assamese in Upper Assam: An Ethnolinguistic Approach, Ph.D Thesis.

[5] Ladefoged, Peter, Vowels and Consonants: An Introduction to the Sounds of Language, Maldern, MA: Blackwell, p. 40.

[6] Ladgefoged an Broden 1957, ‘‘Information conveyed by vowels,’’ J. Acoust. Soc. Am. **29**, 88–104

[7] Peterson, G. E., and Barney, H. L. **1952**. ‘‘Control methods used in the study of the vowels,’’ J. Acoust. Soc. Am. **24**, 175–184.

[8] Hindle, D. **1978**. ‘‘Approaches to formant normalization in the study of natural speech,’’ in Linguistic Variation, Models and Methods, edited by D. Sankoff Academic, New York.

[9] Labov, W. 2001. Principles of Linguistic Change: Vol. II: Social factors Blackwell, Oxford

[10] Pols, L. C. W., Tromp, H. R. C., and Plomp, R. 1973. ‘‘Frequency analysis of Dutch vowels from 50 male speakers,’’ J. Acoust. Soc. Am. 53, 1093– 1101.

[11] Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. The Journal of the Acoustical Society of America, 49.2:606-608..

[12] H Wang and V. J. Ven Heuven, "Acoustical Analysis of English Vowels produced by Chinese, Dutch and American speakers"(2006)

[13] Traunmuller, H., "Articulatory and perceptual factors controlling the age- and sex- conditioned variability in formant frequencies of vowels", Speech Communication 3,1(1984): 49-61.

[14] L. Rabiner and B. Juang, Fundamentals of speech recognition. Prentice Hall, 1993.

[15] H. Widom, "Toeplitz Matrices" in Studies in Real and Complex Analysis, edited by I.I. Hirschmann, Jr., MAA Studies in Mathematics, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[16] S. McCandless,An algorithm for automatic formant extraction using linear prediction spectra. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-22.