International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-9)

Research Article | **September 2015**

Spectral and Spectrographic Analysis with Reference to Some Languages of North East India

Dr. Rashmi Dutta
Associate Professor, Department of Physics
North Gauhati College, Guwahati 781031, Assam, India

Abstract:

Spectral and Spectrographic analysis is an emerging area of research in Physics and has vast application in the study of phonemes of different languages . For our study, we take up three important link- languages of North East India, namely Assamese, Bodo and Rabha. This paper is concerned with two objectives: the first objective deals with the study of power coefficients [2,5,8] and the second is devoted to the spectrographic analysis [1,4,7,9] of the Assamese, Bodo and Rabha vowels , and estimations and analysis of different spectrograms with their impact on the overall structural features of these three languages. In fact, this paper represents the pattern congruity of a few Assamese , Bodo and Rabha phonemes in terms of spectral envelope. Initially, the acoustic waveforms, plotting amplitude against time (time spectra) are obtained for the selected `Speaker dependent' Assamese, Bodo and Rabha vowels, and a comparative and constructive study of these selected vowels of Assamese , Bodo and Rabha speech is made parallel to the linguistic interpretation [3,4,7].

Key Words: Spectral Analysis / Phonemes / Spectrograms/ linguistic interpretation

I. INTRODUCTION

The Data Set:

When investigating the properties of speech, it is important to break phrases down into their constituent building blocks. Phrases are broken down into words and then words are broken down into small segments, each of which is unambiguously distinguishable. These segments are called **phonemes** . For connected speech, the length of time that a speaker typically sustains and individual vowel is extremely short. Performing an accurate analysis of such small segments of data is highly impractical and therefore one has to compromise the naturalness of the speech in order to produce sustained vowel sections. Therefore, in this analysis we have asked each speaker to say isolated vowel sounds for approximately 1 second. The recording equipment consists of a Pentium IV , 300 MHz HP personal computer with an Ultrasound Max Sound Card employing a sampling rate of 8 KHz and a 16- bit A/D converter. A head-mounted microphone was placed to the side of the subject's mouth to reduce breath noise. Special acquisition software, Goldwave Version 4.26 enabled the capture of clean vowel sounds . To reduce possible noise contamination the subject was placed away from the computer in a sound-proofed booth. Consequently, we can consider the data to be stationary. In total 60 subjects (20 each, 10 males and 10 females, for Assamese , Bodo and Rabha) were recorded. They were speakers in the age group 25-40 years, chosen from Assamese , Bodo and Rabha dominated sampled areas respectively, having educational qualification graduate or more. Each subject was asked to read their respective set of vowel phonemes, each repeated 5 times. Before discussing spectrograms, we first pay our attention to finding the power coefficients:

II. DETERMINATION OF POWER COEFFICIENTS

In signal modeling, power measure (s) is important for parametric representation [1,3,5]. Power is computed by the formula :

$$\mathbf{P(n)} = \frac{1}{\mathbf{N}} \sum_{\mathbf{m}=0}^{\mathbf{N_s}-1} \left(\boldsymbol{\omega}(\mathbf{m}) \mathbf{s}\left(\mathbf{n} - \frac{\mathbf{N_s}}{2} + \mathbf{m} \right) \right)^2 \qquad (1)$$

where , $\mathbf{N_s}$ is the number of samples used to compute the power,

$\mathbf{s(n)}$ denotes the signal , $\boldsymbol{\omega}(\mathbf{m})$ denotes a weighting function and

n denotes the sample index (discrete time) of the center of the window.

The weighting function in (1) is referred to as a window function. The purpose of the window is to weigh, or favor, samples towards the center of the window and obtain smoothly varying parametric estimates. In speech recognition systems, $\boldsymbol{\omega}(\mathbf{n})$ is taken as

$$\boldsymbol{\omega}(\mathbf{n}) = \frac{\boldsymbol{\alpha}_{\boldsymbol{\omega}} - (1 - \boldsymbol{\alpha}_{\boldsymbol{\omega}}) \cos(2\pi \mathbf{n} / (\mathbf{N_s} - 1))}{\boldsymbol{\beta}_{\boldsymbol{\omega}}} \quad \text{for } 0 \le \mathbf{n} < \mathbf{N_s}, \qquad (2)$$

and $\boldsymbol{\omega}(\mathbf{n}) \equiv 0$ elsewhere. $\boldsymbol{\alpha}_{\boldsymbol{\omega}}$ is defined as a window constant in the range [0,1] and $\mathbf{N_s}$ is the window duration in samples. To implement a **Hamming window** [3,8,9] , $\boldsymbol{\alpha}_{\boldsymbol{\omega}}$ = 0.54 . $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ is a normalization constant defined so that the root mean square (rms) value of the window is unity. Mathematically, $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ is defined as

$$\beta_{\omega} = \sqrt{\frac{1}{N_s}\sum_{n=0}^{N_s-1}\omega^2(n)} \qquad (3)$$

Power like most parameters in a speech recognition system including fundamental frequency is computed on a frame-by-frame basis.

Frame duration T_f is defined as the length of time (in seconds) over which a set of parameters is valid. In equation (2), **n** is updated by the frame duration in samples. T_f typically ranges between 20 and 10 msec., in practical systems. N_s is known as the window duration (in samples). Window duration T_w is normally measured in units of time (seconds). Frame duration and window duration are normally adjusted as pair : for T_w = 30 msec., we take T_f = 20 msec., and for T_w = 20 msec., we take T_f = 10 msec.

III. MAIN FINDINGS

Experimental Results:

3.01: Objective : To calculate power coefficients with the help of a computer program in MATLAB.

Database : Mentioned in Section 2

3.01.1 Algorithm for power :

```
x = ('input signal')
[b, a] = firl (1.95);
y = filter (b,a,x);
y11 = y (1: 240);
w = hamming (240);
y11w = y11. *w;
p11 = (y11w). ^2;
y12 = y1 (81 : 320);
y12w = y12. *w ;
p12 = (y12w). ^2;
y13 = y1(161:400)
y13w = y13.*w;
p13 = (y13w).^2;
fs = 8000;
p1 = (p11 + p12 +p13)/(240*3);
stairs ((1:240)/240*fs, (10. ^8) *p1);
Total Power
 w= hamming (8000);
yw= y.*w;
p= yw.*2;
power = sum(p) / 8000;
```

After computing, we get the following results:

Table 3.1 : Table for Power Coefficients for Assamese and Bodo vowels

Assamese vowels	Male	Female	Bodo Vowels	Male
/a/	0.0017	0.0029	/a/	$8.9342e^{-0.05}$
/aa/	0.0048	0.0055	/e/	$1.7945e^{-0.04}$
/e/	$1.3345e^{-0.04}$	$3.9157e^{-0.04}$	/i/	$2.1674e^{-0.04}$
/i/	$4.6541e^{-0.05}$	$6.9510e^{-0.05}$	/o/	$1.7990e^{-0.04}$
/ea/	0.0021	0.0052	/u/	$3.2393e^{-0.04}$
/o/	0.0022	0.0011	/w/	$2.8337e^{-0.04}$
/u/	$3.8998e^{-0.04}$	$9.3518e^{-0.04}$		
/w/	0.0018	0.0042		

Power coefficients of Rabha vowels

Vowel	**Female Speaker**	**Male Speaker**
/a/	0.0071	0.0013
/aa/	0.0072	0.0070
/i/	0.0028	0.0047
/e/	0.0055	0.0034
/u/	0.0046	0.0044
/w/	0.0048	0.0072

3.03 Spectrograms

Speech waveform consists of sequence of different events. The time variation corresponds to highly fluctuating spectral characteristics over time[2,3,5,6]. A single Fourier transform of the entire acoustic signal cannot capture the time varying frequency contents for all the harmonics present. In order to capture the time varying nature of the speech signal, another Fourier Transform, called Short-time Fourier Transform (STFT) is used. It consists of a separate Fourier Transform on a piece of the waveform under a sliding window, which is represented by $w[n, \tau]$, where τ is the position of the window centre and n is the number of sample per window.

The Fourier Transform of the windowed speech waveform, i.e., STFT [4] is given by

$$x\ (\omega,\tau) = \sum_{\mathbf{n}=-\infty}^{\infty} \mathbf{x}[\mathbf{n},\tau]\exp[-\mathbf{j}\omega\mathbf{n}] \qquad \text{---------- (4)}$$

where $x[n, \tau]=w[n, \tau]x[n]$ represents the windowed speech segments as a function of the window centre at time τ. The spectrogram is a two dimensional representation of the time dependent spectrum in which the vertical dimension on the paper represents frequency and the horizontal dimension represents time. The spectrogram magnitude is given by

$$s[\omega,\tau]=|x(\omega,\tau)|^2 \qquad \text{---------- (5)}$$

Basically, there are two types of spectrograms - (a) **wideband spectrogram** and (b) **narrowband spectrogram.** The difference between these two types of spectrograms is the length of the window $w[n, \tau]$. The wideband spectrogram displays good temporal resolution and poor frequency resolution. On the other hand, the narrowband spectrogram display the good frequency resolution and poor time resolution.

For voiced speech, the output of a linear time-invariant system with impulse response h[n] and with a glottal flow input given by convolution of the glottal flow over one cycle g[n], with the impulse train is given by

$$P[n] = \sum_{\mathbf{k}=-\infty}^{\infty} \delta[\mathbf{n}-\mathbf{kp}] \qquad \text{---------- (6)}$$

In the windowed speech waveform the result can be expressed as

$$x[n, \tau] = w[n, \tau]\{(p(n)*g(n))*h(n)\}$$

$$= w[\mathbf{n},\tau]\left(\mathbf{p}[\mathbf{n}]*\bar{\mathbf{h}}[\mathbf{n}]\right)^{-} \qquad \text{---------- (7)}$$

where, the glottal waveform, over a cycle, and vocal tract impulse response are lumped into $\bar{\mathbf{h}}[\mathbf{n}] = \mathbf{g}[\mathbf{n}]*\mathbf{h}[\mathbf{n}]$. Using Multiplication and Convolution theorem, the Fourier Transform of the speech is given by

$$X(\omega,\tau) = \frac{1}{\mathbf{p}}\mathbf{W}(\omega,\tau)\otimes\left[\mathbf{H}(\omega)\mathbf{G}(\omega)\sum_{\mathbf{k}=-\infty}^{\infty}\delta(\omega-\omega_{\mathbf{k}})\right]$$

$$= \frac{1}{\mathbf{p}}\sum_{\mathbf{k}=-\infty}^{\infty}\mathbf{H}(\omega_{\mathbf{k}})\mathbf{G}(\omega_{\mathbf{k}})\mathbf{W}(\omega-\omega_{\mathbf{k}},\tau)$$

$$= \frac{1}{\mathbf{p}}\sum_{\mathbf{k}=-\infty}^{\infty}\bar{\mathbf{H}}(\omega_{\mathbf{k}})\mathbf{W}(\omega-\omega_{\mathbf{k}},\tau) \qquad \text{……..(8)}$$

where $\bar{\mathbf{H}}(\omega_{\mathbf{k}}) = \mathbf{H}(\omega_{\mathbf{k}})\mathbf{G}(\omega_{\mathbf{k}})$ and $\omega_{\mathbf{k}} = \frac{2\pi\mathbf{k}}{\mathbf{p}}\ \mathbf{and}\ \mathbf{d}\frac{2\pi}{\mathbf{p}}$

is the fundamental frequency.

Therefore, the spectrogram of x[n] can be expressed as

$$S(\omega,\tau) = \frac{1}{\mathbf{p}^2}\left|\sum_{\mathbf{k}=-\infty}^{\infty}\bar{\mathbf{H}}(\omega_{\mathbf{k}})\mathbf{W}(\omega-\omega_{\mathbf{k}},\tau)\right|^2 \qquad \text{---------- (9)}$$

The wideband and the narrowband spectrograms display a great deal of information about the properties of a speech utterance. The characteristics of the consonants can be better represented and separate from each other by spectrogram analysis [7, 8].

Usually, the formant frequencies are greater than the corresponding pitch or fundamental frequency for particular speech signal. While encoding, synthesizing and recognizing the speech signal, the formant frequencies and pitch or fundamental frequencies find extensive use. The time-frequency analysis of speech signal is also used extensively in the study of human speech [1, 5]. The spectrograms which are nothing but the squared magnitude of the STFT, plays important role while visualizing the time-varying frequency content of a speech signal [8]. Of course, the STFT has limited resolution. This limitation is subsequently overcome the use of mixed time-frequency signal representation, which is substantially different from the spectrogram. This was first proposed by Wigner 1932 and Ville, 1958 and the technique is known as Wigner-Ville distribution (WVD). The WVD is the FT of the autocorrelation of the signal obtained from the Hilbert transform of the original speech signal. A major problem with WVD technique is the interferences between two signal components located at different regions in the time-frequency plane [4].

3.04 Assamese, Bodo and Rabha Vowel's Spectrograms :

The speech signal, though taken as non-stationary signals, is assumed as stationary one for speech periods (10ms to 50ms) [107]. The basic software tools used in the formant study is MATLAB (version7.1). Other C++ programs were specially developed to check and validate the accuracy of our computational results. Low-pass and band-pass filtering operations were implemented with MATLAB filter design functions. This typical window length is chosen for the computation of the spectrum of the vowels utterances in the present study. The spectrograms of the eight Assamese , six Bodo and six Rabha vowels utterances corresponding to male and female speakers have been shown in Figs- 3.2(a), Fig.-3.2(b), Fig.-3.2(c) , Fig.-3.2.(d), Fig.-3.2(e), and Fig.-3.2 (f). From the graphs , the pitch of the vowels can be determined from the first line along the frequency axis. However, these frequencies are observed, very roughly, while formants are hardly seen directly from the graphs as depicted below:

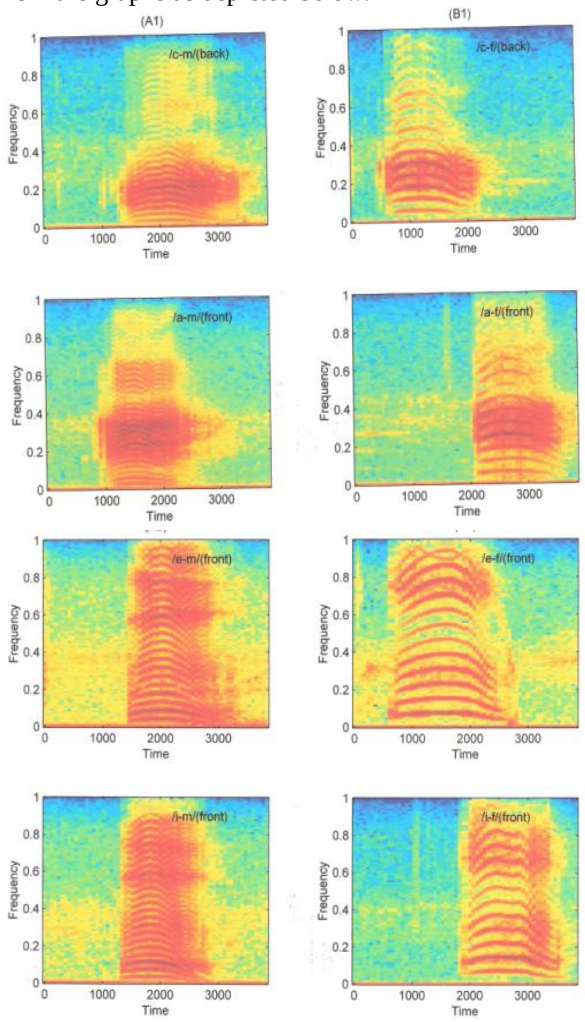
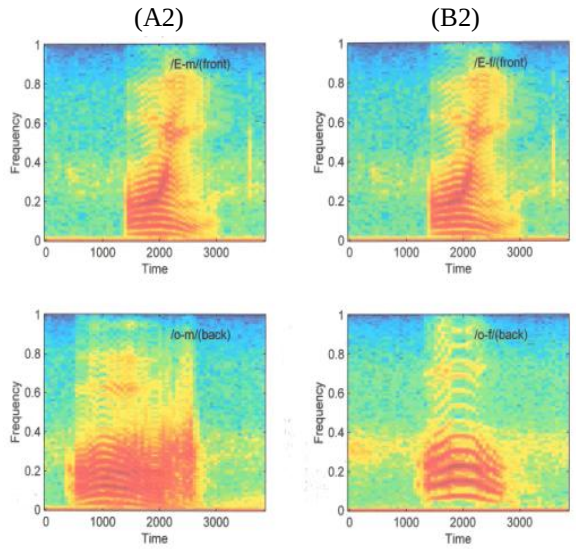


Fig 3.4(a) :Spectograms of Assamese vowels: (A1) Male and (B1) Female



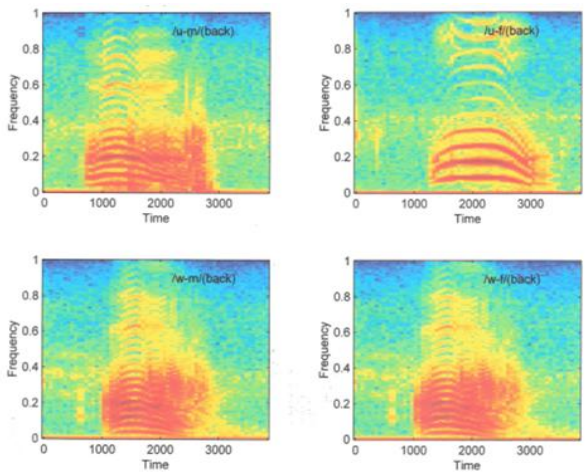


Fig 3.4(b) :Spectograms of Bodo vowels: (A2) Male and (B2) Female

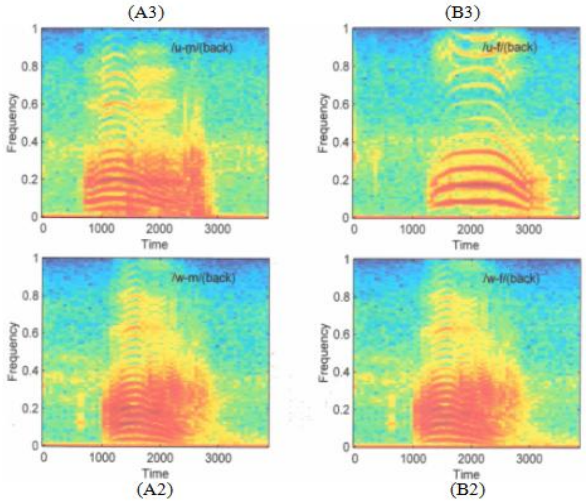


Fig 3.4 (c) :Spectograms of Rabha Vowels: (A3) Male and (B3) Female

IV. CONCLUSION

From the power coefficients (Table 3.1), we get the parameter power for each vowel signal, which is an important feature for application in speech detection synthesis. In general , it is inferred that power coefficients of females are higher than those of males in most vowels .

It is seen from the spectrograms that in almost all cases, the pitch and first formant frequency and second formant frequency are very distinct corresponding to both male and female informants. However, this can be seen very clearly up-to a maximum frequency of 1 KHz, beyond which the spectrograms are hazy and noisy. Thus, low pass filter (LPF) seems more relevant in the present study of pitch and formant frequency analysis of the vowel's speech utterances. From the present analysis and study of different vowels spectrograms, the following observations have been made:

(i) For Assamese Male Speakers, the frequency components are located about
100 Hz $\leq$ f(Hz) $\leq$ 350 Hz for different vowels .

(ii) For Assamese Female Speakers, the frequency components are located about
50 Hz $\leq$ f(Hz) $\leq$ 350 Hz for different vowels .

(iii) In case of Bodo vowels, the spectrograms show some unusual characteristics. Here, the lower frequency region , including the pitch and first , second ,third, fourth and up to fifth formant frequency, is very clear lying within the range of 100Hz to 2KHz **.**

(iv) For Rabha male speakers, the frequency components are located about
100Hz $\leq$ f(Hz) $\leq$ 500Hz irrespective of the vowels.

(v) However, in case of female Rabha informants, 100Hz $\leq$ f (Hz) $\leq$ 1000Hz.
In both the cases, the higher frequency components are faded out or noisy.

Further, unlike the Rabha vowel's spectrograms, the higher components of the formant frequency of Bodo vowels lie within the range 3 KHz (vowel/e/) and 3.5 KHz (vowel/a/). This is a very uncommon characteristic observed in case of the Bodo vowel utterances.

However, in the frequency range 500Hz $\leq$ f $\leq$ 2 KHz, there is total disappearance of spectra of the vowels. This is found common in case of all vowel's spectra. It is seen from the Fig.3.2(a,b,c,d,e,f,g and h) that the frequency-Time spectrograms corresponding to both the male and female speakers, the spectrograms within the time scale ranging from 0.2 sec. to 0.3 sec., are totally viewed as scattered dots. The same patterns found repeated within 0.5sec.to 0.7sec.

The pitch and formant frequency characteristics of the vowel's spectras of all the three languages are found in agreement, in principle, with the WVD method as mentioned earlier. However, the only expectation in the present study with respect to the formant and pitch detection, as proposed by Zaho et al [85] using the WVD technique, is that the gradual blackout of the entire speech spectra, mostly obtained in case of female speakers, may be attributed to the irregular vocal dynamics and asymmetric vocal fold characteristics of the speakers concerned.

The asymmetric behaviour of the vocal fold dynamics may arise due to drift of various vocal fold parameters i.e. effective length of vocal folds, mass and tension of the vocal fold which are controlled by muscle action etc.. As the control of these quantities is much slower than the vibration of the folds, so this may produce the uneven and scattered part of the spectrograms.

REFERENCES

[1] Ahad A., Fayyaz A., and mehmood T., "speech recognition using Multi-layer perception" , Proceedings of IEEE students conference, ISCON-02, Vol 1, pp103- 109, 2002.

[2] Barra R, Montero J M, Macias J, dharo LF, San-sagundo R and De Cordoba R,"Prosodic and Segmental Rubrics in Emotion Identifications", Proc International Conference on Acoustic Speech and Signal Processing 206, Toulouse, pp 1085-1088, 2006.

[3] Bassi A., Becerra Yoma N., and Loncomilla P., "Estimating Total Prosodic Discontinuities in Spanish using Hidden Model", Speech Communication 48, 9: pp 1112- 1125, 2006.

[4] Bhattacharya P.C., "aspect of North-East Indian Languages", 2003

[5] Basumatary P., "An Introduction to the Boro Language" , mittal Publication, 2005

[6] Candless Mc S, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", IEEE Trans. Acoust, Speech, signal Processing, Vol ASSP-22, pp 135-141, 1974.

[7] Hakacham U R, "rabha Bhasa Aru sahitya" First Edition, Sept 1997

[8] Markel P, Herzel H, et al, "Irregular Vocal Fold vibration- High Speed observation and Modeling", Acoustical Society of America, Vol 108 (6), July 2007.

[9] Snell R C and Milinazzo F, "Formant Location from LPC Analysis Data", IEEE Trans Speech Audio Processing, Vol 1, pp129- 134, April 1993 ..

