Proceedings of the 5th WSEAS Int. Conf. on SIGNAL, SPEECH and IMAGE PROCESSING, Corfu, Greece, August 17-19, 2005 (pp138-142)

Acoustic Space of Bangla Vowels

SYED AKHTER HOSSAIN
Dept of Computer Science and Engineering
East West University
43 Mohakhali C/A, Dhaka 1212
BANGLADESH

M LUTFAR RAHMAN
Dept of Computer Science and Engineering
University of Dhaka, Dhaka
BANGLADESH

FARRUK AHMED
Dept of Computer Science and Engineering
North South University, Dhaka
BANGLADESH

Abstract: - Acoustic space of Bangla Vowel based on articulatory properties of vocal tract plays a significant role in Bangla speech synthesis and recognition. A vowel system is essentially a way of dividing up the "vowel space" into distinct vowel phonemes. In general, languages make the most efficient use of the vowel space. This paper synthesizes vowel space illustration by graphically showing where a speech sound for Bangla language, such as a vowel, is located in both "acoustic" and "articulatory" space of the vocal tract tube. Measurements of the spectral characteristics and the formant frequencies were made for each vowel from isolated Bangla word spoken by both male and female speakers. All these measurements are tested in synthesis of isolated utterance.

Key-Words:- Speech Processing, Formants, Energy, Voiced, Speech Synthesis, Phoneme

1 Introduction

Speech signals are composed of a sequence of sounds. These sounds and the transitions between them serve as a symbolic representation of information. In speech production process the sub-glottal system composed of lungs, bronchi and trachea serves as a source of energy for the production of speech. Speech is simply the acoustic wave that is radiated from this system when air is expelled from the lungs and the resulting flow of air is perturbed by a constriction somewhere in the vocal tract [1,2,3].

Speech sounds are classified into three distinct classes according to their mode of excitation. *Voiced sounds* are produced by forcing air through the glottis with the tension of the vocal chords adjusted so that they vibrate in a relaxation oscillator, thereby producing quasi-periodic pulses of air which excite the vocal tract. In English language, voiced segments are labeled /U/, /d/, /w/, /i/ and /e/. Forming a constriction at some point in the vocal tract, and forcing air through the constriction at a high enough velocity to produce turbulence generates fricative or unvoiced sounds. Plosive sounds result from making a complete closure in front of the vocal tract, building up pressure behind the closure, and abruptly releasing it [1,2,3].

1.1 Vowel Space Illustration

Vowels are produced by exciting a fixed vocal tract with quasi-periodic pulses of air caused by vibration of the vocal chords. The resonant frequency of the tract (formants) varies with the cross sectional area and the dependence of cross-sectional area upon distance long the tract, which is known as area function is determined primarily by the position of the tongue, but the positions of the jaw, lips, and, to a small extent, the velum also influences the resulting sound.

The chief characteristic of the vowels is the freedom with which the air stream, once out of the glottis, passes through the speech organs. The supra-glottal resonators do not cut off or constrict the air; they cause only resonance, that is to say, the reinforcement of certain frequency ranges. A vowel's timbre (or quality) depends on the following elements:

- the number of active resonators (among the oral, labial, and nasal cavities);
- the shape of the oral cavity;
- the size of the oral cavity.

There are three possible resonators involved in the articulation of a vowel: the oral cavity, the labial cavity, and the nasal cavity. If the soft palate is raised,

the air does not enter the nasal cavity, and passes exclusively through the oral cavity; if the soft palate is lowered, however, air can pass through nose and mouth simultaneously. If the lips are pushed forward and rounded, a third, labial resonator is formed; if, on the other hand, the lips are spread sideways or pressed against the teeth, no labial resonator is formed.

The sound of a vowel is governed by many factors, of which the position of the tongue and rounding of the lips are the most important. The "vowel space" is thus defined as the total area over which the tongue position ranges, along the two axes of height and backness.

Thus, each vowel sound can be characterized by the vocal tract configuration (area function) that is used in its production.

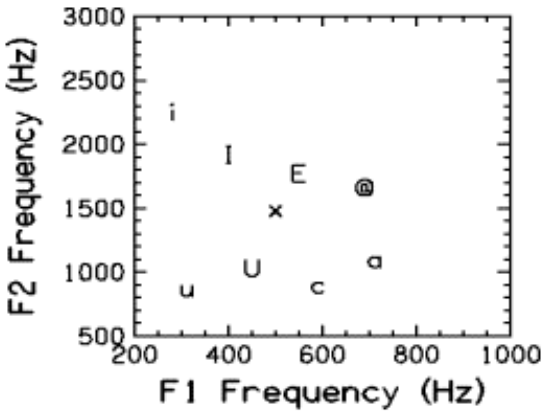

Fig.1 Vowel Space representation of English

The vowel space illustration shown above provides a graphical method of showing where a speech sound, such as a vowel, is located in both "acoustic" and "articulatory" space. The illustration shows an acoustic vowel space based on the first two formants for vowels (formants are the bands of energy that correspond to the resonance of the vocal tract for particular shapes). The vertical axis represents the frequency of the first formant (F1). The horizontal axis shows the frequency gap between the first two formant (F2-F1)[2,5,6].

This 2-dimensional representation corresponds, to a certain degree, to tongue body position, with indications of high vs. low and front vs. back positions -- an articulatory space [3].

2 Analysis Method

The speech data used in this study is captured in a noise free environment using the sound capture card in Speech Filing System environment. For each utterance of Bengali vowels, one of the most commonly used words were captured for different male and female speakers and the Table-1 shows the detail of the recording for each of the following utterance:

Bengali Word	Corresponding Vowel	Avg. Duration (msec)
আম	/আ/	132
অজগর	/অ/	83
ইদুর	/ই/	72
ঈগল	/ঈ/	92
এভারেষ্ট	/এ/	129
ঐক্য	/ঐ/	115
ওরাং ওটাং	/ও/	134
ঔষধ	/ঔ/	172
উল	/উ/	208
ঊষা	/ঊ/	169

Table 1. Recording duration of utterance

The digitized speech was filtered to remove any additional noise and was segmented visually to apply formant extract procedure through Fast Fourier Transform (FFT) and Linear Predictive Coding (LPC).

There are many approaches to formant extraction, which are based on various (typically LPC-based) signal-processing techniques. Such approaches to formant extraction are normally fully automated although provision is often made for intervention after the process has been completed especially editing of formant tracks [8,9,10].

A further limitation of most formant tracking procedures is that formant gain and bandwidth extraction is either absent or very rudimentary. It is extremely difficult, for example, to extract meaningful bandwidth information as the extracted values are dependant on both the actual formant bandwidth and also on the extraction process (e.g. variations in the number of LPC coefficients result in variations in bandwidth).

It is also difficult to define gain in a useful way. Again, in LPC extraction procedures, the relative peak height of the formants can vary with the number of coefficients used.

Further, is it desirable for raw formant peak height be measured or is it more meaningful to measure formant gain after removing the effect of source spectral slope. For many applications are may not be necessary to accurately measure formant bandwidth and gain. For example, many speech analysis tasks of relevance to acoustic phonetic research only require accurate formant frequency information. Such tasks typically most often examine formants in vowels and (to a lesser extent) vowel-like consonants.

In vowels, gains and bandwidths are fairly predictable (Fant, 1960). Vowel formant bandwidth is also readily calculated utilizing a simple relationship between formant center frequency and formant bandwidth [11,12,13].

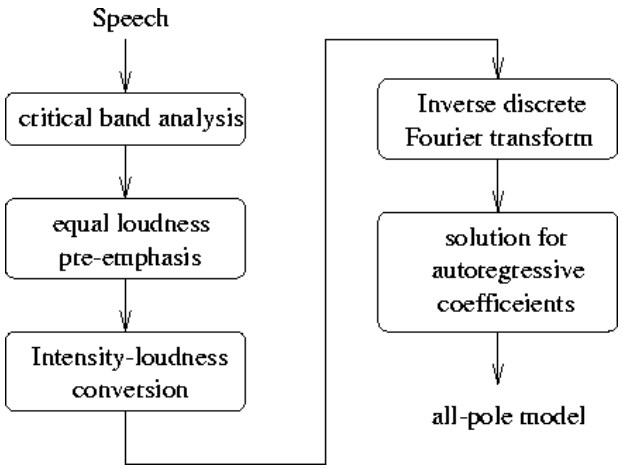

Fig.2 Perceptual Linear Prediction

In Perceptual Linear Prediction method as shown in Fig. 2, the typical window length is 20ms. For 10 kHz sampling frequency, 200 speech samples are used, padded with 56 zeros, hamming windowed, FFT and converted to a power spectral density. After applying the pre-emphasis filtration, the IFFT was applied for cepstrum calculation and a suitable peak detection algorithm was used to determine formants.

For all the processing of speech samples, the first three formants are recorded and the corresponding formants are used to produce synthetic speech for the accurate estimation of the formant frequencies. All these formants are grouped together to plot in two-dimensional space to produce vowel space representation considering the first two formants [5].

3 Results and Discussion

The speech corpus was conditioned and required pre emphasis and de-emphasis were carried out and labeled to segment the phoneme boundary of interest. The extracted speech segment was taken for Linear Predictive Coding (LPC) scheme with 14 as number of poles. The maximum formant considered was 5000 Hz and the analysis width of the window was 25msec with an overlap of 10msec for the accurate resolution of the formants. Among the existing methods, the Burg method was applied and along with the formant extraction, the vocal tract response characteristics were extracted for each of the phoneme in the vowel space. The results are postulated here in the following figures:

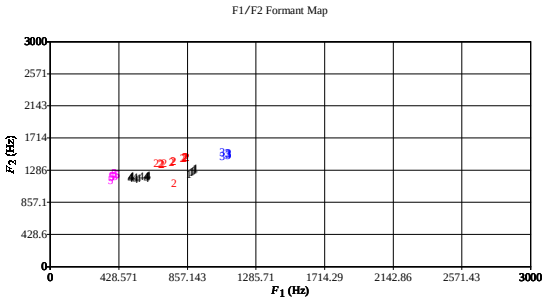

Fig.3 (a) Vowel Space representation of /আ/ in আম

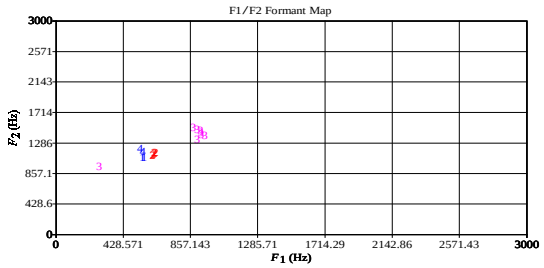

Fig. 3 (b) Vowel Space representation of /অ/ in অজগর

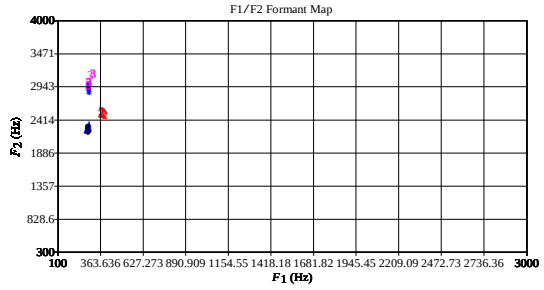

Fig.3 (c) Vowel Space representation of /ই/ in ইদুর

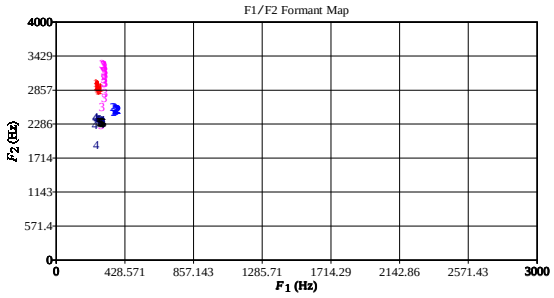

Fig.3 (d) Vowel Space representation of /ঈ/ in ঈগল

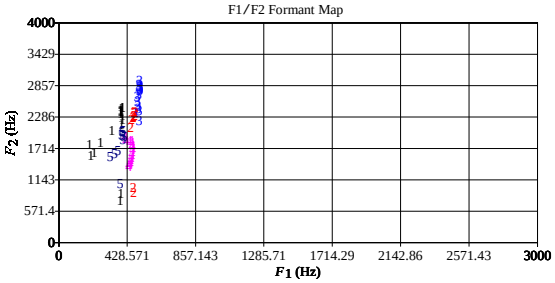

Fig.3 (e) Vowel Space representation of /এ/ in এভারেষ্ট

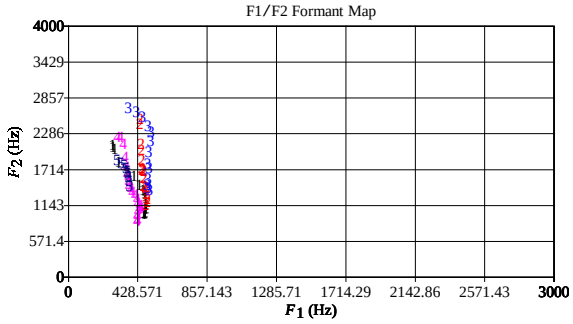

Fig.3 (f) Vowel Space representation of /ঐ/ in ঐক্য

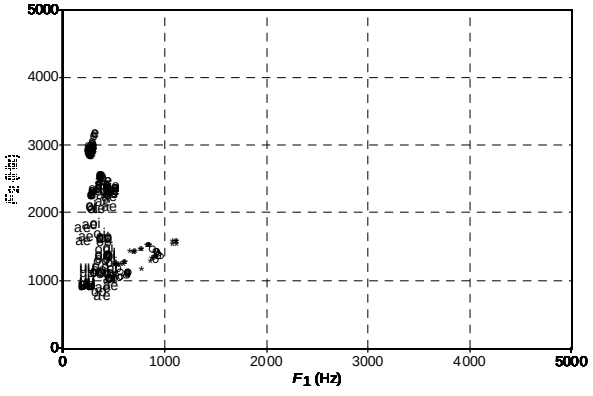

Fig.4 Combined Vowel Space representation of /অ/ through /ঊ/

The vowel space distribution of the vocal tract responses for Bangla vowel utterance is shown in the above figures from 3(a) through 3(f) which indicates the spectral property as well as the vocal tract response during the production of the vowel [4].

The vowel space representation indicates a range for the formants in order to be recognized or for the purpose of synthesis but the vocal tract response as shown in case of /ভ/ and /ঊ/ and a similar pair of phoneme /ই/ and /ঈ/ seems very much indistinguishable.

4 Conclusion

Bengali Vowels synthesis and recognition with respect to vowel space requires more methodical and careful study of the linguistic phonetics. The vowel space representation postulated in this paper is used to produce synthetic vowel sounds. The synthesis result has shown good degree of correlation only with the exception of the phoneme boundary detection issues, which would require further enhancement of the algorithm used for the feature extraction.

References:

[1] Rabiner L. R., Schafer R. W., "Digital Processing of Speech Signals", Trentice-Hall Inc, Englewood Cliffs, 1978

[2] John R Deller, John G Proakis, John H L Hansen, "Discrete-Time Processing of Speech Signals", Macmillan Publishing Company, 1993

[3] G.Fant, *Acoustic Theory of Speech Production*, 's-Gravenhage, The Netherlands: Mounton and Co., 1960.

[4] Muhammad Abdul Hai, *Dhvani Vijnan O Bangla Dhvani-Tattwa*, Mullick Brothers, 2000.

[5] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Procs.*, pp. 208-211, Apr. 1979.

[6] S.Blumstein and K.Stevens, "Acoustic invariance in speech production," *J. Acoust. Soc. Am.*, vol. 66, pp. 1001-1017, 1979.

[7] S. Blumstein and K. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Am.*, vol. 67, pp. 648-662, 1980.

[8] S.Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.27, pp. 113-120, Apr. 1979.

[9] B. Gold and N. Morgan, *Speech and audio signal processing*, Wiley, 2000.

[10] S Akhter Hossain, M Lutfar Rahman , Farruk Ahmed "Vowel Space Identification of Bangla Speech", Dhaka University Journal of Science, 51(1): 31-38 2003(January)

[11] S Akhter Hossain, Md Faruk Ahmed, Mozammel Huq Azad Khan, M A Sobhan, and Md Lutfar Rahman, "Analysis by Synthesis of Bangla Vowels", 5th International Conference on Computer and Iinformation Technology Proceeding, 2002, pp. 272-276

[12] S Akhter Hossain, M A Sobhan, Mozammel Huq Azad Khan, "Acoustic Vowel Space of Bangla Speech", International Conference on Computer and Iinformation Technology 2001 Proceeding, pp. 312-316

[13] S Akhter Hossain & M Abdus Sobhan, "Fundamental Frequency Tracking of Bangla Voiced Speech" –1st National Conference on Computer and Information System Proceeding 1997, pp. 302-306