ISSN 2278-3091
Volume 10, No.1, January - February 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse371012021.pdf
https://doi.org/10.30534/ijatcse/2021/371012021

Generating Vowel Nasality for a Rule-Based Bangla Speech Synthesizer

Shahina Haque[1], Md. Hanif Ali[2], A. K. M. Fazlul Haque[3]
[1,2] Department of CSE, Jahangirnagar University, Savar, Dhaka, Bangladesh,
[1,3]Department of ETE, Daffodil International University, Dhaka, Bangladesh
Corresponding author: oceanuniverse2017@gmail.com, shahina@daffodilvarsity.edu.bd

ABSTRACT
Bangla is a useful language to study nasal vowels because all the vowels have their corresponding nasal vowel counterpart. Vowel nasality generation is an important task for artificial nasality production in speech synthesizer. Various methods have been employed by many researchers for generating vowel nasality. Vowel nasality generation for a rule-based speech synthesizer has not been studied yet for Bangla. This study discusses several methods using full spectrum and partial spectrum for generating vowel nasality to use in a rule-based Bangla text to speech (TTS) system using demisyllable. In a demisyllable based Bangla TTS 1400 demisyllables are needed to be stored in database. Transforming the vowel part of a demisyllable into its nasal counterpart reduces the speech database size to 700 demisyllables. Comparative study of the e□ciency of the methods for nasal vowel production is evaluated through listening test. It is observed that although some of the methods have higher efficiency, vowel nasality produced by simple sine curve model is perceived to be acceptable to the native speakers of Bangla and could be easily used in a rule-based Bangla TTS system.

Key words: Bangla, Cepstrum, Nasality, Neural Network, Speech synthesis.

1. INTRODUCTION

Nasality is a distinctive feature of Bangla [1]. All the 7 vowels of Bangla have their corresponding nasal vowel counterparts as in Portuguese. The contrast in spectral characteristics of Bangla oral vowel /i/ and nasal vowel /˜i/ is shown in Figure 1. For /˜i/, the nasal zero-pole is seen around 1kHz. Until now, very few noticeable works have been reported on Bangla vowel nasality [2].

Nasality is produced by lowering the velum so that part of the sound wave is free to pass through the nasal cavity. The nasal and oral cavities resonate together and lower the energy of the whole spectrum, as well as they broaden F1 bandwidths

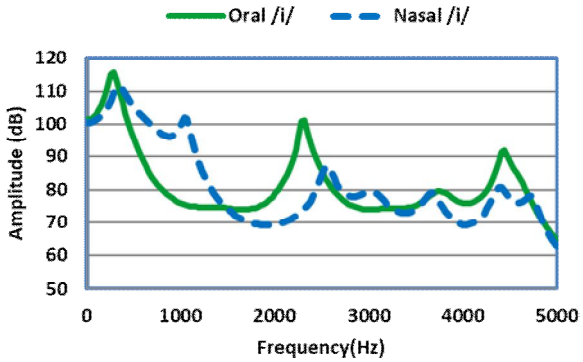


Figure 1: Contrast in vocal tract transfer function of Bangla vowel /i/ and /˜i/

and introduce nasal poles and zeros in the spectrum [3]. If the velopharyngeal opening is large, nasal pole shows increased spectral prominence [4]. Even with a small velopharyngeal opening for the vowel /˜i/, a prominence of 13.5dB can be introduced with a pole around 810Hz. A large velopharyngeal opening for the vowel /˜a/ can introduce a prominence as large as 12.1dB with a pole at 810Hz [5]. The zeros of nasal vowels depend on the nasal tract characteristic originating from the velum [6]. The vocal tract shape and amount of velar opening interact to determine the position of spectral zeros. Another characteristic of vowel nasalization may be the cancellation of the 3rd formant [3].

The acoustics of nasal vowels is a complex phenomenon, which has been the subject of numerous studies [6, 7, 9–11]. Among the important features found in various studies of nasal vowel synthesis, nasal pole-zero pairs around 250-450Hz [7], around 1kHz [4] and around 2-2.5kHz [6] are reported to be contributing to the e□ect of nasality perception. Synthetic nasal vowels were produced by either reducing the amplitude of the first formant (F1) [8] or by adding a pole-zero pair near 1kHz [4]. Modifying the oral vowel spectrum around F1 has been proved to create the perception of nasality although essentially for non-high vowels [4]. The nasalization e□ect was found to be enhanced by adding another pole-zero pair around 250-450Hz [7]. In another

study of vowel nasalization, nasal sounds are modeled as a dynamic trend from an oral configuration toward a /ŋ/ like configuration corresponding to the pharyngonasal tract [10], F1 bandwidth and other F1 profile criteria, and the number of peaks above a threshold 40dB below signal peak, and two criteria relating the amplitude of the first formant to the first harmonic. Today's most well accepted acoustic parameters for vowel nasality in speech are:

(i) the standard deviation around center of mass in the band below 1kHz, and the percentage of time of observed extra poles at low frequencies

(ii) the frequency of the nasal extra poles P0 and P1 with respect to the frequency of F1 (iii) the amplitudes of the extra poles with respect to the amplitude of F1

(iv) F1 bandwidth and other F1 profile criteria, and the number of peaks above a threshold 40dB below signal peak, and two criteria relating the amplitude of the first formant to the first harmonic.

Although the method used in the previous studies reported to model nasal vowel spectrum quite accurately, the methods did not seem to be used easily in the rule-based speech synthesizer. Therefore, several methods for vowel nasality generation is discussed.

Relative evaluation of the used methods show that sine curve (SC) model to be suitable for the present study. Therefore, simple, parametric and rule-based SC model, which produces enough nasality perception, is observed to be suitable to use in rule-based speech synthesizer.

So initially, the whole spectrum is taken into account for vowel nasality generation. Then for simplicity, only the e□ect of nasality of low frequency part of spectrum is considered. It is observed that low frequency nasality information of nasal vowel spectrum is su□cient to perceive nasality of vowels. It is observed through the listening tests that the proposed SC model acted e□ciently for vowel nasality generation of Bangla. However, the result is least satisfactory for high back vowel /˜u/. SC model may be used to reduce the size of the database of the rule-based speech synthesis system to around 50%. The speech unit used in the rule-based speech synthesizer concerned in this study is demisyllable. The number of demisyllables to be stored in the database is around 1400. To reduce the number of demisyllables to half, oral vowel part of a syllable is transformed to its corresponding nasal vowel counterpart. As an initial phase of this work, vowel's nasality is transformed. Cepstral method [6] and log magnitude approximation (LMA) filter [11] approximates e□ectively both poles and zeros of the vocal tract.

The rest of the paper is organized as follows:
Section II describes about speech materials and speech analysis synthesis method. Section III discusses about the nasality generation methods. Section IV discusses about results and discussions. Section V concludes the paper.

2. SPEECH ANALYSIS SYNTHESIS METHOD

The aim of this section is to describe how the speech samples are acquired. All the seven Bangla vowels and their nasal counterpart are selected as the speech samples of this study. The experimental part consists of recording each of the isolated Bangla oral-nasal vowels at a normal speaking rate three times in a quiet room in a DAT tape at a sampling rate of 48 kHz and 16 bit value. The best one of these three speakers sample data is used for the study. These digitized speech sounds are then down sampled to 10kHz and normalized for the purpose of analysis. Short-term cepstral analysis method is used to extract the speech parameters. The speech wave is segmented to 25.6ms frame length. A time-domain Blackman window is used. Frame shifting time is 10ms. Cepstrum for spectral parameter, which is the inverse Fourier transform of the short time logarithm amplitude spectrum of the speech waveform [5]. The resulting parameters of the speech unit include the number of frames and for each frame, voiced/unvoiced decision, pitch period and cepstral coe□cients.

The speech synthesis section shown in Figure 2 is designed on the source filter model [6]. The vocal tract features can be suitably represented for all speech sounds by the pole-zero LMA filter [11] proposed by Imai. LMA filter is a part of a homomorphic vocoder. In our speech synthesizer, LMA filter is made from a cascade of 30 elemental second order filters. LMA filter representing the vocal tract is driven by an adequate excitation source. In case of producing voiced speech, the excitation source is a series of unit impulses separated by fundamental period intervals. In case of producing unvoiced speech, the excitation source is noise having unit amplitude and random polarity. This Bangla speech synthesis system [12] is based on general speech synthesis system for Japanese [13].

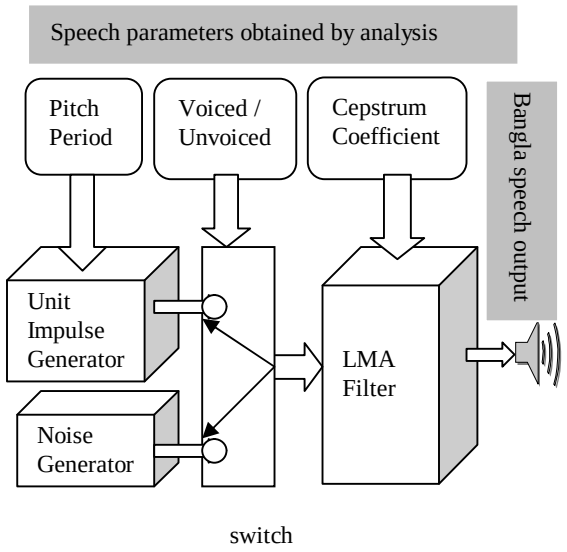


Figure 2: Speech synthesis sub-system

3. VOWEL NASALIY TRANSFORMATION AND GENERATION METHODS

The aim of the experiments used in this section is to describe di□erent methods for vowel nasality generation and then to choose the most suitable method to use in the rule based speech synthesizer. The chosen configuration should produce acceptable perception of nasality to the native speakers of Bangla for vowels /˜i/, /˜e/,/˜æ/,/˜a/, /˜ɔ/,/˜o/, and /˜u/. Therefore, initially the whole spectra is taken into account for transformation then for simplicity only the low frequency portion is transformed to obtain the e□ect of nasality. It is observed that the low frequency part of a nasal vowel spectrum is su□cient for human ear to perceive vowel nasality.

A. Nasal vowel generation using full spectra:

(1) Inverting distinctive feature (IDF): In this method, three distinctive feature outputs (high, back and nasal) from a single layer neural network is trained by backpropagation method. Cepstrally smoothed log spectrum of all the 14 Bangla vowels are fed to the input of the neural network. For each of the 14 Bangla vowels, thirty spectral data are used in this section. Input spectrums are normalized using Eq. 1 [14]. In Eq.1, x_i are components of m dimensional spectral parameter in each frame.

$$x_i^{'} - x_i / \sqrt{\sum_{i-1}^{m} x_i^2} \quad (1)$$

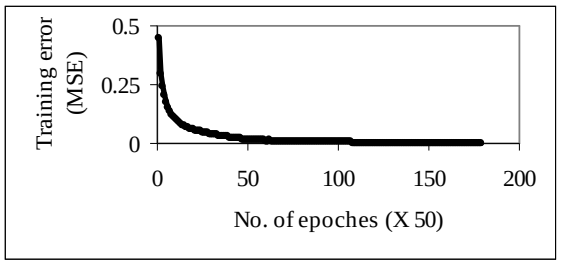
In Eq. 1

$$\sqrt{\sum_{i=1}^{m} x_i^2} = 1 \quad (2)$$

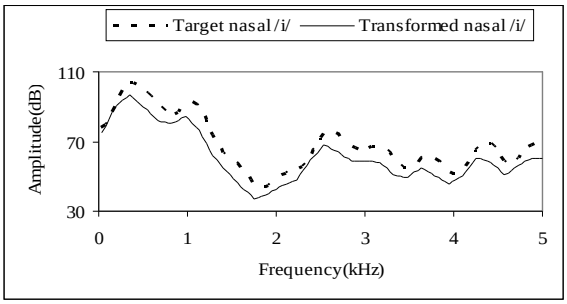
The nasal detector neuron outputs obtained by the oral spectra input is inverted to make the nasal features. The inverted nasal detector outputs and other two outputs are fed to three inputs (high, back and nasal) of two layer weight neural network. The number of neurons used in this network for the first layer is 3, for second layer 100, and 32 for the output layer. The output is desired to be the corresponding nasal vowel spectra.

(2) Neural network (NN): The neural network used in IDF method has two layer weight. In order to make it more effective, backpropagation method is adopted to train a three layer weight neural network. The neural network is trained by giving oral vowel spectra as the input, and the target is the corresponding nasal vowel spectra. The mean square error of training is shown in Figure 3(a). A data-open test example of a transformed spectrum /˜a/ is shown in Figure 3(b).

(3) Spectrum difference model: Direct method to transform oral vowels to nasal vowels is to use the spectral difference between orals and nasals. Each vowel spectrum difference (EVSD) is calculated for each oral-nasal vowel pair in log frequency domain. Then EVSD for all vowels is averaged to form the average spectrum difference (ASD) of vowels. ASD



(a)



(b)

Figure 3: (a) Training error of neural network (b) Spectrum of /˜a/ transformed by neural network

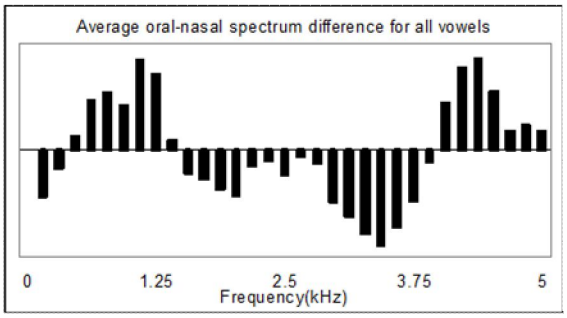


Figure 4: ASD over Bangla vowels

is shown in Figure 4. Both EVSD and ASD are subtracted from oral vowel spectra to obtain nasal vowel spectra. ASD gives a measure of the average difference between all nasal and oral vowel spectrum (normalized). The positive part of Figure 4 shows the frequencies at which ASD of all nasal vowels is greater than the ASD of all oral vowels. So, the nasal feature of vowels must be seen in the positive part of ASD.

B. Nasal vowel generation using partial spectra

In this section, generation of vowel nasality is done by least possible simple change made in the oral vowel spectra. Region of spectra where the perceptual characteristic of nasality is concentrated for each vowel is focused. It is already found that nasality for perception is mainly in the region near

F1. So in this section, resulting sound by appropriately changing the low frequency region of oral vowel spectra is checked. Then a very simple parametric SC model is used to model nasality.

(1) Partial ASD model (PASD): In this method, ASD used in full spectra transformation method is made equal to zero at higher frequency than 2kHz to form PASD. Then PASD is subtracted from oral vowel spectra to make nasal vowel spectra.

(2) Sine curve (SC) model: From nasal vowel spectra, it is observed that the nasal PZ or ZP has a pattern like sine curve. Simple method to use partial spectrum is to adopt such a typical pattern. In SC model, nasal pole (P) and zero (Z) of the vocal tract transfer function is modeled by using a symmetrical mathematical function such as a single period sine curve. So sine curve is added in log spectrum domain to transform oral vowel spectra to nasal vowel spectra. The strength and width of such a curve is directly proportional to the effect of vowel nasality. PZ and ZP of SC model rule are

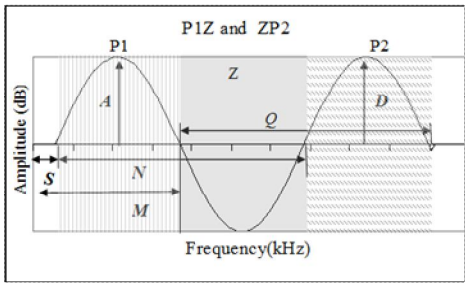


Figure 5: PZ and ZP rule for nasality generation by SC model

Table 1: Parameters of SC model for a certain degree of nasalization of a male speaker

Vowel	A1 (dB)	P1 (kHz)	Z (kHz)	P2 (kHz)	A2 (dB)
/i/			0.790	1.090	13.28
/e/			0.796	1.058	8.7
/æ/			0.990	1.252	5.8
/a/	11.29	0.272	0.480		
/ɔ/			0.561	0.981	3.95
/o/			0.756	1.176	5.64
/u/			0.561	0.981	11.29

shown in Figure 5. Parameters of SC model are amplitude and position of nasal pole or zero from the origin. Amplitude of PZ = A1[dB]. Position of pole of PZ from the origin = P1[kHz]. Position of Z from the origin = Z [kHz]. Amplitude of PZ = A2[dB]. Position of pole of ZP from the origin = P2[kHz]. SC model itself is general. It can be used for different person and different degree of nasalization by varying the parameters of the model. In this work, use of SC model for a particular male speaker is shown. For each vowel of the male speaker and for certain degree of nasalization, values of parameters of SC model are given in Table 1. SC method can be applied adopting its proper parameters for each speaker.

As clarified in many past studies, there are a number of poles and zeroes in nasal vowel spectrum [4, 6, 7]. But adding pole-zero or zero-pole pair near F1 of non-nasal spectra is reported to produce enough nasality perception [4]. In this study, SC model near F1 using the same phase as used by the previous study [4] for each vowel. For transforming each vowel, the strength, width and position of the sine curve is varied systematically in steps at regions near F1. For a certain degree of nasalization of a speaker, appropriate frequency points are located in the spectrum where this model can change each oral vowel to its nasal vowel counterpart giving the best synthetic sound. For transforming the low vowel /a/ by this model, nasality is modelled by PZ. This is because /a/ has high F1, so nasal pole-zero at frequency lower than F1 is important. For transforming other Bangla vowels, nasality is modeled by ZP.

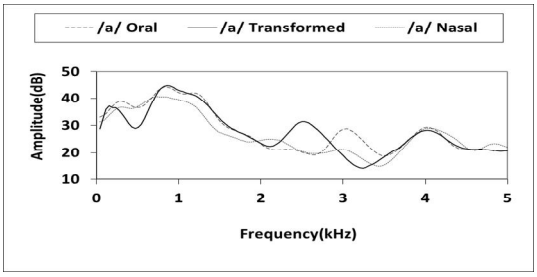


Figure 6: Transformed spectrum of /˜a/ by SC model

An example of transformed spectrum /˜a/ by SC model is shown in Figure 6. As a preliminary test, the result is checked by adding a second sine curve representing pole-zero pair around 2-2.5kHz. Nasality is not so much affected with this addition. So, it can be observed that although the nasal vowel spectra seem to be different more or less by this higher sine curve, but the most noticeable range for perception is the low frequency portion.

In isolated speech, the parameters of SC model [15] can be applied directly to the oral vowel spectrum for transforming it to nasal vowel spectrum. For a speaker and for a certain degree of nasalization, the strength of the pole-zero is controlled with the power of the speech signal. To use the model in continuous speech, the vowel portion can be selected where the vowel is to be nasalized, then apply the rule for nasalization in the same way as described for isolated speech. However, this work is remained as the future work.

4. LISTENING TEST AND RESULT

Vowel nasality generation methods are evaluated by perception and by spectral distance i.e. by the listening test and by measuring Euclidean distance between the

transformed spectra and the original nasal vowel spectra. Then the listening test and spectral distance result are compared. Seven native speakers of Bangla with normal hearing ability participated in the listening test. At first, they were introduced to the listening test system and were presented to hear a few examples of data. Then the test started. The listening test is done in a soundproof room. Ten data sets involving original, analysis-synthesis, and transformed data sets, were presented to the listeners. Fourteen phoneme data (nasal and oral vowels) were involved in each data set giving a total of 140 data. Each data is played randomly three times. So the total number of data presented to each listener were =140*3=420. Each data is played in 3 second interval, which is sufficient for listeners to answer what they heard before the next phoneme is played. The listeners were forced to select them as one of 14 (7 oral and 7 nasal) vowels.

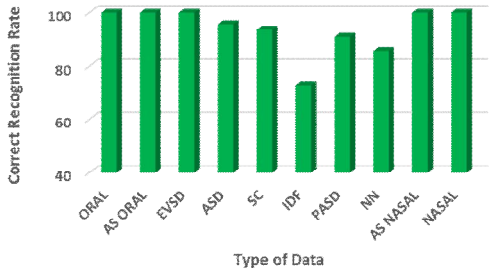


Figure 7: Average nasal vowel recognition score of vowel nasality generation methods.

The result of listening test of different type of data involved in vowel transformation is shown in Figure 7. Original and analysis-synthesis data of oral and nasal vowels were found to give a 100% recognition score. EVSD data is almost similar to the analysis synthesis data of nasal vowel, so it has a 100% recognition score. ASD gives a 95.6% nasal vowel recognition score and so contains the necessary nasality information of general nasality over all vowels which can transform appropriate nasality to each vowel. SC model gives a 93.5% nasal vowel recognition score, proving that appropriate parametric modelling is almost sufficient for nasality generation. PASD also seems to give enough nasal vowel perception to all the listeners, giving a 90.8% recognition score. NN and IDF methods give recognition scores of 85.7% and 72.7% respectively. It has already been demonstrated for voice conversion that a linear conversion of frequency axis is not sufficient to transform the spectrum from one speaker to another [16]. The case of transforming vowel nasality by linear transformation is also checked using Pseudo inverse matrix. Due to the poor recognition score obtained by listening test, it is ignored in this paper.

5. DISCUSSION

This study aimed at producing a rule-based Bangla text to speech synthesis system with a compact database. Although it is not so easy to get highly natural speech output from rule-based speech synthesizers, it is even more difficult to get a natural nasal vowel sound. The reason is that a nasal vowel spectrum has additional poles and zeros compared to its oral counterpart. Nasal zero is one of the important characteristics for nasality. The cepstral model and LMA filter approximates both poles and zeros of the vocal tract transfer function. The LMA filter is a pole-zero filter which efficiently models the vocal tract features for all speech sounds. So the adoption of cepstral method for analysis and LMA filter for synthesis is appropriate for this work.

In section 3, different methods are used for vowel nasality generation to make a compact database for a demisyllable based speech synthesizer. Starting with full spectra transformation and heading towards partial spectra transformation methods, it is concluded that: although a nasal vowel spectrum seems to be different more or less throughout as compared to its oral vowel spectrum, nasality for perception is concentrated at the low frequency region of the spectrum. Although as seen from listening test result, EVSD and ASD method has better nasality recognition score and so are the best way to transform the spectrum, but are not simple as a rule based one. So the simple rule based SC model may be chosen which produces enough nasality perception to use in a rule based speech synthesizer.

In listening test, most of the listening test error occurred for high back vowel /˜u/ for data produced by SC model. This may show that other changes in the spectrum are needed for sufficient nasality perception for high-back vowels. By SC model, while searching for the appropriate frequency point for vowel nasality generation, it is kept in mind that the SC model should generate vowel nasality by keeping the oral formants intact.

Euclidean distance is calculated between the transformed spectrum and their target nasal vowel spectrum. As a result, the average Euclidean distance is lower by NN method (4dB) than that of SC model (7.2dB). But, the listening test result of SC model gave better recognition score (93.5%) than NN method (85.7%). This may be explained as follows: by NN transformation, error is distributed all over the transformed spectrum and overtraining may occur resulting in phoneme confusion. However the important nasality information of nasality perception is concentrated in some frequency region where the PZ and ZP rule is applied. So, the recognition score of the listening test is greater for SC model. The errors in NN and IDF method using a neural network occurred due to phoneme confusion between /˜a/ and /˜ɔ/, /˜e/ and /˜æ/. This is because of the similarity of the spectra of the respective pairs of vowels.

6. CONCLUSION

This paper has focused on the generation of vowel nasality in a rule-based speech synthesizer using Bangla vowels. The result of this study may be summarized as follows, (1) For Bangla vowels, nasality perception is located in regions where PZ and ZP rules are applied. (2) Nasality can be modeled by sine curve (SC) model parameters. (3) This model can be easily added in log spectral domain to transform an oral vowel to a nasal vowel and control the degree of vowel nasalization by parameters. (4) Use of this model to produce nasal vowels gave a correct recognition score of 93.5% to the native speakers of Bangla.

Some methods of vowel nasality transformation and generation are discussed. These methods may be used to reduce the size of database of speech synthesis system to half of the numbers needed. Listening test results of the used nasality generation methods had a reasonable correct rate. Among the different nasality synthesis methods, sine curve model is best due to its simplicity, easily exploitable parametric property and nasality producing quality. Amount of vowel nasalization can be controlled by sine curve model parameters. So, it can be used to produce weak or strong nasal vowels by controlling the parameters. Some Central American languages have phonemic contrast between weak and strong nasal vowels. Even in such cases sine curve model will be useful.

REFERENCES

1. A. Hai, ***Dhvani-Vignan O Bangla Dhvani Tattwa***, June, 1985. (In Bangla)
2. Shahina Haque, Chapter Name: **Bangla Speech Analysis, Synthesis and Vowel Nasality**, In the *Technical Challenges and Design Issues in Bangla Language Processing*, IGI Global. 701 E. Chocolate Avenue Hershey PA 17033-1240, USA. April, 2013
3. J.R. Deller and J.H.L. Hansen, ***Discrete-Time Processing of Speech Signals***, IEEE Press, pp. 137, 2000.
4. S. Hawkins and K.N. Stevens, **Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels**, *J. Acoust. Soc. Am.,* 77(4), pp. 1560-1575, April, 1985.
5. M.Y. Chen, **Acoustic correlates of English and French nasalized vowels**, *J. Acoust. Soc. Am.,* 102(4), pp. 2360-2370, October, 1997.
6. S. Furui, ***Digital Speech Processing, Synthesis, and Recognition,*** Second Edition, Marcel Dekker, Inc., pp. 30–31, 2001.
7. S. Hattori, K. Yamamoto, and O. Fujimura, **Nasalization of vowels in relation to nasals,** *J. Acoust. Soc. Am.,* 30, pp. 267-274, 1958.
8. A.S. House and K.N. Stevens, **Analog studies of the nasalization of vowels,** Journal of Speech and Hearing Disorders, Vol. 21, pp. 218-232, 1956.
9. O. Fujimura and J. Lindqvist, **Sweep-tone measurements of vocal-tract characteristics,** *J. Acoust. Soc. Am.,* 49(2), pp. 541-558, 1971.
10. G. Feng and E. Castelli, **Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization,** *J. Acoust. Soc. Am.*,99(6), pp. 3694-3706, 1996.
11. S. Imai, **Log Magnitude Approximation (LMA) filter,** *Trans. of IECE Japan*, J63-A, 12, pp. 886–893 (1980). (In Japanese)
12. S. Haque, T. Takara, **Rule Based Speech Synthesis by Cepstral Method for Standard Bangla,** *in Proceedings of 18th International Congress on Acoustics*, ICA 2004, Kyoto, Japan, 4-9 April, 2004. Th.P3.19, IV-3341.
13. T. Takara and T. Kochi, **General speech synthesis system for Japanese Ryukyu dialect**, in *Proc. of the 7th WestPRAC*, pp. 173-176, Oct. 2000.
14. T. Takara and S. Imai, **Vowel Recognition Based on Mel-Sone Spectrum,** *Trans. IECE*, J65-A, pp. 818-825, No. 8, 1982. (In Japanese)
15. S. Haque, T. Takara, **Nasality Perception of Vowels in Different Language Background**, in *Proceedings of INTERSPEECH 2006 - ICSLP*, 17-21st September, 2006, Pittsburgh, Pennsylvania, USA, page 869-872.
16. G. Baudoin and Y. Stylianou, **On the transformation of the speech spectrum for voice conversion,** in *Proceedings of ICSLP 96*, Vol. 3, pp. 1405-1408, 1996.