Rule Based Speech Synthesis by Cepstral Method for Standard Bangla

Shahina HAQUE, Tomio TAKARA

Faculty of Information Engineering
University of the Ryukyus, Okinawa

shahina@iip.ie.u-ryukyu.ac.jp, takara@ie.u-ryukyu.ac.jp

Abstract
In the first phase of this paper, we describe the construction of a Bangla speech synthesizer. In the second phase, we discuss our work on Bangla nasal vowel. Nasality is one of the distinctive characteristic of Bangla phonemes. We discuss methods employed for transforming Bangla oral vowel to the corresponding nasal vowel counterpart and its application to our speech synthesizer. The perceptual evaluation result of the system at present, oral vowel 100%, nasal vowel 90%, oral-nasal detection 100%, initial consonant 87% and final consonant is 82%, 86% for non-linear transformation and 91 % for sine curve model.

1. Introduction
Bangla is a language of more than 120 million people of Bangladesh. There are many dialects spoken in Bangladesh. This system is based on the standard colloquial dialect which is commonly spoken in offices, educational institution and broadcast. Bangla is a language which has no stress, tone or accent i.e. variation of pitch does not have significance effect in Bangla[1]. Bangla is heard in more or less in monotone. But the interesting point in Bangla is nasality. All the 7 vowels in Bangla have their corresponding nasal counterparts. Nasalization of vowel changes meaning of some words in Bangla. The contrast lies in the spectrum of nasal and oral vowels as shown in Fig. 1.

The purpose of speech synthesizer is to produce speech with high intelligibility and naturalness. We aim to construct a complete Bangla TTS system with compact database. The number of demisyllables needed to be stored in the database is around 1400. To reduce the number of demisyllables to half, we can transform the vowel part of a syllable to its corresponding nasal vowel part. As an initial phase of this work, we have used two methods to transform oral vowel to its corresponding nasal vowel counterpart. The work of transforming the vowel part of a syllable is being in progress.

2. Bangla speech synthesis system
The design of Bangla speech synthesis system as shown in Fig. 2 is obtained from the general speech synthesis system [2]. The synthesis system has two main components, a text-to-segment conversion component and a segment-to-speech synthesizer. The input to the system is text either from a text file or the characters are typed in to the system and the system produces corresponding speech output. The database contains demisyllable parameters consisting of voiced/unvoiced decision parameter, pitch period, cepstral coefficients etc. The system rule section contains demisyllable connection rules, interval rules and tonation rules. The text analysis system segments the text into demisyllables and the necessary information is inputted to the synthesis sub-system for segment-to-speech conversion. Synthesis sub-system takes in the required information from the database section and produces corresponding text to speech output.

Figure 1: Spectrum of Bangla Oral and nasal vowel /i/.

Figure 2: Block diagram of Bangla speech synthesis system
sound by concatenating the given speech units according to the system rule.

2.1. Analysis - synthesis technique

2.1.1. Analysis Technique

The characteristic of nasal vowel spectrum has additional pole-zero pair as compared to its oral counterpart as studied by many researchers. The cepstral model approximates both the formants and the antiformants of the vocal tract, so adoption of this method is appropriate for our work.

The experimental part consists of recording each of the pronounced syllables at a normal speaking rate in a quiet room by a male native Bangla speaker in a DAT tape at a sampling rate of 48 kHz and 16 bit value. These digitized speech sound are then down-sampled to 10 kHz for the purpose of analysis. In this work demisyllable is chosen as the basic speech unit as it has a reasonable number of speech units for Bangla which is suitable for good synthesized speech quality as well as memory required for storage. Then syllables were segmented to get the required speech unit. In the analysis phase, short-term cepstral analysis method is used to extract the speech parameters. The speech wave is segmented to 25.6ms frame length and Blackman window of the same length as the frame length is used.

Frame shifting time of 10ms is used. Cepstrum is the inverse Fourier transform of the short time logarithm amplitude spectrum of the speech waveform [3]. From cepstrum, the voiced unvoiced decision, pitch period, vocal tract parameters etc. of a speech segment are obtained and are subsequently stored in the database.

2.1.2. Synthesis Technique

The speech synthesis section shown in Fig. 3 is designed on the source-system model [3]. The vocal tract features can be suitably represented for all speech sounds by the pole zero LMA filter [4] proposed by Imai. LMA filter is a part of homomorphic vocoder. It is made from cascade of 30 elemental second order filters in this system. In this method of speech synthesis, the LMA filter representing the vocal tract is driven by an adequate excitation source. If the speech to be produced is unvoiced, the excitation is noise having unit amplitude and random polarity. In case of producing voiced speech, unit impulses separated by pitch period interval is used.

2.2. Listening test result

Listening test result of the system shown in Fig. 4, was done by seven native Bangla speakers with normal hearing ability. The 3 vowels /i/, /a/, and /u/ were selected and syllable data of their combination with all the initial and final consonants were prepared. Each sound was played randomly in a quite room in 3 seconds interval. For consonant detection most of the confusion occurred between aspirated and its non-aspirated counterpart.

3. Transformation of oral vowels to their corresponding nasal counterparts

The spectrum of nasalized vowels has been studied by many researchers [3],[5],[6],[7]. The characteristic of all nasal vowel spectrum as compared to their oral counterparts cannot be described by a single rule. Nasal vowels are produced by coupling of the nasal tract with the oral tract acting as a side branch resonator. This coupling shifts the center of gravity of the formants and introduces additional poles and zeroes in the spectrum in the vicinity of the first formant and in mid-frequency region, which may or may not appear in the smoothed spectrum (depends on the vowel, amount of vowel nasalization, nasal cavity structure of individual) but increase of formant bandwidth is clearly seen. Another characteristic may
be the cancellation of 3rd formant with higher amplitude.

The principle and most consistent consequence on the acoustic spectrum as well as perception appears to be in the vicinity of first formant. Nasal zero always lies between the nasal and oral pole. In the spectrum of high vowels a zero-pole appears after the first formant and for low vowels the nasal pole-zero appears before the oral pole. In any event the end result is the replacement of the first formant with a pole-zero-pole combination [6].

As a first step to transform the oral vowel part of syllable to their corresponding nasal vowel counterpart we transformed oral vowels to nasal vowels. The transformation of log frequency spectral parameters was done by two methods (1) by non-linear transformation and (2) using Sine curve model.

3.1. Non-linear transformation using neural network

Artificial neural networks are modeled after neuron with weighted links interconnecting the units together. If a multi-layer network is trained properly by backpropagation algorithm using weight adjustment based on the sigmoid function, it tends to give reasonable answers when presented with inputs that they have never seen. In our work, we have adopted the backpropagation method of training a three layer weight neural network. We also checked the same work of transformation with increasing no of hidden layers. The transformation result is not improved to a detectable range, so we have kept the network architecture as described above. Neural network was also used for voice conversion [8] in earlier works.

The neural network was trained by giving oral vowel spectrum in the input and the target was the corresponding nasal vowel spectrum. The mean sum square error of training is shown in Fig. 5. The weights of this training is used later to produce data open test nasal vowel from the oral vowel.

![Figure 5: Training error of neural network.](image)

3.2. Transformation using Sine curve model

We want to use a generalized rule in our rule based speech synthesizer for transforming all the seven Bangla vowels to their nasal counterpart. So we have added the model that the first formant of oral vowel is replaced by sine curve of one and a half cycle representing the pole-zero-pole(PZP) combination as shown in Fig. 7. In the mid frequency region, we have added another sine curve of one cycle representing a pole-zero(PZ) combination as shown in Fig. 8. Fig. 6 shows the transformed vowel spectrum as compared to the oral vowel and nasal vowel spectrum.

![Figure 6: Original oral, nasal and transformed spectrum for vowel /a/.](image)

Amplitude of PZP = ±A(dB). Amplitude of PZ = ±B(dB). Centre point of PZP from 0 kHz = M kHz. Width of PZP from its centre point = N kHz. Centre point of PZ from 0 kHz = P kHz. Width of PZ from its center point = Q kHz. Values of necessary parameters for transforming vowels are given in Table 1.

![Figure 7: PZP rule for spectrum modification.](image)

![Figure 8: PZ rule for spectrum modification.](image)

As a preliminary test, we also checked the result without the second pole-zero pair addition, the nasal effect is not so much affected with the removal. So, we may say that, although the nasal vowel spectrum is seen to be changed more or less throughout, as compared to its oral vowel counterpart, the most noticeable range is the low frequency for perception.
Table 1: Values of necessary parameters used for vowel transformation.

<table>
<thead>
<tr>
<th></th>
<th>PZP</th>
<th></th>
<th></th>
<th>PZ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A (dB)</td>
<td>M (kHz)</td>
<td>N (kHz)</td>
<td>B (dB)</td>
</tr>
<tr>
<td>/i/</td>
<td>8.6</td>
<td>0.335</td>
<td>0.546</td>
<td>10.8</td>
</tr>
<tr>
<td>/e/</td>
<td>8.6</td>
<td>0.335</td>
<td>0.351</td>
<td>10.4</td>
</tr>
<tr>
<td>/æ/</td>
<td>8.6</td>
<td>0.335</td>
<td>0.429</td>
<td>10.8</td>
</tr>
<tr>
<td>/a/</td>
<td>8.8</td>
<td>0.343</td>
<td>0.429</td>
<td>13.0</td>
</tr>
<tr>
<td>/ç/</td>
<td>11.7</td>
<td>0.457</td>
<td>0.390</td>
<td>13.8</td>
</tr>
<tr>
<td>/o/</td>
<td>13.0</td>
<td>0.507</td>
<td>0.664</td>
<td>13.0</td>
</tr>
<tr>
<td>/u/</td>
<td>8.6</td>
<td>0.335</td>
<td>0.429</td>
<td>13.0</td>
</tr>
</tbody>
</table>

3.3 Comparison of the two transformation methods

In this section we evaluate the two transformation methods by Listening test and by measuring Euclidean distance between the transformed spectrum and the original nasal vowel spectrum. The overall result of listening test of different type of data involved in vowel transformation is shown in Fig. 9. Fig. 10 shows the calculated average Euclidean distance between the transformed spectrum and their target nasal vowel spectrum. Although the average Euclidean distance is lower by data open test obtained by neural network method (6dB) than that obtained by sine curve method (4dB), but listening test result of sine curve method gave better recognition score (91%) than neural network (86%). This may be explained as follows. By neural network transformation, error is distributed all over the transformed spectrum and overtraining may occur resulting in phoneme confusion. The important information for nasalization is concentrated in at some frequency region where the PZP and PZ rule is applied. So, the recognition score of listening test is greater for Sine curve transformation method.

4. Discussion

We aim at producing rule based Bangla text to speech synthesis system with compact database giving natural like speech output. Although from rule based speech synthesizers it is not so easy to get highly natural speech output, it becomes even more difficult to get natural like nasal vowel sound. So we have adopted cepstral analysis technique which can model both pole and zero of the vocal tract transfer function as nasal vowel has both pole and zero. Listening test result of the system and the transformed vowels shows an acceptable recognition score. Most of the errors in system listening test were due to confusion between aspirated consonants and its non-aspirated counterpart. In case of transformation methods, confusion between the mid-front, mid-back vowels gave errors.

5. Conclusions

In this work we have presented a rule based speech synthesis system for Bangla. One important feature of Bangla language which is nasalization is also studied. Some methods are discussed to reduce the database half the numbers needed. Listening test score of the system and transformed phonemes gave acceptable result.

6. References