ACOUSTIC ANALYSIS OF BANGLA CONSONANTS
Firoj Alam, S.M. Murtaza Habib, Mumit Khan
Center for Research on Bangla Language Processing, BRAC University
{firojalam, habimurtaza, mumit}@bracu.ac.bd

ABSTRACT
This paper describes the acoustic characteristics of Bangla consonants, obtained by analyzing the recordings of male and female voices. First, the duration of each phoneme was identified by averaging both the male and female voice data; then, formant were measured and formant comparison was made for controversial phonemes, which also served to resolve the controversies in the existing phoneme inventories; and finally, a consonant phoneme inventory was designed.

Index Terms — Phoneme inventory, Speech Synthesis, Speech Recognition

1. INTRODUCTION
The goal of this paper is to determine the total number of consonant phonemes and their acoustic properties in Bangla language. This analysis is an essential component in linguistic of a language and in the diphone concatenation technique where proper duration and prosodic characteristics are needed to synthesize natural sounding speech. With nearly 200 million native speakers, Bangla (exonym: Bengali) is one of the most widely spoken languages of the world (it is ranked between four and seven based on the number of speakers). There have been quite a few articulatory investigations on Bangla phonemes during the last several decades. These analyses have resulted in a set of phoneme inventories, each with several phonemes that are controversial. One of the great motivations of this work is to solve these controversies. This study is focused on acoustic evidences of Bangla phonemes based on the acoustic cues of Pickett [7], durational characteristics and comparison of controversial phonemes. Acoustic evidence is the perfect cue to identify aspiration, voicing, duration and other important features in a phoneme. The durational characteristics, closure, release burst, turbulent, gliding and the formant transitions in the onset and offset were measured for each class of consonant phonemes. Articulatory investigation like minimal pair testing did not consider in this study because it does not help in controversial phonemes in Bangla. A list of dictionary words containing all possible phonemes was selected for recording. The carrier sentences were designed with /cTj/ and /aCa/ patterns. These sentences were then recorded by both professional and non-professional male and female speakers. The recorded data was analyzed with the help of Praat [8].

2. LITERATURE REVIEW
There have been several studies in the past, mostly based on articulatory phonetics, of the articulatory and acoustic properties of Bangla consonants. It is showed in [1] that Bangla have 20 stops (B/ŋ/, B/k/, B/g/, B/ŋg/), 3 nasals (B/n/, B/m/, B/ŋm/), 2 glides (B/ʃ/, B/ʒ/), 5 fricatives (B/s/, B/z/, B/ʃ/, B/ʒ/, B/ʃʃ/), 7 nasals (B/n/, B/ŋ/, B/m/, B/ŋm/), 1 lateral (B/l/), 1 trill (B/r/), 2 flaps (B/ɬ/, B/ɮ/), 1 glide (B/ʎ/) as a total of 36 consonants. Incidentally, Abdul Hai [1] claimed that the sound produced by three letters B/ʃ/, B/ʒ/, B/ʃʃ/ is represented by a single phoneme /ʃ/. Daniul Huq [2] showed 21 stops (B/k/, B/ŋg/), 3 nasals (B/n/, B/m/, B/ŋm/), 5 fricatives (B/s/, B/ʃ/, B/ʒ/, B/ʃʃ/), 2 flaps (B/ɬ/, B/ɮ/) as a total of 35 consonants. It is explained in [6] that Bangla have 20 stops (B/k/, B/ŋg/), 3 nasals (B/n/, B/m/, B/ŋm/), 2 glides (B/ɬ/, B/ɮ/) as a total of 35 consonants. A brief literature review is given in section 2, followed by a description of the methodology in section 3. The analytical results are presented and discussed in section 4. A summary and conclusions of the study are given in section 5.

This study seeks to find the answer of the following three questions in order to develop a Bangla phoneme inventory: (1) to find the comparison between controversial phoneme pairs; (2) to determine the durational values; and (3) to determine the formant characteristics of the consonants in spoken Bangla utterances for the phoneme inventory. A list of dictionary words embedded in carrier utterances was chosen for the analysis. Different patterns were selected for identifying the list of words. This data was recorded by a number of speakers and then analyzed by the Praat software.

3.1. Recording material

The list of words selected for this investigation consists of all possible phonemes with the following two patterns: vCv [iCi] and vCv [aCa], embedded in carrier words to form utterances. The reason for recording consonants in a carrier utterance is so that the context remains the same. So a total of 35x2 (35 possible phonemes x 2 patterns) utterances were selected for recording of the following form.

1. aCa pattern
   अम্রা কাজ পাই → k
   1P Pl work get.pres [We get work.]

2. iCi pattern
   আমি কেছু পাই → k
   1Sg some get.pres [I get something.]

The first character of the second word is the target phoneme in both patterns. The same carrier words were used for all target phonemes. Pickett [7, pp-87] summarizes from Umeda (1977) that the duration of consonants are also affected by the syllable stress, emphasis and position of the consonant in a word. So the utterances were selected in such a way such that the prosodic variation (such as stress, tone, emphasis and vocal effort) and feature dependent segment duration do not have any effect on the target phoneme. Also, the manner of articulation was considered when these utterances were collected, as the manner of articulation is the usual first basis for segmentation or duration calculation. All words are phonetically symmetrical i.e. the grapheme to phoneme correspondence is regular, an assertion that was confirmed by linguists.

3.2. Speaker selection

Both professional and non-professional male and female speakers were selected by considering different ages, heights and the speakers' locality in Bangladesh.
Unfortunately, we were unable to include any speaker from the Indian State of West Bengal in this analysis. Four male and four female speakers, with equal numbers of professional vs. non-professional male speakers, were selected. The professional speakers’ ages ranged from 52 to 54 and non-professional speakers’ ages ranged from 25 to 29. Each speaker was given flash cards containing the utterances, and was asked to record each utterance in straight tone/pitch level and without assigning any stress in a word.

3.3. Recording

The recording of the utterances was done using the Nundo speech processing software. A professional recording studio was chosen to record the utterances. The equipment consisted of an integrated Tascam TM-D4000 Digital-Mixer, a high fidelity noise free Auditechnica microphone and two high quality speakers. The recorded waveform files were used for acoustic analysis with Praat version 4.6.27. The speakers were asked to keep a distance of 10-12 inches from the microphone. The speech data was digitized at 44100 Hz at 24-bit resolution and stored as wave format. After each recording, the moderator checked for any wrong pronunciation during the recording, and if so, the affected utterances were re-recorded.

3.4. Analysis

Total 35x2x8 = 560 (35 possible phonemes x 2 patterns x 8 speakers) segment were analyzed in this study. Scarborough R. [9] showed a set of segmentation criteria (i.e beginning and ending position of each classes of phoneme), in her lecture which was considered when the duration was calculated using Praat. The duration for each phoneme (both closure and leg VOT-Voice onset time) was computed from the recorded voice. The start position of the plosive is the end of preceding vowel and the end position of the plosive is the beginning of the voicing of the following vowel. The frication and aspiration after the release burst of plosives was considered a part of the consonants. The average value was then calculated for each phoneme. Durations for both male and female speakers were compared. Onset and offset formant movement of consonant phonemes was observed and noted to identify place of articulation. Acoustic cues of all phonemes were classified by place and manner of articulation. Then the comparisons between the controversial phonemes were identified as is explained in the result section.

Pratat settings: The Praat spectrogram and formant settings were maintained for both male and female speakers. In spectrogram settings the window length is 0.005 second, the window shape Gaussian and the view range up to 5000 Hz. The Fourier analysis method is used in spectrogram analysis settings. In formant settings, we used maximum formant 5500 Hz for five formants with window length 0.025 second. Pitch range was used 75 Hz to 500 Hz which covers both male and female speakers.

4. RESULTS

The durations, controversial phoneme comparison and consonant phoneme inventory were computed and identified in this analysis. The output of the analysis is presented and discussed in this section. The durational characteristics of Bangla consonant phonemes are shown in Table 1 and Table 2. Table 3 shows the consonant phoneme inventory of Bangla. According to the different formant value we classified and identified the place and manner feature of consonant phonemes which is explained in section 4.3. The following controversial phonemes /ʃ/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/ solved in two steps. Formants value were observed for these phonemes /ʃ/, /ʃ̩/, /ʃ̩/, /ʃ̩/ and it is identified that the phoneme /ʃ̩/, (fricative) has strong frequency above 4500 Hz and frication of the palatal phoneme /ʃ̩/ start at 1600 Hz and strong high frequency above 3000 Hz. So the two phonemes were eliminated in this step. In second step, formant comparison were made for these phoneme /ʃ/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/, /ʃ̩/ explained in section 4.2 and then we identified three phonemes among the six phonemes. So finally we concluded 30 phonemes from the 35 phonemes with their acoustic values.

4.1. Durational characteristics

In Table 1, column 1 shows the plosive phonemes and the rest of the column shows the average duration of male and female, total average of both male and female and standard deviation of both male and female. Closure and VOT were calculated separately in each case. It was observed that the duration of the aspirated sound is longer than the unaspirated one and VOT (due to aspiration) is longer than closure. The duration of glide, trill, flap identification was the most difficult part during analysis as start and end segment could not be identified easily. Table 2 shows the durational value of fricative, nasal, trill, flap and glide.
4.2. Controversial comparison

There is a controversy in terms of the sound produced between the following pairs of letters: "N/n/-"I/n/; "T/͸/-"B/͸; "j/ӫ/-"k/ӫ. Formants were measured in whole durations for each pair of sounds by Praat formant listing. Praat uses 25ms window length, five formants and maximum 5500 Hz in formant settings. Each pair is then compared as shown in the graph. The same procedure is applied for every pair. The graph shown here is the calculated value of one male voice. Other speaker comparison values are similar, which has not been shown. The spectrographic analysis of letter "N/n/ in ^LN\]U/ and letter "I/n/ in ;I\]ka/ shows that both are the same phoneme /n/ as shown in Figure 1. The average duration of "N/n/ is 69.69 (msec) and "I/n/ is 68.44 (msec).

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Male (Avg)</th>
<th>Female (Avg)</th>
<th>Male (Std)</th>
<th>Female (Std)</th>
<th>Total (Avg)</th>
<th>Total (Std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/n/</td>
<td>105.11</td>
<td>109.53</td>
<td>9.08</td>
<td>7.62</td>
<td>107.32</td>
<td>57.55</td>
</tr>
<tr>
<td>I/n/</td>
<td>86.51</td>
<td>83.54</td>
<td>10.22</td>
<td>15.06</td>
<td>84.93</td>
<td>21.48</td>
</tr>
<tr>
<td>T/͸/</td>
<td>82.86</td>
<td>84.56</td>
<td>29.49</td>
<td>7.66</td>
<td>83.71</td>
<td>22.45</td>
</tr>
<tr>
<td>B/͸/</td>
<td>77.98</td>
<td>74.64</td>
<td>29.57</td>
<td>11.47</td>
<td>76.31</td>
<td>23.93</td>
</tr>
<tr>
<td>j/ӫ/-</td>
<td>93.55</td>
<td>78.30</td>
<td>19.22</td>
<td>9.23</td>
<td>85.93</td>
<td>20.91</td>
</tr>
<tr>
<td>k/ӫ/</td>
<td>72.88</td>
<td>70.30</td>
<td>11.02</td>
<td>30.10</td>
<td>71.61</td>
<td>11.76</td>
</tr>
<tr>
<td>/c/</td>
<td>67.50</td>
<td>68.10</td>
<td>13.70</td>
<td>10.73</td>
<td>67.80</td>
<td>3.98</td>
</tr>
<tr>
<td>/d/</td>
<td>64.83</td>
<td>57.17</td>
<td>10.23</td>
<td>19.15</td>
<td>61.00</td>
<td>9.81</td>
</tr>
<tr>
<td>/p/</td>
<td>130.22</td>
<td>96.56</td>
<td>12.03</td>
<td>4.41</td>
<td>113.39</td>
<td>14.24</td>
</tr>
<tr>
<td>/t/</td>
<td>102.34</td>
<td>83.97</td>
<td>12.85</td>
<td>14.87</td>
<td>93.15</td>
<td>17.64</td>
</tr>
<tr>
<td>/k/</td>
<td>88.58</td>
<td>83.49</td>
<td>10.27</td>
<td>2.92</td>
<td>86.03</td>
<td>12.29</td>
</tr>
<tr>
<td>/h/</td>
<td>86.16</td>
<td>71.81</td>
<td>9.79</td>
<td>32.90</td>
<td>78.98</td>
<td>19.72</td>
</tr>
<tr>
<td>/l/</td>
<td>131.75</td>
<td>120.35</td>
<td>12.13</td>
<td>5.84</td>
<td>126.05</td>
<td>14.72</td>
</tr>
<tr>
<td>/r/</td>
<td>139.98</td>
<td>89.10</td>
<td>15.93</td>
<td>17.34</td>
<td>114.54</td>
<td>54.19</td>
</tr>
<tr>
<td>/m/</td>
<td>100.35</td>
<td>85.77</td>
<td>11.76</td>
<td>33.84</td>
<td>98.52</td>
<td>48.95</td>
</tr>
<tr>
<td>/n/</td>
<td>197.02</td>
<td>103.95</td>
<td>22.37</td>
<td>13.74</td>
<td>175.14</td>
<td>23.66</td>
</tr>
<tr>
<td>/o/</td>
<td>65.81</td>
<td>72.32</td>
<td>18.36</td>
<td>38.25</td>
<td>69.07</td>
<td>30.49</td>
</tr>
<tr>
<td>/u/</td>
<td>62.35</td>
<td>67.07</td>
<td>9.30</td>
<td>13.63</td>
<td>64.71</td>
<td>11.09</td>
</tr>
<tr>
<td>/v/</td>
<td>108.30</td>
<td>93.10</td>
<td>19.36</td>
<td>10.22</td>
<td>100.70</td>
<td>14.47</td>
</tr>
<tr>
<td>/g/</td>
<td>81.64</td>
<td>69.77</td>
<td>10.22</td>
<td>30.46</td>
<td>75.70</td>
<td>24.71</td>
</tr>
<tr>
<td>/z/</td>
<td>156.17</td>
<td>185.43</td>
<td>149.87</td>
<td>87.49</td>
<td>120.80</td>
<td>55.29</td>
</tr>
</tbody>
</table>

Table 1: Average duration of Stops in millisecond

Table 2: Average duration of fricative, nasals, trill, lateral, flap and approximant in millisecond

4.2. Controversial comparison

There is a controversy in terms of the sound produced between the following pairs of letters: "N/n/;"."I/n/; "T/͸/; "B/͸/; "j/ӫ/; "k/ӫ/. Formants were measured in whole durations for each pair of sounds by Praat formant listing. Praat uses 25ms window length, five formants and maximum 5500 Hz in formant settings. Each pair is then compared as shown in the graph. The same procedure is applied for every pair. The graph shown here is the calculated value of one male voice. Other speaker comparison values are similar, which has not been shown. The spectrographic analysis of letter "N/n/ in ^LN\]U/ and letter "I/n/ in ;I\]ka/ shows that both are the same phoneme /n/ as shown in Figure 1. The average duration of "N/n/ is 69.69 (msec) and "I/n/ is 68.44 (msec). Different
linguists claimed that the two letters "র"/j/ and "ঈ"/j/ represent two phonemes in Bangla. But in our analysis we identified that both are same phoneme as is shown in Figure 2. The average duration of "র"/j/ is 67.33 (msec) and "ঈ"/j/ is 68.27 (msec). The average duration of "ঝ"/c/ is 50.90 (msec) and "ঞ"/c/ is 48.29 (msec). The formant graph of the sound produced by the letters "ঝ"/c/ in ঝাক/scarf/ and "ঞ"/c/ in ঝল্লা/jhalla/ is shown in Figure 3. It shows some differences in F2 formants. More data may be required for exhaustive analysis. For now we consider these as a single phoneme /c/.

![Figure 1: Comparison of sound produced by the letter ছ and জ](image1)

Another finding of our analysis is that the sound produced by the letters "র"/j/ and "ঈ"/j/ is the phoneme /j/. But the letter "র"/j/ and "ঈ"/j/ also produce another sound which is identified as phoneme /j/.

![Figure 2: Comparison of sound produced by the letter ছ and জ](image2)

4.3. Consonants phoneme inventory

This study identified 30 consonant phonemes in Bangla as shown in table 3, including 20 plosives, 3 nasals, 1 trill, 1 flap, 3 fricatives, 1 lateral and 1 approximent. The plosive sound in each place composed of two voiced versus two voiceless and two aspirated versus two in-aspirated. Each aspirated sound is represented by superscript (ʰ).

**Bilabial:** The average formant value of the following vowel of bilabial plosive shows that, F2 is upward. Another significant acoustic cue is the weak and diffuse spectrum in the release burst. The four bilabial plosives composed of two voiced (৩্চ/শ, ৩্চ/শʰ) versus two voiceless (৩্চ/শ, ৩্চ/শʰ) and one bilabial nasal (৩্চ/শ). The bilabial nasal also has an upward F2, strong frequency up to about 600Hz and very weak mid-formants.

**Dental:** The four dental plosives (২্চ/শ, ২্চ/শʰ, ২্চ/শ, ২্চ/শʰ) have downward F2 of the following vowel and have high frequency energy in release burst.

**Alveolar:** There are four alveolar plosives (৩্চ/শ, ৩্চ/শʰ, ৩্চ/শ, ৩্চ/শʰ) of Bangla which have a downward F2 of the following vowel. Bangla has alveolar nasal, trill, flap, fricative and lateral. The spectrogram of the alveolar nasal phoneme shows that, the F2 of the following vowel is downward. The phoneme trill ৩্চ/শ has high intensity on F1, which gradually decreases in F2, F3 and F4, low freq strong up to 835 Hz, mid freq stronger than nasals and following vowel F2 is upward. The phoneme ৩্চ/শ (flap) follow vowel transition on F1, F2 and F3. The phoneme ৩্চ/শ (lateral) have high frequency at 400 Hz, 1350 Hz, and 3300 Hz, the intensity of the formants gradually decrease after the first formants. The phoneme ৩্চ/শ (fricative) has strong frequency above 4500 Hz.
Post-alveolar: It is observed that the spectrogram of the post-alveolar plosives (ץ/c/, ש/ɡ/ , ו/ʃ/ , ז/ʒ/) shows downward F2 transition and generally stronger high frequency energy after release burst.

Palatal: Frication of the palatal phoneme ח/υ/ starts at 1600 Hz. Strong high freq above 3000 Hz. The phoneme ח/υ/ has different spectrographic patterns in different context.

Velar: The spectrogram of the velar phonemes (4 plosives צ/k/, ש/ɡ/ , ו/ʃ/ , ז/ʒ/) and 1 nasal ח/υ/ shows the divergent F2 and F3 of the preceding and following vowel. The phoneme ח/υ/ have high intensity at 300 Hz, 1300 Hz, 2300 Hz and 3800 Hz. Release transient on F1 and F2 of the following vowel.

Glottal: The fricative ח/י/ phoneme have the strongest resonance at about 1100 Hz and go upward.

5. CONCLUSION

Here we described the duration of each consonant phoneme and we also identified the place and manner of articulation of consonants in the phoneme inventory. We also conclude that Bangla consonant phoneme inventory consist of 30 phonemes. This phoneme inventory can be used in all Bangla linguistic components and to develop speech application in Bangla. It may also help in diphone database for speech synthesis, speech recognition as well as speech processing such as speech-to-speech translation.

6. ACKNOWLEDGMENTS

This work has been supported in part by the PAN Localization Project (www.pan110m.net), grant from the International Development Research Center (IDRC), Ottawa, Canada, administrated through Center for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences, Pakistan. We would also like to thank Dr Sarmad Hussain (NUCES), Sameer Ud Daula (UCLA), Naira Khan (Dhaka University) and BRAC University students who helped by providing their speech for analysis.

7. REFERENCES