Bengali Diphone Duration Modeling for Bengali Text to Speech Synthesis System

Labiba Jahan[a], Umme Kulsum[b], Abu Naser[c]
[a, b, c] Shahjalal University of Science & Technology
Sylhet, Bangladesh
labibasu@gmail.com, Kulsum51@gmail.com, Abu.naser002@gmail.com

Abstract. This paper elaborates the analysis and implementation of a durational model of diphone for Bengali Text To Speech. Our analysis focused on duration of diphone according to several categories of consonant. Here we have proposed and implemented a durational model of diphone based on pronunciation place of consonant. This durational model is convenient to any diphone based Bengali Text to Speech Synthesizer. We have implemented our proposed durational model of diphone on Bengali Text To Speech synthesis software "Subachan". Outcomes of this implementation is satisfactory. We have enhanced the overall performance of "Subachan" successfully with new diphone set.

Keywords: Consonant, Corpus Selection, Diphone Preparation, Durational Model.

1 INTRODUCTION

Text to speech (TTS) synthesizer is a modern computer system which converts normal language text into its speech by applying some linguistic rules & algorithm. It has a broad research and application area in the modern Human Computer Interaction (HCI) systems (Human computer interaction on Wikipedia).It has significant and widespread applications in education, entertainment, business, and especially for the people with visual impairment and dyslexia (Abu Naser *et al.* 2010).

Fig. 1. Block diagram of text-to-Speech Synthesis System (Muhammad Masud Rashid *et al.* 2008).

The two primary technologies for speech synthesis are formant synthesis (CRBLP online) and concatenated synthesis (Festival online). Formant synthesis converts text to its speech based on an acoustic model. The intelligibility of formant synthesis is relatively high than its naturalness. It is widely used for its high speed & low memory (Abu Naser *et al.* 2010, CRBLP online).On the other hand, concatenated synthesis converts text to its speech based on human speech samples. The naturalness of concatenative synthesis is relatively high. The three main sub types of concatenative synthesis are unit selection synthesis, diphone synthesis and domain-specific synthesis (Accents, Symbols & Foreign Scripts online).

BRAC developed a Bengali TTS synthesis system (Firoj Alam *et al.* 2007). "KOTHA" using festival (A. G. Ramakrishnan *et al.* 2002). Festival is a multilingual speech synthesis system which provides general framework for building speech synthesis. This too big & slow system used 4355 diphones. It takes two seconds to generate a ten second utterance. Moreover, we can't implement Bengali character directly in festival. Another Bengali TTS is "BANGLA VAANI" which is synthesis system of Kolkata based Bengali language. This is a syllable based synthesis system. SUST developed a diphone preparation technique for Bengali TTS synthesis named "SUBACHAN" which is written entirely in Java programming language. It used 527 diphones. It uses a minimum diphone set for Bengali Text to Speech Synthesis. It takes 45 ms to generate a ten second utterance (Abu Naser *et al.* January 11-13).

In this paper we have described the whole process of preparation and implementation of a durational model of diphone based on pronunciation place of consonant.

2 RELATED WORKS

There have been several studies of duration analysis in the past, mostly based on articulatory and acoustic properties of Bengali vowel & consonant.

BRAC University measured duration of each vowel phoneme as well as identified acoustic features of Bengali vowel phoneme inventory. They also described that Bengali vowels consist of 14 monophthong and 21 diphthong phonemes. The duration of each phoneme was identified by averaging both the male and female voice data (Firoj Alam *et al.* 2008).

BRAC also measured duration of each consonant phoneme as well as identified the place & manner of articulation of Bengali consonant phoneme inventory. They described that Bengali consonants consist of 30 phonemes. The duration of each phoneme was identified by averaging both the male and female voice data (Firoj Alam *et al.* 2008).

CDAC-Kolkata presented a method of duration modeling based on the nucleus vowel duration of Bengali. They explained the study of duration variation with respect to syllable position, type and context of occurrence etc. (Rajib Roy *et al.* 2008).

All previous durational analysis was based on duration of vowel or consonant but we have focused on duration of diphone which is never studied before.

3 METHODOLOGY

3.1 Diphone analysis

In phonetics, a diphone is an adjacent pair of phones. It is usually used to refer a recording of the transition from middle of one phone to middle of another phone (Abu Naser *et al.* 2010).

We have worked with all diphones related with 29 consonants (mainly) and 7 vowels (total 464 diphones). The table listed all diphones we have mapped and categorized.

Table 1. Diphone list

Diphone Type	Number
{C	29
C}	29
CV	203
VC	203
Total	468

{C = Starting of consonant e.g. {ক, {চ, {ট etc.
C} = Ending of consonants e.g. ক}, চ}, ট} etc.
CV = Consonant + Vowel e.g. কও, চই, টঅ etc.
VC = Vowel + Consonant e.g. ওম, ইচ, অট etc.

3.2 Consonant analysis

According to Wikipedia (Firoj Alam *et al.* 2008) Bangla has 29 consonants with 20 stops (ক, খ, গ, ঘ, চ, ছ, জ, ঝ, ট, ঠ, ড, ঢ, ত, থ, দ, ধ, প, ফ, ব, ভ), 3 nasals (ঙ, ন, ম), 3 fricative (শ, স, হ), 3 liquids (ল, র, ড়).It is showed in[1] that Bangla have 20 stops(ক, খ, গ, ঘ, চ, ছ, জ, ঝ, ট, ঠ, ড, ঢ, ত, থ, দ, ধ, প, ফ, ব, ভ), 7 fricatives (শ, স, ষ, য, ব, হ, ঃ), 4 nasals (ঙ,ং, ন, ণ, ম), 1 lateral (ল), 1 trill (র), 2 flaps (ড়, ঢ়), 1 glide (য়) as a total of 36 consonants. Incidentally, Abdul Hai (Human computer interaction, Wikipedia) claimed that the sound produced by three letters শ, স, ষ is represented by a single phoneme. Daniul Huq (Muhammad Masud Rashid *et al.* 2008) showed 21 stops(ক, খ,

গ, ঘ, চ, ছ, জ, ঝ, ট, ঠ, ড, ঢ, ত, থ, দ, ধ, প, ফ, ফ',ব, ভ), 5 fricatives (ব, স, জ, শ, হ), 3 nasals (ঙ, ন, ম), 1 lateral (ল), 1 trill (র), 2 flaps (ড়, ঢ়), 2 glides (ব, য়) (total of 35 consonants).

We have considered three categories of consonant for our analysis.

1. According to air emerges from lungs during the time of pronunciation.

Small portion of air emerges from lungs during the time of pronunciation – ক, চ, ট, ত, প, গ, জ, ড, দ, ব।

Large portion of air emerges from lungs during the time of pronunciation – খ, ছ, ঠ, থ, ফ, ঘ, ঝ, ঢ, ধ, ভ।

2. According to vibration of vocal.

Vocal doesn't vibrate during the time of pronunciation – ক, চ, ট, ত, প, খ, ছ, ঠ, থ, ফ।

Vocal vibrates during the time of pronunciation – গ, জ, ড, দ, ব, ঘ, ঝ, ঢ, ধ, ভ, ঙ, ঞ, ণ, ন, ম।

3. According to their pronunciation place.

Type: stops

Velar (Pronunciation place: Vocal) – ক, খ, গ, ঘ, ঙ

Post alveolar (Pronunciation place: Hard Palate) – চ, ছ, জ, ঝ, ঞ

Alveolar (Pronunciation place: Soft Palate) – ট, ঠ, ড, ঢ, ণ

Dental (Pronunciation place: Teeth) – ত, থ, দ, ধ, ন

Bilabial (Pronunciation place: Lips) – প, ফ, ব, ভ, ম

Type: fricatives

Palatal (Pronunciation place: Palate) – শ, স, ষ

Type: Approximant

Palatal (Pronunciation place: Palate) – য়

Type: flap

Alveolar (Pronunciation place: Soft Palate) – ড়, ঢ়

Type: trill

Alveolar (Pronunciation place: Soft Palate) – র

Type: lateral

Alveolar (Pronunciation place: Soft Palate) – ল

Type: fricatives

Glottal– হ (Bengali alphabet on Wikipedia)

3.3 Corpus selection

Corpus is a large and structured set of texts from where diphones are extracted (Abu Naser *et al.* 2010).

E.g. Corpuses for letter 'ক': কক কাক কিক কুক কেক কোক ক্যাক

We took 4 types of corpus with different frequency (mono 16 bit 44 100 Hz sample rate). Then we generated some words, sentences with them and made a statistic based on given opinion of 50 listeners according to intelligibility & naturalness.

Category 1: high frequency, 1 corpus occurs 3 times. Maximum times we used 2nd corpus.

Category 2: low frequency, 1 corpus occurs 3 times. Maximum times we used 2nd corpus.

Category 3: high frequency, 1 corpus occurs once.

Category 4: low frequency, 1 corpus occurs once.

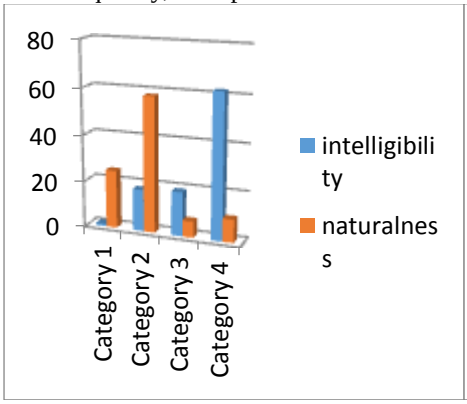

Fig. 2. Bar diagram for word level performance

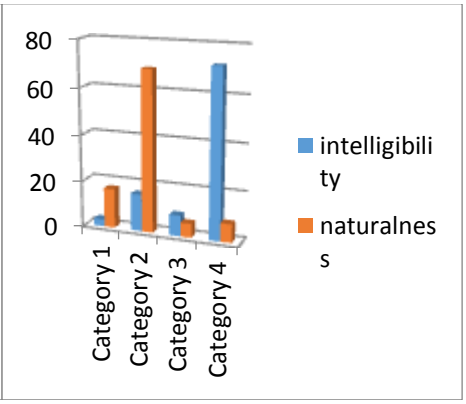

Fig. 3. Bar diagram for sentence level Performance

It is observed in above figures that despite of more natural than category 4, category 2 has low intelligibility rate. We took corpus 4 for our analysis as we focused more on intelligibility rate than naturalness.

3.4 Duration mapping of diphone

At first we took each corpus and cut them in equal duration of diphone.
E.g. Corpus: কাক

Diphones: {ক + কআ + আক +ক}

We took each diphone and chose a word where it places at first position. Then we made five copy of this diphone with different random durations close to the first duration gapping 5 units and applied them to make the word. We synthesized all diphones in a software named "wavelab". After hearing all same five words with different durations in "wavelab" we have selected one word which sounded better than others. We increased or decreased some units of duration when necessary. We also selected another two durations for this diphone in same process for middle and last position in word. Finally we averaged these three durations and fixed the average value for the diphone.
E.g. Corpus: মাম

Diphones: {ম+ মআ + আম+ ম}

Diphone want to map: মআ

Duration after cutting corpus equally: 97 ms
Five random durations: 90ms , 95 ms , 100 ms , 105 ms , 110 ms
Word for first position: মালা

Fixed duration for first position: 107 ms
Word for middle position: আমার

Fixed duration for middle position: 112 ms
Word for last position: জামা

Fixed duration for last position: 111 ms
Final average value for মআ: (107 + 112 + 111)/3 = 110 ms

3.5 Durational model of diphone

We categorized duration of diphone according to three categories of consonant which are - air emerges from lungs when we pronounce consonant, vibration of vocal during the time of pronunciation of consonant & pronunciation place of consonant. For first two categories of consonant there is no congruency among the average durations of diphones. But the last category has an exception. After observations we have represented a durational model of di-phones according to pronunciation place of consonants which is best suited.

Table 2. Average durations of diphone in ms (types: {C, CV, C})

	start	অ	আ	ই	উ	এ	ও	এ্যা	**end**
ক/খ/গ/ঘ	45	100	150	120	90	108	108	93	143
চ/ছ/জ/ঝ	130	103	98	108	101	115	90	89	143
ট/ঠ/ড/ঢ	40	60	93	102	112	120	140	80	143
ত/থ/দ/ধ/ন	45	102	115	123	119	113	121	138	143
প/ফ/ব/ভ/ম	50	115	110	140	100	120	110	110	143
শ/স	50	172	191	214	210	213	212	203	140

র/ল/য	50	170	160	112	127	124	160	162	143
ড়	50	176	168	120	134	132	169	171	143
হ	100	155	150	145	160	139	140	131	140

Table 3. Average durations of diphone in ms (type: VC)

	ক/খ/গ/ঘ	চ/ছ/জ/ঝ	ট/ঠ/ড/ঢ	ত/থ/দ/ধ/ন	প/ফ/ব/ভ/ম	শ/স	র/ল/য	ড়	হ
অ	100	145	45	130	105	180	179	185	123
আ	147	147	72	143	84	182	160	169	154
ই	126	153	77	149	115	166	162	175	129
উ	93	149	91	152	78	209	124	136	148
এ	107	150	99	135	102	200	138	149	130
ও	105	139	123	142	87	192	160	172	140
এ্যা	90	147	63	160	86	211	164	178	129

3.6 System selection

We selected a diphone based Bengali Text to Speech synthesis system “Subachan” to implement our durational model as “Subachan” is only diphone based Text to Speech synthesis system in Bengali language (Abu Naser *et al.* 2010). In “Subachan” they have calculated 527 diphones according to 6 unique vowels and 32 unique consonants.

3.7 Error analysis of system

We took 11838 unique words from a book by Dr Muhammed Zafar Iqbal named “Amar Bondhu Rashed” and found 3322 errors after applying them on “Subachan”. Total error 28.06 %

After detecting errors we have categorized them and found problems with ‘ও-কার’- 60%, problems with ‘এ-কার’-12%, durational problem of diphone - 19%, others - 9%.

It is noticed that the main problems of “Subachan” are durational problem of diphone & problem with ‘ও-কার’ but a large portion of ‘ও-কার’ problem occurs for durational problem of diphone. So, duration of diphone is the main factor of error occurrence in “Subachan”.

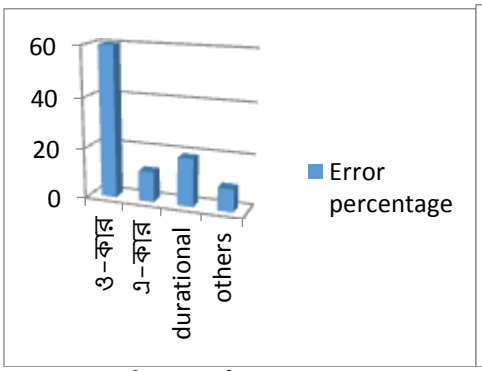

Fig. 4. Bar diagram of error categorization

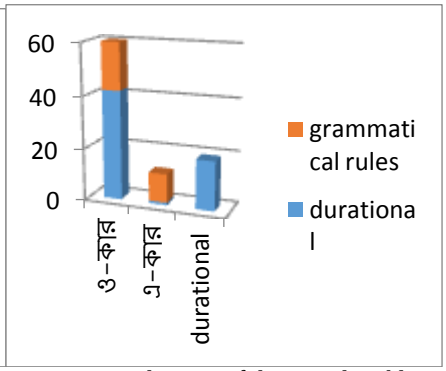

Fig.5. Bar diagram of durational problem

3.8 Performance measure

We took 50 words from the error list which pronounced wrongly for durational problem. After applying our durational model based on pronunciation place of consonant to them we found 37 words correct which is 74%.

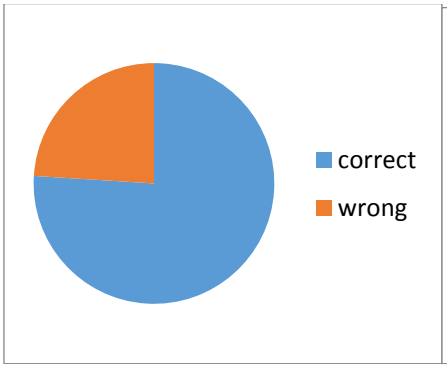

Fig. 6. Pie diagram of performance applying durational model on wrong Words

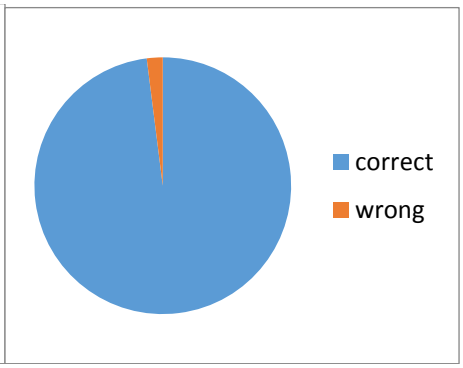

Fig.7. Pie diagram of performance applying durational model on correct words

Again we took 50 correct words. After applying our durational model based on pronunciation place of consonant to them we found 48 words correct which is 98%.

CONCLUSION

The above experimental study leads to the findings that we have increased the accuracy of "Subachan" and reduced the error percentage. With some exceptions, Subachan can work in any situation and produce better performances with this diphone set after applying our durational model. Moreover we need not to map duration for each diphone as we group them in a discipline way. We need future analysis on duration of diphones related with vowel like {V (start of vowel), VV (vowel + vowel), V} (end of vowel).We have to focus on joint letter also.

References

Human computer interaction. Retrieved from http://en.wikipedia.org/wiki/Human%E2%80%93computer_interaction.

Abu Naser, Devojyoti Aich, Md. Ruhul Amin. (18-20 Dec. 2010). Implementation of Subachan: Bengali Text To Speech Synthesis Software.ICECE-2010.

Muhammad Masud Rashid, Md. Akter Hussain, M. Shahidur Rahman. (2008).Text Normalization and Diphone Preparation for Bangla Speech Synthesis. ICCIT-2008, Dhaka, Bangladesh.

Center for Research on Bangla Language Processing. Retrieved from http://crblp.bracu.ac.bd/

Festival Speech Synthesis, Speech Tools & documentation. Retrieved from http://www.festival.org/

Accents, Symbols & Foreign Scripts. Retrieved from http://symbolcodes.tlt.psu.edu/bylanguage/bengalichart.html

Firoj Alam, Promila Kanti Nath and Mumit Khan. (2007).Text To Speech for Bengali Language using Festival. Proc. of 1st International Conference on Digital Communications and Computer Applications (DCCA 2007), Irbid, Jordan.

A. G. Ramakrishnan, G. L. Jayavardhana Rama, R. Muralishankar and R Prathibha. (2002). A Complete Text-To-Speech Synthesis System In Tamil, Proceedings of IEEE Workshop on Speech Synthesis, 191-194.

Abu Naser, Devojyoti Aich and Md. Ruhul Amin. (January 11-13).Architectural Design of Bengali Text to Speech Synthesis Software with Sentence Analysis using Advanced Linguistic Processing Modules: Stemming, Phrase Analysis and Expansion Rules. CERIE- 2010, Sylhet, Bangladesh.

Firoj Alam, S.M. Murtoza Habib, Mumit Khan. (2008). Acoustic analysis of Bangla vowel inventory. BRAC University.

Firoj Alam, S.M. Murtoza Habib, Mumit Khan. (2008). Acoustic analysis of Bangla consonants. BRAC University.

Rajib Roy, Tulika Basu, Arup Saha, Joyanta Basu, Shyamal Kr Das Mandal. (November 12-14, 2008).Duration Modeling for Bangla Text to Speech Synthesis System. International Conference on Asian Language Processing, Chiang Mai, Thailand.

Bengali alphabet on Wikipedia. Retrieved from http://en.wikipedia.org/wiki/Bengali_alphabet#Consonants

Labiba Jahan is a lecturer at Metropolitan University, Sylhet, Bangladesh. She received her BS in Computer Science & Engineering from Shahjalal University of Science & Technology, Sylhet, Bangladesh.

Umme Kulsum is a software engineer at TigerIT Bangladesh Limited, Dhaka, Bangladesh. She received her BS in Computer Science & Engineering from Shahjalal University of Science & Technology, Sylhet, Bangladesh.

Abu Naser is a lecturer at Shahjalal University of Science & Technology, Sylhet, Bangladesh. He received his BS in Computer Science & Engineering from Shahjalal University of Science & Technology, Sylhet, Bangladesh.