

Bangla Speech Synthesizer System for Bangladesh

Mohammad Nurul Huda, Sabbir Arif Siddique, Ishteaque Alam

Department of Computer Science and Engineering

United International University, Dhaka, Bangladesh

mnh@cse.uju.ac.bd, sabbir.a.siddique@gmail.com, ishteaque.ark@gmail.com

Abstract

This paper illustrates the design and implementation of Bangla (widely used as Bengali) Text to Speech (TTS) system from the very raw level without using any third party speech synthesis tool. For constructing the system we have considered two directions, where one is based on phoneme and another one is on syllable. In this study, our proposed system comprises some stages. At first stage audio sounds are recorded for each of the Bangla phonemes and three thousand out of 250000 syllables in Bangla, and then noise is reduced to obtain high quality sounds for each phoneme and syllable. Second stage searches for longest possible matching of the syllables if it is available in the input text, and if not, then searches for the phonemes to match with the corresponding graphemes. For further improvement, we also added the complex conjuncts which need to be handled separately. It is observed from the experiments that the syllable based method provides the better quality speech for the input text in comparison with the method based on phoneme.

Keywords: text to speech; speech synthesis; phoneme; syllable; graphemes

Résumé

এই গবেষণাটি বাংলা লেখ্য ভাষাকে কথ্য ভাষায় রূপান্তর করে। এই কাজটি দুইভাবে করা যায়, একটি হলো ধ্বনির মাধ্যমে অন্যটি হলো শব্দাংশ এর মাধ্যমে। এই গবেষণাটিতে, আমাদের প্রস্তাবিত সিস্টেমটি কয়েকটি স্তর নিয়ে গঠিত। প্রথম পর্যায়ে প্রতিটি বাংলা ধ্বনির জন্য অডিও শব্দের রেকর্ড করা হয় এবং বাংলায় ২৫০০০০ শব্দাংশের জন্য তিন হাজার শব্দাংশ রেকর্ড করা হয় এবং তারপরে প্রতিটি ধ্বনির এবং শব্দাংশের জন্য উচ্চমানের শব্দ পাওয়ার জন্য নয়েস কমিয়ে আনা হয়। দ্বিতীয় পর্যায়ে শব্দাংশের দীর্ঘতম সম্ভাব্য মিল খুঁজে পাওয়ার চেষ্টা করা হয় যদি এটি ইনপুট টেক্সট এ পাওয়া যায় এবং যদি না পাওয়া যায়, তবে ধ্বনিযুক্ত শব্দগুলিকে অনুরূপ গ্রাফিম এর সাথে মিল রেখে অনুসন্ধান করা হয়। অধিক উন্নত সিস্টেম পাওয়ার জন্য, আমাদের যুক্তাক্ষর গুলোকে আলাদাভাবে চিন্তা করা দরকার। পরীক্ষাগুলি থেকে এটি পর্যবেক্ষণ করা হয়েছে যে শব্দাংশ ভিত্তিক পদ্ধতি ধ্বনি ভিত্তিক পদ্ধতির তুলনায় ভাল মানের ফলাফল প্রদান করে।

1. INTRODUCTION

Speech synthesis is the automatic production of human speech, where a computer system used for this purpose is called a speech synthesizer, which can be implemented in software or hardware products, but a text to speech (TTS) system converts normal language text into speech [1]. Mute people cannot talk, but they will be able to talk using the TTS system, where they will type their desired words or sentences, and the TTS system will convert them into human speech. On the other hand, blind people cannot see, but they can hear the sound. If a mute want to communicate with a blind, the blind cannot see mute's sign language, but he/she can hear the speech that has been produced by the TTS system. Therefore, a mute can converse with a blind using the TTS system [2, 3].

Several attempts [4-8] had been made to develop and stimulate the process of development of the Bangla TTS synthesis system. In [4], epoch synchronous non overlap add (ESNOLA) method based concatenative speech synthesis system for Bangla was developed by Shyamal Kumar Das Mandal, et. al, in which authors described a system for concatenative speech synthesis using ESNOLA technique. Again, S. K. D. Mandal, et. al [5] showed some practical applications of Bangla TTS system using Epoch Synchronous Non Overlap Add (ESNOLA) technique. On the other hand, some important aspects of Bengali Speech Synthesis System proposed by A. Bandyopadhyay [6] used phonemes to develop voice database and used Epoch Synchronous Overlap Add (ESOLA) technique to concatenate the phonemes. Besides, T. Sarkar, et. al [7] described about grapheme to phoneme conversion, optimal text

selection, automatic segmentation tools and shown their experiment results. Moreover, Firoj Alam, et al [8] described the development process of Bangla (widely used as Bengali) TTS using a speech synthesis tool named Festival. But very few literatures are found in Bangla spoken by Bangladeshi people.

In this paper, we have proposed a system that shows the design and implementation of Bangla Text to Speech (TTS) system from the very raw level without using any third party speech synthesis tool. Two proposed systems in this study based on phonemes and syllables comprises two stages, in which the first stage audio sounds are recorded for each of the Bangla phonemes and three thousand out of 250000 syllables in Bangla, and then noise is reduced to obtain high quality sounds for each phoneme and syllable; and the second stage searches for longest possible matching of the syllables if it is available in the input text, and if not, then searches for the phonemes to match with the corresponding graphemes. For further improvement, we also added the complex conjuncts which need to be handled separately.

2. SPEECH SYNTHESIS

Speech synthesis is the computer-generated simulation of human speech, which is used to translate written information into aural information where it is more convenient [9]. The generation of a sound waveform of human speech from a textual or phonetic description is called speech synthesis [10]. To generate speech output from a given text, first, the input text is analyzed deeply. Then grapheme to phoneme conversion is carried out using pronunciation and letter to sound rule. Same phoneme or syllable may have different pronunciations depending on the grammatical and pronunciation rules. So the

pronunciation of the phonemes and syllables are detected by analyzing those rules. After identifying the corresponding sounds of matched syllable and phonemes, they are concatenated and played to generate expected speech output. Fig. 1 shows the workflow of speech synthesis.

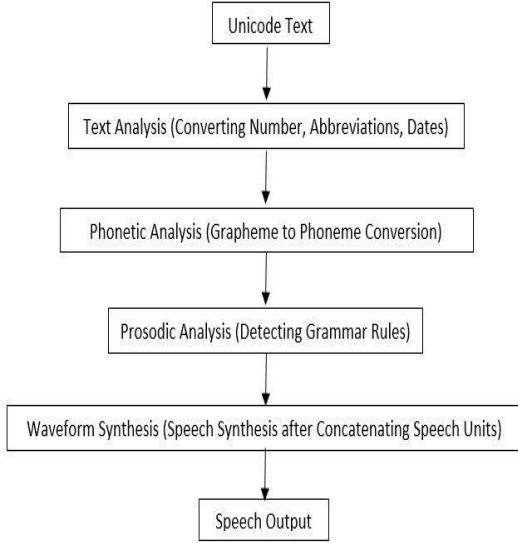


Fig. 1. Work flow of Speech Synthesis.

3. SPEECH SYNTHESIS METHODOLOGY

Different phases of our speech synthesis methodology are described below.

3.1 Text Analysis

In this level, the input text is analyzed deeply to convert it into pronounceable sounds. Bangla text contains the following alphabets and symbols. Here, IPA symbol for phoneme is also shown here

Bangla Vowel (বাংলা স্বরবর্ণ)

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
a	ā	i	ī	u	ū	ɳ	e	ai	o	au
[ɔ, o]	[ɑː]	[i, e]	[iː]	[u, o]	[uː]	[ɳ]	[e, æ]	[oi]	[o]	[ow]
ক	কা	কি	কী	কু	কূ	ক্	কে	কৈ	কো	কৌ
ka	kā	ki	kī	ku	kū	kʰ	ke	kai	ko	kau

Bangla Consonants (বাংলা ব্যঞ্জনবর্ণ):

ক	ka	[kɔ]	খ	kha	[kʰɔ]	গ	ga	[gɔ]	ঘ	gha	[gʰɔ]	ঙ	nga	[ŋɔ]
চ	ca	[tʃɔ]	ছ	cha	[tʃʰɔ]	জ	ja	[dʒɔ]	ঝ	jha	[dʒʰɔ]	ঞ	ña	[ɲɔ]
ট	ta	[tɔ]	ঠ	tha	[tʰɔ]	ড	da	[dɔ]	ঢ	dha	[dʰɔ]	ণ	na	[ɳɔ]
ত	ta	[tɔ]	থ	tha	[tʰɔ]	দ	da	[dɔ]	ধ	dha	[dʰɔ]	ন	na	[ɳɔ]
প	pa	[pɔ]	ফ	pha	[pʰɔ]	ব	ba	[bɔ]	ভ	bha	[bʰɔ]	ম	ma	[mɔ]
য	ya	[dʒɔ]	র	ra	[rɔ]	ল	la	[lɔ]						
শ	śa	[ʃɔ/ʂɔ]	ষ	ṣa	[ʃɔ]	স	sa	[sɔ/ʂɔ]	হ	ha	[ɦɔ]			
য়	ya	[ʃɔ]	ড়	ṛa	[rɔ]	ঢ়	ṛha	[rʰɔ]						

Modifier Symbols:

ক্	hasanta - mutes inherent vowel	ক্	k [k]
কৎ	khanda-ta - final unaspirated dental	কৎ	Kat [kɔt]
কং	anusvāra - final velar nasal	কং	kaṅ [kɔŋ]
কঃ	visarga - adds voiceless breath after vowel	কঃ	kaḥ [kɔh] / [kɔ]
কঁ	chandra-bindu - nasalises vowels	কঁ	kñ [kɔ̃]

Post-consonantal vowel signs:

- আ = া (akar) as in কা (should be after consonant)
- ই = ি (hrossikar/ikar) as in কি
- ঈ = ী (dirghikar/ikar) as in কী (should be after consonant)
- উ = ূ (hrossukar/ukar) as in কু (should be under consonant)
- ঊ = ু (dirghukar/ukar) as in কূ (should be under consonant)
- ঋ = ্ (rikar) as in ক্ (should be under consonant)
- এ = ে (ekar) as in কে
- ঐ = ৈ (oikar) as in কৈ
- ও = ো (okar) as in কো
- ঔ = ৌ (oukar) as in কৌ

First, total number of characters in input text is counted. From the beginning of the input text character, longest possible match for the synthesized sound unit is searched. Words are differentiated with space so the search continues to find longest sub-word match until a space is found. If search pointer reaches to a space, new word begins. The search continues until the end of the input text.

To illustrate the analyzing process let us take a sample text and analyze it:

Sample Bangla text: "আমি ভাত খাই"

Here, total number of characters = 11

Total spaces = 2

Total pronounceable characters = 11 - 2 = 9

Total words = 2

আমি = আ + ম +

ভাত = ভ + া + ত

খাই = খ + া + ই

The Non Standard Words (NSW) need to be normalized. Non Standard Words include abbreviations, acronyms, currency, dates, numbers (year, time, ordinal, cardinal, floating point). These Non Standard Words needs to be converted to standard words or syllables.

Example of Non Standard Words (NSW):

Example of currency: ১০০০/- = এক হাজার টাকা

Example of date: ২২-৪-২০১৪ = বাইশ চার দুই হাজার চৌদ্দ

Example of number: ১ = এক, ২ = দুই, ৩ = তিন, ৪ = চার, ৫ = পাঁচ, ৬ = ছয়, ৭ = সাত, ৮ = আট, ৯ = নয়, ০ = শূন্য

Example of time: ১২:৪০ = বারোটা চল্লিশ

There are several conjuncts in Bangla language. Some of them are shown below:

Bangla Conjuncts (বাংলা যুক্তবর্ণ): [3.2]

ক্ক = ক + ক; Example- আক্কেল, টেক্কা

ক্ট = ক + ট; Example- ডক্টর (Comment: Basically used in English/foreign debt words)

ক্ট্র = ক + ট + র; Example- অক্ট্রয়

ক্ক্ত = ক + ত; Example- রক্ক্ত

3.2 Phonetic Analysis

The method of finding pronunciation of input text is analyzed in this level. In our TTS system, we have used the following techniques for phonetic analysis:

3.2.1. Phoneme based technique for phonetic analysis: At first, we used the phonemes of Bangla language in our system. We have recorded the phonemes and implemented in our system by grapheme to phoneme conversion to generate speech output, but the result was not satisfactory. Let us give an example of the technique:

The Bangla word "তোমার" will be pronounced as following, if the phonemes are used only:

তোমার = ত্ + ও + ম্ + আ + র্

which is: /t/ + /o/ + /m/ + /a/ + /r/

This pronunciation is not good enough to understand. So, we decided to think in another direction. Then we found that if we can use the syllables, the output will be more satisfactory. So, then we started working with the syllables.

3.2.2. Syllable based technique for phonetic analysis:

We started recording the syllables and implemented them in our system by searching longest possible match with the syllables from the beginning of the words. But, later on we observed that longest possible match is not the best option always. So, we started analyzing the syllables and implemented them after finding and sorting them in the order which suits the best combination for pronunciation. To illustrate the syllable analyzing process, let us take a sample text and analyze it to find the best combination of syllable for pronunciation :

Sample Bangla Text: তোমার - তোমাদের - তোমরা

Here, the word "তোমার" can be a good example for analyzing Bangla text. If we select "তোমা" as a syllable and "র" as another syllable as we have searched for the

longest possible match from the beginning of the input text first, the word "তোমার" will be pronounced as following:

তোমা + র্

This pronunciation is not good enough for "তোমার" but this "তোমা" will be good for pronunciation of "তোমাদের"

If we select "তোম" as a syllable and "ার" as another syllable, the word "তোমার" will be pronounced as:

তোম্ + আর্

This pronunciation is also not good for "তোমার" but this "তোম" will be good for pronunciation of "তোমরা" .

If we select "তো" as a syllable and "মার" as another syllable, the word "তোমার" will be pronounced as:

তো + মার্

This pronunciation is good for "তোমার" and this "তো" will be also good for "তোমাদের".

So, তো + মার is the best combination to make pronounceable "তোমার".

Thus, the analysis process is done for "তোমার - তোমাদের - তোমরা".

3) Syllable and Phoneme based technique for phonetic analysis: While working with the syllables, we have come to know that there are more than 200000 syllables in Bangla language. The more syllables we can use, the more the performance will increase. But, it was not possible to work with almost 250000 thousand syllables. So, we used the most common three thousand syllables and phonemes together according to our requirement.

First, the system will search for longest possible match of the syllables. If there is no such syllable found, then it will search for the phonemes to match with the corresponding graphemes. We observed that the performance was getting better.

Let us take the following example to illustrate the syllable + phoneme technique:

Suppose we have the syllables "তো", "মার", and "দের" in our voice database as we have seen in the previous technique that this is the best combination of syllables to pronounce "তোমার" and "তোমাদের". And suppose, we do not have the syllable "তোম্", but we have "তো" and "রা" in our database.

Yet the word "তোমরা" can be pronounced as the following as we have used the phonemes with the syllables:

তোমরা = তো + ম্ + রা। So, in this technique, we need all the phonemes and the basic common syllables of a language to develop its TTS system. By this technique, we can cover the whole language in our system with better pronunciation.

3.3 Prosodic Analysis

Same phoneme or syllable may have different pronunciations depending on its prior and post characters or even its position in the word. The pronunciations will be according to the grammatical and pronunciation rules of Bangla language. To apply those rules, prosodic analysis is required. This prosodic analysis process is discussed below using examples.

Let us take a look at the following simplest Bangla pronunciation rules to understand the prosodic analysis process:

Bangla Pronunciation Rule #1 (a):*

If a consonant letter appears in the beginning of a Bangla word and if the post character of this consonant is a consonant letter, the first consonant will be pronounced as¹¹:

Phoneme of the consonant + "অ"

Bangla Pronunciation Rule #1 (b):*

And if a consonant letter appears in the beginning of a word and if the post character of this consonant is a vowel, then the consonant will be pronounced as:

Phoneme of the consonant + "ব"

For example, let us take the consonant ÓeÓ.

The word "বক" has got first character "ব" and its post character is "ক" which is a consonant. So, according to Bangla pronunciation rule #1 (a), the word "বক" will be pronounced as:

বক = ব্ + অ + ক্

In the word "বই", the first character is also "ব" but its post character is "ই" which is a vowel. So, according to Bangla pronunciation rule #1 (b), the word "বই" will be pronounced as:

বই = ব্ + ও + ই

Similarly, in the word "বল", the first character is "ব" and its post character is "ল" which is a consonant. So, according to Bangla pronunciation rule #1 (a), the word "বল" will be pronounced as: বল = ব্ + অ + ল্

And in the word "বউ", the first character is also "ব" but its post character is "উ" which is a vowel. So, according to Bangla pronunciation rule #1 (b), the word "বউ" will be pronounced as:

বউ = ব্ + ও + উ

In the same way,

মগ = ম্ + অ + গ্ but, মই = ম্ + ও + ই

কম = ক্ + অ + ম্ but, কই = ক্ + ও + ই

3.4 Waveform Synthesis

The waveform is synthesized step by step. The steps are given in Fig. 2.

1) *Record Sound*: The sound is first recorded for corresponding phonemes and syllables. Fig. 3 shows

recorded sound wave with noise. The recording should be in a sound proof place, otherwise there will be so much noise and interference in the recording, which cannot be reduced.

2) *Convert Sound to Wave Signal*: After recording the sounds, they are converted to wave signals. The wave signal of a recorded sound is shown below.

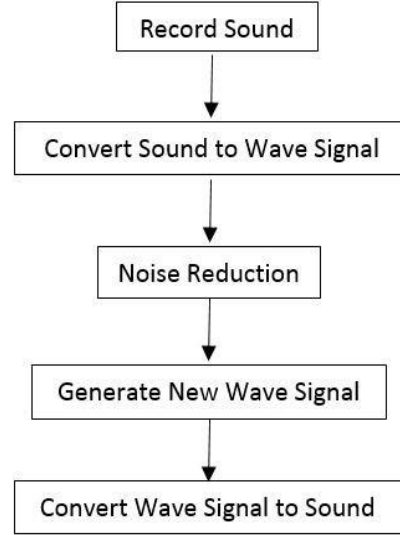


Fig. 2. Waveform Synthesis

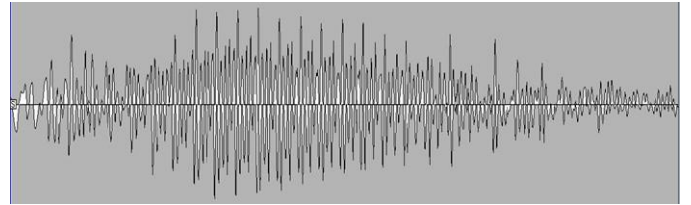


Fig. 3. Wave Signal of a Recorded Sound (with noise)

3) *Noise Reduction*: The noise in the wave signal is cut down to reduce the noise of the sound. The noise of the above wave signal is cut down to reduce the noise of the sound.

4) *Wave signal by reducing the noise*: After reducing the noise from the wave signal, new wave signal is generated and is shown in Fig. 4. The new noise free wave signal of the above wave signal is shown below.

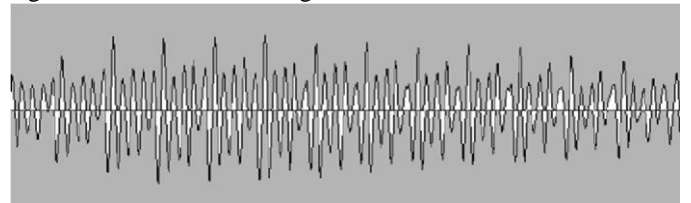


Fig. 4. Wave Signal after Noise Reduction

5) *Convert Wave Signal to Sound*: After generating the new noise free wave signal, the signal is converted to sound. This sound is the required noise free sound which can be used to generate speech output of a given text.

Concatenate Sound Units to Generate Output: The noise free sound units are concatenated by the application according to the given input text to generate the desired speech output.

4. EXPERIMENTS

4.1 Corpus

We have prepared the voice database by renaming the audio files by their corresponding syllable or phoneme names by which we can search easily to find and concatenate them to generate the desired speech output. O

4.2 Experimental Setup

The application has two methods to take input text. User can write on the given textbox or can open a text file to be read. Texts in the above picture are inputted by opening a text file tested for experiment. The output was good enough to be understood.

5. EXPERIMENTAL RESULT ANALYSIS

As we worked with the phonemes first, then the syllables, we got different results for phonemes and syllables. The result of the experiments with phonemes and syllables are shown below with a sample text. Suppose, the database has the following phonemes:

ত্, ও, ম্, আ, র্, দ্, এ

and the following syllables:

তো, মার, মা, দে, তোম্

So, with these phonemes and syllables the result with the sample text will be shown in Table 1.

As we told earlier that Bangla language has huge number of syllables, so we could not include them all in our system, but we found a way- that is use all phonemes and most common syllables together. Now, suppose our database does not have the syllable "তোম্", but has all the phonemes. So the result will be shown in Table 2. So the result of the different phonetic analysis technique can be compared and shown in Table 3.

TABLE I. EXPERIMENTAL RESULTS FOR PHONEMES AND SYLLABLES

Sample text: "তোমার - তোমাদের - তোমরা"			
Pronunciation	তোমার	তোমাদের	তোমরা
Using Syllables	তো + মার	তো + মা + দে	তোম্ + রা
Using Syllables + Phonemes	তো + মার	তো + মা + দে	তো + ম্ + রা

TABLE II. EXPERIMENTAL RESULTS FOR PHONEME + SYLLABLES

Sample text: "তোমার - তোমাদের - তোমরা"			
Pronunciation	তোমার	তোমাদের	তোমরা
Using Phonemes	ত্ + ম্ + ও + র্ + আ +	ত্ + ম্ + ও + আ + দ্ + এ + র্	ত্ + ম্ + ও + র্ + আ
Using Syllables	তো + মার	তো + মা + দে	তোম্ + রা

TABLE III. COMPARISON OF DIFFERENT TECHNIQUES USING PHONETIC ANALYSIS

Using	Pronunciation	Coverage of input text
Phonemes only	Cannot be understood properly	Covers All input text
Syllables only	Good enough to be understood	Do not cover all input text
Syllables + Phonemes	Good enough to be understood	Covers All input text

6. CONCLUSION

This paper has showed a technique for Bangla text to speech and concludes the following:

- This research was done from the very raw level, starting from using our own voice recordings to create phonemes and syllables.
- More than 3000 syllables and phonemes were used during the development process.
- Syllable based method showed high quality speech than the phoneme based method

In near future the author would like to do synthesis by covering the whole Bangla grammar and doing text normalization for larger context. The author would also work on conjuncts and more syllables in future. Besides, the experiments for the existing system to compare with our proposed method are not presented here. The author would like to these experiments for the future study.

REFERENCE

- Speech Synthesis
Website: http://en.wikipedia.org/wiki/Speech_synthesis
- Crowdsourcing helps Bangladesh's blind pupils
Website: <http://www.dw.de/crowdsourcing-helps-bangladesh-blindpupils/a-17744891>
- Bangla text to Speech using Festival by Firoj Alam
Website: https://www.academia.edu/2955759/Bangla_Text_to_Speech_using_Festival
- Shyamal Kumar Das Mandal and Asoke Kumar Datta, "Epoch Synchronous Non Overlap Add (ESNOLA) Method based Concatenative Speech Synthesis System for Bangla," Centre for development of Advanced Computing (C-DAC), Kolkata, India
- S.K.D. Mandal and B. Pal, "Bengali Text to Speech Synthesis System: A Novel Approach for Crossing Literacy Barrier," CSI-YITPA(E), 2002
- A. Bandyopadhyay, "Some Important Aspects of Bengali Speech Synthesis System," IEMCT, 2002.
- T. Sarkar, V. Keri, M. Santhosh and K. Prahallad, "Building Bengali Voice Using Festvox," CLSI 2005
- The Festival Speech Synthesis System
Website: <http://www.cstr.ed.ac.uk/projects/festival/>
- Speech Synthesis
Website: <http://whatis.techtarget.com/definition/speech-synthesis>
- Speech Synthesis
<http://dictionary.reference.com/browse/speech+synthesis>