A Novel Training Based Concatenative Bangla Speech Synthesizer Model

Firoz Mahmud, MD. Abdullah-al-MAMUN, Mumu Aktar, Shyla Afroge
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh.
fmahmud.ruet@gmail.com

***Abstract*— In the modern era of information technology, information is carried out in various ways to lead human life easily. Information can be exchanged among people in various ways and speech is the primary communication process among human beings. A TTS (Text-to-Speech) is used to convert input text to speech, and it's very popular application for computer users. Although different types of speech synthesis technologies are available for the English, France, Chinese and so many other languages, but in Bengali language, it's so scarce. This paper represents the implemented process of training based Concatenative Bangle Speech Synthesizer System and its performance. The synthetic utterances are built by concatenating different speech units selected from recorded database from the training session for concatenative speech synthesizer system. Here training based means any person can train his/her voice and that will be stored on database and next time that person will input a text to convert speech and listen according to his/her trained voice. So this process is known as independent voice. And to train the voice a set of Bengali keyword is stored on the database as segmented audio file. At last the performance of this Bangla speech synthesizer system implemented by the concatenative speech synthesizer technology is analyzed which has provided 85% accuracy to listener to identify the sentence.**

***Keywords*— TTS; Training Based Synthesizer; Bangla Speech Synthesizer; Bangla keyword set; Concatentive Synthesis.**

I. INTRODUCTION

The most powerful and common method of human communication is the oral mode. In our daily life we communicate each other via speech. Now-a-days computer is the most vital part of our life. So it is natural for people to expect to be able to carry out spoken dialogue with computers. This involves the integration of speech technology and language technology. A text to speech synthesizer is now an important part of information technology because it has integrated language and speech for human-computer interaction. Creation of synthetic voice from text is usually referred to as the general term 'text-to-speech' though it requires a wide range and variety of procedures. Speech synthesis is the artificial reproduction of natural speech. Spoken texts are generated by a computer. Rather than being played from a previously recorded body of texts, each sentence is individually generated [2].Speech synthesis is also known as Text-to-Speech (TTS). A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic into speech [3]. Speech technology can differ in size from the stored voice. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output [4]. The block diagram of TTS system is given in the figure 1. Various TTS (Text-to-Speech) works have already been done for different language like English, Arabic, France, Turkey, and German and so on [1, 2, 7, 8, 9, and 11]. Although different TTS systems have been introduced in different languages efficiently but in Bangla, TTS system is not so rich.

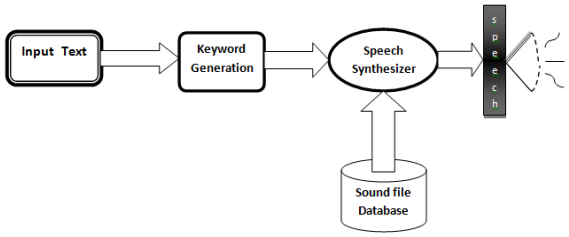

Figure 1: A general speech synthesizer model

A large database is needed to build a TTS system. So it's very difficult to build a TTS system. In past an attempt was made by C-DAC, Kolkata which had developed a complete Bangla TTS system named Bangla Vaani [10]. Very recently, CRBLP of BRAC University has released another Bangla TTS, Katha [12], which is built under the Festival framework using unit selection. A complete system has been shown here [12].

In this paper, Bangla Speech Synthesizer System is implemented by concatenative synthesizer technology. Here a system is represented which can convert input text to speech. The text contains character which needs to convert as normal text. In concatenative speech synthesis, a set of recorded speech units are selected from a database and are concatenated to create synthetic utterances [8]. This database contains prerecorded voice according to the keyword.

II. CONCATENATIVE SYNTHESIZER OVERVIEW

Several types of synthesizer technologies exist on Text-to-Speech system like concatenative synthesis, formant synthesis, HMM (Hidden Markov Model) based synthesis and sine wave synthesis [1], etc. The formant synthesis uses fundamental frequency, voicing, noise levels instead of human speech samples to create a synthetic waveform of speech and the concatenative synthesis uses segments of recorded human speech [2]. HMM-based synthesis is a synthesis method based

on hidden Markov models. In this system, the frequency spectrum, fundamental frequency and duration of speech are modeled simultaneously by HMMs. An overview of HMM based synthesis is shown in [13-14]. Sine wave synthesis is a technique for synthesizing speech by replacing the formants (main bands of energy) with pure tone whistles. Concatenative synthesis is based on the concatenation of segments of recorded speech. It uses large databases of recorded speech. During creation of database, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases which are known as keywords. Speech synthesized technology can be created by concatenating number of recorded voice (according to the keyword) that are stored in a database as audio file.

A. *Keyword*

A keyword is a basic unit of a word. It is a unit consisting of uninterrupted sound that can be used to make up words. So, a Keyword is a unit of organization for a sequence of speech sounds. In English it is known as 'syllable'. For example, the word "বাংলা" consist of two keywords, one is "বাং" and other is "লা".

B. *Classification of Bangla keyword*

According to the combination of number of letters (বর্ণ) Bangla keywords can be classified by the following way which is shown in figure.

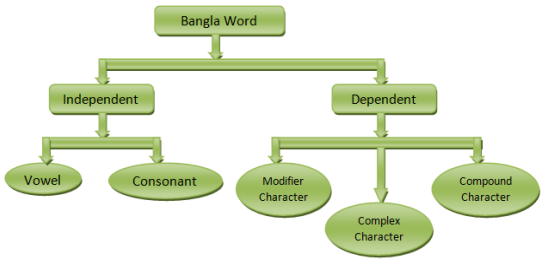

Figure 2: Classification of Bangla keyword

1) Independent keyword: If a keyword is constructed by only one letter then it is known as independent keyword. There are two kinds of independent keywords.

a) Vowel(স্বরবর্ণ): A speech sound that is produced by comparatively open configuration of the vocal tract, with vibration of the vocal cords but without audible friction and is a unit of the sound system of a language that forms the nucleus of a syllable. For example, অ, আ, ই, ঈ, উ, ঊ, ঋ, এ, ঐ, ও, ঔ।

b) Consonant(ব্যঞ্জনবর্ণ): A basic speech sound in which the breath is at least partly obstructed and which can be combined with a vowel to form a syllable. For example, ক, খ, গ, ঘ, ঙ, চ, ছ, জ, দ, ঝ, ঞ, and so on.

2) Dependent keyword: If a keyword is constructed by one or more consonant with combining kar(কার) (smallest term of vowel. i.e,া ি ু ৈ and like this) or fola(ফলা) (smallest term of consonant. i.e, ক্ক ; here ক is fola) then it is known as dependent keyword. Dependent keyword can be classified in the following way.

a) Modifier Character: A keyword that is constructed by one consonant with kar (কার)is known as modifier character. For example, কা, ঢে, তা, রু, নূ, লে, গো, ঐী and like this.

b) Compound Character: A keyword which is the combination of two or more consonants is known as compound character. For example, ক্ক, ঙ্ক, ত্ত, দ্ধ, ণ্ঠ, ঙ্ঘ, প্ত, গ্ম, হ্ম and link this.

c) Complex Character: If a keyword is the combination of both modifier character and compound character then it is called complex character. For example, ক্কা স্তা ঙ্কো ত্তা ণ্ঠো ঙ্ঘা and so on.

C. *Bangla keyword set*

There are many keywords in Bangla language. From the anlaysis of Bangla literacy, we have found about 1200 keywords. To detect the keywords we have used following four Bangla literacy books: রিক্তের বেদন(Riktar Badon) Written by কাজী নজরুল ইসলাম(Kazi Nazrul Islam); দুগেশনন্দিনী (Durgashnandini)written by বঙ্কিমচন্দ্র চট্টোপাধ্যায় (Bumkimchandro Chittopadhai); শেষ প্রশ্ন (Shas Prasno) written by শরৎচন্দ্র চট্টোপাধ্যায় (Shratchandro Chittopadhai); মেঘনাবদ কাব্য (Magnabod Kabbo)written by মাইকেল মধুসূদন দত্ত (Maikal Modhosudon Dotto).
Following section shows the sample keywords set for Bangla language.

1) For Vowel set: There are 11 charcters in vowel set. They are অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

2) For Consonant set: There are 35 charcters in consonant set. They are ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ভ ম য র ল শ ষ স হ ড় ঢ় য়

3) For Numeric set: There are 10 charcters in numeric set. They are ০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯.

4) For Modifier Character set: In our Bangla language there are huge collection of modifier characters. Collection of all modifier characters is called modifier character set. Modifier character set is shown in the following table (Table-1).

Table 1: Modifier Character set sample

For letter	Possible character set
ক	কা কি কী কু কূ কে কৈ কো কৌ ক্র ক্য কৎ কং
খ	খা খি খী খু খূ খে খৈ খো খৌ খ্র খ্য খং খৎ
গ	গা গি গী গু গূ গৃ গে গৈ গো গৌ গ্র গ্য গৎ গং
ঘ	ঘা ঘি ঘী ঘু ঘূ ঘৃ ঘে ঘৈ ঘো ঘৌ ঘ্র

	ঘ্য ঘৎ ঘং
ঙ	None
চ	চা চি চী চু চূ চে চৈ চো চৌ চৎ চং চ্য
ছ	ছা ছি ছী ছু ছে ছৈ ছো ছৌ ছৎ ছং ছ্য
জ	জা জি জী জু জূ জে জৈ জো জৌ জৎ জং জ্র জ্য
ঝ	ঝা ঝি ঝী ঝু ঝে ঝৈ ঝো ঝৌ ঝং ঝ্য
ঞ	ঞা ঞি ঞী
	. . .
য	যা যি যী যু যূ যে যৈ যো যৌ যৎ যং য্র য্য
র়	রা রি রী রু রূ র্ রে রৈ রো রৌ রং রৎ র্র
ল	লা লি লী লু লূ লে লৈ লো লৌ লৎ লং ল্য
শ	শা শি শী শু শূ শৃ শে শৈ শো শৌ শৎ শং শ্র শ্য
ষ	ষা ষি ষী ষু ষূ ষে ষৈ ষো ষৌ ষ্য
স	সা সি সী সু সূ সৃ সে সৈ সো সৌ সৎ সং স্র স্য
হ	হা হি হী হু হূ হে হৈ হো হৌ হৎ হং হ্র হ্য
ড়	ড়া ড়ি ড়ী ড়ু ড়ূ ড়ে ড়ো ড়ৈ
ঢ়	ঢ়া ঢ়ি ঢ়ী ঢ়ু ঢ়ে ঢ়ৈ ঢ়ো
য়	য়া য়ি য়ী য়ু য়ূ য়ে য়ো য়ৈ য়ৌ

5) For Compound Character set: Like modifier Characters there are huge collection of compound characters in our Bangla language. Collection of all compound characters is called compound character set. Compound character set is shown in the following table (Table 2).

Table 2: Compound character set sample

For letter	Possible character set
ক	ক্ক ক্ট ক্ত ক্ল ক্ষ ক্ষ্ম ক্র ক্ল ক্ষ ক্ষ ক্ষ
খ	খ্য স্খ ল্খ
গ	গ্ধ গ্ব গ্ম
ঘ	ঘ্ধ ল্ঘ
ঙ	ঙ্ক ঙ্গ ঙ্ঘ
চ	চ্চ চ্ছ চ্ছ্ব
ছ	চ্ছ্ব ঞ্ছ শ্ছ জ্ছ
জ	ঞ্জ ল্জ জ্ঞ জ্জ স্জ
ঝ	ঝ্ঝ ঞ্ঝ ঝ্ব
ঞ	ঞ্চ ঞ্জ ঞ্ছ ঞ্ঝ
	. . .
ম	ম্ম ম্ল ম্ক ম্গ ম্ভ ম্প ম্ভ ম্ল
য	None
র়	None
ল	ল্ম ল্স ল্ক ল্খ ল্ল ল্ঘ ল্ঙ ল্চ ল্জ ল্ঝ ল্ট ল্ঠ ল্ড ল্ঢ ল্ণ ল্ত ল্দ ল্ধ ল্ন ল্প ল্ব ল্শ ল্হ
শ	শ্ক শ্খ শ্গ শ্ঘ শ্ঙ শ্ঝ শ্ট শ্ঠ শ্ড শ্ত শ্দ শ্ধ শ্ন শ্প শ্ব শ্ম শ্ল শ্শ শ্স শ্হ
ষ	ষ্ট ষ্ঠ ষ্ণ ষ্ম ষ্ত ষ্চ
স	স্ক স্খ স্গ স্ঙ স্জ স্ঝ স্ঞ স্ট স্ঠ স্ড স্ঢ স্ণ স্ত স্দ স্ধ স্ন স্ব স্ম স্ল স্শ স্ষ স্হ
হ	হ্ণ হ্ন হ্ম হ্ল হ্ল
ড়	None
ঢ়	None
য়	None

6) For punctuation character set: Punctuation marks are symbols that indicate the structure and organization of written language, as well as intonation and pauses to be observed when reading aloud. Every symbol has a predefined meaning. If we use punctuation character in a wrong place then the meaning of that sentence may significantly be changed. Collection of all punctuation characaters is called punctuation characater set. Punctuation characater set is shown by the following table (Table 3).

Table 3: Punctuation character set sample

Punctuation	Means	Description
।	Full-stop	To stop the sentence
।।	Twice-stop	To repeat the sentence twice
,	Coma	To stop a few time within the sentence
;	Semicolon	To stop few time within the sentence
:-	Colon-Dash	To show the example for a topic
?	Question Mark	To question of a sentence
!	Exclamation mark	To exclamation of a sentence
-	Hipen	To combine different between two sentences

/	Forward Slash	Means or operation between two words
“	Quotation start	To start the quotation
”	Quotation end	To end the quotation
(	First bracket start	To start the first bracket
)	First bracket close	To close the first bracket

D. Training Keyword modeling

By modeling a sequence of sentences, keywords can be separated. If the word of a sentence is consist of only two keywords that can be easy to separate. As a result, here each word of a sentence is consisting of only two keywords. For example, “আমি ভাত খাই”| this sentence has three words: আমি, ভাত and খাই| from these words, easily keywords can be detected. আমি word consists of two keywords আ & মি | respectively, ভাত and খাই words consist of ভা, ত, খা, ই |
So to modeling the training keywords, we can create a passage that each word is consisting of only two keywords. And then separate each keyword from the word of the passage.

E. Recording voice

Any person who wants to create his/her voice to convert Text-to-Speech (TTS) first needs to login then reads the passage (here, passage contains all bangle keywords set) and corresponding audio file will be stored on the database. Special notification is that each word for the passage is shown one by one, so starting and ending position of each audio file for a word can easily be detected.

F. Segmentation of recording voice

From each word of corresponding audio file keyword can be separated by the signal ratio analysis. For example, the word “আমি” has two keywords, আ and মি | Here from the audio file of “আমি”, its audio signal ratio of আ: মি=2:1. That means if the total audio file length of আমি is 1000ms then first 666.67ms audio signal is keyword for “আ” and next 333.33ms audio signal is keyword for “মি” | An audio file for each keyword from the segmenting portion for each word has been created. Figure 3 is shows the original speech signal “আমি” and figure 4 is shows segmentation of the original signal.

G. Creating database

Database has been created by collecting all keywords of audio file from segmented recording voice. Database is a collection of audio files for all of the Bangla keywords set. That database is used for retracting keyword audio file corresponding to the input string of keyword. For .NET framework resource folder or any database server (i.e. SQL server) can be used to store keyword audio file or database.

III. Steps for Synthesizer Model

A text-to-speech system (or "engine") is composed of two parts [5] a front-end and a back-end. The front-end converts inputted text into the equivalent keyword set. This process is known as text normalization, pre-processing, or tokenization. The front-end then assigns phonetic transcriptions to each word and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The progression of conveying phonetic transcriptions to words is called text-to-phoneme. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end often referred to as the synthesizer then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations) [6] which are then imposed on the output speech.

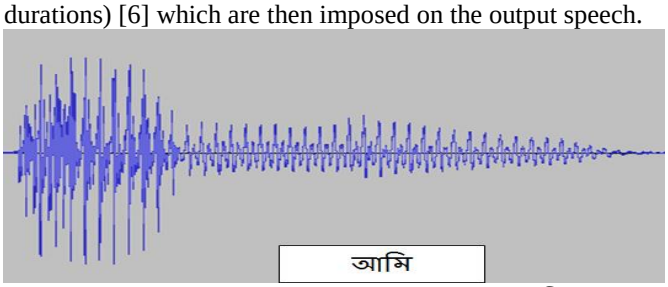

Figure 3: Original speech signal for “আমি”

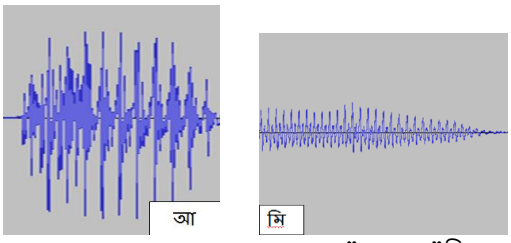

Figure 4: Splitting two keywords “আ” and ”মি” from original signal “আমি”

A. Text as input

A Bangla text is inputted to convert the Speech. For example, Input “আমাদের দেশের নাম বাংলাদেশে ।এই দেশের আয়তন ১৪৭৫৭০ বগকিলোমিটার। প্রাকৃতিক লীলানিকেতনের এই দেশের দিকে দৃষ্টিপাত করলে দেখা যায় সবুজ মাঠ, ফুলফলময় বৃক্ষ, তৃণগুল্ম শোভিত অরণ্য আর শস্য-শ্যামল ক্ষেত্রের মনোরম শোভা ।” as a text.

B. Text normalization rule

From the inputted text the string is normalized according to the punctuator and thus at first text is divided into three parts according to the fullstop (দাড়ি). First part is “আমাদের দেশের নাম বাংলাদেশে”, second part is “এই দেশের আয়তন ১৪৭৫৭০ বগকিলোমিটার” and third part is “প্রাকৃতিক লীলানিকেতনের এই দেশের দিকে দৃষ্টিপাত করলে দেখা যায় সবুজ মাঠ, ফুলফলময় বৃক্ষ, তৃণগুল্ম শোভিত অরণ্য আর শস্য-শ্যামল ক্ষেত্রের মনোরম শোভা”. The first and second parts contain no punctuator further within them but the third part contains some punctuators (i.e, কমা(,) and হাইপেন(-)). So third part will be divided according to the

coma (কমা(,)) and hipen(হাইপেন(-)) which contains another four parts. First part is: প্রাকৃতিক লীলানিকেতনের এই দেশের দিকে দৃষ্টিপাত করলে দেখা যায় সবুজ মাঠ, Second part is: ফুলফলময় বৃক্ষ, Third part is: তৃণগুল্ম শোভিত অরণ্য আর শস্য, and last part is:শ্যামল ক্ষেত্রের মনোরম শোভা. Total text normalization process is shown in figure 5.

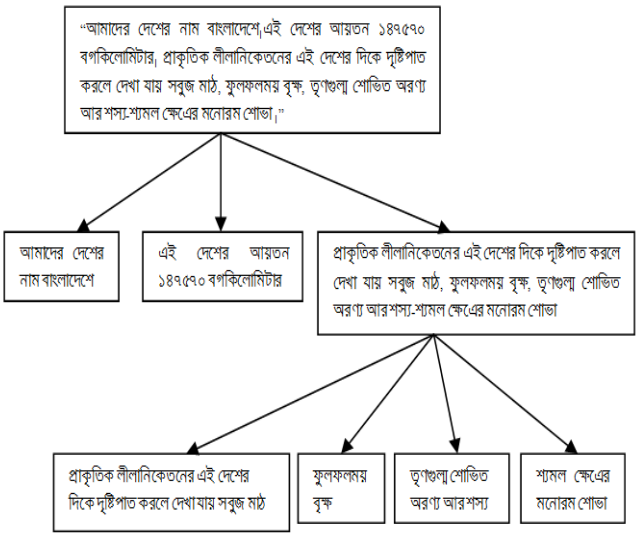

Figure 5: Input text normalization process.

Again to convert the number, each is placed according to its degree. Then it is converted from number to word. For example, ১৪৭৫৭০ will be converted to "এক লক্ষ সাতচল্লিশ হাজার পাঁচশত সত্তর"

C. Text-to-keyword rule

Speech synthesis technology uses text-to-keyword rule to determine the pronunciation of a word based on its spelling which is often called text-to-phoneme or grapheme-to-phoneme conversion (phoneme is the term used by linguists to describe distinctive sounds in a language). So normalized text is split to generate a set of keywords. For example, a normalized text is "আমাদের দেশের নাম বাংলাদেশ". So the keyword set is generated from the normalized text as following:

আ	মা	দে	র	[]	দে	শে	র
[]	না	ম	[]	বাং	লা	দে	শ

Here, [] represent the space.

D. Synthesizer

Synthesizer is a system of taking keyword as input and returning the output as speech. To return the speech, a keyword is taken as input and searched from the database. If keyword corresponding audio file is matched from the database then the keyword's audio file is retracted from database. And that file is played as output. This is a continuous process to where audio file is played one by one corresponding to the keywords sequence. This is seemed to be speech corresponding to the text.

IV. SYNTHESIZER COMPLEXITY ANALYSIS

Several types of complexities are associated with the concatentive speech synthesizer technology which are described in the following sections.

A. Ratio problem to segmenting recorded voice

For segmenting recorded voice, we used the ratio between two keywords. But in this process we cannot fully detect each keyword starting and ending point from the audio file. We considered the ratio 2:1 for segmenting a recorded word. The following table (Table 4) shows the file segmentation example for various keywords. Here, L represents total length of the audio file, L_A represents length of a keyword after segmentation and L_O represents length of a keyword before segmentation. Figure 6 shows error rate for different keywords before and after segmentation.

Table 4: Time variation for different keywords before and after segmentation

Word (শব্দ)	**Total audio file length L(ms)**	**2:1 ratio length for letter "আ" L_A(ms)**	**Original length for "আ" L_O (ms)**	**Error(%) $\frac{\lvert L_A - L_O \rvert}{L}$**
আমি	1120	746.67	810	5.65%
আজ	960	640	710	7.29%
আর	1050	700	670	2.85%
আম	940	626.67	690	6.73%

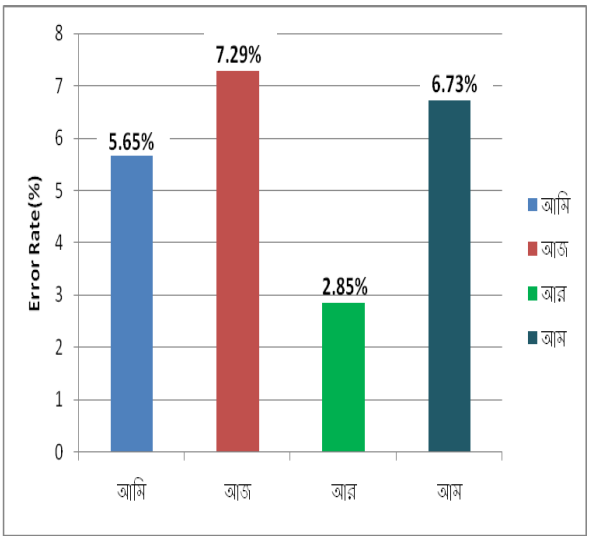

Figure 6: Error rate for different keywords before and after segmentation.

From the table (Table 4) it is easy to understand that a letter audio file length can be varied according to different keywords.

B. Speaker variation problem

If we use ratio only for segmenting keywords then segmentation process can't properly divide each keyword because the tone of a word can differ from person to person. As a result, the resultant output speech can be little bit unnatural. For example, the following table (Table 5) shows how audio file length varies for same word "আজ". Figure 7 shows error rate for different speakers for same word "আজ".

Table 5: Time variation for different speakers

Speaker	Audio file length for Speaker S (ms)	2:1 ratio length for letter "আ" S_A (ms)	Original length for letter "আ" S_O (ms)	Error(%) $\frac{\lvert S_A - S_O \rvert}{S}$
Speaker_1	1150	766.67	820	4.63%
Speaker_2	980	653.34	690	3.74%
Speaker_3	1070	713.34	780	6.22%
Speaker_4	1020	680	750	6.86%

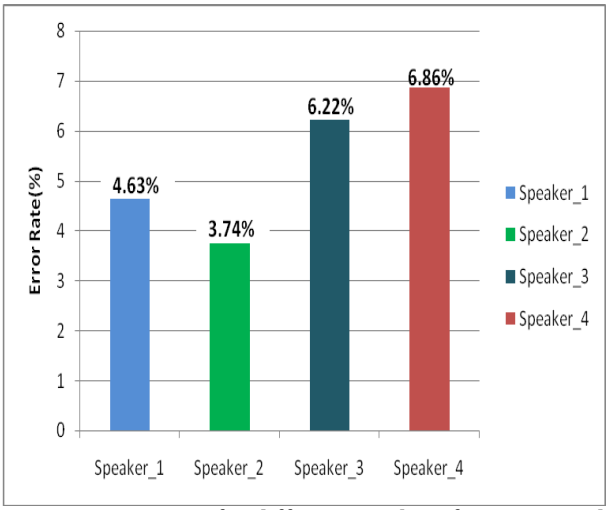

Figure 7: Error rate for different speakers for same word "আজ".

C. Confusing letter problem

Some types of Bangla keyword utterance can't be detected properly which are called the confusing letters. For example, some confusing letters are হসন্ত(্), চন্দ্রবিন্দু(ঁ), বিসর্গ(ঃ) ।For example, the word তারা and তাঁরা both utterance are same. So the keyword তা and তাঁ has no difference.

D. Number to Text Conversion problem

Number conversion is another problem in TTS systems. A TTS system often infers how to expand a number based on surrounding words, numbers and punctuation and sometimes the system provides a way to specify the context if it is ambiguous.

E. Universal Keyword Set

The consistent evaluation of speech synthesis systems may be difficult because of a lack of universally agreed objective evaluation criteria. Different organizations often use different speech data. The quality of speech synthesis systems also depends to a large degree on the quality of the production technique (which may involve analogue or digital recording) and on the facilities used to replay the speech. Evaluating speech synthesis systems has therefore often been compromised by differences between production techniques and replay facilities.

V. Performance

In spite of large improvements, Speech Synthesis can still sound a little unnatural. Since HMM-based TTS system can be applicable to pretense against speaker verification systems due to its having an ability to synthesize speech with arbitrarily given text and speaker's voice characteristics. The performance of this Bangla speech synthesizer system was measured by listening tests. Listening of single vowel, consonant and digits was fully identified. So in this case it provides 100% accuracy to identify the words correctly. But in another case, when we took a sentence as an input (text) to synthesis that time listener couldn't fully identify all of the words. It was found that the listeners identified about 85% of the words correctly from the text.

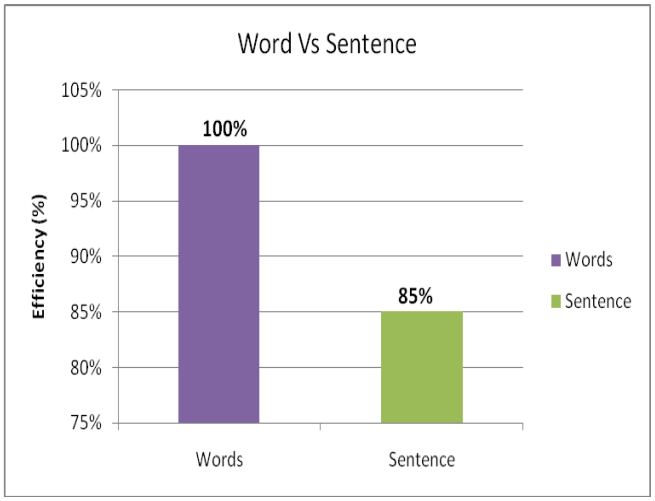

Figure 8: Accuracy comparison to identify words or sentences from input.

VI. Conclusion

Concatenation speech synthesis algorithm worked by the units recorded in different phonetic contexts has to reduce the audio waveform discontinuities and the phantom mismatches at the borders. The prosodic outline of the units is modified to the desired qualifications given by a prosody making block whose input is the text to be well-defined by the system. In this work, we developed a Bangla speech synthesizer system

for computers and present the implementation process and performance of this Bangla speech synthesizer system that is implemented by the concatenative speech synthesis architecture. Our goal is to develop a Bangla text-to-speech (TTS) application that can produce almost real-time speech efficiently from the input text for human-computer interaction. But concatenative speech synthesis technology is no longer providing the great accuracy for the text-to-speech model. On the other hand, Bangla speech training model has some lacks to develop the Bangla synthesizer model.

So, in future we will try to improve the accuracy about this Bangla speech synthesizer model or Text-to-Speech model that will be implemented by the concatenative speech synthesizer technology.

References

[1] Dutoit, "*An Introduction to Text-To-Speech Synthesis*," Kluwer Academic Publishers, 1997.

[2] Carvalho, P., Trancoso, I.M., and Oliveira, L.C.,"*Automatic Segment Alignment for ConcatenativeSpeech Synthesis in Portuguese*", Proc. of the 10thPortuguese Conference on Pattern Recognition,RECPAD'98, pages 221-226, Lisbon, March, 1998.

[3] Allen, Jonathan, Hunnicutt, M. Sharon, Klatt, and Dennis (1987), "*From Text to Speech: The MITalk system,"* Cambridge University Press. *ISBN 0-521-30641-8*.

[4] Rubin, P.; Baer, T.; Mermelstein, P. (1981), "*An articulatory synthesizer for perceptual research*," Journal of the Acoustical Society of America **70** (2): 321–328. DOI:10.1121/1.386780.

[5] Van Santen, Jan P. H. Sproat, Richard W. Olive, Joseph P. Hirschberg, and Julia (1997), "*Progress in Speech Synthesis*," Springer. ISBN 0-387-94701-9.

[6] Van Santen, J. (April 1994), "*Assignment of segmental duration in text-to-speech synthesis*," Computer Speech & Language **8** (2): 95–128. DOI:10.1006/csla.1994.1005.

[7] Bernd M Obius and Jan P. H. van santen, "*Modeling Segmentation Duration in German Text-to-Speech Synthesis,"* In International Conference on Spoken Language Processing (ICSLP), pages 2395{2399, Philadelphia, 1996.

[8] Daniel Erro, Asunción Moreno, and Antonio Bonafonte, "*Flexible Harmonic/Stochastic Speech Synthesis,"* Proc. 6th ISCA Speech Synthesis Workshop, 2007.

[9] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, *The AT&T next-gen TTS System*. Online::http://www.research.att.com/projects/tts, 26th, Jan, 2010.

[10] *C-DAC: "Research & Development-Speech Research*," Online:www.kolkatacdac.in/html/texttospeech.htm, 26th Jan, 2010.

[11] Marc Schr ¨ oder and J¨ urgen Trouvain, *The German Text-to-Speech Synthesis System MARY:A Tool for Research,"* Development and Teaching. Institute of Phonetics, University of the Saarland, Saarbr ¨ ucken, Germany.

[12] Firoj Alam, S.M. Murtoza Habib, and Mumit Khan, "*Text normalization system for Bangla*," Proc. of Conf. on Language and Technology, Lahore, pp. 22-24, 2009.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "*Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis*," Proc. of EUROSPEECH, vol.5, pp.2347–2350, 1999.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "*Speaker Interpolation in HMM-Based Speech Synthesis System*," Proc. of EUROSPEECH, vol.5, pp.2523–2526, 1997.