A Novel Bengali Text-to-Speech System to Achieve Both Intelligibility and Naturalness Using a Small Voice Database

Subhajyoti Barman[1], Uttam Kumar Roy[2]
Department of Information Technology
Jadavpur University (Saltlake Campus), Kolkata 700098, India
Email: [1]barman.subhajyoti@gmail.com, [2]royuttam@gmail.com

Abstract— **One of the key objectives for a Text-to-Speech (TTS) system is to achieve intelligibility and naturalness. However, designing such system using a small memory footprint becomes a challenging task.**

This paper proposes a concatenative Bengali TTS system using very small voice database. It mainly uses syllables as the building blocks of the concatenation processes. To limit database size, frequently-used context-sensitive syllables are identified and stored in a database. The similarity between target syllable and its closest matched syllable in the database is calculated. If it is higher than a certain threshold, matched syllable is used for concatenation; else the system switches to the diphone-based synthesis. This dual synthesizing model ensures that the system will never suffer from completeness issue. Since, maximum portion of the artificial speech comes from pre-recorded syllables, which contains higher amount of prosodic information, speech output sounds quite natural. The experimental results also prove its correctness.

Keywords— Text-To-Speech; Bengali; Sylabale; Unit selection, Diphone;

I. INTRODUCTION

Text-to-Speech synthesis is one of the intricate research topics in the field of **N**atural **L**anguage **P**rocessing (**NLP**) and **H**uman **C**omputer **I**nteraction (HCI). In this domain, knowledge of linguistics and **D**igital **S**ignal **P**rocessing (**DSP**) techniques are combined to add speaking capability to the computer system. By virtue of this technology, a computer system can assist us by generating speech instead of dull text.

Most of the current approaches have focused on synthesizing English speech. Although there are few language-independent frameworks for TTS conversion, but implementing a new language with all its linguistic features is not a trivial task. Moreover, most of the existing systems do not sound natural due to the lack of prosody. There are several methods for prosodic incorporation in TTS system, but those techniques increase the resource requirement of the system drastically. So design of a natural sounding TTS system with limited resources is still an open issue of this domain.

Current trends of research indicate that the concatenative speech synthesis is the most popular among researchers as the naturalness of its synthesized speech is more superior to other paradigms. In this approach, pre-recorded speech samples get concatenated to generate a synthetic speech signal for an arbitrary sentence. These voice units could be very basic voice units like phone, diphone or higher level voice unit like syllable or a whole word. It is observed that the presence prosodic information is more prominent in higher level voice units than lower level voice unit. It is clear that if concatenation is done with higher-level voice units, then the system generates more realistic speech. The key difficulty with higher-level voice unit is the large requirement of the memory to store those pre-recorded voice samples. Additionally, it is impossible to ensure that the database covers all possible units. In the absence of a particular voice unit, system will suffer in completeness or coverage issue. On the other hand, if a system deals with the low-level voice units, then its memory footprint becomes smaller but the synthesized speech sounds mechanical due to the lack of prosody.

This proposed system handles the tradeoff among naturalness, memory footprint and coverage issue by an optimal blend of high-level and low-level concatenative approach. It considers syllables as high-level voice units and diphones as the low-level voice units. To keep the database concise, a relatively small voice corpus is prepared which covers most of the frequently used syllable of Bengali language. **C**lassification **A**nd **R**egression **T**ree (CART) based voice database is formed for all unique syllables present it the corpus. Each CART stores different contextual variation of a particular syllable. To handle the completeness issuer, a separate diphone database is also designed for Bengali diphone inventory. During concatenation syllable based database is searched initially to identify the most appropriate match for a particular target syllable. Using a fitness function, this system calculates the contextual similarity index between the selected syllable and target syllable. A threshold is defined for the similarity index that determines whether the selected syllable is suitable for the concatenation or not. Whenever the system fails to fetch the proper contextual match for a target syllable, it switches to diphone based synthesis mode to generate the syllable. This technique helps to overcome the completeness issue. Moreover optimally selected speech corpus not only bound the size of its memory footprint, but also ensures that the most of the target syllables will get selected from the syllable database and that will help to ensure naturalness of the speech output.

This paper is organized as follows. Sections-II illustrates the different approaches for Bengali TTS systems. Section-III provides linguistic information about Bengali language. Section-IV demonstrates proposed model, Section-V analyzes the experimental results, and Section-VI concludes the paper.

II. Related Works

Text to speech conversion is a vast domain, several attempts were made to generate realistic artificial speech signals since 1939 [1]. This section highlights some of the significant works which has been already done in the field of Bengali TTS.

For an accurate development of any TTS system for a particular language, standardized linguistic information is very much needed. IPA stander for Bengali Language is can be found in [2]. Acoustic analysis on Bengali Vowels and Consonants is done by Firoj Alam et.al. [3][4] can be considered as a crucial work for Bengali linguistic.

Grapheme sequence to Phoneme sequence mapping is one of the key tasks for any TTS system. Shyamal at.el. proposed a rule based conversion for Bengali language [5]. To achieve a higher degree of accuracy Part-of-Speech (POS) tagging is done prior to apply rules on the grapheme sequence. Moreover to speed up the conversion and to manage exceptional words a dictionary was introduced along with the rules.

For Bengali POS tagging there exist several proposals using different approaches like Global Linear Model [6], HMM and Maximum Entropy [7], Support Vector Machine [8], Conditional Random Fields [9]. In contracts to rule-based POS tagging approach, in data-driven approach no morphological information is required. Krishnendu et.al. proposed a memory based G2P conversion technique [10]. This kind of approach is suitable if precise morphological information is unavailable for a particular language.

In existing Bengali TTS systems, concatenative approach based on Festival [11] architecture is predominating. In Festival prerecorded speech are concatenated either using a diphone model or using a unit selection technique. Diphone data base preparation procedure can be found in Muhammad Masud Rashid el. al.'s [12] work. Epoch Synchronous Non Overlap Add (ESNOLA) based Diphone can be found in [13], it describes the technique of prosodic inclusion in a diphone based synthesizer. More realistic sound can be obtained from a unit selection database as it attaches higher level of sound units together. N. P. Narendra et.al. proposed a syllable dependent concatenation technique for Bengali language [14][15]. For unit-selection strategy choice of optimum text from speech corpus is a vital task, this issue is addressed in [16]. This work uses a giddy algorithm to select an optimal set of phonetically balanced sentences from a large set of sentences.

In recent days Hidden Markov Model based statistical parametric models are also applied for Bengali TTS system [17][18]. Though it has many advantages like speaker adaptation, prosody incorporation, small memory requirement, etc., but it produces muffled speech output due to over-smoothing of the synthesized spectra.

III. Bengali Lingustic Information

Bengali is an Eastern Indo-Aryan language which was evolved from the Magadhi Prakrit and Pali during 1000–1200 AD and also influenced by Vedic Sanskrit. The Bengali writing system uses a script called as Bangla lipi, a derived from the Siddhaṃ script, which belongs to the Brahmic script family. The orthography of this script is from left to right, just like other western scripts. This script goes through several evolutions and currently it consists of 39 consonants grapheme and 11 grapheme symbols for vowel among them 10 appears in diacritic form.

Table 1: Vowels in Bengali with IPA

Grapheme Symbol	Grapheme Name	Diacritic form	Diacritic form Name	IPA transcription
অ	স্বর অ *sbôrô ô*	-	-	**/ɔ/**
আ	স্বর আ *sbôrô a*	া	আ কার *a kar*	**/a/**
ই	হ্রস্ব ই *hrôsbô i*	ি	হ্রস্ব ই কার *hrôsbô i kar*	**/i/**
ঈ	দীর্ঘ ই *dirghô i*	ী	দীর্ঘ ই কার dirghô i kar	**/i/**
উ	হ্রস্ব উ *hrôsbô u*	ু	হ্রস্ব উ কার *hrôsbô u kar*	**/u/**
ঊ	দীর্ঘ উ *dirghô u*	ূ	দীর্ঘ উ কার *dirghô u kar*	**/u/**
ঋ	হ্রস্ব ঋ *hrôsbô ri*	ৃ	হ্রস্ব ঋ কার *hrôsbô ri kar*	**/ṛi/**
এ	স্বর এ *sbôrô e*	ে	এ কার *e kar*	**/æ/**
ঐ	স্বর ঐ *sbôrô ôi*	ৈ	ঐ কার *ôi kar*	**/o/+/i/**
ও	স্বর ও *sbôrô u/o*	ো	ও কার *u/o kar*	**/o/**
ঔ	স্বর ঔ *sbôrô ôu*	ৌ	ঔ কার *ôu kar*	**/o/+/u/**

Table 2: Consonants in Bengali with IPA

	Stop								Nasal		Approximant		Fricative			
	Generic sounds															
Voicing	Voiceless				Voiced								Voiceless		Voiced	
Aspiration	Un-aspirated		Aspirated		Un-aspirated		Aspirated		Un-aspirated						Aspirated	
Vocal	ক	**kô /kɔ/**	খ	**khô /kʰɔ/**	গ	**gô /gɔ/**	ঘ	**ghô /gʰɔ/**	ঙ	**ngô /ŋɔ/**					হ	**hô /ɦɔ/**
Palatal	চ	**chô /tʃɔ/**	ছ	**chhô /tʃʰɔ/**	জ	**jô /dʒɔ/**	ঝ	**jhô /dʒʱɔ/**	ঞ	**ñô /nɔ/**	য	**zô /dzɔ/**	শ	**shô / ʃɔ/**		
Post-Dental	ট	**ṭô /ʈɔ/**	ঠ	**ṭhô /ʈʰɔ/**	ড	**ḍô /ɖɔ/**	ঢ	**ḍhô /ɖʰɔ/**	ণ	**ṇô /nɔ/**	ড়	**rô /rɔ/**	ষ	**shô / ʃɔ/**		
Dental	ত	**tô /t̪ɔ/**	থ	**thô /t̪ʰɔ/**	দ	**dô /d̪ɔ/**	ধ	**dhô /d̪ʰɔ/**	ন	**nô /nɔ/**	ল	**lô /lɔ/**	স	**sô /sɔ/**		
Labial	প	**pô /pɔ/**	ফ	**fô /ɸɔ/**	ব	**bô /bɔ/**	ভ	**bhô /bʱɔ/**	ম	**mô /mɔ/**						

Other Consonants	ড়	**ṛô /ɽɔ/**	ঢ়	**ṛhô /ɽʱɔ/**	য়	**yô /e̯ɔ/**	ৎ	**t /t̪/**	ংং	**ônusbar /ŋɔ/**	ঃ	**bisôrgô**	ঁ	**chôndrôbindu**

Apart from this, in Bengali script there are many compound graphemes due to consonant conjuncts. When more than one consonant appear consecutively without any separation of inherent vowel, are represented by a typographic ligature called a "consonant conjunct" (যুক্তাক্ষর juktakkhôr). For example word "স্ প স্ ট" is written as "স্পস্ট" (spastô). This script is a shallow orthographic script, i.e. phonetic information is embedded in the grapheme sequence. One-to-One Grapheme-to-Phoneme correspondence can be seen in most of the cases, though exceptions do happen in some exceptional cases. This language contains 41 phones among them 34 are for consonants only 7 for vowels. It is observed that for each vowel phones, there exists a nasal variation. All the grapheme symbols and their probable phones are shown in the following two tables, table I for vowels and table II for consonants.

IV. PROPOSED WORK

The system design of this Bengali TTS system is distributed in two parts as i) Text Analyzer and Syllable Identification Subsystem ii) Speech Concatenation Subsystem. The first module can be considered as the front-end of the system. Its main objective is preprocessing of the input text and generation of the syllable based phonetic representation of each word of the input text. The second one is the core unit of this system which responsible for the speech signal generation. This core module is comprised with two voice databases and a cost estimation function. The primary database is based on syllable specific unit-selection paradigm and the secondary database is for Bengali diphones. The cost estimating function is designated for the appropriate database selection. This function calculates the cost for marching a target syllable in the pre-recorded unit-selection database, if the calculated cost remains higher than a predefined threshold, then the matched syllable unit is selected for the concatenation, else the target syllable is synthesized using the diphone model. Cost of this matching is based on five contextual parameters as i) Position of the syllable in a word, ii) Position of the syllable in a phrase, iii) Type of the sentence to which syllable is associated, iv) Phonological context of the previous syllable, V) Phonological context of the next syllable. Internal block diagram of this system is illustrated in Figure 2 and implementation details of each sub-subsystem are described in following subsection.

A. *Text Normalization*

This is one of the basic tasks for any TTS system. Like other existing system Scheme language based regular expression ware used to identify Nonstandard Words (NOS) like acronyms, dates, currency, abbreviations, numbers etc. and converted to their proper utterance. This is essential as the pronunciation of NOS word is not directly depended on their grapheme sequence. For example, if the given text is "১৯৯৫ সালে পৃথিবীর গড় তাপমাত্রা ২°C বৃদ্ধি পাইয়াছে!" then it is required to expend the year and temperature to its actual pronounceable form, which is "উনিশো পঁচানব্বই সালে পৃথিবীর গড় তাপমাত্রা দুই ডিগ্রি সেলসিয়াস বৃদ্ধি পাইয়াছে!". Rules for this kind of expansion are well always language specific and they are well defined for Bengali language. Special treatment is required for abbreviation and acronyms, as it is impossible to resolve them using some rules, so a separate dictionary is used to store actual utterance corresponding to all abbreviations and acronyms supported by our system.

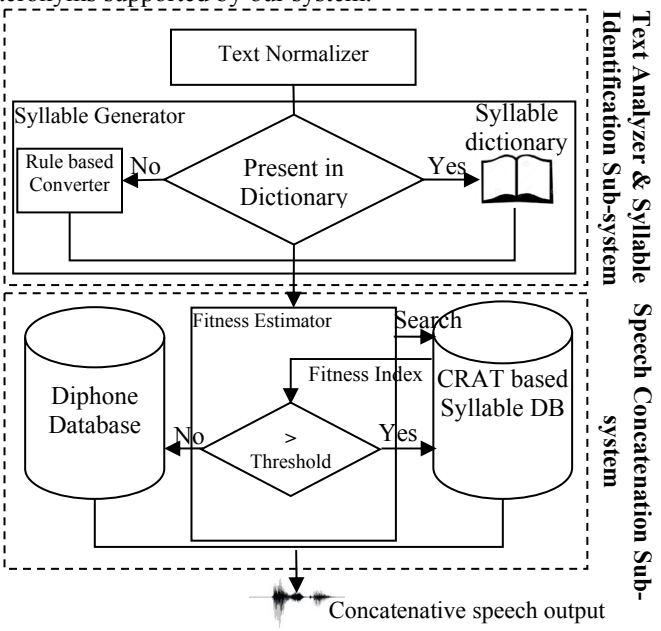

Figure 1: Internal block diagram of the System

B. *Syllablic Analysis*

This phase can be considered as an enhanced phonetic analysis. In phonetic analysis, an orthographic representation of a particular word is mapped to a string of phonetic symbols which represents the pronunciation of that word. This is also known as Grapheme-to-Phoneme (G2P) conversion. In Syllable Conversion, something more needs to be done than just a G2P mapping. In this phase after G2P conversion phone sequence are clustered depending on their syllable representation of each word. Primarily a Bengali phonetic dictionary is used for G2P conversion, for non-dictionary words Bengali linguistic rule based approach as described in [5] is used. After this conversion graphemes sequence is cluttered using syllable formation rules as described in [19]. For example, if input word is 'সময়' then after G2P mapping it will become a stream on phones as **/sɔ//o//mɔ//eɔ/**. As this word consists of two syllables so the outcome of this phase will be two clusters of phones as **/sɔ//o/-/mɔ//eɔ/**, hare syllables are separated by the '-' symbol.

C. *Voice Database Creation*

For a concatenative TTS synthesizer a voice database is nothing but a data structure which stores actual voice signal corresponding to a symbolic representation of that voice unit. It is already discussed that two separate voice database is used in this system along with their importance. This section is mainly focused on designing issues of Unit-selection database and Diphone database.

*a) **Unit-selection Database***

Preliminary phase for any Unit-selection database is text collection. Around 70,000 sentences are collected from e-newspapers, movies' subtitle file and Bengali novels, which are mainly written in the Bengali colloquial language.

Next, all of those sentences are converted to their equivalent syllable sequence. As pronunciation of every syllable varies with its context, so the extracted syllables are labeled with their context of occurrence. From this context labeled syllable sequence; it is possible to find out the total number of unique syllables present in the collected text. To achieve this goal, Algorithm 1 is proposed which calculates the total number of unique syllables present in the speech corpus based on the context of the syllables. If a Syllable appears multiple times in the corpus with varying context, all of them are treated as a unique syllable. Contextual parameters that are considered for syllables and their probable context are shown in Table 3.

Table 3: Contextual Parameters

Context	Value
Syllable position in a word	Initial, Medial, Final
Syllable position in a phrase	Initial, Medial, Final
Type of the phrase	Affirmative, Negative, Interrogative
Phonological Context of the previous syllable	Silence or other syllable
Phonological Context of the next Syllable	Silence or other syllable

Algorithm 1: Unique Syllable Finding

Input: C[]: Array of context labeled syllable representation of all Sentences
Output: Y: Set of <Syllable, Context> duplet
Start
  Set Y:=NULL
  For each **Sentence st** of **C**
    For each **Syllable sy** of **st**
      Context cx ←getContext(sy)
      If <sy,cx> not in **Y**
        Add **<sy,cx>** to **Y**
      End if
    End For
  End For
  Return (Y)
End

Next an optimal text selection is done on the corpus to minimize its size without violating syllable diversity. For this minimization a giddy algorithm is used which is described in Algorithm 2. This algorithm selects minimum number of sentences from the corpus that covers all unique syllables present in the corpus.

Algorithm 2: Optimal Text Selection (without violating syllable diversity)

Input: C[]: Array of context labeled syllable representation of all Sentences, **Y** set of Syllable, Context duplet (Output of Algorithm 1)
Output: Optimal selected text **O** (subset of C)
Start: Perform sorting on Y depending on frequency of the occurrence of each elements of Y in the corpus.
While Y not Empty
  Select the list frequent <Syllable, Context> duplet **sl**
  Add a sentence **st** which content maximum number of unique syllable-context duplet along with **sl,** to the set **O.**
  Remove **sl** from **Y** along with all other syllables sl' present in **st**
Wend
End

As an outcome of this phase 9803 sentences are obtained which contain 24424 words and 19743 syllables.

- In the next phase those optimally selected texts are recorded by a speaker. Sampling frequency of the recording was 16 KHz and stored at 16 bit PCM format.

- Recorded utterances are getting labeled and clustered in Festvox voice database builder of Festival architecture. For each phonetically seemlier syllable are grouped together in a CART tree, where each leaf level of the tree represents pronunciation of the syllable in a particular context. In the other hand intermediate nodes of this decision tree store branching information. A sample CART tree structure for our proposed scheme is shown in the following Figure 3.

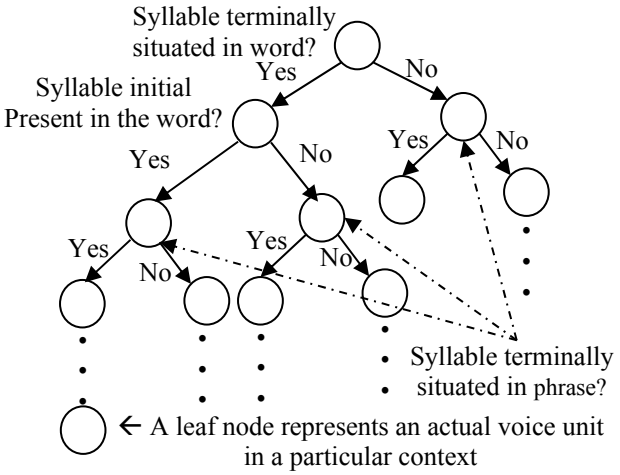

Figure 2: CART Decision Tree's Structure

*b) **Diphone Database***

- Diphone database is the secondary database of this system. In the absence of a particular target syllable in its proper context in unit-selection database, it gets synthesized using diphone database.

- This database contains all possible phones to phone transition for Bengali language. As there are 48 phones,

including all nasal variations of vowels, so mathematically 48x48 numbers of diphone possible.

- In practical few diphones never arise in Bengali (like **/ɔ/- /ɔ/,** /n̪/-/ŋ̪/ etc.), eliminating those impractical diphones a set of tetra-syllabic nonsense word is formed which covers all vowels and non-nasal consonants. Moreover to handle /n/ /ŋ/ two phones separately a set of octa-syllabic nonsense word is formed which covers all combinations of /n/, /ŋ/ in conjugate form with other consonants as disparity of pronunciation of these two phones only happen during conjugation with other consonants.

- All this nonsense words are also recorded and labeled as done in unit-selection technique. To get better output it is required to perform power-normalization and frequency normalization before performing the labeling process.

- To synthesize any arbitrary syllable it uses the ESNOLA algorithm as described in [13]. This algorithm not only concatenates diphones optimally, but also manipulates duration, intonation and power of the signal depending on the position of the diphone in the input phone sequence. As this database is designed to generate a syllable instated of an entire sentence or word, so a small modification is done in the existing ESNOLA algorithm. In the modified algorithm we are just passing previous and next syllable information along with the target syllable.

D. *Database Selection*

Selecting the most appropriate Database is the prime requirement of this system. For selecting the most suitable database for a particular target syllable, it uses a cost estimating function. This function identifies the best match for a particular target syllable from the Unit-selection database depending on contextual parameter matching. It uses the same contextual parameters which have been used during Unit-selection database creation. A weight is associated with each contextual parameter which is shown in Table 4. Cost of the matching is calculated by the cumulative sum of matched parameters' weight.

Table 4: Weight of Contextual Parameters

Context	Value	Weight
Syllable's position in a word	Initial, Medial, Final	0.5
Syllable's position in a phrase	Initial, Medial, Final	0.2
Type of the phrase	Affirmative, Negative, Interrogative	0.18
Phonologic Context of the previous syllable	Silence or other syllable	0.07
Phonologic Context of the next Syllable	Silence or other syllable	0.05

Let assume target syllable is SL^T with context $<c_1,c_2,c_3,c_4,c_5>$ and its best matched syllable is SL^M with context $<c'_1,c'_2,c'_3,c'_4,c'_5>$, so the cost of this matching can be calculated using following Formula 1.

$$C(S^{LT} \rightarrow S^{LM}) = \sum_{i=1}^{5} W(c'_i) * K \qquad \dots (1)$$

$$If\ c_i = c'_i \text{ then } K = 1, \text{else } K = 0$$

If the cost remains higher than a threshold value ***T*** then it selects the Unit-selection database, else it selects the diphone database to generate the syllable. The value of this threshold is chosen as 0.75 for this implementation. This value ensures that there exist at least three contextual matching between target syllable and selected syllable and one of the context is the 'Syllable's position in the word'. As a syllable with less similar contextual matching with target syllable may degrade the quality of the synthesized speech, so in the absence of a proper syllable in the Unit-selection database, the system generates the target syllable using diphone concatenation processes.

V. Evaluation of The Proposed Work

Evaluation of a speech synthesis is quite a tricky task, as it is difficult to predict the quality of the synthesized speech by visualizing its amplitude envelope or it is frequency spectrum. The quality of the speech is associated with too many psychoacoustic parameters and it is difficult to predict those parameters using computer automation. That is why to evaluate this proposed system we rely on human perseverance of the sound signal. A group of 30 peoples ware selected to evaluate this system from the age group of 18 to 40 years having equal gender ratio. All subjects are native speakers of Bengali language and have sufficient Bengali linguistic knowledge. For this testing each volunteer provides 20 different Bengali text input of their own to the system and listen the generated sound output. They rated this system on the basis of completeness, naturalness and correctness within a 0 to 5 rating scale. From this rating **M**ean **O**pinion **S**core (MOS) is calculated for these three criteria. Overall rating of this system is given in the following Bar-graph (Figure 4).

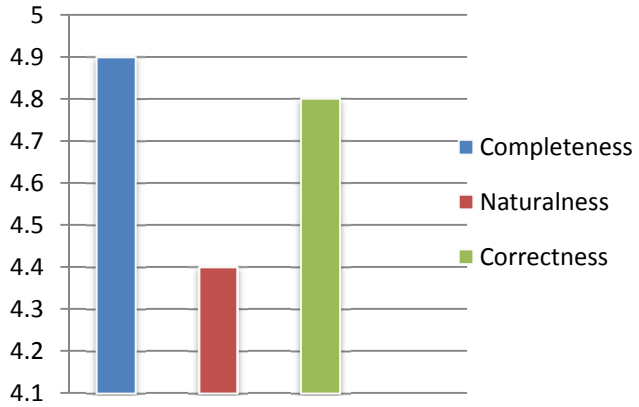

Figure 3: MOS for the Proposed System

VI. Conclusion

This proposed system combines the features of syllable based unit-selection technique and diphone based voice generation technique. It overcomes the limitation of one

technique by the strength of the other one. Due to this reason the overall performance of this system gets improved significantly as compared to the other existing systems and that is also clear from the MOS ratting of the end users. This result is too much encouraging for this research. In future, it can be further improvised by incorporating more sophisticated modules for batter prosodic modeling.

Acknowledgment

We sincerely acknowledge Department of Information Technology, Jadavpur University and Department of Science and Technology (DST), Govt. of India for providing the funding support and facilitate to carry out our experiment.

Reference

[1] H. Dudley, "TheVocoder," BellLab. Re.17. 122 (1939).

[2] Sameer ud Dowla Khan, “Bengali (Bangladeshi Standard)”, Journal of the International Phonetic Association / Volume 40 / Issue 02 / August 2010, pp 221 – 225

[3] Firoj Alam, S.M. Murtoza Habib, Mumit Khan, “Acoustic analysis of Bangla vowel inventory” Citeseer Jan 1, 2008 BRAC University CRBLP Technical report doi=10.1.1.173.651

[4] Firoj Alam, S.M. Murtoza Habib, Mumit Khan “Acoustic analysis of Bangla consonants” Center for Research on Bangla Language Jan 1, 2008 doi=10.1.1.173.1310

[5] J. Basu, T. Basu, M. Mitra and S. K. D. Mandal, "Grapheme to Phoneme (G2P) conversion for Bangla," Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on, Urumqi, 2009, pp. 66-71.

[6] S. Mukherjee and S. K. Das Mandal, "Bengali parts-of-speech tagging using Global Linear Model," 2013 Annual IEEE India Conference (INDICON), Mumbai, 2013, pp. 1-4.

[7] S. Dandapat, S. Sarkar, A. Basu, “Automatic part-of-speech tagging for bengali: an approach for morphologically rich languages in a poor scenario,” Proceedings of the Association for Computational Linguistic, pp. 221-224, 2007.

[8] A. Ekbal, S. Bandyopadhyay , “Part of speech tagging in bengali using support vector machine”, ICIT-08, IEEE International Conference on Information Technology, pp. 106-111, 2008.

[9] A. Ekbal, R. Haque and S. Bandyopadhyay “Bengali Part of Speech Tagging using Conditional Random Field”, In Proceedings of the 7th International Symposium on Natural Language Processing (SNLP-07), 131-136, Thailand, 2007.

[10] K. Ghosh and K. S. Rao, "Memory-based data-driven approach for grapheme-to-phoneme conversion in Bengali text-to-speech synthesis system," 2011 Annual IEEE India Conference, Hyderabad, 2011, pp.1-4.

[11] The Festival Speech Synthesis System
Website: http://www.cstr.ed.ac.uk/projects/festival/

[12] M. M. Rashid, M. A. Hussain and M. S. Rahman, "Diphone preparation for Bangla text to speech synthesis," Computers and Information Technology, 2009. ICCIT '09. 12th International Conference on, Dhaka, 2009, pp. 226-230.

[13] Shyamal Kumar Das Mandal, Asoke Kumar Datta “Epoch Synchronous Non Overlap Add (ESNOLA) Method based Concatenative Speech Synthesis System for Bangla”, 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24,2007.

[14] N. P. Narendra and K. Sreenivasa Rao “Syllable specific target cost formulation for syllable based text-to-speech synthesis in Bengali” International Conference on Computer & Communication Technology (ICCCT)-2011

[15] N.P. Narendra, K. Sreenivasa Rao, Krishnendu Ghosh, Ramu Reddy Vempada and Sudhamay Maity “Development of syllable-based text to speech synthesis system in Bengali” Int J Speech Technol (2011) 14:167–181

[16] Sandipan Mandal, Biswajit Das, Pabitra Mitra, Anupam Basu “Developing Bengali Speech Corpus for Phone Recognizer Using Optimum Text Selection Technique” 2011 International Conference on Asian Language Processing

[17] S. Mukherjee, S. K. Das Mandal, ”A Bengali HMM Based Speech Synthesis System” International Conference on Speech Database and Assessments (Oriental COCOSDA), pp.255-259, 9-12 Dec. 2012

[18] A. Pradhan, A. Prakash, S. Aswin Shanmugam, G. R. Kasthuri, R. Krishnan and H. A. Murthy, "Building speech synthesis systems for Indian languages," Communications (NCC), 2015 Twenty First National Conference on, Mumbai 2015, pp. 1-6 doi:10.1109/NCC.2015.7084931

[19] Mallik, B. P. (1960). Phonemic analysis of the consonant clusters in standard colloquial Bengali. Bulletin of the Philological Society of Calcuta 1(2), 37-46.