BUILDING BENGALI VOICE USING FESTVOX

ABSTRACT

This paper describes about work done in Building Bengali Text to Speech system. Also describes with experimental results on various methodology which we used for Building. The methodology takes into account the new optimal text selection algorithm was proposed to select appropriate text database, grapheme to phoneme converter that transliterates text into its phonetic equivalents and for automatic segmentation process a FESTVOX framework was used.

1. INTRODUCTON

Voice User Interfaces for IT applications and services have become more and more prevalent for languages like English, and are valued for ease of access, especially in telephony-based applications. In a country like India, where the majority of the population comfortable using English and given the relatively lower rates literacy, local language speech interfaces can provide access to IT applications and services, through internet and/or telephones, to the masses.

Given the increased availability of digital content in local languages, and the advent of Digital Library portal of India, appliances such as PCTVT for illiterate and common people, there is a real need and a set of real users asking for speech synthesis systems in all of the Indian languages.

Following our previous work in building Hindi, Telugu and Tamil voices, we have continued to add more Indian languages and in the process we have built a unit selection voice for Bengali. In this paper we describe the nature of Bengali scripts, letter to sound rules and the development of unit selection voice for Bengali. This work is done within the FESTVOX voice building framework, which offers general tools for building unit selection synthesizers in new languages. FESTVOX offers a language independent method for building synthetic voices, offering mechanisms to abstractly describe phonetic and syllabic structure in the language.

This paper is organized as follows: Section 2 describes the nature of Bengali scripts and Phone set. Section 3 discusses the system architecture. Section 4 discusses the grapheme to phoneme conversion. Section 5 describes about the Optimal Text Selection algorithm. Section 6 discusses the creation of TTS databases for Bengali. Section 7 discusses the Bengali speech data collection. Section 8 describes the speech segmentation and Section 9 discusses about the unit clustering and synthesis.

2. BENGALI ORTHOGRAPHY

2.1 Nature of Bengali Scripts:

The basic units of the writing system in Indian languages are characters which are an orthographic representation of speech sounds. A character in Indian language scripts is close to

a syllable and can be typically of the form: C*V C*, where C is a consonant and V is a vowel. There is fairly good correspondence between what is written and what is spoken. Bengali has a rich set of 665 graphemes to represent the different sounds in the language. And there are fewer characters than many of the other Indian languages.

FIgure.1. shows the vowels and consonants of Bengali used in our text to speech system.

2.2 Phone Set:

This Diagram shows the vowels and consonants of Bengali used in our text to speech system. There are 13 vowels and 52 consonants characters.

Vowels:

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ	ং	ঃ
a	**aa**	**i**	**ii**	**u**	**uu**	**r**	**e**	**ai**	**o**	**au**	**n'**	**:**

Consonants:

ক	ka	খ	kha	গ	ga	ঘ	gha	ঙ	ṅa
চ	ca	ছ	cha	জ	ja	ঝ	jha	ঞ	ña
ট	ṭa	ঠ	ṭha	ড	ḍa	ঢ	ḍha	ণ	ṇa
ত	ta	থ	tha	দ	da	ধ	dha	ন	na
প	pa	ফ	pha	ব	ba	ভ	bha	ম	ma
য	ÿa	র	ra	ল	la				
শ	śa	ষ	ṣa	স	sa	হ	ha		

Figure1. *Vowels and Consonants of Bengali along with Transliteration Scheme*

3. SYSTEM ARCHITECTURE

Figure 2 shows a block diagram of the Bengali TTS system which is implemented within the Festival framework. Optimal Text Selection, linguistic processing's such as letter to sound rules and word pronunciation. Segmentation, Unit Selection and Synthesis are done by using FESTVOX and FESTIVAL framework.

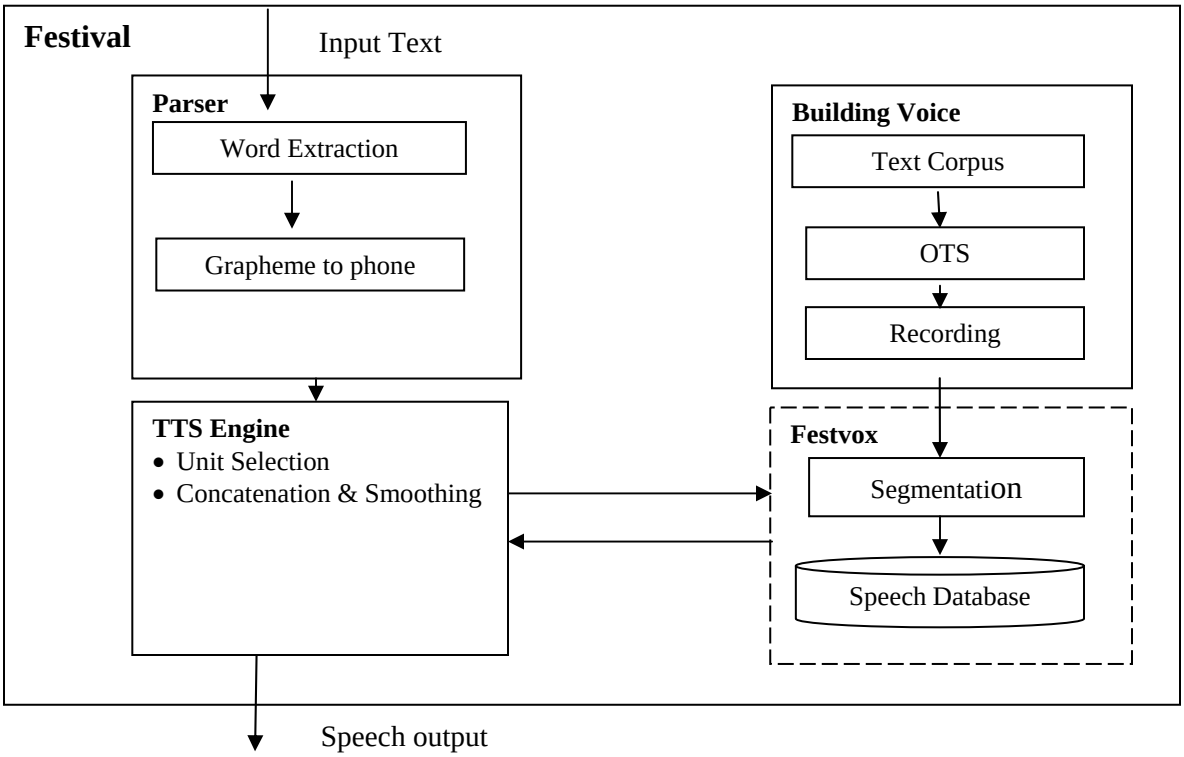

Figure 2. *System architecture of Bengali TTS*

4. LETTER TO SOUND RULE

Letter to Sound rules for Bengali can be build using rule-based methods. In this work we have used a set of rules to map a letter to sound

Unlike other languages based on Devanagari script, the inherent schwa in Bengali is /au/ as opposed to /a/. Some specific rules that are applicable only for Bengali scripts are as follows:

4.1 Gemination of consonants:

- Prominent clusters taking part in gemination are /Cy/, /Cb/, /ksh/, /shm/, /sm/, /tm/ where C stands for consonant. The prerequisite for gemination of these consonants is that all these clusters must be preceded by a vowel.
- In case of word initial cluster the second consonant is dropped.
- In case of clusters like Cyaa or byaa, /aa/ is modified to /ae/.

5. OPTIMAL TEXT SELECTION

The problem of selecting a set of phonetically rich and balanced sentences from a text corpus is a tough task. Literature on this shows that the greedy algorithm performs marginally better and is the most frequently used algorithm for text selection.

The algorithm selects sentence recursively, selecting the sentences with the largest unit first. All the units covered are then removed from the list of units and the next sentence

covering the largest number of the remaining units is selected next. This process is continued until all the covered or the corpus is exhausted.

The units to be covered can be assigned weights, which maybe related to frequency of their occurrence in the corpus. All sentences in the corpus can then be cored based on these units and the sentence with the highest score be selected first. After selection the sentences are rescored based on the remaining to be covered if the corpus covers all the units possible, then the weights assigned can be related inversely to the frequency of occurrence. This ensures that algorithm focus on the rare units first; the logic behind that the most frequent units would be covered alongside. However, if it known that complete coverage is not possible then the weights may be related to the actual frequency of the units.

6. CREATION OF TTS DATABASE FOR BENGALI

A common trend in concatenative approach for TTS system is to use large database of phonetically and prosodically varied speech

6.1 Text Selection for Bengali

The design of a database that contains as much variations as possible in terms of prosody and phonetics is crucial for concatenative synthesis approaches. However, the database size is also an important consideration in this because of the need to minimize the preparation effort of this database as well to maximize the efficiency of the unit selection at the time of synthesis.

With this perspective, to ensure coverage of all Bengali phones in their context (diphones) is a difficult task as Bengali as a large phone set. Thus Optimal Text Selection is extremely important for Bengali.

The first step for text selection involves, parsing html tagged UTF8 into UTF8 format and intern converting it in to ITRANS-3 (Transliteration Scheme).

Linguistics rules like, those for schwa-addition were incorporated into the language-specific module. Finally this raw ITRANS-3 corpus is given to the optimization module.

Text selection was carried out using the BBC Corpus obtained by web crawler.

Total Initial Diphones	1322
Total Sentences Analyzed	1057014
Count of Selected Sentences	719
Count of Selected Words	2666
Total Diphones covered	913
Total Diphones not covered	409

Table 1: *Results of Bengali Text Selection*

7. SPEECH DATA COLLECTION

Presently the speech data collection in the process of selecting a suitable speaker to record the voice

Specifications for high quality speech database are:

- Male/ Female native Bengali speaker.
- The recording has to be made in a sound proof studio using a condenser microphone.
- The speech data should be recorded at 44 KHz, mono channel at 16 bits per sample.

- The high sampling has to be preferred because it would be easier to down sample the speech with least degradation of quality.

As it is not possible to complete the recording in a single session, to ensure consistency, the recordings were made at the same time everyday.

8. SPEECH SEGMENTATION

One of the most important tasks in building speech databases is the annotation of speech data with its contents (labeling) and the time alignment between labeling and speech (segmentation). Phonetic segmentation and labeling are highly desirable and useful for TTS as this information is used for classifying the speech units that helps to select and concatenate the right units in terms of linguistic and acoustic features.

The most precise way to annotate speech data is manually by linguistic experts. However, manual phonetic labeling and segmentation are very costly and require much time and effort. Even well trained, experienced phonetic labelers using efficient speech display and editing tools require about 200 times real time to segment and align speech utterances. To reduce this effort considerably and aid the phonetic labelers, an automated segmentation tool was developed at University of Edinburgh.

For automatic phonetic segmentation, we have been using the most frequently used Tool – FESTVOX based phonetic recognizer for the task of automatic phonetic segmentation.

9. UNIT CLUSTERING AND SYNTHESIS

Given these segments, the unit selection algorithm in FESTVOX clustered the phones based on their acoustic differences. These clusters are then indexed based on higher level features such as phonetic and prosodic features. During synthesis, the appropriate clusters are sought using phonetic and prosodic features of the sentence. A search is then made to find a best path through the candidates of these clusters. Though the units used here are phones, the acoustic frame of previous unit is used during clustering as well as for concatenation.

10. CONCLUSION

In this paper, we have described the development of Bengali TTS system using Data driven approach. In the above sections, the methods and tools developed for collection of speech data has been described. Optimal Text Selection algorithm, Grapheme to Phoneme converter and Automatic Segmentation tools has been discussed and their implementations in the creation of database for Bengali TTS have been described.

Though the paper describes the methodology and the use of tools for Bengali TTS, the tools and the methodology are in themselves languages independent and can be easily customized for any Indian languages.

The work is in the progress of selecting the speaker for recording and building the voice with respect to the specifications given in the Section 7.

11. REFERENCES

[1] “Digital Library of India, http://dli.iiit.ac.in/,” 2005.

[2] Raj Reddy, “PCtvt: A multifunction information appliance for illiterate people, http://www.rr.cs.cmu.edu/pctvt.ppt,” in ICT4B retreat at UC Berkeley, August 26, 2004.

[3] S.P.Kishore and Alan W Black, “A data-driven synthesis approach for indian languages using syllable as basic unit,” in Proceedings of Eurospeech, Geneva, Switerzland, 2003.

[4] Alan W Black and Kevin Lenzo, “Building voices in the festival speech synthesis system, www.festvox.org/festvox/index.html,” 2000.

[5] N. Udhyakumar, C.S. Kumar, R. Srinivasan, and R. Swaminathan, “Decision tree learning for automatic
grapheme-to-phoneme conversion for Tamil,” in SPECOM-2004, St. Petersburg, Russia, 2004.

[6] C.S.Kumar, V.Shunmugom, N. Udhyakumar, and R. Srinivasan, “Rule-based automatic grapheme to phoneme conversion for Tamil,” in Proc. ICSLT,
Delhi, India, 2004.

[7] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, “Spoken Language Processing: A Guide to Theory, Algorithm and System Development”, Prentice Hall PTR, 2001.

[8] Jurafsky, Daniel, and James H Martin, “Speech and Language Processing”, Prentice-Hall, 2000.

[9] “Optimal Data Selection for Unit Selection Synthesis” Alan W Black and Kevin A. Lenzo Carnegie Mellon University and Cepstral, LLC.

[10] “Experiments with Unit Selection Speech Databases for Indian Languages”, S.P.Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal.