# A Bengali Speech Synthesizer on Android OS

**Sankar Mukherjee, Shyamal Kumar Das Mandal**
Center for Educational Technology
Indian Institute of Technology Kharagpur
sankar1535@gmail.com
sdasmandal@cet.iitkgp.ernet.in

## Abstract

Different Bengali TTS systems are already available on a resourceful platform such as a personal computer. However, porting these systems to a resource limited device such as a mobile phone is not an easy task. Practical aspects including application size and processing time have to be concerned. This paper describes the implementation of a Bengali speech synthesizer on a mobile device. For speech generation we used Epoch Synchronous Non Overlap Add (ESNOLA) based concatenative speech synthesis technique which uses the partnemes as the smallest signal units for concatenations.

## 1 Introduction

Technologies for handheld devices with open platforms have made rapid progresses. Recently open-platforms Android is getting momentum. Mobile devices with microphone and speaker, video camera, touch screen, GPS, etc, are served as sensors for experiencing with augmented reality in human life. Speech synthesis may become one of the main modalities on mobile devices as the screen size and several application scenarios (e.g., driving, jogging) limits the use of visual modality. Optimizing a speech synthesis system on mobile devices is a challenging task because the storage capacity and the computing performance are limited. Even if the storage capacity of the device is quite high, it is unlikely that users will let e.g., the half of their storage for speech synthesis purposes. So it is necessary to have small footprint. Until now, text-to-speech applications have been developed on many platforms, such as PC, electronic dictionary and mobile device. However, most applications are for English language. Early works on developing a TTS system on a mobile device focused mainly on migration of an existing TTS system from a resourceful platform to a resource-limited platform (W. Black and K. A. Lenzo, 2001; Hoffmann, R et al., 2003). Most of the effort was spent on code optimization and database compression. Since the space was quite limited, only a small diphone database could be utilized which reduced the quality of synthesized speech. To improve the output speech quality some researchers attempted to apply a unit selection technique on a resource limited device. (Tsiakoulis, et al, 2008) used a database small enough for an embedded device without much reduction in speech quality.  (Pucher, M. and Frohlich, 2005) used a large unit selection database but synthesize an output speech on a server and then transferred the wave form to a mobile device over a network.

Bengali TTS systems have been already developed and produced reasonably acceptable synthesized output quality on PC, as Shyamal Kumar Das Mandal and Asoke Kumar Datta (2007). However the same has not yet been implemented for resource-limited or embedded devices such as mobile phones.

The goal of our research is to develop a Bengali speech synthesizer that can produce an acceptable quality of synthesized output in almost real-time on mobile device.

*Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments*, pages 43–46,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

## 2 Speech Synthesis Techniques

Speech synthesis involves the algorithmic conversion of input text data to speech waveforms. Speech synthesizers are characterized by the methods used for storage, encoding and synthesis of the speech. The synthesis method is determined by the vocabulary size, as all possible utterances of the language need to be modeled. There are different approaches to speech synthesis, such as rule-based, articulatory modeling and concatenative technique. Recent speech research has been directed towards concatenative speech synthesizers. We develop our synthesizer based on Epoch Synchronous Non Overlap Add (ESNOLA) concatenative speech synthesis method, as Shyamal Kumar Das Mandal and Asoke Kumar Datta (2007).

ESNOLA allows judicious selection of signal segment so that the smaller fundamental parts of the phoneme may be used as unit for reducing both the number and the size of the signal elements in the dictionary. This is called Partnemes. Further the methodology of concatenation provides adequate processing for proper matching between different segments during concatenation and it supports unlimited vocabulary without decreasing the quality.

## 3 TTS System Based on ESNOLA Method

A schematic diagram of the speech synthesis system is shown in Figure 1.

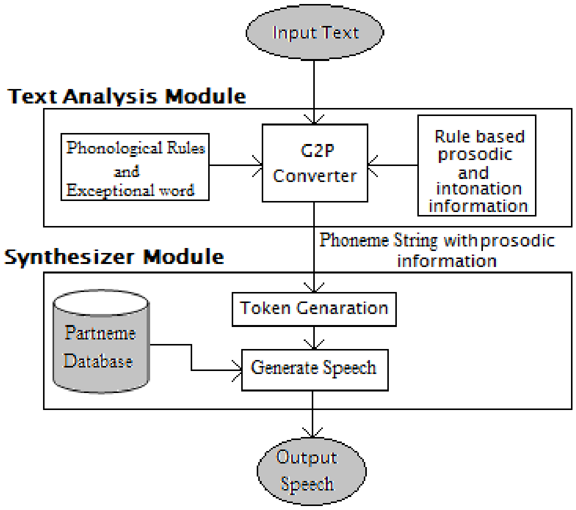

Input Text
Text Analysis Module
Phonological Rules and Exceptional word
G2P Converter
Rule based prosodic and intonation information
Phoneme String with prosodic information
Synthesizer Module
Token Genaration
Partneme Database
Generate Speech
Output Speech

Figure 1: Text-to-Speech process diagram

The Text-to-Speech system consists of two main parts: Text analysis module and Synthesizer module.

### 3.1 Text Analysis Module

The text analysis module has two broad sections one is the phonological analysis module and other is the analysis of the text for prosody and intonation. Bangla is a syllabic script, phonological analysis i.e. Grapheme to Phoneme conversation is a formidable problem (Suniti Kumar Chatterji, 2002; Sarkar Pabitra, 1990) specially found in case of two vowels /a/ and /e/ and some consonant clusters. A set of phonological rule including exception dictionary is developed and implemented, as (Basu, J et al., 2009). The phonological rules also depend upon POS and semantics. But due to its requirement of language analysis it is taken care by exception dictionary.

### 3.2 Synthesizer Module

Synthesizer module has two parts. First generate token and second combine splices of pre-recorded speech and generate the synthesized voice output using ESNOLA approach as in Shyamal Kr Das Mandal, et al. (2007). In ESNOLA approach, the synthesized output is generated by concatenating the basic signal segments from the signal dictionary at epoch positions. The epochs are most important for signal units, which represent vocalic or quasi-periodic sounds. An epoch position is represented in Figure 2.

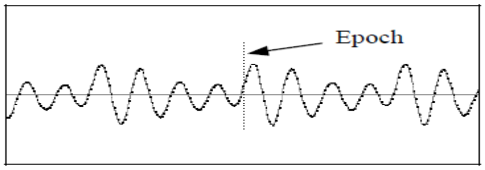


Figure 2: Epoch position of a speech segment

Steady states in the nucleus vowel segment of the synthesized signal are generated by the linear interpolation with appropriate weights of the last period and the first period respectively of the preceding and the succeeding segments. The generated signals require some smoothing at the point of concatenation. This is achieved by a proper windowing of the output signal without hampering the spectral quality.

The synthesized voiced output for the name “ভারত” is shown in Figure 3.

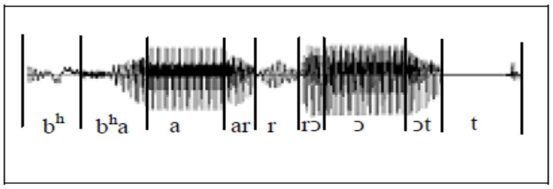


Figure 3: Represent a synthesized voiced output for a given text input / $b^h$arot /

# 4 Implementation on Android

The exact system specification is shown on Table 1. An Android application will run on a mobile device with limited computing power and storage, and constrained battery life. Because of this, it should be efficient. Following actions are taken to run the application on Android –

Table 1: System Specifications

| Features | LG Optimus One P500 |
|---|---|
| Operating System | Android OS, v2.2 |
| Processor | ARM 11 |
| CPU speed | 600 MHz |
| RAM | 512 MB |
| Display | 256K colors, TFT |
| Input method | Touch-screen |
| Connectivity | USB |

## 4.1 Memory Management

On Android, a Context is an abstract class which is used for many operations but mostly to load and access resources. But keeping a long-lived reference to a Context and preventing the GC (Garbage Collection) from collecting it causes memory leaks issues. But in here this system has to have long-lived objects that needed a context. So to overcome this Application-Context Class is used. This context will live as long as your application is alive and does not depend on the activities life cycle. It is obtained by calling *Activity.getApplication()*. Apart from that the partneme database is kept in external storage card. Owing to memory constraints, the speech output file is deleted after the speech is produced.

## 4.2 Optimizing the Source Code

On Android virtual method calls are expensive, much more so than instance field lookups. So common object-oriented programming practices are followed and have getters and setters in the public interface.

All total 596 sound files are stored in the partneme database. Total size of the database is 1.0 Mb and application size is 2.26 Mb.

The TTS system has two major functionality. Firstly, it can read the Bengali massage stored in the phones inbox and secondly, user can generate Bengali speech by typing the Bengali word in English alphabet format.

The input text in English alphabet can be keyed in the provided text box (Figure 3). The ‘Speak to me’ button generates the speech file corresponding to the text keyed in and plays the audio file generated. Graphical user interface is shown in Figure. 4-5.

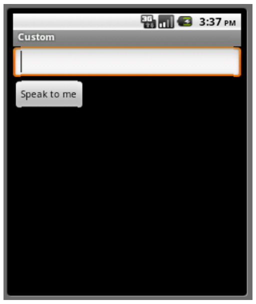
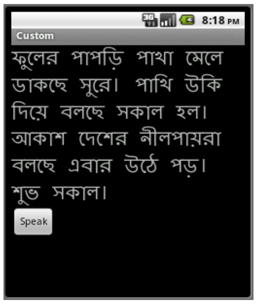


Figure 4



ফুলের পাপড়ি পাখা মেলে
ডাকছে সুরে। পাখি উকি
দিয়ে বলছে সকাল হল।
আকাশ দেশের নীলপায়রা
বলছে এবার উঠে পড়।
শুভ সকাল।

Figure 5

Fig. 4-5 is the internal interface of the application

Application can be distributed to end user directly by developer website or by Android Market an on device application store. This TTS application has not published yet but it can be downloaded on the android device connected to a desktop PC through the USB port.

# 5 Performance And Quality Evaluation

## 5.1 Processing Speed Test

Measurement of processing speed is done by counting the synthesis time manually. We started measuring the time when a "speak" button (Figure 5) is pressed until the first speech sound is pronounced. Results are shown in Table 2.

Table 2 speed time test

| Utterance (words) | No. of syllables | Processing Speed [in sec.] |
|---|---|---|
| 2 | 6 | 0.45 |
| 3 | 8 | 0.56 |
| 4 | 11 | 0.86 |
| 5 | 15 | 1.19 |

## 5.2 Speech Quality Evaluation

To measure the output speech quality 5 subjects, 3 male (L1, L2, L3) and 2 female (L4, L5), are selected and their age ranging from 24 to 50. All subjects are native speakers of Standard Colloquial Bangla and non speech expert. 10 original (as uttered by speaker) and modified (as uttered with android version) sentences are randomly presented for listening and their judgment in 5 point score (1=less natural – 5=most natural). Table 3 represents the tabulated mean opinion scores for all sentences of each subject for original as well as modified sentences.

Table 3 result of listing test

| | | L1 | L2 | L3 | L4 | L5 |
|---|---|---|---|---|---|---|
| Modified Sentences | Avg | 3.82 | 1.76 | 2.62 | 2.73 | 3.5 |
| | Stdev | 0.73 | 1.15 | 0.82 | 0.81 | 0.5 |
| Original Sentences | Avg | 4.91 | 4.33 | 4.82 | 4.76 | 4.8 |
| | Stdev | 0.11 | 0.23 | 0.83 | 0.42 | 0.3 |

The total average score for the original sentences is 4.72 and the modified sentence is 2.88. Figure 6 graphically represents the mean opinion score to visualize the closeness of the employed prosodic rules to the original sentences.

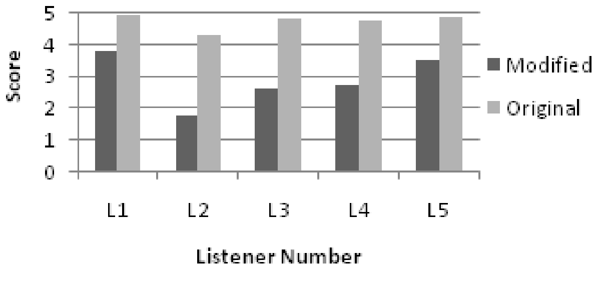


Figure 6: Bar chart of the listening test

## 6 Conclusion And Future Works

In this paper, we describe our implementation of a Bengali speech synthesizer on a mobile device. Our goal is to develop a text-to-speech (TTS) application that can produce an output speech in almost real-time on an average smart phone. Our synthesizer is based on Epoch Synchronous Non Overlap Add (ESNOLA) suitable for implementing a fast and small TTS application. We modified several components in ESNOLA to make it run on android device. As for the output sound quality of TTS, there is plenty of room for improvement. We also plan to develop a more complete text analysis module which can handle the prosody at the sentence better way.

## References

Basu, J., Basu, T., Mitra, M., Mandal, S. 2009. Grapheme to Phoneme (G2P) conversion for Bangla. Speech Database and Assessments, Oriental COCOSDA International Conference, pp. 66-71.

Chatterji Suniti Kumar. 2002. The Original and Development of the Bengali Language. Published by Rupa.Co, ISBN 81-7167-117-9, 1926.

Das Mandal Shyamal Kr, Saha Arup, Sarkar Indranil Datta Asoke Kumar. 2005. Phonological, International & Prosodic Aspects of Concatenative Speech Synthesizer Development for Bangla. Proceeding of SIMPLE-05, pp56-60.

Hoffmann, R et aL. 2003. A Multilingual TTS System with less than 1: MByte Footprint for Embedded Applications. Proceeding of ICASSP.

M. Pucher, and P. Frohlich. 2005. A User Study on the Influence of Mobile Device Class, Synthesis Method, Data Rate and Lexicon on Speech Synthesis Quality. Inter Speech.

P. Tsiakoulis, A. Chalamandaris, S. Karabetsos, and S. Raptis. 2008. A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis. ICASSP, Las Vegas, Nevada, USA.

Sarkar Pabitra. 1990. Bangla Balo. Prama prakasani.

Shyamal Kumar Das MandaI and Asoke Kumar Datta,. 2007. Epoch synchronous non-overlap-add (ESNOLA) method-based concatenative speech synthesis system for Bangla. 6th ISCA Workshop on Speech Synthesis, Germany, pp. 351-355.

W. Black and K. A. Lenzo. 2001. Flite: a small fast nm-time synthesis engine. 4th ISCA Workshop on Speech Synthesis.