C-DAC Kolkata

Name of the Technology: Bengali Text to Speech Synthesis System

Nature of Technology: It converts the given written Bengali text (in ISSCI) into unintonated phonetically clear speech

Level: Technology

Technical Description:

Speech generation is the process, which allows the transformation of a string of phonetic and prosodic symbols into a synthetic speech signal. The quality of the result is a function of the quality of the string, as well as of the quality of the generation process itself

In the past few decades, various Researchers have worked in the area of Speech Synthesis and Recognition and have developed different algorithms and methodologies for different speech technology development. In area of Speech Synthesis there are a number of different methodologies like Formant, Articulatory, Sinusoidal and Concatenation Synthesis.

In the last decade there has been a significant trend for development of speech synthesizers using Concatenative based Synthesis techniques. There are a number of different methodologies for Concatenative Synthesis like TDPSOLA, PSOLA, MBROLA and Epoch Synchronous Non-Over Lapping Add (ESNOLA).

In TDPSOLA method based di-phone concatenative technique has inherent problem in introducing intonation and prosody. In manipulating pitch for introduction of intonation the phonetic quality is often seriously compromised as because it is only pitch synchronous. Only pitch synchrony does not guaranty the preservation of phonetic quality.

MBROLA has the same problems. Along with those, to accommodate intonation a large multiplicity of the diphones is required. This is a major problem in building all necessary elements of the signal dictionary.

ESNOLA technique provides the complete control on implementation of intonation and prosody. It allows judicial selection of signal segment so that smaller fundamental parts of the phonemes may be used as units reducing both the number and the size of the signal elements in the dictionary. Further the methodology of concatenation provides adequate processing for proper matching between different segments during concatenation. The use of special type of basic signal segment makes the size of signal dictionary very small so there is a possibility of its implementation in low-cost, general-purpose electronic devices.

Recently CDAC (Kolkata) has produced user friendly complete TTS for Bangla using the ESNOLA technique with very high intelligibility and naturalness of phonetic quality. It was possible because the ESNOLA supports introduction of jitter, shimmer and complexity perturbations. Recent development in ESNOLA technique has also shown the capability of dealing with the complexity mismatch and pitch mismatch across the concatenation boundary. A tentative demonstration model of intonated Bangla speech is also ready at CDAC (Kolkata). The final version awaits development of intonation and prosodic rules, which are not available in any Indian languages. The methodology of developing rule bases has also been completed. Absence of adequate databases has held up further development.

Since indigenous technology of adequate refinement for development of TTS in Indian Languages is already available we do not see any advantage in wasting time over experimenting with the technologies available for European Languages, particularly when we know their deficiencies vis-à-vis ESNOLA technique and the fact that the phonetic structure of Indian Languages differ significantly from that of European Languages. We

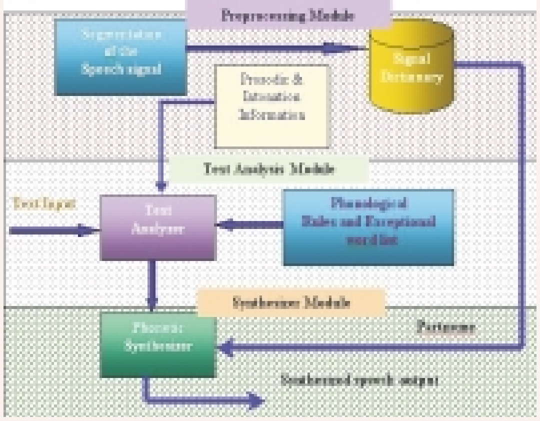

Preprocessing Module
Segmentation of the Speech signal
Signal Dictionary
Prosodic & Intonation Information
Text Analysis Module
Text Input
Text Analyzer
Phonological Rules and Exceptional word list
Synthesizer Module
Partneme
Phonetic Synthesizer
Synthesized speech output

Figure 1. Basic Block Diagram of TTS System using ESNOLA Technique

intend to develop TTS for all the major Indian Languages either at CDAC (Kolkata) or to fully support these developments any where through technology transfer. The estimated time for the complete development is around one year provided the rule for phonology; intonation and prosody are carried out simultaneously.

The above block diagram (Fig.-1) describes the basic part of the ESNOLA technique for the development of text-to speech synthesis system.

It consists of there part 1. Preprocessing module 2. Text analysis module 3. Synthesizer module.

Preprocessing module: In this module the required speech segment database is created from the pre-recorded natural speech signal. In our system we called the segment as pratneme. The advantage of using partnemes as the basic unit is the simplicity of introducing intonation and prosodic rules into the synthesized speech signals. Though prosody and intonation have not been implemented in the present developed system due to the lack of intonation and prosodic rule but the implementation methodology development is tested.

Text analysis module: The Text analysis module is the front-end language processor of the Text-to-Speech System, which accepts input text and generates corresponding phoneme string and stress markers. On many occasions the Text Analyzer consists of a natural language processing module (NLP), capable of producing a phonetic transcription of the text read, together with the desired intonation and rhythm (often termed as prosody).

Synthesizer module: It is the task of the Synthesizer module to combine splices of pre-recorded speech and generate the synthesized voice output. A sequence of segments is first deduced from the phonemic input of the synthesizer. If required, the prosodic events may be assigned to individual segments based on the information extracted by the NLP.

The Synthesizer Module functions in the following way:

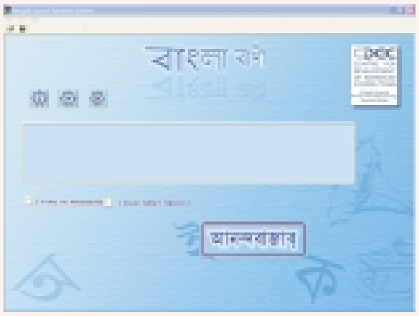
- The Phoneme string input from the Text Analyzer is assigned tokens based on the indexing of the segmented partneme voice signals.
- Modification of pitch, amplitude and duration of the vowels to implement the prosodic and intonational data may be done.
- The selected segments are concatenated to get the raw output signal.
- Spectral smoothing is performed on the concatenation points to remove mismatch and other spectral disturbances to generate the final voice output.

Specification: unlimited, flat, phonetically clear Bengali concatenative synthesizer using ESNOLA technique.

O/S: Windows and NT.

Front-end:

It also available in dll from which can be integrated with other application.



Expandability: Technology can be used for development of text to speech synthesis system in other Indian language.

Intonation and prosodic incorporation can be done for naturalness.

Portability: this system can be easily integrated with other application for information disbursing in local languages i.e. telephone address enquiry system (197 pariseba)

Availability of documentation: documentation is available

Name and address of the Resource Person:

A. B. Saha (amiya.saha@erdcical.org)
Shyamal Das Mandal
(shyamal.dasmandal@erdcical.org)
C-DAC Kolkata 700 091