# Nepali Text to Speech Synthesis System using ESNOLA Method of Concatenation

Bhusan Chettri
Sikkim Manipal Institute of Technology
Majitar, Sikkim, India

Krishna Bikram Shah
Dream Tech
Kalimpong, West Bengal, India

## ABSTRACT

This paper confer the tools and methodology used in developing a Nepali Text to Speech Synthesis System, which is based on concatenative approach employing Epoch Synchronous Non Overlap Add Method (ESNOLA), which uses signal dictionary having raw sound signal representing parts of phonemes as a speech database. The developed system is an unintonated (flat) TTS system where the pitch of the pre-recorded speech signal remains same throughout, while taking care of aspects such as naturalness, personality, platform independence and quality assessments. Some of the applications and problems encountered with TTS systems are also discussed.

## Keywords

TTS, ESNOLA, Partneme, Speech, Synthetic Synthesis.

## 1. INTRODUCTION

The vocalized form of human communication is termed as Speech. It is the primary means of communication among human beings, though it is said that human can communicate through eye and some time heart as well. The approach is based upon the syntactic combination of lexical and names that are drawn from very large vocabularies generally consisting of more than 10,000 words. Each spoken word is created out of phonetic combination of a limited set of vowel and consonant speech, which are the sound units in speech synthesis. The current age of information technology, information exchange methodologies demands overcoming the barrier of human limitations, and has gained noticeable importance. A text to speech system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. [1]

Synthetic speech is created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored units. A system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely “synthetic” voice output [2]

It is apparent that there is steady development in Synthetic or artificial speech during the last decades though there remain much to be done in this regard. Speech synthesis may be categorized as restricted (messaging) and unrestricted (text-to-speech) synthesis. The first one is suitable for announcing and information systems while the latter is needed for example in applications for the visually impaired. The text to speech procedure consists of two main phases, usually called Natural Language Processing (NLP) or high level and Digital Signal Processing (DSP) or low level synthesis. In high level synthesis the input text is converted into such form that the low level synthesizer can produce the output speech. The three basic methods for low level synthesis are the formant, concatenative, and articulatory synthesis. The formant synthesis is based on the modeling of the resonances in the vocal tract and is perhaps the most commonly used during the last decades. However, the concatenative synthesis which is based on playing pre-recorded samples from natural speech is becoming more popular. The most accurate method is articulatory synthesis which models the human speech production system directly, but it is also the most difficult approach. Since the quality of synthetic speech is improving steadily, the application field is also expanding rapidly.

Different methodologies for Concatenative Synthesis are there like TDPSOLA, PSOLA, MBROLA and ESNOLA. The Nepali text to speech synthesis system described here is based on Concatenative technique that uses Epoch Synchronous Non Overlap Add Method (ESNOLA) where the speech signal dictionary is created by cutting the pre-recorded natural speech such that the speech segments begin and ends all at the positive zero crossing to eliminate mismatch or distortion during concatenation. The system uses partneme as the smallest signal units for concatenation.

ESNOLA technique provides the complete control on implementation of intonation and prosody [2, 10]. It allows judicious 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 that there is a possibility of its implementation in low-cost, general-purpose electronic devices.

## 2. TEXT TO SPEECH

A true text to speech system should be able to accept any input text in the chosen language including new words. Text-to-speech synthesis is the automated transformation of a text into speech that sounds, as closer as possible, as a native speaker of the language reading the text. All the TTS systems take the text in digital format, such as ASCII for English, Unicode for Nepali. It is possible to build a TTS system which will work in combination with Optical Character Recognition system so that the system can read from printed text; but this does not affect the point, the output of the OCR system will be finally coded in corresponding digital text that is served as input for the TTS system. It is considered that a complete Text-to-speech system for any language must be able to handle text written in its normal form in that language, using its standard script or orthography. Hence a TTS which will accept romanized input to speak out is not considered as a true

TTS system. Normally, a text contains many inputs besides ordinary words. A typical Nepali text may contain numbers in different contexts; amount of money or phone number, percentage, acronyms and so on. A full text-to-speech system should be able to handle all these kinds of inputs with reasonable tolerance. [17]

A closer examination of TTS system problem of converting text into voice output can be sub-divided into two problems. The first one is proper analysis of text with linguistic rules in that language. This analyses the input text and converts the orthographic representation of the text into its linguistic representation. The linguistic representation gives information about grammatical categories of words, their tonal properties and most importantly pronunciation of the words. The second problem is proper speech synthesis, i.e. generation of speech waveform from the internal linguistic representation. Speech can be synthesized by creating an environment of oral track simulation (articulatry synthesis) or by creating an acoustic model (formant synthesis) or by manipulating the prerecorded speech units (concatenative synthesis). [17, 19]

Researchers have been studying for centuries for artificial production of speech. The effort has transited from mechanical modeling of human speech production system to electrical speech synthesizers and now to different modern synthesis techniques of concatenating recorded speech with text analysis to obtain more natural sounding voice outputs. The earliest efforts of producing artificial sound were with different mechanical devices which model the human speech production system. The earliest mechanical models had different music instrument like devices capable of producing only five long vowels each (a, e, i, o, and u). To produce voice from these devices air had to be blown with air. Hence the production of voice was limited to very basic vowels and the process of speech production was not automatic. These were the acoustic resonators modeling the human vocal tract. After a few generations other parts of the machine were improved, like pressure chamber for lungs, vibrating reed to act like vocal cords and leather tube for the vocal tract. With appropriate manipulation of the shape of the leather tube these machines were able to produce vowel sounds and consonants were simulated by separate constricted passages and controlled by the fingers. To simulate more sounds other components were added to the machine, e.g. movable lips, and tongue. Such machines were a good model of the human speech system but were limited to produce phones and limited set of words only, but not long sentences.

First full electrical synthesis devices had a buzzer as excitation and resonant circuits to model the acoustic resonances of the vocal tract. These were able to generate single vowel sounds and no consonants. The first true speech synthesizer was introduced by Homer Dudley in 1939, called VODER (Voice Coder). The VODER consisted of wrist bar for selecting a voicing or noise source and foot pedal to control the fundamental frequency. Source signal was passed through ten band pass filters whose output levels were controlled by fingers. Only skilled operator of VODER could produce a sentence of speech from the device. The demonstration of VODER showed that artificial production of speech was possible and increased more interest towards speech synthesis. Further study on speech signal and its decomposition invented new technique of speech synthesis called formant synthesis with proper prediction of parameters representing the signal. Formant synthesis does not produce natural sounding speech when operated in fully automated mode to predict the signal parameters. With the advent of digital representation of digital sounds, availability of cheap and powerful computer hardware, and different digital signal processing techniques speech generation method shifted from fully synthetic to concatenation of natural recorded speech. Speech generated from concatenation method resulted more closer to natural voice than the fully synthesized voice. In addition to this, since memory cost has been dramatically decreased and the processing speed has been exponentially increased the developers are nowadays interested in different concatenative approaches. [5, 4]

## 3. METHODOLOGY

The development of ESNOLA based Nepali Text to speech synthesis system is basically divided into three phases. Phase-I: Database creation. Phase-II: Text Analysis. Phase-III Synthesis. The first two phases are language dependent while the third phase is language independent. The entire process is shown in Fig. 1.

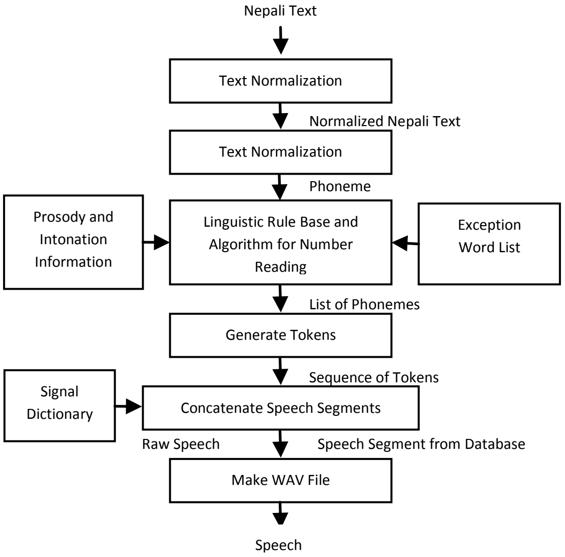

**Figure 1: Basic Flow Diagram of Nepali TTS Synthesis System based on ESNOLA method of Concatenation**

### 3.1 Database Creation

In this phase the basic speech segment (partneme) database is created from the pre-recorded natural speech in Nepali. The advantage of using partneme [2] 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 system, the scope for further enhancement of the system using these techniques still remains. The following steps are followed for making the database [11]

Step 1: Nonsensical words of the form CVCVCV (C: Consonant, V: Vowel) are uttered by a speaker and recorded. Recording format: 16 bit PCM, mono, sampling frequency 22.05 KHz. The recorded words should be stress free. From these words the best VCB syllables, which are also stress free are selected.

Step 2: Pitch normalization is performed on the samples to avoid pitch mismatch. The average pitch of all the signals is found and the corresponding pitch of all the signals is modified accordingly by changing the sampling frequency, Sf2 using the equation:

Sf2 = (Signal frequency) / (Average frequency) * Sampling frequency.

Step 3: Amplitude normalization is performed with respect to the amplitude of the vowels.

Finally partneme are extracted from the VCV nonsensical words as shown in figure2. The initial part is the consonant which gives way to the transition from the consonant to the vowel (CV), as can be seen by the rising waveform. The steady state period which follows is the vowel. The vowel is extracted by taking a single period, which is pasted, in a repeated fashion during concatenation. Finally the falling part of the waveform represents the transition from the vowel to the consonant (VC) [1]

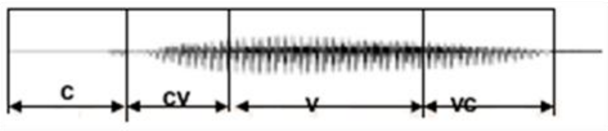

**Figure 2: Segmentation of recorded speech signal**

## 3.2 Text Analysis

The text analysis phase is the front end language processor of the Text to Speech System, which accepts input text and generates corresponding phoneme string and stress markers. Here we take the input text in Unicode format (Devanagari script) from a front end written in Visual C++. Text in Nepali may contain many non standard words like abbreviations, currencies, digits, date and acronyms. This text needs to be converted from nonstandard words into standard words. On many occasions the Text Analyzer consists of a natural language processing module, capable of producing a phonetic transcription of the text read, together with the desired intonation and rhythm [2, 11].

A text analysis module is necessary as a front end, since the TTS system should in principle be able to read any text, including numbers, abbreviations, acronyms and idiomatic, in any format. It also performs the apparently trivial task of finding the end of sentences in the input text. It transforms the input sentences into manageable lists of word-like units and stores them in the internal data structure.

## 3.3 Synthesis

The input to the final phase, synthesis is the sequence of phonemes from the text analysis phase. A sequence of segments is first deduced from the phoneme input of the synthesizer. The synthesis phase basically works in the following way. First the phoneme input from the text analyzer is assigned tokens based on the indexing of the segmented partneme voice signals. The selected segments are concatenated to get the raw output signal. Finally spectral smoothing is performed on the concatenation point to remove mismatch and other spectral disturbances to generate the final voice output. In concatenation, the major problem is the matching of complexity, pitch and amplitude across the boundary. The last two problems are solved by normalization of signal units before putting them into the segment dictionary. This produces only flat pitch. The complexity mismatch is yet a problem. The following rules are followed for generating token for the sequence of phonemes received from text analyzer:

CVCV = C +CV+V+VC+C+V+Vo

VCV = Vi+V +VC+C+CV+V+Vo

CVYV = C +CV+V+VY+YV+Vo

CVV = C+CV+VV+Vo

Where Vi, Vo, V and C represent fade-in vowel, fade-out vowel, Medial vowel and consonant respectively. The fade in and fade out operation is applicable for the terminal vowels only. In non-terminal cases Vo and Vi are to be treated as V. In ESNOLA approach [2, 11], 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 3.

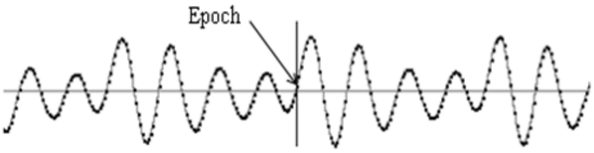

**Figure 3: EPOCH position of a Speech Segment**

In this section we discuss how the system produces speech for the inputted Nepali text taking an example of two different Nepali sentences.

Example1:

Input Text:

मेरो नाम कमल हो। (meaning my name is Kamal in English)

The Nepali text (grapheme) is first converted to some phonetic representation usually called phoneme, through a grapheme to phoneme conversion. On these, set of phonemes phonological rule base, which contains set of linguistic rules, is applied to introduce some naturalness in the final speech.

After application of Grapheme to Phoneme rule on the input, we get the following phoneme

ME3RO3 NA4M KML HO3/

After application of various phonological rules like Add Vowel A, Delete Halant, rule of Word Final, the following set of phonemes is obtained, which will be sent to the Synthesis module.

#MA2RO NA4M KAMAL HO/

Now the synthesis module will generate the following tokens from the above set of phonemes:

#M MA2 A2 A2R R RO O N NA4 A4 A4M M K KA A AM M MA A AL L H HO O/

{C CV V VC C CV V C CV V VC C C CV V VC C CV V VC C C CV V}

For the sequence of tokens generated, pre-recorded speech segments will be extracted from the speech signal dictionary and the segments are concatenated together to get a final waveform which is the required output speech for the given input text.

Example2:

Input Text:

मेरोमा १० रूपया छ। (I have ten rupees in English)

After Text Normalization the input Nepali Text is transformed to following:

मेरोमा दस रूपया छ।

Grapheme to Phoneme conversion: #ME3R03MA4 DS RB3PYA4 C1/

Application of various phonological/Linguistic rules yields the following set of phonemes: #MA2ROMA4 DAS RUPYA4 C1A/

The synthesis module will now generate the following tokens:

#M MA2 A2 A2R R RO O OM M MA4 A4 D DA A AS S R RU U UP P Y YA4 A4 C1 C1A A/

{C CV V VC C CV V VC C CV V C CV V VC C C CV V VC C C CV V C CV V/}

For the sequence of tokens generated, pre-recorded speech segments will be extracted from the speech signal dictionary and the segments are concatenated together to get a final waveform which is the required output speech for the given input text.

The system was also tested with many other input Nepali sentences. In a room with a capacity of 50 people, the system was tested. It was given various input text and the synthesizer was run on the given inputs. The audiences present in the room were asked about the clarity, audibility and naturalness in the speech produced from the computer speaker, wherein the following feedbacks were given by the audience.

The synthetic speech was clear and understandable to the audience in the room.

The speech sounded quite natural as spoken by a human.

The curiosity among the audience was in learning about the variations in pitch that was missing and the prosodic parameters as the sound produced was flat, since the pitch of all the recorded voice segments while creating the speech database were kept same. In the present system we are not considering prosody and intonation. So the speech produced would sound flat.

## 4. CONCLUSION

In this paper we have described the various steps involved in developing the Nepali text to speech synthesis system based on ESNOLA method of concatenation. Partneme has been used as the smallest signal unit for preparing the signal dictionary. In the current system linguistic feature such as intonation and prosody has not been implemented. Further enhancement of the system incorporating these features still remains. However the system produces flat speech for any inputted Nepali text, which sounds as natural as spoken by a human. This system can help to overcome the literacy barrier of common mass, can also empower the visually impaired population and increase the possibilities of improved man-machine interaction through on-line newspaper reading from the Internet. The system can be extended to include more features such as more emotions, improved tokenization and use of minimal database.

## 5. REFERENCES

[1] Jonathan Allen, M. Sharon Hunnicutt, Dennis Klatt, “From Text to Speech The MITalk System”, Cambridge University Press, 1987.

[2] Mandal, Shyamal Kumar Das and Datta, Asoke kumar, “Epoch Synchronous non-overlap-add (ESNOLA) method based concatenative speech synthesis system for Bangla”. ISCA workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007.

[3] CDAC: Research & Development – Speech Research. Online: Access Date: 4th July, 2012.

[4] Thierry Dutoit. An introduction to Text-To-Speech Synthesis. Kluwer Academic Publishers. 1997.

[5] Speech Synthesis. Online: http://en.wikipedia/wiki/Speech_Synthesis. Access Date: 4th July, 2012.

[6] Building Synthetic Voices. Online: http://www.festvox.org/bsv. Access Date: 4th July, 2012.

[7] Muhammad Masud Rashid, Md. Akhter Hussain, M. Shahidur Rahman, “Diphone preparation for Bangla text to Speech Synthesis”, Proc. Of International Conference on Computer Sciences and Information Technology, pp. 226-230, Dhaka, November, 2009.

[8] Firoj Alam, S.M. Murtoza Habib, Mumit Khan, “Text normalization System for Bangla”, Proc. of Conference on Language and Technology, Lahore, pp. 22-24, 2009.

[9] Firoj Alam, Promila Kanti Nath and Mumit Khan, “Text to Speech for Bangla language using Festival”, Proc. of Intl. Conf. on Digital Communications and Computer Applications, Irbid, Jordan, 2007.

[10] Tanuja Sarkar, Venkatesh Keri, Santhosh Yuvaraj, Kishore Prahalad, “Building Bengali Voice using Festival”, Proc. of ICLSI 2005, Hyderabad, India, 2005.

[11] Das Mandal S.K, Datta A.K, Gupta B. “Spectral Matching of Epoch Synchronous Non-Over lapping Add (ESNOLA) Method based Concatenative Synthesizer”, International Conference on Communication Devices and Intelligent System (CODIS-2004), Jadavpur University, 2004, pp. 729-732.

[12] Das Mandal Shyamal Kr, Saha Arup, Sarkar Indranil, Datta Asoke Kumar, “Phonological, International & Prosodic Aspects of Concatenative SpeechSynthesizer Development for Bangla”, Proceedings of SIMPLE-05, February 2005, pp. 56-60, 2005.

[13] Nepali TTS. Online: http://www.bhashasanchar.org/pdfs/NepaliTTS_%20manual.pdf. Access Date: 4th July, 2012.

[14] Nepali Language. Online: http://en.wikipedia.org/wiki/Nepali_language. Access Date: 4th July, 2012

[15] Nepali fonts. Online: http://www.explorenepal.com/fonts. Access Date: 4th July, 2012.

[16] J. Acharya, A Descriptive Grammar of Nepali And An Analyzed Corpus, Georgetown University Press, Washington, DC, 1991.

[17] M.J. Liberman, K.W. Church, "Text Analysis and Word Pronunciation in Text-to-Speech Synthesis", Advances in Speech Signal Processing, S.Fumy, M.M. Sondhi eds, Dekker, New York, pp. 791-831, 1992.

[18] Narasimhan B, Sproat R and Kiraz G. Schwa-deletion in Hindi Text-to-speech synthesis. In workshop on computational linguistic in South Asian Languages, 21st SALA, October 2001, Konstanz.

[19] Hunnicut S., "Grapheme-to-Phoneme rules: a Review", Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR 2-3, pp.