Epoch Synchronous Non-Overlapping Add (ESNOLA) Approach Concatenative Text to Speech Synthesis-A Technical Report

1. INTRODUCTION:

The most common mode of human communication is the oral mode. While man to machine communication in speech mode is yet to reach the level of large-scale implementation and use in practical systems, machine to man communication has attained a reasonable level for direct practical application. In the modern times, speech syntheses in various European languages are commercially available. However in Indian languages, speech synthesis is yet to attain the level for direct large-scale applications.

Speech synthesis 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.

For a text-to-speech (TTS) system two quality criteria are usually proposed. The first one is intelligibility, which can be measured by taking into account several kinds of units (phonemes, syllables, words, phrases). The second one, more difficult to define, is often labeled as *pleasantness* or *naturalness*. Actually the concept of naturalness may be related to the concept of realism: the goal is not to restitute the reality but to suggest it. Thus, listening to a synthetic voice must allow the listener to attribute this voice to some *pseudo-speaker* and to perceive some kind of expressivities as well as some indices characterizing the speaking style and the particular situation of elocution. For this purpose the corresponding extra-linguistic information must be supplied to the system.

Most of the present TTS systems produce an acceptable level of intelligibility, but the naturalness dimension, the ability to control expressivities, speech style and pseudo-speaker identity still are poorly mastered.

In the past few decades, in the area of Speech Synthesis different algorithms and methodologies have been developed, 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).

ESNOLA technique provides the complete control on implementation of intonation and prosody. 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 there is a possibility of its implementation in low-cost, general-purpose electronic devices.

C-DAC, Kolkata has developed a 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.

2. METHODOLOGY

Preprocessing Module

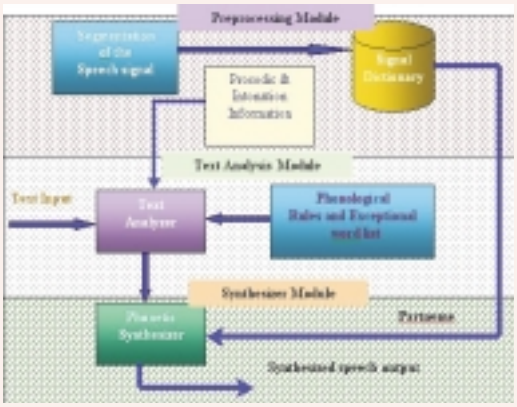

Figure 1. Basic Block Diagram of TTS System using ESNOLA Technique

The above figure-1 describes the basic Block diagram of ESNOLA technique based text-to speech synthesis system. It consists of there module A) Preprocessing module B) Text analysis module and C) Synthesizer module. The module A & B are language dependent modules and the module C is language independent.

2.1 Preprocessing module:

In this module the basic speech segment (pratneme) database known as signal dictionary is created from the pre-recorded natural speech signal for a particular language. The advantage of using partnemes as the basic unit is the simplicity of introducing intonation and prosodic rules into the synthesized speech signals.

For the building of signal dictionary the following steps are to be followed:

Step1. Creation of nonsense word set

This set of words must contain acoustic phonetic characteristics of all phonemes. A set of tetra-syllabic nonsense words of the forms CVCVCVCV, CVVCVVCVVCVVC is used for normal consonants and vowels. However, as /n/-/K/ distinction in case of Bangla is not ascertained except in conjunction with appropriate consonant, an additional 8 syllabic form CVNVCVNVCVNVCVNV for two nasals /n/ and /K/ are included. The choice of tetra syllabic words in case of Bangla is necessary because Bangla being a bound stress language with stress occurring normally at the first syllable (Annexure-I). Another set of words has to be collected from the normal lexicon of the language where the different vowel-vowel combinations occur. Usually all possible combination may not be easily available. For the unusual combinations appropriate sentences are created where such combination occur at word juncture.

Step2: Recording

The above-created set of words is to be uttered by a professional speaker with good voice quality in a noise free environment. The recorded words should be 'Stress Free'. Care must be taken to ensure that the pitch of the recorded word remains almost same throughout the recording. Recording format:- 16 bit PCM mono, sampling frequency 22050Hz.

Step3. Pitch Normalization

All signal segments should preferably brought to exactly same fundamental frequency. This is necessary to avoid pitch-mismatch. However adjustment of pitch by manipulation of sampling frequency may be used only when the pitch difference does not exceed 10% of the original value.

Step4. Amplitude Normalization

Amplitude normalization is performed with respect to the amplitude of vowels. It is known that the vowels of equal amplitude do not sound equally loud. This is known intrinsic loudness. For Bangla the values are:

Step5. Segmentation

Figure2 above shows how partnemes are extracted from the VCV nonsensical words. For plosives the partneme consists of occlusion and burst (C) and for affricate the friction after the plosion is also included. The co-articulation between the vowel and consonant include the voice-onset-time (VOT) and the consonant vowel transition (CV). Vowel to consonant transition begins at the end of the steady state of the vowels and at the beginning of the occlusion of the next consonant or any other consonant marker. Only one pitch period of the steady state of the vowel is taken to represent the vowel.

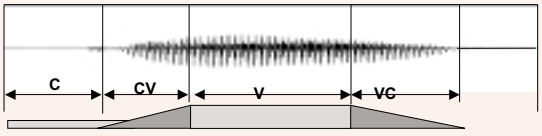

Figure 2: Shows the Segmentation procedure of the Recorded Voice Signal

2.2 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).

A preprocessing (or text normalization) module is necessary as a front-end, since the TTS systems should in principle be able to read any text, including numbers, abbreviations, acronyms and idiomatic, in any format. The preprocessor also performs the 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. The NLP module also includes a morpho-syntactic analyzer, which takes care of part-of-speech and phrase/clause

Tagging and organizes the input sentence into syntactically related groups of words. So the txt analysis module has two broad section one is the phonological analysis module and other is the analysis of the text for prosody and intonation. Bangla has a syllabic script. Grapheme-to-phoneme (phonological analysis) conversion is a formidable problem.

2.2.1 Phonological analysis

The study of phonology of Bangla has a long history since the early part of twentieth century. Since the pioneering work of Suniti Kumar Chatterji, a large number of eminent linguists of West Bengal and Bangladesh have contributed to the development of phonological rules of Bangla. The phonological problems are mainly found in the pronunciation of the two vowels /a/ and /e/ as well as a number of consonant clusters. Even the semantic and the parts of speech of a word sometimes play a significant role in pronunciation. The linguist has enumerated a large number of rules to tackle these problems. In most cases the rules have a large number of exceptions. It is very difficult to ascertain that the number of rules and exceptions are indeed exhaustive. Furthermore, the rules are generally given in verbose forms, which are not directly implementable in computer algorithms. It is therefore necessary to reduce the rules and exceptions in a form, which could be usable in an algorithm.

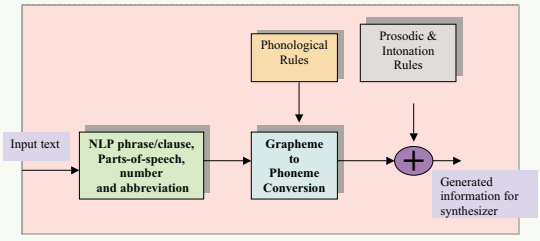

Figure-3. Block Diagram for the Linguistic Analysis of Inputted Text

The graphemes for consonants and consonant-clusters without ligature represent the syllable CV where the hidden V is . If one wants to represent a consonant without a vowel then one has to attach the sign . But some times this vowel is not pronounced and some times this vowel is pronounced as . For example a consonant cluster without ligature in word medial or word final position the hidden // is pronounced as /o/. There are some other rules for consonant cluster and the changing some vowel grapheme to other from. As Rules of [/y-ligature/], Rules of /dz??/, Rule of [/b-ligature/], Rule of [/m-ligature/] etc.

The above rules are implemented in software to make the phonological change of the input text. In some of the cases the above rules have an exception those exceptional words are store as a exceptional data base. The suffix problem is solved by storing some of the suffix as a list and subtract it first in search time. The most difficult problem is the deletion of vowel at the juncture point of compound word this problem is not yet solved. This problem can be shorted out by the root word analysis of the languages.

2.2.2 Prosodic and Intonational rules

The naturalness of the synthesized speech out put depends on the suprasegmental feature (prosodic and intonation feature) of the speech signal mainly pitch variation, syllabic duration variation, amplitude variation and pause. The implementation of the variation of suprasegmental feature in synthesized speech depend on the two factor one generation of intonation and prosodic rule along with the development of text parser for intonation and prosodic marking and the implementation of the suprasegmental feature variation in the synthesizer.

The intonation prosodic rule, which can be implemented, is yet not available for any Indian languages. This requires the study of the suprasegmental feature variation of the natural speech signal along with some linguistic references.

Intonation, stress, rhythm, duration etc. are the prosodic or suprasegmental elements, which give naturalness to the synthesized speech. Their physical correlates are fundamental frequency, segmental duration, energy, and to some extent complexity variation. In human speech, the prosody of an

utterance often depends not only on its words, but also on its intended meaning, its intended audience, the emotional or physical state of the speaker, and many other factors. Several of these factors are present even while reading a text, because humans generally interpret and understand the text that they are reading. Thus, it is likely that TTS systems will perform as nicely as humans only when it too can understand the input text, using some form of artificial intelligence. So, finding correct intonation, stress and duration from written text is one of the most challenging problems. Introduction of prosody in synthetic speech requires the determination of phrase boundaries and the segmental durations and the specification of fundamental frequency contours.

Method of intonation rule generation

The general tendency of the voice sound is to begin with a moderate pitch value and lower the medium pitch line during the sentence. This is going on up to a syntactic boundary, like phrase, clause or the end of the sentence. For phrases or clauses, the pitch value again reset to a moderate higher value and go on decreasing again following another declination line. This is the declination reset. So there is two types of variation of the F0 profile of any sentence one is local variation (word label) and other is global variation (sentence label).

It is difficult to measure declination reliably. Ideally, one would like to have the declination line, which could be a reference line to which local pitch movements can be scaled. Liberman et al calculated three different declination measures, by fitting linear regression lines to local peaks (top line), local valleys (base line) and all F0 points. In fig 2 the top line, base line and the line through all pitch values are drawn by the regression method. The line to be fitted has the equation,

$$y = a + bx \ldots \ldots \ldots (3)$$

where, $a = (\Sigma_i y_i - b\Sigma_i x_i)/N$, and

$b = (N\Sigma_i x_i y_i - \Sigma_i x_i \Sigma_i y_i)/(N\Sigma_i x_i^2 - (\Sigma_i x_i)^2)$.

Here N is the total number of pitch data for a syllable and 'b' indicating the steepness of the line as well as whether it is a rising intonation or a falling intonation.

The average pitch values are taken from the approximated pitch values, those are obtained by linear regression method, for each syllable in the sentence. A line is again fitted through the average points by linear regression method.

The syntactic boundary for a sentence is either the phrase or the clause or the sentence as a whole. Declination resets also break the sentence into such syntactic boundary. To find out intonation classes for a language it would be justified to find the class for parts of the sentences those can be classified by declination reset rule. The pitch movements within one declination line can be segmented at the word boundaries since words are the basic units of semantics. The intonation pattern within the declination is comprised of the intonation pattern of the individual word in it. Now the words contain a number of nucleus vowels, which represents a syllable. The syllables are taken as the basic unit of the local pitch movements. So, the total pitch movement within a declination can thus be seen as the sum of the pitch movement in the syllabic level. Classifying the intonation pattern in the syllabic level thus would be able to describe the intonation pattern of the part of a sentence classified by a declination line. The pitch movement for each syllable is represented by a straight line, which is fitted using linear regression method as described earlier. After fitting the straight line, we get the close copy syllabic stylization contour for the whole pitch movement within the declination.

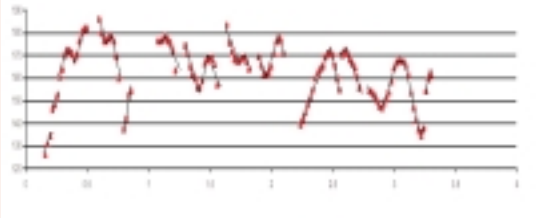

Figure 4 Close copy stylization of a Bangla sentence

Figure 4 show the close copy stylization of the Bengali sentence but if use the close copy stylization then information available for the local pitch variation is very high witch can be simplified by using syllabic stylization. The pitch movement for each syllable is approximated by a straight line, the syllabic intonation pattern will be either rising intonation or falling

intonation. As the words are comprised of a number of syllables, their intonation pattern would be a combination of series of rising and falling intonations. Figure 5 shows this close copy syllabic stylized intonation pattern for a Bengali sentence.

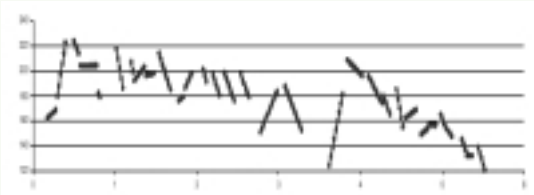

Figure-5 Syllabic Stylization of Bengali Sentences

Using the above method we can analyze the number of different types of sentences and generate a statistical intonational rule based and it can be implemented in the synthesizer. The durational and amplitude variation in syllable label can also be study for different types of sentences for generating the implementable rule.

For the above purpose a speech database have been developed for prosodic and intonational study of Bangla Language. The database consists of different types of sentences, which content most of the prosodic pattern of the Bangla.

2.3 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 text analysis module.

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 prosody and intonation. 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

Rules for Token generation:

CVCV → C +CV+V+VC+C+V+Vo

VCV → Vi+V +VC+C+CV+V+Vo

CVYV → C +CV+V+VY+YV+Vo

Vi → Fade-in Vowel Vo → Fade-out Vowel.

The fade in and fade out operation is applicable for the terminal vowels only.

Epoch

In ESNOLA approach the synthesized out put is generated by concatenating the basic signal segments from the signal dictionary. The epochs are most important for signal units, which represent vocalic or quasi-periodic sounds. An epoch position is represented in Figure 6.

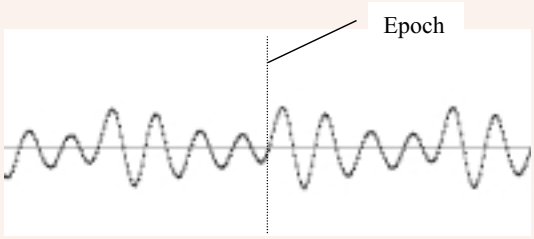

Figure 6 Epoch position of a speech segment

The basic segment are concatenate at the epoch position as in Figure 7

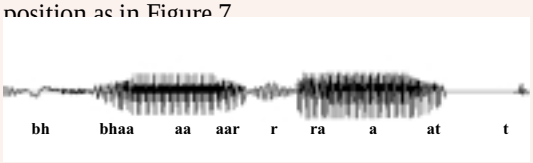

Figure 7. Synthesis of Voice Segments using ESNOLA technique

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 out put signal with out hampering the spectral quality. The equation of the window is as given below.

W(n) = 1/2(1-cos(pi*n/N); where 0<n<0.125N
= 1where 0.125N<n<0.625N
= 1/2(1+cos (pi*n/N); where 0.625N<n<N

2.3.1 Implementation of naturalness in synthesizer:

Intensity modification (amplitude modification): this done by multiplying each of the sample value of the segment by the value specified by Amplitude parameter of the corresponding code.

Duration modification: This operation in the present system is performed on steady state vowel segment. Repeat the vowel segment for M number of times, where duration of each repeating segments and the number M depend on the pitch contour pattern specified by the F0 parameter. The number M in this case is an integer.

F0 modification: Pitch (F0) modification of the synthesized signal is one of the important aspects to introduce intonation in the synthesized speech signal. In concatenative speech synthesis system partnemes consist of a) Consonant (C) b) Consonant to Vowel transition (CV) c) Vowel to Consonant transition (VC) d) Vowel- Vowel transition (VV). In our segment dictionary the signal whose pitch have to be modified are the CV, VC, VV, nasal murmurs and laterals. Pitch modification is done by the time scale modification of the each period. Time scale pitch modification is done by changing the length of the period of the original signal.

The basic principle of modifying pitch is to cut the required period T1 corresponding to the new pitch and suitably fade out the trail. If T1 is greater than the original period T of the signal then additional part is to be added. If T1 is less than T then cut the required period from the original signal with out loosing the quality of the original segment. Due to proper segmentation process each segment contain an integral number of periodic waveform.

Let y(n) be the segment whose pitch has to be changed according to some defined pitch profile information as given in the Figure 8.

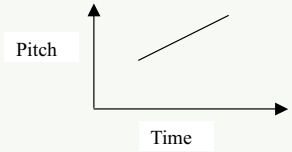

Figure 8 Syllabic pitch pattern

First we need to sub-segment each period of the segment with the help of epoch position identification as shown in Figure 9.

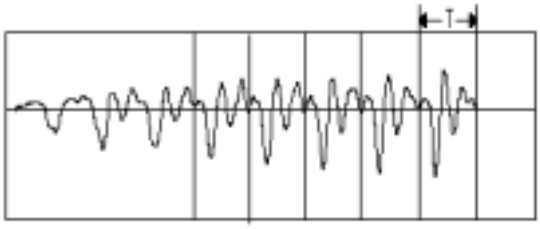

Figure 9. Sub-segmentation at epoch position

Let x(n) be each sub-segment and x(n) has N number of sampling point and its period is T. Also let the required period is T1. When T1 is greater than T, i.e., target pitch is lower than the original pitch, and the sub-segment x(n) is extended by making a intermediate signal such that $x_{int}(n)$ is the concatenation of signals x(n) and á x(n), where 0< á < 0.25. Thus we get a signal whose period is 2T having pitch half of the original and contains 2N numbers of sampling point. Now we define a window W(n) whose length is equal to the desired pitch period of the intermediate signal. The window characteristic is shown in Figure 10.

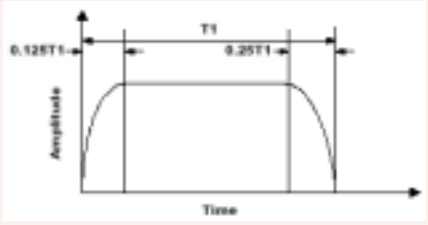

Figure 10 Window characteristics

Window characteristics can be expressed by the following formula,

W(n) = sin (π*t)/2*T1 where 0<t<0.125T1
= 1 0.125T1<t<0.75T1
= cos (π*t)/2*T1 0.75T1<t<T1

Now concatenating those changed pitch periods generate the required segment. This process creates a prominent striation and produces a perceptible mechanical horn like sound over and above the normal quality of the voice. This is because such concatenation produces exactly periodic wave instead

of quasi-periodic ones. Normal human voice is not perfectly periodic. Two successive pitch cycles do not produce exactly the same pressure waves. The variations are random in nature and occur for pitch, amplitude and complexity, which are referred to as jitter, shimmer and complexity perturbations respectively. An optimum value of these produces natural sound. An excess of the perturbations makes the quality of sound rough or hoarse. Absence of these perturbations again produces an unnatural horn like sound. Addition of jitter and complexity perturbation almost removes the defect. A random variation of 2-3% in pitch period is introduced for jitter by introducing appropriate modification of T1. the complexity perturbation is introduce by randomly varying the sample value by ±1% .

3. CONCLUSION

Speech Synthesis is assuming importance in the Indian IT scenario with immediate application potential. But the signal processing issues have not been perfected. To day most of the present TTS systems produce an acceptable level of intelligibility, but the naturalness dimension, the ability to control expressivities, speech style and pseudo-speaker identity still are poorly mastered. Let us mention however that users demands vary to a large extent according to the field of application: general public applications such as telephonic information retrieval need maximal realism and naturalness, whereas some applications involving professionals (process or vehicle control) or highly motivated persons (visually impaired, applications in hostile environments) demand intelligibility with the highest priority.

REFERENCE

1. Datta A.K, Ganguli N.R and Mukherjee B. “Intonation in segment concatenated speech” Proc. ESCA Workshop on speech synthesis, France, pp.-153-156, Sep 1990.
2. Dan T. K and Datta A. K. “Speech synthesis: a time domain model”. Proc. International Workshop on recent trends in speech, music and allied signal processing, 9-11 Dec, 1991, Delhi, India, pp. 39-45.
3. Datta A. K, Ganguly N R, Mukherjee B. “Intonation in segment-concatenated-speech”. Proc. ESCA Workshop on speech synthesis, Sep 1990, France, pp. 153-156.
4. Dan, T. K., Mukherjee B & Datta A. K. (1993). “Temporal approach for synthesis of singing (Soprano1).” SMAC 93, pp. 282-287, 1993.
5. Das Mandal S. K, Datta A.K, Gupta B. “Spectral Matching of Epoch Synchronous Non-Over Lapping Add (ESNPLA) Method Based Concatenative Synthesizer”, *International Conference on Communications Devices and Intelligent System (CODIS-2004)*, Jadavpur University, 2004, pp 729-732.
6. Chowdhary Soumen, Datta A. K. and Chowdhary B.B, “Multilingual TTS System in Indian context”, Proceeding of IWSMSP-2001,pp 56-64.
7. Chatterji Suniti Kumar (1926) “The Original and Development of the Bengali Language” *Published by Rupa.Co, 2002*, ISBN 81-7167-117-9.
8. Datta A.K, Ganguli N.R and Mukherjee B. “ Intonation in segment concatenated speech” Proc. ESCA Workshop on speech synthesis, France, pp.-153-156, Sep 1990.
9. Sarkar Pabitra, “Bangla Balo” *Published by Prama prakasani1990*
10. S. Deketelaere, O. Deroo, T. Dutoit, “Speech Processing for Communications: What’s New?”(*) MULTITEL ASBL, 1 Copernic Ave, Initialis Scientific Park, B-7000 MONS(**) Faculté Polytechnique de Mons, TCTS Lab, 1 Copernic Ave, Initialis Scientific Park, B-7000 MONS
11. P.H. Low, S. Vaseghi, “Synthesis Of Unseen Context And Spectral And Pitch Contour Smoothing In Concatenated Text To Speech Synthesis”, ICASSP, Florida, USA, Vol. 1, pp. 469-472, 2002