

# Analysis and Synthesis of $F_0$ Contours for Bangla Readout Speech

*Shyamal Das Mandal*  
CDAC, Kolkata  
*shyamal.dasmandal@cdackolkata.in*

*Anal Haque Warsi*  
CDAC, Kolkata  
*anal.warsi@cdackolkata.in*

*Tulika Basu*  
CDAC, Kolkata  
*tulika.basu@cdackolkata.in*

*Keikichi Hirose*  
University of Tokyo  
*hirose@gavo.t.u-tokyo.ac.jp*

*Hiroya Fujisaki*  
University of Tokyo  
*fujisaki@alum.mit.edu*

## Abstract

*It is well known that the  $F_0$  contour plays an important role in conveying prosodic information, but the process of synthesizing the  $F_0$  contour from the underlying linguistic information has not been elucidated for Bangla. This paper defines the prosodic units of Bangla on the basis of  $F_0$  contour analysis using the command-response model by Fujisaki et al. For the study, 200 Bangla declarative sentences spoken in the readout mode are analyzed. Based on the analysis, rules are constructed for predicting both phrase and accent command parameters of the model for generating the  $F_0$  contours of Bangla readout speech. A perceptual evaluation test for the naturalness of prosody shows that there is no significant difference between synthetic speech with model-generated  $F_0$  contours and the original natural speech.*

**Index Terms**  $F_0$  contour, the command-response model, prosodic units, Bangla readout speech.

## 1. Introduction

In India, speech synthesis is considered to be of primary need to empower not only disabled people, but also the functionally illiterate population. Various speech synthesis systems are now appearing for some of the major Indian languages. There is, however, no synthesis system for Bangla that can handle unlimited vocabulary and at the same time can produce natural sounding speech.

Prosody plays an important role in bestowing both intelligibility and naturalness in synthesised speech. In addition to its important function in communicating linguistic information concerning word meaning, sentence structure and discourse structure, prosody is also important in the transmission of para-linguistic and non-linguistic information such as speaker's intention, emotion and idiosyncrasy [1]. Therefore, the technique

of utilizing knowledge of prosody is extremely important in machine-generated speech.

Since the naturalness of the synthesized speech output of a text-to-speech (TTS) system depends predominantly on its prosodic characteristics, it is necessary to construct models or rules which well-relate the prosodic characteristics of output speech and the linguistic content of the text [1]. In the case of data-driven synthesis, the bridge between linguistic and prosodic information for the input text comes solely from the speech database and needs to be acquired during the building-up of the speech database. Hence, building of any unlimited vocabulary TTS system using such a method may demand a huge coverage of training speech corpus. Therefore, an effective alternative will be to have an exclusive prosody model for the concerned language, as a control layer on top of the wave concatenative synthesis system.

The contour of the voice fundamental frequency ( $F_0$ ) plays an important role in expressing prosodic information of an utterance. An  $F_0$  contour generally consists of slowly-varying components corresponding to phrases and clauses and rapidly-varying components corresponding to word accents or syllable tones. Several rule-based and corpus-based methods were developed for the generation of  $F_0$  contours for many of the foreign languages [2] as well as for some of the Indian languages [3].

Among them the Hidden Markov Model (HMM) is now commonly used for speech synthesis for many languages. However it needs a large training corpus to maintain the speech quality. This method handles the generation of  $F_0$  contour in a frame-by-frame manner, which is inadequate for the generation of supra-segmental features like  $F_0$ . For synthesis of larger units like phrases and sentences, a model based on the generation process of  $F_0$  contours has been shown to produce better results [4]. Such a model, also called the

command-response model, has been developed by Fujisaki and his coworkers, firstly for common Japanese [4, 5], and has since been successfully applied to many languages of the world [6]. The model expresses the contour of logarithm of  $F_0$  as the sum of three types of elements: the phrase components, the accent components, and a baseline. The exact relationship between these components of an  $F_0$  contour and the underlying linguistic information have also been formulated for common Japanese [7]. It has been widely shown that the command-response model can generate very close approximations to observed  $F_0$  contour from a relatively small number of parameters representing the linguistic information, and is therefore quite useful in speech synthesis.

In Bangla, tone is not phonemically significant which means changing the tone of the word doesn't change its meaning. In declarative sentences most words in Bangla is said to carry a rising tone with the exception of the last word of the utterance, which carries only a low tone [8].

On the other hand, intonation plays an important role in Bangla. In declarative sentences with neutral focus, the intonation pattern is falling. In sentences involving focused words or phrases, the rising pattern lasts until the right edge of the focused word; the remaining portion of the sentence carries a falling pattern. WH sentences follow the same intonation pattern as the sentences involving focused words, but in Yes-No sentences the intonation pattern is rising [8].

This paper presents some of the most recent results of our investigation to elucidate the intonation pattern of Bangla declarative sentences. It describes a detailed analysis of  $F_0$  contours of Bangla based on the command-response model, and derives a set of rules for predicting the command parameters of the model from linguistic information of the text, with an aim of developing an intonated TTS system for Bangla. .

## 2. Speech material

Bangla is a part of the Indic group of the Indo-Aryan (IA) branch of the Indo-European family of languages. It is the official state language of the Eastern Indian state of West Bengal and the national language of Bangladesh. With nearly 230 million total speakers, Bangla is one of the most spoken languages (ranking fifth or sixth) in the world. The present study is based on the official dialect of West Bengal, i.e. Standard Colloquial Bengali (SCB) [9].

The speech material used for the present study is the readout speech of 200 declarative sentences of Bangla spoken by one female speaker, whose

utterances are also used for the development of an ESNOLA-based Bangla TTS system [10]. The age of the speaker is 27 years and her speech rate is 6 syllables/sec. The speech is recorded in a studio environment and digitized at a sampling rate of 22,050 Hz with an accuracy of 16 bits/sample. These 200 sentences were randomly selected from a large (10,000 sentences) Bangla text corpora.

The model parameters are initially extracted using a method for automatic extraction of  $F_0$  contour model parameters, and the results are manually corrected.

## 3. Methods for $F_0$ contour generation of Bangla readout text

### 3.1. $F_0$ contour model

The  $F_0$  contour model is a command-response model that describes an  $F_0$  contour in the logarithmic scale as the superposition of a baseline value, phrase components, and accent components as given by Equation (1) [6].

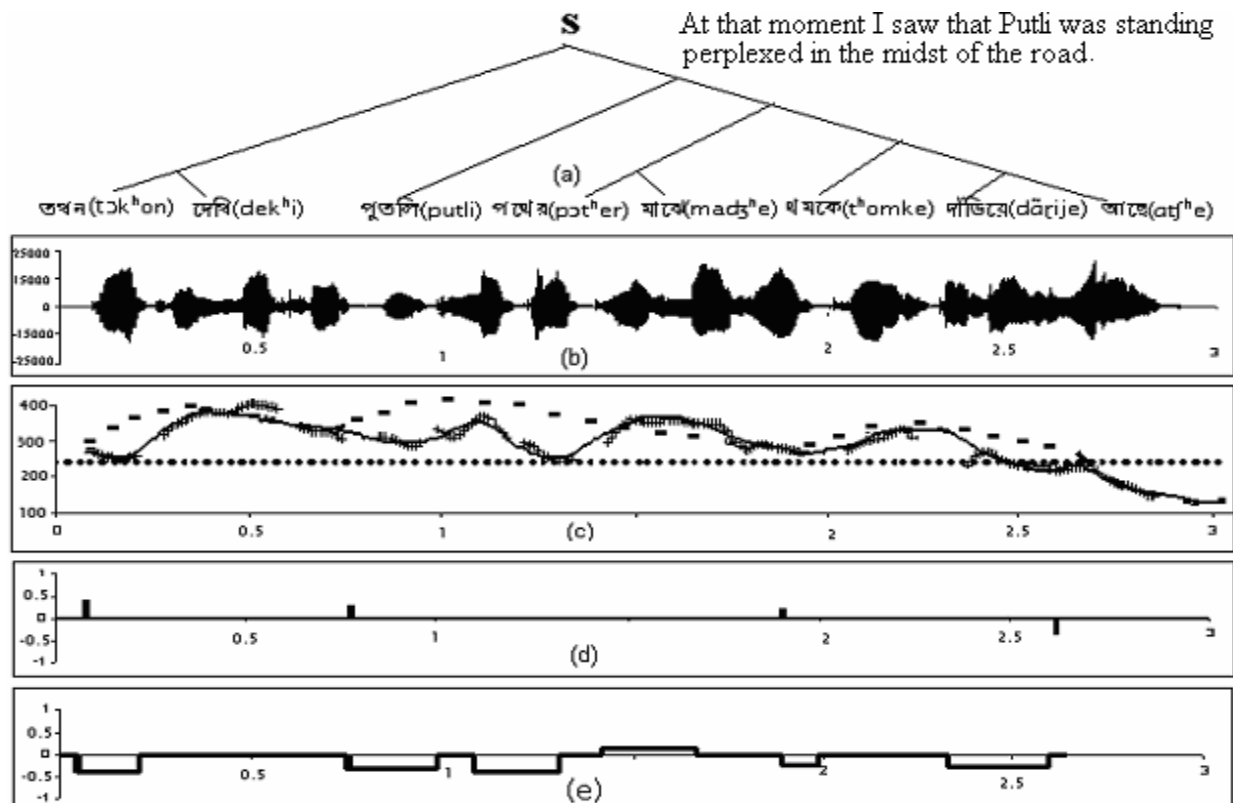
$$F_0(t) = \ln(F_b) + \sum_{i=1}^I A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^J A_{aj} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \quad (1)$$

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (2)$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t) \exp(-\beta t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (3)$$

Equation (2) represents the second-order, critically damped linear filter in response to an impulse called phrase command, and Equation (3) represents the second-order, critically damped linear filter in response to a step function called accent command. The parameters  $\alpha$  and  $\beta$  are the time constants for the phrase and accent control mechanisms respectively, and  $\gamma$  is the ceiling parameter. Since these parameters are tightly related to the mechanical system of the speaker's larynx they are considered to be constant for all the utterances. For the analysis of  $F_0$  contours of Bangla their values are fixed at 3.0, 20 and 0.9 respectively.

In the above equation (1),  $F_b$  is the baseline frequency and its value is around 233 Hz for the present utterances.  $I$  and  $J$  are the numbers of phrase and accent commands, respectively.  $A_{pi}$  is the magnitude of the  $i$ th phrase command and  $A_{aj}$  is the amplitude of the  $j$ th accent command.  $T_{0i}$  is the time initiation of the  $i$ th phrase command, while  $T_{1j}$  and  $T_{2j}$  are the onset and offset times of the  $j$ th accent command.



**Figure 1. A Bangla sentence and the analysis-by-synthesis of its  $F_0$  contour. (a) the syntactic tree, (b) speech waveform, (c) measured  $F_0$  values (+ symbols), model-generated best approximations (solid line), the phrase component (dashed line), the baseline frequency (dotted line), (d) phrase command position and its magnitude, (e) accent command position and its amplitude.**

### 3.2. Relationship between $F_0$ model parameters and linguistic information

Figure 1 shows an example of the analysis-by-synthesis of the  $F_0$  contour of the Bangla declarative sentence ‘তখন দেখি পুতলি পথের মাঝে থমকে দাঁড়িয়ে আছে’ (At that moment I saw that Putli was standing perplexed in the midst of the road).

The figure shows, from top to bottom, the syntactic tree, the speech waveform, measured  $F_0$  values (+symbols), model-generated best approximations (solid line), the phrase component(dashed line), the baseline frequency (dotted line), the phrase commands (impulses), and the accent commands (pedestal functions). The dashed lines indicate the contributions of the phrase components, and the differences between the  $F_0$  contour and the phrase components correspond to the accent components.

Analysis of the utterances of 200 Bangla declarative sentences indicates that the model can always generate very good approximations to the measured  $F_0$  contours,

if the timing and magnitude/amplitude of the commands are optimized.

It is observed from the study that the placement of phrase commands for the readout speech of Bangla declarative sentences is related to the syntactic structure, but they occur mostly at deeper syntactic boundaries. Following a prior work on Japanese [7], the prosodic phrase of Bangla can be defined on the basis of the phrase command. The prosodic phrase may consist of more than one syntactic phrase as shown in the example in Figure 1. The placement of the phrase command also depends on the length of the phrase. There is a certain limit on the distance between two phrase commands. When an utterance has more than one phrase commands, their magnitudes have a general tendency of decreasing from the start of an utterance towards its end. It is to be noted that the initiation of the phrase command always leads the segmental onset of the corresponding prosodic phrase by a few hundred milliseconds.

In the same vein, the prosodic word of Bangla can be defined on the basis of the accent commands. Our

analysis shows that most of the word-level variations of  $F_0$  in Bangla can be interpreted by a single negative accent command at the beginning of a word. Sometimes a group of words is pronounced as if they were one word. In such a case, these words are considered to constitute a single prosodic word. For example, the words /tɔk<sup>h</sup>on/ and /dek<sup>h</sup>i/ form a prosodic word in Figure 1. The syntactic constituents of a prosodic word may be a sequence of one or more content words followed by one or more function words sharing an  $F_0$  contour of a single word accent type [7]. Analysis of the current data shows the existence of two accent types in Bangla readout speech. Accent Type 1 contains only one negative accent command at the beginning, while Accent Type 2 contains one negative accent command at the beginning followed by a positive accent command towards the end. It is also to be noted that the onset of the negative accent command may lead the segmental onset of the corresponding prosodic word by a few hundred milliseconds. The rule for the formation of prosodic word from a given text is described in section 3.3.

### 3.3. Rules for formation of prosodic words

Since the purpose of the study is the implementation in TTS, the goal is to generate rules for prosodic word formation from textual information. From the  $F_0$  contours of 200 Bangla sentences the following rules are formulated on the basis of Parts-of-Speech (POS) information of the constituent words. These rules are tested over a set of 100 new sentences, and the accuracy of correct predictions of prosodic words is 85%.

**Rule 1:** Hyphenated words and repeated words always form a prosodic word.

**Rule 2:** Two consecutive proper nouns within the same prosodic phrase form a prosodic word.

**Rule 3:** If a common noun (length  $\leq 3$  syllables) is preceded by an adjective (length  $\leq 3$  syllables) then they are combined together to form a prosodic word.

**Rule 4:** A common noun and a verbal noun join together to form a prosodic word.

**Rule 5:** An arbitrary word followed by a postposition together forms a prosodic word.

**Rule 6:** A verb (main or auxiliary) and the following particle together form a prosodic word.

**Rule 7:** A main verb and the following auxiliary verb (viz., a compound verb) combine together to form a prosodic word.

**Rule 8:** A common noun (or an adjective or a verbal noun) and a verb form a prosodic word.

### 3.4. Synthesis of phrase components

The extracted phrase components of the 200 Bangla utterances are analyzed to determine rules for the prediction of phrase command parameters, viz., magnitude and position. Table 1 shows the mean and the standard deviation of the command magnitude at various positions in the utterance. The table shows that the magnitude of the phrase command is largest at the utterance-initial position, and decreases towards the end of the utterance. For the purpose of speech synthesis, the mean magnitude of each phrase command can be used since the standard deviation is very low. It is also observed that if a phrase command is preceded by a pause of greater than 100ms then the phrase command magnitude will be increased by 15% of its mean value. In addition to this, a negative phrase command may be inserted to represent the gradual (or sometimes rapid) downfall of  $F_0$  towards the end of the utterance. This negative phrase command has nothing to do with phrasing or with word accent, but approximates the process of relaxation of the vocal folds.

**Table 1. Phrase command magnitude.**

Phrase position	Command magnitude ( $A_{pi}$ )	
	Mean	Standard deviation
Utterance-initial	0.327	0.009
2nd phrase	0.234	0.03
3rd phrase	0.185	0.03
$\geq 4$ th phrase	0.161	0.02
Utterance final	-0.221	0.06

Table 2 shows the phrase command lead-time from the segmental onset time of each prosodic phrase. It is evident from the study that the lead-time is directly proportional to the length of the preceding pause. There is no significant relationship between the lead-time of the phrase commands and other factors such as the length of the phrase, syntactic type of the phrase

**Table 2. Phrase command lead-time [in sec].**

Phrase position	Lead time		
	Pause > 0.25s	0.1s $\leq$ Pause $\leq$ 0.25s	Others
Utterance-initial	-	-	0.227
2nd phrase	0.371	0.262	0.186
3rd phrase	0.334	0.260	0.182
$\geq 4$ th phrase	0.361	0.258	0.187
Utterance final	-	-	0.150

and the segmental properties of the phrase-initial syllable. For the generation of  $F_0$  contours the mean values given in Table 2 are used for the lead times of the phrase commands at their respective positions. It is observed from the study that placement of the phrase command depends on the syntactic type and the length of the prosodic phrase in terms of number of syllable. Based on the analysis, the following rules have been formulated for the generation of phrase components. The magnitude and the lead-time of the command are selected by referring to Table 1 and Table 2 respectively.

**Rule 1:** A phrase command is always placed at the utterance-initial position.

**Rule 2:** If a syntactic phrase is preceded by a pause of greater than 100 ms, then a new phrase command is placed.

**Rule 3:** If the previous syntactic phrase is of length greater than three syllables and the current syntactic phrase is of length less than three syllables and the phrase type is other than VP (Verb Phrase), then a new phrase command is placed at the beginning of the current syntactic phrase. If it is a VP, then it is merged with the previous phrase.

**Rule 4:** If the current syntactic phrase is preceded by a syntactic phrase of length less than three syllables, then the current phrase is merged with the previous phrase.

**Rule 5:** If the resulting prosodic phrase is of length greater than six syllables then it is split in such a way that none of the resulting phrases are bi-syllabic.

### 3.5. Synthesis of accent components

There are three important parameters for the generation of the accent components, *viz.*, accent command amplitude,  $t_1$ , and  $t_2$ , where  $t_1$  is the distance between the onset time of the syllable and the onset time of the accent command, while  $t_2$  is the distance between the offset time of the syllable and the offset time of the accent command. It is to be noted that  $t_1$  and  $t_2$  are different from  $T_{1j}$  and  $T_{2j}$  used in Eq. (1) though they are closely related.

Based on the analysis of about 2000 words taken from utterances of 200 Bangla declarative sentences it is observed that 91% of the prosodic words in Bangla have only one negative accent command (Accent Type-1), while the remaining 9% of the prosodic words have one negative accent command followed by one positive accent command (Accent Type 2). It is also observed that Accent Type 2 occurs only in the case of polysyllabic words that are emphasized or that contain suffix in prosodic word-medial positions. The

amplitude of the accent command is dependent only on the position of the prosodic word within an utterance.

**Table 3. Accent command amplitude.**

Position		Accent command amplitude ( $A_{aj}$ )	
		Negative	Positive
Utterance-initial		-0.325 ( $\sigma = 0.06$ )	0.232 ( $\sigma = 0.05$ )
Utterance Medial	Phrase initial	-0.317 ( $\sigma = 0.05$ )	
	Phrase medial	-0.261 ( $\sigma = 0.08$ )	
Utterance final		-0.328 ( $\sigma = 0.09$ )	-

From the analysis of  $t_1$  and  $t_2$  it is observed that  $t_1$  is dependent upon the position of the prosodic word in the utterance, presence/absence of a pause and presence/absence of voicing at the syllable-initial position.  $t_2$ , on the other hand, depends on the length of the prosodic word and type of the initial syllable.

**Table 4. Value of  $t_1$  [in sec].**

Position	$t_1$		
Utterance-initial	0.148		
Utterance-medial	Pause $\geq 0.10$ s	0.147	
	Pause $< 0.10$ s	Voiced	0.071
		Unvoiced	0.111

**Table 5. Value of  $t_2$  [in sec].**

Word Length	$t_2$	
Monosyllabic	0.176	
Polysyllabic	Syllable Type- 'V', 'VC', 'CV'	0.037
	Syllable Type-'others'	0.071

Thus the following steps are required for the generation of accent components.

- Identification of each prosodic word.
- Determination of the accent type on the basis of the position and length (number of syllables).
- Selection of appropriate amplitude by referring to Table 3.
- Selection of  $t_1$  and  $t_2$  values from Table 4 and Table 5 respectively.

### 4. Evaluation of Synthesis Results

The synthesized  $F_0$  contour is evaluated using two methods: a) the objective method, and b) the subjective

method. Another set of 50 Bangla declarative sentences (not included in the analysis data) is used for the evaluation.

In the objective evaluation, the original  $F_0$  contour is compared with the synthesized  $F_0$  contour and the root mean squared value of the difference is calculated in  $\log_e F_0$ . The root mean squared difference per sample of the synthesized and original contour for the above 50 sentences is 0.041.

In the subjective evaluation, the same set of test sentences is used. For the evaluation 5 subjects - 2 males (L1, L2) and 3 females (L3, L4, and L5) - are selected. All subjects are native speakers of Standard Colloquial Bangla and are not speech experts. In this experiment, two sets of the above 50 test sentences, one having the model-generated  $F_0$  contour and the other with the original  $F_0$  contour are randomly mixed and presented to the subjects for giving their judgment on the naturalness of the sentence on a 5-point scale (5: very good, 4: good, 3: neutral, 2: poor, and 1: unacceptable).

**Table 5. Mean opinion scores for the original and the synthesized speech.**

Score		Subject				
		L1	L2	L3	L4	L5
Original	Avg	4.64	4.52	4.58	4.70	4.62
	Stdv	0.56	0.58	0.67	0.54	0.49
Model generated	Avg	4.52	4.44	4.52	4.64	4.56
	Stdv	0.71	0.68	0.71	0.62	0.67

Table 5 shows the mean opinion scores for all sentences of each subject for the original and the synthesized sentences. The scores are analyzed using ANOVA (single factor test). It is observed that the calculated F-value is much smaller than the  $F_{crit}$  value at all the levels. Thus, it is evident that the difference in naturalness between the original and model generated sentences is not significant at any level.

## 5. Conclusion

The  $F_0$  contours of 200 Bangla declarative sentences read by a good native speaker are analyzed on the basis of the command-response model to extract phrase and accent commands, and the results are used to define prosodic units of Bangla read-out speech. Rules are then formulated to identify the prosodic phrases and prosodic words from the text and to assign appropriate parameter values to their respective commands for use in TTS system. A subjective evaluation test of synthesized speech shows that the rules for phrase command generation are quite satisfactory, while those for accent command

generation still need further improvement. Further work is in progress to obtain and utilize information from the prosody of the speaking style that is not necessarily derivable from the text.

## 6. References

- [1] Fujisaki, H., "Prosody, Models, and Spontaneous Speech", In *Computing Prosody*, (Y. Sagisaka, N. Campbell, N. Higuchi, eds.), New York: Springer-Verlag, pp. 27-42, 1996.
- [2] Fujisaki, H., Hirose, K., Hallé, P. and Lei, H., "Analysis and Modeling of Tonal Features in Polysyllabic Words and Sentences of the Standard Chinese", *Proceedings of 1990 International Conference on Spoken Language Processing*, vol. 2, pp. 841-844, 1990.
- [3] Sreenivas, K. and Yegnanarayana, B., "Intonation Modeling for Indian Languages", *Journal of Computer Speech and Language*, Volume 23, Issue 2, Pages: 240-256, 2009.
- [4] Fujisaki, H. and Nagashima, S., "A Model for the Synthesis of Pitch Contours of Connected Speech", *Annual Report of the Engineering Research Institute*, University of Tokyo, vol. 28, pp.53-60, 1969.
- [5] Fujisaki, H. and Hirose, K., "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese", *J. Acoust. Soc. Japan* (E), vol.5, no.4, pp.233-242, 1984.
- [6] Fujisaki, H. "Information, Prosody, and Modeling", *Proceedings of Speech Prosody 2004*, Nara, Japan, pp.1-10, 2004.
- [7] Fujisaki, H., Hirose, K. and Takahashi, N., "Manifestation of Linguistic Information in the Voice Fundamental Frequency Contours of Spoken Japanese", *IEICE Trans, Fundamentals*, vol. E76-A, No. 11, pp. 1919-1926, 1993.
- [8] Hayes, B. and Lahiri, A., "Bengali Intonational Phonology", *Natural Language and Linguistic Theory*, Springer Science, pp 56-58, 1991.
- [9] Bhattacharya, K., "Bengali Phonetic Reader" published by Central Institute of Indian Languages, 1999.
- [10] Das Mandal, Shyamal Kr., Datta, A. K., "Epoch Synchronous Non-OverLapping Add (ESNOLA) Method Based Concatenative Synthesis System for Bangla", *Proc. of 6th ISCA Workshop on Speech Synthesis*, University of Bonn, Germany, pp. 351-355, 2007.