

Chapter 2

Epoch Synchronous Overlap Add (Esola) Algorithm

2.1 Introduction

Concatenative speech synthesis, the recent trend for synthesizing speech, uses real speech signal units for constructing utterances. Since speech is a continuous process one has to keep in mind the dynamics of the continuity so that it is not jeopardised at the junction points. It implies that constant watch has to be maintained on the dynamics of the three fundamental properties of sound namely pitch, amplitude and timbre. A concatenative TTS (Text-To-Speech) synthesis system produces synthesized speech, as specified by the input text, by conjoining the stored speech units. The choice of signal unit set for any concatenative synthesizer is the cornerstone for producing a good synthesized output, for retaining the natural sounding quality after prosodic modification, and for implementation in a portable system. Since the advent of concatenative synthesis techniques for unlimited vocabulary, several kinds of synthesis units, such as, diphone, syllable, demi-syllable, phone, CV (Consonant-Vowel) sequence, VCV (Vowel-Consonant-Vowel) sequence, CVC (Consonant-Vowel-Consonant) sequence, and tri-phone (context dependent phones) (Brownman 1980; Courbon et al. 1982; Dixon et al. 1968; Fries 1993; Hakoda et al. 1995; Saito et al. 1968; Santen et al. 1997; Tohkura et al. 1989), have been used. These choices seem to be dictated by the demand of naturalness, size of the signal dictionary, ease and extent of manipulation and the domain of use. As a general rule the naturalness and the size of the signal dictionary increase with the increase in the length of signal units. On the other hand ease and extent of manipulation seem to be better when the lengths of the units are small. One of the challenges for the development of a synthesizer is to improve naturalness of acoustic quality in concatenative synthesis using small units, at least to the level of, if not better than, those using larger

units. In any TTS system, the stored signal units and the prosody modification algorithm are the most important factors to determine the quality of the synthetic speech produced (Stylianou et al. 1995). In conventional concatenative systems, insufficient variations of speech unit cause artificial sounding speech.

This chapter presents a set of new speech synthesis units together with appropriate prosody modification procedures. Prosody modification is basically a pitch, amplitude and duration modification algorithms of speech. These contribute to develop a high-quality TTS system for SCB (Standard Colloquial Bengali). The core and one of the novelties of the new concatenative TTS (Text-To-Speech) system for SCB (Standard Colloquial Bengali) presented in this book is the new signal unit ‘partnemes’ (Chowdhury 2002; Chowdhury et al. 2000b) at the sub-phonetic level. The partnemes i.e., part of the phones, which are the smallest units, so far, being used as the signal units for the concatenation. A partneme is in no way meant to be associated with the meaning of a word of which it is a part.

The ESOLA (Epoch Synchronous Overlapping Add) algorithm is primarily developed for concatenation, regeneration as well as for pitch and duration (prosodic) modification. It may be noted that the prosody of the stored units is often not consistent with that of the target utterance and must be altered at the time of synthesis. Furthermore, several types of mismatches can occur at unit boundaries of the synthesized signal, which have to be properly truncated and matched. The problems related to combining signal units (such as prosody control, spectral mismatch) for producing natural speech output are analyzed and appropriate solutions are given in this chapter. ESOLA (Epoch Synchronous Overlap Add) technique is shown to preserve phonetic quality even when pitch is modified by an octave. It is reported that the speaker’s identity is also preserved in such level of F_0 modification. The different operations of concatenation for producing unlimited set of proper utterances in Standard Colloquial Bengali (Bangla) are also included. Listening tests confirming that the new synthesis units yield synthetic speech with high intelligibility and naturalness are presented in Chap. 5. The advantages of a partneme-based synthesizer using the epoch synchronous method are also discussed. In this chapter we have given the full Bengali phone set along with their IPA symbols. Thus, this chapter is devoted to the basic design of the TTS system based on ESOLA technique along with the description of the different units of the synthesizer system and their interdependencies.

Before detailing the development of ESOLA a critical review of the recent popular relevant synthesis system may be in order. The different types of signal units have their own merits and demerits. Among these units, the sizes of databases for diphones, syllables or demi-syllables are uncomfortably large for unlimited speech. A well-known synthesis technique, which is based on pitch synchronous waveform processing and uses diphones as synthesis units, is TD-PSOLA (Moulines et al. 1990). Although TD-PSOLA provides good quality speech synthesis, its limitation lies in spectral mismatch at segmental boundaries and tonal

quality degradation when prosodic modifications are applied on the acoustic units. This technique has a relatively narrow range of prosody modification wherein naturalness is retained; speech distortion is evident if prosody modification is large. MBROLA (Dutoit et al. 1993) tries to overcome these concatenation problems by re-synthesizing voiced parts with constant phase and constant pitch. This artificial processing produces buzz in the output signals. Some sinusoidal models (Crespo et al. 1996) perform concatenation by making use of glottal closure instants. Often what they care most about is a very precise estimate of pitch. In some systems the quality of the output signal is poor because of phase mismatch at segment boundaries. But some more successful systems do not have problems with phase mismatch at segment boundaries.

The current chapter presents the core of the concatenative speech synthesis system and a time scale modification technique for the SCB (Standard Colloquial Bengali) using ESOLA technique. ESOLA algorithm is based on the result that the phonetic quality including speaker's identity remains almost intact in case of sonorants if the first part of the signal corresponding to that of the glottal period (Dasmandal et al. 2003) (i.e., the signal starting from the epoch position) is retained. This chapter presents the ESOLA framework, its mathematical analysis and analytical results, in detail. The primary need in building the segment dictionary for concatenative synthesis is to record natural speech so that all used units in all possible contexts (allophones) are included. We have also presented the speech signal inventory used in the speech synthesis, their organization and methods for their preparation. The discussion on the supremacy of the ESOLA over the existing concatenative synthesizers is also presented in a section.

The core concatenative approach, ESOLA, is described in this chapter along with a mathematical analysis of it. The pitch modifier, an essential element, is described in this chapter. Experimental results are presented showing that the pitch modification technique keeps the spectrum almost identical to the original one for up to \pm one octave. Listening tests show the clarity and naturalness of the synthesized output speech. The main block diagram of the proposed speech synthesis system is given and described in this chapter. Basically the total speech synthesis system is the hybridization of the two separate units, one is the text pre-processing and corresponding rule base generation unit and other one is the low-level synthesizer unit, i.e., the actual synthesis unit. In the low-level synthesis unit, the speech is produced by taking the phone string along with information about intonation and prosody as input. The aforesaid information comes from the other unit, which is the high level part of the synthesizer. The sub units that build up the two units are also described in detail. A syllable-breaking algorithm is included in the chapter. The details of the partname dictionary, how it is to be built up from the recorded signal is described here. I have also described here the recording process and what should be the utterances from where the signal dictionary is to be prepared. The method for amplitude normalization and pitch normalization of the signal unit is also described here.

2.2 Basic Principles of ESOLA

A concatenative TTS system may generally be said to have two basic units, (a) A language processing unit for the input text analysis at the front end and (b) A signal processing unit, that takes care of proper concatenations of the basic units and at the same time modifying them, if necessary, to incorporate pitch, duration and amplitude changes. Both the parts are important for the production of synthesized speech of good quality.

At the very outset the text to be synthesized is first pre-processed. Text pre-processing is a complex language dependent problem (Sproat 1997). The pre-processing deals with the numerals, abbreviations and acronyms present in the text. These are converted into text form. In Bengali texts, the uses of abbreviations or acronyms are not abundant. The abbreviations are generally followed by a ‘dot’ and rarely followed by a colon. Since in Bengali, dots are not used for punctuation its presence is a robust indicator that the preceding group of characters is an abbreviation. The conversion to the text then is a simple look up operation in the table consisting of the abbreviations and corresponding full forms. The other problem in pre-processing is with numerals. These are not normally read digit by digit. Instead they are generally converted into a corresponding set of words, e.g., the sequence of digits (1,25,336) will be read as /æk lokkho potʃiʃ hədʒertinso tʃhottriʃ/. The commas in the digit strings indicate different units like /lokkho/ (lakh), / hədʒer / (thousand), /so/ (hundred) etc. The scanning and conversion of these are very simple.

The next step is to convert the input grapheme string into the corresponding linguistic or phone representations. In a language, the grapheme form of a word does not always follow exactly the indicated phone at the time of its pronunciation. These mappings are not only language specific but also depend on the dialects present in the particular language. These deviations from the standard pronunciation of a word are guided by the phonological rules of the particular dialect for a language. Thus, to get the natural speech i.e., to synthesize the usual pronunciations of a word for a particular dialect from the text, proper grapheme to phone conversion is required. This grapheme to phone conversion requires the comprehensive set of phonological rules for the particular selected dialect as every dialect has its own phonology. Phonology of each Bengali spoken dialect has its own large number of rules and corresponding exceptions. However to imitate natural speech one must not compromise phonetic clarity of output speech. The details of phonological rules and the method to incorporate them at the time of text processing are discussed in Chap. 4. The present exercise uses the dialect known as standard colloquial Bengali (SCB), called simply as ‘Bangla’.

There are many aspects of naturalness. One is the acoustic quality. Another is the prosody. For the acoustic quality, attention must be paid to the signal dictionary and to the random perturbations associated with quasi-periodic parts of the speech signal. The structure of the signal dictionary, used in the present synthesizer system, has been discussed in the next section of this chapter. For the introduction of

prosody it is necessary to develop a set of rules for supra-segmental i.e., intonation and prosodic rules for the particular language. For a TTS system, one of the major problems is that of finding the rules for appropriate intonation, stress, duration and amplitude profile from the written text. Their physical correlates are fundamental frequency, segmental duration, and energy, whereas, melody, rhythm and emphasis respectively are their perceptual associations. The introduction of supra-segmentals needs syllabification of the words. A syllable-breaking algorithm is therefore included in this chapter. In human speech, the prosody depends not only on its words, but also on its intended meaning (i.e., whether it is neutral, imperative or interrogative), its intended audience, emotion (anger, happiness or sadness etc.) or physical (sex and age) state of the speaker, and many other factors. Many of these factors are present even in normal reading, because a human being generally interprets and understands the text that they are reading out. Thus, a TTS system will perform as well as humans only when it too can understand the input text, using some form of artificial intelligence. This kind of analysis of text is beyond the scope of the present book. The details of finding out the intonation rules for normal text readings will be discussed in Chap. 5. The durational rules for the Bengali syllables are also discussed in the same chapter. Introduction of stress can be easily accomplished by proper adjustment of the pitch contour and the amplitude and the duration of syllables.

The basic synthesis engine for concatenative synthesis, as envisaged here, is concerned with the production of continuous speech signal by concatenating appropriate signal elements in the signal dictionary after due transformations dictated by the rules of prosody, random perturbations i.e., shimmer, jitter and complexity perturbation. Detailed discussions on the shimmer jitter and complexity perturbation is presented in Chap. 6.

2.2.1 *Partneme: Sub-Phonemic Signal Inventory*

In concatenative speech synthesis as already pointed out, the smallest speech signal units may have the range from a single waveform to a stretch of phones, diphones or vowel-consonant-vowel segments, syllables, demi-syllables. The present speech engine is developed on the basis of the smallest speech unit, namely, the partnemes. There are certain limitations for using phones, diphones, syllables, demi-syllables as the smallest signal units. Though syllables are linguistically appealing unit, there are thousands of different syllables in any language. For the case of phones, SCB consists of thirty-four segmental phones. Among these, seven are vowels and twenty-seven are consonants. But all efforts to make synthesizing speech by concatenating the phone string failed because of the well-known co-articulatory effects between adjacent phones that cause substantial changes to the acoustic manifestations of a phone depending on context. The minimal co-articulatory influences at the acoustic center of a phone lead to the idea to use the diphones as the smallest signal unit. There are altogether 34 times 34 numbers of diphones possible in SCB,

though all do not occur. But the main problem in using diphones is to incorporate stress and intonation in the synthesized speech. Changing the duration as per the prosodic rules, though problematic in the case of diphones, could be taken care of through appropriate technique like PSOLA. Those are true for the case of syllables also. The number of diphone units and also the syllable units increases very much to handle these issues. These problems can be tackled easily with the use of part-nemes as the smallest units. Besides these, the potential disadvantage of the diphone approach is that discontinuities may appear right in the middle of vowels if the two abutting diphones do not reach the same vowel target. This type of problem may also occur in the case of partnemes. This has been taken care of in this investigation by generating some portion of the CV or VC transition. Introduction of stress also becomes very handy in the case of partneme by lengthening or shortening the CV transitory portion. So, handling the change of the fundamental frequency, duration and stress do not require storing extra signal units.

It has been discussed before that for the production of high quality synthesized output speech, both for flat as well as broad range prosodic modification, the choice of speech inventory constituting the signal dictionary plays an important role. In general, our speech inventory is chosen keeping the following criteria in view:

1. Number of units should be small.
2. The definition of the units in signal space should be precise.
3. Average size of a unit should be small.
4. The units should either contain all necessary co-articulatory and anticipatory influences in the signal domain or have the possibility of creating them easily during synthesis.
5. The units leave scope of modification of signal during synthesis to accommodate the demands of supra-segmental features.

In this context the two most popular candidates are phones and diphones. Phones have been the smallest units of spoken language used by the linguists for centuries. Their numbers are also the smallest in any language. From the point of view of production and perception they are very well defined. However, in the signal domain their definitions are vague and imprecise. Except for sibilants and un-aspirated intermittents, a consonant does not have a well-defined boundary in the signal domain. For example an aspirated plosive has a small segment after plosion, which is really a part of the consonant. But it also represents the anticipatory influence of the following vowel and thus is referred to as aperiodic transition. For all vocalic phones, except nasal murmurs, a phone is not a single consistent phenomenon. It has a nucleus whose boundaries are fuzzy but has a nature of its own. Additionally, it has parts, which bear strong influence of the contiguous phones. Their presence can be ignored neither in terms of production nor in terms of perception. Though they play extremely important role in perception of speech, they have no existence in the list of phones.

Diphones are considered to be one of the most promising candidates for concatenative synthesis (Moulines et al. 1990). They have well defined boundaries, though not minimal in size in signal domain. The set of diphones is complete for the

production of continuous speech with unlimited vocabulary. However it is not economic in the sense that many portions of a signal are repetitive resulting in an unnecessary increase in the size of signal dictionary. One such example is that for a particular C and all V's the C is repeated and usually the segment length for the C is of the same order as that of the rest of the signal. Furthermore, all possible consonant clusters are included in the dictionary, which, as we shall see later, is avoidable. This also lends to the increase of the size of the dictionary. In this set up, at the time of introducing intonation and prosody, an additional complication is introduced to detect the vocalic portion of the signal as against the non-vocalic region.

Some of the limitations mentioned above led to the choice of “partnemes”, which are sub-phonemic by their nature, as the speech inventory. For deciding on the inventory of the signal dictionary we assume that the speech signal may be generally divided into two groups without any loss of clarity or significant loss of naturalness, in synthetic speech. The first groups are those, which have their own separate identity. These are the segments corresponding to the phones. The other groups represent the co-articulation between adjacent phones, like CV, VC and VV etc. It must be noticed that even the phonetic quality of the phones are not always the same even for the same speaker. They are also highly dependent on the context, mood, etc. The co-articulatory and anticipatory influences are known to occur with distant phone event. However, such exactitude is not considered at the present level of speech synthesis and therefore we shall assume that representative partnemes would be quite sufficient for almost a natural quality of synthetic speech. Partnemes include identifiable portions unique for phones as well as the segments representing co-articulation. The set of partnemes is divided into two sub-groups. The first group consists of the segments of occlusion or voice-bar along with the plosion or affrication, sibilants, semivowels and diphthongs etc. and a single perceptual-pitch-period for the steady vocalic regions like nucleus vowel, nasal murmurs and laterals. The second group has all CV, VC, and VV co-articulatory regions. It may be noticed that though VOT (Voice Onset Time) is an integral part of the plosives and affricates, it is not included in the consonantal parts for these phones. This is because during the VOT, particularly for aspirated counterpart of these phones with long VOT, strong co-articulatory influences of the succeeding vowels are manifested in terms of aperiodic transitions. It is therefore, judicious to keep them in the corresponding CV transitions.

The plosives and affricates can be broken down into two acoustical subdivisions as:

Plosive: Occlusion/Voice-bar + Plosion

Affricate: Occlusion/Voice-bar + Affrication

The segments corresponding to this group are taken from the starting of the occlusion to the completion of the release of closure. It may be noted here that even for the aspirated counterparts of these phones, this definition holds good. As already mentioned, the aspirated portion is considered as a part of the CV transition.

Figures 2.1 and 2.2 show the partnemes respectively for plosives and affricates according to aforesaid definition. The upper part of each figure shows the time

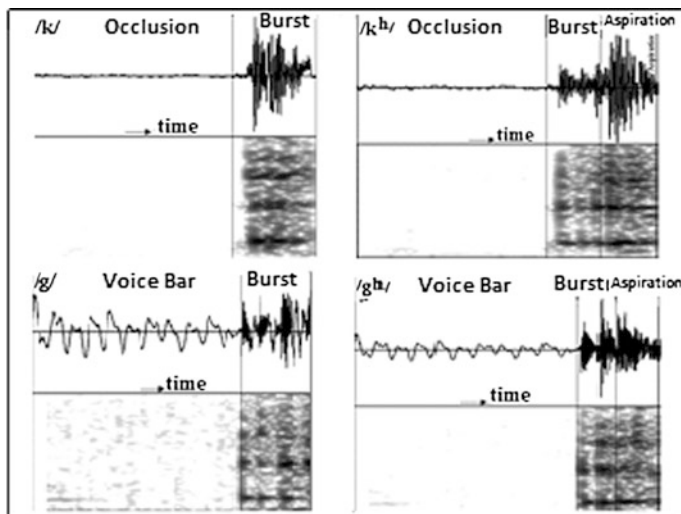


Fig. 2.1 Examples of partnames for plosives

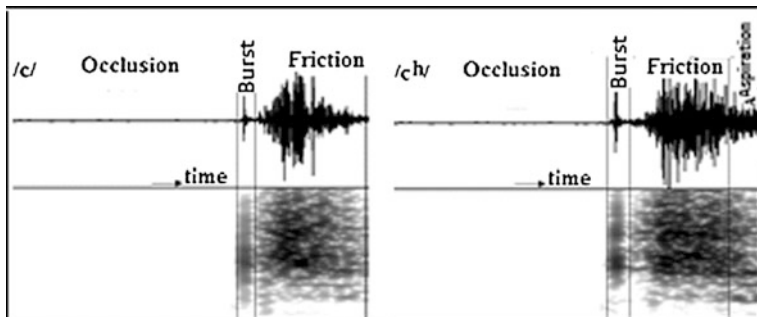


Fig. 2.2 Examples of partnames for affricates

domain representation of the respective consonant and the lower part is its spectrographic representation. In these figures, for the time domain representation, the sample values are plotted along Y-axis and for the spectrographic representation, the frequencies in kHz unit are plotted along Y-axis. In all cases, times in second are plotted along X-axis. The occlusion or the voice bar and the burst portions are indicated clearly in each figure.

The other elements of the first group are sibilants, trills, semi vowels and diphthongs. These are easily and unambiguously identifiable. Signal corresponds to the complete consonantal parts of the phones of largest possible duration. These form the comparatively longer units of the dictionary. Figures 2.3 and 2.4 show respectively the consonants /h/ and /s/. Similar as in the above pictorial representations of the consonants, each figure has two parts.

Fig. 2.3 Consonant /h/

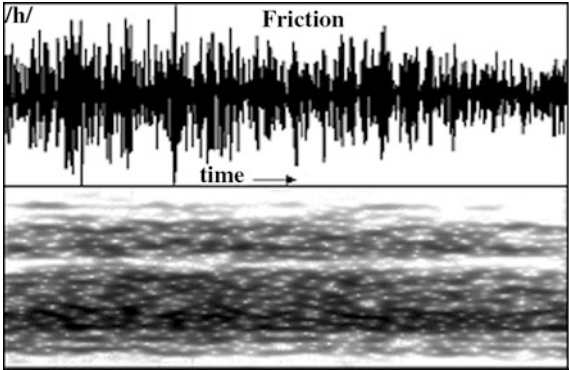
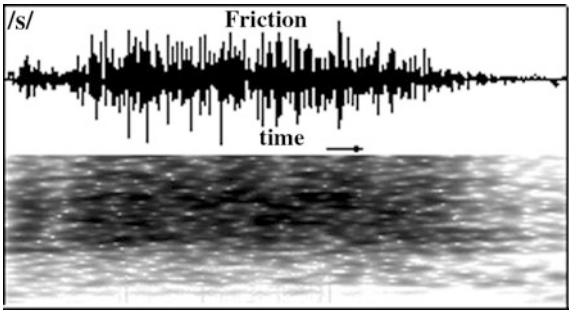
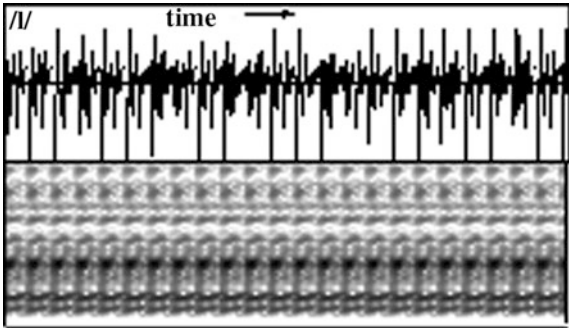


Fig. 2.4 Consonant /s/



Other members of the first group are the vowels, nasal murmur and laterals. For each of the vowels, only a single perceptual-pitch-period (see Sect. 1.4, Chap. 1) from the steady state of each of the vocalic portion is kept as segment for the signal dictionary. Signals corresponding to the complete consonantal parts of the steady portion of the phones of largest possible duration are kept for the nasal murmur and laterals. Figure 2.5 shows the lateral /l/ and its spectrographic representation.

Fig. 2.5 Consonant /l/



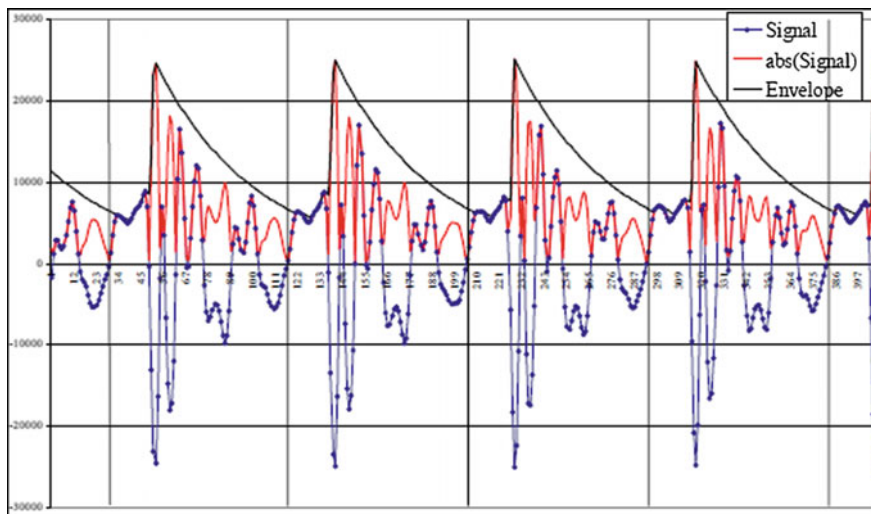


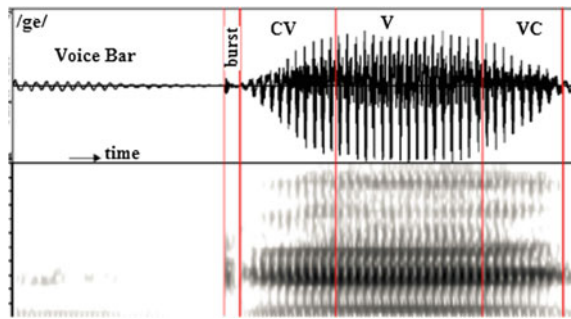
Fig. 2.6 Perceptual-Pitch-Period (PPP) for the vowel /æ/

For all the vowels only a single perceptual-pitch-period from the steady state of each of the vocalic portions is kept as segment for the signal dictionary. The Fig. 2.6 shows the PPP (Perceptual Pitch Period) for the vowel /æ/ in between two vertical lines. The definitions of epoch points and Perceptual Pitch Period are provided in the later Sect. 2.4.

The segment corresponding to the transition class can be thought of as the intermediate part between the aforesaid identifiable parts of the consecutive interacting phones and are the elements of the group 2. These consist of (i) all CV transitions, (ii) all VC transitions and (iii) all VV transitions and glides. The segments under group (i) start from the point where release of the occlusion is complete up to the beginning of the steady state of the vowel, where the coarticulation effect is just stabilized. The segments under group (ii) are extracted in the reverse way i.e., from the end of steady state of the vowel part up to the beginning of a consonant. The segments corresponding to group (iii) start from the end of the steady state of the preceding vowel up to the beginning of that of the following vowel.

The Fig. 2.7 shows the signal and its spectrographic representation for /ge/. The figure clearly shows the CV, VC and steady V regions along with the voice bar and burst for the consonant /g/.

Fig. 2.7 Signal /ge/: the upper part represents the signal and the lower part spectrographic representation



2.3 Structure of Esola

Figure 2.8 below is the schematic diagram of the proposed partname based TTS (Text To Speech) synthesis system. The whole system consists of two main blocks, block A, the high level part of the synthesizer and block B the low-level synthesizer. Block A consists of the units for the language processing. It may be noted here that the structure of this part may differ for different languages. The units constituting the block B are for the purpose of signal processing. The block B actually generates synthesized speech after getting language dependent information corresponding to the input text.

The block A consists of a text input device, a text analyzer, a TLP¹ unit and a unit for phonological, prosodic and intonation rule bases. It may be noted here that though the TLP (Textual Language Processing) unit has been shown in the present system, the development of this unit is not a part of the present book. The information, which is expected from the TLP unit, has been tagged manually with the input text to test the performance of the present system. The input text, essentially a string of characters corresponding to the Bengali grapheme string, may be data from a word processor, standard ASCII (American Standard Code for Information Interchange) from e-mail, a mobile text message or a scanned text from newspapers. After pre-processing the input text for the numerals and acronyms, the next job for the text analyzer is to analyze the input grapheme string and convert the grapheme string into the corresponding ASCII string according to Tables 2.1, 2.2 and 2.3 given below. Thus the text analyzer in block A is basically a grapheme to ASCII string converter with some added features. The output of the text analyzer is a string of ASCII representation of Bengali grapheme with some additional information, such as the word number, syllable number and special emphasis, if any, in the input text. This additional information is required to get the respective rules for intonation, duration and stress for the input text. The word number, syllable number and

¹In this book we shall be using the term TLP in place of the normally used term 'NLP'. As I have said before, the use of the term 'NLP', which normally processes textual language not a natural language, may be confusing for the purpose of the book.

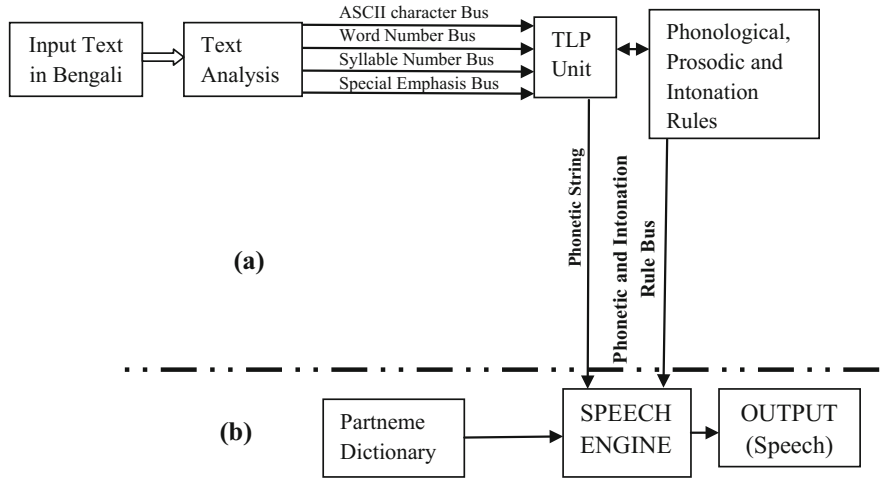


Fig. 2.8 Schematic diagram of partneme-based synthesizer

Table 2.1 ASCII representation of consonantal phones and their IPA notations

	Un-voiced un-aspirated		Un-voiced aspirated		Voiced un-aspirated		Voiced aspirated		Nasal	
	IPA	ASCII	IPA	ASCII	IPA	ASCII	IPA	ASCII	IPA	ASCII
Velar Plosive	k	K	k ^h	KH	g	G	g ^h	GH	ŋ	NG
Palatal Affricate	tʃ	C	tʃ ^h	CH	dʒ	J	dʒ ^h	JH	ɳ	NI
Alveolar Retroflex Plosive	ɖ	T0	ɖ ^h	TH0	ɗ	D0	ɗ ^h	DH0	ɳ	N0
Dental Plosive	t	T	t ^h	TH	d	D	d ^h	DH	n	N
Labial Plosive	p	P	p ^h	PH	b	B	b ^h	BH	m	m
Trill					r	R				
Trill Retroflex					ɽ	R0	ɽ ^h	RH0		
Lateral					l	L				
Alveolar Sibilant	s	S1								
Dental Sibilant	ʃ	S								
Palatal Sibilant	ç	SH								
Glottal Sibilant			h	H						

the special emphasis, if any, in the input text are fed into the TLP unit along with the ASCII string. The TLP unit then analyzes the ASCII string for parts of speech and for clausal/phrasal boundaries. After getting all this information, the unit for phonological, intonation and prosodic rule bases through close interaction with the TLP generates corresponding rules for these suprasegmentals. The phonemic string

Table 2.2 Three-, two- and one-character representations of Bengali vowels and their IPA notations

		Back			Central			Front		
		Oral	Nasal	Murmured	Oral	Nasal	Murmured	Oral	Nasal	Murmured
High	IPA	u	ũ	ɯ				i	ĩ	ĩ
	ASCII	U	U0	U+				I	I0	I+
Mid	IPA	o	õ	ɔ				e	ẽ	ẽ
	ASCII	O	O0	O+				E	E0	E+
Low	IPA	ɐ	ẽ	ɐ	ɔ	õ	ɔ	æ	æ̃	æ̃
	ASCII	AA	AA0	AA+	A	A0	A+	EE	EE0	EE+

Table 2.3 Representations of Bengali semi-vowels and their IPA notations

Semi-vowel/Glides	IPA	j	w	ɥ	ɥ
	ASCII	J0	W	I0	V

bus is the output from the TLP unit, and the corresponding prosodic and intonation rule buses are the output from the unit of phonological, intonation and prosodic rule bases. All these are fed into the speech engine unit in block B for the synthesis of output speech corresponding to the input text.

After getting the phonemic string and the corresponding language dependent rules from block A, the speech engine generates the necessary sequence of tokens for partnemes. The details of the token generation method would be discussed in a later section. After taking the required partnemes from the partneme dictionary, the concatenation and signal processing tasks, as dictated by the rule buses, are done using the ESOLA method to produce synthesized output.

2.3.1 Signal Units Representation

The performance of a synthesizer depends on, to some extent, the method of representing the smallest units at the time of storing them electronically in the computer. In the present case, the smallest speech signal units are partnemes. The partneme representation should be such that the computer can process it easily at the time of synthesis. One such representation is ASCII characters since ASCII code is standardized to facilitate transmitting text between computers or between a computer and a peripheral device.

In the present system, partnemes are represented by some combinations of the ASCII characters while storing them in computer in Windows “wav” format. The Tables 2.1, 2.2 and 2.3 show the three-character, two-character and one-character ASCII representations of the Bengali consonants, vowels and semi-vowels with

their IPA notations. This set of phones is used in the present synthesis system for the generation of unlimited vocabulary. The ASCII representations are used in the ESOLA speech synthesis system for the token generation. The co-articulatory representations between two phones are expressed simply by the combination of the two strings used to represent the two phones, e.g., if one phone is represented by the string X and another is represented by the string Y, then the co-articulatory representation between the two phones would be simply XY. Here, X and Y might be one, two or three character representations of the Bengali phones. In the tables, the dashes mean that those kinds of phones are not present in SCB (Chatterji 1926). All phones in SCB can be represented by one, two and three character representations.

2.3.2 Word Number Bus: Word Segmentation

One of the information buses resulting from the text analyzer unit, after analyzing the input text, is the word number bus. Word boundary detection from text is a relatively simple task for Bengali. These are indicated by the presence of a space character or one of the punctuation marks like stop, comma, semi-colon, colon, question mark, exclamation mark and double quotation. The apostrophe marker is not a word boundary marker.

The word number, i.e., the position of a word in a sentence is important for the introduction of the intonation and prosody in the synthesized speech. The general tendency of the voice sound is to begin with a moderate pitch value and lower the median pitch line during the sentence. This goes up to a syntactic boundary, like phrase, clause or the end of the sentence (Pike 1945). Thereafter, the pitch value again resets to a moderate higher value and the process repeats. This lowering of pitch during continuous speech is called declination and the reset of pitch at the syntactic boundaries is called the declination reset. To introduce the declination over the pitch contour within a sentence the knowledge of the word position is necessary. The position of the declination reset point of the pitch contour is obtained by finding the sentential or clausal/phrasal boundary from the written text. Sentence end is indicated by well defined punctuation marks like stop, question mark etc. and clausal or phrasal boundary is indicated by the punctuation mark comma, semi-colon etc. present in the input text.

2.3.3 Syllable Number Bus: Syllable Breaking Algorithm

Another output of the text analyzer is the syllable number bus for a word of the input text. This syllable marking of a word is necessary for the introduction of intonation and prosody in the word level. The word intonation pattern (Chap. 5) signifies the syllabic level intonation pattern. So, the variation of the pitch in the word level is accomplished by introducing the variation in the syllabic level. Also,

the prosodic variation due to duration is based on knowing the syllable markers for a word. Thus for making the output synthesized speech more natural, syllable marker is one of the most important parameters.

In a language, words are nothing but a string of the combination of consonants (C), and vowels (V). To automate the process of getting the syllable positions of a word in Bengali, the following algorithm is developed. For the purpose of syllabification semi-vowels are considered as consonants.

1. If there is a consonant-consonant cluster ...XXVCCVXX... in a word (here 'X' is either C or V), then first break the CC cluster and go to rule 4 for both the parts.
2. If there is a vowel-vowel cluster ...XXCVVCXX... in a word then first break the VV cluster and go to rule 4 for the first part and rule 5 for the second part.
3. If the vowel-vowel clustering is at the beginning of the word like VVCVCV... then the cluster VV should be treated as a separate syllable and go to rule 1 and process for the remaining portion of the string.
4. If the word is a CVCV... chain, then simply break it in the units CV, CV, ... and mark them by 1, 2, ...
5. If the word is starting with a single vowel (V), like VCVCV... then first treat V as a first syllable and then apply rule 1.

This algorithm was applied to a set of 5000 Bengali words and no misclassification was found.

2.3.4 *Special Emphasis Bus*

The fourth information, given by the text analyzer unit, is the special emphasis bus. It gives the indication when there is any special emphasis that has to be given in the synthesized speech. Getting the information from the input text string about the special emphasis is not an easy task. For this only the syntactic analysis is not sufficient. Some sort of semantic analysis is necessary to get this information from the written text. The semantic analysis is beyond the scope of the present book. In the proposed system, to put emphasis in the synthesized speech, diacritical markers are introduced manually into the written text where special emphasis has to be given.

2.3.5 *Textual Language Processing (TLP) Unit*

TLP (Textual Language Processing) is an important part of a TTS system. It is the part that can find out the clauses, phrases, and parts of speech from the given text input. The output from this drives the phonological, and prosodic rules unit. As we

shall see in later relevant chapters that while clause and phrase boundaries need to be known in the context of prosodic and intonation rules (Chap. 5), the parts-of-speech tags are often required for the phonological rules (Chap. 4).

The TLP is altogether a separate and vast area of language research and analysis technique in text form. For the Bengali language this is well researched domain. Thus, no attempt has been made for the development of this unit.

2.4 Speech Engine

The units described above constitute the high level part of the synthesizer involved in the language processing job for the synthesizer. The units constituting block B are for signal processing. This block consists of the speech engine and the signal dictionary. For ESOLA the signal dictionary consists of partnemes as the basic acoustic signal units.

2.4.1 *Epoch Synchronous Overlap Add (ESOLA) Technique*

Concatenative synthesis uses basic units of signal to be joined together to form speech or music. To retain the intelligibility of speech or music it is necessary that appropriate modification of pitch, duration and amplitude of the units are made without disturbing the timbral quality, particularly the phonetic quality and the identity of the original informant. The listener must feel that the sound he is hearing is coming from a human source. This is a difficult problem particularly for the quasi-periodic region of the signal. ESOLA allows pre-recorded voiced speech signal samples of the dictionary to be smoothly concatenated and at the same time it provides good control over pitch, duration and loudness keeping the natural quality almost undisturbed. From the algorithmic point of view, this is a windowing technique that can modify the signal as well as can regenerate some portion of the signal in between two given voiced segments. This will be discussed later. The novelty of this windowing process is the way of placing of the window on the signal. Each time the windowing uses the epoch position of the signal for aligning itself.

2.4.2 *Epoch Points for Voiced Speech Signals and Perceptual Pitch Period (PPP)*

The quasi-periodic vibration of the vocal folds is the source for generation of the voiced speech. An air pressure difference is created across the closed vocal folds by

contraction of the chest and/or diaphragm muscles. When the pressure difference becomes sufficiently large, the vocal folds are forced apart and air begins to flow through the glottis; this is the abduction phase of the glottal cycle. When the pressure difference between the sub-glottal and supra-glottal passages is sufficiently reduced, airflow begins to reduce and the glottis begins to close. This is the adduction phase of the glottal cycle. It has been observed that the adduction occurs more rapidly than abduction. The glottis quickly closes, resulting in the closed phase of the glottal cycle. The Fig. 2.9 shows schematically the glottal volume velocity function. It is to be noted that the actual excitation of the vocal tract is generated by the pressure changes associated with the cyclic variation in volume velocity. The shape of the pressure variation function is similar to the volume velocity function as shown in the figure. Pressure increases during the abduction phase, drops sharply during the adduction phase, and returns to zero during the closed phase of each glottal cycle. These phases are clearly shown in the Fig. 2.9.

The glottal pressure pulses are responsible for generation of voiced speech. In glottal pulses, the positive rate of change of pressure corresponds to abduction of vocal folds whereas negative change corresponds to the adduction of vocal folds. The maximum of slope of the first one occurs at the epoch positions where the major excitation of the glottal pulses coincides (Ananthapadmanabha et al. 1979) for the voiced speech signal. Each of the glottal pulse acts as an impulse and the air column in the oral-nasal cavities begins to oscillate. The oscillation dies down exponentially during the adduction phase of the vocal folds. In normal voice, the next pulse appears before the oscillation die out. This produces the voiced speech signal (Fig. 2.10). It may be noted here that if the next glottal pulse starts before the previous pulse has

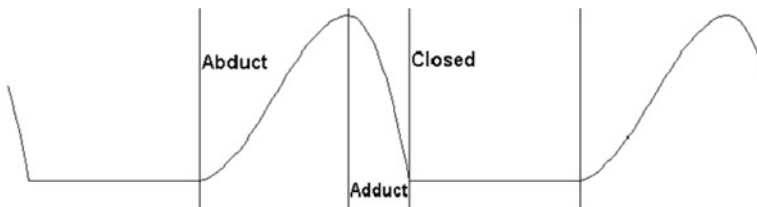


Fig. 2.9 Model of glottal volume velocity function

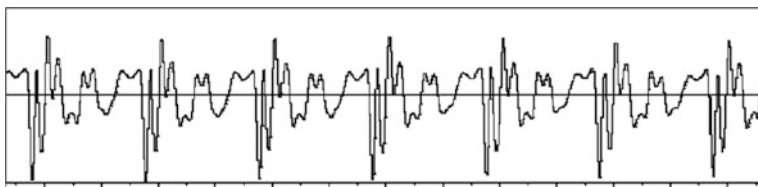


Fig. 2.10 Vowel /æ/As an example of voiced speech signal

decayed sufficiently, the voice will be breathy. Otherwise, if the next pulse starts after the previous glottal pulse has decayed, then the voice will be creaky.

The Fig. 2.11a below shows the time plot of the vowel /æ/ for four consecutive pitch periods (shown as “Signal” in figure), the time plot of the absolute values of the same [shown as “abs(Signal)” in figure] and time plot of the sequence (shown as “Envelope” in figure) defined as below:

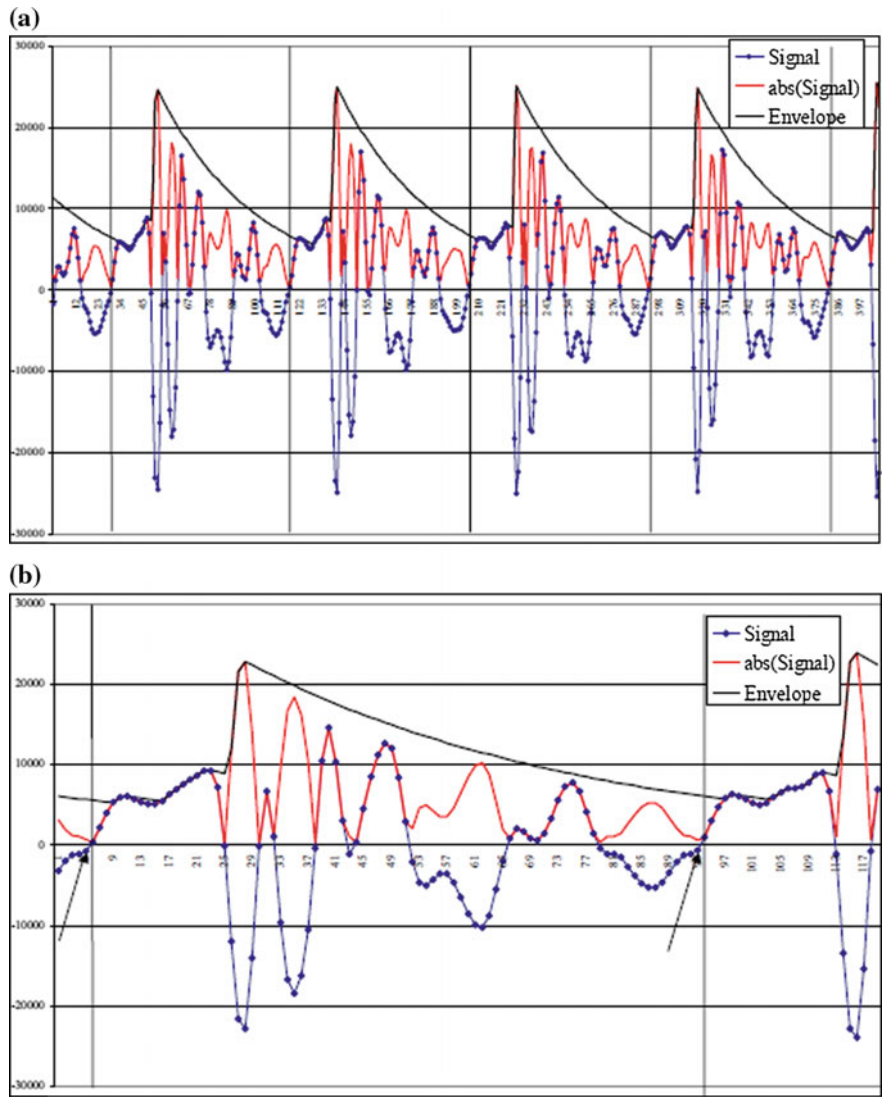


Fig. 2.11 a Vowel /æ/ and Epoch positions. b A single PPP for Vowel /æ/ and Epoch positions

The new sequence $x(n)$ of the ‘abs (signal)’ is constructed from the sequence $y(n)$, representing the speech signal, such that

$$x_i = |y_i|.$$

To get the envelop, the sequence $x(n)$ is modified in the following way:

$$\begin{aligned} x'_i &= x_i \text{ if } x_i > (x_{i-1}) * C \\ &= (x_{i-1}) * C \text{ if } x_i \leq (x_{i-1}) * C. \end{aligned}$$

The modified sequence $x'(n)$ is the envelope, C is the time constant and in the present case its value is 0.98.

The aim to construct the envelope of the voiced speech signal is to get the portion of speech signal that corresponds to the decaying portions of the glottal pulses. For this, analysis similar to the full-wave rectifier circuit of ac current has been done here. The Fig. 2.11b shows the time plots of the three sequences for a single pitch period. In the Figs. 2.11a. b, the “Envelope” plots correspond to the smooth rectified version of the speech signal. The fluctuation of amplitude within a period can be seen easily within a period. We define epoch point as the point of zero crossing closest to the minima of envelope. When the zero crossing near to the minima is not obtained, we take the minima point itself as an approximation to the epoch point (Chowdhury et al. 2001).

In the Fig. 2.11a, we have shown the epoch positions in a portion of the speech signal for the vowel /æ/. The four vertical lines are passing through the epoch points in the segment of the signal.

For any periodic signal, any portion equal to the pitch period if repeated would have the same timbre quality. However for voiced speech, if a portion of the signal significantly smaller than a pitch period is repeated it will produce a sound of different pitch depending on the length of the piece. Interestingly the phonetic quality will be different depending on the position of the piece relative to the epoch on the original signal. However, if the beginning of such a period coincides with the epoch point, defined above, the phonetic quality is retained Dan et al (1993b). The particular period of a speech signal which begins from the epoch is named hereafter as Perceptual Pitch Period (PPP).

In the Figs. 2.11a. b, the vertical lines pass through the epoch points and the pitch periods in between two consecutive epoch positions represents the PPP's.

2.4.3 ESOLA Framework

The ESOLA synthesis scheme involves three steps. These are (1) generation of short-time signals from original speech waveform, (2) epoch synchronous modification brought to the short-term signals, and finally, (3) the synthesis by the concatenation of the modified signals. These three steps are described below.

1. Generation of Short-Time (ST) Signals

Let $x(t)$ be the digitized speech waveform and let e_m : $m = 1, 2, \dots$ represent the successive epoch positions in the signal. The intermediate representation of $x(t)$ is a sequence of short-time (ST) signals $x_m^n(t)$, defined by

$$x_m^n(t) = w_p(t) x(t - pT) \quad \text{for } 0 \leq t < nT \quad (2.1)$$

Here $w_p(t) = (1/\alpha)^p - 1$ for positive integers p, n such that the value of p runs from 1 to n for each ST signal and α is an empirically chosen constant and it is greater than 0. T is the time interval between epoch positions e_{m-1} and e_m . In the Eq. 2.1, the value of p is 1 for the range $0 \leq t < T$, the value of p is 2 for the range $T \leq t < 2T$... the value of p is n for the range $(n-1)T \leq t < nT$. The physical implication of Eq. 2.1 is that the m th ST signal for the m th epoch points of the original signal constitute of n numbers of intermediate signals, constructed from the same PPP in between $(m-1)$ th and m th epoch points, but each time the amplitude is diminished by the factor $(1/\alpha)^p - 1$ with increasing value of p . The length of the ST signal depends on the value of n .

The Fig. 2.12 shows the three consecutive epoch positions and let us denote the three as e_1, e_2 , and e_3 from left to right. Figure 2.13 shows the ST signal for the epoch e_1 the original signal. The ST signal is for $n = 3$ and $\alpha = 4$. The ST signal constitute of generated signal. The part of the signal, left to the left vertical line is for $p = 1$, that in between the two vertical line is for $p = 2$ and the right most one is for $p = 3$. It is to be noted that the number of generated ST signals is equal to the number of epoch points in the original signal.

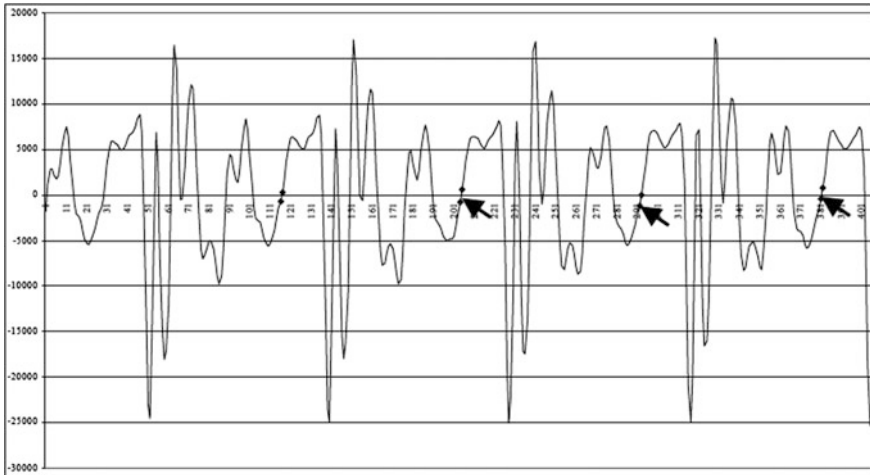


Fig. 2.12 Epoch positions indicated by arrows

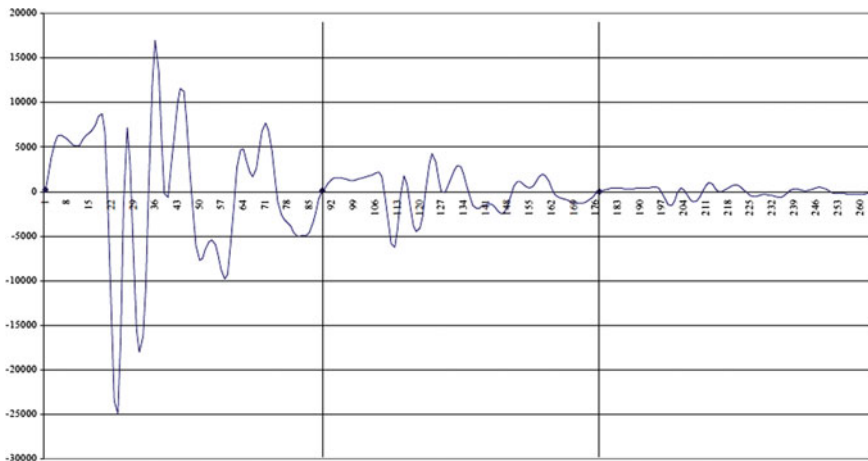


Fig. 2.13 ST signal for e_1 in Fig. 2.12 for $n = 3$ and $\alpha = 4$

It is obvious that if α is chosen a large value, then the amplitude of the generated signals for $p > 1$ become negligibly small. The effect of it in the synthesized signal would be like that a glottal pulse is generated much after the dying down of the previous glottal pulse. This condition would create a creaky voice. Similar, if the value of α is much lower, then the effect of it in the synthesized signal would be like that a glottal pulse is generated much before the dying down of the previous one. Thus, this will create a breathy voice. Empirically the value of α is obtained as 0.25 for the production of good synthesized output. From this ST signal, the smallest pitch that can be generated is

$$f_m = 1/nT$$

Each Short-Time signal is generated for the production of a single PPP of the synthesized speech signal. The value of n depends on the required pitch value of the synthesized signal. After generating the ST signal for a particular epoch points of the original signal, all the parameters are reset and we shift to the next epoch point for the generation of ST signal for that.

2. Epoch Synchronous Modification (ESM) of Short-Time signals

Epoch synchronous modification of $x_m^n(t)$ is described below.

During pitch modification, the stream of Short-Time signals $x_m^n(t)$ is converted into modified stream of synthesized signals by placing a window appropriately and giving rise to a new set of epoch marks e_{sm} . Let $\{e_{sm}: m = 1, 2, \dots\}$ denote the epoch positions of the synthesized speech signal. The algorithm works out a mapping $f: \{e_m: m = 1, 2, \dots\} \rightarrow \{e_{sm}: m = 1, 2, \dots\}$ between original and synthesized epoch marks such that the time difference between two consecutive epochs

equals the required synthesis pitch period. The modified stream of synthesized signals can be represented as:

$${}_sX_m(t) = W_m^n(t) X_m^n(t) \quad (2.2)$$

In the above equation, the left side represents the synthesized speech signal for the m th ST-signal and $W_m^n(t)$ represents the window function for it. Note that this window is defined for every t less than or equal to the modified pitch period and it is zero beyond the pitch period. Selection of W_m^n and its consequence on are described below.

3. Mathematical Analysis of Windowing Processes

(a) Bell Function

The fundamental window function that may be used for ESM is the Bell Window (function, which is defined below.

$$\begin{aligned} w(t) &= 1/2[1 - \cos(2\pi t/T_1)], \text{ for } 0 \leq t \leq T_1 \\ &= 0, \text{ otherwise.} \end{aligned} \quad (2.3)$$

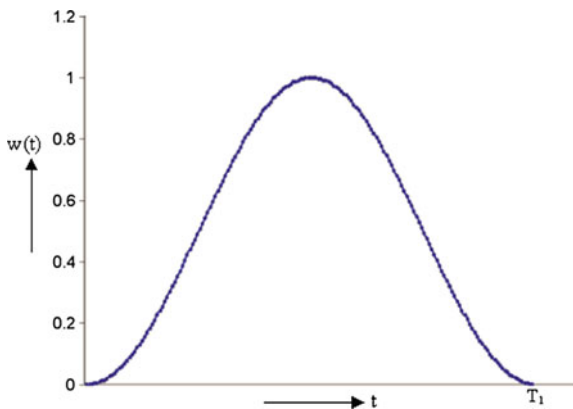
In the Eq. 2.3, T_1 is the pitch period of the modified signal and its value has to be less than or equal to nT . The Fig. 2.14 shows the graphical representation of the Bell Window. In the figure, time is plotted along X-axis and the function value is plotted along Y-axis (Figs. 2.15, 2.16, 2.17, 2.18).

Analytically the speech signal wave $f(t)$, having the period T ($T \geq T_1$), could be expressed as,

$$f(t) = \sum_k a_k \sin\left(\frac{2\pi kt}{T}\right) \quad (2.4)$$

where, a_k 's are the amplitude corresponding to k th harmonics.

Fig. 2.14 Graphical representation of bell function



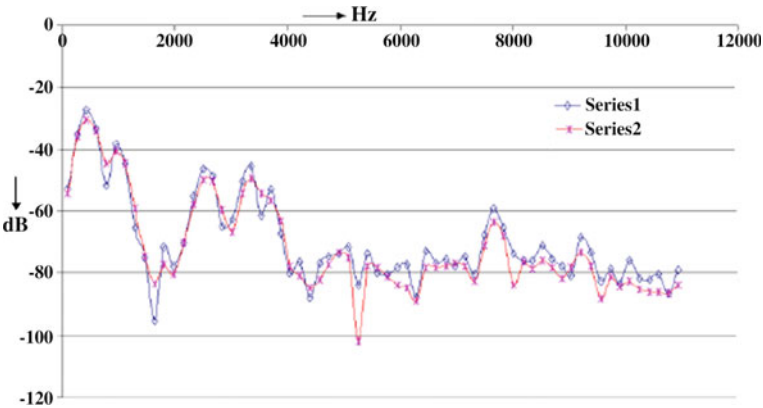


Fig. 2.15 Spectrum sections for vowel /u/

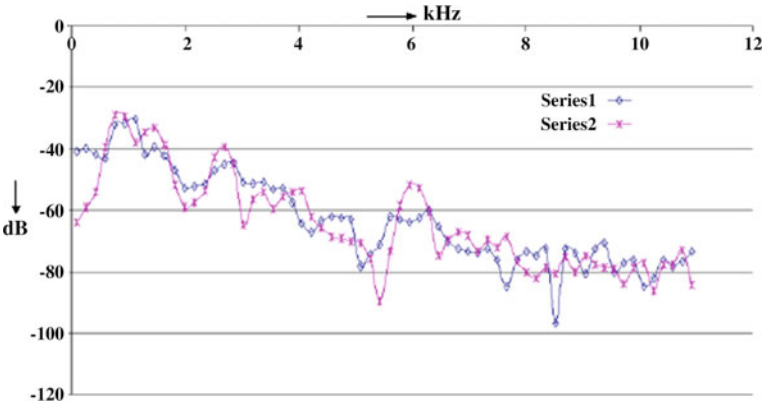


Fig. 2.16 Spectrum sections for vowel /a/

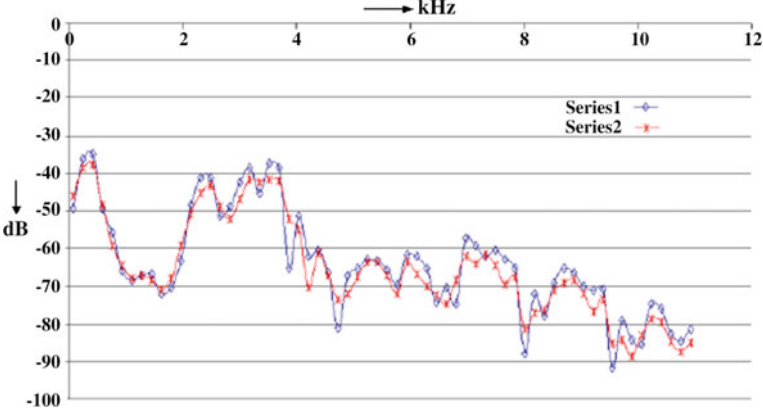
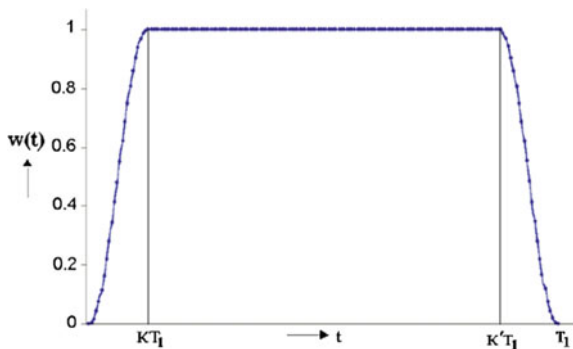


Fig. 2.17 Spectrum sections for vowel /i/

Fig. 2.18 Graphical representation of extended bell function



The windowing method gives the resultant signal $fw(t)$ as:

$$(f_w)t = (f)t(w)t \quad (2.5)$$

In Eq. (2.5) is defined in the region $0 \leq t \leq$ and zero outside of it. The Eq. 2.2 is equivalent with the Eq. 2.5. Assuming that has the same period of the window length, by substituting the values of $(f)t$ and $(w)t$ in the Eq. 2.5 and simplifying, we get the functional form of the synthesized signal as:

$$f_w(t) \frac{1}{2} \sum_k a_k \sin\left(\frac{2\pi kt}{T_1}\right) - \frac{1}{4} \sum_k a_k \sin\left(\frac{2\pi(k+1)t}{T_1}\right) - \frac{1}{4} \sum_k a_k \sin\left(\frac{2\pi(k-1)t}{T_1}\right) \quad (2.6)$$

where a_k 's, are the amplitude corresponding to k th harmonics.

Equation (2.6) yields the amplitude A_k of the k th component of the synthesized speech as

$$A_k = \frac{a_k}{2} - \frac{a_{k-1}}{2} - \frac{a_{k+1}}{2} \quad (2.7)$$

(b) **Preservation of Monotonic Properties of Harmonics in the case of Bell Function**

During the speech production, the glottal pulses are modulated by the resonating property of the vocal cavities i.e., by the response curve of the vocal cavities. The harmonics present in the glottal pulses are changed according to the response curve. Let us find out the conditions under which the monotonic properties of the response curve are also preserved in this windowing process.

2.4.4 Monotonic Properties

Let the response curve of the vocal cavities possess the monotonic increasing properties where the relation between the harmonics is $a_{k-1} < a_k < a_{k+1}$. This condition implies that $a_k - a_{k-1} > 0$ and $a_{k+1} - a_k > 0$. Now it is to be found whether $A_k - A_{k-1} > 0$ and $A_{k+1} - A_k > 0$ hold after the windowing process. Using the Eq. 2.7 we get,

$$A_k - A_{k-1} = 1/4[3(a_k - a_{k-1}) - (a_{k+1} - a_{k-2})] \quad (2.8a)$$

$$A_{k+1} - A_k = 1/4[3(a_{k+1} - a_k) - (a_{k+2} - a_{k-1})] \quad (2.8b)$$

The right side of the above two equations will be positive only when the following inequalities hold.

$$3(a_k - a_{k-1}) > (a_{k+1} - a_{k-2}) \quad (2.9a)$$

$$3(a_{k+1} - a_k) > (a_{k+2} - a_{k-1}) \quad (2.9b)$$

Combining the above two inequalities we get the following condition, which is to be satisfied between the harmonics for holding the monotonic increasing property.

$$\frac{a_{k+1} + a_{k-1}}{a_{k+2} + a_{k-2}} > 0.5 \quad (2.10a)$$

Let us suppose that the response curve has the monotonic decreasing property, i.e., at that point the relation between the harmonics is $a_{k-1} > a_k > a_{k+1}$. This condition implies that $a_{k-1} - a_k > 0$ and $a_k - a_{k+1} > 0$. Similarly it can be seen that, to satisfy $A_k - A_{k-1} > 0$ —and $A_k - A_{k-1} > 0$, the following condition is to be satisfied by $a_{k+1}, a_{k+2}, a_{k-1}$ and a_{k-2} .

$$\frac{a_{k+1} - a_{k-1}}{a_{k+2} - a_{k-2}} < 0.5 \quad (2.10b)$$

Inequalities 2.10a, b give the relations, which are to be maintained among the harmonics of the original speech signal units in order to preserve the monotonic increasing and decreasing properties among harmonics. In normal speech signal the harmonics generally satisfy these inequalities. The Figs. 2.19, 2.20 and 2.21 support this.

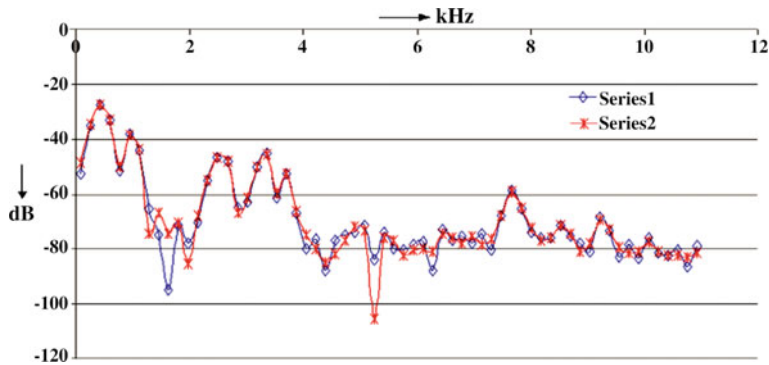


Fig. 2.19 Modification by extended bell function spectrum sections for vowel /u/

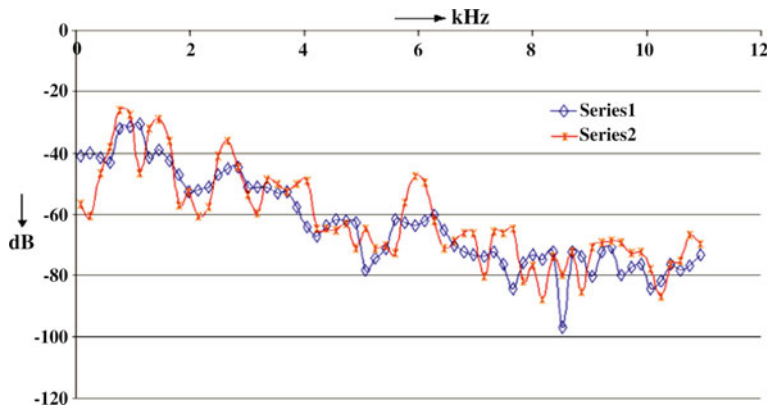


Fig. 2.20 Modification by extended bell function spectrum sections for vowel /a/

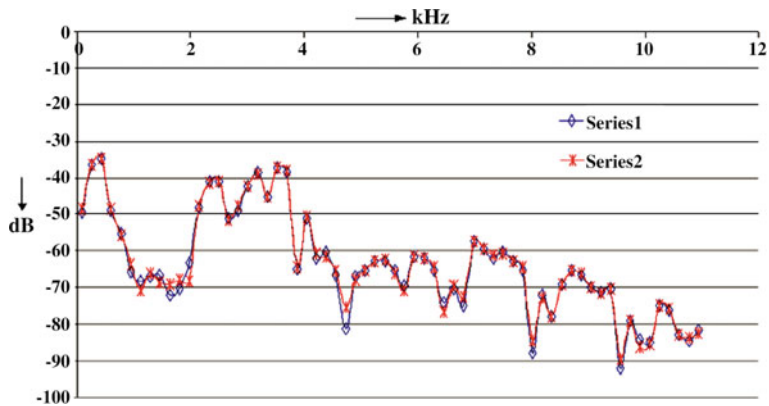


Fig. 2.21 Modification by extended bell function spectrum sections for vowel /i/

2.4.5 Properties Related to Peak

At the peak in the response curve of the vocal cavities, the condition between the harmonics would be $a_k - a_{k-1} > 0$ and $a_k - a_{k+1} > 0$. As similar to the above process, let us now find the conditions for which $A_k - A_{k-1} > 0$ and $A_k - A_{k+1} > 0$ hold after the windowing. Using the same method as above, the obtained condition is as below.

$$\frac{2a_{k-1} + 2a_{k+1} + 3a_k}{a_{k-2} + a_{k+2}} \quad (2.10c)$$

Similar to the above, 2.10c also gives the relation that has to be maintained among the harmonics of the original speech signal units at the peak, in order to maintaining the same condition among the harmonics of the modified signals.

2.4.6 Properties Related to Valley

At the valley in the response curve of the vocal cavities, the condition between the harmonics would be $a_k - a_{k-1} < 0$ and $a_k - a_{k+1} < 0$. As similar to the above process, let us now find the conditions for which $A_k - A_{k-1} < 0$ and $A_k - A_{k+1} < 0$ hold after the windowing. Using the same method as above, the obtained condition is as below.

$$\frac{2a_{k-1} + 2a_{k+1} + 3a_k}{a_{k-2} + a_{k+2}} > 0.5 \quad (2.10d)$$

Thus, the relation in 2.10d has to be maintained among the harmonics of the original speech signal units at the valley in order to preserve the same property among the harmonics of the modified signals.

Figures 2.15 through 2.17 clearly show that the peaks and valleys of the harmonics for the original signals are preserved in the case of synthesized signal. Here series 1 is the original and series 2 is the modified. These figures also show respectively the spectrum sections for vowels /u/, /a/ and /i/. In the figures, series1 represents the spectrum section of the signal obtained by concatenation of the one Perceptual Pitch Period for several times and series 2 represents the spectrum section for the signal generated by concatenating the same signal for the same number of times after modifying with the Bell window function. It may be seen that in accordance with theoretical consideration, the spectrum reveals that the positions of extrema in the spectrum are not shifted except in a few rare instances, with

respect to frequency. However for the vowel /a/, there is some shift after 6 kHz which is of no significance for phonetic quality and of little significance in voice quality. Similar results have been obtained for other voiced signal also. It is to be noted that in all the above cases the pitch is not being modified.

(c) The Extended Bell Function

In practical applications the window function that we have used in the present purpose for concatenation and for modifying the pitch is the Extended Bell Window function, which is defined below:

$$\begin{aligned}
 W(t) &= 1/2[1 - \cos(\pi t/KT_1)] && \text{for } 0 \leq t \leq KT_1 \\
 &= 1 && \text{for } KT_1 \leq t \leq K'T_1 \\
 &= 1/2 \left[1 + \cos \left\{ 1 + \frac{\pi t}{(1-K')T_1} + \pi \left(\frac{2-3K'}{1-K'} \right) \right\} \right] && \text{for } K'T_1 \leq t \leq T_1
 \end{aligned} \quad (2.11)$$

Here, T_1 is the modified pitch period and its value must be less than or equal to nT . K and K' are constants such that $K + K' = 1$. In a practical situation, the value of K is chosen to be 0.125 and K' to be 0.875, i.e., it is a symmetric extended Bell function. The Fig. 2.18 shows the graphical representation of the symmetric extended Bell Window function. In that figure time 't' is plotted along the X-axis and the function value is plotted along the Y-axis.

If $K = K'$, then the Eq. 2.11 reduces to the Bell Function. Thus, the Bell Function is a special case of the Extended Bell Function. The mathematical analysis for this Extended Bell Function is more cumbersome and an analytic solution could not be obtained for the harmonics as we have done in the case of Bell Function. In the case of Extended Bell Function, the wave signal is modified only in a small region. More precisely, the small region is on both sides of the point of concatenation. In the present case, only 25% of the total Perceptual Pitch Period is modified if Extended Bell Function is used while for the case of Bell Function it is 100%. Thus, if we use the extended Bell function as the window function at the time of concatenation, a very small portion of the original signal is being doctored. Thus, it can be assumed intuitively that the harmonics present in the signal will also be modified only within the tolerance limit as in the case of Bell Function.

The Figs. 2.19 through 2.21 show the spectrum sections for the vowels /u/, /a/ and /i/ respectively, series 1 is the original and series 2 is the modified. In each of them, series1 represents the spectrum section of the signal obtained by concatenation of the one Perceptual Pitch Period for several times and series 2 represents the spectrum section for the signal generated by concatenating the same signal for the same number of times after modifying with the Extended Bell Window function (pitch is not modified here). The study of the two series in the above said three figures reveal that the monotonic increasing and decreasing property, property at the peaks and valleys for the harmonics of the original signal are also preserved among the harmonics of the concatenated signals using the extended Bell function. Thus

the windowing does not modify the harmonics present in the signal too much. As in the case of Bell Function, in all the above said cases the pitch is not being modified.

2.4.7 Pitch Modification Using Extended Bell Function

As detailed in the last section, the Extended Bell Function have been used, in the proposed approach for modifying pitch. Figures 2.22, 2.23, 2.24, 2.25, 2.26 and 2.27 show the spectrum sections separately for the vowels /u/, /a/ and /i/ respectively, for the cases of when the pitch is the double of the original PPP, and when the pitch is half of the original PPP. In each figure the spectrum section of the modified signal is given along with the original one to show that the formant structures remain almost same for all cases.

After a careful study of all the spectrum sections, it can be concluded that the amplitudes of the harmonics are generally reduced from the previous values. The fundamental frequency component is relatively boosted, whereas the formant positions are normally preserved. Similar observations are found for other vocalic signals.

In case of sonorants, it has been observed experimentally that the phonetic quality including speaker's identity remains within a small region (about 1.5 ms.) of PPP starting from the epoch position (Chowdhury 2006; Chowdhury et al. 2000a). The principle of ESOLA technique depends on this. The success of this method lies in the fact that the acquired signal retains the phonetic quality including the speaker's identity (Dan et al. 1993a).

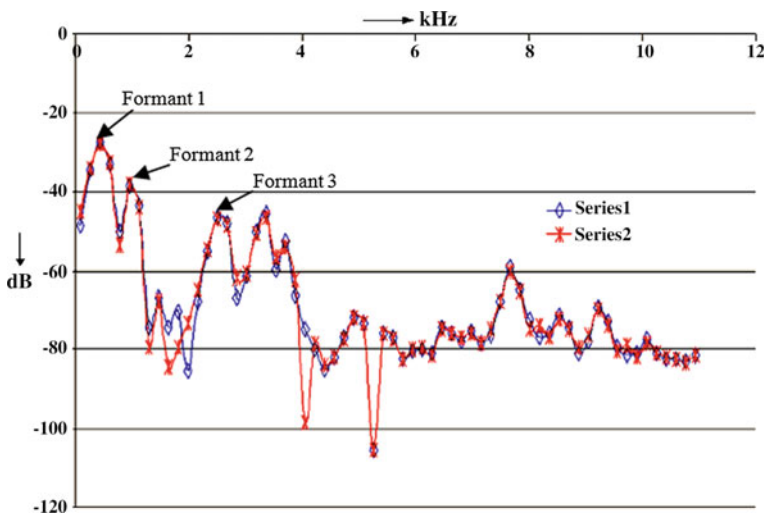


Fig. 2.22 Spectrum sections for vowel /u/, Original Pitch (series 1) and Half Pitch (series 2)

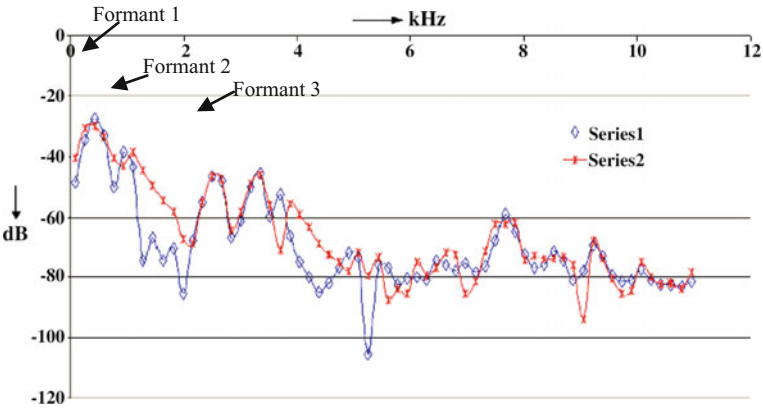


Fig. 2.23 Spectrum sections for vowel /u/, Original Pitch (series 1) and Double Pitch (series 2)

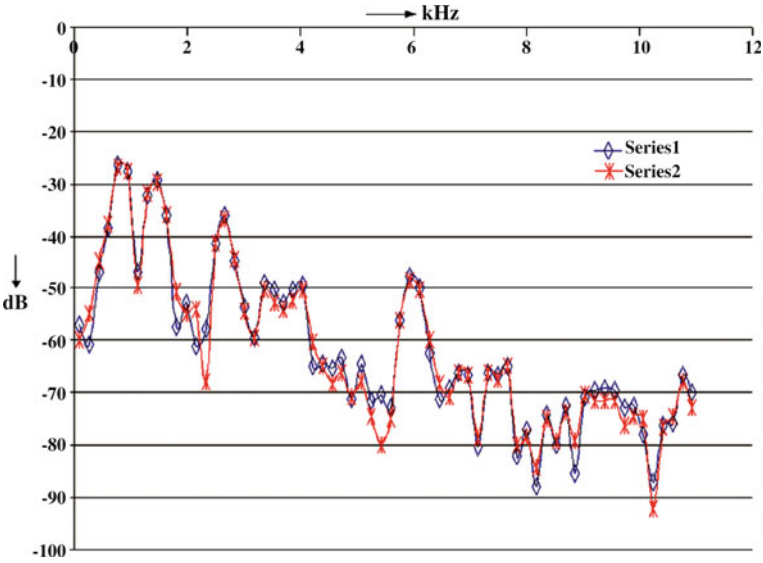


Fig. 2.24 Spectrum sections for vowel /a/, Original Pitch (series 1) and Half Pitch (series 2)

2.5 Preparation of Signal Dictionary

For a concatenative speech synthesizer, construction of a ‘good’ dictionary is the cornerstone of the whole process since its quality determines the quality of the output speech, particularly with respect to phonetic clarity and natural quality of the timbre. The primary need in building the segment dictionary is to record natural utterances. The structure of the signal dictionary, i.e., the nature of the speech

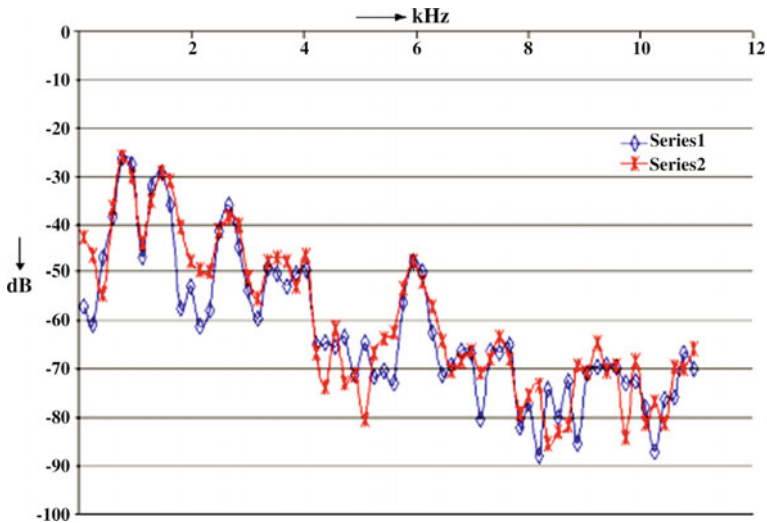


Fig. 2.25 Spectrum sections for vowel /a/, Original Pitch (series 1) and Double Pitch (series 2)

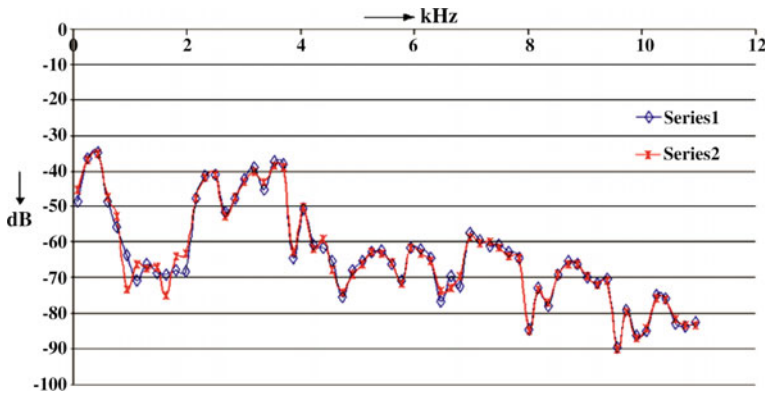


Fig. 2.26 Spectrum sections for vowel /i/, Original Pitch (series 1) and Half Pitch (series 2)

inventory constituting the signal dictionary, plays a dominant role in the selection of natural utterances from where the signal units are to be obtained. These should be such that they include all units used in all possible local contexts. Besides, the selection of goal, economy of design consideration (the size of dictionary, complexity of the rules for the picking of segments from the analysis of text), etc. are also considered at the time of selection of the natural utterance. As for example, co-articulatory and anticipatory phenomena between even non-contiguous phones are the facts in natural speech. But accepting these as a goal in the name of providing naturalness increases the size of the segment dictionary manifold leading, possibly, to an unwieldy size. This, therefore, has to be carefully weighed against

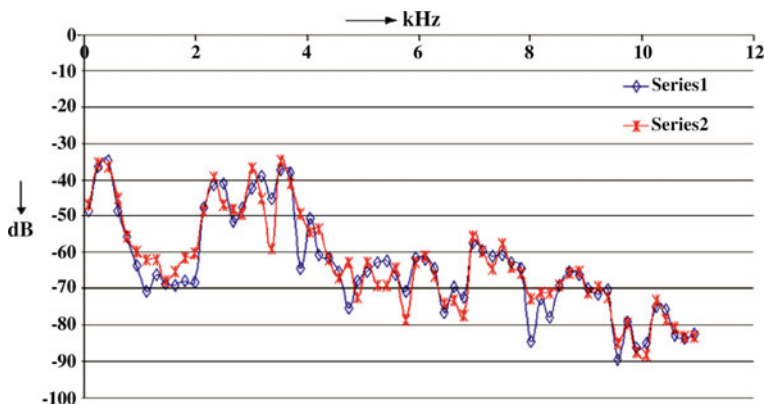


Fig. 2.27 Spectrum sections for vowel /i/, Original Pitch (series 1) and Double Pitch (series 2)

the economy aspect. In the same vein, the naturalness of the output speech, a commendable criteria, must be carefully weighed against clarity, which is considered to be of utmost importance in synthetic speech. It must be noted that while words from continuously spoken sentence possess high degree of naturalness they often lack clarity.

In the proposed system, supra-segmental features, like intonation, duration, stress etc. are introduced at the time of synthesis. Thus, the recordings of natural utterances must be free from these features, i.e., there should not be any variations in intonation, duration and stress at the timing of recording. At the time of utterance of a dictionary word by a native speaker, these features may be introduced unconsciously. At the time of the preparation of signal dictionary, it has been found that the SCB informants sometimes put stress at the first syllable of the utterances. It has been also found that at the time of utterances, the last syllables are sometimes weak or incomplete.

Considering all these situations it is decided to use nonsense words for building the segment dictionary. For getting the correct and signal elements with adequate clarity, nonsense words of the form CVCVCVCV and CVVCVVCVVCVVC are chosen. Here, C's are the elements of the set of all Bengali consonants and V's are the elements of the set of all Bengali vowels. As already mentioned the first and the last syllable in a four syllabic non-sense word may be potentially defective, one would be left with only two syllables left to cut the partnemes from. From these the best, stress free VCV and CVVC syllables, are selected through careful listening. Partneme dictionary has been prepared from these selected syllables. This dictionary contains altogether 1221 number of distinct signal units where the total number of (1) Consonants (C) + Semi-vowels (C) is 36, (2) Vowels (V) is 21 (7 nasals + 7 non-nasals + 7 murmured), (3) CV is 540 (252 nasals + 252 non-nasals + 36 murmured), (4) VC is 504 (252 nasals + 252 non-nasals + 36 murmured), and (5) VV is 84 (42 nasals + 42 non-nasal). With 22.05 kHz, 16 bits format, the size of the signal dictionary would approximately be 3–4 mega bytes.

2.5.1 *Recording*

In the proposed synthesis system, nonsense isolated words are used for extraction of different unit segments. The reading of the words should be flat and free from emphasis. Maintenance of constant pitch is another important requirement at the time of recording. To achieve all these as well as to obtain a good voice quality, a professional speaker was chosen as the informant for the recording of the nonsense utterances.

At the time of recordings, the order of the nonsense words is important. We have put the nonsense words, where the same vowel is combined with all consonants, in a group. The number of such groups is equal to the total number of vowels in Bengali. At a single sitting, the recordings are done for the nonsense words of a group, i.e., for a single vowel. The order of vowels is taken from back-vowels to front-vowels. At first the recordings of non-nasal vowels are done followed by the recordings of nasal and aspirated vowels. All these will enable the speaker to maintain constant phonetic quality for the same vowel throughout reading the words in the list.

After the preparation of the partname dictionary, normalizations in amplitude and pitch are done for each of the partname units. These normalizations are required to reduce the pitch and power mismatch at the junction of two signal units as far as possible. Besides, these two types of mismatches, there is a third one, which is the spectral mismatch at the boundaries. The overlap-add method of concatenation is supposed to take care of it. The details of these three types of normalizations are described below:

2.5.2 *Pitch Normalization*

At the time of concatenating two sounds, there must be a close match in pitch across the junctions. Otherwise, a mismatch will generate audible warbles at the background. This will decrease the quality of the output speech. To bring the pitch of the all the voiced signals in the signal dictionary to a single value, the pitch normalization has been done for all of them. Before going for pitch normalisation, the following factors are kept in mind. These are: (1) the formant frequencies are not rigid values. The formant frequencies of vowels in continuous pitch have in general a large spread around the mean values. In natural speech, a shift of formant frequencies within $\pm 10\%$ of the mean values does not perceptually affect phonetic quality of the vowels in any significant manner. (2) Pitch modification using change in the sampling rate changes the resonance frequencies proportionately. However this neither changes the interrelationship between the Fourier components nor introduces any unwanted characteristics usually present in time domain manipulation of pitch. (3) For a professional informant, it is not difficult for him/her to maintain the pitch of his/her utterances within $\pm 10\%$.

Under these considerations the average pitch for the entire vocalic region is first determined for the raw digital signal files of the nonsense words and let this be P . Then, the pitch of the individual partname unit is normalized to this value by over-sampling or under-sampling method. Now, for pitch normalization of a partname unit, the average pitch value of the corresponding signal unit is measured and let this value of pitch be P_1 . After this, using the Eq. 2.12 the new sampling rate is found out.

$$S_{\text{new}} = S \times \frac{P_1}{P} \quad (2.12)$$

where S_{new} is the new sampling rate and S is the original sampling rate.

To change the pitch value from P_1 to P , the signal unit is resampled using the new sampling rate S_{new} but saving it as the old sampling rate S . From the Eq. 2.12, it is clear that to increase the pitch from the previous value, the signal unit has to be under-sampled from the previous sampling rate and to decrease it has to be over-sampled.

2.5.3 Amplitude Normalization

At the time of concatenating two sounds, there must be a close match in instantaneous power across the junction. Otherwise, a mismatch will generate audible clicks at the background. This will decrease the quality of the output speech. The power mismatch across the junction is eliminated by normalizing the amplitudes. There are two aspects for amplitude normalization, one is the already spoken power mismatch at the junctions and other one is to adjust the intrinsic loudness of the vowels. The different vowels having same amplitude do not produce equal loudness. This effect is known as intrinsic loudness of vowels. For Bengali vowels the data for intrinsic loudness is available (Datta 1998). Thus, the amplitude normalization also requires to conform all signals containing vowels to these required values so that the output has equal loudness over the continuous sentences. This normalisation is necessary for putting proper amplitude as required by prosodic considerations later.

Amplitude normalization of the signal is done in the following way. Let the segment of the discrete speech signal whose amplitude is to be normalised be $y(n)$ [$1 \leq n \leq N$], where N is the total number of sampling points in the signal and n is an integer. The amplitude normalization factor α is defined as

$$\alpha = K/(\max - \min), \quad (2.13)$$

where ‘ K ’ is a positive integer and \max and \min are respectively the maximum and the minimum values of $\{y(n)\}$.

The amplitude normalized signal $y'(n)$ is obtained as follows:

$$y'(n) = \alpha * y(n) \quad (2.14)$$

where, $1 \leq n \leq N$. The property of intrinsic loudness of vowels (Datta 1989) is introduced by changing the normalizing factor α , different for different vowels, by varying the values of 'K'.

2.5.4 Complexity Matching: Regeneration of signal

The source of mismatch of complexity between two component signals lies in the fact that complexity can not exactly be kept equal while reading the long list of words. This is of particular significance with steady state for vocalic signals, which are represented by a single unit, chosen from a large number of possible candidates. While for a vowel V this is fixed for the CV and VC elements in a particular CVC context these could be all different, at least to some degree, for a different consonantal context. This means that complexity at the target end is likely to be different for different elementary units involving the same target vowel. Thus it is necessary to evolve a method to minimize the complexity mismatch by the manipulation of the complexity at the target end for all the CV and VC transitions.

In normal speech, when there are two adjacent phones in an utterance, there is a continuous change in the articulators' position going from the first phone to the second. This is revealed in the formant structure, i.e., the complexity pattern, for this utterance. Thus we get a transition part, which correspond the dynamic change of the articulators in going from the shape to produce first phone to that shape to produce next one. The transitory movement of the spectral structures particularly the formants is generally non-linear and a non-linear manipulation is complex and requires elaborate study of the non-linearities. One way to circumvent this is to use linear manipulation and subject these to perceptual tests. This section deals with a signal domain approach to solve the problem.

First, an attempt is made to regenerate the whole of CV, VC and VV transitions from the two terminal pitch-periods. The basic principle is simply to mix these two terminal waveforms with suitable weights. Let $Y_1(n)$ and $Y_2(n)$ [$1 \leq n \leq N$] be the two given discrete speech signals, where N is the total number of sampling points in each of the waveforms. The numbers of sampling points are equal here since the signals in the signal dictionary are pitch normalized.

Also let, X_i [$1 \leq i \leq M$] be the i th waveform in between $Y_1(n)$ and $Y_2(n)$. Then, the j th sampling point of X_i will be given by,

$$X_i(j) = Y_1(j) * \frac{M - i + 1}{M} + Y_2(j) * \frac{i}{M} \quad (2.15)$$

where, $1 \leq j \leq N$.

The results of the aforesaid method of regeneration are shown in Figs. 2.28, 2.29, 2.30, 2.31, 2.32, 2.33 and 2.34. These figures represent syllables of the form VCV where the figures having the numbers with ‘a’ and ‘b’ show the generated signals and the original signals respectively. In the generated signals, the original CV and VC transitions are replaced by the recreated counterparts. In the examples, seven non-nasal vowels are taken with the combination of that consonant which produces large transitory movements in the formant structures. The combinations are chosen such that the positions of articulations of one vowel and the corresponding taken consonants differ largely. Thus, these combinations show long transitory movements in their formant structures.

Comparisons of the formant structures between the recreated VCV syllables with the original one reveal extremely smooth reconstruction of the transition. Perceptually, the recreated signals also showed good agreements with their counterparts. These examples definitely show that the linear generation of the whole transitions could be the alternatives of the original one. However for the purpose of matching the transition with target in the signal dictionary it is not required to recreate the whole transition. It would be sufficient to recreate the last two to three

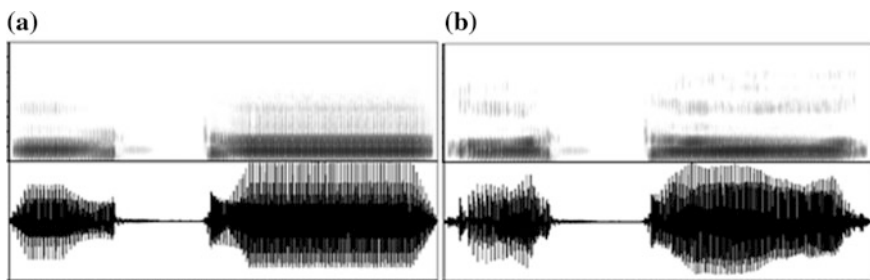


Fig. 2.28 Spectrogram of the transitions **a** Generated and **b** Original for /ata/

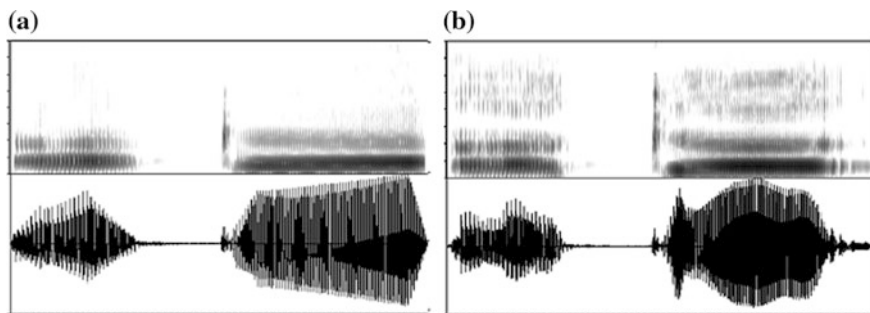


Fig. 2.29 Spectrogram of the transitions **a** Generated and **b** Original for /ækæ/

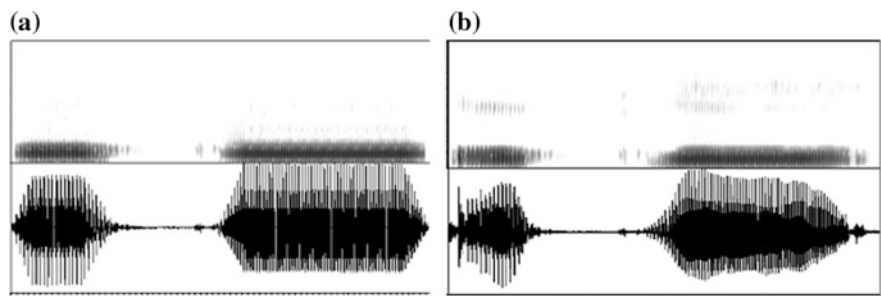


Fig. 2.30 Spectrogram of the transitions **a** Generated and **b** Original for /ɔkɔ/

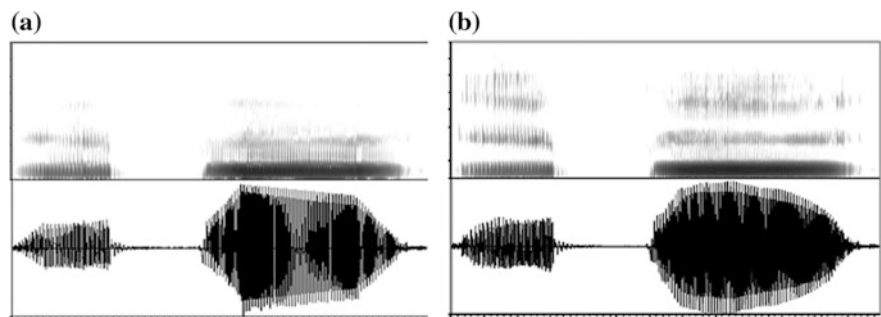


Fig. 2.31 Spectrogram of the transitions **a** Generated and **b** Original for /epe/

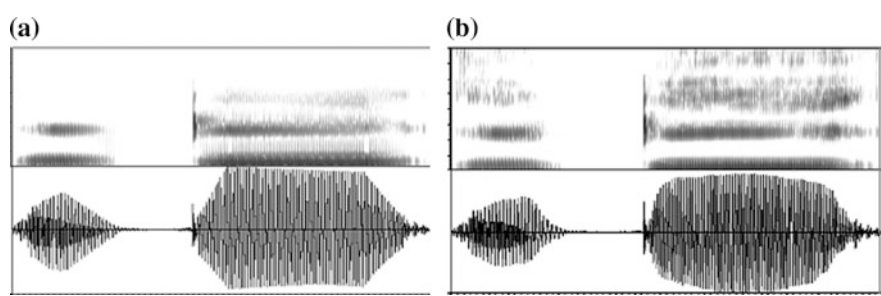


Fig. 2.32 Spectrogram of the transitions **a** Generated and **b** Original for /iti/

pitch periods to obtain a match with the target vowels. It may be noted that the amplitude profile of the original signals were not imitated as they are not relevant in the perception of the place of articulation of the consonants.

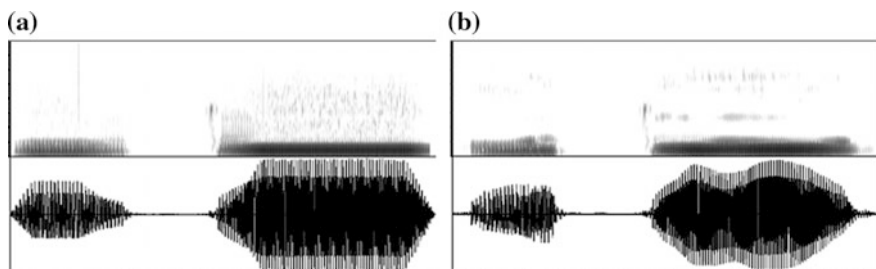


Fig. 2.33 Spectrogram of the transitions **a** Generated and **b** Original for /oto/

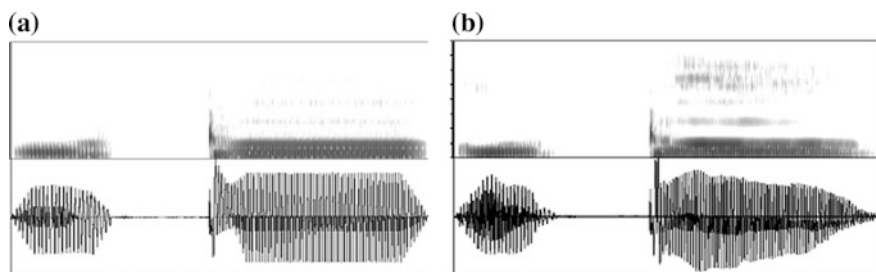


Fig. 2.34 Spectrogram of the transitions **a** Generated and **b** Original for /utu/

2.6 Synthesis Procedures

In this section the steps in succession necessary for speech synthesis in the ESOLA approach are described. The details of the techniques and methodologies for each step are already described in the previous sections. The input text is pre-processed in the Text Analyzer Unit resulting in the (1) ASCII string corresponding to input Bengali grapheme, (2) Word Number Bus, (3) Syllable Number Bus, and (4) Special Emphasis Bus, if any present in the input text. These are fed into the TLP unit of the synthesizer. The TLP unit generates a phonetic string that consists of five-tuple codes of the form (Token, F0, Ampl, Dur, Tag). The token field may be one of the following: the consonant-vowel transition (CV), the vowel-consonant transition (VC), the vowel-vowel transition (VV), the vowel (V), the consonant (C) and pause (P).

2.6.1 Rules for Token Generation

1. $/C_1VC_2/ \rightarrow /C_1/ + /C_1V/ + /V/ + /VC_2/ + /C_2/$
2. $/C_1C_2/ \rightarrow /C_1/ + /C_2/$

$$3. /N_1V_2/ \rightarrow /N_1/ + /N_1V_2/ + /V_2/$$

$$4. /,/, /./, /?/, /:/, /:/ \rightarrow /P/$$

The phonetic string bus in ASCII form is parsed in accordance with the above rules to produce the token for the synthesis operation. For any punctuation marks the token 'P' is generated. The /V/, /CV/ and /VC/ tokens are characterised by the 'F0' and 'Ampl' fields. And the 'Dur' field only characterises the /V/ token. Proper values of these fields are obtained from the 'Phonological Prosodic and Intonation Rules' unit. For the /C/ and /P/ token these three fields have the value 'null' that means nothing is to be done with these fields. The 'Tag' field provides information to the synthesis unit about what has to be done with the signal unit corresponding to the token. For the /V/ tokens, the 'Tag' field contains the string 'ext', which means that extension operation has to be performed with the signal unit corresponding to /V/ token. It is to be noted here that the nasals and the lateral are considered as /V/ tokens. At the time of this operation the other fields namely 'F0', 'Ampl' and 'Dur' are used. 'Tag' field contains 'con' string for the /C/, /CV/ and /VC/ tokens, which imply that simple concatenation have to be performed with the signal unit corresponding to these tokens. Again for /C/ token, the other three fields are null. This means that for this time there would not be any type of manipulation on the signal unit corresponding to /C/ token. But for /CV/ and /VC/ tokens the 'F0' and 'Ampl' field may contain values and in that case the signal units corresponding to these tokens are modified accordingly. For the /P/ token, the 'Tag' field contains the amount of pause that has to be incorporated at the time of synthesis according to the punctuation marks obtained at the time of text pre-processing. All these are fed into the synthesis unit for the actual synthesis and proper signal processing.

2.6.2 Synthesis Operations

The synthesis unit performs the actual concatenation and signal processing operations for each of the signal units corresponding to each token and the values in the fields characterizing the token. It should be noted here that corresponding to each token there exists a unique signal unit in the partname database and therefore a signal unit is always selected each time corresponding to any token.

2.6.3 Signal Processing Aspects

The following signal processing operations are done on the concatenating signal according to the instruction obtained from the 'Tag' field and the values obtained from the 'F0', 'Ampl' and 'Dur' fields.

(a) **Intensity Modification (Amplitude Modification)**

The intensity modification is nothing but manipulation of the amplitude of the signal unit that can be achieved by increasing or decreasing the sample values in that signal unit. This is done by multiplying each of the sample values of the selected segment by the value specified by ‘Ampl’ field parameter of the corresponding token.

(b) **Duration Modification**

The “duration modification” manipulates the syllabic duration. In any syllable, there is no scope to increase or decrease the duration in the /CV/ or /VC/ segments since their lengths are fixed according to their definitions. Thus the only way to increase or decrease the syllabic duration is by changing the steady state duration of the vowel. This is done by calculating the number of the perceptual pitch period of the corresponding vowel that has to be concatenated to obtain the required duration. The value of duration is obtained from the ‘Dur’ tag corresponding to the /V/ token. It may be noted the duration can be changed only by a whole pitch period.

(c) **F₀ or Fundamental Frequency Modification**

To introduce intonation in the synthesized speech pitch has to be modified. The values of the pitch corresponding to a token is obtained from the ‘Phonological Prosodic and Intonation Rules’ Unit and stored in the ‘F₀’ field. For CV, VC, and VV transitional segments and vowels (V), nasal murmurs and laterals successive periods are modified according to the value of pitch specified by the F₀ field. The pitch modification is done using the ESOLA technique already described. It may be noted the F₀ may contain a sequence in case of changing intonation.

(d) **Introduction of Shimmer, Jitter and Complexity Perturbation (CP)**

Normal human voice is not perfectly periodic, it is quasi-periodic in nature. In normal speech two consecutive periods differ in pitch, amplitude and complexity and these variations are random in nature. They are known as jitter, shimmer and CP (Complexity Perturbation). Absence of these parameters produces a perceptible mechanical horn like quality over and above the normal quality of the voice. To eliminate this, proper values of jitter, shimmer and CP have to be introduced into the synthesized speech signal. The introduction of jitter means the incorporation of random pitch change with certain percentage over and above the normal pitch value in the synthesized speech. The introduction of shimmer means that random changes in amplitude with certain percentage over and above the normal amplitude value have to be incorporated. This, to some extent, is automatically done when one introduces the complexity perturbation into the synthesized speech. The detailed discussion on jitter, shimmer and CP are done in Chap. 6. In that chapter, the exact values of those parameters are found which have been introduced in our present system to improve the quality of the synthesized output.

Finally, to remove striation in the synthesized speech, a smoothing is performed. The smoothing basically block the higher harmonic components of the synthesized

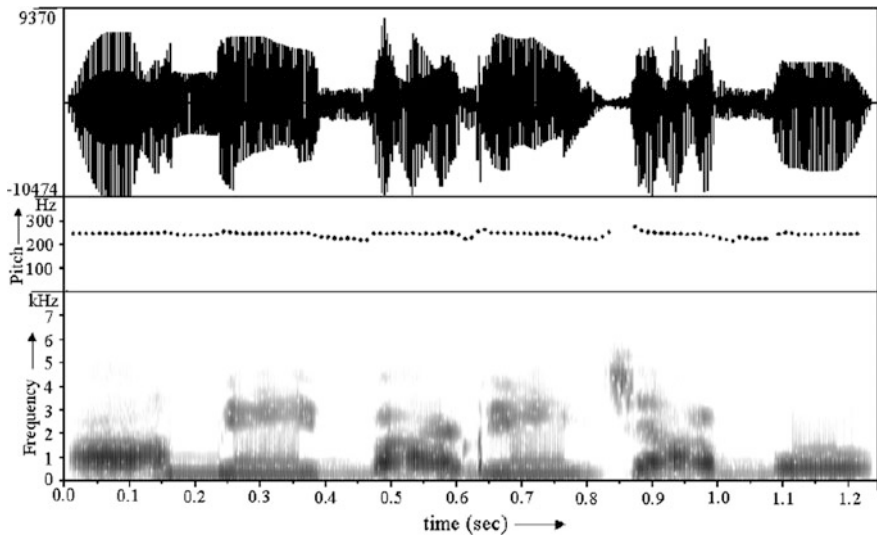


Fig. 2.35 Synthesized output for /আমি বাড়ি যাব/

signal, those are introduced at the junctions due to concatenation and also due to manipulation of the signal for the introduction of prosodic features. The applied smoothing algorithm is defined below:

Let the segment of the discrete speech signal on which smoothing has to be applied be $y(n)$ [$1 \leq n \leq N$], where N is the total number of sampling points in the signal. The modified smooth signal $y_M(n)$ is given by,

$$y_M(i) = [y(i) + 2y(i+1) + 2y(i+2) + y(i+3)]/6 \quad (2.16)$$

where, $y(i)$ and $y_M(i)$ are the i th sample of the original and the modified signals respectively.

Using the described techniques, we have synthesized several sentences in SCB. In all the cases, the quality of the synthesized speech is found to be good. As an example the Fig. 2.29 below shows the waveform and spectrographic representation (the upper and lower part of the figure respectively) of a synthesized signal for a Bengali sentence /আমি বাড়ি যাব/ (I home shall-go.) /ami badi dzabo/. The figure also shows the calculated pitch values (middle part of the figure) for different parts of the sentence (Fig. 2.35).

2.7 Esola and Other Concatenative Approaches

A popular concatenative technique that is being widely used now for synthesizing speech is PSOLA (Pitch Synchronous OverLap Add). In the PSOLA method, the pitch markers determine the placing of window function. The pitch markers could

start from anywhere in the signal. But, PSOLA method typically sets the pitch markers at the signal maximum positions and places the centre of the window function there. This assures that the epoch positions will lie well inside the window and degree of attenuation will be less at epoch positions. But when pitch is changed by too large a degree the possibility of modification increases (Bunnell et al. 1994). Generally TD-PSOLA provides acceptable quality speech synthesis. However spectral mismatch is reported at segmental boundaries leading to tonal quality degradation when prosodic modifications are applied on the acoustic units (Stylianou et al. 1995). Naturalness is retained only in a limited range of prosody modification.

It is reported that the perception of phonetic quality depends only on a small segment (about 1.5 ms) of the pitch-period measured from the epochs (Dan et al. 1993b). This epoch lies close to the beginning of the corresponding glottal cycle. Thus if this modification is large enough, that might distort the phonetic quality of the synthesized speech. In contrast to PSOLA, in ESOLA method, the windows used for modification of pitch and duration as well as for generation of steady states are aligned with this epoch. The epoch markers are determined either by manual inspection of the speech signal or automatically by some epoch detection methods. An offline measurement of the epoch positions and keeping them properly could reduce the time complexity of the ESOLA based synthesizer.

In concatenative speech synthesis, the small speech signal units might have the range from a single waveform to a stretch of phone, diphone or vowel-consonant-vowel segments, syllable, demi-syllables. As pointed out in Sect. 2.1 there are certain limitations in using phone, diphone, syllable, demi-syllables as the smallest signal unit. Though syllable is a linguistically appealing unit, there are thousands of different syllables in any language. In the case of phones, SCB consists of thirty-four segmental phones (Chatterji 1926). Among these, seven are vowels and twenty-seven are consonants. But efforts to synthesize speech by concatenating the phone string create problems because of the well-known co-articulatory effects between adjacent phones. However CHATR is a phone-based synthesizer and in some respects it can be considered quite successful. Co-articulations cause substantial changes to the acoustic manifestations of a phone depending on the context. The minimal co-articulatory influences at the acoustic center of a phone led to the idea of using the diphones as the smallest signal units (Klatt et al. 1982). The number of possible diphones in the language Bengali is 342, though all may not occur. The main problem in using diphones is the incorporation stress and intonation in the synthesized speech. Changing the duration as per the prosodic rules is also complex in the case of diphones. This is because the change in duration means the shortening or lengthening the steady vowel portions. Since the steady part of the vowels are the part of the diphone signal units, to change their length would require extra efforts, though in practice it is not too hard achieve. These are true for the case of syllables also. The increase in the number of diphone units (or syllable units) is too steep to handle the entire gamut of feature space. Besides these, the potential disadvantage of the diphone approach is the appearance of discontinuities in the middle of vowels of the two abutting diphones. This is

because the spectral dynamics of the two steady regions may not reach the same target value. However there are a number of approaches that have been developed to remove discontinuities at diphone boundaries (Dutoit 1994).

The aforesaid limitations of diphones and others speech units can be handled very easily with the use of partnemes. The problem of discontinuities between two abutting vocalic units may also occur in the case of partnemes. But, in the previous sections, we have shown that this problem is tackled by generating some portion of the CV or VC transition by ESOLA technique. It is found experimentally that the stress on a syllable decrease the length of the corresponding CV or VV transitions (Ganguly et al. 1998). CV and VV are well defined units in the partneme dictionary. Thus, only by lengthening or shortening the CV transitory portion stress can be handled. It may be noted here that the CV or VV transition portions also constitute of a number of PPP. To shorten the length of CV or VV transitions, we have to first detect the epoch positions in the CV or VV transitory regions and then depending on stress one or two PPP has to be eliminated from steady vowel target side of the CV or VV transitions. The lengthening of the transitory portion is nothing but regeneration of one or two PPP in the transitory regions.

So, handling the change of the fundamental frequency, duration and stress do not require storing extra signal units. This essentially reduces the size of the signal dictionary. Since, in partneme dictionary, the consonants are well defined, the gemination of consonants and clustering of consonants can be done easily by concatenating appropriate consonantal segments one after another. Apart from the size of signal dictionary, for SCB which is 3,431,768 bytes in 22.05 kHz and 16 bits format, the choice of partnemes as the basic building blocks has twofold advantages over the standard diphone units. Some of the redundancies associated with standard diphone dictionaries are removed from the database. For example the consonantal segments are not replicated in all CV and VC combinations. This enables significant reduction in the size of segment dictionary. The second advantage is the ease of controlling the prosody by the use of ESOLA framework. These advantages give rise to the choice of partnemes as the speech inventory for an epoch synchronous based synthesizer.

For any portable device application, till now, the size of memory is a matter of concern. Text-to-speech would have wide range of applications in any portable or mobile system. To reduce the memory requirement as well as the time complexity, partnemes based synthesizer could be a good choice.

2.8 Conclusions and Discussion

There are two novelties in ESOLA, (1) the use of partnemes used as the smallest signal units and (2) the epoch-synchrony in signal manipulation. The theoretical analysis of the ESOLA technique shows that the manipulation of the signal, even in the case of pitch modification, does not cause any perceptible corruption. The graphemic forms of the Bengali consonants and vowels are also given with their

IPA representations. The details of the partname dictionary have been described with their signal and spectrographic representations. The method of preparation of partname dictionary from nonsense utterances has also been described. The advantages of a partname-based synthesizer using the ESOLA technique for concatenation are also presented.

It may be noticed that the ESOLA engine is language independent. It is also known that the language dependent portion of the major Indian dialects including phone set and the prosodic and intonation rules belong to the same class. The success of its application with SCB indicates the potential for its use in other major dialects of India.

The ESOLA framework and partname inventories altogether give a simple approach for the production of high quality synthesized speech, particularly useful for intonated concatenative synthesis system. Using only the epoch information of the voiced speech signal, the pitch and prosody can be manipulated by keeping the quality intact. The attractiveness of the present approach is its computational simplicity for pitch and duration manipulations. For prosody modification, it is also necessary to manipulate the pitch and duration in the CV, VC, murmur and laterals portions of the stored signals. The epoch detection algorithm is necessary for manipulating pitch and duration in these cases. But this can be avoided by an offline detection of the epochs and storing them in files.

References

- Ananthapadmanabha T, Yegnanarayana B (1979) Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust. Speech, and Signal Processing (ASSP27)*, vol 27. August 1979, pp 309–319
- Brownman C (1980) Rules for demisyllable synthesis using LINGUA, a language interpreter. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 561–564
- Bunnell HT, Yarrington D, Barner KE (1994) Pitch control in diphone synthesis. In: *Proceedings of 2nd European Speech Communication Association (ESCA)/IEEE Workshop on Speech Synthesis*, Mohonk Mountain House, New Paltz, New York, USA, September 12–15, pp 127–130
- Chatterji SK (1926) *The origin and development of the Bengali language*. Calcutta University, Calcutta
- Chowdhury S (2002) Multilingual TTS system in Indian context. In: *Proceedings of the Indo-European Conference on Multilingual Communication Technologies (IEMCT)*, Tata McGraw-Hill Publishing Company Limited, New Delhi, India, pp 101–115
- Chowdhury S (2006) *Concatenative Text-To-Speech synthesis: a study on standard colloquial Bengali*. Ph.D. thesis, Indian Statistical Institute, Kolkata, India
- Chowdhury S, Datta AK, Chaudhuri BB (2001) Concatenative synthesis for a group of languages. In: *17th International Congress on Acoustics*, Rome, Italy, September 2001
- Chowdhury S, Datta AK, Chaudhuri BB (2000a) Pitch detection algorithm using state phase analysis. *J Acoust Soc India* 28(1–4):247–250
- Chowdhury S, Datta AK, Chaudhuri BB (2000b) On the design of universal speech synthesis in Indian context. In: *Proceedings of the Fifth International workshop on Recent Trends in*

- Speech, Music and Allied Signal Processing (IWSMSP), Thiruvananthapuram (India), 14–15 December, pp 14–25
- Courbon JL, Emerand F (1982) SPARTE: a text to speech machine using synthesis by diphones. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1597–1600
- Crespo M, Velasco P, Serrano L, Sardina J (1996) On the Use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech. In: Van Santen JPH, Sproat RW, Olive JP, Hirschberg J (eds) Progress in Speech Synthesis. Springer-Verlag, New York, pp 57–70
- Dan T, Datta AK (1993a) PSNOLA approach to synthesis of singing. In: Proc, PC Mahalanobis Birth Centenary, Volume IAPRDT3, Indian Statistical Institute, Calcutta, pp 388–394
- Dan TK, Mukherjee B, Datta AK (1993b) Temporal approach for synthesis of singing (Soprano 1). In: Proceedings of the Stockholm Music Acoustics Conference (SMAC93), pp 282–287
- Dasmandal S, Datta AK, Chowdhury S (2003) Speech coding a new approach. IEEE TENCON-2003, Bangalore, India, October 14–17
- Datta AK (1989) Manner-based phonetic labeling of speech signal for amplitude information. J Acoust Soc India 17:319–322
- Datta AK (1998) Some studies on acoustic correlation of loudness of vowels. J Acoust Soc India 26(3–4):514–519
- Dixon NR, Maxey HD (1968) Terminal analog synthesis of continuous speech using the diphone method of segment assembly. IEEE Trans Audio Electroacoustics (AU-16), pp 40–50
- Dutoit T, Leich H (1993) MBR-PSOLA: Text-To-Speech Synthesis based on an MBE Re-synthesis of the segments database. Speech Commun 13(3–4):435–440
- Dutoit T (1994) High quality Text-To-Speech synthesis: a comparison of four candidate algorithms. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 565–568
- Fries G (1993) Phoneme-depended speech synthesis in the time and frequency domains. In: Proceedings of Eurospeech 93, vol 2, pp 921–924
- Ganguly NR, Datta AK, Mukherjee B (1998) Acoustic correlates of perceptual stress in Bengali text reading. In: Proc Int Conf On Computational Linguistics, Speech and Document Processing, ISI, Calcutta, pp B68–B71
- Hakoda K, Hirokawa T, Tsukada H, Yoshida Y, Mizuno H (1995) Japanese Text-To-Speech software based on waveform concatenation method. In: Proceedings of International Conference of Applied Voice Input/Output Society(AVIOS), pp 65–72
- Klatt DH (1982) The KLATTalk Text-To-Speech conversion system. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1589–1592
- Macon MW (1996) Speech synthesis based on sinusoidal modeling. Ph.D. thesis, Georgia Institute of Technology, Oct 1996
- Moulines E, Charpentier F (1990) Pitch-synchronous waveform processing techniques for Text-To-Speech synthesis using diphone. Speech Commun 9:453–467
- Pike KL (1945) The intonation of American English. MI: University of Michigan Press, Ann Arbor
- Saito S, Hashimoto S (1968) Speech synthesis system based on interphoneme transition unit. In: Kohasi Y (ed) Reports of the 6th International Congress on Acoustics. International Council of Scientific Unions, Tokyo, pp 195–198
- Sproat R (1997) Multilingual text analysis for Text-to-Speech synthesis. Nat Lang Eng 2(4):369–380
- Stylianou Y, Laroche J, Moulines E (1995) High-quality speech modification based on a Harmonic + Noise Model. In: Proceedings of Eurospeech 95, Madrid, Spain, pp 451–454
- Tohkura Y, Sagisaka Y (1989) Synthesis by rules using CV syllables. In: Proceedings of the fall meeting of the Acoustical Society of Japan, vol 3-4-3, pp 623–624
- Van Santen JPH, Sproat RW, Olive JP, Hirschberg J (eds) Progress in speech synthesis. Springer-Verlag, New York Inc.

Epoch Synchronous Overlap Add (ESOLA)

A Concatenative Synthesis Procedure for Speech

Datta, A.K.

2018, XII, 197 p. 88 illus., 32 illus. in color., Hardcover

ISBN: 978-981-10-7015-0