EFFECTS OF PITCH CONTOURS STYLIZATION AND TIME SCALE MODIFICATION ON NATURAL SPEECH SYNTHESIS

Asok Bandyopadhyay, Shyamal Kr. Das Mandal and Barnali Pal
Speech and Signal Processing Group,
ER&DCI, Calcutta
b_ashoke@hotmail.com

ABSTRACT
This paper describes the method of generation of intonated speech for natural speech synthesis using prosody generation model. The effect of pitch modification through pitch contour stylization for parameter extraction and time scale modification for its implementation has been mentioned. An approach for close-copy syllabic stylization has been described. In the latter part, algorithm for implementation of time scale modification with necessary approximation for sinusoidal signal has been mentioned. Experimental results of applying the technique for pitch modification on Bengali sentence have been shown. The output shows satisfactory performance of sound quality after necessary pitch modification to make synthetic speech natural.

1. INTRODUCTION
One of the major problems in text-to-speech synthesis systems consists in the automatic generation of natural prosody. Prosody is described in terms of fundamental frequency contours (for voiced portion of speech), duration of speech segments (mainly phonemes or syllables) and the amplitude of the speech segments. Speech synthesizers having correct prosodic information can produce very high quality synthetic speech. Various approaches have been proposed for the generation of intonation in speech synthesis systems. The prosodic models used in these systems are quite different, which are defined in terms of (1) pitch target values derived directly from phonological representation, (2) voice source commands, (3) standardized pitch movements, obtained using some stylization of F0 contours [1]. We will discuss about the latter approach. Both in synthesis and analysis the stylized pitch contour is a meaningful level of representation. For synthesis, the stylized pitch controls the synthesizer and for analysis the stylization based on perceptual criteria is a meaningful representation through transcription of the prosodic auditory events in the utterance. In concatenative synthesis method signal segments of nonsense words are joined to produce synthesized output. Concatenated segments having constant pitch produces unintonated and flat output. To introduce intonation and prosody in the synthesized speech signal the pitch profile is to be modified. Amplitude and duration of the segments are also to be changed as per the pitch pattern generated by the text analyzer.

2. PROSODY GENERATION
The term prosody refers to certain properties of the speech signal, which are related to audible changes in pitch, loudness, and syllable length. The most apparent effect of prosody is that of focus. For instance, there are certain pitch event that makes a syllable stand out within the utterance and indirectly the word or syntactic group it belongs to will be highlighted as an important or new component in the meaning of that utterance.
The presence of a focus has various effects, such as contrast, depending on the place where it occurs, or the semantic context of utterance. Prosody generation is typically achieved in two main steps: An abstract symbolic description of prosody is first derived on the basis of its syntactic structure. The amount of syntactic analysis varies considerably from one system to another, but it is most often very superficial (shallow parsing) or even crude.

This symbol string is then converted into an acoustic description of prosody (phoneme duration and F0 curve)

2.1. Stylization

Stylization is a way to reduce the amount of information contained in the fundamental frequency tracing in such a way as to retain only those parts of the pitch curve that have a linguistic function in speech communication, and hence are necessary for the synthesis of prosody. It is really a difficult task to determine which parts of F0 contours are relevant, and how to determine them, there are several approaches to intonation stylization. We can decompose the overall process into three successive components. The first is F0 determination, as F0 is a major input to the stylization algorithm. The second component is the actual stylization. The result is a simplified pitch curve, whatever procedure is used to obtain this curve. This component can be followed by a classification step, in which parts of the pitch curve are recognized as instances of discrete units with in a particular intonation model. The last two components can be merged with in a single step when the intonation model is seen as the set of (normalized) pitch movements.

2.2. Effects of stylization

The syntactic boundary for a sentence is either the phrase or the clause or the sentence as a whole. Declination resets also splits the sentence into such syntactic boundary. To find out intonation classes for a language we have to find out the class for parts of the sentences those can be classified by declination reset rule. The pitch movements within one declination line can be segmented at the word boundaries since words are the basic units of semantics. The intonation pattern within the declination is comprised of the intonation pattern of the individual word in it. Now the words contain a number of nucleus vowels, which represents a syllable. If we take the syllables as basic unit of the pitch movements, then total pitch movement within a declination can thus be seen as the sum of the pitch movements in the syllabic level. Classifying the intonation pattern in the syllabic level thus would be able to describe the intonation pattern of the part of a sentence classified by a declination line. The pitch movement for each syllable is represented by a straight line, which is fitted using linear regression method. After fitting the straight line, we get the close copy syllabic stylization contour for the whole pitch movement within the declination [2].

2.3. Pitch Modification Methods

Pitch modification of the synthesized signal is one of the important aspects to introduce intonation in the synthesized speech signal. In concatenative speech synthesis system partnemes [3] consist of a) Consonant (C) b) Consonant to Vowel transition (CV) c) Vowel period (V) d) Vowel to Consonant transition (VC) e) Vowel- Vowel transition (VV). One period of each vowel is used and total vowel segment is generated by repeating the vowel period. In our segment dictionary the signal whose pitch have to be modified are the CV, VC, VV, V, nasal murmurs and laterals. For V segment the period of successive repetitions are adjusted along with the duration modification. Pitch modification can be done by the two following methods.

• Phase Scale modification.
• Time Scale modification.
2.3.1. Phase Scale modification

The pitch of the segment can be modified by changing the sampling rate of the segment. Here, the envelop of the segmented signal is shifted according to the change of pitch. As a result, large change in pitch can affect the quality of the modified signal.

2.3.2. Time Scale modification

In this method, pitch change is done by changing the period (T) of the signal. The basic principle of modifying pitch is to cut the required period T1 corresponding to the new pitch and suitably fade out the trail. If T1 is greater than the original period T of the signal then additional part is to be added. If T1 is less than T then cut the required period from the original signal with out loosing the quality of the original segment. Due to proper segmentation process each segment contain an integral number of periodic waveform. In case of sonorant, if the first part of the signal corresponding to a glottal period is present, the phonetic quality, including speaker’s identity remains almost intact.

Let y(n) be the segment whose pitch has to be changed according to some defined pitch profile information as given in the Figure 1.

First we need to cut each period of the segment with the help of epoch position identification as shown in Figure 2. The epoch is determined as the minima of the envelope.

Let x(n) be each sub-segment and x(n) has N number of sampling point and its period is T. Also let the required period is T1. When T1 is greater than T, i.e., target pitch is lower than the original pitch, and then glottal period is extended by making a intermediate signal such that \( x_{int}(n) \) is the concatenation of signals x(n) and \( \alpha x(n) \), where \( 0 < \alpha < 0.25 \). Thus we get a signal whose period is 2T having pitch half of the original and contains 2N numbers of sampling point.

Now we define a window W(n) whose length is equal to the desired pitch period on the intermediate signal. The window characteristic is shown in Figure 3.

Now concatenating those changed pitch periods generate the required segment. This process creates a prominent striation and produces a perceptible mechanical horn like sound over and above the normal quality of the voice. This is because such concatenation produces exactly periodic wave instead of quasi-periodic ones. Normal human voice is not perfectly periodic. Two successive pitch cycles do not produce exactly the same pressure waves. The variations are random in nature and occur for pitch, amplitude and complexity, which are referred to as jitter, shimmer and complexity perturbations respectively. An optimum value of these produces natural sound. An excess of the perturbations makes
the quality of sound rough or hoarse. Absence of these perturbations again produces an unnatural horn like sound. Addition of jitter and complexity perturbation almost removes the defect. A random variation of 2-3% in pitch period is introduced for jitter. Adding integer of proper maximum with zero mean and required amplitude to the T1 may do this. Although the introduction of the jitter to the signal will produce some spectral discontinuity. To remove this effect spectral smoothing is done through out the synthesized signal. Smoothing algorithm, which has been implemented is described here.

Let \( y(n) \) be the signal on which smoothing should be applied. According to the algorithm the ith sampling point of the signal is given by,

\[
Y(i)=(y(i)+2y(i+1)+2y(i+2)+y(i+3))/6
\]

for 4 point smoothing .

3 RESULT

A useful method for changing the pitch of each period of a segment has been discussed here. An example of time scale modification of segment ‘ka’ is shown in Figure 4. We have seen that the spectral representation of the modified signal is quite good. It is experimented that the modification of the pitch up to first octave can be done by this method with out loosing the quality significantly. A female voice was used for segmentation and creation of signal dictionary for synthesis.

Figure 4a.
Segment of ‘ka’ transition (250Hz)

Figure 4b.
Pitch Modified ‘ka’ transition (320Hz)

4 CONCLUSION

Pitch contour stylization in general is a powerful tool for designing prosodic models in speech synthesis. Stylistization process is time consuming, and ad hoc decisions regarding the relevant features are need to be taken. Also, the pitch movements obtained could be dependent on the specific characteristics of the speech corpus used and are not strongly linguistically guided. But automatic stylization based on perception, is fast and efficient as the procedure separates the linguistic and perceptual-acoustic aspects of F0 contours. The pitch movements approach is an efficient representation for designing intonation rule for synthesis having above-mentioned limitations. A method of pitch modification has been described here. Output quality of the pitch modified signals showed no degradation and can generate synthesized voice with required prosodic features.

Applying smoothing at the concatenated end unit boundaries improves quality of synthesized voice output. This can be done by conventional LPC smoothing, by ‘fusing’ adjacent units based on Line Spectrum Pair (LSP) or by application of HMM techniques [4].
5 REFERENCES