PROSODY ADAPTATION IN TEXT-TO-SPEECH SYNTHESIS

Say-Wei FOO, *Senior Member, IEEE,* Teng-Chuan POH, and Huat-Chye YEO

***Abstract*-- A prosody adaptation text-to-speech system based on concatenation of spoken English words is presented in this paper. To make the resulting sentences sound natural, modifications are made to the prosody of the speech units using the Time-Domain-Pitch-Synchronous-Overlap-Add method. A prosody generator is built to generate the intonation pattern for the synthesised speech. An approach of smoothing the distortion at the boundaries of the concatenated units is proposed. Listening tests show that the synthesised speech has a high level of intelligibility and naturalness.**

***Index terms*-- speech, text, synthesis**

I. INTRODUCTION

A Text-to-Speech (TTS) system takes in an input text and produces a voiced speech. There are several approaches to speech synthesis. One of the approaches is to concatenate spoken words to form the complete sentences. For this approach, the basic speech units are the spoken words. This approach requires the storage of the sound units of all the spoken words that will be used. However, simple concatenation of spoken words does not give natural sounding speech.

In this paper, an improved synthesis method is presented. The proposed approach makes use of the sinusoidal model [1] for speech parameterisation and the Time-Domain-Pitch-Synchronous-Overlap-Add (TD-PSOLA) [2] for prosody modification. A prosody generator is used to provide pitch contour modification to generate a natural intonation for the synthesised speech. A suitable pitch contour is selected to suit the type of sentence: e.g. question, exclamation or narration.

Informal listening tests show that the proposed method is able to produce natural sounding speech. The block diagram of the proposed methods is presented in Section 2.

Say-Wei FOO and Teng-Chuan POH are with the Department of Electrical and Computer Engineering, National University of Singapore. Huat-Chye YEO is with DSO National Laboratory.

This is followed by detailed description of the various sub-modules in Sections 3 and 4 and discussion of the performance tests in Section 5.

II. THE PROPOSED TEXT-TO-SPEECH SYSTEM

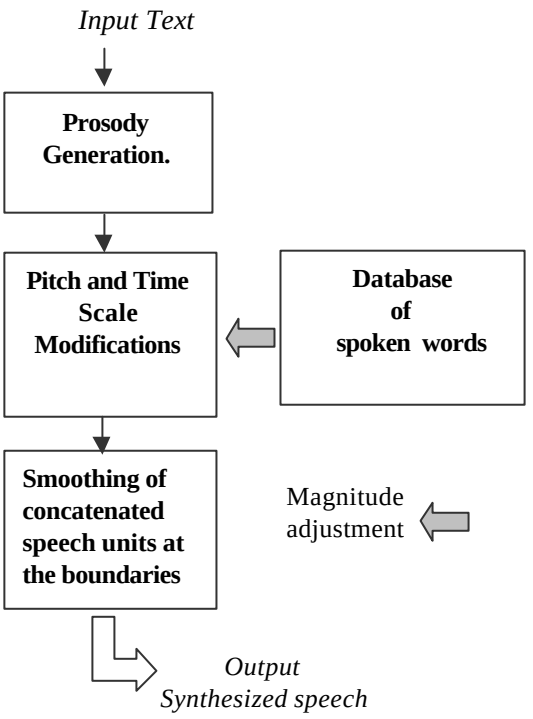


Figure 1. Block diagram of the proposed TTS

The block diagram of the proposed system is presented in the following figure.

A database of speech units is first established. The speech units are parameterised using the sinusoidal model of speech representation.

For a given input sentence of text, the type of intonation for the sentence is determined according to the nature of the sentence. The corresponding pitch contour is then generated. The speech units required to form the sentence are extracted from the database. Pitch and time scale modifications of the speech units are carried out in accordance to the pitch contour. Magnitude interpolation is made to smoothen the joints. Finally, smoothing of the concatenated speech units is performed to produce the output speech.

III. The Sinusoidal Model For Speech Representation

A parametric signal model is often used to extract the features of a signal into a form that is more easily or efficiently processed, and requires substantially less storage than the original signal. For the proposed model, the sinusoidal model [1] is used to represent and synthesise the speech units as the pitch of the sound units can readily be modified using this model of representation. In most practical cases, the synthesised speech sound is essentially indistinguishable from the sound of the original speech.

A. Analysis

For the sinusoidal model, a segment of speech signal *s(n)* is modelled as the sum of a number of sinusoids with time

varying amplitudes and frequencies given by

$$s(n) = \sum_{l=1}^{L} A_l \cos[n\omega_l + \phi_l] \qquad (1)$$

where A_l is the amplitude, ω_l is the angular frequency and ϕ_l is the phase of the lth sinusoidal component.

In implementation, the speech signal is first filtered at 4 kHz and sampled at 10,000 samples per second. Pitch synchronous analysis is used as the sinusoidal model works best with such mode of analysis. The average pitch period for a particular segment of speech is determined using the Maximum Likelihood Pitch Estimator (MLE). Samples in that segment are then divided into frames of width about 2.5 times the pitch period. The Hamming coefficients *w(n)* for this pitch adaptive window are then computed and normalised according to

$$\sum_{-N_k/2}^{N_k/2} w(n) = 1 \qquad (2)$$

The parameters of the model, A_l, ω_l and ϕ_l are found using the Short Time Fourier transform (STFT) i.e. the DFT of Hamming weighted windowed frames of the signal.

The peaks of the spectral magnitude for each frame are identified. Typically, the number of peaks in each frame ranges between 20 and 80 depending on the pitch period.

The frequencies of these peaks change from frame to frame. The time variation of frequencies of spectral peaks of a segment of speech signal is presented in Figure 2. To account for such rapid movement in the spectral peaks, the concept of “birth” and "death" of sinusoidal components is introduced to monitor the change of frequency of spectral peaks from frame to frame.

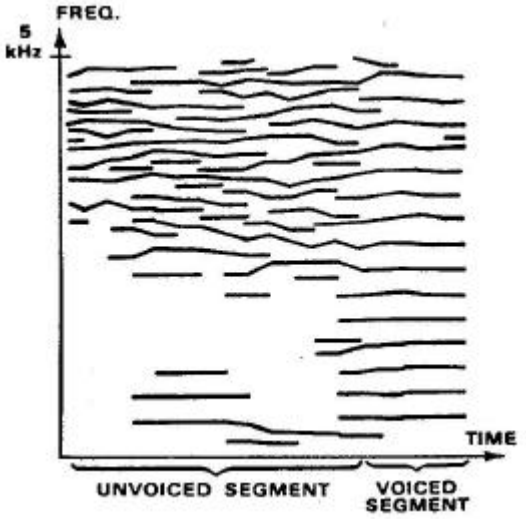


Figure 2. Frequency tracks for a segment of speech

The general idea is to compare the frequency of each selected spectral peak of the k^{th} frame to the frequency of spectral peaks in the following frame. If the difference is smaller than a specified limit, say δ, then it is considered a match. If no candidate is found, the frequency is considered “dead” in the $(k+1)^{th}$ frame. On the other hand, the frequency of a spectral peak selected in the $(k+1)^{th}$ frame that cannot find a match in the k^{th} frame is considered “born”.

It is found that a value of δ between 80-120 Hz provides the least distortion in the final synthesised waveform and the results of frequency tracking.

B. Synthesis

A straight forward approach to synthesis will be to determine the value of each sample in a frame using the amplitudes, frequencies and phases estimated for all peaks in that frame.

For example, if the length of the k^{th} frame is N_k, all the samples $\hat{s}(n)$, $n=0,1,\ldots(N_k-1)$, may be obtained using

$$\hat{s}(n) = \sum_{l=1}^{L(k)} \hat{A}_l^k \cos[\, n\hat{w}_l^k + \hat{q}_l^k \,] = \sum_{l=1}^{L(k)} \hat{A}_l^k \cos[\, \hat{f}_l^k(n) \,] \qquad (3)$$

where $\hat{A}_l^k, \hat{w}_l^k, \hat{q}_l^k$ are the amplitude, frequency and phase of the l^{th} peak of the k^{th} frame. The over-script $\wedge$ is used to denote values obtained through STFT of signal of the frame.

For simplicity, the variable in the cosine function may be grouped into one term, $\hat{f}_l^k(n)$. Note that the number of peaks in a frame, $L(k)$, is frame dependent.

However, for smooth transition, the synthesis of the k^{th} frame would need to take into account the parameters of the k^{th} frame and those of the $(k+1)^{th}$ frame.

To derive the amplitudes of the underlying sine components in transition, a simple linear interpolation method is used. Let $(\hat{A}_l^k, \hat{w}_l^k, \hat{q}_l^k)$ and $(\hat{A}_l^{k+1}, \hat{w}_l^{k+1}, \hat{q}_l^{k+1})$ denote the successive sets of parameters for the l^{th} frequency track, then the amplitude $A_{k,l}(n)$ for $s(n)$ in the k^{th} frame is estimated as

$$\tilde{A}_{k,l}(n) = \hat{A}_l^k + \frac{\hat{A}_l^{k+1} - \hat{A}_l^k}{N_k} n \qquad (4)$$

where $n = 0, 1, \ldots\ldots, N_k-1$. The symbol ~ is used to denote values obtained by interpolation for samples.

For the value $\tilde{f}_{k,l}(n)$, interpolation is based on a cubic phase function, which is set to fit to the set of measured phases at the frame boundaries [1]. In short, the following steps are taken.

Find x^* using the following expression

$$x^* = \frac{1}{2p}\left[(\hat{q}_l^k + \hat{w}_l^k T - \hat{q}_l^{k+1}) + (\hat{w}_l^{k+1} - \hat{w}_l^k)\frac{T}{2}\right] \qquad (5)$$

where T is the duration of the k^{th} frame.

Let M^*=integer of x^*, find α and β from

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{bmatrix} \begin{bmatrix} \hat{q}_l^{k+1} - \hat{q}_l^k - \hat{w}_l^k T + 2pM^* \\ \hat{w}_l^{k+1} - \hat{w}_l^k \end{bmatrix} \qquad (6)$$

The value $f_{k,l}$ is then determined from

$$\tilde{f}_{k,l}(t) = \hat{q}_l^k + \hat{w}_l^k t + at^2 + bt^3 \qquad (7)$$

where t is the time counting from the start of the k^{th} frame. Through a simple conversion from t to n, the sample number counting from the beginning of the frame, we obtain $f_{k,l}(n)$.

The final synthesised waveform is given by

$$\tilde{s}(n) = \sum_{l=1}^{L(k)} \tilde{A}_{k,l}(n) \cos[\, \tilde{f}_{k,l}(n) \,] \qquad (8)$$

C. Performance Of The Method For Single Spoken Words

For all experiments reported in this paper, speech samples from the TIMIT (Texas Instruments Massachusetts Institute of Technology) database are used. In Figure 3, the waveform of the original speech signal of a single word and the re-synthesised speech signal are presented.

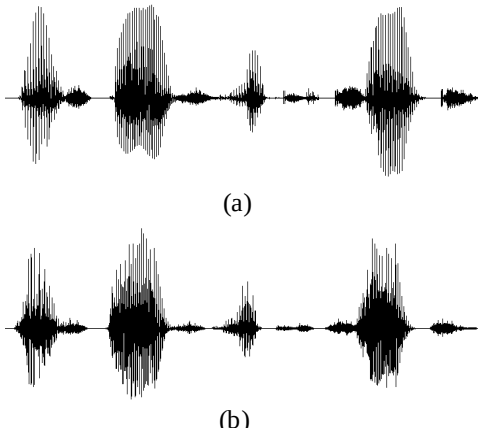


Figure 3. Waveform of "ice baths, electric shocks" (a) Original signal (b) Synthesized signal

Although the waveform shows some variation in the fine details, listening test shows that the intelligibility and speaker identity are well preserved in the sound produced.

IV. Pitch And Time Scale Modifications

The naturalness of the synthesised sentence depends very much on the prosody of the speech. Prosody is a function of the pitch, duration and intensity of the speech units. Straightforward concatenation of speech units will not in general produce naturally sounding sentences. Hence, there is a need to modify the component sound units by a pitch contour suitable for the nature of the sentence, to adjust the duration of the component words and to take care of the discontinuities at the joints of the spoken units.

The synthesis pitch markers t_s are determined from the analysis pitch markers t_q according to the desired pitch-scale and time-scale modifications. A mapping of $t_q \rightarrow t_s$ is determined, specifying which short time analysis signal $x_m(n)$ should be selected for a given time instants.

Lastly, the synthesised signal $y(n)$ is obtained by combining the synthesised waveforms synchronised on the stream of synthesis pitch markers. Least square overlap add synthesis procedure [3] is used for this purpose.

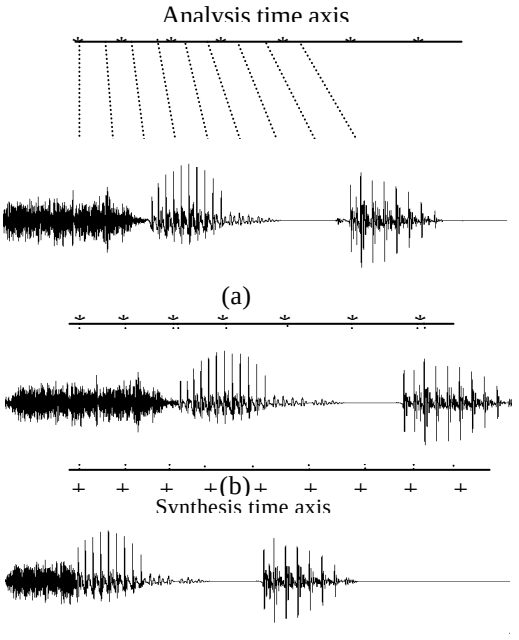
A. Time-Scale Modification

For a constant time-scale modification factor α, the $t_s \rightarrow t_q$ pitch-mark mapping associates t_s with the analysis pitch mark, t_q, being nearest to the instant αt_s. A graphical illustration of the time scale pitch-marker mapping is given in Figure 4.

However, There is a limit to which the speech rate can be increased or decreased since there must be a minimum number of waveforms to represent a voiced region. Listening test shows the minimum scaling factor to be about 0.4. On the other hand, if the speech unit is lengthen by a factor of above 1.7, a tonal quality is noted in the unvoiced region of the synthesised speech. For scaling between the limits, negligible acoustical distortion is introduced by TD-PSOLA. Although there are limits in the scaling factor when performing pitch or time scale modifications, these limits are well above most Text-to-Speech applications.

Figure 5 gives a graphical illustration of the time scale modification of the word "simple".

B. Pitch-Scale Modification



using the TD-PSOLA algorithm.

Figure 5. Time Scale Modification.
(a) The original speech signal of the word "simple"
(b) Waveform with time scaleing factor of 1.4.
(c) Waveform with a time scaling factor of 0.6.

For pitch-scale modification, analysis pitch marks t_q are first determined and the short time signals formed from the windowed samples of the speech signals based on the analysis pitch markers. Depending on the pitch-scaling factor, the appropriate synthesis pitch marks t_s are formed.

Figure 6 illustrates the pitch scaling operation on a segment where the scaling factor is 0.8. Since pitch-scaling will inherently result in time scale modification, compensatory steps are taken to ensure an independent pitch modification.

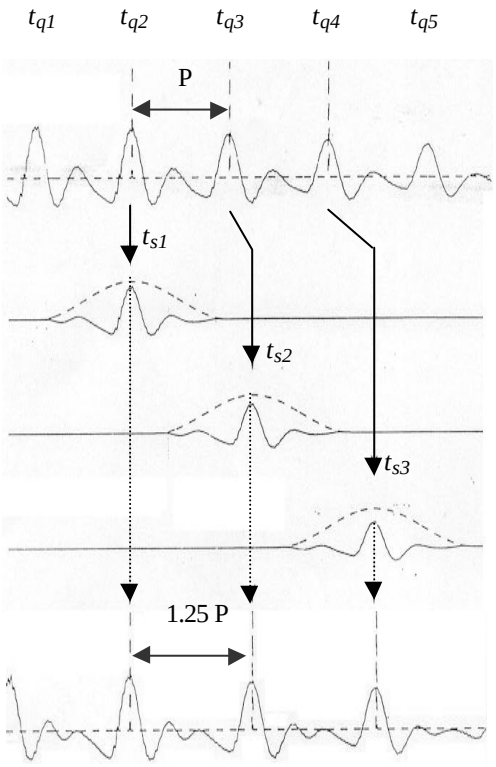
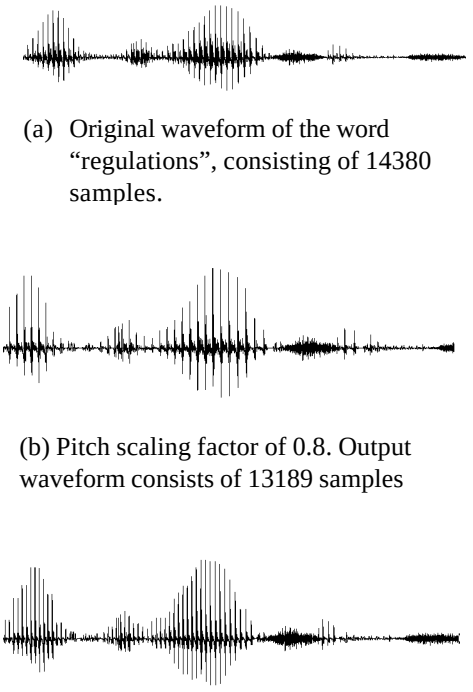


Figure 6. Pitch scale modification with a pitch-scaling factor of 0.8

To achieve this, the proximity of t_{qi} and t_{si} are checked periodically during the concatenation of the short-term analysis signals. If t_{si} is too far ahead of t_{qi}, the next analysis pitch-mark will not be used for the synthesis concatenation. If the t_{si} is too far behind of t_{qi}, then the previous analysis pitch-marker will be reused for synthesis.

In Figure 7, the pitch scale modification of the word "regulations" is illustrated. Figure 7 (a) shows the original waveform of sample size 14380. The pitch-modified segments show a change in the sample size of about 2%-8% of the original sample size.



(a) Original waveform of the word "regulations", consisting of 14380 samples.

(b) Pitch scaling factor of 0.8. Output waveform consists of 13189 samples

(c) Pitch scaling factor of 1.25. Output waveform consists of 140494 samples.

Figure 7. Pitch Scale Modifications of the word "regulations".

C. Smoothing of the Concatenated Segments

The smoothing of the individual speech segments before concatenation is a critical step to reduce the distortion that arises from direct concatenation. In particular, when the segmentation of the acoustical units are not properly done, or if the pitch of at the 2 boundaries differ greatly, or if one frame is voiced and the adjacent frame is unvoiced, the distortion introduced could be significant.

Here, additional processing of the concatenated segments using PSOLA is proposed to reduce this type

of distortion. The average value of the pitch periods of the two frames is determined and the synthesis window length is chosen to be twice the pitch period. A Hamming window is applied to the samples with the joint as the centre and the TD-PSOLA method is again applied. The additional processing results in the reduction of distortion in the concatenation process. This is demonstrated in Figure 8.

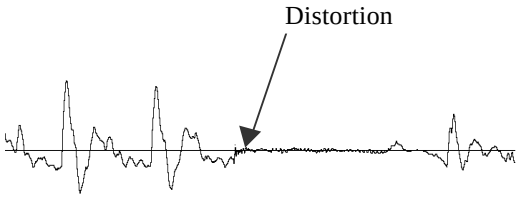


Figure 8(a) Distortion introduced as a result of direct concatenation.

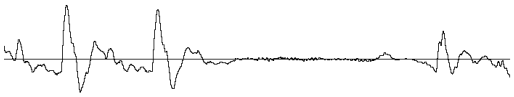


Figure 8(b) Distortion removed by further processing.

V. TESTS AND RESULTS

Several sentences were constructed using spoken words from different original sentences taken from the TIMIT database. Two of the sentences constructed were:

i. Oily like a taxicab.
ii. She just had a new project.

Two pitch contours were selected for the sentences: the inquisitive tone and the narrative tone. The pitch contours together with the contour of pitches of the original spoken words are presented in Figure 9.

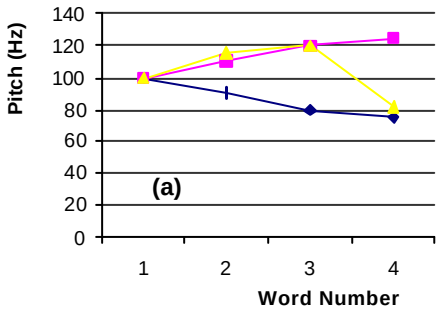
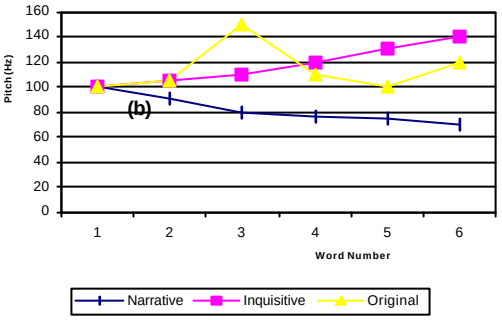


Figure 9(a). Pitch contours for "Oily like a taxicab"

Figure 9(b) Pitch contours for "She just had a new project".

Some slight distortion was discernible after the time and



pitch scale modification. However, the sentences produced with modified prosody sound much more natural compared with the speech obtained from simple concatenation.

VI. CONCLUSION

In this paper, a method to concatenate spoken English words to form natural sounding sentence is proposed. The method makes use of the sinusoidal model for parameterisation and the Time-Domain Pitch-Synchronous-Overlap-Add method to modify the duration and pitch of the words to fit the sentence. A novel approach of smoothing the concatenated segments further improves the quality of the synthesised speech.

Listening tests conducted reveal that the final synthesised speech demonstrates a high level of intelligibility and a significant level of naturalness.

VII. REFERENCES

[1] R.J. McAulay and T.F. Quatieri, "Speech Analysis/synthesis based on a sinusoidal representation," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-34, pp 744-754, August 1986.

[2] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Comm*., 9:453-467, Dec1990.

[3] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-32, No 2, pp 236-243, 1984.

[4] James D. Wise, James R Caprio and Thomas W. Parks, "Maximum Likelihood Pitch Estimation," IEEE Transaction on Acoustics, Speech and Signal Processing, vol. ASSP-24, No 5, October 1976.

[5] W.B. Kleijn and K.K. Paliwal, Speech Coding and Synthesis. Elsevier Science B.V., 1995.