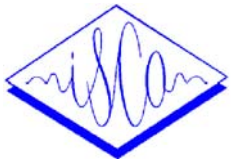


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A NOVEL MODEL TD-PSPTP FOR SPEECH SYNTHESIS

HUANG Yan XU Bo

National Laboratory of Pattern Recognition, Institute of Automation

Chinese Academy of Sciences, P. O. Box 2728, 100080, Beijing, P. R. China

E_mail: hy@nlpr.ia.ac.cn xubo@nlpr.ia.ac.cn

Tel: +86-10-82614495, Fax: +86-10-62551993

Abstract

In this paper, a novel approach based on time-domain pitch-synchronous point-to-point (TD-PSPTP) model for speech synthesis is presented. Compared to TD-PSOLA, which is currently one of the most popular concatenation methods, TD-PSPTP model provides a wider range of pitch and time modification. The quality of synthesized speech by TD-PSPTP shows to be high, especially its capability of overcoming reverberation, existing in TD-PSOLA when there is a drastic prosodic modification. The computational expense of TD-PSPTP model is no higher than that of TD-PSOLA. It provides an efficient way for the real time implementation of synthesis system.

10.21437/Eurospeech.1999-502

1. Introduction

One of the goals of speech synthesis is to enable a machine to transmit information orally to a user in a man machine communication context [1]. This requires the synthesized speech be natural or pleasant, which is always the most difficult subject in text-to-speech synthesis. In general, there are two key points in developing a high-quality speech synthesis system: one is how to find out the prosodic model, which can be used to control the prosodic feature; the other is how to build a flexible speech synthesizer, which permits to modify the prosodic features effectively. Here we focus on developing a flexible speech synthesizer.

A number of strategies have been forwarded for developing speech synthesizer. Sinusoidal model [2][3], harmonic plus noise model (HNM)[4], harmonic plus stochastic model (HSM)[5], and MBE model [6] are all effective representations in TTS, which have comparatively broad scope of modification and produce high quality synthesized speech. But the computational expense of these models is high. TD-PSOLA [7] is currently one of the most widely used concatenation approaches, which is very simple in computation and capable of producing synthesized speech with high quality. However, experiments show that the quality of synthesized speech based on TD-PSOLA will sharply decline as drastic prosodic modification is made. It has limitation in the range of prosodic modification. It is ideal to make least signal processing and at the same time guarantee the flexibility in prosodic modification in speech synthesis. One extreme is based on large database without signal processing at all, in which several samples of each pronunciation unit are recorded and then are concatenated. But the task of building such a huge database is really tough. So we think of developing a new algorithm, which has wider scope of time and pitch modification and meanwhile keeps the computation simplicity. TD-PSPTP model meets such goal in some sense.

In following sections a novel approach of speech synthesis based on time-domain pitch-synchronous point-to-point (TD-PSPTP) algorithm will be introduced, including pitch-synchronous analysis, time scale modification, pitch scale modification, time scale and pitch scale modification in one step, and concatenation. An experiment of mandarin speech synthesis based on TD-PSPTP is shown in section 3. Section 4 is a brief conclusion.

2. The Time-Domain Pitch-Synchronous Point-to-Point Model for Speech Synthesis

The TD-PSPTP synthesis scheme involves the following three steps: the first step is to make analysis of the original speech waveform to gain a sequence of pitch synchronous short-term signals, the second step is to do pitch and time scale modification, in which the short-term analysis waveforms are extended or(and) stretched to fulfil the corresponding time or(and) pitch scale modification, the last step is to concatenate short-term synthesis waveforms generated by step 2 with some kind of boundary smoothing.

2.1. Pitch-synchronous analysis

A series of short-term waveforms are obtained by multiplying the signal by a sequence of pitch-synchronous analysis windows $h_m(n)$:

$$x_m^a(n) = h_m(n - t_m)\,x(n)$$

Instants t_m are called pitch-marks, which are set at a pitch-synchronous rate on the voiced portions and at a constant rate on the unvoiced portions. $h_m(n)$ is a rectangle window, synchronous with pitch-marks. So actually $x_m^a(n)$ is a short-term analysis signal derived from original waveform signal based on the sequence of pitch-marks.

2.2. Time-scale modification

In speech synthesis the object of time-scale modification is to alter the rate of articulation without affecting the spectral content.

Compared to pitch-scale modification, time-scale modification is a comparatively easier problem. The most natural thought is to rewrite the original waveform and keep the phase continuity at the same time.

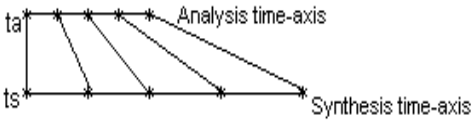
From the stream of analysis time-instants t_a^m and the desired time-scale modification factor β(t), the synthesis time-instants t_s^m will be determined. The mapping $t_a^m \rightarrow t_s^m$ is also obtained. The relation of t_s^m and t_a^m can be denoted by MT(t) :$t_a^m \rightarrow t_s^m$. MT(t) is referred to as the time-scale warping function which is defined as the integral of β(t):

$$MT(t) = \int_0^t \beta(\varphi)\, d\varphi$$

As to constant time-modification rate time, that is β(t)= β, then warping function:

$$MT(t) = \beta * t$$

Figure1 illustrates an example of the synthesis pitch marks computation for time scale modification by 2.0.



Time-scale modification by 2.0

Figure1

The following approach is based on the assumption of short-term periodicity. When a sequence of t_s^m are mapped to t_a^m, a series of short-time waveforms of original signal are acquired. Then recurrence and repetition are used to implement the time scale, $x_m^s(t)$ denoting synthesis short-term waveform:

$$x_m^s(t) = x_m^a[(t + t_{start}^m) \bmod (T_a^m)],\ T_a^m = t_a^{m+1} - t_a^m$$
$$0 <= t < T_s^m,\ T_s^m = t_s^{m+1} - t_s^m,$$

t_{start}^m is a parameter used to keep the time synchronization of successive synthesis waveform. It is defined as:

$$m = 0,$$
$$t_{start}^m = 0;$$
$$m > 0,$$
$$t_{start}^{m+1} = [(T_s^m + t_{start}^m) \bmod (T_a^m)] * \gamma, T_s^m = t_s^{m+1} - t_s^m;$$
$$\gamma = T_a^{m+1} / T_a^m, T_a^m = t_a^{m+1} - t_a^m;$$

Experiments show that it preserves original speech quality well even when a major time-scale modification is made. Parameter t_{start}^m shows to be effective in maintaining time synchronization between two successive short-term signals. By informal objective test, the quality of synthesis speech keeps high with time-scale modification factor β ranging from 0.5 to 2.5.

Following are examples of synthesized syllables based on TD-PSPTP model. Figure2. is the spectrum of the original waveform of [*Liu1*](a syllable in mandarin); Figure3 is the spectrum of the synthesis waveforms with time-scale modification factor β=0.6, 0.7, 0.8, 0.9 ; Figure 4 is the spectrum of the synthesis waveforms with time-scale modification factor β =1.1, 1.2, 1.3, 1.4.

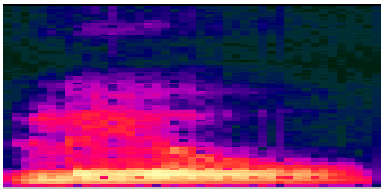


Figure 2.original spectrum of [*liu1*]

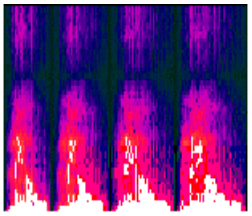
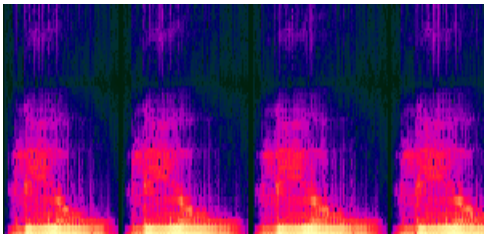


Figure 3.synthesis spectrum with time-scale modification β<1(β=0.6,0.7,0.8,0.9)

Figure 4 synthesis spectrum with time-scale modification



β>1 (β=1.1,1.2,1.3,1.4)

2.3. Pitch-scale modification

Pitch scale modification is very important in speech synthesis, especially in mandarin speech synthesis, which is a tonal language. An effective algorithm for pitch scale modification will surely improve the quality of synthesis. The goal of the pitch-scale modification is to alter the fundamental frequency of a speaker, which is relatively difficult especially when the pitch-scale modification factor α deviates much from 1.

From the stream of the analysis time-instants t_a^m, the series of synthesis time-instants t_s^m are also determined. Here assuming only dealing with pitch-scale modification and keeping the duration unchanged, we have $t_s^m=t_a^m$. There are two steps to fulfil the pitch-scale modification. Firstly, point-to-point mappings based on the pitch-scale modification factor α are made to get the short-term synthesis waveforms. Secondly, recurrence and repetition similar to that used in above mentioned time-scale modification are carried out to meet the requirement of duration. As to point-to-point mapping, interpolation algorithms should be taken into account. Linear interpolation and two-step polynomial interpolation are adopted, in which the result of two-step polynomial interpolation appears better than that of linear interpolation. The whole procedure can be described as following:

(1) point-to-point mapping and interpolating. Here xx(t) represents the new pitch-synchronous waveform of the synthesized speech, on which pitch scale modification is made. Set t_1=int(t*α), . The formation of function in($\bullet$) is determined by which interpolation algorithm is used, such as linear interpolation or two-step polynomial interpolation, etc.

$$0 <= t < T_S^m, T_S^m = t_S^{m+1} - t_S^m$$
$$if\ (t * \alpha = t_1)$$
$$xx_m^s(t) = x_m^a(t_1)$$
$$else$$
$$xx_m^s(t) = in[x_m^a(t_1), x_m^a(t_1 + 1)]$$

(2) recurrence and repetition

$$x_m^s(t) = xx_m^s[(t_s^m + t + t_{start}^m) \bmod (t_a^{m+1}/\alpha - t_a^m/\alpha)]$$
$$0 <= t < T_s^m, T_s^m = t_s^{m+1} - t_s^m$$
$$t_a^m = t_s^m$$

t_{start}^m is used to maintain the time synchronization of the synthesis waveform. Its definition is similar to the above mentioned in 2.2.

Experiments show that pitch modification factor β can range from 0.6 to 1.8. The quality of two-step polynomial interpolation is higher than that of linear interpolation and its modification range appears a little wider than the latter. Figure5 is the spectrum of the synthesis waveforms with pitch-scale modification factor α=0.6, 0.7, 0.8, 0.9; Figure6 is the spectrum of the synthesis waveforms with pitch-scale modification factor α=1.1, 1.2, 1.3, 1.4.

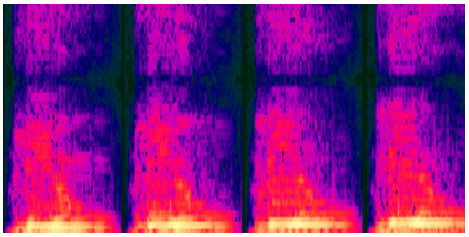


Figure 5: synthesis spectrum with pitch-scale modification α<1 (α=0.6, 0.7, 0.8, 0.9)

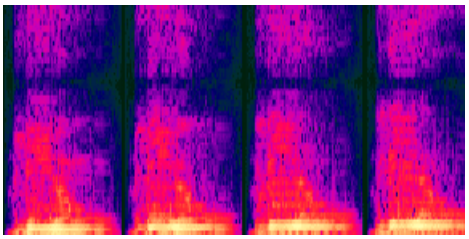


Figure 6: synthesis spectrum with pitch-scale modification α>1(α=1.1, 1.2, 1.3, 1.4)

2.4 Time-scale and Pitch-scale in one step

Practically, the task in speech synthesis always requires making both time and pitch scale modification. Of course we can do them separately. But it is unnecessary. These two modifications can be jointed into one step.

From 2.2 and 2.3 we can see that pitch-scale modification factor β determines the synthesis short-term waveform $xx_m(t)$, while time-scale modification factor α determines recurrence and repetition of $xx_m(t)$. Thus the procedure of making time-scale and pitch-scale in one step can be described as following:

(1) point-to-point mapping and interpolating.

Set t_1=int(t* α).

$$0 <= t < T_S^m, T_S^m = t_S^{m+1} - t_S^m$$
$$if\ (t * \alpha = t_1)$$
$$xx_m^s(t) = x_m^a(t_1)$$
$$else$$
$$xx_m^s(t) = in[x_m^a(t_1), x_m^a(t_1 + 1)]$$

(2) recurrence and repetition

$$x_m^s(t) = xx_m^s[(t_s^m + t + t_{start}^m) \bmod (t_a^{m+1}/\alpha - t_a^m/\alpha)]$$
$$0 <= t < T_s^m, T_s^m = t_s^{m+1} - t_s^m$$
$$t_s^m = t_a^m * \beta$$

Here the time and pitch scope modification is made in one step. Now a series of short-time synthesis waveforms are obtained, which can be concatenated to get synthesis speech.

2.5 Concatenation

Finally, the series of short-term waveforms $x_m^s(t)$ are to be concatenated. Some kind of boundary smoothing is also necessary here.

3 TD-PSPTP Model Used in Mandarin Speech Synthesis

A mandarin speech synthesis system is built based on TD-PSPTP. We select syllable, which is a natural unit in mandarin, as the synthesis unit. The sound file of the following sentence synthesized by our system can be referred. (h005PTP.WAV) :

[Zhou1 Wu3 Cheng2 Fei1 Ji1 Qu4 Huang2 Shan1, Qie3 Fei4 Yong4 Da4 Yu2 San1 Qian1 Si4 Bai3,Wo3 Men2 You3 Tuan2, Qi2 Jia4 Ge2 Shi2 San1 Qian1 Si4 Bai3 Ba1 Shi2.]

4 Conclusion

TD-PSPTP model is a high quality and low computational complexity strategy for speech synthesis. By informal subjective test the quality of synthesized syllables is satisfying, as time scale modification ranges from 0.5 to 2.5, or pitch scale modification ranges from 0.6 to 1.8,

The idea presented in this paper is to develop a flexible synthesizer with comparatively broad range of modification, but requiring least signal processing to keep the original speech information and satisfy real time implementation. TD-PSPTP model shows to be a promising approach for speech synthesis.

Reference

[1]L.R.Rabiner, “Application of Voice Processing To Telecommunications, ” *Proc.IEEE,* vol. 82,pp199-228, Feb 1994.

[2]Quatieri. T.F. and McAulay, .J.(1992), “Shape invariant time-scale and pitch-scale modification of speech”, IEEE Transactions on Signal Processing, March 1992, vol. 40, (no.3)”497-510.

[3]McAulay, R.J. and Quatieri, T.F, “Speech analysis / synthesis based on a sinusoidal representation”, *IEEE Trans.* ASSP-34. No.4 pp 744-754, 1986

[4]Y.Styliano, Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. Ph.D. thesis,

[5]Y.Styliano, “Decompositions of speech signals into a deterministic and a stochastic part”, *International* Conference on Speech and Signal Processing 1996, Oct. 1996

[6]T.Dutoit and H.Leich. “Text-to-speech synthesis based on a MBE re-synthesis of the segments database”, Speech Communication, 13: 435-440,1993

[7]E.Moulines and F.Carpentier, “Pitch-synchronous wave-form processing techniques for text-to-speech synthesis using diphones, ”*Speech Communication,* vol.9, pp.453-467, Dec 1990.