A HYBRID APPROACH TO SYNTHESIZE HIGH QUALITY CANTONESE SPEECH

Chu Min and P.C. Ching
Dept. of Electronic Engineering, Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: mchu@ee.cuhk.edu.hk, pcching@ ee.cuhk.edu.hk

ABSTRACT

Synthesizing high quality speech necessitates an intelligent modification algorithm to adjust the important prosodic features of the pre-stored speech units to meet the desired output requirements, such as smoothness, naturalness and pleasantness. The time domain pitch-synchronous overlap and add (TD-PSOLA) scheme is a simple but effective method of varying the pitch and time-scaling of speech and it can produce high quality synthetic output. However, when the prosodic pattern requires a drastic modification in the spectral content of the stored units, TD-PSOLA often generates speech with reverberant sound. This paper develops a new hybrid synthesis method based on TD-PSOLA and shape-invariant sinusoidal technique to alleviate the problem of reverberation. It is particularly useful for the generation of Cantonese speech, since it can cope with the rapidly changing of the pitch profile of Cantonese, which is a mono-syllabic and tonal language. The proposed method has been applied to construct a Cantonese synthesizer which is shown to be capable of producing high quality Cantonese speech without reverberation.

1. INTRODUCTION

Text-to-speech systems based on the concatenation of short speech units taken from a prerecorded inventory are widely used nowadays, mainly because they have a good compromise between complexity, flexibility and performance. These systems try to overcome the lack of knowledge of the human phonation process by considering the speech sounds in many different phonetic contexts and then applying a prosodic modification algorithm to adjust the prosodic features (pitch, duration and energy) of the stored speech units to satisfy the output perceptual requirements. In order to perform prosodic modifications without introducing unnatural-sounding artifacts, signal processing techniques, such as the popular PSOLA [1], are employed. The time domain version of PSOLA (TD-PSOLA) is a simple and effective method of varying the pitch and time-scaling of speech and can produce high quality synthetic speech. Several high quality TTS systems based on TD-PSOLA method have been reported[2][3][4]. However, TD-PSOLA method is known to suffer from spectral and phase distortions whenever the prosodic pattern requires a severe modification in the spectral content of the stored units. As a consequence, the synthetic speech often sounds reverberant [5]. To overcome this disadvantage, a hybrid synthesis method based on TD-PSOLA and shape-invariant sinusoidal technique[6] is developed in this paper. The new method can provide very smooth amplitude and phase transition between successive synthetic frames, which thus reduce the possible artifact of reverberation.

Cantonese is a widely used dialect in southern part of China and is well known of its very rich tonal contents. It is very sensitive to pitch movements as well as duration variations. This makes a severe demand on prosody modification if one is to generate perceptually pleasant Cantonese speech. Conventional synthesizers are usually unable to cope with the rapid changes of pitch profile in Cantonese whereas the proposed method provides a viable solution. A Cantonese synthesizer based on the newly developed hybrid technique is constructed with an attempt to demonstrate its capability in producing high quality speech.

This paper is organized as follows. In Section 2, the hybrid synthesis method based on TD-PSOLA and shape-invariant sinusoidal technique is developed, while the implementation of the Cantonese synthesizer is presented in Section 3. Section 4 gives a brief conclusion.

2. THE NEW SYNTHESIS METHOD

2.1 Brief Introduction of TD_PSOLA Method

The basic idea of TD-PSOLA is to obtain a pitch-modified version of a voiced speech sound by summing a sequence of windowed data $s_i(n)$ that is extracted pitch-synchronously from the original signal $x(n)$, and changing the time-shift between windows from the original pitch period T_0 to the desired one T. The transformed speech segment can be expressed as

$$s(n) = \sum s_i(n - i(T - T_0))$$

where

$$s_i(n) = x(n)w(n - iT_0)$$

For time-scale modification, some of the windowed data $s_i(n)$ might be deleted or repeated in order to obtain the desired duration.

When the distance between successive samples of $s_i(n)$ is changed, the continuity of phases between neighboring segments is also disrupted. If the change of pitch period is small, the discontinuity of phase can be smoothed by applying an overlap-add process to the successive data. However, if a drastic modification is desired, it is difficult, if not impossible, to resolve the phase problem. As a result, reverberation is often perceived in the synthetic speech .

2.2 The basic model of the new method

In order to obtain the smoothest possible amplitude and phase transition between successive pitch period, the windowed data of $s_i(n)$ are represented by a sinusoidal model. Mathematically, they can be denoted by summing a series of sine waves, viz.

$$s_i(n) = \sum_{l=1}^{L_i} A_l \cdot \cos[\omega_l n + \theta_l]$$

where A_l, ω_i and θ_l are the amplitudes, frequencies and phases of the sinusoidal components of $s_i(n)$, and L_i is the number of sine-waves being used to compose $s_i(n)$.

All these parameters can be obtained by taking DFT of $s_i(n)$ as described in [7]. In order to maintain high resolution of the periodogram, the window that is used to get $s_i(n)$ has the width of two and a half times of the local pitch period, and is normalized according to,

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

where N is the window width.

Since the analysis window is centered around an excitation point where the phases of all excitation sinusoids are assumed to be integer multiples of 2π , the respective phases at each peak frequency are, in principle, the phases of the vocal system. In natural circunstances, the excitation signal has very flat magnitude spectrum, and the magnitudes of excitation at all peak frequencies can be assumed to be unity. The measured magnitudes at peak frequencies are, therefore, closely related to the system amplitudes. Hence, the original speech can be represented by a set of parameters at each pitch onset time.

The target duration and pitch profile of the synthetic speech are decided by the prosody model, from which the pitch onset times of the synthetic speech are deduced. Before synthesis, a time warping algorithm is used to map the original pitch onset times, t_i, to the synthetic ones, t_j, according to the congruence between the original pitch profile and duration and the target values. There are three typical situations need to be considered (see Figure 1): two neighboring targets map onto two successive original onset times (Figure.1.(a)); one target pitch period maps to several successive original pitch periods (Figure.1(b)); and several successive target pitch periods map to one original pitch period (Figure.1.(c)). The pitch scale factor and time scale factor change all the time. Thus, it is a nonlinear process. The detail of the underlying fundamental of the warping process is given in [8].

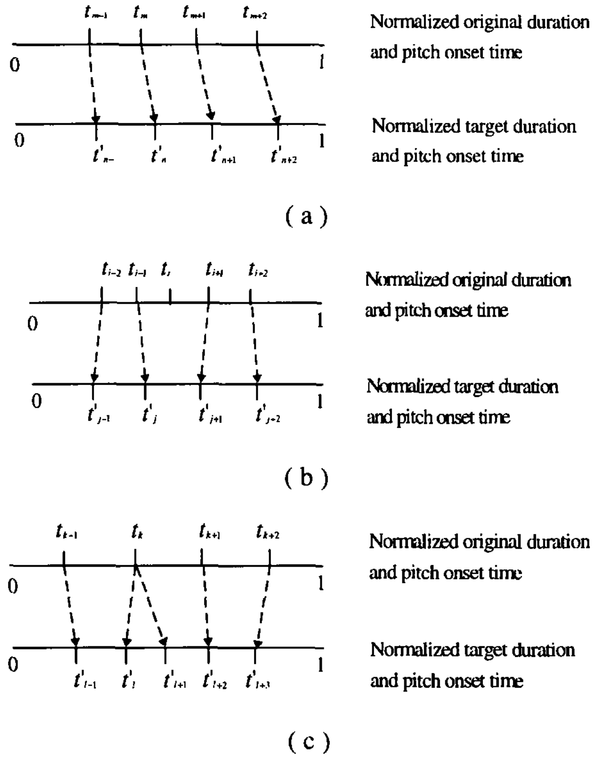
t_{m-1} t_m t_{m+1} t_{m+2}
0
1
Normalized original duration and pitch onset time
0
1
t'_{n-} t'_n t'_{n+1} t'_{n+2}
Normalized target duration and pitch onset time
(a)
t_{i-2} t_{i-1} t_i t_{i+1} t_{i+2}
0
1
Normalized original duration and pitch onset time
0
1
t'_{j-1} t'_j t'_{j+1} t'_{j+2}
Normalized target duration and pitch onset time
(b)
t_{k-1} t_k t_{k+1} t_{k+2}
0
1
Normalized original duration and pitch onset time
0
1
t'_{l-1} t'_l t'_{l+1} t'_{l+2} t'_{l+3}
Normalized target duration and pitch onset time
(c)

Figure 1. Three typical situations of time warping algorithm between original pitch onset times and target ones.

After time warping, the original peak frequencies are scaled by the local pitch scale factor to obtain the new peak frequencies for the target onset times. Then, a set of modified sinusoidal parameters for each target onset time are calculated from the corresponding parameters by sampling the amplitude and phase envelop at the new peak frequencies. During the synthesis process, amplitudes and phases are interpolated along frequency tracks to keep the smoothest transition between neighboring onset times and all sinewaves are summed together to get the synthetic speech with the desired pitch profile and duration. This step is very similar to the procedure adopted in the shape-invariant synthesis method except the following. In shape-invariant method, the end points of a synthetic frame is not identical with pitch onset time. If a cubic interpolation function is used with three restrictive conditions at times t=0 (i.e. the start point of a synthesis frame), t=T, (i.e. the end point of the synthesis frame) and t=Z (i.e. the pitch onset time of the synthetic frame) to interpolate the excitation phase, there is an unstable region for Z/T. One solution to the problem is to use a quadratic interpolation function[9], yet the algorithm is very complex. In the proposed algorithm, the problem is solved by forcing Z=T, i.e. the synthesis procedure is done pitch-synchronously.

The block diagram of the new method is given in Figure 2. It can produce very natural speech without reverberation. However, since extraction and interpolation of parameters together with the synthesizing process all requires heavy computation, the synthesizer in Figure 2 can not be realized in real time. We have investigated various means to reduce system complexity such that real time implementation is feasible.

2.3 Algorithm Optimization And Data Compression

To save computation time, some of the tasks should be done off-line. The sinusoidal parameters at the original pitch onset times can be calculated well in advance. However, there are too many parameters for each onset time, which requires a substantial amount of storage space than the original waveform. Data compression is necessary and is performed in the following steps.

> First, system amplitude envelop and phase envelop are interpolated from values at peak frequencies. The amplitude envelop is represented by 12 order cepstral coefficients. The phase envelop is decomposed by cosine analysis. All the cepstral vectors of system amplitude and the cosine coefficient vectors of system phase are vector quantized by LBG algorithm separately. The size of each codebook is 1024, and all amplitude and phase vectors are subsequently encoded by a table looking-up procedure.

After compression, the total storage space is less than 10MB for a syllable based Cantonese synthesizer. The data being stored include: two VQ codebooks (about100KB); two-byte codes for all amplitude and phase vectors (about 330KB); and frequency at onset times (about 8300KB, assume average number of frequency peaks is about 50).

On synthesis, amplitude and phase envelops at the original pitch onset times are reconstructed from their feature vectors, while the parameters at target onset times are obtained by resampling these envelops. The remaining steps are the same as those described in Section 2.2. The data compression causes no perceivable quality distortion.

According to the above setup, more than 90% of memory space is used for storing peak frequencies. If s higher data compression ratio is needed, harmonic frequencies of the fundamental frequency could be used instead. By so doing , the storage space would be reduce to about 170KB. However, the synthetic speech will sound a little machine like.

3. THE CANTONESE SYNTHESIZER

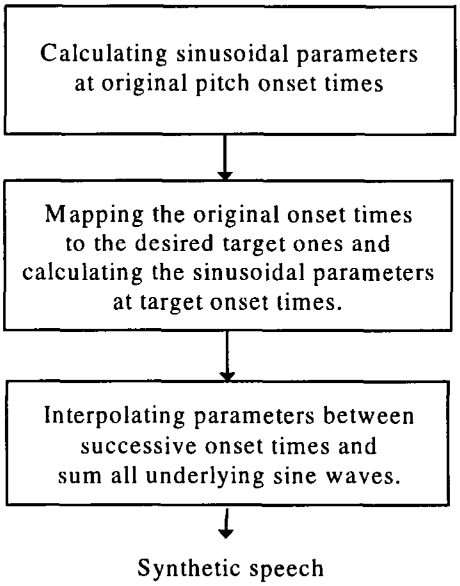
Calculating sinusoidal parameters at original pitch onset times
Mapping the original onset times to the desired target ones and calculating the sinusoidal parameters at target onset times.
Interpolating parameters between successive onset times and sum all underlying sine waves.
Synthetic speech

Figure 2. The block diagram of the proposed synthesis method

Cantonese is a widely used dialect in southern part of China and is well known of its very rich tonal contents. Conventional synthesizers are usually unable to cope with the rapid changes of pitch profile in Cantonese. The proposed method has been applied to construct a Cantonese synthesizer with an attempt to produce natural output speech.

The block diagram of the Cantonese synthesizer is given in Figure 3. At the present stage, the input of the synthesizer is Cantonese transcriptions. There is a simple text pre-processing module which scans the input text and extracts phonetic and structural information (such as the phonetic structure of the syllable, its tone and position in the sentence) for the prosody model. The prosody model decides the target duration and pitch profile of the synthetic syllable. More details about the prosody model can be found in [8]. About 1600 syllables are used as basic synthetic units. These syllables are preprocessed according to section 2.3 to obtain a 10MB syllable parameter dictionary. Through informal listening tests by several native Cantonese speakers, the synthetic speech produced by the synthesizer is confirmed to be clear, fluent and natural.

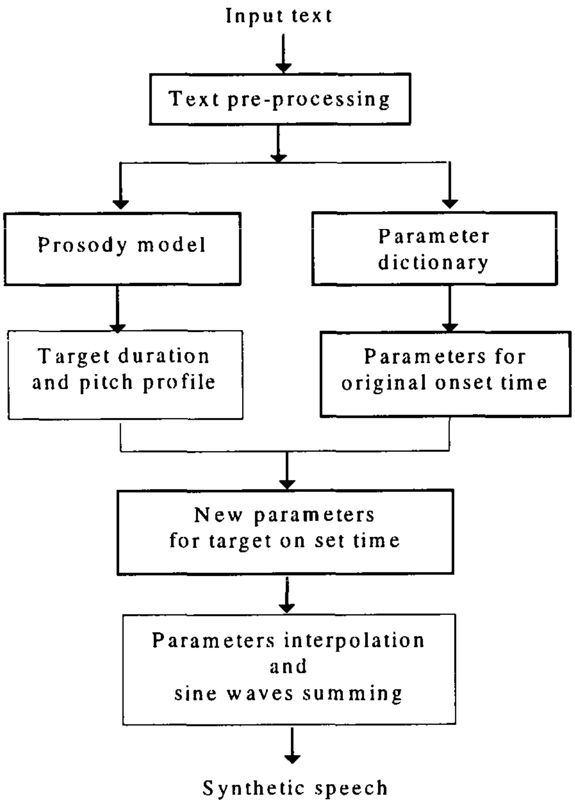
Input text
Text pre-processing
Prosody model
Parameter dictionary
Target duration and pitch profile
Parameters for original onset time
New parameters for target on set time
Parameters interpolation and sine waves summing
Synthetic speech

Figure 3. The block diagram of the Cantonese synthesizer

4. SUMMARY

This paper presents a hybrid synthesis method that attempts to eliminate the reverberation problem caused by phase discontinuity in TD-PSOLA. The new method incorporates sinusoidal representations into TD-PSOLA and keeps smooth transition between neighboring synthetic segments by amplitude and phase interpolation. It has been applied to construct a Cantonese Synthesizer. The output speech quality has been informally evaluated by several native Cantonese speakers, who confirm the synthetic speech to be clear and natural and perceived no reverberation. A formal speech quality evaluation is planning, in which the synthetic speech generated by the new method will be compared and contrasted with the one generated by the TD-PSOLA method.

REFERENCE

[1] Moulines, E. and Charpentier, F.(1990), "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Dophones", Speech Communication 9 (1990), P.453-467

[2] Moulines, E., Emerard, F., Larreur, D., Le Saint Milon, J.L., et al.(1990), "A Real-Time French Text-To-Speech System Generating High-Quality Synthetic Speech", Proc. of ICASSP-90, P.309-312.

[3] Bigorgne, D., et al.(1993), "Multilingual PSOLA Text-to-Speech System", Proc.of ICASSP-93, P.(II-187, II-190).

[4] Chu, M. and Lu, S.N.(1995), "High Intelligibility and Naturalness Chinese TTS System and Prosodic Rules", Proc. of XIII ICPhS, P.2:334-2:337

[5] Eduardo R. Banga and Carmen Gercia-Mateo (1995), "Shape-invariant pitch-synchronous text-to-speech conversion, Proc. of ICASSP,95, Vol 1, P 656-659

[6] Quatieri, T.F. and McAulay, R.J.(1992), "Shape invariant time-scale and pitch modification of speech", IEEE Trans on Signal Processing Vol.40, p. 497-510, 1992

[7] McAulay, R.J. and Quatieri, T.F(1986)., "Speech analysis/synthesis based on a sinusoidal representation", IEEE Trans. ASSP-34, No.4 P.744-754, 1986

[8] Chu Min and P.C. Ching, "A Cantonese Synthesizer Based on TD-PSOLA Method", to be presented in the 1997 ISMIP, Taipei, Taiwan.

[9] Pollard, M.P., Cheethman, B.M.G., Goodyear, C.C. and Edgington, M.D. (1996), "Phase interpolation methods for pitch and tine-scale modification of voiced speech", Institute of Acoustics Autumn Conference 1996 (Speech & Hearing).