NOTES FOR THE SYLLABLE-SIGNAL SYNTHESIS METHOD: TIPW

Hung-Yan GU
Department of EE, National Taiwan University of Science and Technology
43 Keelung Road, Section 4, Taipei 106
E-mail: root@guhy.ee.ntust.edu.tw FAX: 886-2-27376699

ABSTRACT

In this paper, the drawbacks found in PSOLA is briefly discussed. To eliminate these drawbacks, the syllable-signal synthesis method TIPW that was proposed in another work of us is recommended here. The processing steps of TIPW will be briefly described. Besides largely reducing the drawbacks of PSOLA, TIPW also provides a new factor (in addition to duration and pitch contour) for timbre control. Nevertheless, it has its own minor problems, i.e. occasional clicks and slower signal-synthesis speed. In this paper, these two problems are studied. The results are that occasional clicks can now be fully prevented and the speed of signal synthesis is nearly doubled.

1. INTRODUCTION

TIPW (time-proportioned interpolation of pitch waveforms) is a Mandarin-syllable signal synthesis method proposed by us [1,2]. It is a time-domain processing method and may be viewed (in some scenes) as derived from PSOLA [3,4]. The design of TIPW was motivated by the drawbacks found in PSOLA. For examples, the effects of reverberation and dual-tones (or called chorus) are often heard when syllable signals are synthesized with PSOLA. The effect of dual-tones means that two different tones (one low and one high) are simultaneously heard. This occurs when the pitch contour of a synthetic syllable is much higher or lower than the pitch contour of its original syllable. The cause that results in such effects is the lack of pitch-length synchronization between a synthesized syllable and its original syllable, i.e. signal windows' lengths are determined only according to the pitch lengths in the original syllable waveform without the pitch lengths in the synthesized syllable being considered. In fact, the synchronization considered in PSOLA is just pitch-location synchronization, i.e. signal windows are placed centrally around pitch peaks. In addition, there is another serious problem with PSOLA. The formant frequency traces will be nonlinearly warped (or have discontinuities generated) when the tone of a synthetic syllable is different from the tone of its original syllable or the duration of a synthesized syllable is set to be longer or shorter than the duration of its original syllable.

On the contrary, the effects of reverberation and dual-tones will not be heard if TIPW is adopted. Also, interference incurred by syllable duration and pitch contour in formant frequency traces will be largely reduced. Furthermore, a new control factor, vocal-track length, is provided in TIPW. By independently setting the parameters of pitch contour and vocal-track length in a reasonable value range, many distinct timbres (no the phenomenon of a male mimicking a female) of a child, a female, and a cartoon actor can be synthesized from the original syllable signals collected from a male adult. Therefore, the method of TIPW not only can eliminate the drawbacks of PSOLA but also can support more independent control of pitch contour, duration, and timbre. For testing, some example signal files synthesized by TIPW can be got from http://guhy.ee.ntust.edu.tw. Although TIPW is better than PSOLA from the viewpoints mentioned, it does have its own minor problems. In this paper, the observed minor problems are studied. The problems are occasional clicks and slower synthesis speed. In Sections three and four, each problem and the proposed solution method are described while the processing steps of TIPW are briefly described in Section two.

2. THE METHOD OF TIPW

In this section, the method of TIPW will be briefly described. The details are referred to another work of us [1]. In TIPW, the signal of a Mandarin syllable is considered to be the concatenation of an unvoiced part and a voiced part. If a syllable is entirely periodic, the part of its signal preceding the first pitch peak is considered to be the unvoiced.

2.1 Synthesis of Unvoiced Part

Before synthesizing a syllable's signal, the durations of the unvoiced and voiced parts must be determined first. Note that these two parts are not linearly extended when a syllable is pronounced slower than normal. In addition, the unvoiced part of a syllable must be classified first before its duration can be determined. In TIPW, two classes of unvoiced parts are defined, called short-unvoiced and long-unvoiced. The class short-unvoiced is intended to include those syllables with initial phonemes which are non-aspirated stop, nasal, glide, liquid, or vowel. On the other hand, the class long-unvoiced is intended include those syllables with initial phonemes which are fricative, aspirated or non-aspirated affricate, or aspirated stop. If the unvoiced part of an original syllable is short-unvoiced, the signal portion preceding the first pitch peak will be directly copied to the synthesized syllable to form its unvoiced part. On the other hand, if the unvoiced part is long-unvoiced, the duration of this part in a synthesized syllable will be determined according to the time proportion of its corresponding part in the original syllable. Then, the assigned duration is checked to see whether it is greater than the duration limit, 1.5 times of the duration of the unvoiced part in the original syllable. The assigned duration of the unvoiced part will be changed to the value of the duration limit when it is greater than this limit.

After the duration of the unvoiced part (long-unvoiced) is determined, the signal waveform of this part is synthesized in two steps. First, the leading 300 signal samples of the original syllable (under the sampling rate 11,025Hz) are directly copied to the leading portion of the synthesized syllable. This step is intended to reserve the initial stop characteristics of the affricate phonemes. Secondly, the remaining signal samples of the unvoiced part are synthesized by means of time-proportioned mapping and interpolation. Suppose that Tx is the number of samples in the synthesized unvoiced part, Ty is the number of samples in the original unvoiced part, x is a sample point within the synthesized unvoiced part, and y is the sample point in the original unvoiced part to be mapped from x. Then, y is computed as ((x-300) / (Tx-300)) (Ty-300) + 300. After y is computed, the sample value in the position x is computed by linearly interpolating the two adjacent samples around y.

2.2 Synthesis of Voiced Part

To synthesize the voiced part of a syllable, the lengths (in sample points) of all the pitch periods in this part are computed first according to the given parameters for pitch-contour control. Then, the signal samples in successive pitch periods are synthesized in order. In fact, the name TIPW is derived from the procedure used to synthesize the signal samples of a pitch period. This procedure has five steps as described below.

2.2.1 Finding Two Corresponding Pitch Periods

In TIPW, a pitch period is meant the signal portion bounded by two adjacent pitch peaks. Also, the time position of a pitch period is defined by the time position of the central sample point within it. According to these definitions for pitch period and time position, two adjacent pitch periods, in the original syllable, corresponding to the pitch period to be synthesized are found with the criterion of time proportion. That is, the normalized (divided by the duration of the voiced part) time positions of the two pitch periods found must surround the normalized time position of the pitch period to be synthesized.

2.2.2 Re-sampling

If only the pitch contour is raised, the speech synthesized by using the original signal waveform collected from a male will be heard as a male mimicking a female's speech. To solve this problem, we had studied and proposed a new control factor, i.e. vocal-track length control. By independently setting the pitch contour and vocal-track length, many distinct timbres can be synthesized. In fact, vocal-track length control is achieved by resampling the signal samples of the two pitch periods found in 2.2.1. If a woman's or a child's timbre is intended, the numbers of samples in each of the pitch periods must be decreased (under the same sampling rate) to shorten the vocal track. That is, the n'th sample point in the resampled waveform is mapped to the m'th point in the original waveform with $m = c \cdot n$ and c being a constant of value greater than 1. On the contrary, if an old man's timbre is intended, the mapping constant must be set to a value less than 1. This also increases the lengths of the two pitch periods.

2.2.3 Weighting Two Pitch Periods

Because the two pitch periods found will be weighted and combined to synthesize the waveform of a synthetic pitch period, the weights for the two pitch periods must be determined beforehand. Here, the weights are computed according to time proportion. Suppose that α and β are the normalized time positions of the two pitch periods found, and γ is the normalized time position of the pitch period to be synthesized. Then, the weight for the first pitch period found is computed as $w1 = (\beta-\gamma) / (\beta-\alpha)$ and the weight for the second pitch period is computed as $w2 = (\gamma-\alpha) / (\beta-\alpha)$. In terms of these weights, the amplitudes of the signal samples in the first pitch period are scaled by w1 and the signal amplitudes in the second pitch period are scaled by w2.

2.2.4 Windowing and Aligning

In general, the lengths of the synthesized pitch period and the two original pitch periods are mutually different. Therefore, the signal waveforms in the two original pitch periods must be windowed. Also, the length of the window function must be carefully determined in order to prevent the effects of dual-tone and reverberation. Because the signal waveform of the voiced part will be synthesized as the concatenation of pitch periods and a pitch period is bounded by two pitch peaks, here two half window functions, Wl and Wr, are used to window an original pitch period. Wl represents the right half of a Hanning window and its peak is placed and aligned with the left boundary of a pitch period, and Wr represents the left half of a Hanning window and its peak is placed and aligned with the right boundary of a pitch period. If the length of the original pitch period under windowing is greater than the length of the synthesized pitch period, both Wl and Wr will be set to have the length of the synthesized pitch period. Otherwise, both Wl and Wr will be set to have the length of the original pitch period. After windowing, the waveform portion windowed by Wl will be placed and aligned with the left boundary of the synthesized pitch period while the waveform portion windowed by Wr will be placed and aligned with the right boundary of the synthesized pitch period.

2.2.5 Overlapping and Adding

Because two original pitch periods are found for a pitch period to be synthesized and two half window functions are used for an original pitch period, there are four windowed waveform portions after Step 2.2.4. So, in this step, these four waveform portions are overlapped and added to form a synthesized pitch-period's waveform.

3. OCCASIONAL CLICKS

In some synthetic syllables, the phenomenon of waveform discontinuity is occasionally seen at the boundary point between two adjacent pitch periods under some combinations of parameter values (pitch contour, duration, and vocal-track length). A discontinuity is an abrupt amplitude change between two adjacent signal samples and is usually heard as a click added upon a normal syllable voice. According to our analysis, the causes that may lead to a discontinuity are: (1) At least one pitch period within an original syllable can be found, which has large amplitude difference between its left and right boundary samples; (2) The pitch contour of a synthetic syllable is raised (or lowered) two or more times of the pitch contour of its original syllable, or the duration of a synthetic syllable is extended more than two times of the duration of its original syllable. With the second cause, the normalized (divided by duration) time proceeded per pitch period in the synthesized syllable is less than one half of the one in its original syllable. When this is combined with the first cause, it may occur that a pitch period with large amplitude difference between its left and right boundary samples becomes a dominator in synthesizing some two adjacent pitch periods of a synthetic syllable. Then, a waveform discontinuity located between the two adjacent synthesized pitch periods may be generated. Note that in TIPW, a synthesized pitch period can somewhat be viewed as the weighting sum of its two corresponding adjacent pitch periods in the original syllable.

To eliminate such kind of discontinuities, we have studied it here and proposed a method, called PDGC (pitch-wise dynamic gain control). With this method, the annoying occasional clicks can be eliminated while signal clarity is kept. PDGC is consisted of two processing steps as described below.

3.1 Determine Boundary Samples' Amplitudes

Before the processing steps in TIPW are

performed to synthesize signal samples of a pitch period, the final amplitude values of the left and the right boundary samples are determined first. The determination method here is also based on the idea of time proportion. To explain it more concretely, let Tu and Tv in Fig. 1 be the total numbers of sample points in the original and synthesized syllables respectively, Ta, Tb, and Tc be the sample points of the first, second, and third pitch peaks in the original syllable, Ts be the first pitch peak in the synthesized syllable, and Tp and Tq be the boundary sample points of the pitch period to be synthesized. Then, the corresponding points, Tx and Ty, in the original syllable for Tp and Tq are computed according to time proportion as

$$Tx = \frac{Tp - Ts}{Tv - Ts}(Tu - Ta) + Ta$$
$$Ty = \frac{Tq - Ts}{Tv - Ts}(Tu - Ta) + Ta \qquad (1)$$

Suppose that Tx is located between the two pitch-peak points, Tb and Tc, and the sample amplitudes at Tb and Tc are Ab and Ac respectively. Then , the final signal amplitude at the point Tp, denoted as Ap, is defined according to linear interpolation as

$$Ap = \frac{Tx - Tb}{Tc - Tb}(Ac - Ab) + Ab \qquad (2)$$

Similarly, the final signal amplitude at the point Tq, denoted as Aq, can be computed.

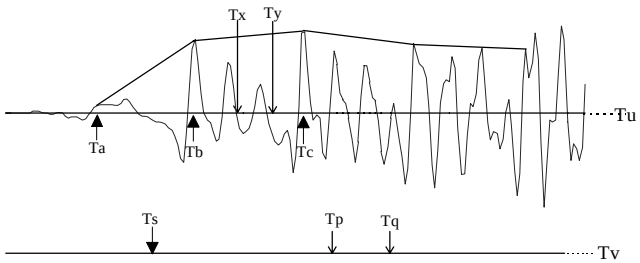

Fig. 1 Example waveform for demonstrating boundary-sample amplitude determination.

3.2 Dynamic Gain Computation

Before TIPW is used to synthesize the signal samples between Tp and Tq in Fig. 1, the amplitudes at the points Tp and Tq are first computed by TIPW. Suppose the computed amplitudes at Tp and Tq are Bp and Bq respectively. In general, Bp ≠ Ap and Bq ≠ Aq. To adjust Bp to Ap and adjust Bq to Aq, a method of pitch-wise dynamic gain control that can satisfy this requirement is hence proposed. In details, let $S(t)=S_{tipw}(t)\cdot G(t)$ where S(t) is the final signal amplitude at point t and $S_{tipw}(t)$ is the signal amplitude computed by using TIPW. Then, the time-varying gain function G(t) is defined as

$$G(Tp) = Ap / Bp , \qquad G(Tq) = Aq / Bq ,$$
$$Gd = (G(Tq) - G(Tp)) / (Tq - Tp) ,$$
$$G(t) = G(Tp) + Gd*(t - Tp), \qquad Tp < t < Tq \qquad (3)$$

We have programmed this method into software and the annoying clicks are not heard now. Also, no notable side-effects are heard.

4. SIGNAL SYNTHESIS SPEED

Because more computations are performed in TIPW, PSOLA is faster than TIPW. For example, to synthesize a signal sample, only one or two calling to cosine function are needed for PSOLA but four calling are needed for TIPW. Values of cosine functions are computed because Hanning windowing is used in both TIPW and PSOLA. Inspecting the processing procedure of TIPW, we find that the most time consuming computations are cosine function evaluation and re-sampling processing using quadratic polynomial approximation. By speeding up these kinds of computations, we think that the difference in signal-synthesis speed between TIPW and PSOLA can be reduced a lot.

To save time spent in computing cosine function values, a method of table-lookup is used. That is, the cosine function values that may be used are all computed once at program launching time and saved in a table with two indices, denoted as CosTab(I1, I2). Suppose that the sampling rate adopted is 11,025Hz, and the range of accepted fundamental frequencies is from 30Hz to 500Hz. Then, the possible integer values of pitch-period lengths, in sample points, are 22 (11,025/500), 23, ..., 368 (11,025/30). These values of pitch-period lengths are used as the first index. In addition, with the symmetry characteristics, only one forth of a period of Hanning window values needed to be

saved, i.e. the possible second index values are 0, 1, 2, ..., I1/4, where I1 represents the first index value. For examples, cos(-x) = cos(x) and cos(π/2 + x) = -cos(π/2 - x). With CosTab(I1, I2), a cosine function value can then be directly looked up.

In TIPW, to synthesize the signal samples of a pitch period, two corresponding adjacent pitch periods in the original syllable must be found first. Suppose that Q_i and R_i represent the two corresponding adjacent pitch periods for P_i, Q_{i+1} and R_{i+1} represent the ones for P_{i+1}, and P_i and P_{i+1} represent two adjacent pitch periods to be synthesized. Then, Q_{i+1} will usually be R_i. If pronunciation speed is slowed down, it may occur that Q_{i+1} equals Q_i and R_{i+1} equals R_i (However, the weights for Q_{i+1} and Q_i will be surely different). These indicate that re-sampling processing made for R_i can be used to synthesize both P_i and P_{i+1}, i.e. redundant re-sampling processing can be prevented to save time by buffering re-sampled samples.

We have programmed the ideas mentioned and practically compared the time spent in three conditions, denoted as Orig (original), CosT (with cosine table), and CosT+PRR (cosine table and preventing redundant re-sampling). The measured average time spent are as listed in Table 1. In this table, the number at the left of a cell represents the numbers of seconds needed to synthesize one second of speech samples, and the number of percentage at the right represents the relative time consumed within a row. From the first row, it can be seen that for a 486-33 personal computer, the processing time can be reduced from 1.121sec. to 0.515sec., i.e. 54% time saving and rendering a non-real-time processing into a real-time processing. Also, from the second row, it can be found that the processing time 0.133sec. is reduced to 0.0773sec., i.e. 42% time saving.

Table 1 CPU time spent in different conditions.

CPU	Orig	CosT	CosT+PRR
486-33	1.121, 100%	0.761, 67.9%	0.515, 45.9%
Pentium-133	0.133, 100%	0.105, 78.9%	0.0773, 58.1%

5. CONCLUSION

In this paper, the Mandarin-syllable signal synthesis method TIPW is recommended in order to reduce the drawbacks found in PSOLA. According to our study, the effects of dual-tone and reverberation found in PSOLA can indeed be eliminated by TIPW. In addition, the control factor of vocal-track length newly provided by TIPW can indeed be used to synthesize distinct timbres. Although TIPW is better than PSOLA in the viewpoints mentioned, it does has its own minor problems, i.e. occasional clicks and slower synthesis speed. These two problems are therefore studied here and the results are: (1) occasional clicks can now be fully prevented by the proposed method of pitch-wise dynamic gain control; (2) the speed of signal synthesis is nearly doubled by table-lookup of cosine function values and preventing redundant resampling processing.

6. ACKNOWLEDGMENT

This work was supported by the National Science Council under the contract number NSC 86-2213-E011-066.

7. REFERENCES

[1] Gu, Hung-Yan and Wen-Lung Shiu, "*A Mandarin-Syllable Signal Synthesis Method with Increased flexibility in Duration, Tone and Timbre Control*", Proceedings of the National Science Council, R.O.C., Part A: Physical Science and Engineering, Vol. 22, No. 3, pp. 385-395, May 1998.

[2] Gu, Hung-Yan, "*A Mandarin-Syllable Signal Synthesis Method with Increased Flexibility in Independent Control of the Parameters and the Capability to Generate Many Timbres*", R.O.C. Patent No. 087899, Nov. 1997.

[3] Charpentier, F. and M. Stella, "*Diphone synthesis using an overlap-add technique for speech waveform concatenation*", IEEE Int. Conf. ASSP (Tokyo, Japan), pp. 2015-2018, 1986.

[4] Modulines, E. and F. Charpentier, "*Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones*", Speech Communication, Vol. 9, pp. 453-467, 1990.