Proc. Natl. Sci. Counc. ROC(A)
Vol. 22, No. 3, 1998. pp. 385-395

A Mandarin-syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control

HUNG-YAN GU AND WEN-LUNG SHIU

Department of Electrical Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan, R.O.C.

(Received July 19, 1997; Accepted October 7, 1997)

ABSTRACT

The syllable is commonly adopted as a synthesis unit in Mandarin text-to-speech systems. Therefore, the problem of syllable signal synthesis is studied and a new synthesis method is proposed in this paper. When compared with other time-domain synthesis methods, our method not only can synthesize clear speech signals, but also provides more flexibility in control of the duration, tone (pitch contour), and timbre. The duration of a syllable should be adjustable to reflect the influence of pronunciation speed and other relevant factors. The tone should be changeable because it is desired that syllables of other tones can be synthesized from the First-tone syllable to save memory. The vocal track length intrinsic in a syllable's waveform should be adjustable because it is desired that the speech of females (or cartoon actors) synthesized from the original syllable signals of a male be perceived as more natural. We find that many distinct timbres could be synthesized if both the vocal track length and pitch-contour's height were adjusted simultaneously. Here, the vocal track length is a newly studied factor and timbre control becomes realizable using this factor. Although the ability to control a syllable's duration and pitch contour is also provided by other time-domain synthesis methods, our method provides more flexibility and largely decreases the interference the traces of formant frequencies incurred due to these two factors.

Key Words: speech synthesis, Mandarin speech, waveform interpolation

I. Introduction

There are two main components of a text-to-speech system. One of the components is called the prosodic processing unit. Its work is to find the sequence of syllables corresponding to an inputted sentence and to then assign values to the prosodic parameters of each syllable. The syllable is commonly adopted as a unit because there are only about 410 different syllables (if different tones are not distinguished) in Mandarin speech, and a First-tone syllable can be used to synthesize syllables of the same phonemes but of other tones. The main prosodic parameters of a syllable include the pitch contour (tone and intonation), duration, intensity, and pause preceding a syllable. The other component in a text-to-speech system is called the signal synthesis unit. Its work is to synthesize the corresponding signal waveform according to the syllable and prosodic parameters given by the prosodic processing unit. It not only has to satisfy the values of the prosodic parameters, but also has to synthesize speech signals as clearly as possible. This paper will focus on the synthesis of speech signals, i.e., the signal synthesis unit of a text-to-speech system.

Although many Mandarin text-to-speech systems have been proposed in the past (Lee *et al.*, 1993; Chiou *et al.*, 1991; Chen *et al.*, 1992; Wu *et al.*, 1995), these efforts mainly focused on the prosodic processing unit or did not provide a signal synthesis unit which is as flexible as that provided here. With regard to speech signal synthesis, a few techniques have been proposed previously, e.g., techniques based on LPC (linear prediction coding) (Makhoul, 1975; Markel and Gray, 1976) or formant synthesis (Klatt, 1980; Holmes, 1983). An important drawback of the LPC based techniques is that the synthesized signal is not very clear. Although formant synthesis based techniques can synthesize speech signal more clearly, the acoustic control parameters have to be tried and recorded manually (there is still no good automatic procedure). Recently, a technique based on time-domain waveform processing was proposed (Charpentier and Stella, 1986; Hamon *et al.*, 1989; Modulines and Charpentier, 1990), which is called PSOLA (pitch-synchronous overlap and add).

It can synthesize speech signals very clearly, and the required signal-synthesis parameters, pitch peaks, are well defined and can nearly be automatically detected. Nevertheless, it has a drawback in synthesizing Mandarin speech; i.e., formant frequency traces are nonlinearly warped when the values of the factors duration and pitch contour must be satisfied. More recently, a few variants of PSOLA have been proposed (Dutoit and Leich, 1993; Kawai *et al.*, 1994; Galanes *et al.*, 1994). However, they either do not solve the problem mentioned above or they induce side effects. For example, the technique LP (linear prediction)-PSOLA (Galanes *et al.*, 1994) induces small noise signals that make the synthesized signal noisy. To understand the drawback of PSOLA in more detail, suppose the signal of a Forth-tone (falling tone) syllable /ai/ is to be synthesized from the signal waveform of a First-tone (level tone) syllable /ai/. For convenience of explanation, let the two phonemes /a/ and /i/ of the First-tone syllable /ai/ be pronounced with equal duration. For the tone of the synthesized syllable to be heard as falling with time, its pitch periods must become longer with time. Then, in the synthesized Forth-tone syllable, the /i/ portion will be significantly longer than the /a/ portion if the number of pitch periods is kept the same. If the /i/ portion is longer than the /a/ portion, this indicates that the formant frequencies trace is nonlinearly warped. If the duration is to be kept the same, the number of pitch periods must be decreased. In PSOLA, a few pitch periods are directly removed from the original signal waveform in order for the synthesized signal to have the same duration and for its tone to fall with time. If pitch periods are directly removed, this also means that the formant frequency trace is nonlinearly warped (inducing discontinuity actually).

In order to eliminate the drawback of PSOLA, a Mandarin-syllable signal synthesis method has been studied and is proposed in this paper, called TIPW (Time-Proportioned Interpolation of Pitch Waveform). It also processes the recorded time-domain waveform directly, and the synthesized speech signals are clear enough. In addition, interference incurred by the duration and pitch contour in formant frequency traces is largely reduced. Therefore, our method can provide more flexibility in control of the duration and pitch contour. Furthermore, another freedom of control, i.e., the vocal track length, is also investigated here, which has not been used in other time-domain synthesis methods. Vocal track length is a very important factor because it can be used to prevent the phenomenon where a synthesized speech signal is heard as a man mimicking a woman's voice when a speech signal of a female is synthesized by only raising the pitch contours of the original syllable signals collected from a male. Therefore, we think that there are other factors that enable us to distinguish a male's speech from a female's speech, and that the vocal track length is an important factor. This can be clearly seen because the average length of females' vocal tracks is shorter than that of males' (O'Shaughnessy, 1987; Rabiner and Juang, 1993). In addition, we find that many timbres can be synthesized if the two factors, the pitch-contour height and vocal track length, are simultaneously adjusted. Furthermore, in our syllable-signal synthesis method, the three factors of duration, pitch contour, and vocal track length can be independently controlled within a reasonable range of parameter values. This is shown by practically implementing a prototype text-to-speech system.

In the following, the signal-synthesis procedure for the unvoiced part of a syllable will be given in Section II while the procedure for the voiced part will be given in Section III. In Section IV, a prototype text-to-speech system will be briefly described, which is implemented in order to verify our syllable-signal synthesis method. Also, results of spectrogram analysis for synthesized syllable-signals will be explained. Finally, a conclusion will be given in Section V.

II. Synthesis of Unvoiced Part

The signal of a Mandarin syllable is considered to be the concatenation of an unvoiced part and a voiced part. The unvoiced part which has a random waveform corresponds to one of the following phonemes: stop, fricative, and affricate consonants. The voiced part which has a periodic waveform corresponds to one of the following phonemes: nasal, glide, liquid, and vowels. If a syllable's waveform is entirely periodic, it can still be considered as having an unvoiced part; i.e., the part of the signal preceding the first pitch peak is considered to be unvoiced. As a result, all Mandarin syllables can be considered as having the UV structure, where U and V represent the unvoiced and voiced parts, respectively.

Before synthesizing a syllable's signal, the durations of the unvoiced and voiced parts must be determined first according to the duration limit of the unvoiced part and the syllable-duration parameter received from the prosodic processing unit. A limit for the duration of the unvoiced part is required because the durations of the unvoiced and voiced parts are not linearly extended when a syllable is slowly pronounced. For example, the unvoiced part /p/ of the syllable /pa/ is not extended at the same rate as the voiced part /a/ when /pa/ is slowly pronounced. In addition, before

the duration of the unvoiced part can be determined, the unvoiced part must be classified. For example, we should not set both of the phonemes /b/ and /p/ to have the same duration because the phoneme /p/ is aspirated but /b/ is not and an aspirated phoneme has longer duration in practice. Therefore, in our synthesis method, the unvoiced part is classified first. Then, the durations of the unvoiced and voiced parts are determined in order. When the durations are known, the signal waveform can then be synthesized. The synthesis method for the voiced part will be given in the next section. In the remainder of this section, the synthesis method for the unvoiced part will be described.

In our synthesis method, we define two classes into which the unvoiced part of each syllable should be classified, called short-unvoiced and long-unvoiced. The class short-unvoiced is intended to include those syllables with initial phonemes which are non-aspirated stop (e.g. /ba/), nasal (e.g. /na/), glide (e.g. /wa/), liquid (e.g. /la/), or vowel (e.g. /a/). On the other hand, the class long-unvoiced is intended to include those syllables with initial phonemes which are fricative (e.g. /ha/), aspirated or non-aspirated affricate (e.g. /tsa/, /dza/), or aspirated stop (e.g. /pa/). In practice, it is found that an original syllable's unvoiced part can be correctly classified by simply defining a time threshold and then checking whether the time position of the first pitch peak is greater than this threshold. This idea can be understood when inspecting the typical waveforms for short-unvoiced, as shown in Fig. 1, and for long-unvoiced, as shown in Fig. 2. Here, the threshold is set to be 300 points under a sampling rate of 11,025 Hz.

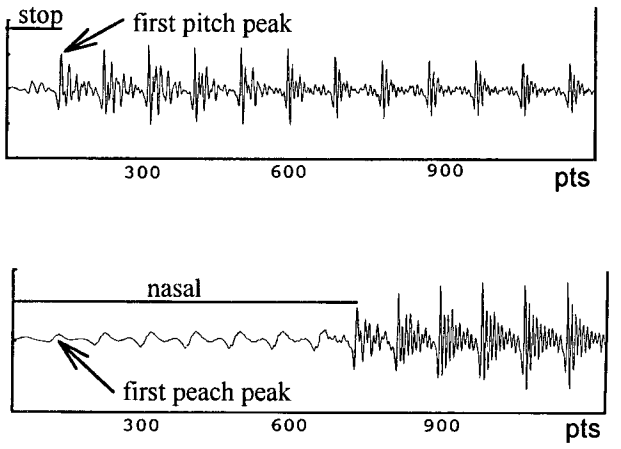

Fig. 1. Signal waveforms with short-unvoiced part.

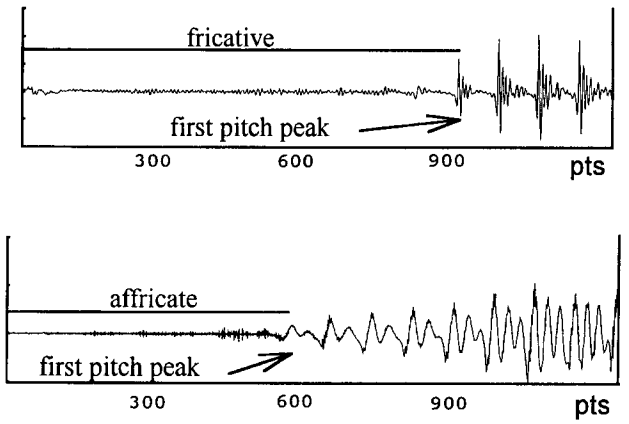

Fig. 2. Signal waveforms with long-unvoiced part.

After the unvoiced part of an original syllable is classified, the durations of the two parts of the corresponding syllable to be synthesized can then be determined accordingly. If the unvoiced part of the original syllable is short-unvoiced, the signal portion preceding the first pitch peak of the original syllable will be directly copied to the synthesized syllable to form its unvoiced part. Then, the duration of the voiced part to be synthesized is simply the value of the syllable-duration parameter minus the duration of the unvoiced part. On the other hand, if the unvoiced part of the corresponding original syllable is long-unvoiced, the durations of the two parts will be determined in two steps. First, the value of the syllable-duration parameter, Ts, is divided into two values, $T1$ and $T2$, according to the proportion of the durations of the two parts, Tu and Tv, in the original syllable, i.e., $Ts=T1+T2$ and $T1/T2=Tu/Tv$. Then, the assigned duration, $T1$, of the unvoiced part is checked to see whether it is greater than the duration limit, $1.5 \cdot Tu$, where the number 1.5 is empirically set. If $T1$ is greater than $1.5 \cdot Tu$, $T1$ is reassigned a value $1.5 \cdot Tu$, and $T2$ must then be set to the value $Ts-T1$.

After the duration of the unvoiced part (long-unvoiced) is determined, the signal waveform of this part is then synthesized. Our synthesis method for the long-unvoiced part has two steps. First, the leading 300 signal samples of the original syllable are directly copied to the leading portion of the synthesized syllable. This step is used 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. To show this in detail, let x in Fig. 3(b) be a sample position, Tx be the number of samples in the synthesized unvoiced part, and Ty be the number of samples in the original unvoiced part. Then, a position y, shown in Fig. 3(a), in the original unvoiced part is mapped from x by means of time-proportioning, i.e.,

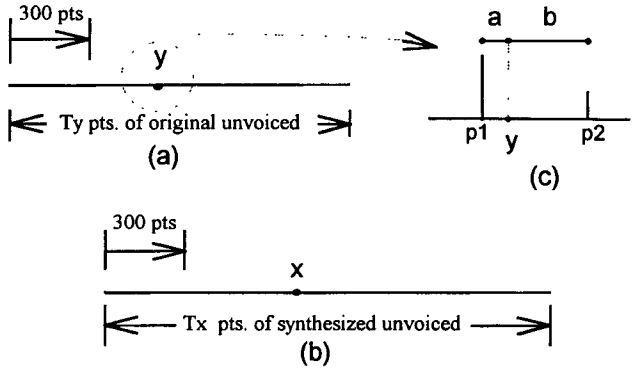

Fig. 3. Mapping and interpolation for synthesis of an unvoiced sample.

$$y = \frac{x-300}{Tx-300} \bullet (Ty-300)+300 . \quad (1)$$

In general, the value of y may not be an integer. Suppose it lies between the two integers $p1$ and $p2$, and that $a=y-p1$ and $b=p2-y$ as shown in Fig. 3(c). Then, the sample value in position x is interpolated as

$$x\_sample = p1\_sample \bullet \frac{b}{a+b} + p2\_sample \bullet \frac{a}{a+b} . \quad (2)$$

Although simple linear formulas are used to synthesize unvoiced samples, the resulting signal waveform is clear and comprehensible when heard.

III. Synthesis of Voiced Part

In the previous section, the method for the determination of the duration of the voiced part was explained. The next step is to determine the lengths (in sample points) of all the pitch periods in the voiced part and to then synthesize the signal samples in each pitch period. Basically, these two tasks, pitch-period length determination and signal sample synthesis, can be carried out separately. Our procedures for these two tasks are given in Section III.1 and III.2, respectively.

1. Determination of Pitch-Period Lengths

To enable a synthesized syllable to have a specific pitch contour which satisfies a required tone, timbre, or intonation, the prosodic processing unit will send out pitch-contour parameters to control the speech signal synthesis unit. In our method, eight pitch-contour parameters are used to approximate a syllable's pitch contour. Actually, the pitch contour is approximated using seven connected line segments, and each segment occupies one seventh of the duration of the voiced part. Apparently, the desired pitch contour can be more precisely approximated when more line segments are used.

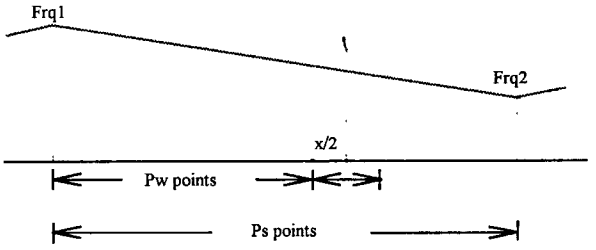

Fig. 4. Illustration of the determination of the pitch period length.

To explain the proposed procedure for determining a pitch-period's length, suppose Pw points are passed in the current line segment that has in total Ps points, that the next pitch period has x points, and that $Frq1$ and $Frq2$ are the frequency values of the two end points as shown in Fig. 4. Then, x can be solved by means of linear interpolation as

$$x = \left(\frac{Pw+\frac{x}{2}}{Ps}\right)\left(\frac{11{,}025}{Frq2} - \frac{11{,}025}{Frq1}\right) + \frac{11{,}025}{Frq1} , \quad (3)$$

where 11,025 is the sampling rate, and 11,025 over a frequency value represents the length (in sample points) of a pitch period located around the position of interest. In this paper, a pitch-period's time location is defined at the time location of its center sample. Therefore, the time location of the next pitch period is computed as $Pw+x/2$ in Eq. (3). When the remaining points, $Ps-Pw$, are shorter than one half of the previous pitch-period's length, these points will be treated as if they are in the next segment.

2. Synthesis of Signal Samples within a Pitch Period

After the length of each pitch period is determined, the signal samples within each period can be computed. To compute these signal samples, a procedure based on time-proportioned interpolation of corresponding pitch waveforms is proposed. The name TIPW, representing our Mandarin-syllable signal synthesis method, is derived from this procedure. In the following, this procedure, which includes four steps, will be explained.

A. Step (1): Find Two Corresponding Pitch Periods

Suppose the pitch period to be synthesized is located at point c as shown in the upper part of Fig. 5, Ns is the number of sample points to be synthesized

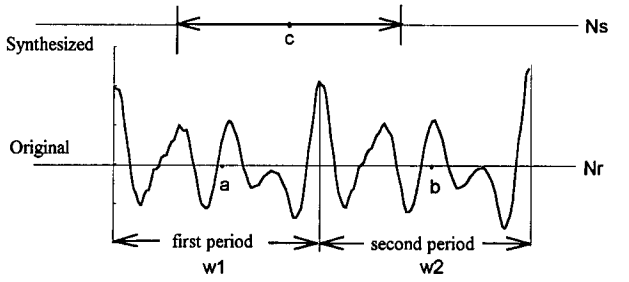

Fig. 5. A synthesized pitch period and its two corresponding pitch periods.

for the voiced part, and Nr is the number of sample points in the voiced part of the original First-tone syllable. Then, the first step is to find two consecutive pitch periods in the original syllable such that

$$\frac{a}{N_r} \le \frac{c}{N_s} < \frac{b}{N_r}, \tag{4}$$

where a and b, as shown in Fig. 5, are the center sample points of two consecutive pitch periods in the original syllable.

B. Step (2): Weight the Two Pitch Periods Found

After two consecutive pitch periods satisfying formula (4) are found, they are used to synthesize the target pitch period by means of interpolation. Two weights, $w1$ and $w2$, as shown in Fig. 5, for the first and second periods, respectively, have to be set. Here, these weights are determined by applying the linear time-proportion relation. That is, let

$$\alpha = \frac{a}{N_r},\ \beta = \frac{b}{N_r},\ \gamma = \frac{c}{N_s}$$

and set

$$w1 = \frac{\beta - \gamma}{\beta - \alpha},\ w2 = \frac{\gamma - \alpha}{\beta - \alpha}. \tag{5}$$

In terms of the weights computed, the signal waveform in the first pitch period found in Step (1) is then weighted using $w1$; i.e., each sample value is multiplied by $w1$, and the second pitch period is weighted using $w2$.

C. Step (3): Windowing

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 before they are copied to the synthesized pitch period. Also, the length of the applied window function must be carefully determined. Before describing our method, let us recall the phenomenon that sometimes the speech signal synthesized using the PSOLA technique is perceived as having two co-existent tones (one high and one low). This phenomenon can be attributed to two facts. First, the synchronization achieved by PSOLA is only the pitch-location's synchronization, not the pitch-period length's synchronization. In PSOLA, a pitch period is defined around a pitch peak (i.e., each pitch peak is the center point of a pitch period), and the time distances to the previous and next pitch peaks are not equal in general, which implies that the length of a window function cannot be simultaneously set to be the integer times of the left and right distances to the adjacent pitch peaks. Secondly, in PSOLA, the length of the window function is set without considering the length of the pitch period to be synthesized.

To synchronize both the location and length of a pitch period, we therefore define a pitch-period's extent as the sampling points bounded by two adjacent pitch peaks, and let the length of the applied window function be dependent on the three lengths of the synthesized and two original pitch-periods. As in PSOLA, the window function adopted is a cosine window, and its formula is

$$W(n) = 0.5 + 0.5 * \cos\left(\frac{2\pi n}{N}\right),\ 0 \le n \le N-1. \tag{6}$$

However, the window function's length, N in Eq. (6), is set to be two times the length of the synthesized pitch period as shown in Fig. 6(a) if the length of the original pitch period is greater than the length of the synthesized pitch period. Otherwise, it is set to be two times the length of the original pitch period as shown in Fig. 6(b). After the length of the window function is determined, two half windows, $Wa(n)$ and $Wb(n)$, are placed on the waveform of the original pitch period as shown in Fig. 6(a) or (b). Then, the waveform multiplied by $Wa(n)$ is copied to the left side and aligned with the left end of the synthesized pitch period while the waveform multiplied by $Wb(n)$ is copied to the right side and aligned with the right end. This manner of placement and alignment of the windowed waveforms is illustrated in Fig. 7(a). Here, assume that the length (e.g., 100 points) of the synthesized pitch period is greater than the length (e.g., 83 points) of the original. Then, when the two waveforms in Fig. 7(a) are overlapped and added, the waveform shown in Fig. 7(b) is obtained.

D. Step (4): Adding Two Processed Original Waveforms

There are two weighted original pitch periods in Steps (1) and (2). In Step (3), for convenience of explanation, only one original pitch period is windowed. In fact, the waveforms of the two weighted

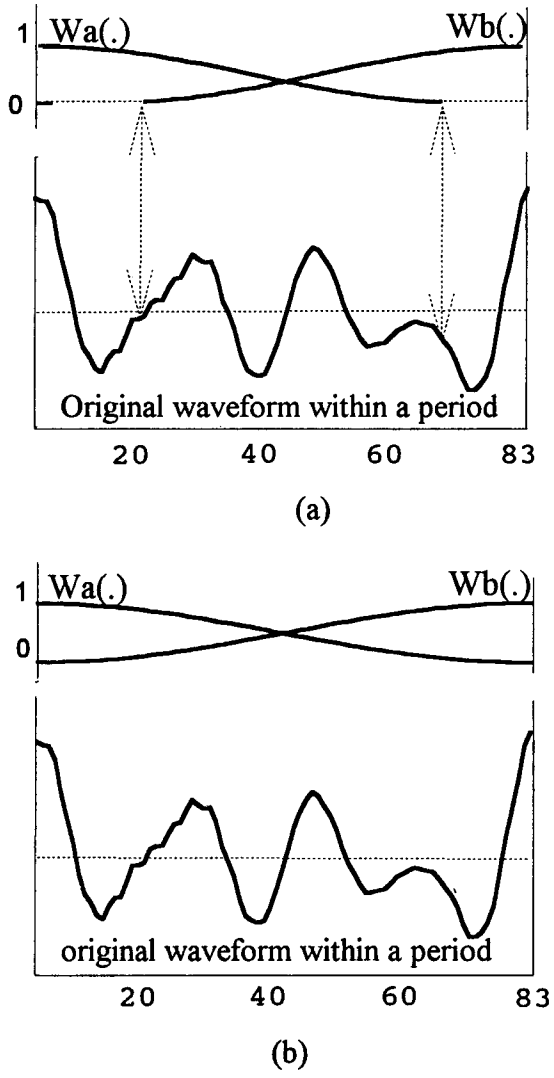

Fig. 6. Window placement and length determination. (a) The synthesized pitch period is shorter than the original. (b) The synthesized pitch period is longer than the original.

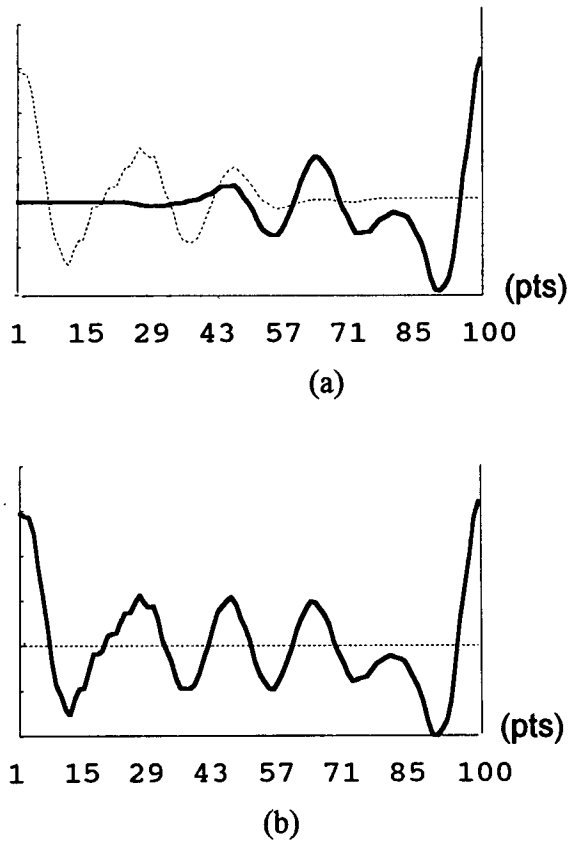

Fig. 7. Windowed waveform. (a) Placed and aligned. (b) Overlapped and added.

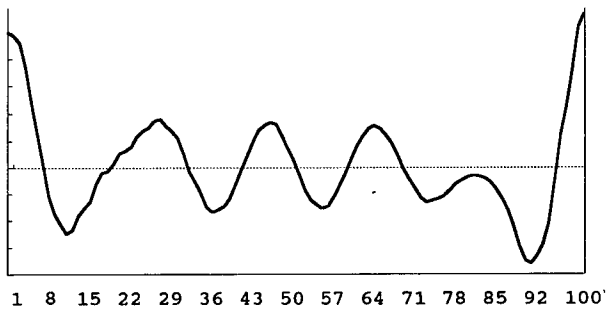

Fig. 8. An example waveform of a synthesized pitch period.

original pitch periods must be windowed in Step (3) separately. Then, the two windowed waveforms are overlapped and added. In Fig. 8, the synthesized waveform of a pitch-period is shown. It can be seen that there are more wave peaks in Fig. 8 than there are in either of the two original pitch periods in Fig. 5. This is because the height of the formant frequencies is kept the same while the length of the synthesized pitch period is extended and becomes greater than the length of the original pitch period.

By applying the four processing steps to the consecutive pitch periods of a pitch contour, the waveform of the voiced part of a syllable can be synthesized. Then, the synthesized voiced part can be concatenated with the synthesized unvoiced part to form the waveform of a synthesized syllable.

3. Vocal Track Length Control

When a female's speech signal is synthesized by only raising the pitch contours of the syllable signals collected from a male, the synthesized speech signal is heard as a man mimicking a woman's speech. Therefore, in addition to the factor of the pitch-contour's height, there are other factors that help us distinguish a female's speech from a male's speech. One of the

important factors is the vocal track length because, on average, a male's vocal track length is longer than a female's (O'Shaughnessy, 1987; Rabiner and Juang, 1993). Control of the vocal track length was, therefore, studied.

From the acoustic model of the vocal track (O'Shaughnessy, 1987; Rabiner and Juang, 1993), it is seen that there is an inverse proportion between the vocal track length and the formant frequency. That is, the vocal track can be imagined as being shortened or lengthened when formant frequencies are raised or lowered, respectively. Let us recall an application of this relationship. To dub a cartoon voice, an adult's voice is recorded beforehand and then played back at a faster speed. Faster playing speeds can raise formant frequencies. When formant frequencies are raised, it sounds like the length of the vocal track is decreased. Therefore, an adult's voice can be made to sound like a child's voice using a faster playback speed. However, there are severe side effects with faster playback speeds; i.e., the pitch contour will be raised, and the duration will be decreased. Therefore, we are motivated to develop a syllable-signal synthesis method that can support independent control of the factors duration, pitch contour, and formant frequency height (or vocal track length).

In Sections III.1 and III.2 , a method supporting independent control (in a reasonable value range) of the factors syllable duration and pitch contour was presented. Here, a method to support independent control of the vocal track length will be described. It can be directly inserted between the two processing steps, Step (2) and Step (3), described in Section III.2, and will not affect the original processing steps. Basically, the same principle as that applied in dubbing a cartoon voice is used. Actually, our approach is to resample the two pitch periods found in Step (1) of Section III.2. For example, suppose the formant frequencies are raised to 1.3 times their original values (i.e., the vocal track length is to be shortened to 1/1.3 of its original value) and the sampling frequency is kept the same; then, the nth sample point in the resampled waveform will be mapped to the point with index $m=1.3 \bullet n$ in the original waveform. Apparently, the value of m may not be an integer. From the theoretical viewpoint, there is no problem because the original analog waveform can be reconstructed from the discrete samples. However, for practical purposes (computation time required), an approximation of quadratic interpolation is adopted here. That is, the three samples $y0$ at $\lfloor m-1 \rfloor$, $y1$ at $\lfloor m \rfloor$, and $y2$ at $\lfloor m+1 \rfloor$ are used to construct a quadratic polynomial, $y=f(x)=A \bullet x^2+B \bullet x+C$. The values of coefficients A, B and C can be computed as

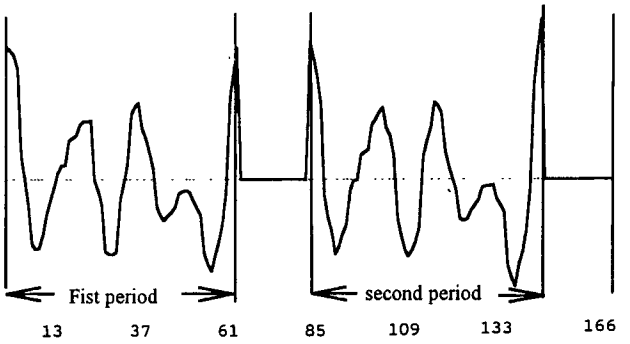

Fig. 9. Waveform resampled at 1.3 points for two original pitch periods.

$$\begin{cases} A = 0.5 \bullet y_2 - y_1 + 0.5 \bullet y_0 \\ B = -0.5 \bullet y_2 + 2 \bullet y_1 - 1.5 \bullet y_0 , \\ C = y_0 \end{cases} \tag{7}$$

where the numbers 0.5, -1.5, ..., etc. are obtained by replacing the sample indices $\lfloor m-1 \rfloor$, $\lfloor m \rfloor$ and $\lfloor m+1 \rfloor$ with 0, 1, and 2. Then, the sample value at point m is interpolated as $f(m-\lfloor m \rfloor+1)$. After resampling in the way described, the resulting waveforms for the two original pitch periods in Fig. 5 will be as shown in Fig. 9. From this figure, it is seen that the pitch periods' lengths are shortened to 1/1.3 of their original lengths. Therefore, in Step (3) of Section III.2, the length of original pitch period must be modified accordingly to reflect the effect of the resampling process.

When the values of the factors pitch-contour height and vocal track length are independently assigned, many distinct timbres can be synthesized. Suppose the original syllable signals are collected from a male speaker. Then, a female's voice can be synthesized by raising the pitch contour to a height that is two times its original value and by shortening the vocal track length to 0.87 times its original length (i.e., walk about 1.15 point each time in resampling). Also, the voice of a cartoon actor or child can be synthesized by shortening the vocal track length to about 0.8 times its original length while the height of pitch contour can be raised or kept. Besides controlling these two factors independently, it is usually convenient to define a dependency relation between them. That is, when the pitch contour is raised or lowered, the vocal track length is automatically modified. With this dependency relation, it is desired that the synthesized timbre will be smoothly changed from that of a male to that of a female when the pitch contour is slowly raised and will not exhibit the mimicking phenomenon (where a male mimics a female's voice). In fact, we have

studied this problem and found a useful dependence relation:

$$WalkPoints = \frac{1 + 0.15 \cdot (\frac{NewF0}{120} - 1)}{1 + 0.15 \cdot (\frac{OldF0}{120} - 1)}, \quad (8)$$

where *WalkPoints* is the factor to be multiplied for index mapping (such as the number 1.3 mentioned near Eq. (7)) in the resampling process, *OldF0* is the average fundamental frequency of the original syllable signals collected, *NewF0* is the desired average fundamental frequency of the synthesized speech, 1.15 is the ratio of the average F1 formant frequency of the phoneme /∂/ for females to that for males, and 120 is the average fundamental frequency for males. This equation is derived from the idea of linearly interpolating the normalized average-F1-frequencies for males and females (i.e., 1 and 1.15) to obtain the normalized F1-frequencies for the fundamental frequencies *NewF0* and *OldF0*, where the average-fundamental-frequencies for males and females (i.e., 120 Hz and 240 Hz) are normalized (to 1 and 2) and used as the reference positions.

IV. Verification of Our Synthesis Method

We assert that the proposed synthesis method can support independent control of the factors syllable duration, pitch contour, and vocal track length in a reasonable value range. To show that our assertion is reasonable, it is necessary to use the proposed synthesis method to implement a practical signal synthesis unit. Note that a signal synthesis unit can only synthesize a syllable's signal each time it is invoked and cannot synthesize an entire sentence's signal without a prosodic processing unit's guidance. To evaluate whether the synthesized speech is clear and whether distinct timbres can be synthesized, it is necessary to hear an entire sentence's signal rather than hear a sequence of independent syllable-signals. Therefore, we built a prototype text-to-speech system to show that our Mandarin-syllable signal synthesis method is practically applicable and versatile. To build the text-to-speech system, we had a male pronounce 409 First-tone syllables in isolation and recorded the signal waveforms at a sampling rate of 11,025 Hz and a resolution of 16 bits/sample. The memory space required to store the signal waveforms was about 2.24 Mbytes. Then, the time locations of the pitch peaks within each syllable's voiced part were detected and saved as the signal synthesis unit's parameters. For the prosodic processing unit, a rule-based approach (Lee *et al.*, 1993; Chiou *et al.*, 1991) was adopted after a few changes were made.

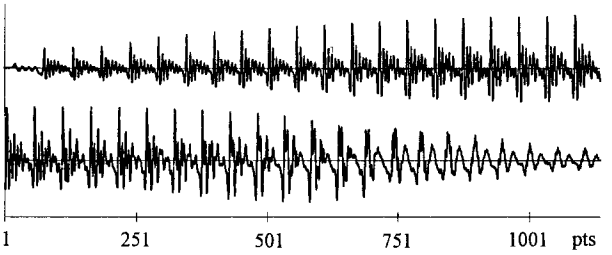

Fig. 10. The waveform of the original First-tone syllable /ai/.

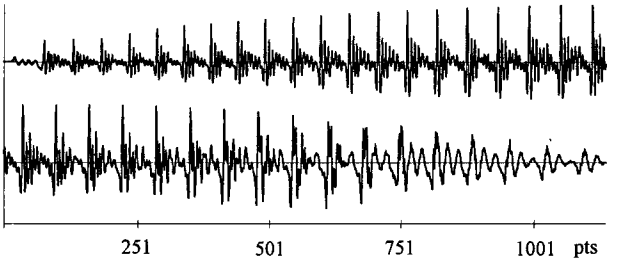

Fig. 11. The waveform of the synthesized Forth-tone syllable /ai/.

The prototype text-to-speech system has been implemented in real-time execution. It is text driven; i.e., the control commands for the pronunciation speed, fundamental frequency or pitch-contour height, and vocal track length are represented as recognizable text and directly placed within the ordinary text to be synthesized. Therefore, a dialog involving many persons (i.e., many timbres) can be synthesized in real-time. Concerning the quality of the synthesized speech, initial evaluation shows that our method's output signal is as clear as other time-domain synthesis methods. As far as comprehensibility and fluency are concerned, we think they cannot be used to evaluate the signal synthesis unit separately because they are mainly affected by the prosodic processing unit and our prosodic processing unit is not very good. With regard to the characteristics of independent control of the factors duration, pitch contour, and vocal track length, we found that a syllable's duration could be changed from 0.5 to 4 times its original duration (0.25 sec. on average) without notable side effects when the other two factors are kept unchanged. Similarly, a syllable's fundamental frequency could be changed from 0.5 to 2.5 times its original value and a syllable's intrinsic vocal track length could be changed from 0.7 to 1.6 times its original length. In the following two subsections, the results of spectrogram analysis of changes in pitch contour and vocal track length will be discussed.

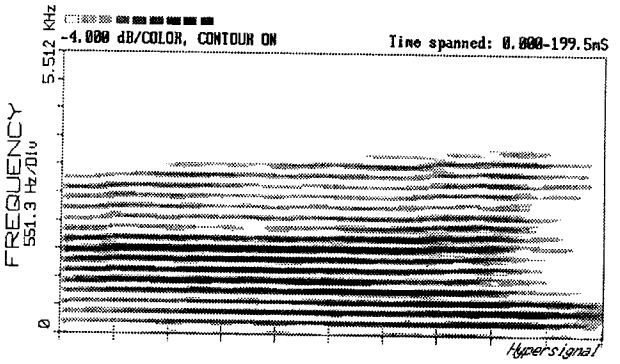

Fig. 12. Spectrogram of the original First-tone syllable /ai/.

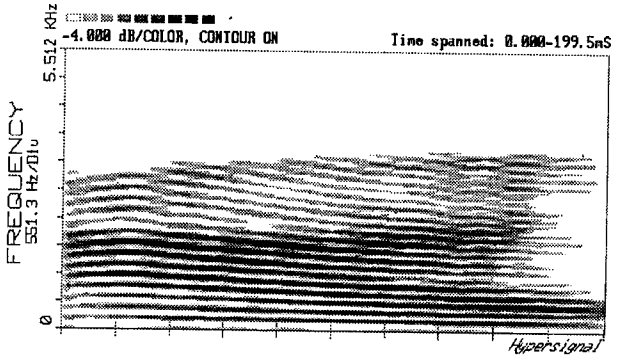

Fig. 13. Spectrogram of the synthesized Forth-tone syllable /ai/

1. Change of Tone

Because the original syllables were all of First-tone, it may be suspected that the formant frequency trace would be affected when a syllable of another tone was synthesized. Here, the waveform of the original syllable /ai/ was used to synthesized the waveform of a Forth-tone syllable /ai/ while the duration and vocal track length were not changed. The original and synthesized waveforms are as shown in Figs. 10 and 11. After analysis was conducted using signal analysis software (Hyperception, 1991), the waveforms' spectrograms were drawn as shown in Figs. 12 and 13. From Fig. 11, it can be seen that the pitch periods became longer with time. Therefore, in the spectrogram in Fig. 13, the fundamental frequency and its harmonics become lowered with time. However, the formant frequencies in Fig. 13 progress in the same manner as in Fig. 12; i.e., they are not affected by the falling pitch contour.

2. Change of Vocal Track Length

A syllable's intrinsic vocal track length can be changed by resampling its waveform. Here, we exam-

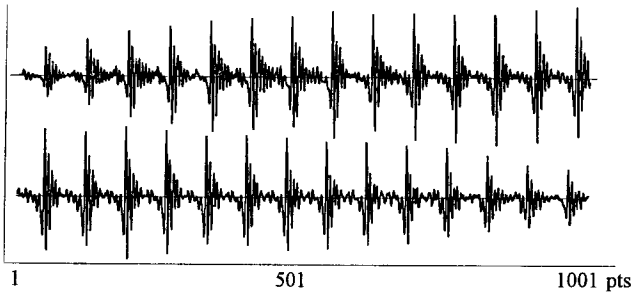

Fig. 14. Synthesized waveform of /a/ by walking 1.3 points each time.

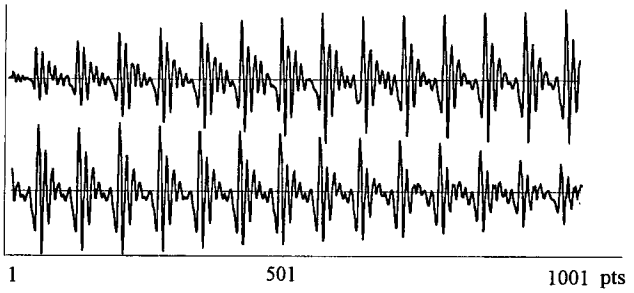

Fig. 15. Synthesized waveform of /a/ by walking 0.7 points each time.

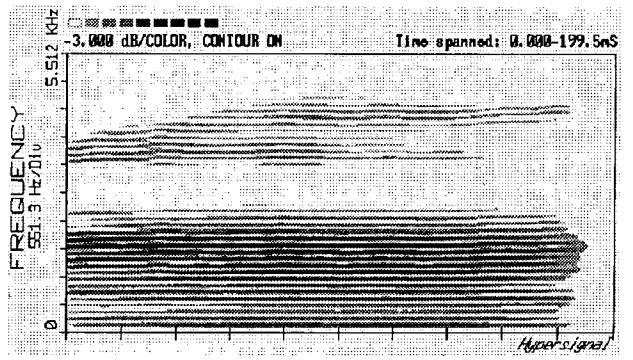

Fig. 16. Spectrogram of the waveform with resampling conducted at 1.3 points.

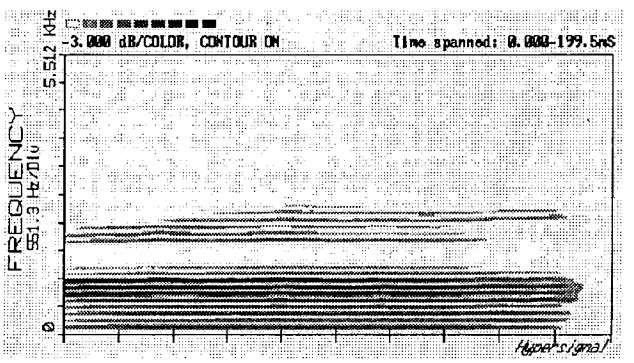

Fig. 17. Spectrogram of the waveform with resampling conducted at 0.7 points.

ined this relation using spectrographic analysis. As an example, the original waveform of the First-tone syllable /a/ was used here to synthesize those waveforms in which the vocal track length changed but in which the other two factors, duration and pitch contour, were kept the same. The waveform synthesized by walking 1.3 points each time in resampling is shown in Fig. 14. Similarly, the waveform synthesized by walking 0.7 points each time is shown in Fig. 15. After the two waveforms in Figs. 14 and 15 were analyzed, the spectrograms obtained were as shown in Figs. 16 and 17, respectively. It can be seen that the number of wave peaks in a pitch period of the waveform in Fig. 14 is much greater than that in a pitch period of the waveform in Fig. 15 while the lengths of the pitch periods in the two waveforms are nearly the same. This indicates that the formant frequencies of the waveform in Fig. 14 were much higher than the corresponding formant frequencies in Fig. 15. From the spectrograms in Figs. 16 and 17, it can be seen that this is indeed the case. Therefore, a syllable's formant frequency height, i.e., its intrinsic vocal track length, can be adjusted by resampling each pitch-period's signal waveform.

V. Conclusion

In this paper, a new Mandarin-syllable signal synthesis method has been proposed. It not only can synthesize clear speech signals as well as other time-domain synthesis methods can, but also has greater flexibility in the control of syllable duration, pitch contour, and vocal track length. In particular, the interference that warps a syllable's formant frequency trace is largely decreased when an original syllable's duration or pitch contour is modified. This advantage is provided by the windowing and time-proportioned interpolation methods proposed here. In addition, control of the vocal track length has been newly studied here. When this factor and the pitch-contour height are simultaneously controlled, many distinct timbres can be synthesized, and the mimicking phenomenon can be prevented. Note that the ability to synthesize distinct timbres is necessary to extend the application domain of speech synthesis to, for example, newscasts with multi-newsreaders and dialog synthesis of a novel. To verify the proposed synthesis method, we have implemented a prototype text-to-speech system that can be operated in real-time and controlled in a text-driven mode. The initial evaluation shows that the synthesized speech is clear enough and that the duration, pitch contour, and vocal track length can be independently controlled in a reasonable value range.

Acknowledgment

This work was supported by the National Science Council of the Republic of China under contract number NSC 85-2213-E011-046.

Reference

Charpentier, F. and M. Stella (1986) Diphone synthesis using an overlap-add technique for speech waveform concatenation. IEEE Int. Conf. ASSP, Tokyo, Japan.

Chen, S. H., S. H. Hwang, and C. Y. Tsai (1992) A first study on neural net based generation of prosodic and spectral information for Mandarin text-to-speech. IEEE Int. Conf. ASSP, San Francisco, CA, U.S.A.

Chiou, H. B., H. C. Wang, and Y. C. Chang (1991) Synthesis of Mandarin speech based on hybrid concatenation. *Computer Processing of Chinese and Oriental Languages*, **5**, 217-231.

Dutoit, T. and H. Leich (1993) MBR-PSOLA: text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, **13**, 435-440.

Galanes, F. M., M. H. Savoji, and J. M. Pardo (1994) New algorithm for spectral smoothing and envelop modification for LP-PSOLA synthesis. IEEE Int. Conf. ASSP, Adelaide, Australia.

Hamon, C., E. Moulines, and F. Charpentier (1989) A diphone synthesis system based on time-domain prosodic modification of speech. IEEE Int. Conf. ASSP, Glasgow, Scotland, U.K.

Holmes, J. (1983) Formant synthesizers - cascade or parallel? *Speech Communication*, **2**, 251-273.

Hyperception (1991) *Hypersignal Users Manual*. Hyperception, Dallas, TX, U.S.A.

Kawai, H., N. Higuchi, T. Simizu, and S. Yamamoto (1994) Development of a text-to-speech system for Japanese. IEEE Int. Conf. ASSP, Adelaide, Australia.

Klatt, D. H. (1980) Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, **67**, 971-995.

Lee, L. S., C. Y. Tseng, and C. J. Hsieh.(1993) Improved tone concatenation rules in a formant-based Chinese text-to-speech system. *IEEE Trans. Speech and Audio Processing*, **1**, 287-294.

Makhoul, J. (1975) Linear prediction: a tutorial review. *Proc. IEEE*, **63**, 561-580.

Markel, J. D. and A. H. Gray, Jr. (1976) *Linear Prediction of Speech*. Springer-Verlag, New York, NY, U.S.A.

Modulines, E. and F. Charpentier (1990) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, **9**, 453-467.

O'Shaughnessy, D. (1987) *Speech Communication: Human and Machine*. Addison-Wesley, New York, NY, U.S.A.

Rabiner, L. and B. H. Juang (1993) *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, U.S.A.

Wu, C. H., C. H. Chen, and S. C. Juang (1995) An CELP-based prosodic information modification and generation of Mandarin text-to-speech. ROCLING VIII, Tauyuan, Taiwan, R.O.C.

Syllable Synthesis with Timbre Control

增進音長、音調及音色控制之彈性的國語音節信號合成方法

古鴻炎　許文龍

國立臺灣科技大學電機工程技術系

摘　要

在許多文句翻國語語音的系統裡，都採用音節爲語音合成之單位，因此本文針對國語音節信號合成的問題提出了一個新的合成方法，與其它時域合成方法比較，除了能夠合成出清晰的語音信號之外，還提供了較多的信號控制之彈性，包括音長之控制，以調整說話速度及反映其它因素對音節長度之影響；音調(或基週軌跡)之控制，以便由第一聲音節去合成其它聲調的音節，而能夠節省記憶需求；以及聲道長度之控制，以便使男生原音合成出的女生聲音(或卡通人物的聲音)較爲自然。我們發現當適當地調整聲道長度與音調高低，就可合成出許多互不相同的音色，聲道長度是一項新提出的控制因素，它使得音色的控制成爲實際可行。雖然其它時域合成方法也有提供音調、音長之控制，但是我們的合成方法提供的彈性較高，且已讓這兩個控制因素產生的對共振峰軌跡的干擾降低很多。