VARIABLE TIME-SCALE MODIFICATION OF SPEECH USING TRANSIENT INFORMATION

Sungjoo Lee*, Hee Dong Kim**, Hyung Soon Kim*

*Dept. of Electronics Eng., Pusan National University
**Dept. of Telematics Eng., The University of Suwon
E-mail: {lsj, kimhs}@pusan.ac.kr

ABSTRACT

Conventional time-scale modification methods have the problem that as the modification rate gets higher the time-scale modified speech signal becomes less intelligible, because they ignore the effect of articulation rate on speech characteristics. In this paper, we propose a variable time-scale modification method based on the knowledge that the timing information of transient portions of a speech signal plays an important role in speech perception. After identifying transient and steady portions of a speech signal, the proposed method gets the target rate by modifying steady portions only. The result of subjective preference test indicates that the proposed method produces performance superior to that of the conventional SOLA method.

I. INTRODUCTION

The purpose of time-scale modification is to change the rate of speech while preserving the characteristics of original speech such as formant structure, pitch periods, etc. There are various applications of time-scale modification. For example, one can reduce the bit rate required for medium-rate speech coding by time-scale compression of the input speech, followed by coding and the transmission, followed by time-scale expansion to the original time scale at the receiver. In digital telephone answering devices(TAD), time-scale modification enables to have quicker playback of received voice messages. In special systems for older people and in foreign language education, slower speech is more helpful for understanding. While there are a number of techniques for the time-scale modification of speech[1]-[6], the synchronized overlap and add (SOLA) method is used widely because of its computational simplicity, allowing real-time implementation [4][5]. Although the SOLA method produces a reasonable quality, the rate-changed speech has a lower degree of intelligibility as the amount of rate change increases. In particular, this problem limits very fast playback of speech in such application areas as digital TAD.

Results of research on speech perception show that the timing information of transient portions of a speech signal plays an important role in discriminating among different speech sounds[7]-[9]. Inspired by this fact, we propose a novel scheme for the time-scale modification of speech, in which the timing information of the transient portions of speech is preserved, while the steady portions of speech are compressed or expanded somewhat excessively for maintaining overall time-scale change. To do so, transient and steady portions must be separated in the speech signal. We devise two different methods to identify transient and steady portions; one is the method using LPC cepstral distance and the other using cross-correlation. To evaluate the performance of the proposed scheme, a subjective preference test by human listeners is conducted. The result indicates that the proposed method is superior to the conventional SOLA method. This paper is organized as follows. After a brief review of the conventional SOLA method in section II, we develop an algorithm for variable time-scale modification using transient information in section III. In section IV, we describe two methods for locating transient and steady portions in a speech signal. In section V, we compare the performance with conventional SOLA method.

II. SYNCHRONIZED OVERLAP AND ADD (SOLA) METHOD[4][5]

The key idea of SOLA method is to shift and average overlapping frames of a signal in a synchronized fashion at points of highest cross-correlation. As a result, the time-scale modified signal by SOLA method preserves the time-dependent pitch, the spectral magnitude and phase to a large degree to produce relatively high quality speech. Let $x(n)$ be the input signal and $y(n)$ the time-scale modified signal. Given the frame length of N, we introduce S_a as the analysis interframe interval and S_s as the synthesis interframe interval. Then the ratio of S_s/S_a is the modification factor α. $\alpha > 1$ corresponds to time expansion and $\alpha < 1$ corresponds to compression.

The SOLA method begins with copying the first frame of size N from $x(n)$ to $y(n)$. Then the m-th frame of the input signal, $x(mS_a+j)$, $0\leq j\leq N-1$, is synchronized and averaged with a neighborhood of $y(mS_s+j)$, on a frame-by-frame basis. The synchronization point, k_m, is determined by maximizing the normalized cross-correlation between $x(mS_a+j)$ and $y(mS_s+j)$ as follows:

$$R_m(k) = \frac{\sum_{j=0}^{L-1} y(mS_s+k+j)x(mS_a+j)}{[\sum_{j=0}^{L-1} y^2(mS_s+k+j)\sum_{j=0}^{L-1} x^2(mS_a+j)]^{1/2}},$$

$$-\frac{N}{2} \leq k \leq \frac{N}{2} \qquad (1)$$

where L is the length of overlap between $x(mS_a+j)$ and $y(mS_s+j)$. Once k_m is found, the time-scale modified signal $y(n)$ is updated as follows:

$$y(mS_s+k_m+j)=(1-f(j))y(mS_s+k_m+j)+f(j)x(mS_a+j), \quad 0 \leq j \leq L_m-1$$
$$y(mS_s+k_m+j)=x(mS_a+j), \quad L_m \leq j \leq N-1 \qquad (2)$$

where L_m is the range of overlap of the two signals for the particular k_m involved and $f(j)$ is a weighting function such that $0 \leq f(j) \leq 1$. In this paper we used a linear weighting function of $f(j) = j / (L_m-1)$, $0 \leq j \leq L_m-1$. The SOLA method produces a fine quality speech in spite of its relatively small amount of computation, however, as the amount of time-scale change increases, a time-scale modified signal becomes less intelligible. This problem may be due to the fact that the SOLA method, like most of the conventional time-scale modification methods, uses only a constant time-scale modification factor, α, for all frames of speech, not considering the effect of articulation rate on speech characteristics.

III. VARIABLE TIME-SCALE MODIFICATION OF SPEECH USING TRANSIENT INFORMATION

Speech sounds are characterized by time-varying spectral patterns which contain both transient and steady portions. Transient portions as well as steady portions of the speech signal are considered to play an important role in speech perception. Results of research on speech perception for several decades show that most consonants, including the plosives and nasals, share the common characteristic of containing their main perceptual distinctive feature in their transient portion, i.e., during their coarticulation with adjacent phonemes[7]-[9]. In the identification tests of syllables modified by initial and/or final truncation, Furui found that perceptual critical points, where the correct identification score of the truncated syllable as a function of the truncation position changes abruptly, are related to maximum spectral transition positions[7]. A speech signal of approximately 10 ms in duration that includes the maximum spectral transition position bears the most important information for consonant and syllable perception. It has also been reported that the rapidity of the spectral change is an important feature for discriminating among different classes of speech sounds[8]. In other words, the property that distinguishes some consonants from vowels and glides is not the shape of the spectrum at any particular instant of time but is rather the rapidity of the spectral change.

Therefore, the transient portions, where the spectrum changes rapidly, contain much information for speech perception and it is helpful for comprehension to keep the timing information of transient portions. In this paper, we propose a variable time-scale modification method based on this finding. In the proposed method, the time-scale modification factor depends on whether the speech segment belongs to the transient portion or not. How to separate transient and steady portions from a speech signal will be discussed in next section. After identifying transient and steady portions in a speech signal, we only modify the time scale of steady portions while keeping the transient portions unchanged. As a result, the steady portions of speech are compressed or expanded somewhat excessively to maintain the required overall speech rate. If T_s and T_t are the number of frames of steady portions and transient portions respectively, then the number of total frames of a speech signal, T, is represented as

$$T = T_t + T_s. \qquad (3)$$

And, in the proposed method, the time-scale modification factor for the steady portions, α_s, and the overall time-scale modification factor, α, has the following relationship.

$$\alpha T = T_t + \alpha_s T_s. \qquad (4)$$

From (3) and (4), α_s can be represented as

$$\alpha_s = ((\alpha-1)T+T_s) / T_s. \qquad (5)$$

With introducing new term, β, the ratio of steady portions in a speech signal, or $\beta = T_s / T$, (5) can be rewritten as

$$\alpha_s = ((\alpha-1)+\beta) / \beta. \qquad (6)$$

In order to apply the proposed method, we must separate transient and steady portions from a speech signal and then find out the ratio of transient (or steady) portions to total frames. This process, however, does not efficiently utilize either memory or real time processing. To alleviate this problem, we identify the transient and steady portions of a speech every prespecified time interval(1 second for an example) to perform a variable time-scale modification according to the percentage of each portion in the interval.

IV. SEPARATING TRANSIENT AND STEADY PORTIONS FOR VARIABLE TIME-SCALE MODIFICATION

In this section, we describe two methods to separate the transient and steady portions of speech signal for the proposed variable time-scale modification. The first method utilizes LPC cepstral distance between neighboring frames with somewhat additional complexity, and second method utilizes the cross-correlation function which can be obtained in the process of the conventional SOLA method.

4.1. The LPC cepstral distance method

A spectral distance between adjacent frames can be a good measure for discriminating the transient portions from the steady portions. In this paper, we approximate the spectral distance with the LPC cepstral distance using the finite LPC cepstra extracted from a speech signal[10]. Then we identified transient and steady portions by comparing the LPC cepstral distance between non-overlapping neighbor frames with a proper threshold. A window length of 30 ms and a window shifting length of 10 ms are used for computing LPC cepstra. The LPC cepstral distance of m-th frame, $D(m)$, is represented as follows:

$$D(m) = \sum_{k=1}^{p} [c_{m-2}(k) - c_{m+2}(k)]^2 \qquad (7)$$

where $c_m(k)$, $k=1,...,p$ is k-th LPC cepstral coefficient of the m-th frame. If $D(m)$ is greater than a predefined threshold, this frame would be taken for a transient portion, otherwise a steady portion. An example of separating transient and steady portions using LPC cepstral distance is shown in Figure 1(a)-(c). Given a speech signal shown in Figure 1(a), LPC cepstral distance between neighboring frames with threshold TH1 (=1.3) is shown in Figure 1(b). Figure 1(c) indicates the result of transient/steady portion detection; a "1" represents transient portions and a "0" steady portions. As can be seen from the figure, the performance of the method based on the LPC cepstral distance is fairly good. However, this method requires the computation of LPC cepstra at every frame, which yields increased computational complexity except for the case that it is used in coupled with LPC-based speech coders[5].

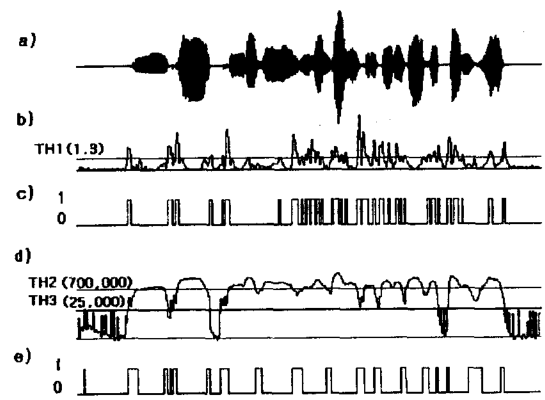

Figure 1. An example of separating transient and steady portions from a speech signal
(a) speech signal
(b) LPC cepstral distance and threshold
(c) transient(1) and steady(0) portions identified by the LPC cepstral distance method
(d) log-scale cross-correlation value and threshold
(e) transient(1) and steady(0) portions identified by the cross-correlation method

4.2. The cross-correlation method

As a computational efficient alternative to the afore-mentioned method, a method using cross-correlation which is employed in the conventional SOLA method is devised. The maximum cross-correlation value at the synchronization point calculated in the process of the conventional SOLA algorithm contains information on the similarity between adjacent frames; steady portions have large cross-correlation values and transient have small ones. Therefore, transient and steady portions of a speech signal can be identified with proper thresholding of the maximum cross-correlation for each frame. The maximum cross-correlation of m-th frame, $C(m)$ is defined as follows:

$$C(m) = \max_k [\frac{1}{L} \sum_{j=0}^{L-1} y(mS_s + k + j) x(mS_a + j)],$$
$$-\frac{N}{2} \le k \le \frac{N}{2} \qquad (8)$$

$C(m)$ is easily computed by using numerator terms of the normalized cross-correlation, $R_m(k)$, in the conventional SOLA algorithm(See equation (1)). However, simple thresholding on $C(m)$ can yield erroneous decision for frames with background silence only. In other words, during the silence intervals, $C(m)$ would be very small, so they could be mistaken for transient portions. To avoid this problem, another threshold is taken. Silent and transient portions can be separated by comparing $C(m)$ with this threshold because $C(m)$ in silent portions are much smaller than those in transient portions. The silent portions are taken for steady portions. Figure 1(d) shows a log-scale representation of the cross-correlation, $C(m)$. In this figure, threshold TH2 is used in order to separate transient and steady portions from speech portions, and another threshold TH3 is used to separate silent and speech portions from a speech signal. Figure 1(e) shows the result of transient/steady detection using the cross-correlation method, and the meanings of "1" and "0" are the same as those in Figure 1(c).

In this paper, a set of threshold values (TH1=1.3, TH2=700000, TH3=2500) were chosen empirically, so that the ratio of transient portions to total frames might be about 20%. Result of a detailed experimentation shows that the accuracy of separating transient and steady portions from a speech signal in the method using cross-correlation is less than the accuracy in the method using LPC cepstral distance, but both methods show good results. Relatively lower accuracy for the method using the cross-correlation is mainly due to the fact that cross-correlation computations in the conventional SOLA method are performed on overlapping neighbor frames for the purpose of synchronization, thereby yielding very smoothed contour. On the other hand, the method using the LPC cepstral distance takes the distance between two non-overlapping neighbor frames to make a fine distinction.

V. EXPERIMENTAL RESULTS

A series of preference test by human listeners was conducted to evaluate the proposed variable time-scale modification algorithm. Both the method using LPC cepstral distance and the method using the cross-correlation were applied to locate the transient portions and steady portions of a speech signal. Speech materials used consist of five phonetically rich Korean sentences, each spoken by a different male speaker in a quiet environment. The speech data were sampled at 8 kHz sampling rate. The window length, N was 30 ms with 240 samples. The analysis interframe interval, S_a was 10 ms with 80 samples. The LPC cepstral coefficients were computed after the preemphasis $1 - 0.95 z^{-1}$. The result of transient and steady portion classification method was smoothed by 5-point median filtering. To compare the two proposed methods with the SOLA method, a listener preference test was done with speech data at five different time-scale modification factors; 0.5, 0.7, 1.3, 1.5, and 1.8. The test was to determine which method provides more intelligible speech. To justify the validity of the experiment, the time-scale modified speech data were presented to listeners in random order. Before the test the listeners were presented speech data at normal speed as a guideline of intelligible speech. Listeners used headphones and took the test individually with the experimenter to minimize distractions. Listeners selected one of the paired samples made by different methods so each result of the evaluation represents how much listeners prefer a method to others at the same speech rate. Tables 1 and 2 summarize the results. The listeners consisted of 18 males and 2 females, with ages from 23 to 31.

Table 1. Result of Subjective preference test between the conventional SOLA method and the proposed method (based on LPC cepstral distance)

Time-scale Modification factor α	Preference		
	Method A	No difference	Method B
0.5	11 %	4 %	85 %
0.7	23 %	19 %	58 %
1.3	21 %	29 %	51 %
1.5	22 %	27 %	51 %
1.8	23 %	24 %	53 %
Average	20.0 %	20.6 %	59.6 %

Method A: Conventional SOLA method
Method B: Proposed method based on LPC cepstral distance

Table 2. Result of Subjective preference test between the conventional SOLA method and the proposed method (based on the cross-correlation)

Time-scale Modification factor α	Preference		
	Method A	No difference	Method C
0.5	16 %	10 %	74 %
0.7	33 %	25 %	42 %
1.3	26 %	28 %	46 %
1.5	30 %	31 %	39 %
1.8	32 %	23 %	45 %
Average	27.4 %	23.4 %	49.2 %

Method A: Conventional SOLA method
Method C: Proposed method based on the cross-correlation

As can be seen from Tables 1 and 2, both of the proposed methods based on the LPC cepstral distance and the cross-correlation show comparatively better performance than the conventional one at every speech rate. Especially at the speech rate of 0.5 the proposed methods show a noticeable improvement. Comparing Table 2 with Table 1 the cross-correlation-based method was not as good as one using the LPC cepstral distance. This might be explained by the failure of the cross-correlation function to detect the exact transient portions of the speech signal as discussed in section IV. The cross-correlation-based method, however, can be implemented with almost the same computational complexity as the conventional SOLA method, while the method using LPC cepstral distance generally requires additional computational complexity due to the LPC parameter extraction and spectral distance computation. It should be noted that, when the time-scale modification is employed in coupled with LPC-based speech coders[6], the extra computational loads for the method using LPC cepstral distance are also negligible.

VI. CONCLUSIONS

The variable time-scale modification proposed in this paper takes advantage of the knowledge that the transient portions of the speech plays a greater role in speech perception. The proposed method identified the transient portions and the steady portions of speech signal and used the SOLA method in a flexible way to modify the time scale of the steady portions of the speech while the transient portions of the speech remain the same. The listener preference test shows that the proposed method is superior to the conventional SOLA method at every speech rate examined. Especially for the case of very fast playback (which requires higher intelligibility than any other rate), the proposed method achieved a significant performance improvement over the conventional one. In the proposed variable time-scale modification, we employed two methods for locating the transient and steady portions of the speech signal, and it was observed that there are trade-offs between the two methods in terms of the amount of improved speech quality and the computational complexity.

REFERENCES

[1] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis." IEEE Trans. Acoustics., Speech, Signal Processing, vol.ASSP-29, no.3, pp.374-390, Jun. 1981.

[2] T. F. Quateri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," IEEE Trans., Signal Processing, vol.40, no.3, pp.497-510, Mar. 1992.

[3] T. F. Quateri and R. J. McAulay, "Speech transformation based on sinusoidal representation," IEEE Trans. Acoustics., Speech, Signal Processing, vol.ASSP-41, no.6, pp.1449-1464, Dec. 1986.

[4] S. Roucos and A. M. Wilgud, "High quality time-scale modification for speech," in Proc. ICASSP, pp.493-496, Apr. 1986.

[5] J. Makhoul and A. El-jaroudi, "Time-scale modification in medium to low rate speech coding," in Proc. ICASSP, pp. 1075-1708, 1986.

[6] E. Moullines and F. Charpentier, "Pitch synchronous waveform processing for text to speech synthesis using diphones," Speech Communication, vol.9 (5/6), pp.453-467, 1990.

[7] S. Furui, "On the role of spectral transition for speech perception," J. Acoust. Soc. Amer., vol.80, pp.1016-1025, Oct. 1979.

[8] K. N. Stevens, "Acoustic correlates of some phonetic categories," J. Acoust. Soc. Amer., vol.68, pp.836-842, Sep. 1989.

[9] H. S. Kim, "A study on the use of perceptual information for speech recognition," Ph. D. Dissertation, KAIST, Feb. 1989.

[10] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, pp.100-117, 1993.