ISCA Archive
http://www.isca-speech.org/archive

7th International Conference on Spoken
Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

COMPUTATIONALLY EFFICIENT TIME-SCALE MODIFICATION OF SPEECH USING 3 LEVEL CLIPPING

Sung-Joo Lee, Hyung Soon Kim***

*Spoken Language Processing Team, Electronics and Telecommunications Research Institute, Korea
**Dept. of Electronics Eng., Pusan National University, Korea
*lee1862@etri.re.kr, **kimhs@pusan.ac.kr

ABSTRACT

Among the conventional time-scale modification methods [1]-[6], the synchronized overlap and add (SOLA) method [4] is used widely because of its good performance with relatively low computational complexity. But the SOLA method still requires much computation in evaluating the normalized cross-correlation function for synchronization procedure [9]. In this paper, we employ 3 level center clipping method in order to reduce the computational complexity of SOLA method. The result of subjective preference test indicates that the proposed method can reduce computational complexity by over 80% comparing with the conventional SOLA method without considerable performance degradation. We also apply the variable time-scale modification method using transient information [7] to the proposed algorithm. By doing so, we can maintain the intelligibility of time-scale modified speech in the case of very fast playback.

1. 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 popular [4][5]. But the SOLA method still requires much computation, because it should evaluate the normalized cross-correlation function for synchronization procedure before the overlap and adding procedure [9].

In this paper, we modify SOLA method in order to reduce its computational complexity. Firstly we apply 3 level center clipping to the speech signal. Secondly we calculate the normalized cross-correlation at the zero-crossing points only. By doing so, the total computational complexity could be reduced by over 80% compared with the conventional SOLA method. We also implement the global and local search time-scale modification (GLS-TSM) method [8] focusing on the computational reduction. 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 close to the conventional SOLA method and superior to the GLS-TSM method while the computational complexity is reduced by over 80% compared with the SOLA method. In order to maintain the intelligibility of time-scale modified speech in the case of very fast playback, we apply the variable time-scale modification method [7] to the proposed scheme. Experimental results using Korean broadcast news data shows that fairly intelligible speech can be obtained in very fast playback case. This paper is organized as follows. After a brief review of the conventional SOLA method in section 2 and GLS-TSM method in section 3, the proposed method is described in section 4. The variable time-scale modification using transient information method is mentioned in section 5. In section 6, the performance evaluation is described.

2. 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, S_s / S_a, is the modification factor α. Here $\alpha > 1$ corresponds to time expansion and $\alpha < 1$ corresponds to compression. The page layout should match with the following rules. 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, $x(mS_a+j)$, $0 \le j \le N-1$, 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:

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

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

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:

$$y(mS_s + k_m + j) = (1 - f(j))\, y(mS_s + k_m + j) + f(j)\, x(mS_a + j)\ , \quad 0 \le j \le L_m - 1$$

$$y(mS_s + k_m + j) = x(mS_a + j), \quad L_m \le j \le N - 1 \tag{2}$$

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 $0 \le f(j) \le 1$. In this paper we used a linear weighting function of $f(j) = j / (L_m - 1)$, $0 \le j \le L_m - 1$. The SOLA method produces a relatively fine quality speech. However, it still requires much computation for the synchronization procedure and has the intelligibility problem due to ignoring the effect of articulation rate on speech characteristics [7].

3. GLOBAL AND LOCAL SEARCH TIME-SCALE MODIFICATION(GLS-TSM) [8]

The GLS-TSM method is originally based on the SOLA method. But the GLS-TSM method uses zero-crossing rate and characteristic vectors instead of the normalized cross-correlation function in order to evaluate global and local similarity between the synthesized speech frame and the analyzed speech frame in the synchronization procedure. Thus, the GLS-TSM method can reduce much computational complexity compared with the conventional SOLA method. That is, in order to evaluate global similarity, the GLS-TSM method uses zero-crossing rate. The k_{global} sample point is found when the difference of the zero-crossing rates between two adjacent frames is minimized. Local similarity is evaluated with the characteristic vectors on the basis of the k_{global} sample point and the local similarity evaluation is performed on the zero-crossing point. So, this characteristic vectors are calculated on the zero-crossing points. The characteristic vector, f used in the local similarity evaluation procedure is as follows:

$$\begin{array}{ll}
f_1 & x(i) - x(i+1) \\
f_2 & |x(i)| \\
f_3 & |x(i+1)| \\
f_4 & \{x(i) - x(i+2)\}/2 \\
f_5 & |x(i+2)| \\
f_6 & \{x(i) - x(i+3)\}/3 \\
f_7 & |x(i+3)| \\
f_8 & \{x(i-1) - x(i+1)\}/2 \\
f_9 & |x(i-1)| \\
f_{10} & \{x(i-2) - x(i+1)\}/3 \\
f_{11} & |x(i-2)|
\end{array} \tag{3}$$

The distance measure for the local similarity evaluation procedure is given by

$$d_{k,i} = \frac{1}{11} \sum_{j=1}^{11} \left| f_x(j) - f_{y,i}(j) \right| \tag{4}$$

where k is the first zero-crossing point of speech signal, $x(n)$. $f_x(j)$ is the j-th element of the characteristic vector on a zero-crossing point of speech signal, $x(n)$. $f_{x,i}(j)$ is the j-th element of the characteristic vector on the i-th zero-crossing point of speech signal, $y(n)$. The GLS-TSM method requires small amount of computation comparing with the conventional SOLA method. But the quality of the synthesized speech is reduced comparing with the conventional SOLA method. We apply the GLS-TSM method to 8 kHz sampling rate speech data in order to compare our proposed methods.

4. TIME-SCALE MODIFICATION OF SPEECH USING 3 LEVEL CLIPPING

In order to reduce the computational complexity of the conventional SOLA method, 3 level center clipping method is applied in the synchronization procedure and the frame similarity evaluation is performed only at the zero-crossing points. 3 level clipping method is widely used for speech signal processing and is a kind of nonlinear spectral flattening method [10]. Therefore, we can omit a normalization procedure with safety. Figure 1 shows the 3-level center clipping function, where CL is center clipping level. As can be seen from figure 1, 3 level clipped speech samples could have one of three values: -1, 0, 1. Therefore, one does not need to concern about the overflow problem in cases such as the implementation of the normalized cross-correlation on the fixed-point DSP or microprocessor.

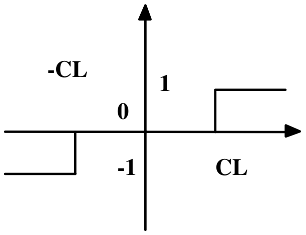

Figure 1: 3-level center clipping function

The equation to find a zero-crossing point is as follows:

$$Z(n) = \frac{\left|\operatorname{sgn}\{x(n)\} - \operatorname{sgn}\{x(n-1)\}\right|}{2} \tag{5}$$

where

$$\operatorname{sgn}\{x(n)\} = \begin{cases} +1, & x(n) \ge 0 \\ -1, & x(n) < 0 \end{cases}$$

If $Z(n)$ =1, then n is the zero-crossing point. The zero-crossing rate (ZCR) is calculated by

$$ZCR = \frac{1}{L} \sum_{n=0}^{L-1} Z(n) = \frac{1}{L} \sum_{n=0}^{L-1} \frac{\left|\operatorname{sgn}\{x(n)\} - \operatorname{sgn}\{x(n-1)\}\right|}{2} \tag{6}$$

The proposed cross-correlation function for the synchronization procedure is given by

$$\hat{R}_m(k_{zero\_cross}) = \frac{\sum_{j=0}^{L-1} \hat{y}(mS_s + k_{zero\_cross} + j)\,\hat{x}(mS_a + j)}{L}, \quad k_{\min} \le k_{zero\_cross} \le k_{\max} \qquad (7)$$

where

$\hat{y}$: 3-level clipped synthesized speech signal

$\hat{x}$: 3-level clipped analyzed speech signal

k_{zero_cross} : zero crossing point

Silence or unvoiced speech frames may have too many zero-crossing points. In this case, we need to reduce the search space for the synchronization procedure between the synthesized speech frame and the analyzed speech frame. We estimate the frame similarity with a certain interval in the case of the high zero-crossing rate frame and find the global synchronization point. And then, the local similarity is estimated among the adjacent points of the global synchronization point.

Table 1: Computational complexity comparison of several time-scale modification method

	SOLA	Proposed Method (1)	Proposed Method (2)	GLS-TSM
Relative Computations	100%	18.4%	16.6%	16.2%

Table 1 shows the computational complexity of several time-scale modification methods. In this table, proposed methods (1) and (2) are the method without and with considering the high zero-crossing rate frames, respectively. As can be seen from the table, the computational complexity of both the proposed methods (1) and (2) is reduced by over 80% comparing with the conventional SOLA method. This test is performed on the 486 100MHz PC platform and the relative computation means a percent value of each processing time which is normalized by the processing time of the conventional SOLA method. Speech signal is sampled at 8kHz and quantized into 16 bits. The analysis interframe interval, S_a, is 10 ms. The time-scale modification factor of 0.5 is applied in this experiment.

5. VARIABLE TIME-SCALE MODIFICATION OF SPEECH USING TRANSIENT INFORMATION [7]

The variable time-scale modification method is based on the idea that the transient portions of the speech signal plays a more important role than the steady portions in human speech perception[7]. Therefore, the transient portions of the speech signal are maintained while the steady portions are modified. This method gets the target rate by modifying steady portions only. In order to identify transient and steady portions of a speech signal, the LPC cepstral distance method and cross-correlation method are introduced [7]. After identifying transient and steady portions in a speech signal, the time scale of steady portions are modified only 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

$$T = T_t + T_s\,. \qquad (8)$$

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.

$$\alpha T = T_t + \alpha_s T_s\,. \qquad (9)$$

From (8) and (9), α_s can be represented as

$$\alpha_s = ((\alpha - 1)T + T_s\,)\,/\,T_s\,. \qquad (10)$$

With introducing new term, β, the ratio of steady portions in a speech signal, or $\beta = T_s\,/\,T$, (10) can be rewritten as

$$\alpha_s = ((\alpha - 1) + \beta)\,/\beta\,. \qquad (11)$$

The variable time-scale modification method specially produces more intelligible speech signal than the conventional SOLA method in the case of very fast playback.

6. EXPERIMENTAL RESULTS

A series of preference test by human listeners was conducted to evaluate the proposed time-scale modification using 3 level center clipping algorithm. 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. Three time-scale modification methods are implemented for this test: original SOLA, GLS-TSM and the proposed method. A listener preference test was done with speech data at two different time-scale modification factors; 0.7 and 1.5. The listeners consisted of 18 males and 2 females, with ages from 24 to 29. Listeners used headphones and took the test individually with the experimenter to minimize distractions.

Table 2: The choice basis for the subjective test

Choice Basis	Contents
1	B is indistinguishable from A
2	B is almost same as A
3	Somewhat different (A is a little better than B)
4	Different (A is much better than B)

A: Conventional SOLA method
B: GLS-TSM, the proposed method (1) or (2)

Table 2 shows the choice basis for the listeners. In order to perform a more reliable test, we played the same utterance twice to all 20 listeners and 8 listeners could not correctly recognize the equality of two utterances. Thus, we decided to exclude 8 listeners from the preference test, and remaining 12 listeners who are more sensitive to speech quality took part in the test.

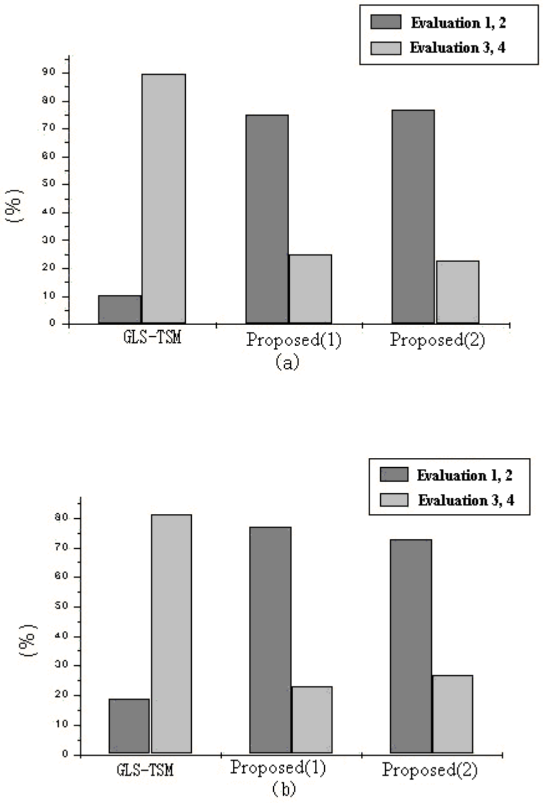

Figure 2: The result of preference test.
(a) $\alpha = 0.7$, (b) $\alpha = 1.5$.

Figure 2 shows the result of subjective preference test. The vertical axis in the figure indicates a percentage value of the choice count collected from 12 listeners. It can be seen from the figure that the performance of the proposed methods (1) and (2) is superior to that of GLS-TSM and is similar to that of the conventional SOLA method in that over 75 % of listeners' evaluation indicate that speech quality of proposed method is indistinguishable from or almost the same as that of the conventional SOLA method.

We also applied the proposed algorithm to the variable time-scale modification method [7]. By doing so, we can maintain the intelligibility of speech data in the case of very fast playback. In this case, we use Korean news speech data for speech materials and the time-scale modification factor is 0.5. LPC cepstral distance method is applied in order to separate transient and steady portions from the speech materials.

7. SUMMARY

The computationally efficient time-scale modification method presented in this paper takes advantage of 3 level clipping method and zero-crossing points of the speech signal in the synchronization procedure of the conventional SOLA method. By using 3 level clipped speech data for the synchronization procedure between the synthesized speech frame and the analyzed speech frame, we can omit the normalization part of synchronization procedure. In addition, we can feel free from the overflow problem in the real-time implementation based on a fixed-point DSP or microprocessor platform. By computing cross-correlation only at the zero-crossing points, we can reduce the search space in the waveform similarity estimation. As a result, we could reduce the overall computational complexity of the SOLA algorithm by over 80%. The result of preference test indicates that our proposed method can produce relatively good speech quality. Over 75 % of listeners' evaluation said that speech quality of proposed method is indistinguishable from or almost the same as that of the conventional SOLA method.

Additionally, we apply the variable time-scale modification algorithm to the proposed method. In this trial, we use Korean broadcast news data. The resulting time-scale modified speech is more intelligible than that produced by the conventional SOLA method in the case of very fast playback.

8. REFERENCES

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

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

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

[4] S. Roucos and A. M. Wilgus, "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.-J. Lee, H. D. Kim and H. S. Kim, "Variable time scale modification of speech using transient information," in *Proc. ICASSP*, pp.1319-1322, 1997.

[8] S. Yim and B. I. Pawate, "Computational efficient algorithm for time-scale modification (GLS-TSM)," in *Proc. ICASSP*, pp.1009-1012, May 1996.

[9] W. Verhelst, M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," Proc. ICASSP 93-II, pp.554-557, April 1993.

[10] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.