Fast Time Scale Modification Using Envelope-Matching Technique (EM-TSM)

Justy W.C. Wong *, Oscar C. Au**, Peter H. W. Wong ***

Department of Electrical and Electronic Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong.

Email: eejusty@ee.ust.hk*, eeau@ee.ust.hk**, eepeter@ee.ust.hk***

Tel.: +852 2358-7053**

ABSTRACT

Time scaling of speech and audio signals is one of the key features of the upcoming MPEG4 standard. Synchronized Overlap-and-Add (SOLA) is a time scaling algorithm known to achieve good speech quality. One problem of SOLA is that it requires a large amount of computation in the search of the best matching point between the analysis and synthesis frames. This is especially serious when multiple time scaling factors are to be supported as in the case of MPEG4. In this paper, we propose a technique called envelop-matching to simplify the computation with effectively the same quality. In envelop-matching, zero crossing locations are used as features for the search. We propose a very fast algorithm for the distortion computation.

1. INTRODUCTION

Time scale modification (TSM) [1-6] is a class of algorithms to change the time scale of a signal. By changing the apparent rate of articulation, TSM can be useful to make degraded speech more intelligible without changing the pitch information. There is an associated parameter, called TSM factor α . When it is one, the signal is unchanged. When α is greater than one, the signal is time expanded. When α is less than one, the signal is time compressed. Some TSM algorithms are time domain techniques such as overlap-and-add (OLA) and synchronized OLA (SOLA) [5]. Some are frequency domain techniques such as Least Square Error Estimation from Modified Short Time Fourier Transform Magnitude (LSEE-MSTFTM) [3].

This paper is concerned about the popular synchronized overlap-and-add (SOLA) which is relatively simple to implement and has good audio quality. The SOLA is based on OLA, which simply overlaps and adds adjacent frames. The analysis (input) frame length and the synthesis (output) frame length are related by the TSM factor. With the simple overlap-and-add operation, OLA may cause undesirable reverberation and clicks. SOLA solves this problem by over-lapping only at the points with highest similarity between the two overlapping

frames by maximizing the normalized cross-correlation between the analysis frame and synthesis frames. One weakness of the SOLA technique is the high computational cost to calculate the normalized cross-correlation function. To time scale one second of an audio clip with sampling frequency of 44.1kHz, approximately 1.8×10^8 or 180 MFLOPS are needed which is very high for the implementation of TSM by a single-processor machine. Many researches were carried out to develop fast algorithms for TSM [6,7]. In [7], a technique called global and local search (GLS-TSM) was developed to reduce the computation complexity by a factor of 40. However, the computation load is still high for a single-processor machine to combine the TSM with multimedia applications.

In this paper, we propose a technique called Envelope-Matching Time Scale Modification (EM-TSM) which further reduces the computation complexity of SOLA. Instead of the cross-correlation function in the SOLA, it matches the envelope between the analysis and synthesis frames using the sign information only. Subjective evaluations of synthesized signal were carried out and we found that the subjective quality of synthesized signals is almost the same as the signals synthesized by the SOLA method.

2. REVIEW OF SYNCHRONIZED OVERLAP-AND-ADD (SOLA)

The input (or analysis) signal $x[n]$ is segmented into overlapping frames of length N that are a distance of S_a apart. The first frame is directly copied to the output (or synthesis) signal $y[n]$. The $(m+1)$ -th frame which starts at $m \times S_a$ slides along the synthesized signal $y[n]$ around the location $m \times S_s$ in the range of $[k_{min}, k_{max}]$ to find a location which maximize the normalized cross-correlation function defined in (1) for the overlapping region.

$$R[k] = \frac{\sum_{i=0}^{L-1} y[m \times S_s + k + i] \cdot x[m \times S_a + i]}{\left[\sum_{i=0}^{L-1} y^2[m \times S_s + k + i] \cdot \sum_{i=0}^{L-1} x^2[m \times S_a + i] \right]^{\frac{1}{2}}} \quad (1)$$

The Sa and Ss are called the analysis and synthesis frame period respectively. The relation between Sa and Ss is defined in (2)

$$Ss = Sa \times \alpha \quad (2)$$

α is called the time scale factor. The signal is time scale expanded when α is greater than one and time scale compressed when α is smaller than one. L is the length of the overlapping region between the shifted analysis frame and synthesized signal. Usually the k_{min} and k_{max} are set to $-N/2$ and $N/2$ respectively. Once the location, which maximizes the cross correlation, is determined, the overlapping region is cross-faded and the remaining of the analysis frame is directly copied.

3. THE ENVELOPE MATCHING TECHNIQUE

3.1. Definition

The envelope matching technique is a modification of SOLA. It modifies the normalized cross-correlation function by using the sign information of the analysis and synthesis signals only. The normalization factor in (1) is replaced by the length of the overlapping region. Although the matching of the sign information is not a good measure of signal similarities, it can be used for signals with high cross-correlation, which is usually the case between the analysis signal and the synthesized signal. The modified function $R[k]$, which we call envelope-matching function (EMF), is defined in (3) and (4).

$$R[k] = \frac{\sum_{i=0}^{L-1} \text{sign}\{y[m \times Ss + i + k]\} \cdot \text{sign}\{x[m \times Sa + i]\}}{L} \quad (3)$$

$$\text{sign}(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ -1 & \text{for } x < 0 \end{cases} \quad (4)$$

After carrying out large amount of simulations, we found that k_{min} can be set to 0 and save the computation load by a half without any audible degradation in the sound quality of the output signal. Since the EMF operates on 1 and -1 only, it can be computed in a very efficient way. In section 3.2, we are going to show how the EMF can be computed in minimum number of operations. In section 3.3, we will show that only the values of k that the zero crossing points coincide are needed to be considered.

3.2. Computation of EMF

Procedures:

- I. Locate the zero crossing points for the overlapping region of the analysis and synthesis frames. Assume there are p and q

zero crossing points in the analysis and synthesis frame respectively.

Let,

$$A_k = \{a_{k_1}, a_{k_2}, \dots, a_{k_p}\}$$

$$B_k = \{b_{k_1}, b_{k_2}, \dots, b_{k_q}\}$$

be the set of zero crossing locations of the overlapping region between the shifted analysis frame and synthesis frame respectively, sorted in ascending order.

- II. Create a set C_k which is the exclusive-OR between A_k and B_k , i.e.

$$C_k = A_k \oplus B_k$$

Assume that the intersection set of A and B consists of r elements which means there are r pairs of zeros crossing locations which coincide, then

$$C_k = \{c_{k_1}, c_{k_2}, \dots, c_{k_{p+q-2r}}\}$$

- III. The $R[k]$ can be computed by

$$R[k] = \frac{\beta_k \left[2 \sum_{j=1}^{p+q-2r} (-1)^{j+1} c_{k_j} + (-1)^{p+q} L \right]}{L}$$

$$\beta_k = \text{sign}\{y[m \times Ss + k]\} \cdot \text{sign}\{x[m \times Sa]\} \quad (5)$$

The efficiency of the above method depends on the number of zero crossings between the analysis frame and the synthesis frame. For common audio signal sampled with frequency of 44.1kHz, the frame size N should be about 1000 (22ms) samples and the number of zeros crossing for a frame is in the order of 10^1 . Assuming an average of 40, the speed up factor is approximately 25.

3.3. The recursive relationship and monotonic property in EMF

Let K_i be the i^{th} smallest k within $[0, k_{max}]$ such that the intersection of A_k and B_k is not empty. If $K_{i+1} > K_i + 1$, then for any k such that $K_{i+1} > k > K_i$, we will show that $R[k+1]$ can be computed recursively from $R[k]$. Furthermore, we will show that $R[k]$ is monotonic within $[K_i, K_{i+1}]$ such that only $R[K_i]$ and $R[K_{i+1}]$ need to be computed in search of the minimum of $R[k]$.

Claim: If $K_{i+1} > K_i + 1$, then for any k such that $K_{i+1} > k > K_i$, $R[k]$ is a monotonic function of k .

Proof:
Let

$$b_{k_i} = c_{f_k(i)} \quad \text{for } i = 1, 2, \dots, q$$

$$\xi_k = \sum_{i=1}^q (-1)^{f_k(i)}$$

$$N[k] = \beta_k \left[2 \sum_{j=1}^{p+q-2r} (-1)^{j+1} c_{k_j} + (-1)^{p+q} L \right]$$

then,

$$b_{(k+\delta)_i} = b_{k_i} - \delta \quad \text{for } i = 1, 2, \dots, q$$

$$\beta_{k+\delta} = \beta_k$$

$$\xi_{k+\delta} = \xi_k$$

Case 1: The length of overlapping region L does not change with k

$$N[k + \delta] = N[k] + 2\delta \xi_k \beta_k$$

$$R[k + \delta] = R[k] + \frac{2\delta \xi_k \beta_k}{L} \quad (6)$$

Since β_k and ξ_k do not change with δ as long as no zero crossing location coincide, $R[k + \delta]$ is a linear function and monotonic.

Case 2: The length of overlapping region L decreases with k

$$N[k + \delta] = N[k] + 2\delta \xi_k \beta_k - (-1)^{p+q} \delta \beta_k$$

$$R[k + \delta] = \frac{R[k]L + 2\delta \xi_k \beta_k - (-1)^{p+q} \delta \beta_k}{L - \delta}$$

Differentiating $R[k + \delta]$ with respect to δ , gives

$$\frac{\partial R[k + \delta]}{\partial \delta} = \frac{L[R[k] + 2\xi_k \beta_k - (-1)^{p+q} \beta_k]}{(L - \delta)^2} \quad (7)$$

Since the sign of the derivatives of $R[k + \delta]$ does not change with δ , $R[k + \delta]$ should be monotonic.

Moreover, by exploiting the monotonic property of the EMF, only the values of k that the zero crossing points coincide with the same sign change need to be calculated. Combining with the speed up factor described in section 3.1, a speed up factor in the order of 10^2 can be obtained.

4. OBJECTIVE MESAURMENT OF SYNTHESIZED SIGNAL QUALITY

To evaluate the quality of the output signal, we employ an objective quality measurement function called the mean square difference between all the overlapping regions, with the smaller in mean square difference indicates better quality of the synthesized signal. The mean square difference is defined as:

$$E = \frac{1}{M} \sum_{m=1}^M \frac{1}{L_m} \sum_{i=0}^{L_m-1} [y(m \times Ss + k_{opt} + i) - x(m \times Sa + i)]^2 \quad (8)$$

k_{opt} is the optimal k that maximize the $R[k]$ in (1) or (3), L_m is the length of overlapping region corresponding to k_{opt} and M is the number frame number that the signal has.

5. SIMULATION AND RESULTS

We tested the EM-TSM technique for two signals. The first one is spoken by a male-speaker called 'Au', which is recorded in room condition with the sampling frequency of 8192Hz, the other one called 'Faye' is a song with background music sung by female singer, sampled at 44100Hz. Different time scale factors were tried and the mean square difference is shown in figure 2. It is assumed that the cross-correlation function is the optimal function for measuring the similarities between the analysis and synthesis frames. The mean square difference of synthesized signal using the cross-correlation function is also shown in figure 2 as a reference. The spectrogram of original 'Au', and the time scaled signals using different functions are shown in figure 1. A lot of subjective testing on the time scaled signal quality were carried out and we found it is very difficult to distinguish between the signals generated by using different functions. It should be noted that the increase in mean square difference using the EM-TSM technique mainly contributed by noise-like or unimportant regions where are less audible.

6. CONCLUSIONS

A fast technique for measuring the signal similarities is proposed for the time scale modification of signal using the SOLA based system. Speed up factor in the order of 10^2 can be obtained with very good speech quality. The reduction in computation complexity allows the time scale modification of signals to be implemented practically by a single-processor machine for multimedia applications.

Figure 1 (a) Top: Spectrogram of the original speech 'Au', (b) Left: Time-scaled 'Au' using cross-correlation Function with $\alpha = 0.5$ (c) Right: Time-scaled 'Au' using EMF with $\alpha = 0.5$

Figure 2 (a) Top: Mean square difference of 'Au' (b) Bottom: Mean square difference of 'Faye'

REFERENCE

- [1] David Malah, "Time Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, Apr. 1979.
- [2] M.R. Portnoff, "Time Scale Modification of Speech based on Short Time Fourier Analysis", *IEEE Trans. on Acoustics, Speech, Signal Processing*, Vol. 9, pp. 374-390, Jun 1981.
- [3] D.W. Griffin, J.S. Lim, "Signal Estimation from Modified Short Time Fourier Transform", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP_32, pp. 236-243, Apr. 1984.
- [4] Quatieri, R.S. McAulay, "Speech Transformation Based on a Sinusoidal Representation", *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Vol. 1, pp. 489-492, Mar. 1985.
- [5] S. Roucos, A.M. Wilgus, "High Quality Time Scale Modification for Speech", *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Vol. 1, pp. 493-496, 1985.
- [6] J. Laroche, "Autocorrelation Method for High-Quality Time/Pitch-Scaling", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 131-134, 1993.
- [7] S. Yim and B. I. Pawate, "Computationally Efficient Algorithm for Time Scale Modification (GLS-TSM)", *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Vol. 2, pp. 1009-1012, 1996.