Generalized Envelope Matching Technique for Time-Scale Modification of Speech (GEM-TSM)

Atsuhiro Sakurai

Texas Instruments Japan, Ltd., DCES Software Lab.
Miyukigaoka 17, Tsukuba-shi, Ibaraki 305-0841, Japan
a-sakurai@ti.com

Abstract
A new time-domain, non-pitch-synchronous method for time-scale modification targeted on broadband speech is proposed. The method is based on the SOLA (synchronous overlap-add) and EM-TSM (envelope-matching time-scale modification) methods, where the sign envelope of the EM-TSM method is replaced by a generalized envelope formed by the highest bits of the samples. (The actual number of bits will depend on word length constraints of the specific hardware.) In addition, a fixed length scheme for calculating cross-correlation is proposed, eliminating the need for normalization after computing each cross-correlation value. With these improvements, the proposed method outperforms EM-TSM both in terms of output quality and computational efficiency.

1. Introduction
TSM (time-scale modification) is an emerging topic in speech and audio signal processing due to the advance of low-cost, high-speed hardware that enables real-time processing by portable devices. Possible applications include intelligible sound in fast-forward playback, foreign language training, and others. TSM algorithms may be pitch-synchronous [1] or not, may operate in the time domain [1][2][3] or in the frequency domain [4], and may use a sinusoidal model [5]. For the case of inexpensive hardware targeted on broadband speech commonly accompanied by background music, non-pitch-synchronous, time-domain methods such as the EM-TSM (envelope-matching TSM) algorithm [3] are preferred due to an attractive trade-off between quality and computational cost, plus the fact that no discrimination needs to be made between speech and non-speech signals.

The EM-TSM algorithm is based on the earlier SOLA (synchronous overlap-add) [2] approach, which requires the calculation of the normalized cross-correlation between the overlapping signals for each possible overlap position around the desired value in order to select the final overlap. The shortcoming of the SOLA method is the high computational cost of the cross-correlation and normalization processes. A number of methods have been proposed to reduce the computational cost of SOLA (e.g. [6], which limits the search to a number of possible candidates). Among them, the EM-TSM method is the one that achieves the highest quality, albeit never higher than SOLA because the intrinsic limitations of SOLA are inherited (see Section 2.3). Moreover, computational complexity still remains an issue even with recent improvements that manage to reduce the average amount of computations [7].

In this paper, a new time-domain TSM method based on the SOLA and EM-TSM approaches is proposed. In the proposed method, the cross-correlation measure is calculated based on an envelope consisting of the n highest bits of the samples instead the sign envelope. The value of n depends on the word length of the specific hardware in question to avoid overflow, considering that the final cross-correlation value will be the result of repeated multiplications and additions along the calculation range. In the present experiments based on Texas Instruments TMS320C55x DSP architecture, n was set to 4. In addition, a fixed length for calculating the cross-correlation measure is proposed which eliminates the need for normalization after the calculation of each cross-correlation value. With these improvements, the proposed method makes full use of certain features of modern DSP architectures such as fast shifting and multiply-and-accumulate instructions and achieves a better trade-off between output quality and computational efficiency when compared to EM-TSM and SOLA. Listening tests indicate that the results obtained by the present method are equivalent to SOLA, outperforming EM-TSM.

This paper is organized as follows: Section 2 describes the SOLA and EM-TSM algorithms along with their limitations; Section 3 describes the proposed generalized envelope-matching scheme; Section 4 reports the results of tests conducted to evaluate the proposed method, and Section 5 contains final conclusions.

2. The SOLA and EM-TSM Algorithms

2.1. The SOLA Algorithm
The SOLA algorithm works as follows:
- Uniform and equidistant frames of the input speech signal are taken in an overlapped fashion.
- Frames are reconstructed in an overlap-add process after changing the overlap factor, i.e., the overlap is reduced to achieve time expansion or increased to achieve time compression.
- The overlap point is adjusted by computing a measure of signal similarity between the overlapping regions around the desired overlap position. The position of maximum similarity is selected.
- In the overlapping region, a fade-in/fade-out windowing process is utilized to smooth out artifacts.

The process is illustrated in Figure 1. $x[i]$ and $y[i]$ are respectively the analysis and synthesis signals (represented as sequences with index i), $N$ is the frame size, $S_a$ is the analysis interval between consecutive frames, $S_s$ is the synthesis interval between consecutive frames that would result in the desired time scaling, and $k$ is a small offset added to $S_s$ to...
obtain the actual output frame interval. $k$ is limited to the range $[-k_{\text{max}}, k_{\text{max}}]$. The role of $k$ is to allow for fine-tuning the overlap point based on a measure of signal similarity between the signals to be overlapped. Thus, $k$ must correspond to the overlap point that provides the maximum similarity between the signals within the allowed range $[-k_{\text{max}}, k_{\text{max}}]$. The maximum overlap offset value $k_{\text{max}}$ must be large enough to accommodate the period of the lowest tone-like frequency component of the signal. The signal similarity measure is obtained by calculating the cross-correlation between the signals to be overlapped according to equation (1), where $L_k$ is the length of the overlap range. This cross-correlation function must be evaluated for all values of $k$ within the range $[-k_{\text{max}}, k_{\text{max}}]$.

$$R[k] = \frac{\sum_{i=0}^{L_k-1} y[mS_x + k + i]x[mS_x + i]}{\left[\sum_{i=0}^{L_k-1} y^2[mS_x + k + i] + \sum_{i=0}^{L_k-1} x^2[mS_x + i]\right]^{1/2}}$$ (1)

Even though the normalized cross-correlation function of SOLA and also the simplified cross-correlation function used in EM-TSM provide a good measure of signal similarity, it is evaluated from a purely geometrical perspective, without taking into account the characteristics of human hearing. For example, it is widely known that short and peaky discontinuities in the signal produce a larger impact on output quality than long but small discrepancies. Figure 2 shows a hypothetical case where $x/f$ contains pulses (A) and (B), and $y/f$ contains pulses (A) and (B'). $x/f$ and $y/f$ must be overlapped at the overlap position that provides maximum signal similarity. The best overlap position is clearly the one that aligns pulses (A) and (A') in order to avoid an audible echo. However, this result cannot be found with the normalized cross-correlation measure of SOLA due to the influence of pulses (B) and (B'). This problem is even more evident in the EM-TSM method, whose cross-correlation function is maximized when pulses (B) and (B') (and not (A) and (A')) are aligned.

2.3. Limitations of the SOLA and EM-TSM Algorithms

2.3.1. The Incorrect Decision Problem

The problem above results from the fact that signal amplitudes are not taken into account. In the SOLA method, the normalization process eliminates the influence of the calculation length of the cross-correlation function (which is different for each overlap value) but the influence of signal amplitudes also disappears. On the other hand, the EM-TSM method normalizes only by the length of the calculation range but the problem persists due to the small dynamic range of the 1-bit sign envelope used in the calculation of cross-correlation. In other words, the limitation of SOLA is inherited by EM-TSM because the sign envelope disconsiders signal amplitudes. Therefore, it seems reasonable to believe that better perceptual results can be achieved by combining the normalization...
scheme used in EM-TSM with a generalized envelope that carries information about signal amplitude in addition to the signs of the samples. The consequence of this change in terms of number of computations compared to the EM-TSM method is to replace a sign detection operation by a right shift, which is advantageous in most DSP architectures.

2.3.2. The Division Problem

Another problem faced by the SOLA and EM-TSM algorithms is the computational cost involved in the division operation shown in (2), since most digital signal processors do not have inbuilt hardware to execute divisions in one machine cycle. Typically, the fastest implementation of 16-bit division requires at least 15 subtractions. Eliminating the need for normalization would result in a significant gain in computational efficiency.

3. The Generalized Envelope-Matching Algorithm (GEM-TSM)

The proposed method includes two solutions to address the problems pointed out in the previous section. The incorrect decision problem is dealt with by making use of a generalized envelope consisting of the highest bits of the samples, and the division problem is avoided by means of a fixed-length calculation range of the cross-correlation function. These two solutions combined constitute the proposed generalized envelope-matching method for time scale modification (GEM-TSM).

3.1. The Generalized Envelope

It is worth noting that the sign envelope can be seen as an envelope where each sample is represented by its most significant bit (after replacing 0’s by 1’s for positive numbers to ensure symmetry). In other words, the EM-TSM method uses just 1 bit from each sample while the SOLA method uses the entire word length, and the normalization scheme accounts for the different dynamic ranges used by each method. However, as seen in Section 3.1, disregarding the amplitude of the signals may lead to incorrect results. In the proposed method, we try to fix that problem by expanding the dynamic range of the signal envelope used in the EM-TSM method to a certain number of bits.

In principle, it is possible to use the full word length of the samples in combination with the same simplified normalization scheme of EM-TSM based only on the length of the calculation range. However, it is advantageous to use just the highest bits in order to avoid overflow problems in view of the limited word length of common DSP devices. Computer simulations indicate that the influence of the dynamic range of the signal envelope decreases rapidly for bit lengths larger than 3. In our case, we set the number of bits of the signal envelope to 4 to suit the particular DSP architecture being used (Texas Instruments TMS320C55x DSP).

In terms of computational cost, an envelope formed by the highest bits of the samples can be obtained by right shifting, which is completed in one machine cycle by most DSP architectures. Considering that the symmetry does not need to be corrected, such a generalized envelope is simpler to obtain than the sign envelope used in the EM-TSM method. The cross-correlation function using the generalized envelope is shown in Equation (3), where the symbol (>>) denotes right-shifting. Here, \( m \) is the shift value corresponding to the desired number of bits of the generalized envelope (e.g., \( m=11 \) for a 4-bit envelope and 16-bit samples).

\[
R^1[k] = \sum_{i=0}^{L_{e}-1} \{ y[mS_y + i + k] \gg \text{m} \} \{ x[mS_x + i] \gg \text{m} \} / L_e
\]

3.2. The Fixed Length Calculation Range for the Cross-Correlation Measure

In order to solve the computational problem related to the division operation executed for every overlap offset value \( k \), a simple solution is proposed: to fix the length of the range where the cross-correlation function is calculated and also the values of vector \( x[k] \). Instead of calculating the cross-correlation function along the entire overlap region, an effective computation region of vector \( x/[k] \) is defined based on the overlap region corresponding to \( k=0 \). Let \( L_{e} \) be the length of the overlap region for \( k=0 \). The effective computation region of \( x/[k] \) has a length \( L_{e} = L_0/2 \) and occupies the center of the overlap region corresponding to \( k=0 \), as shown in Figure 3. For calculating cross-correlation, the same values of vector \( x/[k] \) will be used for all \( k \), and the values of \( y/[k] \) will slide to the left or right according to the value of \( k \) within the range \( -k_{\text{max}} \) to \( k_{\text{max}} \). Note that reducing the computation range of the cross-correlation function does not produce a significant impact on the final selection of \( k \) because the region near the center of the overlap (where the energies of \( x \) and \( y \) have the same order of magnitude) is far more critical than at the borders due to the fade-in and fade-out windows used in the overlap-add process.

![Figure 3: Effective overlap region where the overlap-and-add takes place.](image)

By calculating cross-correlation only inside the effective computation region regardless of the overlap offset value \( k \), it is no longer necessary to normalize the correlation result, resulting in considerable savings in computational complexity. Furthermore, computation is also largely reduced (by roughly half) because of the shorter length of the cross-correlation calculation range, since it is possible to design \( L_{e} \) to be smaller than \( L_0 \) for all \( k \). The final cross-correlation function without
normalization after incorporating the generalized envelope is described in Equation (4).

\[ R^*[k] = \sum_{n=0}^{\infty} \left\{ \sum_{i=0}^{\infty} \left\{ \sum_{m=0}^{\infty} \left( \sum_{i^n=0}^{\infty} \sum_{i^m=0}^{\infty} \sum_{i^i=0}^{\infty} \sum_{i^j=0}^{\infty} \sum_{i^k=0}^{\infty} \sum_{i^l=0}^{\infty} \sum_{i^m=0}^{\infty} \sum_{i^n=0}^{\infty} \sum_{i^o=0}^{\infty} \sum_{i^p=0}^{\infty} \sum_{i^q=0}^{\infty} \sum_{i^r=0}^{\infty} \sum_{i^s=0}^{\infty} \sum_{i^t=0}^{\infty} \sum_{i^u=0}^{\infty} \sum_{i^v=0}^{\infty} \sum_{i^w=0}^{\infty} \sum_{i^x=0}^{\infty} \sum_{i^y=0}^{\infty} \sum_{i^z=0}^{\infty} \right\} \right\} \right\} \]  

(4)

4. Evaluation and Discussion

In terms of computational cost, it is possible to verify the advantages of the GEM-TSM method over the previous EM-TSM method (and obviously over the SOLA method as well) by simply comparing Equation (4) with Equations (1) and (2). As discussed above, the significant simplification results from the shorter size of cross-correlation computation range \((L_s < L_k)\) for all \(k\) by design), the elimination of the division operation, and (depending on the DSP architecture) the more efficient computation of the generalized envelope in comparison with the sign envelope.

In order to compare the quality of the proposed GEM-TSM method against EM-TSM and SOLA, blind listening tests were conducted as described below. The 3 algorithms were executed on 2 wideband (48kHz sampling frequency) input sounds containing respectively male and female speech. Playback speeds were set to 0.5x normal speed (S1) and 0.625x normal speed (S2), resulting in 4 processed utterances per method, and 12 utterances in all. The processed utterances were divided into 4 groups of 3 (containing one representative for each method) and presented to 4 listeners with some background in developing speech and audio-related software (not including the author). The listeners were asked to rank the output quality of the 3 samples within each group without any prior knowledge about them. Points were given to each sample (and its corresponding method) according to the order of preference in the following way: 3 points for the preferred one, 2 points for the second one, and 1 point for the least preferred one. Then, the total number of points for each method across all listeners was obtained. The results are shown in Table 1 and indicate that the quality of GEM-TSM was found to be slightly higher than EM-TSM and equivalent to SOLA. It is worth noting that the differences are extremely subtle, which is not surprising since the difference in quality between the SOLA and EM-TSM methods is known to be small.

Table 1: Results of listening experiments in scores

<table>
<thead>
<tr>
<th>Method</th>
<th>Utterances</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>S1</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>S2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>SOLA</td>
<td>4</td>
<td>11</td>
</tr>
</tbody>
</table>

5. Conclusions

A new high-quality and efficient time-scale modification method (GEM-TSM) is proposed. The proposed method uses a new cross-correlation function as a measure of signal similarity calculated inside a fixed-length range, eliminating the need for normalization after the calculation of cross-correlation. Listening tests indicate that the achieved quality is equivalent to SOLA and higher than EM-TSM in spite of its lower computational cost.

6. References