2004-01-01

Time-scale Modification of Music using a Synchronized Subband/Time-domain Approach

David Dorran
Dublin Institute of Technology, david.dorran@dit.ie

Robert Lawlor
National University of Ireland, Maynooth

Recommended Citation

Dorran, D. & Lawlor, R. (2004) Time-scale modification of music using a synchronized subband/time-domain approach. *IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.4, Montreal, Canada, 2004.*

TIME-SCALE MODIFICATION OF MUSIC USING A SYNCHRONIZED SUBBAND/TIME-DOMAIN APPROACH

David Dorran and Robert Lawlor***

Dublin Institute of Technology*. National University of Ireland, Maynooth**.
david.dorran@dit.ie, rlawlor@eeng.may.ie

ABSTRACT

Time-domain audio time-scaling algorithms are efficient in comparison to their frequency-domain counterparts, but they rely upon the existence of a quasi-periodic signal to produce a high quality output. This requirement makes them unsuitable for direct application to complex multi-pitched signals such as polyphonic music. However, it has been shown that applying time-domain algorithms on a subband basis can resolve this issue. Existing subband/time-domain approaches result in a reverberant/phasy artifact being introduced into the output due to poor synchronization between time-scaled subbands. This paper presents a number of synchronization schemes that greatly reduce the amount of reverberation/phasiness introduced into the time-scaled output by existing subband/time-domain approaches.

1. INTRODUCTION

Altering the time-scale of an audio signal can be achieved in the time-domain or frequency-domain with advantages and disadvantages associated with each approach. Frequency-domain techniques are capable of applying high quality time-scale modifications to a variety of complex audio signals within a wide range of time-scale factors, but their versatility comes at the expense of their computational requirements. Time-domain techniques are comparatively computationally efficient and operate by simply discarding or repeating suitable segments of the audio signal. The discard/repeat process relies heavily upon the existence of a quasi-periodic waveform, making time-domain approaches suitable for speech and monophonic music but unsuitable for their direct application to most polyphonic music due to the generally complex multi-pitch nature of these types of waveforms. However, [1], [2] and [3] have demonstrated that applying time-domain time-scale modification algorithms on a subband basis can resolve this issue.

One problem with the subband/time-domain approach is that a reverberant/phasy artifact is introduced into the time-scaled output due to poor synchronization between time-scaled subbands, as explained in [3]. Lack of synchronization between subbands is also noticeable during 'hard' transients, resulting in the time-scaled transients sounding metallic and harsh. This paper addresses the subband synchronization problem by choosing an appropriate subband segment to discard/repeat by analyzing all subbands collectively; restoring synchronization of subbands during masked and silent regions; and forcing synchronization at transients.

This paper is structured as follows. The synchronized overlap-add (SOLA) algorithm [4] and an efficient variant of SOLA, the variable-parameter SOLA (VSOLA) [5], are outlined in section 2. Section 3 presents an overview of the subband approach and discusses the issues involved with its implementation; the problem of subband synchronization is highlighted. In section 4 a number of schemes are presented that address the synchronization problem. Section 5 presents the results of applying the schemes described in section 4 to a variety of music signals and sections 6 concludes this paper.

2. SOLA AND VSOLA

SOLA segments the input signal x into m overlapping frames, of length N samples, each segment being S_a samples apart. S_a is the analysis step size. The time-scaled output y is synthesized by overlapping successive frames with each frame a distance of $S_s + k_m$ samples apart. S_s is the synthesis step size, and is related to S_a by $S_s = \alpha S_a$, where α is the time-scaling factor. k_m is an offset that ensures that successive synthesis frames overlap in a synchronous manner. k_m is chosen such that

$$R_m(k)=\frac{\sum_{j=0}^{L_m(k)-1} y(mS_s+k+j)x(mS_a+j)}{\sqrt{\sum_{j=0}^{L_m(k)-1} x^2(mS_a+j)\sum_{j=0}^{L_m(k)-1} y^2(mS_s+k+j)}} \tag{1a}$$

is a maximum for $k = k_m$, where m represents the m^{th} input frame and $L_m(k)$ is the length of the overlapping region i.e.

$$L_m(k) = N - S_s + k_{m-1} - k \tag{1b}$$

k is in the range $k_{min} \leq\ k \leq k_{max}$.

$R_m(k)$ is a correlation function which ensures that successive synthesis frames overlap at the 'best' location i.e. that location where the overlapping frames are most similar. Having located the 'best' position at which to overlap, the overlapping regions of the frames are weighted prior to combination, generally using a linear or raised-cosine function (see [5] for details).

Typically, N is fixed at 30ms for speech and 40ms for music, S_a is in the range of $N/3$ to $N/2$, k_{min} is $-N/2$ and k_{max} is $N/2$. However, in [5] (VSOLA) a set of optimum analysis parameters were derived which reduce the number of iterations required for SOLA's implementation. These parameters are given by

$$S_a = (L_{stat} - SR)/|1-\alpha| \tag{2}$$

$$N = SR + \alpha S_a \tag{3}$$

where SR is the search region, which corresponds to two cycles of the longest likely pitch period of the input waveform and L_{stat} is the stationary length, which corresponds to the maximum length of segment that can be discarded/repeated during an iteration of the algorithm. Typical values for L_{stat} and SR, for music, are 33ms and 20ms, respectively; for VSOLA k_{min} and k_{max} are set to 0 and *SR,* respectively.

3. SUBBAND APPROACH

As mentioned in the introduction, time-domain techniques rely upon the existence of a quasi-periodic signal to produce a high quality output. Partitioning a complex multi-pitch signal into appropriate subbands results in a set of signals that are suitable for time-scale modification using time-domain techniques. Time-scaled subbands can then be summed to produce a time-scaled version of the original signal, as illustrated in figure 1. The major issues concerning a subband approach are the partitioning of a complex waveform into subbands of lesser complexity, that are suitable for time-scale modification in the time-domain, and the recombination of the time-scaled subbands in a synchronous manner. The solutions to these issues are diametrically opposite since partitioning a complex waveform into many subbands reduces the complexity of each subband but increases potential subband synchronization problems and vice versa. While [1] and [2] partition the complex signal into subbands using uniform width filterbanks, [3] justified the use of a non-uniform width filterbank based upon the bark scale for the improved time-scale modification of Western tonal music. However, subband synchronization still remains an issue.

Subband synchronization issues arise because time-domain time-scale modification techniques require an offset to ensure that successive synthesis frames overlap in a synchronous manner. Each subband will almost certainly require a different offset, resulting in poorly synchronized subbands. The subband synchronization problem can be simulated by first partitioning the signal into subbands; then passing each subband through a random delay ranging from 0 to some maximum delay, d_{max}. In [6], using a similar model, it was found that delay differences of 0.4ms can be perceived as distortion by trained listeners. To model the use of VSOLA in a subband implementation the d_{max} value would be set equal to *SR* i.e. approximately 20ms.

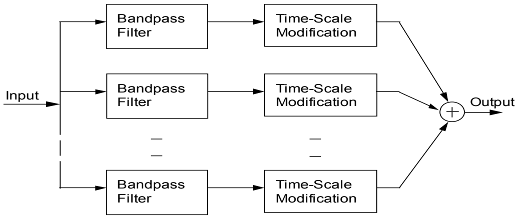

Figure 1. *Subband approach to time-scale modification.*

4. IMPROVED SUBBAND SYNCHRONIZATION

4.1. Choice of subband offsets

When dealing with quasi-periodic signals the correlation function of the SOLA algorithm, $R_m(k)$ generally returns a periodic signal with prominent peaks corresponding to the pitch period of the input signal as illustrated in figure 2 (a); a fact that has been exploited by a number of pitch detection algorithms. The SOLA algorithm chooses a synthesis offset, k_m, related to the most prominent or maximum peak of the correlation function. In general, however, any offset that is related to any of the prominent peaks of the correlation function could be used and would result in a high quality output. While choosing the offset that corresponds to the maximum peak in the correlation function is an obvious choice when SOLA is directly applied to a broadband input signal, for a subband implementation the offset for each subband should be chosen so as to minimize the delay differences between subbands in order to reduce the amount of reverberation/phasiness introduced into the output.

In attempting to determine the ‘best’ offset for each subband synthesis frame a set of suitable offsets must be established. The first step in achieving this aim is to unbias the correlation function so that its magnitude values are not biased toward a large overlap. The effect of unbiasing the correlation function is illustrated in figure 2 (b) with the unbiased correlation function $R'_{m,i}(k)$ given by

$$R'_{m,i}(k) = \frac{R_{m,i}(k)}{L_{m,i}(k)} \tag{4}$$

where the i subscript is introduced to represent the i^{th} subband.

A simple method of determining prominent peaks, and hence suitable offsets, is to first locate all peaks in $R'_{m,i}(k)$, where a peak is defined as a sample that is greater than its two nearest neighbors. Then, any peak of the unbiased correlation function that is within 10% of the maximum peak’s magnitude is considered a candidate peak, from which corresponding candidate offsets can be found. These set of subband candidate offsets are denoted $\{k_{c1,i}, k_{c2,i}, k_{c3,i}, \ldots, k_{cp,i}\}$. An efficient approach to determine the ‘best’ offset from this set of candidates is to provide a global target offset, k_{target}, to which all subband offsets should be focused i.e. for each set of subband candidate offsets the offset that is closest to the global target is used. k_{target} is chosen such that

$$R_{m,sum}(k) = \sum_{i=1}^{J} R'_{m,i}(k) W_i \tag{5}$$

is a maximum for $k = k_{target}$, where J is the number of subbands and W_i is a subband perceptual weighting factor. $R_{m,sum}(k)$ is most influenced by subbands with the greatest energy and W_i provides an additional weight towards those subbands that are perceptually louder. For simplicity the standard ‘A’ loudness-weighting curve is used in calculating W_i for each subband, where the center frequency for each of the J subbands is used to determine the relevant weighting factor from the ‘A’ weighting curve.

Then, as described above, the offset for the i^{th} subband, $k_{m,i}$, is chosen such that

$$D_{m,i}(k_c) = |k_{target} - k_c| \tag{6}$$

is a minimum for $k_c = k_{m,i}$ with k_c being every element in the set of candidate offsets in the i^{th} subband i.e. $\{k_{c1,i}, k_{c2,i}, \ldots, k_{cp,i}\}$.

It should be noted that the approach for determining the ‘best’ offset for each subband described above requires that the same analysis parameters, S_a and N, be applied to all subbands.

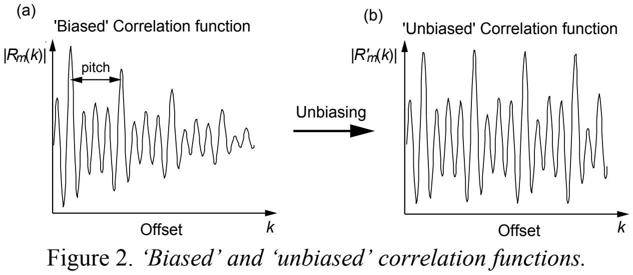

Figure 2. *‘Biased’ and ‘unbiased’ correlation functions.*

4.2. Synchronization during silent/masked regions

Masked regions and regions of silence within subbands can be utilized for subband synchronization purposes. Consider the case where, for an iteration of the VSOLA algorithm, within one of the subands, the energy in the overlapping region of the overlapping synthesis frames falls below the threshold of hearing or the masking threshold, as given in [7]; then any offset could be used to overlap the frames without the introduction of audible distortion (once an adequate overlap for cross-fading is provided so as to eliminate the possibility of clicking). When this situation occurs within a subband, the offset for that suband is set to the global target offset, k_{target}, described above, thereby improving synchronization between subbands.

For the case where a region of silence or masked region occurs in some position along a synthesis frame other than within the overlapping region, as shown in figure 3, some level of synchronization can once again be established. Improved synchronization is achieved by altering the length of the silent/masked region from L_r to $L_r + (k_{target} - k_{m,i})$, where $k_{m,i}$ is the offset used by the subband in the overlapping region, thereby ensuring that all portions of the subband after the silent/masked region are synchronized to the global target. The expansion/compression of the silent/masked region r of length, L_r, assuming $L_r \geq SR$, is achieved by replacing r in the frame with $r_{replacement}$, of length $L_r + k_{target} - k_{m,i}$, where

$$r_{replacement}(j) = r(j), \text{ if } j \leq k_{target} - k_{m,i} \qquad (7a)$$

$$r_{replacement}(j) = (1-f(j))r(j) + r(k_{m,i} - k_{target} + j)f(j), \quad \text{if } k_{target} - k_{m,i} < j < L_r \qquad (7b)$$

$$r_{replacement}(j) = r(k_{m,i} - k_{target} + j), \text{ if } j \geq L_r \qquad (7c)$$

for $1 \leq j \leq L_r + k_{target} - k_{m,i}$,
where

$$f(j) = (j - \max(k_{target} - k_{m,i}, 1)) / (L_r - |k_{target} - k_{m,i}| - 1) \qquad (7d)$$

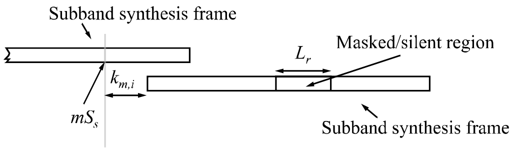

Figure 3. *A masked/silent region within a synthesis frame*

4.3. Synchronization of transients

Transients have posed a problem for all time-scale modification algorithms, both time-domain and frequency-domain, and must be treated differently to other portions of the signal in order to produce a high quality output. Typical artifacts related to time-domain handling of transients are the repetition or skipping of transients and, for a subband approach, the introduction of a harsh metallic effect within the transient portion. In [8] a solution to the transient handling problem within (non-subband) SOLA is proposed in which transient portions of the input are translated to their new time-scaled positions without modification, therefore keeping them in tact, while non-transient portions are time-scaled to a greater degree to ensure that the overall signal is time-scaled to the desired duration. Handling transients in this manner has an added advantage for a subband implementation since, as well as removing any harshness from time-scaled transients, it brings the subsequent subband frames back into a synchronized state. A subband approach requires additional factors to be taken into consideration for transient handling. A description of a time-domain/subband approach to preserving transients is given below.

Having determined the start of each transient $\{t_1, t_2, \ldots, t_q\}$, using the approach set out in [9], the input signal is divided into segments containing at most one transient, with any transient present being positioned at the start of a segment. It should be noted that any transient detection approach could be used and [9] was chosen arbitrarily. Then, given an input x the first segment, s_1, is given by a sequence of samples from x, starting at $x(1)$ and finishing at $x(t_1 - 1)$, i.e.

$$s_1 = x(1 + j), \; 0 \leq j < t_1 \qquad (8a)$$

Subsequent segments are given by

$$s_{w+1} = x(1 + j), \; t_w - ov \leq j < t_{w+1}, \text{ for } 1 \leq w \leq q - 1 \qquad (8b)$$

where ov is a small overlap of approximately 20 ms duration that is used to recombine segments in a synchronous manner during the final synthesis stage and q is the number of transients detected. Each segment s_2 to s_q then contains a transient at the start of its segment. The final segment is given by

$$S_{q+1} = x(1 + j), \; t_q - ov \leq j \leq L_x \qquad (8c)$$

where L_x is the length of the input signal x.

Both the first and last segments, s_1 and $s_{q+1,}$ do not contain transients and can be time-scaled in the usual subband manner and by incorporating the synchronization schemes described above. However, the overlap, ov, provided in the last segment should not be time-scaled. The remaining segments, from s_2 to s_q, are handled slightly differently; the first $ov + L_{tr}$ samples are extracted from each segment prior to time-scale modification, where L_{tr} is the length of the transient portion and is typically set to equate to 10ms. The remaining portion of the segment is time-scaled in the usual subband manner and by incorporating the synchronization schemes described above, however, each subband must be time-scaled to a greater degree to take into consideration the fact that the transient portion is not time-scaled. The updated time-scale factor, γ, should be such that

$$\gamma (L_{seg} - L_{tr}) + L_{tr} = \alpha L_{seg} \qquad (9a)$$

where L_{seg} is the length of the segment.
Therefore,

$$\gamma = (\alpha L_{seg} - L_{tr})/(L_{seg} - L_{tr}) \qquad (9b)$$

Furthermore, the length of the individual time-scaled subbands will generally not be the same; therefore, all subbands must be truncated to the length of the shortest subband before summing. The transient portion, with the overlap, ov, is pre-pended to the time-scaled portion of the segment. These portions will join in a continuous/synchronous manner since VSOLA does not alter the first few samples of the signal to be time-scaled.

Having synthesized the individual time-scaled segments; they must then be recombined to produce a time-scaled version of the original signal. The overlap provided during the segmentation process is used to ensure the individual segments combine in a synchronous manner. A correlation function, similar to that of equation (1a), is used to identify the 'best' overlap position for successive segments. The correlation function is given by

$$R_w(k) = \frac{\sum_{j=0}^{ov-k-1} s_w(L_{sw} - ov + k + j)s_{w+1}(j)}{\sqrt{\sum_{j=0}^{ov-k-1} s_{w+1}^2(mS_a + j) \sum_{j=0}^{ov-k-1} s_w^2(L_{sw} - ov + k + j)}} \qquad (10a)$$

where L_{sw} is the length of the segment s_w, where the w subscript represents the w^{th} segment . The overlap used, ov_w, is given by

$$ov_w = ov - k_w - 1 \tag{10b}$$

where k_w is chosen such that $R_w(k)$ is maximized for $k = k_w$, for k in the range $0 \leq k \leq ov$.

The segments are then linearly cross-faded in the overlapping region to produce a time-scaled version of the original signal with the transients kept in tact and the subband frames immediately following a transient perfectly synchronized.

5. RESULTS

11 evaluation subjects of various age and gender carried out informal blind listening tests. The test comprised of 6 comparisons between a variety of music tracks time-scaled using a subband approach both with and without the synchronization schemes described in this paper. The test used time-scale factors ranging from 0.66 to 2 and all tracks were sampled at 44.1kHz. The 'non-synchronized' tracks were time-scaled using the same parameters given in [3]. The 'synchronized' tracks were partitioned into subbands using the same cutoff frequencies as in [3], with SR and L_{stat} set to 20ms and 33ms, respectively, for all subbands. All thresholds were set assuming the maximum amplitude corresponded to 90dB(SPL).

The results of the listening tests indicate a strong preference for music time-scaled using the synchronization schemes described in this paper. The results of the listening tests are summarized in table 1.

Test subjects indication	**% of total comparisons**
Synchronization (sync) much better than no sync.	12.1 %
Sync slightly better than no sync.	31.8 %
Sync approach equal to no sync.	45.5 %
Sync slightly worse than no sync.	10.6 %
Sync much worse than no sync.	0.0 %

Table 1. *Summary of listening test results.*

Figure 4 illustrates the effect of synchronization upon a small excerpt from a time-scaled oboe signal. It can be seen that the temporal structure of the original waveform is maintained to a greater degree when synchronization techniques are employed. Figure 5 illustrates the effect of synchronization upon an excerpt from a time-scaled signal composed of a guitar and castanets; the transient portion is preserved and is not subject to spreading.

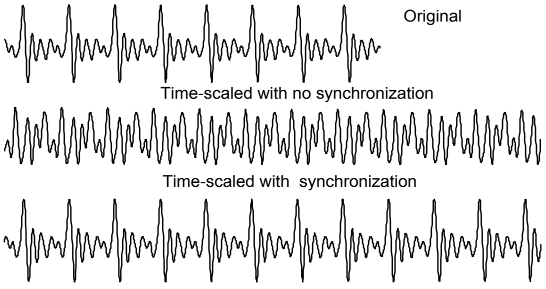

Figure 4. *The effects of synchronization on an oboe signal*

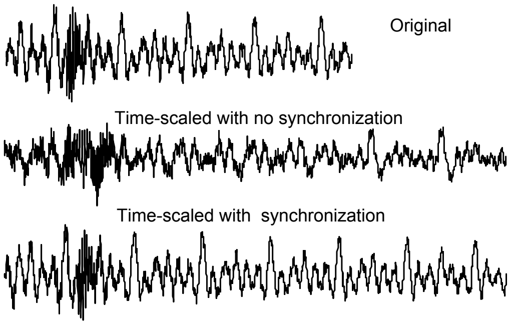

Figure 5. *The effects of synchronization on a guitar and castanets signal*

6. CONCLUSION

Time-domain/subband approaches to time-scale modification are efficient, but result in transients sounding harsh and also introduce a reverberant/phasy artifact into the time-scaled output. These artifacts are caused by the lack of synchronization between time-scaled subbands. This paper presents a number of subband synchronization schemes that greatly reduce the presence of these artifacts. The use of these schemes is supported through subjective listening tests.

7. REFERENCES

[1] Spleesters, G., Verhelst, W. and Wahl, A., "On the application of automatic waveform editing for time warping digital and analog recordings", *Proc. 96th Audio Eng. Soc. Convention*, Amsterdam, preprint 3843, Feb. 1994.

[2] Tan, R.K.C., Lin, A.H.J, "A Time-Scale Modification Algorithm Based on the Subband Time-Domain Technique for Broad-Band Signal Applications", *Journal of the Audio Engineering Society*, vol. 48, pp. 437-449, May 2000.

[3] Dorran D., Lawlor, R., "Time-scale modification of music using a subband approach based on the bark scale", Accepted for *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, Oct. 2003.

[4] Roucos S., Wilgus A.M., "High Quality Time-Scale Modification for Speech", *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, pp. 493-496, Mar. 1985.

[5] Dorran D., Lawlor, R., "An efficient time-scale modification algorithm for use within a subband implementation", *Proc. Int'l Conf. on Digital Audio Effects (DAFx03)*, London, pp. 339-343, Sept. 2003.

[6] Blauert J., Laws P., "Group Delay Distortions in Electroacoustical Systems", *Jour. of the Acoustical Society of America*, vol. 63, no. 5, pp. 1478-1483, May 1978.

[7] Johnston, J.D., "Transform coding of audio signals using perceptual noise criteria", *IEEE Journal on selected areas in communication,* vol. 6, issue 2, pp. 314-323, Feb. 1998.

[8] Lee S., Kim H.D., Kim H.S., "Variable time-scale modification of speech using transient information", *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1319 –1322, Apr. 1997.

[9] Scheirer, E. D., "Tempo and Beat Analysis of Acoustic Musical Signals", *Journal of the Acoustical Society of America*, vol. 103, issue 1, pp. 588-601, Jan. 1998.