A Hybrid Time−Frequency Domain Approach to Audio Time−Scale Modification

Conference Paper · November 2005

DOI: 10.1109/ASPAA.2005.1540224 · Source: IEEE Xplore

David Dorran
Technological University Dublin - City Campus

Eugene Dermot Coyle
Technological University Dublin - City Campus

R. Lawlor
National University of Ireland, Maynooth

A Hybrid Time-Frequency Domain Approach to Audio Time-Scale Modification*

DAVID DORRAN[1], AES Member, **ROBERT LAWLOR**[2], **AND EUGENE COYLE**[1]

[1] Digital Media Centre, Dublin Institute of Technology, Aungier Street, Dublin 2, Ireland.

[2] Department of Electronic Engineering, National University of Ireland, Maynooth, Co. Kildare, Ireland.

Frequency-domain approaches to audio time-scale modification introduce a reverberant/phasy artefact into the time-scaled output. Such artefacts are generally not present within time-domain implementations; however, high quality time-scaling in the time-domain is typically limited to quasi-periodic signals such as speech. A hybrid method of time-scaling is presented which draws upon appealing aspects of both time-domain and frequency-domain implementations. The technique described can be

* This paper collates and extends upon work previously published in [7], [9] and [20].

successfully applied to a wide range of audio and is both robust and efficient. Subjective testing demonstrates that the technique significantly reduces the presence of phasiness.

0 INTRODUCTION

Time-scale modification of audio alters the duration of an audio signal whilst retaining the signals local frequency content, resulting in the overall effect of speeding up or slowing down the perceived playback rate of a recorded audio signal without affecting its perceived pitch or timbre.

There are two broad approaches used to achieve a time-scaling effect i.e. time-domain and frequency-domain; an overview of both can be found in [1]. Time-domain algorithms, such as the synchronized overlap-add (SOLA) algorithm [2], are generally more efficient than their frequency-domain counterparts, but require the existence of a strong quasi-periodic element within the signal to be time-scaled in order to produce a high quality output. This makes them generally unsuitable for their application to complex audio such as multi-pitched polyphonic music. Frequency-domain techniques, such as the phase vocoder [3] and sinusoidal modeling [4], are capable of time-scaling complex audio but tend to introduce a reverberant/phasy artifact into the time-scaled output. This artifact is generally more objectionable in speech than in music. Music recordings typically contain a significantly higher level of reverberation than speech and so the additional reverberation introduced through time-scaling is not as noticeable and, as such, not as objectionable.

In general, time-domain techniques will be applied to speech, while frequency-domain techniques will be applied to more complex sounds. Where the source characteristics are unknown, as in video-cinema frame rate conversion [5] and during the time-scaling of television and radio adverts, frequency-domain techniques will generally be applied. However, the time-domain approach outlined in [6], can be successfully applied to complex sounds for small time-scaling factors (+/- 15%).

A hybrid time-frequency domain approach is presented that takes advantage of certain aspects of each broad approach to realize an efficient and robust time-scaling implementation. The hybrid approach reduces the presence of phasiness associated with frequency-domain implementations [7]. The approach takes advantage of a certain amount of flexibility, which is shown to exist, in the choice of modified STFT phase values. The approach then uses the flexibility derived to improve upon vertical phase coherence, thereby reducing the phasiness effect. Time-domain based techniques are also employed to further improve phase coherence and help bridge the gap that exists between the two broad approaches.

This paper is structured as follows: Section 1 provides an overview of the SOLA time-domain algorithm, which is used within the hybrid implementation; Section 2 outlines the basic operation of the improved phase vocoder [8], which makes use of sinusoidal modeling techniques to improve upon the standard phase vocoder; Section 3 presents an analysis of the phase tolerance which is shown to exist within implementations of the phase vocoder [9] and demonstrates how this tolerance can be used to push/pull phase components of an STFT representation back into a phase coherent state; Section 4 describes the hybrid approach which incorporates both time-domain and frequency-domain features through manipulation of the phase tolerance identified; Section 5 discusses the application of the hybrid algorithm to multi-channel recordings; Section 6 presents a summary of subjective listening tests undertaken which compare the quality of the hybrid approach to the improved phase vocoder; Section 7 concludes.

1 SYNCHRONIZED OVERLAP-ADD

Time-domain algorithms operate by appropriately discarding or repeating suitable segments of the input; with the duration of these segments being typically an integer multiple of the local pitch period (when it exists). Time-domain techniques are capable of producing a very high quality output when dealing with quasi periodic signals, such as speech, but have difficulty with more complex audio, such as multi-pitched polyphonic audio [10]. It should be noted that fewer discard/repeat segments are required the closer the desired time-scale duration is to that of the original duration [10]; Therefore time-domain algorithms produce particularly high quality results for time-scale factors close to one, i.e. when the desired time-scaled duration is close to the original duration. This is due to the fact that significant portions of the output are directly copied, without processing, from the input to the time-scaled output for time-scale factors close to one.

The SOLA algorithm achieves the discard/repeat process by first segmenting the input into overlapping frames, of length N, with each frame 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 $S_s + \tau_m$ 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. τ_m is a offset that ensures that successive synthesis frames overlap synchronously. τ_m is chosen such that the correlation function $R_m(\tau)$, given by

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

is a maximum for $\tau = \tau_m$, where x is the input signal, y is the time-scaled output, L_m is the length of the overlapping region and τ is in the search range $0 < \tau < \tau_{max}$.

Figure 1 illustrates an iteration of this process, whereby an input frame is appended to the current output. From the figure it can be appreciated that, by appropriately changing the overlap between successive frames, two pitch periods of the input are effectively discarded, resulting in the overall reduction of the duration of the signal.

Standard SOLA parameters are generally fixed, with N typically being 20-30 ms and $S_a = N/2$; However in [11] an adaptive and efficient SOLA parameter set is derived, which is also used in the hybrid implementation described in section 5 and is given by

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

$$N = SR + \alpha S_a \qquad (3)$$

where L_{stat} is the stationary length, i.e. the duration over which the audio signal does not significantly change (approx 25-30ms), and SR is the search range over which $R_m(\tau)$, and hence τ_m, is determined. For quasi-periodic signals SR should be set such that its value is greater than the longest likely period within the signal being time-scaled (generally about 12-20ms). If the search range is less than the longest pitch period, the duration of the discard/repeat segment may then be less than a pitch period; This would result in loss of synchronization between overlapping frames and perceptually objectionable discontinuities would be introduced into the time-scaled output if time-domain techniques are applied.

2 IMPROVED PHASE VOCODER

Time-domain techniques maintain ‘horizontal’ synchronization between successive frames by ensuring that a high level of similarity exists between the frames prior to overlap-adding; as such, time-domain techniques require the input to be suitably periodic in nature. When dealing with more complex signals, such as polyphonic music, such a high level of similarity between frames is unlikely to exist, and artefacts will subsequently be introduced.

To resolve this problem phase vocoder implementations operate by maintaining ‘horizontal’ synchronization along quasi-sinusoidal subbands [8]; this approach ensures that no objectionable discontinuities are introduced at a subband level. It follows that if discontinuities do not exist at a subband level then they will not exist at a broadband level either, and the output will be free of the objectionable discontinuity-based artefacts associated with time-domain implementations.

Standard implementations of the phase vocoder make use of uniform width filterbanks to extract the quasi-sinusoidal subbands, typically through the efficient use of a short-time Fourier transform (STFT). Horizontal synchronization (or horizontal phase coherence [8]) is maintained at a subband level by ensuring that the expected phase of each sinusoidal component follows the sinusoidal phase propagation rule i.e.

$$\varphi_2 = \varphi_1 + \omega(t_2 - t_1) \tag{4}$$

where φ_1 is the instantaneous phase at time t_1, ω is the frequency of the sinusoidal component, and φ_2 is the expected phase of the sinusoidal component at time t_2.

During time-scale modification, magnitude values of the sinusoidal subband components are simply interpolated or decimated to the desired duration. In [12] time-scale expansion is achieved by appropriately repeating STFT windows e.g. to time-scale by a factor of 1.5 every second window is repeated; similarly time-scale compression is achieved by omitting windows e.g. to time scale by a factor of 0.9 every tenth analysis window is omitted. The phase propagation formula of equation (4) is then applied to each subband (or discrete Fourier Transform (DFT) bin), from window to window.

In [8] it is recognized that not all subbands are true sinusoidal components, and some are essentially 'interference' terms introduced by the windowing process of the STFT analysis. [8] notes that applying the phase propagation rule to these interference terms results in a loss of 'vertical phase coherence' between subbands which introduces a reverberant or phasy artifact into the time-scaled output. The solution to this problem is to identify 'true' sinusoidal components through a magnitude spectrum peak peaking procedure and applying the phase propagation rule to these components only. The phases of the subband components in the 'region of influence' of a peak/sinusoidal subband are then updated in such a manner as to preserve the original phase relationships [8].

Whilst [8] results in improved vertical phase coherence between a true sinusoidal component and its neighboring interference components, it does not attempt to maintain the original phase relationships that exist between true sinusoidal components. The loss of phase coherence between these components also results in the introduction of reverberation. This problem is addressed in the literature, whereby the phase relationship or ‘relative phase difference’ between harmonically related components of a harmonic signal is maintained through various techniques e.g. [13-15]. These approaches, however, require the determination of the local pitch period. Whilst the techniques of [13-15] attempt to maintain vertical phase coherence through the manipulation of the phase values of harmonically related sinusoidal components, time-domain approaches implicitly maintain vertical phase coherence by virtue of the fact that the broadband signal is not partitioned into subbands in the first place.

3 PHASE FLEXIBILITY WITHIN THE PHASE VOCODER

In [9] the effect of displacing the horizontal phase of a pure sinusoidal component from its ideal/expected value, within a window of the phase vocoder, is considered. It is shown in [9] that such a loss in phase coherence results in a certain amount of amplitude and frequency modulation being introduced into the sinusoidal component. This can be appreciated from Figure 2, which shows the result, c(t), of summing the two sinusoidal components a(t) and b(t), which have been multiplied by alternate halves of a hanning window (as is typically the case within a phase vocoder implementation).

As can be seen from the figure, when a(t) and b(t) are perfectly phase coherent the result is a pure sinusoid. However when a phase difference is introduced the resulting waveform is no longer a perfect sinusoid, but is modulated in amplitude. The waveform is also modulated in frequency, but this is difficult to appreciate from the figure.

In [16] it is empirically shown that the human auditory system does not perceive certain amounts of amplitude and frequency modulations. In an effort to determine the maximum phase tolerance that can be introduced into an STFT representation without introducing audible distortion, a set of equations representing the situation described above is presented below. The phase tolerance identified is used in section 3.1 to push/pull phase values into a coherent state.

The first step in achieving this aim is to describe the above situation through the use of a vector representation. From figure 3, the ramped sinusoidal components are represented by the vectors *a*(*t*) and *b*(*t*), which vary with time, according to the ramping function, but are constantly separated in phase by *θ*, and which sum to produce vector *c*(*t*).

From the well known cosine-rule, the magnitude of *c*(*t*) is given by

$$|c(t)| = \sqrt{|a(t)|^2 + |b(t)|^2 - 2|a(t)||b(t)|\cos C} \tag{5}$$

where $C = \pi - \theta$ radians.

Typically, a hanning window is used within a phase vocoder implementation, therefore, if the magnitude of the original sinusoid is normalized to one, $|a(t)|$ is given by

$$|a(t)| = 0.5(\cos(\pi t / L) + 1) \tag{6}$$

where L is the duration of the overlap and $0 \leq t \leq L$.

The sum of $|b(t)|$ and $|a(t)|$ must be one for perfect reconstruction, therefore

$$|b(t)| = 1\text{-}|a(t)| \tag{7}$$

To determine the maximum variation in $|c(t)|$ the derivative of $|c(t)|$ with respect to t is found, then set to zero and solved for t. It can be shown that when

$$\frac{d|c(t)|}{dt} = 0 \tag{8}$$

t = L/2 provides the only non trivial solution. Therefore, the maximum amplitude variation is given by

$$1-\left|c(L/2)\right| = 1-\sqrt{0.5^2+0.5^2-2(0.5)(0.5)\cos C} \tag{9a}$$

$$= 1-\sqrt{0.5+0.5\cos\theta} \tag{9b}$$

since the magnitude of the original sinusoid has been normalized to one, $C = \pi - \theta$ radians and $|a(L/2)| = 0.5$.

From [16], the human auditory system is insensitive to amplitude variations of tones, within degrees of modulation that are less than 2% for tones that are less than 80dB. It is important to note that the total variation in amplitude from a maximum to a minimum is twice the degree of modulation. This value varies significantly with pressure levels, for example for a pure tone of pressure level 40dB the degree of modulation increases to 4% while at 100dB it decreases to 1%. These values are independent of the frequency of the tone. From [16], these values are dependent on the frequency of modulation, but the values given above are based on the modulating frequency at which human hearing is most sensitive. Also, for white noise the degree of modulation tolerated is 4% for pressure levels greater than 30dB. It can be shown that the amplitude modulation of $c(t)$ is quasi-sinusoidal in nature, with the degree of modulation, D_m, given by, from equation (9b)

$$D_m = \left(1-\sqrt{0.5+0.5\cos\theta}\right)/2 \tag{10}$$

where the divisor of 2 is required since the degree of modulation is half the total variation in amplitude.

By making the assumption that maximum pressure levels of tonal components of the signals being analysed are below 80dB, the degree of modulation of $|c(t)|$ must then be kept below 2%. So, from equation (10)

$$\left(1-\sqrt{0.5+0.5\cos\theta}\right)/2 \le 0.02 \text{ radians} \tag{11}$$

Therefore

$$\theta \le 0.5676 \text{ radians} \tag{12}$$

to ensure no perceivable amplitude modulations are introduced.

It should be noted that the amplitude modulation introduced results in an average decrease in signal amplitude level, however, the decrease is within the just noticeable amplitude level difference, as given in [16], if equation (12) is satisfied.

$B(t)$, see Figure 3, represents the time-varying phase variation between $a(t)$ and $c(t)$ and, from the well known sine-rule, is given by

$$B(t)=\sin^{-1}\left(\frac{|b(t)|\sin C}{|c(t)|}\right) \tag{13}$$

then

$$\frac{dB(t)}{dt}=\frac{\sin C\left(|c(t)|\dfrac{d|b(t)|}{dt}-|b(t)|\dfrac{d|c(t)|}{dt}\right)}{|c(t)|^2\cos B(t)} \tag{14}$$

The frequency f_c of the quasi-sinusoidal component $c(t)$ is given by

$$f_c = f_a+\frac{dB(t)}{dt} \text{ rads/second} \tag{15}$$

where f_a is the frequency of the sinusoidal component $a(t)$.

Since f_a is constant, the derivative of the $B(t)$ with respect to t represents the frequency modulating component of f_c. The maximum frequency modulation is determined by first finding the derivative of f_c with respect to t, setting it to zero and solving for t. Then

$$\frac{df_c}{dt} = \frac{d^2 B(t)}{dt^2} \tag{16}$$

and when (16) is set to zero it can, once again, be shown that $t = L/2$ provides the only non trivial solution. Therefore, it can be shown that the maximum frequency deviation is given by

$$\frac{dB(L/2)}{dt} = \frac{\pi}{L}\tan\left(\frac{\theta}{2}\right) \tag{17}$$

Also from [16], the human ear is insensitive to frequency variations introduced by frequency modulation; for tones greater than 500Hz, modulations less than 0.7% are not perceived and for tones less than 500Hz, a fixed modulation of 3.6Hz is tolerated. Once again, these values are dependent on the frequency of modulation, however the values given above are based on the modulating frequency at which the human ear is most sensitive. Therefore, in order to ensure the ear does not perceive distortion for any frequency, the variation of f_c must be kept below 3.6Hz or 22.62 radians/second. So, from (17) and setting L = 30ms.

$$\frac{\pi}{.03}\tan\left(\frac{\theta}{2}\right) \leq 22.62 \quad \text{radians} \tag{18}$$

Then

$$\theta \leq 0.4255 \ \text{ radians} \tag{19}$$

From (12) and (19) the maximum phase deviation, Ψ_{max}, that can be introduced without introducing audible modulations is

$$\Psi_{max} = 0.4255 \text{ radians} \tag{20}$$

This value only strictly applies to frequencies less than 500Hz, if the dependence of modulations on frequency is considered then Ψ_{max} could be increased to 0.5676 radians for frequencies greater than

$$\left(\frac{\pi}{.03(2\pi)}\tan\left(\frac{0.5676}{2}\right)\right)\frac{100}{0.7} = 694.46\text{Hz} \tag{21}$$

and varied accordingly between 0.4255 and 0.5767 radians for all other frequencies.

In general the maximum phase deviation tolerated θ for a 50% analysis window overlap is given by

$$\theta = \min\{0.5676,\ 2\arctan(3.6W_L)\} \text{ radians} \tag{22}$$

where W_L is the duration of the analysis window in seconds.

The above analysis is carried out based on a single pure sinusoidal tone. However, most audio signals of interest are a sum of quasi-sinusoidal components. This feature is exploited by sinusoidal modeling techniques [13] and is also the underlying assumption of the phase vocoder. It is assumed that the sum of sinusoids that have been amplitude and frequency modulated to the maximum limit, such that they are perceptually equivalent to the original individual sinusoids, results in a signal that is perceptually equivalent to the sum of the non-modulated sinusoids. Informal listening tests in a quiet office environment support this assumption.

The above analysis is also based on an ‘ideal’ horizontal phase shift i.e. vertical phase coherence is maintained. Such a phase shift is straightforward to achieve with synthesized pure sinusoids but is difficult with real audio signals; this difficulty is, of course, the reason for the existence of the phasiness artifact in the first place. However, the above analysis does suggest that a certain amount of flexibility exists in the choice of phase in order to maintain horizontal phase coherence of dominant sinusoidal components. This is further supported by the fact that phase vocoder implementations are capable of producing high quality time-scale modifications even though frequency estimates, used in [3] to determine synthesis phases, are prone to inaccuracies [17], [18].

The derivation of amplitude and frequency modulations introduced due to phase deviation is based on a hop size of half the analysis window length. A similar, albeit more tedious, approach can be used to determine modulations introduced for the case of different hop sizes; a hop size of half the analysis window length is used in this section for its intuitive appeal and mathematical simplicity. The workings for the derivation of equivalent equations for the commonly used 75% overlap are somewhat verbose and can be determined in a similar manner to the methodology outlined above; full details can be found in [19]. For the sake of convenience the equations derived for a 75% overlap are provided here. The maximum phase deviation tolerated θ is given by

$$\theta = \min\{0.27,\ 2\arcsin(2.53W_L)\}\ \text{radians} \qquad (23)$$

It should be noted that (23) is an approximation, valid within 0.2% for values of θ less than 0.27 radians [19].

3.1 Use of Phase Flexibility to Achieve Phasiness Reduction

The phase tolerance established above can be used to push or pull a modified STFT representation into a phase coherent state; the basic principle is briefly explained as follows:

Consider the situation illustrated in Figure 4; assume that the phases of synthesis window 1' are equal to those of analysis window 1; the phases of the repeated synthesis window 2' are then determined such that horizontal phase coherence is maintained between true sinusoidal components (peaks), whilst phases of neighboring components are updated so as to maintain vertical phase coherence. Horizontal phase coherence between the peaks of synthesis windows 1' and 2' can be preserved by keeping the same phase difference between them that exists between analysis windows 1 and 2 [12]; then synthesis window 1' comprises of the magnitudes and phases of analysis window 1 (and is therefore perfectly phase coherent), whilst synthesis window 2' comprises of the magnitudes of analysis window 1 and a set of phases close to those of analysis window 2 (and is therefore generally not perfectly phase coherent). It follows that, in general, synthesis window *n*' comprises of the magnitudes of analysis window *n*-1 and phases close to those of analysis window *n*, for all windows up to the next discard/repeat frame.

In [9] the synthesis phase values of synthesis window *n*' are pushed or pulled toward the phase values of analysis window *n*-1 using the horizontal phase tolerance established. Once the phases of window *n*' equal those of the target phases of analysis window *n*-1 perfect phase coherence is restored. It follows that subsequent windows up to the next discard/repeat window will also be perfectly phase coherent. From Figure 3, once phase coherence is realized (at synthesis window 7' in Figure 4), there is no need for further frequency-domain processing and a segment of the original time-domain input can be simply inserted into the output, in a similar manner to time-domain implementations, as shown in Figure 4. This has the added benefit of reducing the computational costs whilst bringing the time-scaled output into a phase coherent state.

This process requires that a certain number of windows exist before the next discard/repeat operation; for example given a phase tolerance of 0.314 (i.e. π/10) radians, perfect phase coherence is assured to be established for time-scale factors between 0.9 and 1.1, since phase values can be at most +/-π radians from perfect phase coherence. It should be noted that if the phase values of synthesis window 2' were close to those of analysis window 1 then perfect phase coherence would be established quickly; the following section addresses this issue by making use of time-domain techniques in identifying ‘good’ initial phase values, thereby reducing the transition time to perfect phase coherence.

4 HYBRID IMPLEMENTATION

The original motivation behind the SOLA algorithm [2] was to provide a good initial set of phase estimates which would reduce the number of iterations required for the reconstruction of a magnitude only STFT representation of a signal [21]. The same principle is used here to provide a good initial set of phase estimates for use within the procedure outlined in section 3.1, so that perfect phase coherence can be recovered quickly. The remainder of this section describes the approach used to determine the initial phase estimates and their use within the hybrid implementation.

It should be noted that in the following discussion the term window is reserved for the STFT windows used in the approach (which are of fixed duration), whilst the term frame refers to the variable length (α dependent) segment associated with time-domain implementations.

Consider the situation shown in Figure 5, in which a frame extracted from the input is shown overlapping with the current output. As with the standard SOLA implementation the overlap shown is determined through the use of a correlation function. For the m^{th} iteration of the algorithm the offset τ_m is chosen such that the correlation function $R_m(\tau)$, given by equation (1), is a maximum for $\tau = \tau_m$. The optimum frame overlap L_{ov} shown in Figure 4 is then given by

$$L_{ov} = N\text{-}\ S_s - \tau_m \tag{24}$$

Also shown in Figure 5, below the input frame, are the synthesis windows and the synthesis frame; it is this synthesis frame which is appended to the current output within the hybrid approach and not the input frame, as is the case in SOLA, see section 1. The following details the generation of the synthesis frame.

Window b is first extracted from the output y and is positioned such that it has its center at the center of the 'optimum' overlap, as shown in the diagram. More specifically, for the m^{th} iteration of the algorithm, frame b is given by

$b(j) = y(mS_s + \tau_m\ + L_{ov}/2 - L/2 + j).w(j)$ for $0 < j \leq L$ (25)

where w is the STFT analysis window, typically hanning, L is the STFT window length, typically the number of samples which equates to approximately 60ms. Both shorter and longer windows have been proposed in the literature, however 60ms was found to be suitable for an implementation which is intended to cater for both speech and a wide range of polyphonic music.

The window f_1 is extracted from the input x and is positioned such that it is aligned with frame b. Subsequent windows are sequentially spaced by the STFT hop size H. More specifically, for the m^{th} iteration of the algorithm window f_n is given by

$$f_n(j) = x(mS_a + L_{ov}/2 + H.(n-1) - L/2 + j).w(j) \text{ for } 0<j\leq L \tag{26}$$

$F_1^{'}$ the DFT representation of $f_1^{'}$, is then derived using the magnitudes of F_1 and the phase values B, where F_n and B are the DFT representations of f_n and b, respectively; then

$$F_1^{'}(k) = |F_1(k)|\exp(i\angle B(k)) \text{ for all } k \text{ in the set } P_1 \tag{27}$$

where P_1 is the set of peak bins found in $|F_1|$. All other bins are updated so as to maintain the original phase difference between a peak and bins in its region of influence, as described in [8]. The phase values of STFT window B are chosen since they provide a set of phase values that naturally follow the window labeled a in Figure 5 and therefore maintain horizontal phase coherence. Subsequent synthesis windows are derived from

$$F_n^{'}(k) = |\angle F_n(k)|\exp\left(i\left(\angle F_{n-1}^{'}(k) + \angle F_n(k) - \angle F_{n-1}(k) + D(k)\right)\right) \tag{28}$$

for all k in the set P_n, where P_n is the set of peak bins found in $|F_n|$. As above, all other bins are updated so as to maintain the original phase difference between a peak and bins in its region of influence. For the hybrid case perfect phase coherence is achieved when synthesis STFT window $F_n^{'}$ has the magnitude and phase values of window F_n. D is the phase deviation which is used to push or pull the frames into a phase coherent state. D is dependent on the bin number denoted by k and is given by

$$D(k) = \angle F_{n-1}(k) - \angle F_{n-1}^{'}(k) \quad \text{if princarg } \left(\angle F_{n-1}(k) - \angle F_{n-1}^{'}(k)\right) \leq \theta \tag{39}$$

or

$$D(k) = sign\left(\angle F_{n-1}(k) - \angle F'_{n-1}(k)\right)\theta \quad \text{if princarg } \left(\angle F_{n-1}(k) - \angle F'_{n-1}(k)\right) > \theta \tag{30}$$

where θ is the maximum phase tolerance (see section 3).

The number of synthesis STFT windows required is such that an inverse STFT on these windows results in a synthesis frame of duration N+3L/2. This is to ensure that window b is available for the next iteration of the algorithm. It should be noted that the number of the synthesis windows also controls the ability of the algorithm to recover phase coherence; if N is large (which is the case when is α is close to one, see equation (3)) phase coherence is recovered more easily. The synthesis frame x_m is obtained through the application of an inverse STFT on windows F_1', F_2', F_3',.... The output y is then updated by

$$y(mS_s + \tau_m + L_{ov}/2 - L/2 + j) := E(j).y(mS_s + \tau_m + L_{ov}/2 - L/2 + j) + x_m(j) \text{ for } 0 < j \leq L–H \tag{31}$$

$$y(mS_s + \tau_m + L_{ov}/2 - L/2 + j) = x_m(j) \text{ for } L\text{-}H < j \leq N + 3L/2 \tag{32}$$

where := in equation (31) means ‘becomes equal to’ and E is an envelope function which ensures that the output y sums to a constant during the overlap-add procedure.

E is dependent on the STFT hop size H and whether a synthesis window is employed during the inverse STFT procedure. For the case where a synthesis window is employed, and which is equal to the analysis hanning window w, and $H = L/4$

$$E(j) = w^2(H + j) + w^2\,(2H+j) + w^2\,(3H+j) \text{ for } 0<j\leq L–H \tag{33}$$

It should be noted that for the case where the input is perfectly periodic the initial phase estimates provided by STFT window B are assured to be equal to the target phase values of window F_1 and the time-scaled output is always perfectly phase coherent. For quasi-periodic signals, such as speech, the initial phase estimates are generally close to the target phase, and the transition period to perfect phase coherence is generally short.

For the case where more complex audio is being time-scaled, the transition to perfect phase coherence is relatively long, in comparison with speech signals; nevertheless, the reverberant artifact introduced, due to the loss of perfect phase coherence, is perceptually less objectionable in these types of signals, due to the reverberation level generally already present in polyphonic music.

As with time-domain implementations, the quality and efficiency improvements offered by the hybrid approach over frequency-domain approaches are most noticeable for time-scaling factors close to one, with results being particularly good for factors in the range 0.8 to 1.25, see section 6.

Figure 7 illustrates the computational advantage of the phasiness reduction technique; the vertical axis shows the ratio of computations of the standard phase vocoder to the computations of the phase vocoder that utilises the phasiness reduction technique presented in this section. The solid line is plotted for $\theta = 0.4255$ radians and the dashed line is plotted for $\theta = 0.27$ radians, which correspond to the maximum phase deviation allowed for a 50% and 75% overlap given a STFT frame length of 60ms.

5 MULTI-CHANNEL CONSIDERATIONS

In [12] the implications of the application of a phase vocoder based time-scale modification algorithm to stereo recordings are outlined. [12] maintains the stereo image by ensuring that both magnitude and phase differences between related channel components are preserved. Magnitude differences are maintained within standard phase vocoder implementations if the same parameters are used to time-scale each channel, whilst in [12] phase differences are explicitly maintained.

Within the hybrid implementation, segments of different duration could be discarded/repeated from each channel if the channels are time-scaled separately [20]; even if the same algorithm parameters are applied to each channel. This could result in an alteration of the stereo image, since magnitude differences between channels are unlikely to be maintained. The solution to this potential problem is to sum channels before applying the correlation function of equation (1). The offset identified, by finding the maximum of the correlation function, is then applied to both channels for each iteration of the algorithm.

Phase differences are preserved between peaks, at the same bin location, between channels, by first updating the peak with the greater magnitude in the manner described earlier; the peak with the lesser magnitude is updated so as to preserve the original phase relationship. Bins in the region of influence of a peak are updated in the usual manner, i.e. by keeping the same phase difference between each bin and its associated peak that existed during STFT analysis, as described in [8].

6 SUBJECTIVE OUTPUT QUALITY COMPARISON

Fourteen test subjects undertook eight subjective listening tests to compare the quality of time-scaled audio produced by the hybrid algorithm against a phase vocoder implementation. For each test the test subjects were presented with three files; labeled track1, track2 and original. The files labeled track1 and track2 were the time-scaled tracks, and the original was provided for reference. The test subjects were not aware which track was time-scaled by which algorithm and the labeling of tracks was randomised. Test subjects were allowed playback the tracks as often as they wished, and in any order they wished. Test subjects were asked to indicate which track sounded most like the original by selecting one of five options; track1 much better than track2, track1 slightly better than track2, track1 equal to track2, track1 slightly worse than track2, or track1 much worse than track2.

For all tests a 60ms STFT analysis and synthesis window with a 75% overlap were employed, whilst the hybrid algorithm used a search range, *SR*, equal to 20 ms, and the stationary length, L_{stat}, was set to 30 ms.

In a first set of tests, tests were limited to relatively small time-scale factors, in the range of 0.8 – 1.25, and applied only to speech. A summary of the test results are given in Table 1.

Test subjects indication	**% of total**
Hybrid much better than phase vocoder	42.0%
Hybrid slightly better than phase vocoder	44.6 %
Hybrid equal to phase vocoder.	8.0 %
Hybrid slightly worse than phase vocoder.	5.4 %
Hybrid much worse than phase vocoder.	0.0 %

Table 1. Summary of listening test results comparing the use of the hybrid approach against a phase vocoder approach for the time-scale modification of speech for factors in the range 0.8-1.25.

Results of the first set of tests indicate a strong preference for the hybrid implementation. There is also an indication that the improvements are more noticeable within male speech; when only male speakers are considered 58.9% of test subjects results indicate that the hybrid approach is much better than the phase vocoder; this figure is only 25% when only female speakers are considered. One explanation for this is that the harmonic components of female speech are separated by a greater distance in frequency than in male speech, and since bins in the region of influence of a peak are divided evenly between harmonic components, there will be significantly more phase locking between peaks and nearby bins in female speech. It is also the case that there will generally be fewer dominant sinusoidal components in female speech than male and, therefore, that phase coherence will play a more important role in male speech. This explanation is in keeping with the authors' finding that improvements are also more noticeable within gravelly or rough speech; since, in this type of speech a number of additional sub harmonic components are typically present, as found in [22]. From [22] it is also shown that the phase values associated with the sub harmonic components do not adhere to the sinusoidal phase propagation formula, which is used within the standard phase vocoder algorithm; it therefore follows that the standard phase vocoder will produce erroneous results when applied to these sub harmonic components, whereas the hybrid approach will cater for these types of signal more readily, once the search range employed in determining the correlation function is of sufficient duration to encompass the 'growl macro period' [22].

In a second set of subjective listening tests, using the same algorithm parameters and format as outlined above, subjects were requested to compare the quality of time-scaled speech produced by the hybrid algorithm against a phase vocoder implementation for time-scale factors in the range 0.6 to 0.8 and 1.25 to 1.75. A summary of the test results are given in Table 2.

As for the case of smaller time-scale factors, there is a preference for the hybrid approach over the standard phase vocoder implementation; however, results suggest that the preference is less significant. This finding is in keeping with expectations, since the hybrid implementation is more likely to recover 'perfect phase coherence' for time-scale factors close to one.

Test subjects indication	**% of total**
Hybrid much better than phase vocoder	23.7%
Hybrid slightly better than phase vocoder	41.3 %
Hybrid equal to phase vocoder.	30.0 %
Hybrid slightly worse than phase vocoder.	5.0 %
Hybrid much worse than phase vocoder.	0.0 %

Table 2. Summary of listening test results comparing the use of the hybrid approach against a phase vocoder approach for the time-scale modification of speech for factors in the range 0.6-0.8 and 1.25-1.75.

In a final set of subjective listening tests, using the same algorithm parameters and format as outlined above, subjects were requested to compare the quality of time-scaled music produced by the hybrid algorithm against a phase vocoder implementation. Time-scale factors in the range 0.6 to 1.75 were employed. A summary of the test results are given in Table 3.

Test subjects indication	**% of total**

Hybrid much better than phase vocoder	7.5%
Hybrid slightly better than phase vocoder	25.0 %
Hybrid equal to phase vocoder.	42.5 %
Hybrid slightly worse than phase vocoder.	20.0 %
Hybrid much worse than phase vocoder.	5.0 %

Table 3. Summary of listening test results comparing the use of the hybrid approach against a phase vocoder approach for the time-scale modification of music for factors in the range 0.6-1.75.

Results of the subjective test indicate no significant preference for either approach; this is attributed to the fact that there is generally a significant level of reverberation present in music, and that relatively small reduction, or introduction, of reverberation will be difficult to perceive.

It should be noted that the quality of output produced by the hybrid approach, when applied to speech, is slightly inferior to time-domain implementations. However, the hybrid approach has the advantage of being capable of producing high quality time-scale modifications when applied to complex polyphonic music. Future work involves relaxing the phase deviation value during quasi-periodic regions of the signal being time-scaled and/or between harmonically related components in order to further improve vertical phase coherence between these related components. Future work also includes incorporating aspects of the approach described in [6] in order to identify ‘better’ splice locations and, therefore, further reduce, or remove, the transition to perfect phase coherence.

7 CONCLUSION

A hybrid approach to audio time-scale modification is presented. The hybrid technique draws upon the best features of time-domain and frequency-domain implementations and reduces the presence of the reverberant artifact associated with frequency-domain techniques, without the requirement of explicit pitch detection. The technique makes use of a derived amount of phase flexibility present within the phase vocoder to gradually push or pull modified STFT phase components into a phase coherent state. The technique also makes use of SOLA based technique to provide a good initial set of phase estimates, and therefore reduce the transition time to perfect phase coherence.

From subjective testing, the improvement in quality is significant in speech, while no significant improvements are perceived for music recordings; This is attributed to the relatively high levels of reverberation present in polyphonic music.

The algorithm is both robust and efficient and produces high quality results for both speech and a wide range of polyphonic audio. These attributes make it particularly suitable for the time-scale modification of general audio where no prior knowledge of the input signal exists; for example, during the time-scale modification of movies or television/radio adverts, in which both speech and/or music are typically present. It should be noted that the hybrid algorithm does not attempt to resolve the transient smearing problem associated with phase vocoder implementations and that transient handling techniques, such as [23], could be employed to reduce smearing effects.

The algorithm could also be used in conjunction with the approach outlined in [6] to extend its high quality operating range beyond +-15%.

As time-scaling techniques have developed there has been a continual merging between various techniques. For example, the improved phase vocoder draws upon sinusoidal modeling based peak picking to reduce phasiness, and the motivation behind the SOLA algorithm was to provide initial estimates for the iterative STFT reconstruction algorithm of [21]. This paper follows this trend by merging aspects of time-domain techniques with improved phase vocoder techniques. The result is an algorithm which, in effect, pulls together the most appealing aspects of a number of originally isolated techniques i.e. sinusoidal modeling, phase vocoder, time-domain and iterative phase reconstruction techniques.

8 ACKNOWLEDGMENTS

The authors wish to thank Dan Barry for his many fruitful discussions throughout the course of this research.

9 REFERENCES

[1] M. Karhs, K. Brandenburg, Applications of Digital Signal Processing to Audio and Acoustics, Kluwer Academic Publishers, pp. 279-308, 1998.

[2] S. Roucos, A.M. Wilgus, "High quality time-scale modification for speech," IEEE International Conference on Acoustics, Speech and Signal processing, pp. 493-496, 1985.

[3] M. Dolson, "The phase vocoder: A tutorial," Computer Music Journal, vol. 10, pp. 145-27, 1986.

[4] R. McAulay, T Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 34(4), pp.744–754, 1986.

[5] G. Pallone, P. Boussard, L Daudet, P. Guillemain, R. Kronland-Martinet, “A wavelet based method for audio-video synchronization in broadcasting applications”, Proceedings of the International Conference on Digital Audio Effects, Norway, December 1999.

[6] B.G. Crockett, “High quality multi-channel time-scaling and pitch-shifting using auditory scene analysis”, Audio Engineering Society Convention, preprint no. 5948, New York, October 2003.

[7] D. Dorran, R. Lawlor, E. Coyle, “Audio time-scale modification using a hybrid time-frequency domain approach”, Accepted for publication in the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2005.

[8] J. Laroche, M. Dolson, “Improved phase vocoder time-scale modification of audio,” IEEE Transactions on Speech and Audio Processing, vol. 7(3), pp. 323-332, 1999.

[9] D. Dorran, R. Lawlor, E. Coyle, “An efficient phasiness reduction technique for moderate audio time-scale modification,” Proceedings of the International Conference on Digital Audio Effects, pp. 83-88, 2004.

[10] J. Laroche, “Autocorrelation method for high-quality time/pitch-scaling”, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 131 – 134, 1993.

[11] D. Dorran, R. Lawlor, “An efficient time-scale modification algorithm for use within a subband implementation,” Proceedings of the International Conference on Digital Audio Effects, pp. 339-343, 2003.

[12] J. Bonada, “Automatic technique in frequency domain for near-lossless time-scale modification of audio,” Proceedings of International Computer Music Conference, 2000.

[13] T. Quatieri, R. McAulay, “Shape invariant time-scale and pitch-scale modification of speech,” IEEE Transactions on Signal Processing, vol. 40(3), pp 497-510, 1992.

[14] R. Di Federico, “Waveform preserving time stretching and pitch shifting for sinusoidal models of sound,” Proceedings of the International Conference on Digital Audio Effects, pp. 44-48, 1998.

[15] J. Laroche, “Frequency-domain techniques for high quality voice modification,” Proceedings of the International Conference on Digital Audio Effects, pp. 328-322, 2003.

[16] E. Zwicker, H. Fastl, Psychoacoustics: Facts and Models, Springer Verlag, second edition, May 1999.

[17] M.S. Puckette and J.C. Browne, “Accuracy of frequency estimates using the phase vocoder,” IEEE Transactions on Speech and Audio Processing, vol. 6 issue 2, pp. 166-176, March 1998.

[18] S.S. Abeysekera, K.P. Padhi, J. Absar and S. George, “Investigation of different frequency estimation techniques using the phase vocoder,” IEEE International Symposium on Circuits and Systems, vol. 2, pp. 265-268, May 2001.

[19] D. Dorran, “Audio Time-Scale Modification”, PhD Dissertation, Dublin Institute of Technology, Dublin, Ireland, 2005.

[20] D. Dorran, R. Lawlor, E. Coyle, “Multichannel Audio Time-Scale Modification”, AES Convention, preprint no. 6527, New York, October 2005.

[21] D. W. Griffen, J. S. Lim, “Signal estimation from modified short-time Fourier transform”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-32(2), pp.236-243, April 1984.

[22] A. Loscos, J. Bonada, “Emulating rough and growl voice in spectral domain”, Proceedings of the International Conference on Digital Audio Effects, pp. 49-52, October 2004.

[23] C. Duxbury, M. Davies, M. Sandler, “Temporal segmentation and pre-analysis for non-linear time-scaling of audio”, 114th Convention of the Audio Engineering Society, Preprint no. 5812, Amsterdam, April 2003.

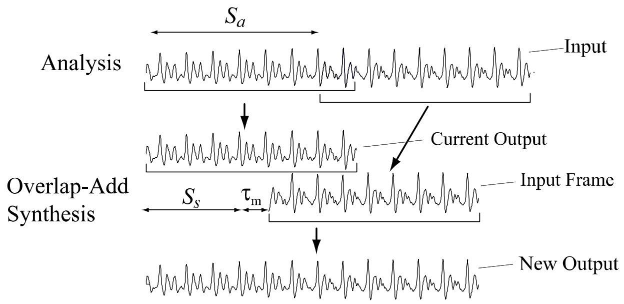

Figure 1: *SOLA iteration for speed up.*

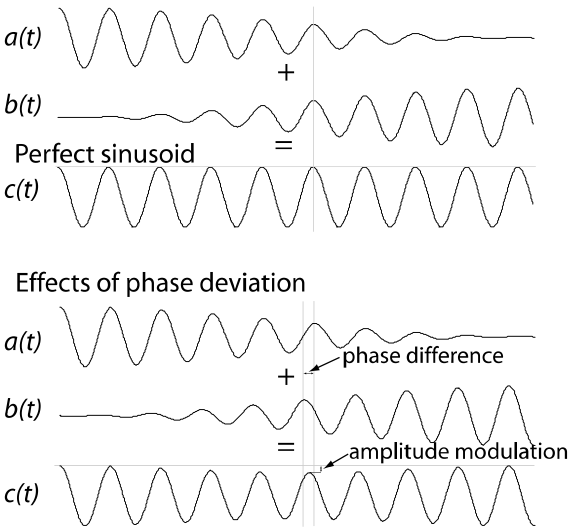

Figure 2: The effects of phase deviation on a pure sinusoid within a phase vocoder implementation.

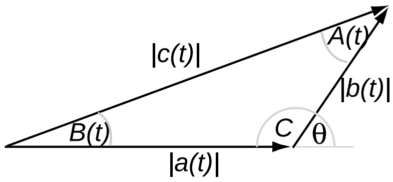

Figure 3 : Vector representation of figure 2

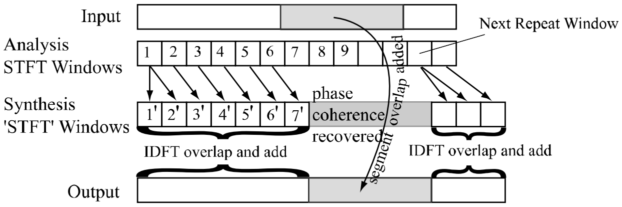

Figure 4: Hybrid time-scaling process

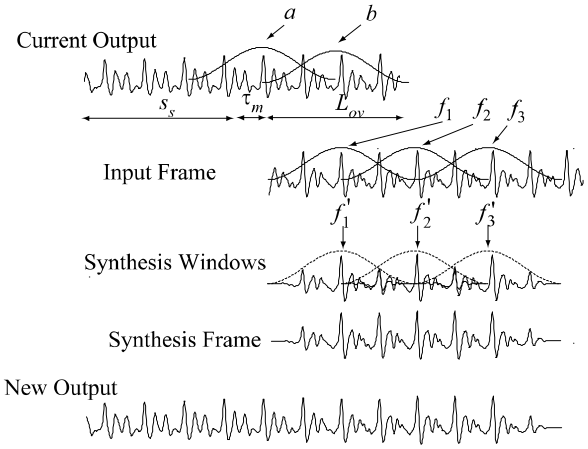

Figure 5: Hybrid iteration

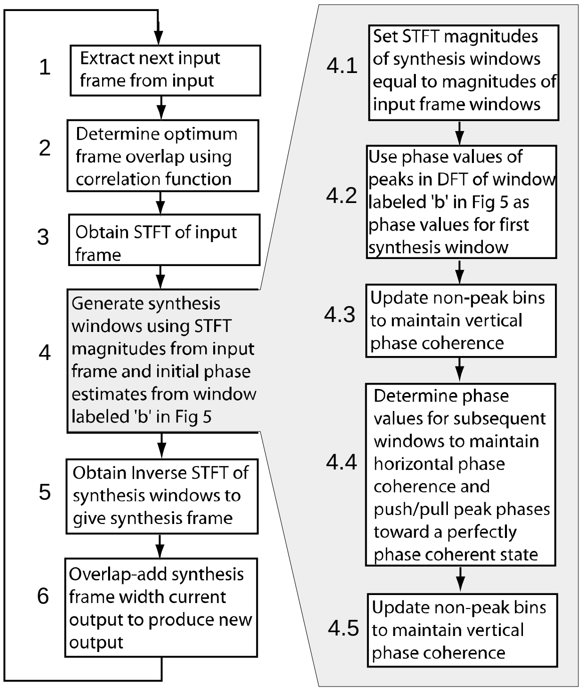
1
Extract next input frame from input
2
Determine optimum frame overlap using correlation function
3
Obtain STFT of input frame
4
Generate synthesis windows using STFT magnitudes from input frame and initial phase estimates from window labeled 'b' in Fig 5
5
Obtain Inverse STFT of synthesis windows to give synthesis frame
6
Overlap-add synthesis frame width current output to produce new output
4.1
Set STFT magnitudes of synthesis windows equal to magnitudes of input frame windows
4.2
Use phase values of peaks in DFT of window labeled 'b' in Fig 5 as phase values for first synthesis window
4.3
Update non-peak bins to maintain vertical phase coherence
4.4
Determine phase values for subsequent windows to maintain horizontal phase coherence and push/pull peak phases toward a perfectly phase coherent state
4.5
Update non-peak bins to maintain vertical phase coherence

Figure 6. Flow diagram outlining the hybrid time-scaling process.

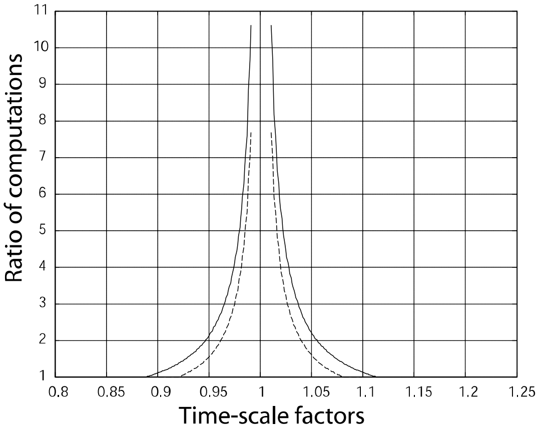

Figure 7. Ratio of computations required for the improved phase vocoder approach to the number of computations required using the hybrid approach.