2004-01-01

An Efficient Phasiness Reduction Technique for Moderate Audio Time-scale Modification

David Dorran
Technological University Dublin, david.dorran@tudublin.ie

Eugene Coyle
Technological University Dublin, Eugene.Coyle@tudublin.ie

Robert Lawlor
National University of Ireland, Maynooth

AN EFFICIENT PHASINESS REDUCTION TECHNIQUE FOR MODERATE AUDIO TIME-SCALE MODIFICATION

David Dorran, Eugene Coyle

Dublin Institute of Technology
Dublin, Ireland.
david.dorran@dit.ie

Robert Lawlor

National University of Ireland
Maynooth, Ireland.
rlawlor@eeng.may.ie

ABSTRACT

Phase vocoder approaches to time-scale modification of audio introduce a reverberant/phasy artifact into the time-scaled output due to a loss in phase coherence between short-time Fourier transform (STFT) bins. Recent improvements to the phase vocoder have reduced the presence of this artifact, however, it remains a problem. A method of time-scaling is presented that results in a further reduction in phasiness, for moderate time-scale factors, by taking advantage of some flexibility that exists in the choice of phase required so as to maintain horizontal phase coherence between related STFT bins. Furthermore, the approach leads to a reduction in computational load within the range of time-scaling factors for which phasiness is reduced.

1. INTRODUCTION

Time-scale modification of audio alters the duration of an audio signal while 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 the quality, pitch, timbre or naturalness of the original signal. This facility is useful for such applications as enhancement of degraded speech, language and music learning, fast playback for telephone answering machines and audio-video synchronization in broadcasting applications.

The phase vocoder is a popular method for time-scaling audio due to its ability to achieve high quality modifications on a variety of signals within a wide range of time-scaling factors. However, the phase vocoder suffers from an artifact known as phasiness that exists predominantly due to a loss of vertical phase coherence between modified short-time Fourier transform (STFT) bins, as explained in [1]. In [1] an improvement to the phase vocoder is presented that reduces the presence of the phasiness artifact by providing a more accurate estimate of the phase of STFT components in the neighborhood of STFT peaks. However, the artifact remains audible and is particularly objectionable in speech.

This paper presents a technique that offers a further reduction in the phasiness artifact for moderate time-scaling, in the range of ± 10%. The approach takes advantage of a certain amount of flexibility that exists in the choice of phase for modified, time-scaled, STFT bins to achieve horizontal phase coherence, and uses this flexibility to improve upon vertical phase coherence, thus reducing the phasiness effect. Section 2 outlines the operation of a phase vocoder implementation that has the same analysis and synthesis STFT hop size, as used in [2]. Section 3 presents an analysis of horizontal phase coherence under 'ideal' conditions, which is then used to determine the amount of flexibility in the phase used so as to maintain horizontal phase coherence. Section 4 demonstrates how the flexibility in the choice of phase can be used to improve vertical phase coherence and outlines the computational benefits associated with the technique. Section 5 discusses the limitations of the approach and the results of informal listening tests. Section 6 concludes this paper.

2. THE PHASE VOCODER

The phase vocoder was first described in [3], with an efficient STFT implementation given in [4]. A tutorial article in [5] provides an excellent insight into the fundamental operation of the phase vocoder and [6] presents some detail of a MATLAB based implementation. The concept and problems of vertical phase coherence are described in detail in [1] and a mathematical description is also provided. In the rest of this section we briefly outline the phase vocoder and how it can achieve time-scale modification, using the same analysis and synthesis STFT hop size, as used in [2].

The first step is to obtain an STFT representation, $X(t_u,\Omega_k)$, of the input, as given in [1]

$$X(t_u,\Omega_k)=\sum_{n=-\infty}^{\infty}h(n)x(t_u+n)e^{-j\Omega_k n} \quad (1)$$

where x is the input signal, $h(n)$ is the analysis window, Ω_k is the center frequency of the k^{th} vocoder channel and t_u is the u^{th} analysis time instant and $t_u = uR$, where R is the analysis (and synthesis) hop size and u is a set of successive integer values, starting at 0.

In [2] time-scale expansion is achieved by appropriately repeating STFT frames e.g. to time-scale by a factor of 1.5 every second frame is repeated, as illustrated in figure 1; similarly time-scale compression is achieved by omitting frames e.g. to time scale by a factor of 0.9 every tenth analysis frame is omitted. Like traditional implementations of the phase vocoder, the magnitudes of the modified, time-scaled, STFT remains unaltered i.e.

$$|Y(t_m,\Omega_k)|=|X(t_n,\Omega_k)| \text{ for all } k \quad (2)$$

where n = round(m/α), m is a set of successive integer values starting at 0, t_n and t_m are a set of analysis and synthesis time instants, respectively.

The phases of the modified STFT, $\angle Y(t_m,\Omega_k)$, are determined so as to maintain both horizontal and vertical phase coherence. To achieve phase coherence, first the peaks, representing the dominant components of each frame are detected. In [1] a peak is defined as

any bin whose magnitude is greater than its four nearest neighbours. In the simplest, most efficient, implementation phases of peaks are updated by maintaining the same phase difference between consecutive synthesis frames that exists between corresponding analysis frames i.e.

$$\angle Y(t_m,\Omega_{k_p}) - \angle Y(t_{m-1},\Omega_{k_p}) = \angle X(t_n,\Omega_{k_p}) - \angle X(t_{n-1},\Omega_{k_p}) \text{ for all } k_p \quad (3)$$

which becomes

$$\angle Y(t_m,\Omega_{k_p}) = \angle Y(t_{m-1},\Omega_{k_p}) + \angle X(t_n,\Omega_{k_p}) - \angle X(t_{n-1},\Omega_{k_p}) \text{ for all } k_p \quad (4)$$

where k_p are the bins of the detected peaks.

Having determined the phases of the synthesis peaks, the phases of bins in each peak's region of influence are updated by maintaining the same phase difference between peaks and the bins in their region of influence that exists in the mapped analysis frame. In [1] the upper limit of the region of influence of a peak is set to the middle frequency between that peak and the next one. Then

$$\angle Y(t_m,\Omega_k) = \angle Y(t_m,\Omega_{k_p}) + \angle X(t_n,\Omega_k) - \angle X(t_n,\Omega_{k_p}) \quad (5)$$

for all k in each peak's region of influence.

A better method for updating phases requires sinusoidal modeling based peak tracking, as explained in [1], however, no advantage was found in using a peak tracking approach when employing the phasiness reduction techniques, described later in section 4, in the range of time-scale factors for which the techniques offer a significant improvement i.e. 0.9-1.1.

A time-scaled version of the original signal is obtained by calculating the inverse STFT of $Y(t_m,\Omega_k)$.

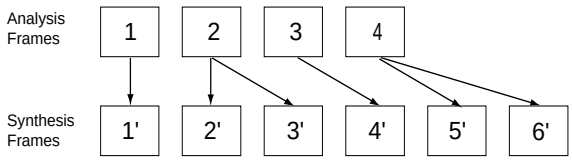

Figure 1 : *Analysis to synthesis frame mapping*

3. FLEXIBILITY OF HORIZONTAL PHASE COHERENCE

The inverse STFT of a given STFT is found by calculating the inverse discrete Fourier transform (IDFT) of each STFT frame. Successive inverse STFT frames are then overlapped and added together to produce the time-domain signal. A single iteration of the overlap and add process is illustrated in the upper three waveforms of figure 2, where two frames of a sinusoidal signal are overlapped and summed together to reproduce a perfect sinusoid. Now consider the case where the overlapping frames are no longer perfectly sychronised i.e. they are slightly out of 'horizontal' phase, as illustrated by the lower three waveforms of figure 2. When the 'out of horizontal phase' sinusoids are summed together the resulting signal is no longer a perfect sinusoid but is a quasi-sinusoidal signal modulated in both amplitude and frequency. As expected intuitively, the greater the relative phase difference between the sinusoidal frames the greater the modulation that is introduced. From [7], human hearing is insensitive to certain amounts of frequency and amplitude modulations, and in an effort to determine the maximum phase difference that can be introduced without introducing audible distortion a set of equations representing the situation described above is derived.

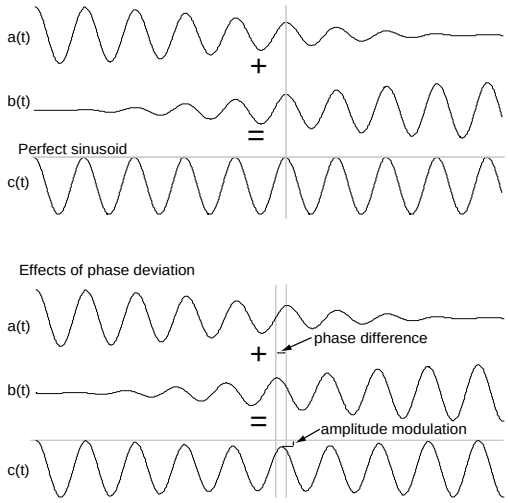

Figure 2 : *Loss of horizontal phase coherence*

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 $c(t)$.

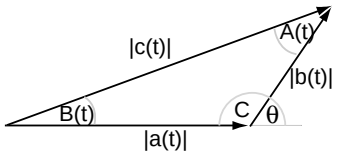

Figure 3 : *Vector representation of figure 2*

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

$$|c(t)| = \sqrt{|a(t)|^2 + |b(t)|^2 - 2|a(t)||b(t)|\cos C} \quad (6)$$

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

$$|a(t)| = 0.5(\cos(\pi t / L) + 1) \quad (7)$$

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

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

$$|b(t)| = 1 - |a(t)| \quad (8)$$

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

$$\frac{d|c(t)|}{dt} = 0 \quad (9)$$

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

$$1 - |c(L/2)| = 1 - \sqrt{0.5^2 + 0.5^2 - 2(0.5)(0.5)\cos C} \quad (10a)$$

$$= 1 - \sqrt{0.5 + 0.5\cos\theta} \quad (10b)$$

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

From [7], the human ear is insensitive to amplitude variations of tones, introduced by sinusoidal amplitude modulation, for 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. It should also be noted that, from [7], 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 (10b)

$$D_m = \left(1 - \sqrt{0.5 + 0.5\cos\theta}\right)/2 \tag{11}$$

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 (11)

$$\left(1 - \sqrt{0.5 + 0.5\cos\theta}\right)/2 \le 0.02 \text{ radians} \tag{12}$$

Therefore

$$\theta \le 0.5676 \text{ radians} \tag{13}$$

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 [7], if equation (13) is satisfied.

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

$$B(t) = \sin^{-1}\left(\frac{|b(t)|\sin C}{|c(t)|}\right) \tag{14}$$

then

$$\frac{dB(t)}{dt} = \frac{\sin C\left(|c(t)|\frac{d|b(t)|}{dt} - |b(t)|\frac{d|c(t)|}{dt}\right)}{|c(t)|^2 \cos B(t)} \tag{15}$$

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

$$f_c = f_a + \frac{dB(t)}{dt} \text{ rads/second} \tag{16}$$

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

$$\frac{df_c}{dt} = \frac{d^2B(t)}{dt^2} \tag{17}$$

and when (17) 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

$$\frac{dB(L/2)}{dt} = \frac{\pi}{L}\tan\left(\frac{\theta}{2}\right) \tag{18}$$

Also from [7], 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 equation (18) and setting L = 23.22ms, which corresponds to half the length of a 2048 point window at a sampling frequency of 44.1kHz.

$$\frac{\pi}{.02322}\tan\left(\frac{\theta}{2}\right) \le 22.62 \quad \text{radians} \tag{19}$$

Then

$$\theta \le 0.3313 \text{ radians} \tag{20}$$

From (13) and (20) the maximum phase deviation, Ψ_{max}, that can be introduced without introducing audible modulations is

$$\Psi_{max} = 0.3313 \text{ radians} \tag{21}$$

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

$$\frac{\pi}{.02322}\tan\left(\frac{0.5676}{2}\right)2\pi = 897.23\text{Hz} \tag{22}$$

and varied accordingly between 0.3313 and 0.5767 radians for all other frequencies.

The above analysis is carried out based on a single pure sinusoidal tone, however, most audio signals of interest are, for the most part, a sum of quasi-sinusoidal components, a feature exploited by sinusoidal modeling techniques [8] and is 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 easy 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 [1] to determine synthesis phases, are prone to inaccuracies [9], [10].

The derivation of amplitude and frequency modulations introduced due to phase deviation was 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. Another commonly used hop size is one quarter of the analysis frame length, for which it can be shown that $\Psi_{max} \approx 0.24$ radians for analysis window lengths of 46.44ms.

4. REDUCTION IN PHASINESS AND COMPUTATIONS

In the previous section it was shown that a certain amount of flexibility exists in the choice of phase required to achieve horizontal phase coherence within a phase vocoder implementation. This flexibility can be used to 'push' or 'pull' modified STFT frames into a phase coherent state; however a set of coherent target phases for each frame are first required. One set of target phases that would guarantee vertical phase coherence are the phases of the original frames that are mapped to each synthesis frame. So, having determined an estimate of the synthesis phases using the procedure described in section 2, the synthesis phases are updated further using the following rules:

If

$$\left|princ\_arg\left(\angle Y\left(t_m,\Omega_k\right)-\angle X\left(t_n,\Omega_k\right)\right)\right| \le \Psi_{\max} \tag{23a}$$

then

$$\angle Y\left(t_m,\Omega_k\right)=\angle X\left(t_n,\Omega_k\right) \tag{23b}$$

else

$$\angle Y\left(t_m,\Omega_k\right)=\angle Y\left(t_m,\Omega_k\right) + sign\left(princ\_arg\left(\angle Y\left(t_m,\Omega_k\right)-\angle X\left(t_n,\Omega_k\right)\right)\right)\Psi_{\max} \tag{23c}$$

where Ψ_{max}, is the maximum deviation in frequency, as determined in section 3, *sign* is a function that returns the sign of the submitted value i.e. 1 or –1 and *princ_arg* returns the principle argument of the submitted value between $\pm\pi$.

For the following paragraphs it is important to be aware of two situations; the first situation is where consecutive analysis frames are mapped to consecutive synthesis frames e.g. in figure 1 the consecutive analysis frames 2, 3 and 4 are mapped to three consecutive synthesis frames 3', 4' and 5', this case can be described more generally as the situation when $t_m \rightarrow t_n$ and $t_{m-1} \rightarrow t_{n-1}$; the second situation covers all other cases.

It should be noted that for the case where consecutive analysis frames are not mapped to consecutive synthesis frames, Ψ_{max} should be reduced to take the likelihood of increased inaccuracies of phase estimates into consideration when using equation (4). Phase estimates of consecutive analysis frames that are mapped to consecutive synthesis frames are likely to be accurate, at least for peaks, since the same phase differences are kept between consecutive analysis frames as consecutive synthesis frames; the same cannot be said for the case where consecutive analysis frames are not mapped to consecutive synthesis frames. It is difficult to determine a precise figure for the inaccuracy of the phase estimate; consequently it is difficult to determine a value for the maximum phase deviation that can be introduced. From experimentation it was found that reducing Ψ_{max} to $\Psi_{max}/2$ is an adequate choice.

It should also be noted that, for the case where multiple consecutive analysis frames are mapped to multiple consecutive synthesis frames, a reduction in phase differences between one synthesis frame and its corresponding, mapped, analysis frame results in the same phase reduction for all consecutive synthesis frames that follow; since from equation (4) the phase modifications are propagated through the remaining synthesis frames. Following from this observation, it can be noted that if $(\pi-\Psi_{max}/2)/\Psi_{max}$ consecutive analysis frames are mapped to $(\pi-\Psi_{max}/2)/\Psi_{max}$ consecutive synthesis frames the phase coherence is guaranteed to be recovered for at least one of the consecutive synthesis frames (the $\Psi_{max}/2$ value represents the phase deviation introduced for non-consecutive synthesis frames). Therefore, the closer the time-scale factor is to one the greater the opportunity to recover phase coherence, since the number of consecutive analysis frames mapped to consecutive synthesis frames, k, is given by

$$k = 1/|1-\alpha| \tag{24}$$

It then follows that phase coherence is guaranteed to be recovered at least once every k frames if

$$\alpha > (\pi - 3\Psi_{max}/2)/(\Psi_{max}/2 - \pi) \text{ for } \alpha < 1 \tag{25a}$$

or

$$\alpha < (\pi + \Psi_{max}/2)/ (\pi - \Psi_{max}/2) \text{ for } \alpha > 1 \tag{25b}$$

Since phase coherence is ensured for some sections of the time-scaled output if equation (25a) or (25b) is satisfied, it follows that these sections are copies of the sections of the input. Therefore, these 'copied' sections do not have to be processed in the frequency domain and can be simply overlapped and added to the time-scaled output; resulting in a reduction in the computational requirements of the approach. This process is illustrated in figure 4, where the analysis frame marked B would achieve phase coherence and the synthesis frame marked A' is almost phase coherent i.e. all STFT bins of frame A' are within Ψ_{max} radians of the phase of the mapped analysis frame marked A.

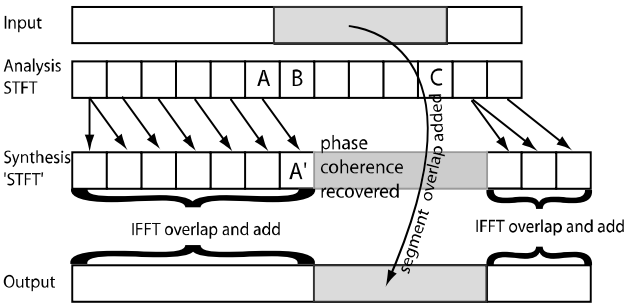

Figure 4 : *Copying a time-domain segment to the output*

The phases of the analysis frame marked C are required to calculate equation (4), therefore, given a set of analysis time instants $t_u = uR$, where u is a set of consecutive integer values starting at 0, the STFT needs only be calculated, at most, for the cases when

$$floor(u|1-\alpha|)/|1-\alpha| -1 \le u \le floor(u|1-\alpha|)/|1-\alpha| + ceil((\pi- \Psi_{max}/2)/\Psi_{max}) \tag{26}$$

where *ceil* and *floor* are functions that return the nearest integer greater than and less than the value submitted, respectively.

Equation (26) provides the maximum number of analysis time instants at which the STFT must be calculated to ensure phase coherence. Further computational savings can be achieved by recognizing that phase coherence can be achieved at any frame within a set of $(\pi-\Psi_{max}/2)/\Psi_{max}$ consecutive synthesis frames. So, given that the synthesis frame mapped to the analysis frame at the analysis time instant $R(floor(u|1-\alpha|)/|1-\alpha| + h)$ is almost phase coherent i.e. all bins are within Ψ_{max} radians of the phase of the mapped analysis frame, then no frequency domain processing is required at the analysis time instants, uR, for u in the range

$$floor(u|1-\alpha|)/|1-\alpha| -1 + h < u < floor((u+1)|1-\alpha|)/|1-\alpha| \tag{27}$$

where h is an integer less than $1/|1-\alpha|$.

By making the assumption that all computations other than calculating the STFT and Inverse STFT are negligible, figure 5 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

utilizes the phasiness reduction technique described in this paper. The solid line is plotted for Ψ_{max} = 0.3313 radians and the dashed line is plotted for Ψ_{max} = 0.24 radians.

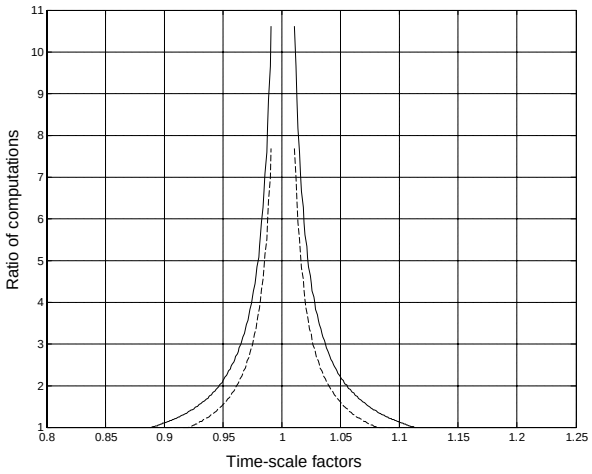

Figure 5 : *Computational advantage of the technique*

5. SUBJECTIVE TESTING AND DISCUSSION

Eight test subjects undertook a number of subjective listening tests. The results indicate that the improvement in the quality of time-scaled output achieved by using this approach is most effective for time-scale factors close to one with a significant improvement noticed for moderate time-scale factors in the range 0.9-1.1. Beyond this limit, the reduction in phasiness is less significant and no improvement in quality was perceived for time-scale factors outside the range 0.85-1.15. The results also indicate a greater improvement for speech signals, due to the fact that the phasiness artifact is more objectionable in speech to begin with. Phasiness appears to be more objectionable in speech because reverberation, which is similar to phasiness, is not often noticeably present in a speech signal, so when it is inadvertently introduced it tends to be obvious; whereas in music reverberation is often noticeably present, and is even synthetically added to music recordings, consequently, when additional reverberation, or phasiness, is introduced into a music signal it is less obvious and therefore less objectionable. The reduction in phasesiness is also particularly noticeable in gravelly type speech. This was attributed to the fact that the phase update procedure proposed in [1] is most applicable to signals composed of strong sinusoidal components and gravelly speech seems to violate this model to a greater degree than other types of speech.

Figure 6 illustrates the effects of the phasiness reduction technique on a speech signal. It should be noted that while the preservation of the waveform shape, i.e. shape invariance, does not ensure phase coherence, the loss of shape invariance can be attributed to a loss of phase coherence.

The range of time-scale factors over which the technique has a significant reduction in phasiness is quite restrictive for many applications, however, it is ideally suited to such applications as audio-video synchronization in broadcasting application, which require time-scale modifications in the range 24/25-25/24 [11].

The phasiness reduction technique described in this paper has similarities with time-domain approaches [12], in that, for moderate time-scaling, certain segments of the time-scaled signal are a copy of the original, as is the case in time-domain approaches; the phase vocoder, however, has the advantage of producing better results for complex polyphonic audio. The technique also has similarities to the synchronised time-domain/subband approach described in [13], where individual subbands are 'pulled' or 'pushed' into a synchronised state by taking advantage of some psychoacoustic properties.

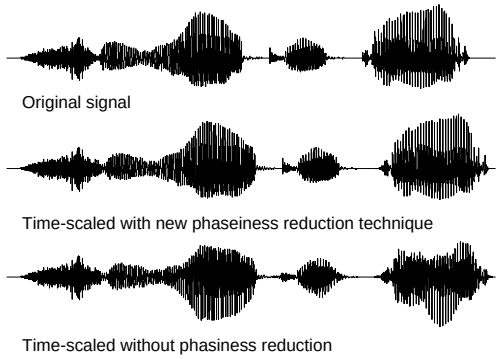

Figure 6 : *The effects of the reduction of phasiness*

6. CONCLUSION

Time-scale modification of audio using phase vocoder based approaches require both horizontal and vertical phase coherence between modified STFT bins to produce a high quality output. In this paper it is shown that some flexibility exists in the choice of phase required to ensure horizontal phase coherence, when psychoacoustic properties are considered. This flexibility in horizontal phase is then used to 'push' or 'pull' the modified STFT into a phase coherent state, resulting in a reduction in the phasiness artifact associated with phase vocoder time-scaling implementations, for moderate time-scale factors in the range 0.9-1.1. It is also shown that the phasiness reduction technique results in a significant reduction in computational overhead for moderate time-scaling.

7. REFERENCES

[1] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, issue 3, pp. 323-332 May 1999.

[2] J. Bonada, "Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio," *Proceedings of International Computer Music Conference*, Berlin, Germany, 2000.

[3] J. L. Flanagan and R. M. Golden, "Phase Vocoder," *Bell System Technical Journal*, pp. 1493-1509, November 1966.

[4] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24 issue 3, pp. 243-248, June 1976.

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

[6] A. De Goetzen, N. Bernardini and D. Arfib, "Traditional implementations of a phase vocoder: the tricks of the trade," *Proceedings of the International Conference on Digital Audio Effects (DAFx00)*, pp. 37-43, Verona, Italy, December 2000.

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

[8] X. Serra and J. O. Smith, "PARSHL: An Analysis /Synthesis Program for Non-Harmonic Sounds based on a Sinusoidal Representation," *Proceedings of the International Computer Music Conference*, pp. 290 – 297, 1987.

[9] 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.

[10] 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.

[11] G. Pallone, P. Boussard, L. Daudet, P. Guillemain, and R. Kronland-Martinet. "A wavelet based method for audio-video synchronization in broadcasting applications," *Proceedings of the COST G-6 Conference on Digital Audio Effects*, Trondheim, Norway, pp. 59–62, December 1999.

[12] D. Dorran and R. Lawlor, "An efficient time-scale modification algorithm for use within a subband implementation," *Proceedings of the International Conference on Digital Audio Effects (DAFx03)*, London, pp. 339-343, September 2003.

[13] D. Dorran and R. Lawlor, "Time-scale modification of music using a synchronized subband/time-domain approach," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV 225 – IV 228, Montreal, May 2004.