Time and Frequency Scale Modification of Speech Signals

Brett Ninness* Soren John Henriksen †

Abstract

This paper presents new and improved methods for independently modifying the time and pitch scale of acoustic signals, with an emphasis on speech signals. The algorithms developed here use parametric (sinusoidal) modelling techniques introduced by other authors, but new ideas are presented here that achieve improved output quality with decreased computational load. In particular, speech quality is improved by using novel ideas to reduce phase dispersion in the scaled signal. The methods described here have been implemented and tested in real-time on a custom-designed portable hardware platform.

Technical Report EE9916, Department of Electrical and Computer Engineering,
University of Newcastle,AUSTRALIA. EDICS Number: 1-ENHA

1 Introduction

There are a number of applications where it is desirable to change the time or pitch scale of an audio signal. A common instance is one in which speech needs to be slowed down in order to make it intelligible; for example during foreign language translations, or for hearing-impaired listeners. In other applications it is also useful to be able to increase the rate of articulation, so that the material may be scanned quickly. In both the aforementioned cases of rate change, it is essential that the pitch and tonal quality of the speaker should remain the same, but in others (recovery of helium-distorted speech) the pitch must be modified while the rate of articulation remains the same.

Perhaps the simplest method of time scaling a sound recording is to just replay it at a different rate. When using magnetic tapes, for example, the tape speed may be varied, but this incurs a simultaneous change in the pitch of the signal. In response to this problem, a number of authors have developed algorithms to independently perform time and pitch scaling; see [6] for a comprehensive survey.

Some of these methods are based on time domain splicing/overlap-add approaches [6, 11, 2, 12], which have the advantage of being computationally cheap, but at the expense of suffering from echos (perceived delayed and diminished amplitude versions of the signal being present in the reconstruction) and other defects [6]. The work in this paper examines a more sophisticated approach, and uses ideas that can be traced back at least to 1981 [7] where a frequency domain approach was used via short-term Fourier Transform calculations.

Since that work, many other methods using the same (and related) frequency domain approaches have been developed [6, 4, 1, 9]. The algorithms involved tend to be computationally intensive, but are capable of providing very high quality output. However, they still suffer from some distortion, mainly due to the effects

*This author is with the Department of Electrical and Computer Engineering, and the Centre for Integrated Dynamics and Control, both at The University of Newcastle, Australia and can be contacted at email:brett@ee.newcastle.edu.au or FAX: +61 2 49 21 69 93

†This author is with the Department of Electrical and Computer Engineering, and the Centre for Integrated Dynamics and Control, both at The University of Newcastle, Australia and can be contacted at email:eesjh@ee.newcastle.edu.au or FAX: +61 2 49 21 69 93

of 'phase dispersion'. That is, while the scaled signal has the same frequency content, the phases between the components change, resulting in a different wave shape.

The contribution of this paper is to develop new frequency domain type time/pitch scale modification methods that provide improved quality output by addressing the phase dispersion problem, while at the same time introducing important modifications that greatly reduce computational burdens. The latter allows the implementation of our new method on cheap and portable hardware, which is what is normally required in applications.

A key point is that the work here is inspired by that of [3, 4], and seeks to extend it by addressing problems of phase-dispersion that the authors of [3, 4] noted. The pre-existing work [8] also addresses this same issue, and here we do not attempt to compete with that contribution nor make claims of superiority. Instead, we offer here an alternative perspective, derivation, interpretation and solution to the phase dispersion problems of [3, 4].

The remainder of the paper is organised as follows. The following §2,3,4.1 and 4.2 provide an overview of the well known sinusoidal based speech analysis and reconstruction methods of [4, 3] which form a basis for the methods of this paper. §4.3 then provides some new insights into a term M_k^m arising during phase interpolation in [4, 3] and the more recent work [8]. §5.1 and 5.2 then provide an overview of pre-existing rate and pitch scale modification methods introduced in [4, 3] and also used in [8]. §5.3 then provides analytical argument (which has been substantiated by listening tests) that the separate source/filter decomposition steps employed in [3, 4, 8] are un-necessary, and hence a significant computational load saving may be made. §6 contains one of the main contributions of the paper in which, via a modification of the earlier profiled work [3, 4, 8], a method that reduces phase dispersion (defined as mismatch between analysis frame and reconstruction frame phase at appropriately matched time instants). As part of this §6.1 introduces a new and simple pitch period estimator which we have found to work well in practice, although this material is not considered core to the papers contribution since other pitch estimators could equally well be employed without altering the main thrust or results of this work. Finally, §7 provides detail on the experimental performance of our method on both real speech and synthetically generated waveforms.

It is important to mention that the authors of the work [3, 4] inspiring this paper, have themselves addressed the phase dispersion problems associated with [3, 4] in [8]. However, as will be profiled in § 6.2, there are several fundamental differences between the solution in [8] and that presented here.

2 Sinusoidal Representation

The algorithms developed in this paper are based on the work [3, 4, 8] in which a source-filter model as depicted in figure 1 is used for the speech production process, and a sum-of-sinusoids description is used for the source model. Within this framework, the excitation signal of the vocal cords is expressed as the linear combination

$$e(t) = \sum_{k=1}^{N} A_k(t) \cos\left(\int_0^t \omega_k(\xi)\,\mathrm{d}\xi + \Phi_k\right), \tag{1}$$

where the $\omega_k(t)$ are the time-varying frequencies of the excitations, which are not necessarily all harmonically related. The vocal tract $H(s,t)$ then performs a filtering operation on the excitation signal before it is radiated at the mouth. The utility of this model is derived from the fact that voice signals (and other acoustic signals) can be modelled to good accuracy with only a (relatively) small number N of sinusoids.

Previous work [3, 4, 8] using this model has considered the excitation and vocal tract responses separately. A key contribution of this paper is to show, via the development of a new algorithm, that this separation is unnecessary (see §5.3), and avoiding it greatly reduces computational overhead.

3 Implementation of Sinusoidal Analysis

In order to use the sinusoidal representation (1) on observed data, it is first necessary to estimate the time varying parameters $A_k(t)$, $\omega_k(t)$ and Φ_k that specify (1) from that data. For this purpose, it is assumed (as in [3, 4]) that

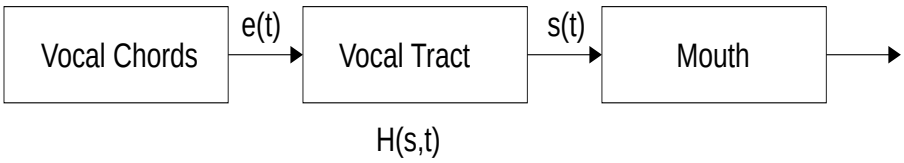

Figure 1: *Source-Filter voice generation model*

the time variation is approximately piecewise constant over sufficiently short durations called 'analysis frames', which are illustrated in figure 2. As shown, the analysis frames are overlapped and the amount of overlap is

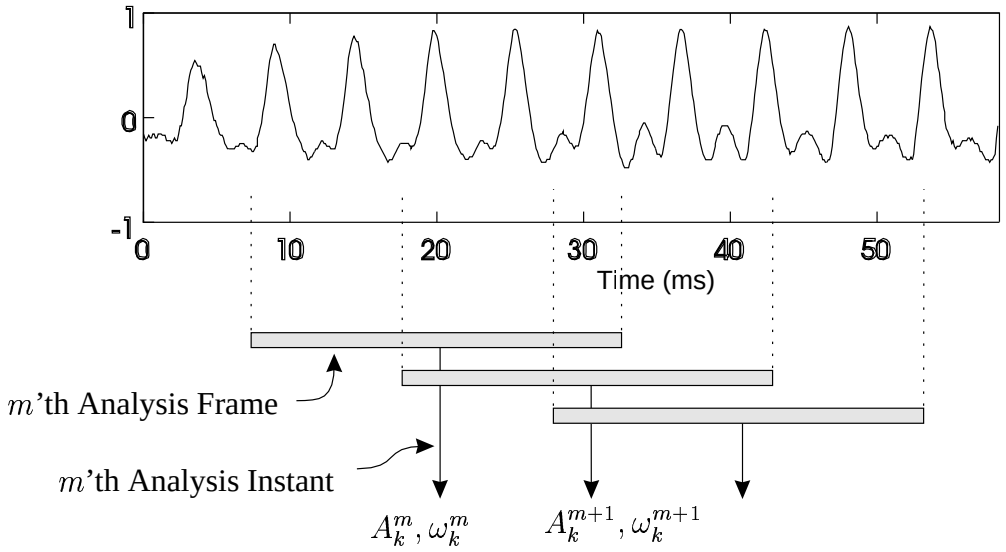

Figure 2: *Calculation of sinusoidal model parameters using overlapping analysis frames.*

a tradeoff between an ability to deal with rapidly varying signals, and computational overhead; the time scale used in figure 2 is only a representative one and is not prescriptive.

On each frame, the spectral content of the signal may be determined through an appropriately windowed Discrete Fourier Transform (DFT). The location of the peaks of the DFT magnitude function are used as estimates of ω_k, the frequencies of the underlying sine-wave components. That is, a frequency estimate is taken as

$$\omega_k \approx 2\pi \times \frac{n}{M} \times f_s \qquad \text{(Hz)} \tag{2}$$

where n is the 'bin number' of the DFT output at which a peak occurs, M is the number of data samples on an analysis frame, and f_s is the sampling frequency (it is assumed that the DFT is implemented as an FFT so that a linear frequency grid ensues. When M is not power of 2, then zero padding or mixed-radix versions of the FFT may be employed [10]), although in this case the relationship (2) between FFT bin-number n and frequency estimate ω_k will need to be modified accordingly.

The magnitude and phase of the Fourier Transform at these measured frequencies are used as estimators (respectively) of A_k and Φ_k. A novel feature of the algorithm presented in this paper is that, as will be detailed presently, it uses the phase estimates of sine-waves in two successive analysis frames to (effectively) generate improved estimates of the frequency.

New model parameters $A_k^m, \omega_k^m, \Phi_k^m$ are calculated for analysis frame number m, and the chosen width of the frame involves a tradeoff between frequency resolution, ability to model time variations, and computational overhead. As in [3], a Hamming window was employed in the test results to be reported on later, and experimentation with a range of speech signals showed that the window width was best set to three times the estimated pitch period.

For good reproduction of speech, it was also found that around 20-40 DFT peaks need to be analysed. A point is chosen to be a peak when its DFT magnitude is larger than the surrounding points, and is above a set threshold; this is illustrated on a typical voiced speech analysis frame in figure 3. The threshold level is adaptively chosen to control the number of peaks detected.

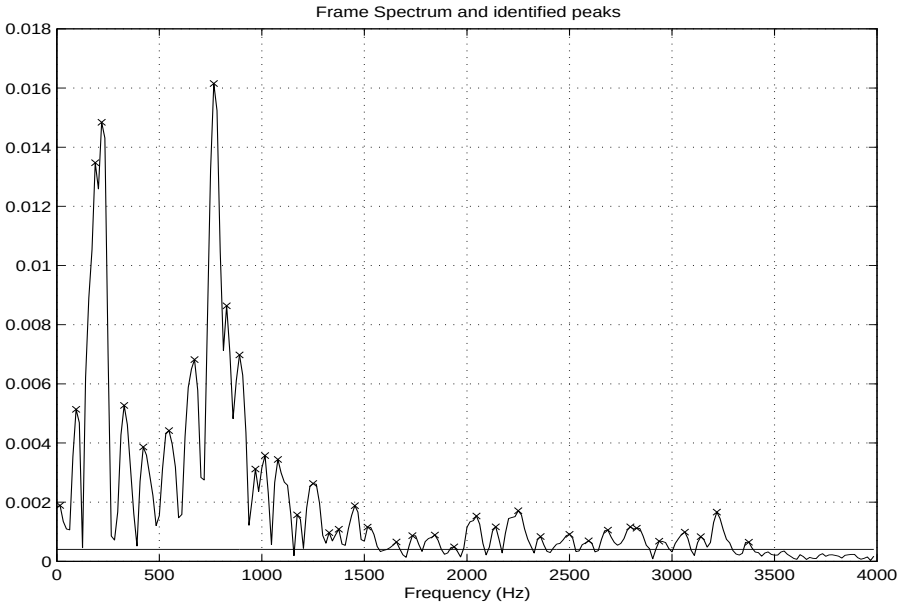

Figure 3: *Windowed DFT spectrum of typical voiced speech analysis frame with identified peaks marked as crosses and lower horizontal line representing threshold level.*

4 Track-Based Reconstruction

The sinusoidal parameter extraction method described above determines sinusoidal models at a rate equal to the analysis frame rate. However, for subsequent signal generation (and hence rate/pitch modifications), what is required are time varying estimates $A_k^m(t), \omega_k^m(t)$ which vary with t discretised at the original signal sampling rate f_s.

The method used here to generate these time varying estimates is based on one developed in [3, 4, 8] in which the concept of 'frequency tracks' is used, and smooth interpolants sampled at rate f_s are created between estimates associated with common tracks occurring at the frame rate.

4.1 Frequency Tracks

At each analysis instant, peak extraction has identified a set of underlying sine-waves. A complication then is that the peaks vary from frame to frame. From one analysis instant to the next, both the number and the location of the peaks change. In terms of the sinusoidal model, this occurs because the frequencies of the sinusoidal components change. Over time, new frequency components will appear and others will disappear.

The idea of 'frequency tracks' is to match certain sine-wave components between frames in the belief that they correspond to a common generating source whose instantaneous frequency has changed. This is illustrated in figure 4 where the broken lines denote these tracks that involve matching between frames. Two frequencies

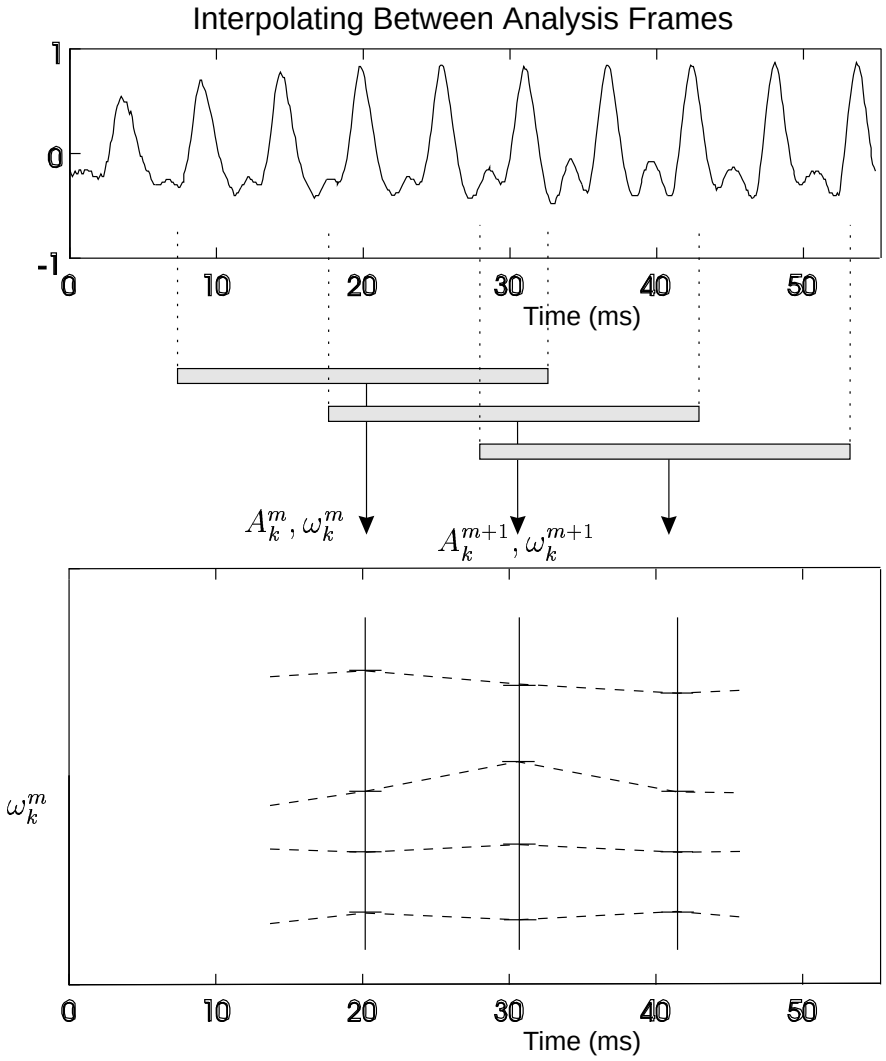

Figure 4: *Origin of "frequency tracks"*

are matched to produce a track if they are within a certain tolerance of one another, with competitions for a match within this tolerance being resolved according to the closest match. In cases where a match cannot be made, a track is declared either 'born' or 'died' depending on whether it lacks a match on a previous or ensuing frame. For further details see [3] where the method was originally proposed.

4.2 Interpolation Between Analysis Instants

In deciding how smooth interpolants of $A_k^m(t)$, $\omega_k^m(t)$ should be generated between estimated values corresponding to matched frequency tracks on adjacent analysis frames, it should first be recognised that a sensible criterion for forming the interpolants is that according to the model (1), they should lead to a successful signal

reconstruction as

$$s_R^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos\left(\phi_k^m(t) + \Phi_k^m\right), \qquad \phi_k^m(t) \triangleq \int_0^t \omega_k^m(\xi)\, \mathrm{d}\xi. \tag{3}$$

In order to achieve this, assuming that the analysis window is of width $t = T_A$, an obvious interpolation requirement is that (assuming that the k'th frequency track on frame m is matched to the j'th track on frame $m+1$):

$$A_k^m(0) = A_k^m, \qquad A_k^m(T_A) = A_j^{m+1}$$

and this may be simply achieved by the linear interpolant

$$A_k^m(t) = \left(\frac{T_A - t}{T_A}\right) A_k^m + \left(\frac{t}{T_A}\right) A_j^{m+1}.$$

For the case of interpolating the phase, the situation is more complicated since the coupling between frequency and phase introduces four constraints

$$\phi_k^m(0) = 0, \qquad \phi_k^m(T_A) + \Phi_k^m = \Phi_j^{m+1} + 2\pi M_k^m, \tag{4}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\bigg|_{t=0} = \omega_k^m, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\bigg|_{t=T_A} = \omega_j^{m+1},$$

where M_k^m is an integer constant which allows for phase unwrapping. Previous developers [3] have recognised that a cubic spline fit possesses sufficient degrees of freedom to satisfy the above interpolation constraints so that $\phi_k^m(t)$ and its derivative are of the following form

$$\phi_k^m(t) = at^3 + bt^2 + ct + d, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t) = 3at^2 + 2bt + c.$$

In this case, the interpolation constraints at $t = 0$ immediately lead to

$$d = 0, \qquad c = \omega_k^m.$$

Using this, the constraints at $t = T_A$ then require the simultaneous solution of

$$\begin{aligned} \Phi_j^{m+1} + 2\pi M_k^m &= aT_A^3 + bT_A^2 + \omega_k^m T_A + \Phi_k^m \\ \omega_j^{m+1} &= 3aT_A^2 + 2bT_A + \omega_k^m \end{aligned}$$

which may be expressed as

$$\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \dfrac{3}{T_A^2} & -\dfrac{1}{T_A} \\ -\dfrac{2}{T_A^3} & \dfrac{1}{T_A^2} \end{bmatrix} \begin{bmatrix} \Phi_j^{m+1} - \Phi_k^m - \omega_k^m T + 2\pi M_k^m \\ \omega_j^{m+1} - \omega_k^m \end{bmatrix}. \tag{5}$$

4.3 A new interpretation of M_k^m

As mentioned above, the integer parameter M_k^m in (4), (5) is included to provide 'phase unwrapping', and previous workers [3, 4, 8] have chosen it according to the value that maximises the smoothness (average second derivative) of $\phi_k^m(t)$.

Here we provide an alternative argument for its choice that has a certain physical significance (missing in a maximally smooth argument) that will subsequently prove useful in understanding the process of time scale alteration.

Firstly, the average frequency of a track on the m'th analysis frame is given by

$$\omega_{av}^m = \frac{1}{T_A} \int_0^{T_A} \omega_k^m(t)\, \mathrm{d}t \tag{6}$$

where $\omega_k^m(t)$, being the instantaneous frequency, may be expressed as the derivative of the phase function $\phi_k^m(t)$ so that (by the fundamental theorem of calculus)

$$\omega_{av}^m = \frac{1}{T_A} \int_0^{T_A} \frac{\mathrm{d}\phi_k^m(t)}{\mathrm{d}t}\, \mathrm{d}t = \frac{1}{T_A} \left[\phi_k^m(T_A) - \phi_k^m(0)\right]. \tag{7}$$

Substituting in the interpolation constraints in (4) then leads to

$$\omega_{av}^m = \frac{1}{T_A} \left[\left(\Phi_j^{m+1} + 2\pi M_k^m\right) - \Phi_k^m\right]. \tag{8}$$

But the average frequency on an analysis frame may equally well be interpreted as the average value of frequency at frame endpoints and corresponding to matched tracks:

$$\omega_{av}^m = \frac{\omega_k^m + \omega_j^{m+1}}{2}.$$

Clearly, in forming a phase interpolant $\phi_k^m(t)$, it is desirable that it be constructed to be consistent with as much measured phase and frequency information as possible, including that of the average frequency variation across a frame. This implies that M_k^m should be chosen by equating the above two expressions for ω_{av}^m and solving for M_k^m:

$$M_k^m = \frac{1}{2\pi} \left[\frac{T_A}{2}(\omega_k^m + \omega_j^{m+1}) + \Phi_k^m - \Phi_j^{m+1}\right]. \tag{9}$$

This is identical to the value for M_k^m obtained in [3] according to a criterion of maximally smooth phase function $\phi_k^m(t)$. Note that since M_k^m must be an integer, the expression (9) is rounded to the nearest integer value.

The point is that the above interpretation of M_k^m shows that the choice for M_k^m of (9) is much more than one that achieves a certain heuristically reasonable, but otherwise seemingly non-essential, goal of maximally smooth phase. In fact, the choice (9) is the only one that makes the nature of the interpolated $\phi_k^m(t)$ completely consistent with the measured frequency information and track matching choices encoded in ω_k^m, ω_j^{m+1}.

Furthermore, the above derivation has physical importance in that it shows that the DFT measurements give two separate measures of the frequency of the track. There are the obvious measures in ω_k^m and ω_j^{m+1}, but the phase information also contains a measure of the frequency. Given a value of M_k^m, the interpolant derivative $\mathrm{d}\phi_k^m(t)/\mathrm{d}t$ can be used to obtain an improved estimate of the instantaneous frequency of the track in the sense that, at any point, it is not constrained strictly by the bin-width of the underlying FFT that via the location of peaks, provides the estimates ω_k^m.

5 Magnitude/Frequency Based Scaling

Once a sinusoidal model of the form (1) has been identified, the pitch and rate may be independently manipulated by altering the time rate of change of the interpolated magnitude and phase functions.

In the following, we give a brief overview of how this may be achieved according to the methods developed in [4]. This is in preparation for work in subsequent sections that will illustrate how pre-existing methods may be improved while also lowering computation overheads.

5.1 Time Scaling

The process of time-scaling a sinusoidally modelled signal is shown in figure 5. As illustrated there, each

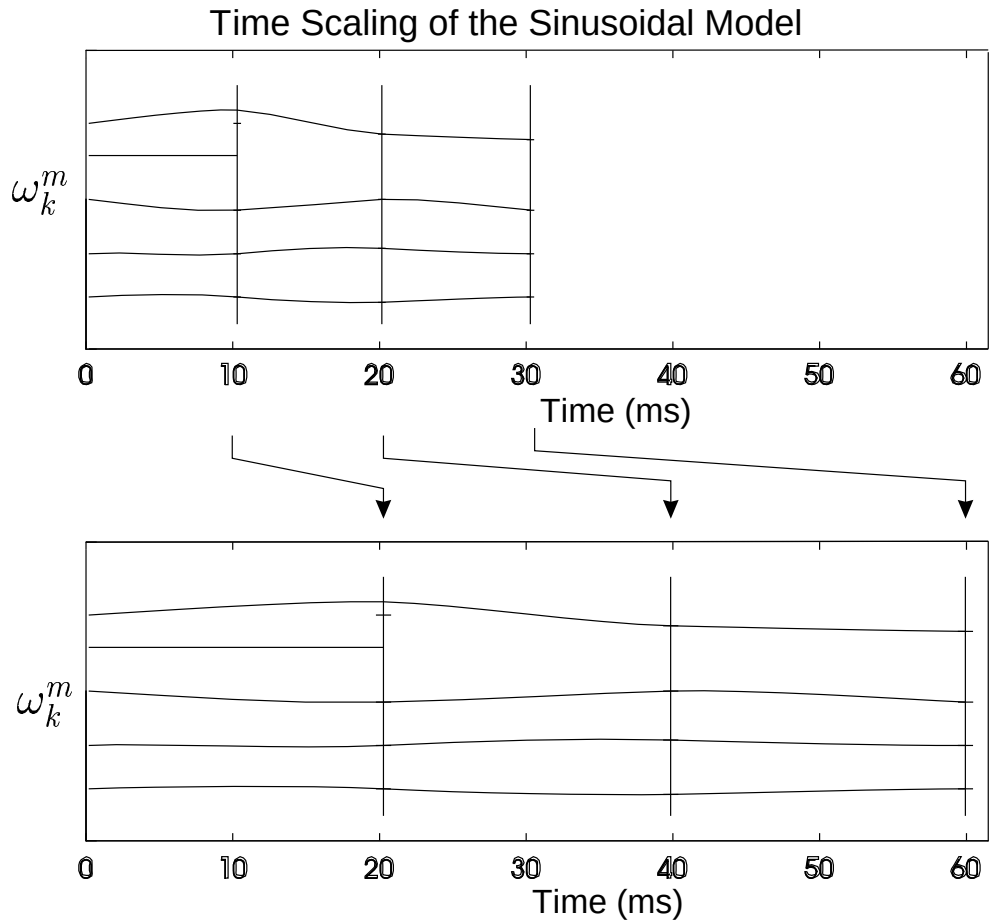

Figure 5: *Time scaling of the sinusoidal model.*

analysis frame (of a representative 10ms duration) has been mapped to a reconstruction frame that is twice the original length, which would correspond to a time-rate change factor of $\rho = 2$. However, it is clear that in the general case of arbitrary ρ, the length T_A of analysis frame and T_R of reconstruction frame should be related as

$$T_R = \rho T_A.$$

We will first consider the problem of maintaining the perceived pitch to be invariant under time-scale modification. Recall that in §4.2, the signal is reconstructed on frame m from interpolated measurements as

$$s_R^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos\left(\phi_k^m(t) + \Phi_k^m\right). \tag{10}$$

By expressing the instantaneous frequency as the derivative of the phase function,

$$\omega_k^m(t) = \frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t),$$

the reconstructed signal (10) may be re-expressed as

$$s_R^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos\left(\int_0^t \omega_k^m(\xi)\,\mathrm{d}\xi + \Phi_k^m\right).$$

For the purposes of time scaling while maintaining an unaltered pitch, the amplitude and frequency information at time t should be mapped to a new time $t' = \rho t$, so that the scaled signal may be represented as,

$$s_S^m(t') = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t'}{\rho}\right) \cos\left(\int_0^{t'} \omega_k^m\left(\frac{\xi}{\rho}\right)\,\mathrm{d}\xi + \Phi_k^m\right).$$

Changing the integration variable to $\mu\rho = \xi$ then allows this to be written as

$$s_S^m(t') = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t'}{\rho}\right) \cos\left(\rho\int_0^{t'/\rho} \omega_k^m(\mu)\,\mathrm{d}\mu + \Phi_k^m\right)$$

which may be expressed in terms of the original interpolating phase function as

$$s_S^m(t') = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t'}{\rho}\right) \cos\left(\rho\,\phi_k^m\left(\frac{t'}{\rho}\right) + \Phi_k^m\right). \tag{11}$$

This function has been derived over one (the m'th) reconstruction frame only, and the polynomial phase function $\phi_k^m(t)$ is only valid for $0 < t < T_A$, where T_A is the analysis frame length.

Of course in practice one would wish to reconstruct, in a time-scale modified manner, a continuous stream of data. This could be achieved most directly by calculating a scaled reconstruction frame for each analysis frame and then simply concatenating successive output frames. Unfortunately, this will result in a strongly degraded result because the time scaling incurs a lack of matching of phases between successive frames.

To see this, note that in the case of unscaled reconstruction, the polynomial phase function $\phi_k^m(t)$, for the k'th track of the m'th frame, was designed so that $\phi_k^m(0)$ and $\phi_k^m(T_A)$ matched the measured DFT phases at the start and end of the frame respectively. If the kth track of the mth frame matches to the j'th track of the $m+1$'st frame then according to (4)

$$\phi_k^m(T_A) + \Phi_k^m = \Phi_j^{m+1} + 2\pi M_k^m. \tag{12}$$

so that continuity of phase between successive frames is ensured since the begin-phase of one frame matches the end-phase of the previous frame modulo 2π.

However, when time scaling by a factor ρ is introduced, then via (12) the argument to the cosine in (11) at $t' = T_R = \rho T_A$ is

$$\begin{aligned}\rho\phi_k^m(T_A) + \Phi_k^m &= \rho\left(\Phi_j^{m+1} + 2\pi M_k^m\right) + (1-\rho)\Phi_k^m, \\ &= \rho\phi_j^{m+1}(0) + \Phi_j^{m+1} + \rho\,2\pi M_k^m + (1-\rho)(\Phi_k^m - \Phi_j^{m+1})\end{aligned} \tag{13}$$

and since, according to the ρ scaled reconstruction (11), the matched j'th starting phase on the $m+1$'st frame will be $\rho\,\phi_j^{m+1}(0) + \Phi_j^{m+1}$, then there is a phase discontinuity of $2\pi\rho M_k^m + (1-\rho)(\Phi_k^m - \Phi_j^{m+1})$ across the concatenated frame boundary.

This abrupt mismatch of phase across boundaries represents a distorting influence that seriously degrades the perceived quality of the time scaled signal.

The solution proposed in [4] is to add a further phase term γ_k^m to the reconstruction (11) so that it becomes

$$s_S^m(t') = \sum_{k=1}^{N(m)} A_k^m \left(\frac{t'}{\rho}\right) \cos\left(\rho\, \phi_k^m \left(\frac{t'}{\rho}\right) + \Phi_k^m + \gamma_k^m\right) \tag{14}$$

where γ_k^m is calculated to eliminate the discontinuity by setting it according to

$$\gamma_j^{m+1} = \rho\phi_k^m(T_A) + \Phi_k^m - \Phi_j^{m+1}. \tag{15}$$

The developers of the above method [4] (there the term γ_k^m is absorbed into a term called Σ_k^ℓ that is defined recursively) make a distinction between it and magnitude-only type schemes. While this is true in the sense that it uses the phase information Φ_k^m calculated via a DFT on each analysis frame, the reconstruction (14) with the γ_k^m offset (designed to preserve phase continuity) destroys this information in the reconstructed signal when $\rho \neq 1$.

The effect of the loss of phase information is called 'phase dispersion' in that the reconstructed signal will contain the same frequency content as the original signal, but the relationship between the phases of the different components will have changed. During passages dominated by voiced speech, the effect of this phase dispersion is to produce an effect that may be qualitatively described as 'chorusing'. The new phase invariant method developed in this paper is specifically designed to addresses this defect and hence improve the perceived quality of the time/pitch scaled signal.

As an aside, note that with instantaneous frequency being the derivative of phase, the addition of a constant phase offset γ_m^k does not affect this frequency. Hence, in a sense, since the inclusion of γ_m^k destroys any absolute phase information, the only real contribution that the measurements Φ_k^m have in forming the interpolant $\phi_k^m(t)$ is as a refinement of the measurements ω_k^m of the frequencies of the tracks.

5.2 Pitch Scaling

The process of pitch scaling is depicted in figure 6 wherein it is illustrated that in this case the duration of the analysis and reconstruction periods are kept the same, but the pitch of each track is changed, in this case scaled up by a factor $\sigma = 2$. The principle is that every frequency track is scaled up in instantaneous frequency by this same constant amount σ so that the reconstructed pitch-scaled signal may be represented as

$$s_P^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos\left(\sigma\phi_k^m(t) + \Phi_k^m\right).$$

Again, if frames are simply concatenated together, phase discontinuities will occur since the interpolation constraint (12) implies that

$$\sigma\phi_k^m(T) + \Phi_k^m = \sigma\phi_j^{m+1}(0) + \Phi_j^{m+1} + \sigma 2\pi M_k^m + (1-\sigma)(\Phi_k^m - \Phi_j^{m+1}).$$

As in the case of time scaling, it is necessary to adjust the phase for continuity across frame boundaries. By the same argument as used before, the reconstructed pitch-scaled signal is actually formed as

$$s_P^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos\left(\sigma\phi_k^m(t) + \Phi_k^m + \gamma_k^m\right)$$

where the phase offset γ_k^m is calculated as

$$\gamma_j^{m+1} = \sigma\phi_k^m(T_A) + \Phi_k^m - \Phi_j^{m+1}.$$

This method has been implemented in real-time, and after extensive listening tests, was found to perform high-quality transformations on speech signals. As in the case of time scaling, it does suffer from some phase dispersion, which can be reduced using the new phase invariant method described in §6.2.

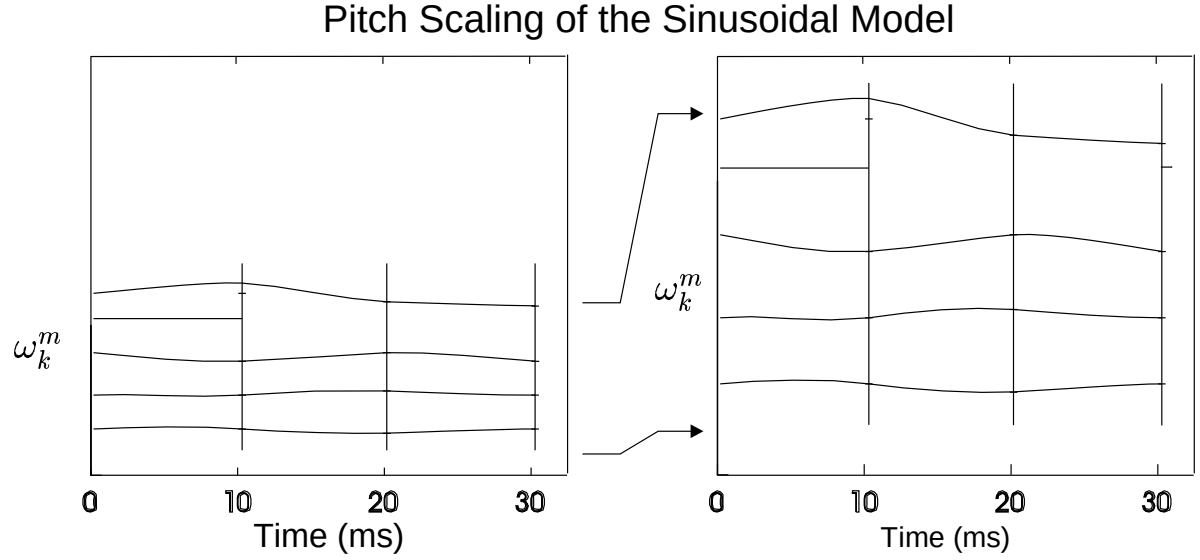

Figure 6: *Illustration of the pitch scaling process.*

5.3 The Role of Source-Filter Decomposition

Previous work [3, 4] on pitch/rate modification using sinusoidal descriptions for vocal excitation has stressed the importance of the model shown in figure 1 wherein a separate model $H(s,t)$ for the vocal tract is employed. This requires that $H(s,t)$ be estimated via homomorphic deconvolution and subsequent Hilbert transform.

By way of contrast, the new pitch/time scale modification method presented in §6 following, and the overview of pre-existing ideas presented in §5.1,5.2 have ignored the effect of the vocal tract. Underlying this is the discovery that, in practice, no perceivable difference exists between strategies of ignoring the vocal tract response, or accounting for it. Since the computational burden of the deconvolution-Hilbert Transform step estimating $H(s,t)$ is quite high (roughly 30% of the whole load), there is a significant advantage involved with being able to dispense with it.

Apart from the empirical evidence indicating that separately accounting for the vocal tract is unnecessary, it may also be argued on physical grounds as follows. Assuming that the time variation of $H(s,t)$ is such that it is approximately constant across the duration T_A of the m'th analysis frame, then its response may be written as

$$|H(j\omega_k^m, t)| = B_k^m, \qquad \angle H(j\omega_k^m, t) = \theta_k^m.$$

Therefore, the total time varying phase $\phi_k^m(t)$ on a reconstruction frame is broken into three parts as $\phi_k^m(t) = \psi_k^m(t) + \theta_k^m(t) + \Lambda_k^m$ where $\psi_k^m(t)$ is the time varying phase contribution due solely to the vocal cord excitation and hence is given as

$$\psi_k^m(t) = \int_0^t \omega_k^m(\xi)\,\mathrm{d}\xi,$$

while Λ_k^m is the starting phase of the vocal cord excitation, and $\theta_k^m(t)$ is the time varying phase of the vocal tract which satisfies boundary conditions of $\theta_k^m(0) = \theta_k^m, \theta_k^m(T_A) = \theta_j^{m+1} = \theta_j^{m+1}(0)$ so that the previously specified phase boundaries of Φ_k^m, Φ_j^{m+1} are refined into two excitation and vocal tract component Λ_k^m and θ_k^m as

$$\theta_k^m(0) + \Lambda_k^m = \Phi_k^m, \qquad \theta_k^m(T_A) + \Lambda_k^m = \theta_j^{m+1}(0) + \Lambda_j^{m+1} = \Phi_j^{m+1}.$$

In this case, the pitch/scale invariant reconstruction (3) on frame m becomes

$$s_R^m(t) = \sum_{k=1}^{n} A_k^m(t) \cos\left(\psi_k^m(t) + \theta_k^m(t) + \Phi_k^m\right).$$

The thinking behind the need for a source/filter decomposition in [3, 4] is that when (for example) a time scale modification occurs, the phase behaviour of vocal cords and vocal tract/mouth need to be separately modified (since pitch changes are affected only by the vocal cords, not the vocal tract) as

$$s_R^m(t') = \sum_{k=1}^{n} A_k^m\left(\frac{t}{\rho}\right) \cos\left(\rho\,\psi_k^m\left(\frac{t}{\rho}\right) + \theta_k^m\left(\frac{t}{\rho}\right) + \Phi_k^m\right).$$

Considering the boundary conditions, this implies a total phase variation Δ_1 across the reconstruction frame of

$$\Delta_1 = \rho(\Lambda_j^{m+1} - \Lambda_k^m) + \rho(\theta_j^{m+1} - \theta_k^m) + \rho 2\pi M_k^m.$$

On the other hand, when the source/filter decomposition is ignored, then according to (13) the total phase variation Δ_2 across a reconstruction frame is

$$\Delta_2 = \rho(\Phi_j^{m+1} - \Phi_k^m) + \rho 2\pi M_k^m.$$

Therefore, since $\Phi_k^m = \theta_k^m + \Lambda_k^m$ and $\Phi_j^{m+1} = \theta_j^{m+1} + \Lambda_j^{m+1}$ then $\Delta_1 = \Delta_2$ and hence the total phase variation across the reconstruction frame is *invariant* to whether or not a source/filter decomposition is used. Furthermore, since the offset component γ_j^{m+1} defined in (15) is added on each reconstruction frame, it is only the phase variation, and not its absolute values that is important.

As a consequence, the only difference between the two approaches of source/filter decomposition or not, is in terms of instantaneous frequency. With the methods reviewed in §5.1, 5.2 the complete time varying phase function $\phi_k^m(t) = \psi_k^m(t) + \theta_k^m(t)$ is cubically interpolated to satisfy the end conditions

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\right|_{t=0} = \omega_k^m, \qquad \left.\frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\right|_{t=T_A} = \omega_j^{m+1}$$

while with the method [3, 4] involving a source/filter decomposition, only the time varying source excitation component $\psi_k^m(t)$ is cubically interpolated to satisfy

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\psi_k^m(t)\right|_{t=0} = \omega_k^m, \qquad \left.\frac{\mathrm{d}}{\mathrm{d}t}\psi_k^m(t)\right|_{t=T_A} = \omega_j^{m+1}$$

while the vocal tract component $\theta_k^m(t)$ is linearly interpolated, and hence no derivative boundary constraints are placed on it. The net result is that for this latter method, the instantaneous frequency implied by the total interpolant $\phi_k^m(t) = \psi_k^m(t) + \theta_k^m(t)$ will in fact (as opposed to the case of §5.1, 5.2 where source/filter decomposition is ignored) be inconsistent with the measured instantaneous frequency information unless the interpolated $\theta_k^m(t)$ happens to be a constant so that $\dot{\theta}_k^m(t) = 0$. In general the inconsistency will be small since the phase estimates θ_k^m are, by the definition of the estimation method used [3, 4, 8], constrained to be slowly varying, so that in practice little discrepancy occurs (it is possible to encounter speech for which imposing the constraint of a slowly varying vocal tract is inappropriate). Nevertheless, there seems to be no reason why one would want to introduce it.

Overall then, the addition of a source/filter decomposition has little effect on the scaling process, both theoretically and in practice, but can involve considerably more computation. Note that in relation to this, the methods of [8] in provide a much simplified (compared to [3, 4]) method of recovering system phase from total phase; in essence all that is required is the removal of a linear phase that is due to pitch pulses, and hence little additional computation is required.

6 An Improved Phase-Invariant Method

Having profiled existing methods of time/pitch scale modifications (together with some new interpretations of various facets of these schemes), the remainder of the paper is devoted to the presentation of a new algorithm.

The motivation here is that while the sinusoidal model based methods just discussed are capable of what is considered high quality time and pitch scale transformations, they do suffer from a number of problems. In particular, during strongly voiced segments, phase dispersion producing distortion that may be described as 'chorusing' is a major problem [8, 6], as is the relatively high computational load involved. To address the chorusing problem, the following §6.1, 6.2 provide a solution that involves reducing phase dispersion.

6.1 A New Pitch Estimator

Key to the development of a new method that reduces distortion due to phase dispersion is the need for an estimate of the so-called 'pitch-period' on a given analysis frame. This period is defined according to an assumption that there is a dominant voicing component on a given analysis frame that produces a strong fundamental and associated harmonic elements. The pitch-period is the period of this dominant fundamental component, and while other non-harmonic sinusoids will be apparent, it is reasoned that the dominant fundamental and harmonic ones will contribute most to perceptual qualities of the complete signal.

In the sequel, distortion that appears as a 'chorusing' effect will be reduced by synchronising phases at certain time instants, and the latter will be defined in terms of this pitch period, so that an estimate of it is essential. Using the sinusoidal model (1), a pitch estimator has also been presented in [5] which, while working well, is too computationally intensive for the 'real-time' implementation aimed at here.

In response to this, a new pitch estimator is derived in this section which, while still based on the sinusoidal speech model (1), allows for a trade-off between accuracy and complexity.

A simple approach would involve approximating the location of the fundamental pitch by simply taking the lowest frequency peak of the DFT on an analysis frame and assuming it to be the fundamental. However, in practice, this strategy proves to be neither robust nor accurate. For example, components (usually small) often appear at a frequency lower than the fundamental, and in some signals it is difficult to ascertain a strong component at a fundamental frequency. Providing a good pitch period estimate therefore requires the use of information contained in the harmonics.

To proceed with the development, denote the peaks of the DFT on a particular analysis frame (see figure 3) by the frequencies

$$f_1, f_2, f_3, \ldots, f_k$$

with associated magnitudes

$$M_1, M_2, M_3, \ldots, M_k$$

respectively. Now for perfectly voiced speech, there should exist a fundamental frequency f_p, such that,

$$f_i = n_i f_p \quad \text{for } n_i \in \mathbf{Z},\ \forall i \in 1 \ldots k.$$

The aim is to estimate the pitch f_p on the given analysis frame, and for this purpose suppose that an initial estimate $f_p^\star$ of it is available; in practice, this initial estimate is the estimated pitch from the previous analysis frame.

The purpose of this initial pitch estimate is to obtain initial estimates $n_1^\star, \ldots, n_k^\star$ of the harmonic spacings as follows

$$n_i^\star = \left\lfloor \frac{f_i}{f_p^\star} + \frac{1}{2} \right\rfloor. \tag{16}$$

where $\lfloor x \rfloor$ denotes the operation of taking the largest integer not greater than x.

Using these quantities, the cumulative squared error in the choice of a pitch f_p on the current analysis frame may be defined as

$$e(f_p) \triangleq \sum_{i=1}^{k} M_i \left(\frac{f_i}{n_i^\star} - f_p \right)^2 . \tag{17}$$

The utility of the specification of an initial estimate $f_p^\star$, and hence initial estimates $n_i^\star$ via (16) is that (17) is

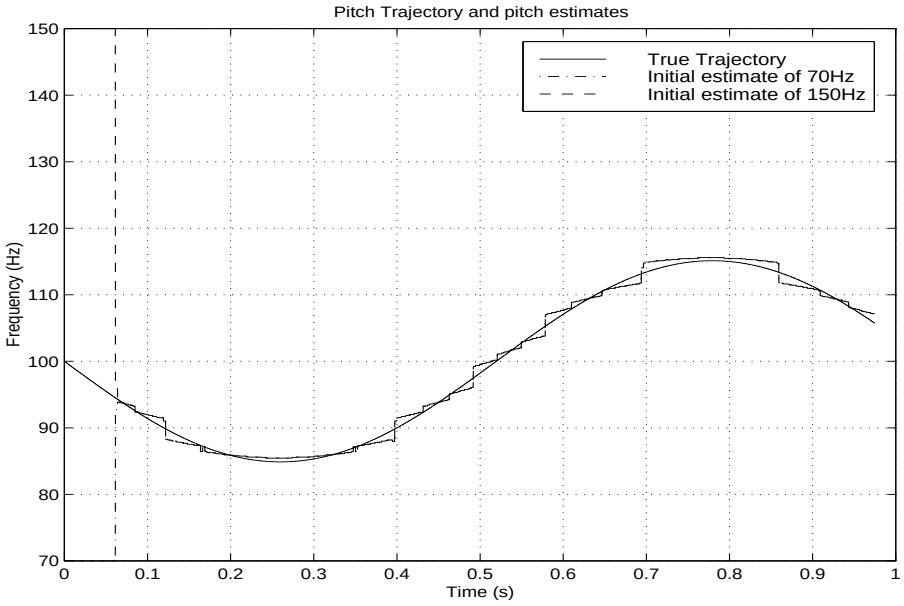

Figure 7: *Simulation example illustrating the performance of the pitch estimator. The solid line is the true time varying pitch of a synthetically generated test signal, while the dashed lines are the estimates generated by the method discussed in § 6.1 and with two (incorrect) starting guesses for the pitch*

quadratic in f_p, so that the value $\widehat{f}_p$ that minimises it (and hence the error of a pitch estimate) may be given in closed form as

$$\widehat{f}_p = \frac{1}{M} \sum_{i=1}^{k} \frac{M_i f_i}{n_i^\star}, \qquad M \triangleq \sum_{i=1}^{k} M_i . \tag{18}$$

It is this value $\widehat{f}_p$ which is taken as the pitch estimate on the analysis frame, with $\widehat{T}_p = 1/\widehat{f}_p$ being the associated pitch period estimate.

In practice, since the high frequency components are often not harmonics of the fundamental, it is better to restrict the number k of harmonics taken to be much fewer than the total number of spectral peaks identified. As well, if an estimation error is to be made, then it is preferable to minimise the error at low frequencies, since the processing of low frequency tracks is more sensitive to the accuracy of the pitch estimate. (See the description of the phase invariant reconstruction method in §6.2 following).

The pitch estimate (18) may be substituted back into the error expression (17), to generate a confidence index κ which is a normalised version of $e(\widehat{f}_p)$ as follows

$$\kappa \triangleq e(\widehat{f}_p) \left(\widehat{f}_p \sum_{i=1}^{k} M_i \right)^{-1} = \left(\widehat{f}_p \sum_{i=1}^{k} M_i \right)^{-1} \sum_{i=1}^{k} M_i \left(\frac{f_i}{n_i^\star} - f_p^\star \right)$$

This confidence index κ is used for a number of purposes in the scaling algorithms:

1. In practice, unless the initial pitch estimate is very good, the harmonic estimates from (16) will not be correct. This, in turn causes error in the pitch estimation. In order to overcome this, the error index κ may be evaluated for a number of initial pitch estimates and the estimate with the best (smallest) confidence index κ used.

2. The pitch estimate $\widehat{f}_p$ on one frame is normally used for the initial estimate $f_p^\star$ on the next frame. However, if the confidence index κ is poor (high), then the previous estimate is retained. This prevents the estimate drifting during unvoiced speech.

3. The index κ may be used as a voicing detector by reasoning that during strongly voiced segments, κ will be small. This voicing decision information may then be used to adapt the time scaling rate dependent on the amount of voicing.

   The motivation here is that consonants, which are unvoiced, are typically articulated at a fixed rate regardless of the rate of the overall speech. On the other hand, vowels which are voiced, have their time scale of articulation dilated or compressed according to the overall speech rate.

   An adaptive scaling process that slows voiced speech more than unvoiced speech can therefore provide a more realistic and intelligible output [8, 6].

In terms of assessing the performance of the pitch estimator presented here, there are manifold aspects which could be investigated. In figure 7 we have focussed on addressing the tracking and acquisition properties of our scheme via a simulation study.

To be more specific, we generated a one second duration record of a $f_s = 8000$Hz sampled synthetic voiced signal with an underlying pitch trajectory wandering around 100Hz (shown as the solid line in figure 7) and with 4 harmonics at 1,2, 20 and 30 times the frequency of the fundamental trajectory. The resultant pitch estimation and tracking abilities of the method presented here are then shown as the dashed lines in figure 7 where, to also assess acquisition performance, two erroneous starting values for the pitch estimate were used. The result appear encouraging, and further theoretical investigation of the properties of this pitch estimator would appear warranted.

6.2 A New Phase-Invariant Method

As explained in §5.1, a key limitation of pre-existing methods for time/pitch scale modification [3, 4] is the distortion introduced by the lack of preservation of phase information between analysis and reconstruction frames.

There is some debate in the literature over how sensitive the human ear is to the phase of acoustic events, but experience has exposed that reducing phase dispersion significantly improves the quality of time/pitch-scale altered acoustic signals.

The new methods presented in this section minimise phase dispersion by constraining the output to match the phases in the original signal at least once in each analysis period. In the design of this new phase-invariant method attention is concentrated on voiced speech because the distorting effects of phase dispersion are most noticeable in this type of speech. By this focus on voiced segments, an assumption of quasi-periodicity is imposed which does introduce some distortion for non-periodic components, but the effects of this have been found to not be generally noticeable in speech.

Without further preliminary comment, the key idea is not to simply ρ and σ scale an analysis frame generated interpolant $\phi_k^m(t)$ on the reconstruction frame (as is done in pre-existing methods [3, 4, 8] reported in §5.1), but rather to directly generate an interpolant $\phi_k^m(t)$ on the reconstruction frame. At first glance, this might appear to be quite simple to achieve; the interpolation considerations that lead to $\phi_k^m(t)$ in §4.2 are simply reproduced, but (for the case of time scaling) with T_A replaced by $T_R = \rho T_A$. However, this ignores an important issue of providing agreement between phase and frequency information.

As has been stressed several times so-far in this paper, when forming the interpolant $\phi_k^m(t)$, an important role of the phase information obtained on analysis frames is in fact to provide a refinement of the frequency track information. For example, the choice (9) of M_k^m is not one that just makes phase interpolants maximally smooth, it is also one that makes the phase variation of $\phi_k^m(t)$ match the average measured frequency information for a track.

In the same manner, when forming an interpolant $\phi_k^m(t)$ on the reconstruction frame, it is essential that it be formed not just to simply match measured phase information on the analysis frame, but also to be consistent with measured frequency information. In particular, on strongly voiced frames (when chorusing due to phase dispersion is most obvious), the main frequency tracks will be harmonically related according to the pitch period on that frame. This means that if, for example, $\phi_k^m(t)$ is formed to match analysis and reconstruction frame phases at the start of the reconstruction frame, then if the frequency tracks are constant (which they approximately are by virtue of how they are defined) the phases will also match at times which are integer multiples of the pitch period.

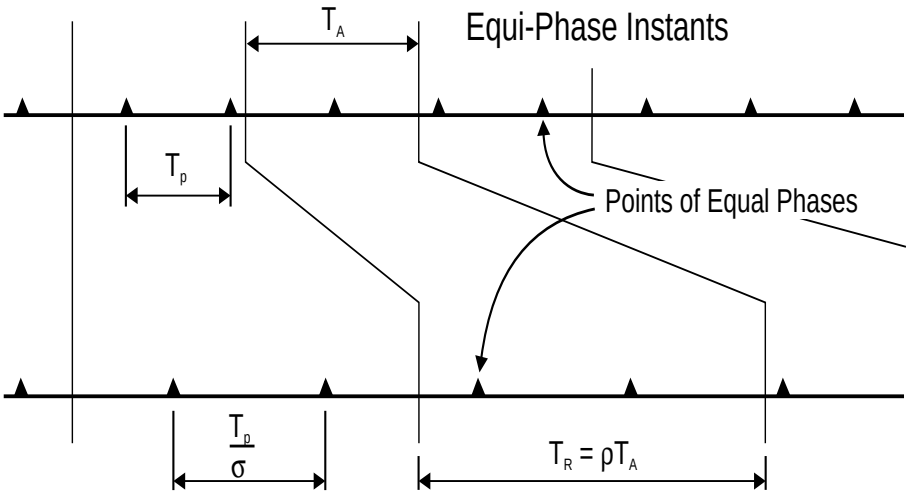

Figure 8: *Example of a set of equi-phase instants*

Such an integer multiple will almost never fall on a subsequent frame boundary, so the measured frequency information does not suggest that phase matching will occur at the end of a frame if matching has been ensured at the start of the frame. However, if $\phi_k^m(t)$ is formed by simply copying the pre-existing method [3, 4] reported in §4.2 but with T_A replaced by T_R, then $\phi_k^m(t)$ will be formed such that this end-of-frame phase-matching does occur. Introducing this discrepancy in $\phi_k^m(t)$ between phase and frequency information leads to distortion.

Instead, it turns out that it is necessary to be more sophisticated in the generation of $\phi_k^m(t)$ by choosing points, that are not equal to frame boundaries, but when used to force phase matching between analysis and reconstruction frames also lead to consistency between measured frequency information and the frequency information implicit in the $\phi_k^m(t)$ generated. These new points are calculated according to a relative offset T_E^m from the start of the m'th reconstruction frame.

To develop this idea, as just highlighted a vital point in reducing phase dispersion on voiced frames is the recognition that phases should match time points which are multiples of the pitch period T_p. In this paper, these points are denoted as 'equi-phase instants' and the idea of them is illustrated in 8. This figure presents the most general case of combined time and pitch scaling whereby the pitch period of the scaled signal on the reconstruction frame has been altered by the pitch scaling factor σ.

The black triangles represent corresponding points on analysis and reconstruction frames which should be

equi-phase. Note that there is no absolute starting location for the equi-phase points, but once one pair of points (between analysis and reconstruction frames) are selected to be equi-phase, this generates a whole family of such points on both the analysis and reconstruction frames; the sets being generated on the analysis frame by adding integer multiples of T_p to the original point, and on the reconstruction frame by adding integer multiples of T_p/σ to the original equi-phase point.

To explain how the analysis and reconstruction phases are ensured to be co-incident at these equi-phase instants, consider the case of a single analysis frame as shown in figure 9 where, to increase the clarity of explanation, the case $\sigma = 1$ of no pitch scaling is considered and it is also assumed that the first equi-phase instant on the m'th analysis and reconstruction frames co-incide with each other, and the frame boundary. Now,

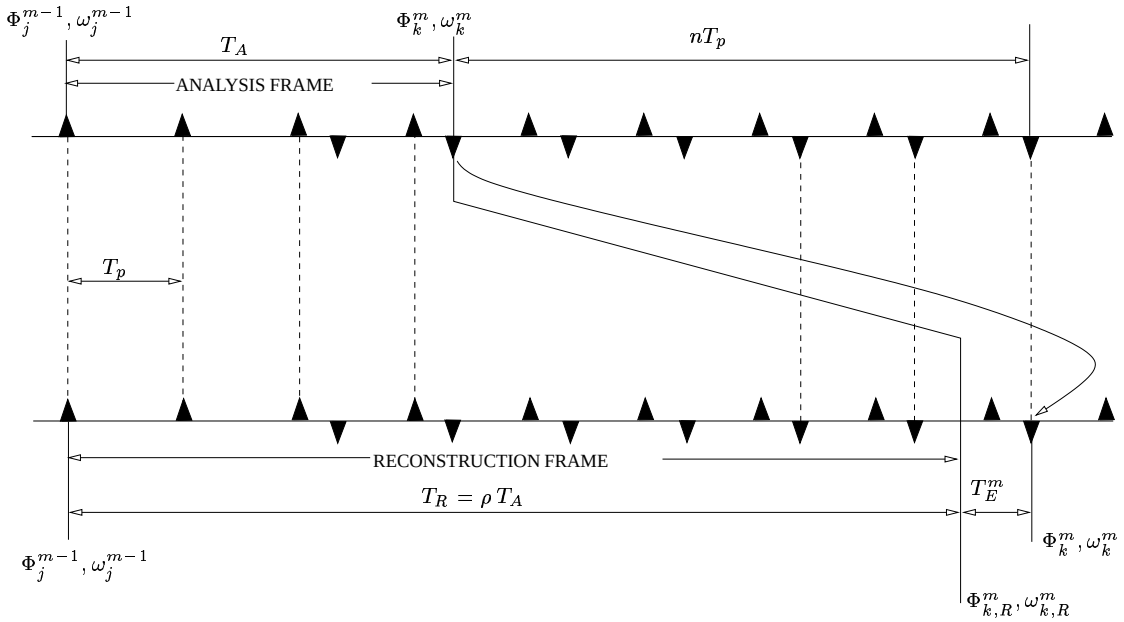

Figure 9: *Interpolation criteria for generation of equi-phase* $\phi_k^m(t)$

this first equi-phase instant on the m'th reconstruction frame boundary generates a family of equi-phase instants at multiples of the pitch period T_p on the reconstruction frame. These are marked with upright black triangles, and since no pitch scaling is involved, they correspond in a one-to-one fashion with equi-phase points (upright black triangles) on the analysis frame.

At the start of the $m-1$'st analysis frame, the measured starting phase Φ_j^{m-1} and starting frequency ω_j^{m-1} for the j'th track (which is matched to the k'th track on the subsequent m'th frame) is available. Since this point happens to be an equi-phase point with one on the reconstruction frame (both are upright black triangles), when generating the interpolant $\phi_k^m(t)$ on that frame, the starting conditions for that interpolant will be the same, and are shown marked that way at the start of the reconstruction frame.

The difficulties begin to arise at the other end of the frames. Specifically, if the analysed (and hence reconstructed) signal is strongly voiced with pitch period T_p, then under the assumption that the frequency track is relatively constant, since we are electing to force the original and reconstructed phases to match at the common equi-phase instant at the start of both frames, they will continue to match at multiples of this pitch period T_p (the matching of phase indicated by the dashed lines connecting the equi-phase points).

However, the end of the analysis frame *is not an equi-phase point.* Therefore, if we try to choose the interpolant $\phi_k^m(t)$ to be such as to match the ending reconstruction phase with the ending phase Φ_k^m on the analysis frame, then there is a fundamental contradiction between this and the measured frequency information.

The solution we have used in our new algorithm is to instead choose an ending phase for the interpolant

$\phi_k^m(t)$ that is consistent with the measured phase on the analysis frame *and* the measured frequency. This is achieved, as shown in figure 9 by identifying the points in the reconstruction frame which are equi-phase with the end of the analysis frame; these are shown as upside-down black triangles. We then choose the interpolant $\phi_k^m(t)$ to match the end-frame phase Φ_k^m at the point on the reconstruction frame which is closest to a point which is equi-phase with the end of the analysis frame.

The reasoning is that an interpolant $\phi_k^m(t)$ formed in this way that is consistent with equi-phaseness at the frame boundaries, is most likely to preserve all the equi-phase instants within the frame boundaries, and hence minimise the total phase dispersion across the whole frame.

The point for matching is specified by an offset T_E^m from the end of the reconstruction frame (it can be negative) which from figure 9 is calculated as

$$T_E^m = T_A + nT_p - T_R$$

where n is the integer that minimises $|T_E^m|$. In fact, the requirement of making $\phi_k^m(T_R + T_E^m)$ match Φ_k^m is only approximately achieved by electing to instead reason that since at the end of the frame

$$\mathrm{d}\phi_k^m(t) \approx \omega_k^m \mathrm{d}t$$

then selecting an end-of-reconstruction-frame phase $\Phi_{k,R}^m$ of

$$\Phi_{k,R}^m = \Phi_k^m - \omega_k^m T_E^m$$

will achieve the required equi-phase result at $t = T_R + T_E^m$ if $\phi_k^m(t)$ is chosen such that $\phi_k^m(T_R) = \Phi_{k,R}^m$.

This explains the fundamental idea of our equi-phase method for a simple case, however in general the start of the analysis and reconstruction frames are not equi-phase; for example, the $m+1$'st frames in figure 9 do not have equi-phase starts. This more general case is illustrated more fully in figure 10. The same methods as used

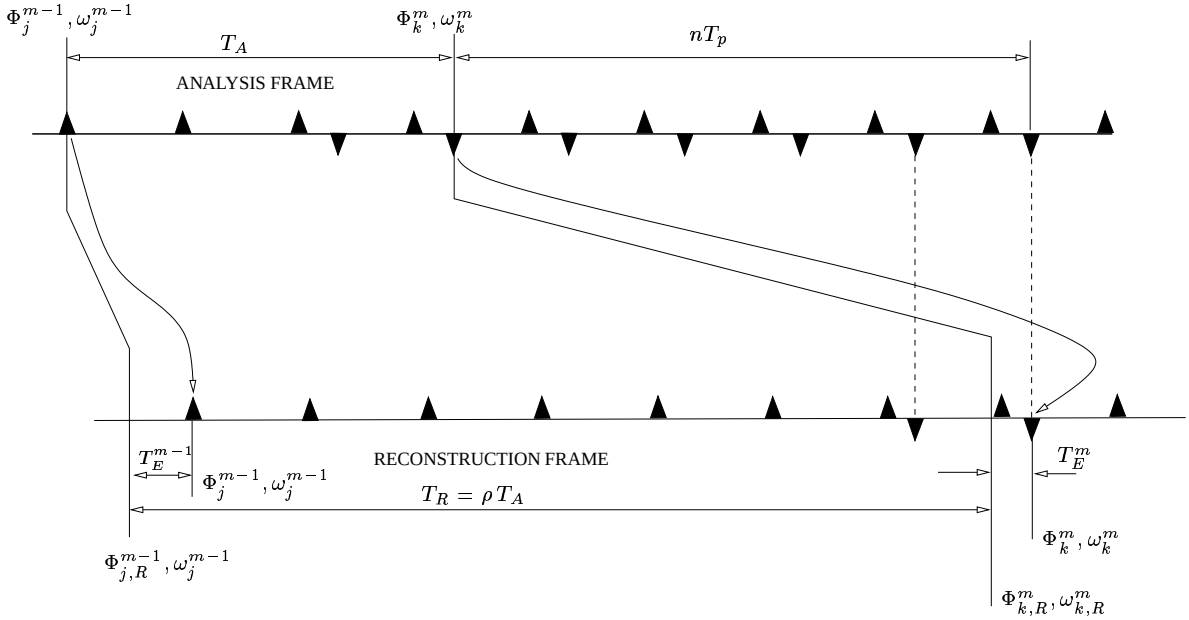

Figure 10: *Interpolation criteria for generation of equi-phase $\phi_k^m(t)$ when initial frame starts are not equi-phase*

in the case illustrated in figure 9 may still be applied save that now the calculation of the offset T_E^m depends on the offset T_E^{m-1} that was used on the previous frame as

$$T_E^m = T_E^{m-1} + T_A + nT_p - T_R$$

and also the interpolation condition for the start of the reconstruction frame becomes $\phi_k^m(0) = \Phi_{j,R}^{m-1}$.

The final level of generality that can be added is to also consider the case where pitch scaling is also required so that $\sigma \neq 1$. Again, the same basic technique is applicable, and it is illustrated in figure 11. There it

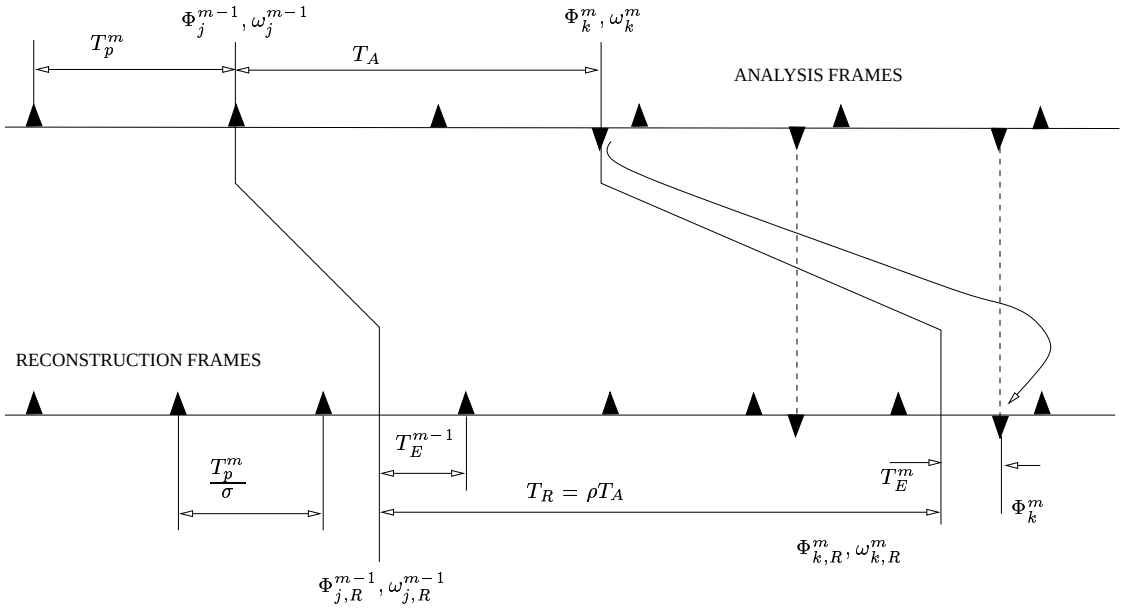

Figure 11: *Interpolation criteria to achieve phase-invariant scaling for the most general case which includes pitch scaling.*

is illustrated that firstly, the (matched track) start and end frequencies $\omega_{j,R}^{m-1}, \omega_{k,R}^m$ on the reconstruction frame are derived simply by applying the pitch scaling factor σ as

$$\omega_{j,R}^{m-1} = \sigma\omega_j^{m-1} \qquad \omega_{k,R}^m = \sigma\omega_k^m.$$

Next, as is illustrated in figure 11, the calculation of T_E^m is modified by the value of σ as

$$T_E^m = T_E^{m-1} + \frac{T_A + nT_P^m}{\sigma} - T_R,$$

where, as before, the integer n is chosen to minimise $|T_E^m|$. Finally, the equi-phase compensated phases at the end frame boundary are calculated as for simpler cases, save that now the pitch-modified frequency is used as

$$\Phi_{k,R}^m = \Phi_k^m - \omega_{k,R}^m T_E^m.$$

In summary then, when using our new phase-invariant method that works by the identification of so-called 'equi-phase' instants, for every k'th track on the m'th frame (matched to the j'th track on the $m-1$'st frame), start and finish amplitudes A_j^{m-1}, A_k^m, start and finish frequencies $\omega_{j,R}^{m-1}, \omega_{k,R}^m$ and start and finish phases $\Phi_{j,R}^{m-1}, \Phi_{k,R}^m$ are all specified for this m'th frame. The time/pitch-scale modified signal is then actually reconstructed using all this information as

$$s_R^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos \phi_k^m(t) \quad 0 \leq t \leq T_R, \tag{19}$$

where, the amplitude $A_k^m(t)$ is a linear interpolant

$$A_k^m(t) = \frac{T_R - t}{T_R} A_j^{m-1} + \frac{t}{T_R} A_k^m \tag{20}$$

while $\phi_k^m(t)$ is a cubic interpolant of the form

$$\phi_k^m(t) = at^3 + bt^2 + ct + d.$$

This is of the same form as reviewed earlier in §4.2, except that now the interpolation constraints at the start of the reconstruction frame are

$$d = \Phi_{j,R}^{m-1}, \qquad c = \omega_{j,R}^{m-1}.$$

while the constraints at the end of the reconstruction frame of $\phi_k^m(T_R) = \Phi_{k,R}^m + 2\pi M_k^m$ and $\mathrm{d}\phi_k^m(T_R)/\mathrm{d}t = \omega_{k,R}^m$ lead to

$$\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \dfrac{3}{T_R^2} & -\dfrac{1}{T_R} \\ -\dfrac{2}{T_R^3} & \dfrac{1}{T_R^2} \end{bmatrix} \begin{bmatrix} \Phi_{k,R}^m - \Phi_{j,R}^{m-1} - \omega_{j,R}^{m-1} T_R + 2\pi M_k^m \\ \omega_{k,R}^m - \omega_{j,R}^{m-1} \end{bmatrix},$$

where M_k^m is the integer closest to

$$M_k^m = \left[\frac{1}{2\pi} \left((\omega_{k,R}^m - \omega_{j,R}^{m-1}) \frac{T_R}{2} + \Phi_{j,R}^{m-1} + \omega_{j,R}^{m-1} T_R - \Phi_{k,R}^m \right) \right].$$

This method of time/pitch scaling was tested and found to provide good quality transformations. There was noticeably less distortion appearing as a chorusing effect than for the original method.

In relation to the developments of this section, the work in [8] has also provided a solution for reducing the phase dispersion that exists in the methods of [4], and the techniques described there also involve estimating an offset (denoted as t_0 in [8]) from the analysis frame boundary. However, beyond these superficial similarities, the actual solutions proposed here and in [7] differ in three important aspects.

Most fundamentally, the work [8] only applies the offset t_0 to modification of the excitation phase. In contrast, this paper applies the offset to the total system phase variation, and the reason for this in terms of ensuring that the (average) instantaneous frequency is consistent with that implicit in the total phase variation has been derived and explained in § 4.3 and §5.3.

Secondly, in[8] the offset t_0 is defined as the time at which 'a pitch pulse occurs when all of the sine waves add coherently' while the algorithms developed here do not depend on the excitation sine waves ever adding coherently. The offset quantity T_E^m while thus appearing superficially similar to t_0 of [8] is in truth quite different in that instead of quantifying a point where an ensemble of sine waves have coherent phase, it quantifies a point where for a single sine wave it is appropriate to attempt phase interpolation in the sense that it is only sensible to try to match phases of sine waves on separate frames at integer multiples of their fundamental frequency.

Finally, the work [8] employs a separate source/filter decomposition so that during rate-change, excitation and system phase trajectories are modified differently. As explained in §5.3 this work argues against such a decomposition and advocates the modification of a single cubic interpolant $\phi_k^m(t)$ for the total phase contribution which, during rate and/or pitch modification, is altered in such a way as to preserve pitch synchronicity and the relationship between excitation frequency variation and total phase variation.

In practice, we have found the methods of [8] provide results that are of a similar quality to those presented in this paper, and both are superior to the pre-existing techniques of [3, 4]. However, it appears that these comparable results are obtained with a computational load which is somewhat higher for the algorithms of [8] than for those of this paper; although this latter difference is only significant during pitch changes in which the methods of [8] estimate separate excitation and system components.

7 Experimental Results

This section profiles the performance of the new phase-invariant algorithm of this paper by illustrating its time scaling performance on a 4 second duration, 8kHz sampled record of the speech of a female news-reader which is shown as the upper plot in figure 12. Shown as the lower plot in that same figure is the speech time-scaled by a factor $\rho = 1.8$. Clearly the key features of the speech are retained, but lengthened under the time scaling operation. Greater detail is shown in figure 13 where only the first 0.4 seconds of the original speech and its $\rho = 1.8$ scaled version are shown.

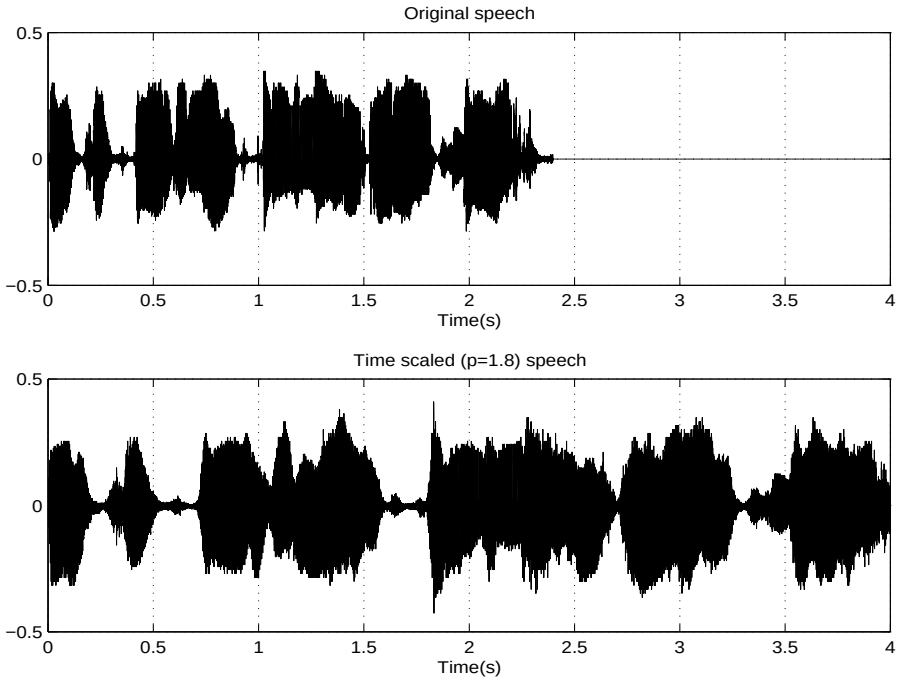

Figure 12: *Top plot is 8Khz sampled speech of a female news-reader saying 'searchers found a note from the men earlier today'. Bottom plot is same sample time scaled by factor* $\rho = 1.8$ *using the new phase-invariant method introduced in this paper.*

In figure 14 the same data as shown in figure 12 is considered, but this time in the frequency domain with the top plot being the spectrogram for the original speech, and the bottom plot being the spectrogram of the $\rho = 1.8$ time scaled speech. Note the clear indication of spectral 'tracks' which have been preserved in frequency, but had their time duration stretched. To further illustrate this concept, the frequency tracks identified by the algorithm over the first forty analysis frames are shown in figure 15. In that figure, a cross indicates an identified frequency component, and lines between crosses represent identified frequency tracks. Tracks (lines) are initiated when an un-matchable component arises by extrapolating it back to a previous frame with an amplitude linearly increasing from zero on the previous frame; hence the occurrence of tracks with no cross on their left most end.

To further highlight the utility of the phase invariant approach pursued in this paper, it is profiled on a synthetic example of a pure tone at 160Hz, which after some time (40 ms) is joined by a second tone at 483 Hz which has a slowly varying phase offset. This test signal is shown in figure 16.

Shown in the left-most diagrams of figure 17 is the $\rho = 1.8$ time scaled version of this signal when the pre-existing sinusoidal model based methods of [3, 4] are employed. Clearly, a lack of perfect reconstruction of the relative phases of the two sinusoids has led to a significant change in overall wave-shape in the time-scaled

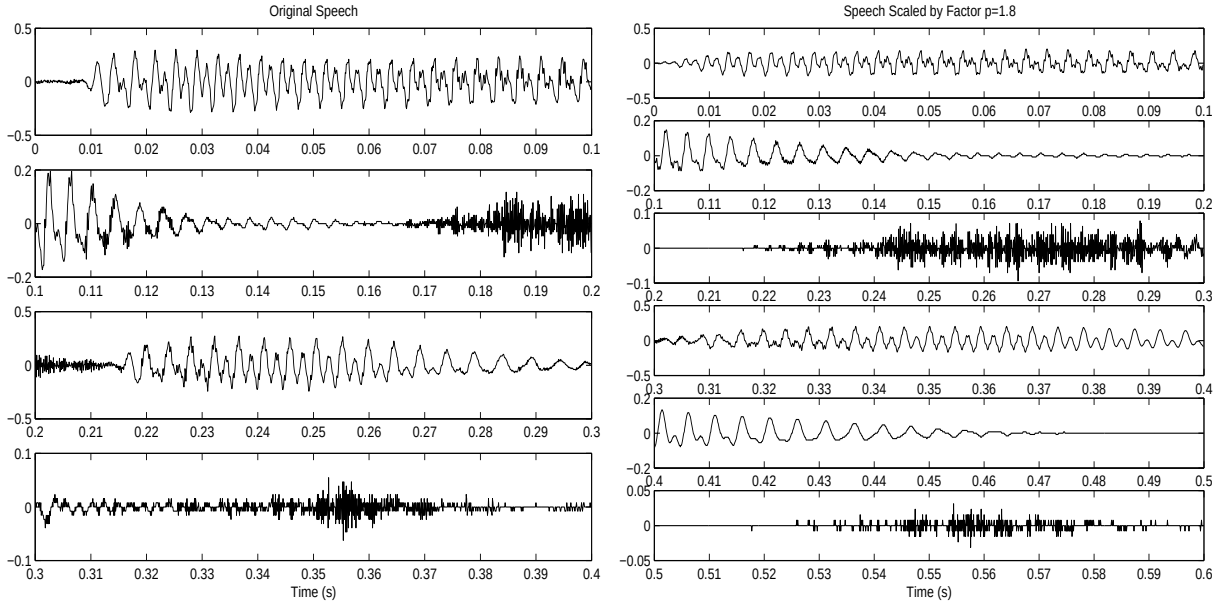

Figure 13: *Same data as in previous figure 12, but first 0.4 (0.7) seconds expanded to show (left) detail of original speech and (right) time scaled by factor* $\rho = 1.8$ *speech.*

signal.

In contrast, the $\rho = 1.8$ time scaled signal which is generated using the phase-invariant technique of this paper is shown in the right-most diagrams of figure 17 to possess a wave-shape that, despite being time scaled, very closely matches that of the original signal of figure 16.

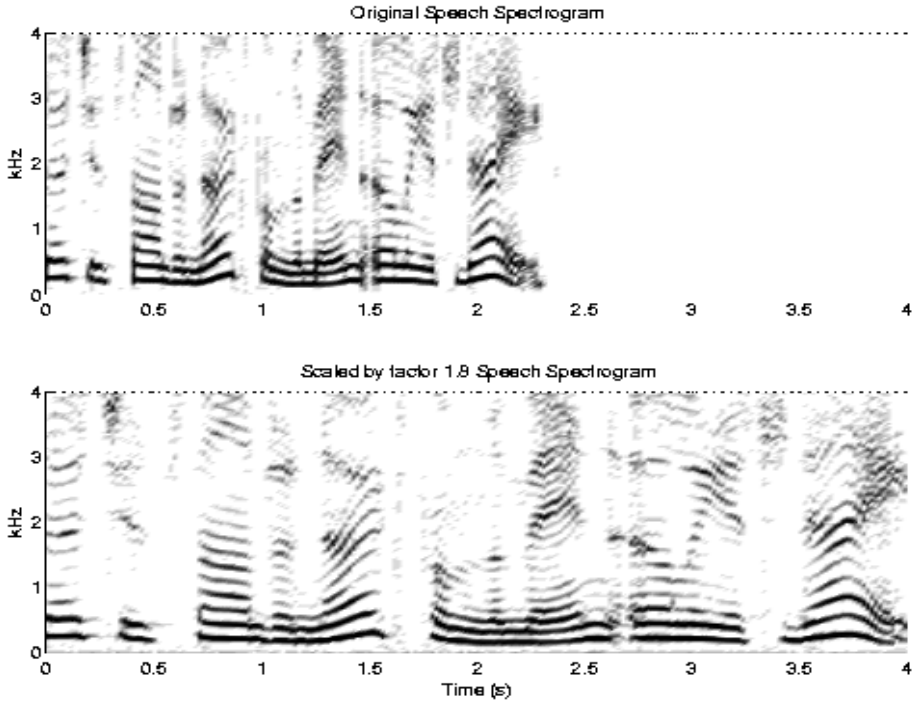

Figure 14: *Spectrograms of original and factor $\rho = 1.8$ time scaled speech. Sample is a female news-reader saying 'searchers found a note from the men earlier today'*

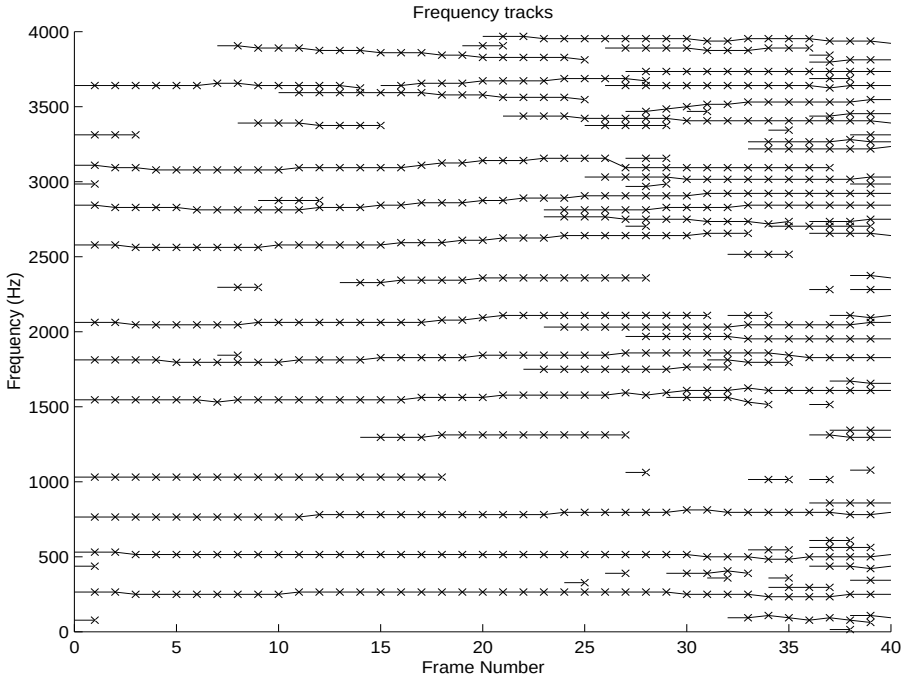

Figure 15: *Frequency tracks identified on first 40 frames of sound sample shown in figure 12*

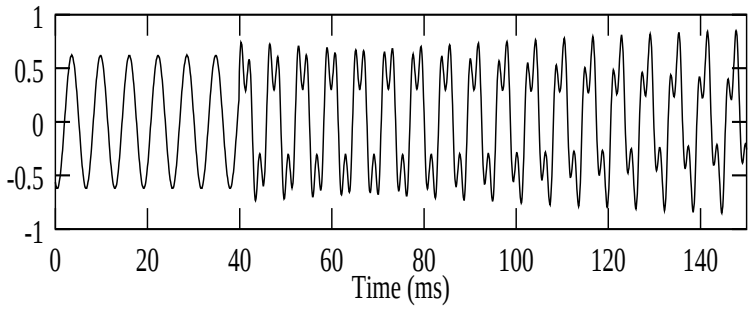

Figure 16: *Synthetically generated test signal consisting of tones at 160Hz and (after 40ms) 483 Hz, with the latter tone having a slowly time varying phase offset.*

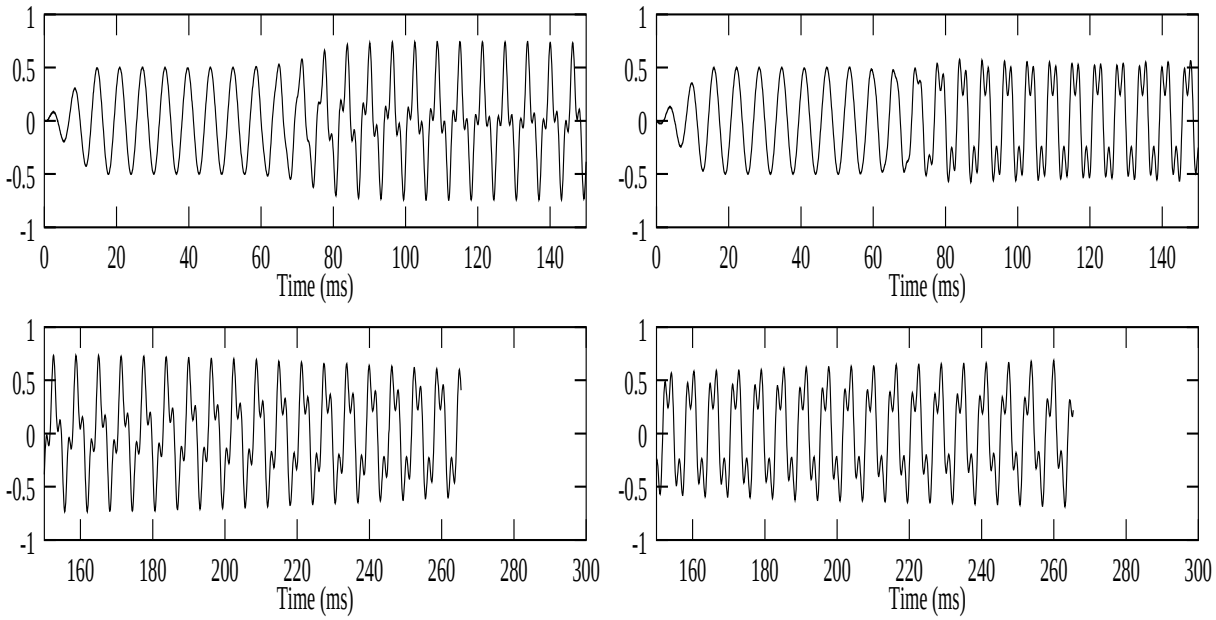

Figure 17: *The two plots on the left show the $\rho = 1.8$ time scaled version of the signal in figure 16 when pre-existing sinusoidal model-based methods are used. The two plots on the right shown that $\rho = 1.8$ time scaled version of the signal in figure 16 when the new phase-invariant method developed in this paper is used.*

8 Conclusion

The scaling method presented in this paper uses parametric modelling techniques to achieve independent time and pitch scaling of audio signals. By addressing the phase dispersion defect in pre-existing frequency domain based scaling methods it provides better quality transformations. This method also has a computational advantage as it does not require the decomposition of the signal into excitation and vocal tract responses. Benchmarking of the algorithm showed that this feature delivered a 50% improvement in execution time. The scaling method has been implemented in real time on a custom designed portable signal processor based on a single Texas Instruments TMS320C31, and this has allowed testing and demonstration of the method in a real-world environment.

References

[1] D. W. GRIFFIN AND J. S. LIM, *Signal estimation from modified short-time fourier transform,* IEEE Transactions on Acoustics, Speech, and Signal Processing, 32 (1984), pp. 236–243.

[2] E. HARDAM, *High quality time scale modification of speech signals using fast synchronised-overlap-add algorithms,* in Proceedings of the IEEE International Conference on Acustics, Speech and Signal Processing, 1990, pp. 409–412.

[3] R. MCAULAY AND T. QUATIERI, *Speech analysis-syntheses based on a sinusoidal representation,* IEEE Trans. Acoust., Speech., Signal Proc., ASSP-34 (1986), pp. 744–754.

[4] ——, *Speech transformations based on a sinusoidal representation,* IEEE Trans. Acoust., Speech., Signal Proc., ASSP-34 (1986), pp. 1449–1464.

[5] ——, *Pitch estimation and voicing detection based on a sinusoidal model,* in Proc IEEE Int. Conf. Acoust., Speech, Sig Proc., Apr. 1990.

[6] E. MOULINES AND J. LAROCHE, *Non-parametric techniques for pitch-scale and time-scale modification of speech,* Speech Communication, 16 (1995), pp. 175–205.

[7] M.R.PORTNOFF, *Time-scale modification of speech based on short-time fourier analysis,* IEEE Trans. Acoust., Speech., Signal Proc., ASSP-30 (1981), pp. 374–390.

[8] T. QUATIERI AND R. MCAULAY, *Shape invariant time-scale and pitch modification of speech,* IEEE Trans. Acoust., Speech., Signal Proc., ASSP-40 (1992), pp. 497–510.

[9] S. ROUCOS AND A. M. WILGUS, *High quality time scale modification for speech,* in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1985, pp. 493–496.

[10] W. SMITH AND J. SMITH, *Handbook of Real-Time Fast Fourier Transforms,* IEEE Press, 1995.

[11] W. VERHELST AND M. ROELANDS, *An Overlap-Add technique based on waveform similarity (wsola) for high quality time-scale modification of speech,* in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1993, pp. 554–557.

[12] J. WAYMAN, R. E. REINKE, AND D. WILSON, *High quality speech expansion, compression, and noise filtering using the SOLA method of time scale modification,* in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1989, pp. 714–717.