Time-Scale Modification of Speech Signals

Brett Ninness and Soren John Henriksen

***Abstract*—This paper presents 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) modeling techniques introduced by other authors, but new perspectives on the role of vocal tract decomposition and maintaining phase relationships between sinusoidal tracks are derived that achieve improved output quality with decreased computational load. Simulation results are provided to illustrate performance, and the algorithms developed here have been demonstrated capable of implementation on simple DSP hardware.**

***Index Terms*—Speech analysis, speech processing, time scale modification.**

I. 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 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 [9] for a comprehensive survey.

Some of these methods are based on time domain splicing/overlap-add approaches [9], [13], [3], [14], 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 [9]. The work in this paper examines another approach, and uses ideas that can be traced back at least to 1981 [10] 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 [9], [7], [2], [12]. 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 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-scale modification methods that provide improved quality output by addressing the phase dispersion problem, while at the same time introducing important modifications that significantly 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.

The phase dispersion issue has been addressed before in [11], with the results obtained there being considered of benchmark quality. There, a key principle is that of "pitch pulse onset time," defined as being when all excitation sine waves add coherently, and "shape invariant" transformations are obtained that preserve these coherencies after appropriate scalings. That is, the input excitation is scaled.

This paper takes a different approach of directly constructing the output. The phase functions of individual frequency tracks are directly constructed in order to minimizephase dispersion via the idea of "equi-phase" instants which are defined on a per-track basis. As will be illustrated on an example, this has been found to achieve results commensurate with those obtainable by [11], but with decreased computational load due to avoiding the need for estimating vocal tract response, and by employing a simple new pitch estimation method.

Finally, it is important to acknowledge that while the work in [11] is most closely related to this paper, the importance of phase dispersion in time and frequency scale modification is widely recognised. See, for example the recent contributions [5], [4], [1] which also discuss the problem and provide solutions alternate to those developed here.

II. Sinusoidal Model Estimation

The algorithms developed in this paper are based on the work [6], [7], [11] in which a source-filter is used for the speech production process, and a sum-of-sinusoids description is used for the source model. Within this framework, the resultant speech signal 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

Manuscript received September 19, 2006; revised August 20, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mariane R. Petraglia.

The authors are with the School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan 2308, NSW, Australia (e-mail: brett.ninness@newcastle.edu.au; eesjh@ee.newcastle.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2007.909350

1053-587X/$25.00 © 2008 IEEE

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.

In order to use this 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 [6], [7]) that the time variation is approximately piecewise constant over sufficiently short durations called "analysis frames" of length T seconds.

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. The magnitude and phase of the Fourier Transform at these measured frequencies are used as estimators, respectively, of A_k and Φ_k.

This procedure determines sinusoidal models at a rate equal to the analysis frame rate. For subsequent signal generation (and hence rate/pitch modifications), time varying estimates $A_k^m(t), \omega_k^m(t)$ valid for the mth analysis frame, and which vary with t discretized at the original signal sampling rate f_s are required. To generate these, this paper uses a method developed in [6], [7], [11] in which the concept of "frequency tracks" is used, and smooth interpolants $A_k^m(t), \omega_k^m(t)$ sampled at rate f_s are created between estimates associated with common tracks occurring at the frame rate.

To generate these, note that they should lead to a successful signal reconstruction as

$$\begin{aligned} s_R^m(t) &= \sum_{k=1}^{N(m)} A_k^m(t)\cos\left(\phi_k^m(t) + \Phi_k^m\right) \\ \phi_k^m(t) &\triangleq \int_0^t \omega_k^m(\xi)\mathrm{d}\xi. \end{aligned} \tag{2}$$

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 kth frequency track on frame m is matched to the jth track on frame $m+1$)

$$A_k^m(0) = A_k^m, \quad A_k^m(T_A) = A_j^{m+1} \tag{3}$$

and this may be simply achieved by the linear interpolant

$$A_k^m(t) = \left(\frac{T_A - t}{T_A}\right) A_k^m + \left(\frac{t}{T_A}\right) A_j^{m+1}. \tag{4}$$

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

$$\begin{aligned} &\phi_k^m(0) = 0, \\ &\phi_k^m(T_A) + \Phi_k^m = \Phi_j^{m+1} + 2\pi M_k^m \end{aligned} \tag{5}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\,|_{t=0} = \omega_k^m, \quad \frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\,|_{t=T_A} = \omega_j^{m+1} \tag{6}$$

where M_k^m is an integer constant which allows for phase unwrapping. Previous developers [6] have recognized 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

$$\begin{aligned} \phi_k^m(t) &= at^3 + bt^2 + ct + d, \\ \frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t) &= 3at^2 + 2bt + c. \end{aligned} \tag{7}$$

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

$$d = 0, \quad c = \omega_k^m. \tag{8}$$

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

$$\Phi_j^{m+1} + 2\pi M_k^m = aT_A^3 + bT_A^2 + \omega_k^m T_A + \Phi_k^m \tag{9}$$

$$\omega_j^{m+1} = 3aT_A^2 + 2bT_A + \omega_k^m \tag{10}$$

which may be expressed as

$$\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \frac{3}{T_A^2} & -\frac{1}{T_A} \\ -\frac{2}{T_A^3} & \frac{1}{T_A^2} \end{bmatrix} \begin{bmatrix} \Phi_j^{m+1} - \Phi_k^m - \omega_k^m T + 2\pi M_k^m \\ \omega_j^{m+1} - \omega_k^m \end{bmatrix}. \tag{11}$$

III. A New Interpretation of M_k^M

As aforementioned, the integer parameter M_k^m in (5) and (11) is included to provide "phase unwrapping," and previous workers [6], [7], [11] 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.

First, the average frequency of a track on the mth analysis frame is given by

$$\omega_{av}^m = \frac{1}{T_A}\int_0^{T_A} \omega_k^m(t)\mathrm{d}t \tag{12}$$

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)

$$\begin{aligned} \omega_{av}^m &= \frac{1}{T_A}\int_0^{T_A} \frac{\mathrm{d}\phi_k^m(t)}{\mathrm{d}t}\mathrm{d}t \\ &= \frac{1}{T_A}\left[\phi_k^m(T_A) - \phi_k^m(0)\right]. \end{aligned} \tag{13}$$

Substituting in the interpolation constraints in (5) then leads to

$$\omega_{av}^m = \frac{1}{T_A}\left[\left(\Phi_j^{m+1} + 2\pi M_k^m\right) - \Phi_k^m\right]. \tag{14}$$

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

$$\omega_{av}^m = \frac{\omega_k^m + \omega_j^{m+1}}{2}. \tag{15}$$

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:

$$M_k^m = \frac{1}{2\pi}\left[\frac{T_A}{2}\left(\omega_k^m + \omega_j^{m+1}\right) + \Phi_k^m - \Phi_j^{m+1}\right]. \quad (16)$$

This is identical to the value for M_k^m obtained in [6] according to a criterion of maximally smooth phase function $\phi_k^m(t)$. Note that since M_k^m must be an integer, (16) 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 (16) is much more than one that achieves a certain heuristically reasonable, but otherwise seemingly nonessential, goal of maximally smooth phase. In fact, the choice (16) also makes the nature of the interpolated $\phi_k^m(t)$ consistent with the measured frequency information and track matching choices encoded in $\omega_k^m, \omega_j^{m+1}$.

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.

IV. Rate and Pitch 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 [7]. This is in preparation for work in subsequent sections that will illustrate how preexisting methods may be improved while also lowering computation overheads.

A. Time Scaling

The process of time scaling a sinusoidally modeled signal by a factor ρ involves the length T_A of analysis frame and length T_R of reconstruction frame to be related as

$$T_R = \rho T_A. \quad (17)$$

This implies that 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

$$s_S^m(t') = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t'}{\rho}\right)\cos\left(\rho\,\phi_k^m\left(\frac{t'}{\rho}\right) + \Phi_k^m\right). \quad (18)$$

This function has been derived over one (the mth) 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. Note that, as will become more evident later, the phase term ϕ_k^m has been multiplied by ρ in order to preserve instantaneous frequency (the derivative of phase) during time scaling.

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 when time scaling by a factor ρ is introduced, then via (5) the argument to the cosine in (18) at $t' = T_R = \rho T_A$ is

$$\begin{aligned}\rho\phi_k^m(T_A) + \Phi_k^m &= \rho\left(\Phi_j^{m+1} + 2\pi M_k^m\right) + (1-\rho)\Phi_k^m, \quad (19)\\ &= \rho\phi_j^{m+1}(0) + \Phi_j^{m+1} + \rho\,2\pi M_k^m\\ &\quad + (1-\rho)\left(\Phi_k^m - \Phi_j^{m+1}\right) \quad (20)\end{aligned}$$

(recall from the interpolation condition (5) that $\phi_j^{m+1}(0) = 0$) and since, according to the ρ scaled reconstruction (18), the matched jth 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.

The solution proposed in [7] is to add a further phase term γ_k^m to the reconstruction (18) so that it becomes

$$s_S^m(t') = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t'}{\rho}\right)\cos\left(\rho\,\phi_k^m\left(\frac{t'}{\rho}\right) + \Phi_k^m + \gamma_k^m\right) \quad (21)$$

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

$$\gamma_j^{m+1} = \rho\phi_k^m(T_A) + \Phi_k^m - \Phi_j^{m+1}. \quad (22)$$

Unfortunately, the reconstruction (21) with the γ_k^m offset (designed to preserve phase continuity) destroys the phase information Φ_k^m in the reconstructed signal when $\rho \neq 1$.

This results in "phase dispersion" (between sines) 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'.

A key contribution of this paper is to develop a new phase invariant method specifically designed to address this defect and, hence, improve the perceived quality of the time/pitch scaled signal.

B. Pitch Scaling

While not studied empirically in this paper, the process of pitch scaling involves every frequency track being scaled by the same constant amount σ so that the reconstructed pitch-scaled signal may be represented as

$$s_P^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos\left(\sigma \phi_k^m(t) + \Phi_k^m\right). \qquad (23)$$

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

$$\begin{aligned}\sigma\phi_k^m(T) + \Phi_k^m = \sigma\phi_j^{m+1}(0) + \Phi_j^{m+1}\\ +\sigma 2\pi M_k^m + (1-\sigma)\left(\Phi_k^m - \Phi_j^{m+1}\right). \quad (24)\end{aligned}$$

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

$$s_P^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos\left(\sigma\phi_k^m(t) + \Phi_k^m + \gamma_k^m\right) \qquad (25)$$

where the phase offset γ_k^m is calculated as

$$\gamma_j^{m+1} = \sigma\phi_k^m(T_A) + \Phi_k^m - \Phi_j^{m+1}. \qquad (26)$$

As in the case of time scaling, this simple existing approach suffers from some phase dispersion, which can be reduced using the new phase invariant method described in Sectioin V-B.

C. The Role of Source-Filter Decomposition

Previous work [6], [7] on pitch/rate modification using sinusoidal descriptions for vocal excitation has stressed the importance of a source/filter representation 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 Section V following, and the overview of preexisting ideas presented in Section IV-A, IV-B 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 mth analysis frame, then its response may be written as

$$\left|H\left(j\omega_k^m, t\right)\right| = B_k^m, \quad \angle H\left(j\omega_k^m, t\right) = \theta_k^m. \qquad (27)$$

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

$$\psi_k^m(t) = \int_0^t \omega_k^m(\xi)\, \mathrm{d}\xi \qquad (28)$$

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

$$\begin{aligned}\theta_k^m(0) + \Lambda_k^m &= \Phi_k^m,\\ \theta_k^m(T_A) + \Lambda_k^m &= \theta_j^{m+1}(0) + \Lambda_j^{m+1} = \Phi_j^{m+1}. \quad (29)\end{aligned}$$

In this case, the pitch/scale invariant reconstruction (2), on frame m becomes

$$s_R^m(t) = \sum_{k=1}^{n} A_k^m(t) \cos\left(\psi_k^m(t) + \theta_k^m(t) + \Phi_k^m\right). \qquad (30)$$

The thinking behind the need for a source/filter decomposition in [6], [7] is that when (for example) a time-scale modification occurs, the phase behavior 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

$$s_R^m(t') = \sum_{k=1}^{n} A_k^m\left(\frac{t}{\rho}\right) \cos\left(\rho\psi_k^m\left(\frac{t}{\rho}\right) + \theta_k^m\left(\frac{t}{\rho}\right) + \Phi_k^m\right). \qquad (31)$$

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

$$\Delta_1 = \rho\left(\Lambda_j^{m+1} - \Lambda_k^m\right) + \rho\left(\theta_j^{m+1} - \theta_k^m\right) + \rho 2\pi M_k^m. \quad (32)$$

On the other hand, when the source/filter decomposition is ignored, then according to (20) the total phase variation Δ_2 across a reconstruction frame is

$$\Delta_2 = \rho\left(\Phi_j^{m+1} - \Phi_k^m\right) + \rho 2\pi M_k^m. \qquad (33)$$

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 (22) 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 Section IV-A, IV-B 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

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\,|_{t=0} = \omega_k^m, \quad \frac{\mathrm{d}}{\mathrm{d}t}\phi_k^m(t)\,|_{t=T_A} = \omega_j^{m+1} \qquad (34)$$

while with the method [6], [7] involving a source/filter decomposition, only the time varying source excitation component $\psi_k^m(t)$ is cubically interpolated to satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_k^m(t)\,|_{t=0} = \omega_k^m, \quad \frac{\mathrm{d}}{\mathrm{d}t}\psi_k^m(t)\,|_{t=T_A} = \omega_j^{m+1} \tag{35}$$

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 Section IV-A, IV-B 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 [6], [7], [11], 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 in [11] provide a much simplified (compared to [6], [7]) 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.

V. An Improved Phase-Invariant Method

Having profiled existing methods of time/pitch scale modifications (together with some new interpretations of some 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 (ones containing only a few dominating sine wave components), phase dispersion producing distortion that may be described as "chorusing" is a major problem, that has been recognized and addressed in [11] and [9]. An attendant relatively high computational load is a further concern. To address the chorusing problem, the following Section V-A, IV-B provide a solution that involves reducing phase dispersion.

A. 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 nonharmonic 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 synchronizing 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 [8] 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 tradeoff between accuracy and complexity.

To proceed with the development, denote the peaks of the DFT on a particular analysis frame by the frequencies

$$f_1, f_2, f_3, \ldots, f_k \tag{36}$$

with associated magnitudes

$$M_1, M_2, M_3, \ldots, M_k \tag{37}$$

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

$$f_i = n_i f_p \quad \text{for } n_i \in \mathbf{Z}, \forall i \in 1 \ldots k. \tag{38}$$

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:

$$n_i^\star = \left\lfloor \frac{f_i}{f_p^\star} + \frac{1}{2} \right\rfloor \tag{39}$$

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

Using these quantities, the (weighted, by component magnitude) cumulative squared error in the choice of a pitch f_p on the current analysis frame may be defined as

$$e(f_p) \triangleq \sum_{i=1}^{k} M_i \left(\frac{f_i}{n_i^\star} - f_p \right)^2. \tag{40}$$

The utility of the specification of an initial estimate $f_p^\star$, and, hence, initial estimates $n_i^\star$ via (39) is that (40) is quadratic in f_p, so that the value $\hat{f}_p$ that minimises it (and, hence, the error of a pitch estimate) may be given in closed form as

$$\hat{f}_p = \frac{1}{M} \sum_{i=1}^{k} \frac{M_i f_i}{n_i^\star}, \quad M \triangleq \sum_{i=1}^{k} M_i. \tag{41}$$

It is this value $\hat{f}_p$ which is taken as the pitch estimate on the analysis frame, with $\hat{T}_p = 1/\hat{f}_p$ 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 minimizethe 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 Section V-B.

The pitch estimate (41) may be substituted back into the error expression (40), to generate a confidence index κ which is a normalized version of $e(\hat{f}_p)$ as follows:

$$\begin{aligned}\kappa &\triangleq e(\hat{f}_p)\left(\hat{f}_p\sum_{i=1}^{k}M_i\right)^{-1}\\ &=\left(\hat{f}_p\sum_{i=1}^{k}M_i\right)^{-1}\sum_{i=1}^{k}M_i\left(\frac{f_i}{n_i^\star}-f_p^\star\right).\end{aligned}\tag{42}$$

Note that this quantity indicates high confidence in voicing when it takes on low values. It 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 (39) 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 $\hat{f}_p$ 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 [11], [9].

B. A New Phase-Invariant Method

As explained in Section IV-A, a key limitation of preexisting methods for time/pitch scale modification [6], [7] is the distortion introduced by the lack of preservation of phase information between analysis and reconstruction frames.

The new methods presented in this section minimize phase dispersion by constraining the set of output phases at one point of time in the reconstruction period to match the phases in the original signal at one point of time 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.

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 preexisting methods [6], [7], [11] reported in Section IV-A), 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 the cubic interpolant $\phi_k^m(t)$ in (7) 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 (16) 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.

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 preexisting method [6], [7] profiled in Section II 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 mth 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 Fig. 1. 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 one family of corresponding points on analysis and reconstruction frames which share the same set of phases, and, hence, are termed equi-phase. Note that every point of time belongs to one these families, and once one

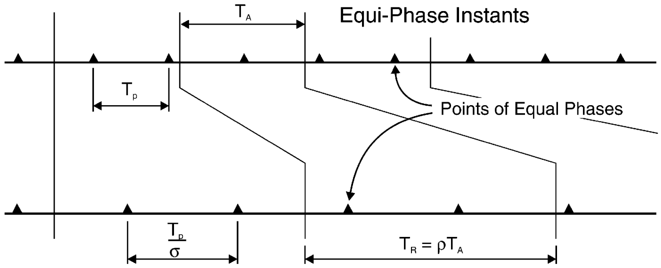

Fig. 1. Example of a set of equi-phase instants.

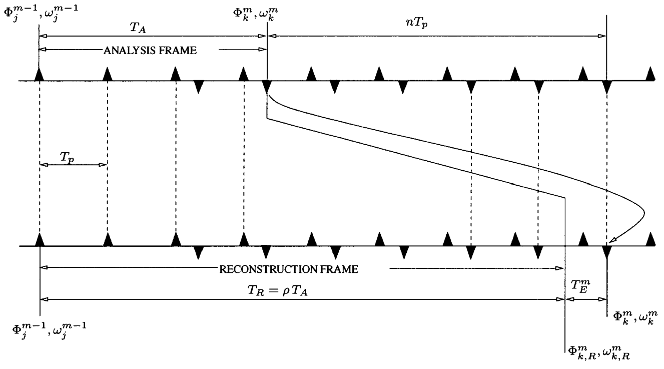

Fig. 2. Interpolation criteria for generation of equi-phase $\phi_k^m(t)$.

pair of points (between analysis and reconstruction frames) are selected to be equi-phase, this generates the 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 coincident at these equi-phase instants, consider the case of a single analysis frame as shown in Fig. 2 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 mth analysis and reconstruction frames coincide with each other, and the frame boundary. Now, this first equi-phase instant on the mth 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 $m-1$th analysis frame, the measured starting phase Φ_j^{m-1} and starting frequency ω_j^{m-1} for the jth track (which is matched to the kth track on the subsequent mth 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 equi-phase* with the end of the reconstruction frame. 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 Fig. 2 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 minimizethe 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 Fig. 2 is calculated as

$$T_E^m = T_A + nT_p - T_R \tag{43}$$

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

$$\mathrm{d}\phi_k^m(t) \approx \omega_k^m \mathrm{d}t \tag{44}$$

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

$$\Phi_{k,R}^m = \Phi_k^m - \omega_k^m T_E^m \tag{45}$$

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 $m+1$th frames in Fig. 2 do not have equi-phase starts.

The same methods as used in the case illustrated in Fig. 2 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

$$T_E^m = T_E^{m-1} + T_A + nT_p - T_R \tag{46}$$

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. First, 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

$$\omega_{j,R}^{m-1} = \sigma \omega_j^{m-1} \quad \omega_{k,R}^{m} = \sigma \omega_k^{m}. \tag{47}$$

Next, the calculation of T_E^m is modified by the value of σ as

$$T_E^m = T_E^{m-1} + \frac{T_A + nT_P^m}{\sigma} - T_R, \tag{48}$$

where, as before, the integer n is chosen to minimize$|T_E^m|$. 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

$$\Phi_{k,R}^m = \Phi_k^m - \omega_{k,R}^m T_E^m. \tag{49}$$

In summary then, when using our new phase-invariant method that works by the identification of so-called "equi-phase" instants, for every kth track on the mth frame (matched to the jth track on the $m-1$th 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 mth frame. The time/pitch-scale modified signal is then actually reconstructed using all this information as

$$s_R^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos \phi_k^m(t) \; 0 \leq t \leq T_R, \tag{50}$$

where, the amplitude $A_k^m(t)$ is the linear interpolant (4) while $\phi_k^m(t)$ is again a cubic interpolant of the form (7) except that now the interpolation constraints at the start of the reconstruction frame are

$$d = \Phi_{j,R}^{m-1}, \quad c = \omega_{j,R}^{m-1} \tag{51}$$

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

$$\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \frac{3}{T_R^2} & -\frac{1}{T_R} \\ -\frac{2}{T_R^3} & \frac{1}{T_R^2} \end{bmatrix} \times \begin{bmatrix} \Phi_{k,R}^m - \Phi_{j,R}^{m-1} - \omega_{j,R}^{m-1} T_R + 2\pi M_k^m \\ \omega_{k,R}^m - \omega_{j,R}^{m-1} \end{bmatrix} \tag{52}$$

where M_k^m is the integer closest to

$$M_k^m = \left[\frac{1}{2\pi} \left(\left(\omega_{k,R}^m - \omega_{j,R}^{m-1} \right) \frac{T_R}{2} + \Phi_{j,R}^{m-1} + \omega_{j,R}^{m-1} T_R - \Phi_{k,R}^m \right) \right]. \tag{53}$$

In relation to the developments of this section, the work in [11] has also provided a solution for reducing the phase dispersion that exists in the methods of [7], and the techniques described there also involve estimating an offset (denoted as t_0 in [11]) from the analysis frame boundary. However, beyond these superficial similarities, the actual solutions proposed here and in [11] differ in three important aspects.

Most fundamentally, the work [11] 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 Section III and Section IV-C.

Secondly, in [11] 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 [11] 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 there the reconstructed signal matches a particular instant in the analysis frame.

Finally, [11] employs a separate source/filter decomposition so that during rate-change, excitation and system phase trajectories are modified differently. As explained in Section IV-C 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.

VI. Experimental Results

This section profiles the performance of the new phase-invariant algorithm of this paper by illustrating its time-scaling performance on a 2–s duration, 11–kHz sampled record of the speech of a female news-reader which is shown as the left hand plot in Fig. 3. Shown as the right-hand plot in that same figure is the speech time-scaled by a factor $\rho = 1.8$. Clearly, the essential features of the speech are retained, but lengthened under the time-scaling operation. In Fig. 4 the same data as shown in Fig. 3 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. Note further, that there are less tracks in the reconstructed signal relative to the original version due to thresholding being applied. No more than 20 tracks were reconstructed, and tracks with amplitude less than 5% of the largest amplitude are ignored. More generous thresholds resulting in more reconstructed tracks were found to yield negligible perceptible benefit.

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 160 Hz, 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 Fig. 5.

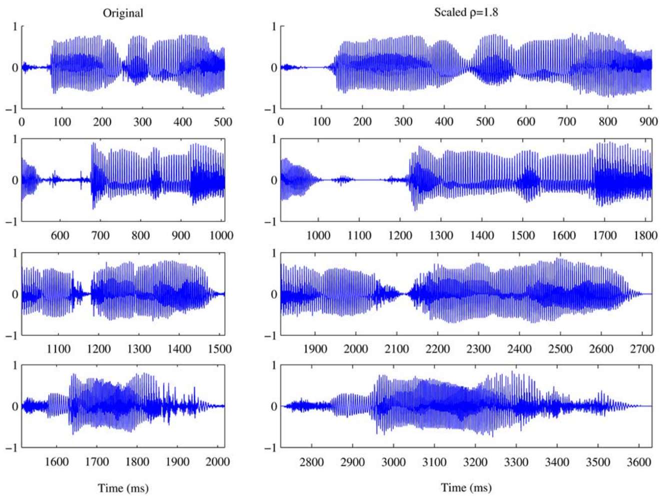

Fig. 3. Left plot is 11 Khz sampled speech of a female news-reader saying "searchers found a note from the men earlier today." Right plot is same sample time scaled by factor $\rho = 1.8$ using the new phase-invariant method introduced in this paper.

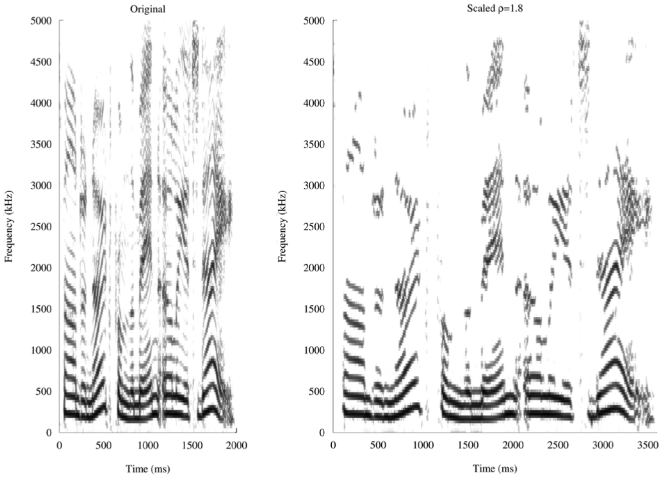

Fig. 4. Spectrograms of original and factor $\rho = 1.8$ time-scaled speech. Sample is a female news-reader saying "searchers found a note from the men earlier today."

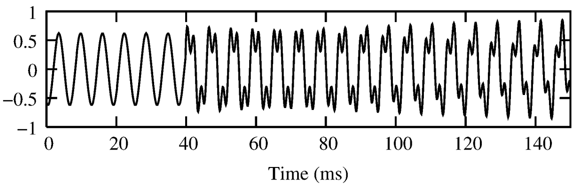

Fig. 5. Synthetically generated test signal consisting of tones at 160 Hz and (after 40 ms) 483 Hz, with the latter tone having a slowly time varying phase offset.

Shown in the top diagrams of Fig. 6 is the $\rho = 1.8$ time-scaled version of this signal when the original sinusoidal model based methods of [6], [7] 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 signal.

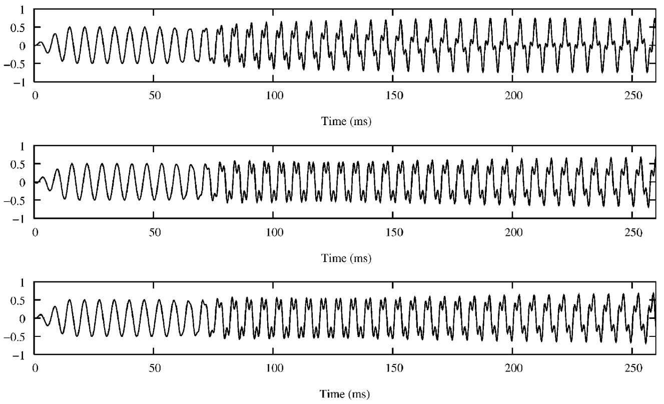

Fig. 6. The plot on the top shows the $\rho = 1.8$ time-scaled version of the signal in Fig. 5 when the original, sinusoidal model-based method of [7] are employed. The middle plot shows the $\rho = 1.8$ time-scaled version using the phase-invariant modification of [7] developed in [11]. The bottom plot shows the $\rho = 1.8$ time-scaled version using the phase-invariant methods developed in this paper.

TABLE I
COMPUTATIONAL BREAKDOWN

Computation	Time
FFT	1.0ms
Peak Extraction	1.0ms
Peak Matching	0.2ms
Spline Coefficients	0.1ms
Reconstruction	0.9ms
Total	3.2ms

In contrast, the $\rho = 1.8$ time-scaled signal which is generated using the phase-invariant methods developed in [11] and shown in the middle plot of Fig. 6 possess a wave-shape that, despite being time scaled, very closely matches that of the original signal of Fig. 5.

The same is true of the $\rho = 1.8$ time-scaled signal produced by the methods of this paper, which is shown as the bottom plot of Fig. 6, and is indistinguishable from the results shown above it.

Finally, the methods developed in this paper were implemented in a portable embedded system using a Texas Instruments TMS320C31 DSP (the intended application is for training medical students in auscultating heard sounds). This is a relatively cheap and low powered device, running at 40 MHz. Using typical parameters for the algorithm, the computation times shown in Table I were found for a single frame of speech processed. The three primary components of the FFT, the peak extraction, and the sinusoidal reconstruction each take about one third of the total computation time. In previous work [11], there has also been a need to separate the response of the excitation and vocal tract. If a Hilbert transform is used to calculate this, an additional FFT is required, which based on our timing measurements profiled in Table I adds approximately a further 30% to the computational requirements.

VII. 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 preexisting 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 30% 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] J. di Martinao and Y. Laprie, "Suppression of phasiness for time-scale modifications of speech signals based on a shape invariance property," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, vol. 2, pp. 853–856.

[2] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[3] E. Hardam, "High quality time scale modification of speech signals using fast synchronised-overlap-add algorithms," in *Proc. IEEE Int. Conf. Acust., Speech Signal Process.*, 1990, pp. 409–412.

[4] J. Laroche and M. Dolson, "Phase-vocoder: About this phasiness," in *Proc. IEEE Workshop on Applicat. Signal Process. Audio Acoust.*, 1997, pp. 1–4.

[5] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 323–332, May 1999.

[6] R. McAulay and T. Quatieri, "Speech analysis-syntheses based on a sinusoidal representation," *IEEE Trans. Acoust., Speech., Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[7] R. McAulay and T. Quatieri, "Speech transformations based on a sinusoidal representation," *IEEE Trans. Acoust., Speech., Signal Process.*, vol. ASSP-34, no. 6, pp. 1449–1464, Dec. 1986.

[8] R. McAulay and T. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal model," in *Proc IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1990, pp. 249–252.

[9] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, 1995.

[10] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech., Signal Process.*, vol. ASSP-29, no. 3, pp. 374–390, Jun. 1981.

[11] T. Quatieri and R. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar. 1992.

[12] S. Roucos and A. M. Wilgus, "High quality time scale modification for speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1985, pp. 493–496.

[13] W. Verhelst and M. Roelands, "An Overlap-Add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1993, pp. 554–557.

[14] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1989, pp. 714–717.

Brett Ninness was born in Singleton, Australia, in 1963. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from the University of Newcastle, Australia, in 1986, 1991, and 1994, respectively.

He has been with the School of Electrical Engineering and Computer Science, University of Newcastle, since 1993, where he is currently a Professor. His research interests are in the areas of system identification and stochastic signal processing, in which he has authored approximately 100 papers in journals and conference proceedings.

Dr. Ninness has served on the editorial boards of *Automatica*, the IEEE Transactions on Automatic Control, and is currently Editor-in-Chief for IET *Control Theory and Applications*. Together with Håkan Hjalmarsson, he jointly organized the 14th IFAC Symposium on System Identification in Newcastle, Australia, in 2006. Further details of his professional activities are available at http://sigpromu.org/brett.

Soren John Henriksen received the bachelor's degree in computer engineering from the University of Newcastle, Australia, in 1996, and the M.S. degree in system identification and control, also from the University of Newcastle in 2001.

He is currently pursuing the Ph.D. degree with his focus on applying statistical methods to system identification.