SHAPE INVARIANT TIME-SCALE MODIFICATION OF SPEECH USING A HARMONIC MODEL

Darragh O'Brien & Alex Monaghan

School of Computer Applications
Dublin City University
Glasnevin, Dublin 9, Ireland
dobrien@compapp.dcu.ie & alex@compapp.dcu.ie

ABSTRACT

A new and simple approach to shape invariant time-scale modification of speech is presented. The method, based upon a harmonic coding of each speech frame, operates entirely within the original sinusoidal model [3] and makes no use of "pitch-pulse onset times" used by conventional algorithms. Instead, phase coherence, and thus shape invariance, are ensured by exploiting the harmonic relation existing between the sine waves to cause them to be in phase at each adjusted frame boundary. Results suggest this approach to be an excellent candidate for use within a concatenative text-to-speech synthesiser [2] where scaling factors typically lie within a range well handled by this algorithm.

1. INTRODUCTION

Preservation of the original waveform shape in time-scaled speech is essential if high quality results are to be obtained. Failure to maintain shape invariance introduces an unnatural reverberant quality into the modified speech [4]. Conventional time-scaling methods, using the sinusoidal model of speech [3], use the notion of a "pitch pulse onset time" in order to retain waveform shape. At each onset time all waves are assumed to be in phase i.e. the phase of each is assumed to be an integer multiple of 2π. By enforcing this restriction at estimated onset times original waveform shape retention is ensured. This "phase synchronisation" may be applied at one onset time per synthesis frame [5], or at several [1] for more accurate waveform shape preservation.

The incorporation of pitch pulse onset times in order to facilitate time-scaling detracts from the simplicity of the original sinusoidal model, adding an undesirable level of complexity. Presented here is an approach which makes no use of onset times but instead adapts the original model to permit time-scaling.

In Section 2 the sinusoidal model of speech is reviewed. The model is adapted in Section 3 to handle time-scale modification. Section 4 contains some experimental results. Conclusions and suggestions for future work are given in Section 5.

2. THE SINUSOIDAL MODEL OF SPEECH

In McAulay and Quatieri's original formulation of the sinusoidal model [3], peaks extracted from the DFT of speech frame k are matched with those of frame $k+1$ using a nearest neighbour algorithm. Let $\{A_l^k, \omega_l^k, \psi_l^k\}$ and $\{A_l^{k+1}, \omega_l^{k+1}, \psi_l^{k+1}\}$ denote the instantaneous amplitude, frequency and phase of the l_{th} sinusoid at the centre of frames k and $k+1$ respectively. Amplitude is interpolated linearly using (1) where T is the time interval from the centre of frame k to the centre of frame $k+1$.

$$A(t) = A_l^k + \frac{A_l^{k+1} - A_l^k}{T} t \tag{1}$$

A cubic polynomial (2) is introduced to model phase interpolation. Given that instantaneous frequency is defined as the derivative of phase, the phase and frequency of each sine wave at any time t are given by (2) and (3) respectively.

$$\tilde{\theta}(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3 \tag{2}$$

$$\dot{\tilde{\theta}}(t) = \gamma + 2\alpha t + 3\beta t^2 \tag{3}$$

Substituting the known boundary values, when $t = 0$, obtained from the DFT analysis into (2) and (3) gives

$$\begin{aligned} \tilde{\theta}(0) &= \zeta = \psi_l^k \\ \dot{\tilde{\theta}}(0) &= \gamma = \omega_l^k \end{aligned} \tag{4}$$

Similarly, substituting the known boundary values when $t = T$ gives

$$\tilde{\theta}(T) = \zeta + \gamma S + \alpha S^2 + \beta S^3 = \psi_l^{k+1} + 2\pi M$$

$$\dot{\tilde{\theta}}(T) = \gamma + 2\alpha S + 3\beta S^2 = \omega_l^{k+1} \quad (5)$$

The target phase ψ_l^{k+1} is measured modulo 2π so phase unwrapping must be performed and the $2\pi M$ term is added to (5) where M is an integer. We now solve for the three unknowns α, β and M. For any M we can solve for α and β from

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} 3/T^2 & -1/T \\ -2/T^3 & 1/T^2 \end{bmatrix} \begin{bmatrix} \psi_l^{k+1} - \psi_l^k - \omega_l^k T + 2\pi M \\ \omega_l^{k+1} - \omega_l^k \end{bmatrix} \quad (6)$$

The value of M is chosen such that a maximally smooth frequency track is obtained. This is achieved by minimising (7) with respect to the continuous variable x.

$$f(x) = \int_0^T [\ddot{\tilde{\theta}}(t;x)]^2 dt \quad (7)$$

The minimising value of x can be shown to be that given by (8). Rounding to the closest integer gives M^*, as shown in (9), where [[ ]] denotes the "nearest integer" operator.

$$x^* = \frac{1}{2\pi}\left[(\psi_l^k + \omega_l^k T - \psi_l^{k+1}) + (\omega_l^{k+1} - \omega_l^k)\frac{T}{2}\right] \quad (8)$$

$$M^* = [[x^*]] \quad (9)$$

Once M^* has been determined we compute $\alpha(M^*)$ and $\beta(M^*)$ from (6) completing the model. Speech may then be re-synthesised from (10) where L_k is the number of waves in frame k.

$$\tilde{s}(t) = \sum_{l=1}^{L_k} A_l^k(t) \cos[\tilde{\theta}_l^k(t)] \quad (10)$$

3. TIME-SCALE MODIFICATION

In the approach adopted here, a pitch estimate is assigned to each speech frame whose DFT is then calculated. For each voiced frame, instead of picking peaks from the DFT however, as in Section 2, the amplitude and phase at each harmonic frequency are coded. For voiceless frames peak-picking applies.

After nearest neighbour frequency matching has been carried out, and during re-synthesis, let us assume a match has been made between the first harmonic (F0) of frame k and the first harmonic of frame $k+1$, each respectively defined by the parameters $\{A_0^k, \omega_0^k, \psi_0^k\}$ and $\{A_0^{k+1}, \omega_0^{k+1}, \psi_0^{k+1}\}$. (A reasonable assumption over voiced speech segments.) Firstly, the original frequency track (3), repeated here as (11), is computed as outlined above.

$$\dot{\tilde{\theta}}(t) = \gamma + 2\alpha t + 3\beta t^2 \quad (11)$$

Time-scaling the first harmonic is straightforward. For a given time-scaling factor, ρ, two parameters need be determined, $M^{*\prime}$ and $\psi^{k+1\prime}$, the new phase unwrapping and target phase values respectively. Let the new time-scaled frequency function be

$$\dot{\tilde{\theta}}'(t) = \dot{\tilde{\theta}}\left(\frac{t}{\rho}\right) \quad (12)$$

The new (unwrapped) target phase is found by integrating (12) over the time interval ρT and adding back the start phase ψ^k,

$$\int_0^{\rho T} \dot{\tilde{\theta}}'(t)dt + \psi^k = \rho T\left(\gamma + \alpha T + \beta T^2\right) + \psi^k \quad (13)$$

By evaluating (13) modulo 2π, $M^{*\prime}$ and $\psi^{k+1\prime}$ are determined. The model is completed by evaluating $\alpha(M^{*\prime})$ and $\beta(M^{*\prime})$. In Section 2, M^* was chosen such that the smoothest possible frequency track was obtained between measured start and target parameters. Here, an original frequency track has been time-scaled in order to estimate a new phase unwrapping parameter, $M^{*\prime}$, and a new target phase, $\psi^{k+1\prime}$. Remaining model parameters are then solved for as in Section 2.

Applying the same procedure to each remaining matched pair of harmonics will, however, lead to a breakdown in phase coherence after several frames as waves gradually move out of phase. To overcome this we first calculate δ from (14).

$$\delta = \frac{\psi^{k+1\prime} - \psi^{k+1}}{\omega^{k+1}} \quad (14)$$

δ represents the amount of time taken for the first harmonic in frame $k+1$ to move from its measured phase value, ψ^{k+1}, to its adjusted phase value, $\psi^{k+1\prime}$, while keeping its frequency, ω^{k+1}, constant. As all waves are harmonically related we are justified in adjusting all target phases by this same amount and we move from one "valid" phase configuration to another equally "valid" one. Thus phase coherence at frame boundaries is guaranteed.

The target phase of each remaining harmonic is adjusted by applying (15).

$$\psi' = \psi + \delta\ \omega \quad (15)$$

Once an adjusted target phase has been determined for each matched pair of harmonics, $M^{*\prime}$ is chosen, not such that the smoothest possible frequency track is obtained but such that the shape of the track matches, as closely as possible, the shape of the original. Let $\vartheta(t)$, given in (16), be the time-scaled version of the original frequency track.

$$\vartheta(t) = \dot{\tilde{\theta}}\left(\frac{t}{\rho}\right) \quad (16)$$

Using a least-squares error criterion, the quantity to be minimised is then given by

$$f'(x') = \int_0^{\rho T} [\dot{\tilde{\theta}}'(t;x') - \vartheta(t)]^2 dt \qquad (17)$$

The value of the continuous variable x', which minimises (17) can be shown to be that given by (18). $M^{*\prime}$ is then computed from (19) where, again, [[ ]] denotes the nearest integer operator. The model is completed by evaluating $\alpha(M^{*\prime})$ and $\beta(M^{*\prime})$.

$$\begin{aligned} x^{*\prime} &= (10\alpha\rho T^2 + 9\beta\rho T^3 - 12\psi^{k+1\prime} + \\ &\quad 11\rho T\omega^k + 12\psi^k + \rho T\omega^{k+1})/24\pi \qquad (18) \\ M^{*\prime} &= [[x^{*\prime}]] \qquad (19) \end{aligned}$$

Obviously, it is necessary to keep track of previous phase adjustments when moving from frame to frame. This is handled by Δ (see Figure 1). In the same way as δ is applied to target phases in (15) so too Δ is applied to both start and target phases prior to the time-scaling of each frame. The complete algorithm is presented in Figure 1.

4. RESULTS

A 22.6ms Hamming window was applied at 11.3ms intervals to speech sampled at a frequency of 10kHz. A pitch estimate was made for each frame and a 4096-point FFT computed. The amplitudes and phases of the harmonics of each frame were then coded.

During re-synthesis McAulay and Quatieri's [3] nearest neighbour matching algorithm was used to match waves from one frame with those from the next. The algorithm presented in Section 3 was then applied with fixed scaling factors, $\rho = 0.6$ and $\rho = 1.3$ (speeding up and slowing down, respectively, the perceived rate of articulation).

Given in Figure 2 is a section of the original speech waveform. Time-scaled versions are shown in Figures 3 and 4. In each case the original waveform shape has been well preserved. Furthermore, the re-synthesised speech from which these examples were drawn was found to be of high quality and free of any reverberation. Phase coherence was found, however, to begin to break down for larger scaling factors, ($\rho > 1.8$). This is to be expected since as the distance between start and target parameters is increased so too is the risk of phase coherence breakdown. The method presented here would then be of most use in concatenative speech synthesisers where scaling factors lie usually within the bounds handled by the algorithm.

$\Delta = 0$
$\delta = 0$
For each Frame
  Begin
    $\Delta = \Delta + \delta$
    For first harmonic
      Begin
        Adjust ψ^k and ψ^{k+1} by Δ
        Compute old frequency track $\dot{\tilde{\theta}}(t)$
        Compute new frequency track $\dot{\tilde{\theta}}'(t)$
        Solve for $\psi^{k+1\prime}$ and $M^{*\prime}$
        Solve for δ
        Compute model parameters
      End
    For remaining harmonics
      Begin
        Adjust ψ^k and ψ^{k+1} by Δ
        Compute old frequency track $\dot{\tilde{\theta}}(t)$
        Adjust ψ^{k+1} by δ
        Compute new frequency track $\dot{\tilde{\theta}}'(t)$
        Solve for $M^{*\prime}$
        Compute model parameters
      End
  End

Figure 1: Time-Scale Modification Algorithm

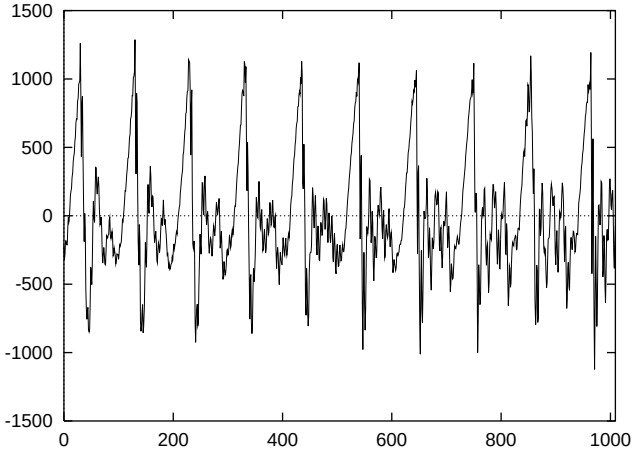

Figure 2: Original speech waveform, $\rho = 1$

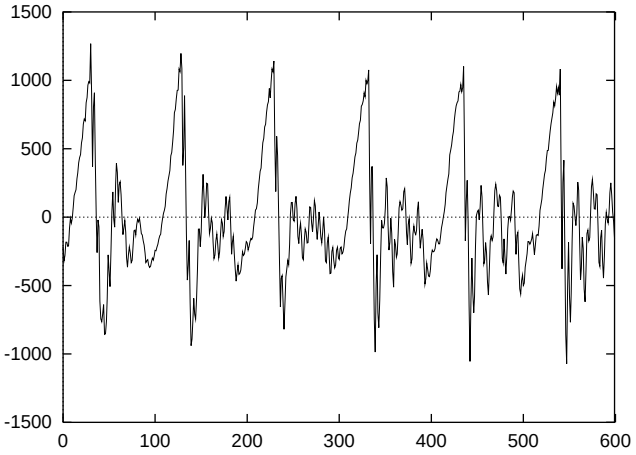

Figure 3: Time-scaled speech waveform, $\rho = 0.6$

5. CONCLUSIONS

A simple method of time-scale modification has been presented which by consistently adjusting phase values in each frame and maintaining, as closely as possible, the original frequency track shapes in the time-scaled speech achieves shape invariance. Importantly, no decoupling (into source and vocal tract models) of the speech production process is necessary.

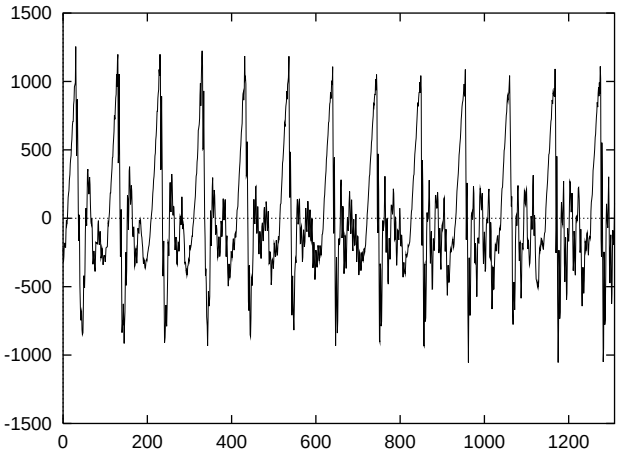

Figure 4: Time-scaled speech waveform, $\rho = 1.3$

The approach works entirely within the original sinusoidal model unlike other methods which use the idea of a "pitch-pulse onset time" in order to keep waveform shape constant. The process of estimating onset times and forcing waves to be in phase at such points contrasts sharply with the simplicity of the original sinusoidal model. That process has been eliminated in the algorithm presented here. As pointed out, shape invariance breaks down for larger scaling factors but the method still gives good results for scaling factors required by a concatenative speech synthesiser.

6. REFERENCES

[1] Pollard M.P., Cheetham B.M.G., Goodyear C.C., and Edgington M.D. Shape-invariant pitch and time-scale modification of speech by variable order phase interpolation. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 919–922, 1997.

[2] Campbell W. N., Isard S. D., Monaghan A., and Verhoeven J. Duration, pitch and diphones in the cstr tts system. In *Proceedings of ICSLP*, pages 825–828, Kobe, Japan, November 1990.

[3] McAulay R.J. and Quatieri T.F. Speech analysis-synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(4):744–754, August 1986.

[4] Quatieri T.F. and McAulay R.J. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(6):1449–1464, August 1986.

[5] Quatieri T.F. and McAulay R.J. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, March 1992.