SHAPE IN Variant PITCH MODIFICATION OF SPEECH USING A HARMONIC MODEL

Darragh O'Brien & A.I.C. Monaghan

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

ABSTRACT

We present a simple but effective approach to pitch modification of speech based on a harmonic model. Building on our time-scaling algorithm [1], pitch modification applies to a harmonically coded glottal wave estimate derived via a simple inverse filtering technique [3]. The modified glottal wave subsequently serves as input to an LPC vocal tract filter and the pitch-scaled speech is generated. Shape invariance is maintained in the glottal wave by exploiting the harmonic nature of the sine waves used to code each frame thus avoiding the need for “pitch pulse onset time” estimation. Furthermore, given its smooth shape it is not necessary to resample the glottal wave spectrum at the new harmonic frequencies. The original spectrum is merely compressed/expanded to produce the desired pitch change.

Keywords: harmonic model, pitch modification

1. INTRODUCTION

Preservation of the original waveform shape in pitch or time-scale modified speech is essential if an unnatural reverberant quality is not to be introduced [5]. Conventional pitch- and time-scaling methods [6] [2], employing the sinusoidal model of speech [4], 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. However the labelling of onset times must be performed either manually or automatically, and in either case errors may be introduced which seriously impair the quality of the subsequently modified speech.

We have therefore chosen to do without pitch pulse markers and to develop algorithms which make minimal assumptions and perform minimal modification of the speech signal. In this fashion, we hope to avoid introducing any unnatural elements.

The sinusoidal model is briefly outlined in Section 2. Our previous work [1] produced an algorithm for time-scale modification which did not require the calculation of pitch pulse onset times: this is briefly revisited in Section 3. In Section 4 the time-scaling method is extended to handle pitch modification. Section 5 contains some experimental results. Conclusions are given in Section 6.

2. THE SINUSOIDAL MODEL OF SPEECH

In McAulay and Quatieri’s original formulation of the sinusoidal model [4], peaks extracted from the DFT of speech frame k are matched with those of frame k + 1 using a nearest neighbour algorithm. Let \( \{ A^k_l, \omega^k_l, \psi^k_l \} \) and \( \{ A^{k+1}_l, \omega^{k+1}_l, \psi^{k+1}_l \} \) 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 \( k + 1 \).

\[
A(t) = A^k_l + \frac{A^{k+1}_l - A^k_l}{T} \tag{1}
\]

A cubic polynomial 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.

\[
\dot{\theta}(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3 \tag{2}
\]
\[
\ddot{\theta}(t) = \gamma + 2\alpha t + 3\beta t^2 \tag{3}
\]

A method for solving for the variables \( \zeta, \gamma, \alpha \) and \( \beta \) is given in [4]. Speech may then be resynthesised from (4) where \( L_k \) is the number of waves in frame \( k \).

\[
\hat{s}(t) = \sum_{i=1}^{L_k} A^k_i(t) \cos[\hat{\theta}^k_i(t)] \tag{4}
\]

3. TIME-SCALE MODIFICATION

During analysis, a pitch estimate is assigned to each speech frame whose DFT is then calculated. For each
voiced frame, the amplitude and phase at each harmonic frequency are coded. Peak-picking applies in voiceless frames [4].

After nearest neighbour frequency matching has been carried out, we assume a match has been made between the fundamental of frame $k$ and the fundamental of frame $k+1$, each respectively defined by the parameters $\{A^k_0, \omega^k_0, \psi^k_0\}$ and $\{A^{k+1}_0, \omega^{k+1}_0, \psi^{k+1}_0\}$. The original frequency track (3), repeated here as (5), is computed.

$$\hat{\vartheta}(t) = \gamma + 2at + 3bt^2$$  \hspace{1cm} (5)

For any given factor, $\rho$, (5) may be time-scaled to give a new frequency function,

$$\hat{\vartheta}(t) = \hat{\vartheta}\left(\frac{t}{\rho}\right)$$  \hspace{1cm} (6)

(6) may be integrated over the interval $\rho T$ and evaluated modulo $2\pi$ to find a new target phase value $\psi^{k+1'}$. Substituting $\rho T$ for $T$ and $\psi^{k+1'}$ for $\psi^{k+1}$ phase and amplitude interpolating functions are computed as in [4].

To keep waves in phase at frame boundaries we calculate $\delta$ from (7).

$$\delta = \frac{\psi^{k+1'} - \psi^{k+1}}{\omega^{k+1}}$$  \hspace{1cm} (7)

$\delta$ represents the amount of time taken for the first harmonic in frame $k+1$ to move from its measured phase value, $\psi^{k+1}$, to its adjusted phase value, $\psi^{k+1'}$, while keeping its frequency, $\omega^{k+1}$, constant. The target phase of each remaining harmonic is adjusted by applying (8).

$$\psi' = \psi + \delta \omega$$  \hspace{1cm} (8)

Once an adjusted target phase has been determined for each matched pair of harmonics, the phase and amplitude interpolating functions may be calculated. (Note this differs from the approach in [1] where the difference between the original and new frequency tracks was minimised. Here the original is never computed.) Obviously, it is necessary to keep track of previous phase adjustments when moving from frame to frame. This is handled by $\Delta$ i.e. the sum of all previous $\delta$ values. In the same way as $\delta$ is applied to target phases in (8) so too $\Delta$ is applied to both start and target phases prior to the time-scaling of each frame.

\section{4. PITCH MODIFICATION}

In order to perform pitch modification the speech is first inverse filtered using a simple algorithm proposed by Alku [3] to estimate the glottal wave. A pitch estimate is assigned to each frame of the estimated glottal wave and its DFT calculated. After matching we assume the fundamental of frame $k$ has been matched with that of frame $k+1$ where each is respectively defined by the parameters $\{A^k_0, \omega^k_0, \psi^k_0\}$ and $\{A^{k+1}_0, \omega^{k+1}_0, \psi^{k+1}_0\}$. The original frequency track is again computed.

$$\hat{\vartheta}(t) = \gamma + 2at + 3bt^2$$  \hspace{1cm} (9)

Let $\mu^k$ and $\mu^{k+1}$ be the pitch modification factors associated with frames $k$ and $k+1$ respectively. The integral of the new pitch-scaled frequency track over the interval $T$ can be estimated by time-scaling the original frequency track by a factor $\lambda$ where,

$$\lambda = \frac{\mu^k + \mu^{k+1}}{2}$$  \hspace{1cm} (10)

The new frequency function is integrated over $\lambda T$ and evaluated modulo $2\pi$ to give a new target phase value $\psi^{k+1'}$. Substituting $\psi^{k+1'}$ for $\psi^{k+1}$, $\omega^{k'}$ for $\omega^k$ and $\omega^{k+1'}$ for $\omega^{k+1}$ the phase interpolating function across $T$ is computed where

$$\omega^{k'} = \omega^k \mu^k$$  \hspace{1cm} (11)

$$\omega^{k+1'} = \omega^{k+1} \mu^{k+1}$$  \hspace{1cm} (12)

As in time-scaling, $\delta$ is calculated and used to estimate the new target phase value for each of the remaining harmonics. Phase interpolation functions are then calculated using scaled frequency values. $\Delta$ is again used to keep track of previous phase adjustments as we move from frame to frame.

The amplitude of each harmonic after pitch-scaling is left unchanged i.e. the glottal wave spectrum is not resampled at the new harmonic frequencies but simply expanded/compressed to effect the desired pitch change. We are firstly assuming that the glottal wave spectrum is relatively flat and secondly that although we alter its spectrum we preserve the shape of the glottal waveform in the time-domain and so voice quality should remain constant.

Furthermore, it should be pointed out that combining pitch and time-scale modification using the approach presented is straightforward. A single algorithm allows the independent modification of pitch and duration. For a given frame with associated time-scale and pitch modification factors $\rho$ and $\lambda$ respectively the net scaling factor $\kappa$ is given by

$$\kappa = \rho \lambda$$  \hspace{1cm} (13)

The fundamental frequency track of each frame is scaled by $\kappa$ and $\delta$ calculated. New phase and amplitude interpolating functions may then be computed over the interval $\rho T$ with start and target frequencies given by $\mu^k \omega^k$ and $\mu^{k+1} \omega^{k+1}$ respectively. The complete algorithm for combined pitch and time-scale modification is presented in Figure 1.
\[
\Delta = 0 \\
\delta = 0 \\
\text{For each Frame} \\
\text{Begin} \\
\Delta = \Delta + \delta \\
\kappa = \rho \lambda \\
\text{For first harmonic} \\
\text{Begin} \\
\text{Adjust } \psi^k \text{ and } \psi^{k+1} \text{ by } \Delta \\
\text{Compute original frequency track } \hat{\theta}(t) \\
\text{Compute new frequency track } \hat{\theta}'(t) \\
\text{Solve for } \psi^{k+1} \\
\text{Solve for } \delta \\
\text{Compute phase interpolation function } \theta(t) \\
\text{End} \\
\text{For remaining harmonics} \\
\text{Begin} \\
\text{Adjust } \psi^k \text{ and } \psi^{k+1} \text{ by } \Delta \\
\text{Adjust } \psi^{k+1} \text{ by } \delta \\
\text{Compute phase interpolation function } \theta(t) \\
\text{End} \\
\text{End}
\]

Figure 1: Pitch- and Time-Scaling Algorithm

5. RESULTS

Speech sampled at 16kHz was inverse filtered using Alku’s algorithm [2] to generate a glottal wave estimate. A 21.6ms Hamming window was applied at 10.8ms intervals to the glottal wave and a pitch estimate assigned to each frame. For each a 4096-point FFT was computed and the amplitudes and phases of the harmonics coded if voiced, if voiceless peak-picking was used.

The pitch-scaling algorithm presented here was then applied to the glottal wave estimate with fixed scaling factors \( \lambda = 0.7 \) and \( \lambda = 1.6 \), lowering and raising the pitch respectively. The speech was resynthesised using an LPC filter of order 16 to model the vocal tract transfer function. Given in Figure 2 is a section of the original speech waveform. The pitch-scaled versions are presented in Figures 3 and 4. In each case the original waveform shape is well preserved and the speech from which these samples are drawn was found to be of high quality. Presented in Figures 5 and 6 are examples of combined pitch and time-scale modification.
A deliberately simple time-scaling algorithm using a harmonic model of speech has been extended to allow pitch modification. Importantly, neither requires pitch pulse onset time estimation. These algorithms produce high quality results for scaling factors of the order required for concatenative speech synthesis. At the moment a crude voiced/voiceless decision is made for each speech frame and results are expected to improve when a finer discrimination is incorporated.

Figure 5: Pitch- and time scaled, \( \rho = 0.7, \lambda = 0.7 \)

Figure 6: Pitch- and time scaled, \( \rho = 1.6, \lambda = 1.6 \)

7. REFERENCES


