ENHANCED SHAPE-INVARIANT PITCH AND TIME-SCALE MODIFICATION FOR CONCATENATIVE SPEECH SYNTHESIS

M. P. Pollard, B. M. G. Cheetham*, C. C. Goodyear* , M. D. Edgington[#] and A. Lowry[#]*

*Department of Electrical Engineering and Electronics, The University of Liverpool, LIVERPOOL. L69 3BX. U.K.

[#]B.T. Laboratories, Martlesham Heath, IPSWICH. IP5 7RE. U.K.

ABSTRACT

To preserve shape-invariance when pitch or time-scale modifying sinusoidally modelled voiced speech, the phases of the sinusoids used to model the glottal excitation are made to add coherently at estimated excitation points. Previous methods achieve this by estimating excitation phases at synthesis frame boundaries, disregarding the frequency modulation that may occur between the frame boundary and the nearest modified excitation point. This approximation can produce a significant mis-alignment of the excitation phases, leading to distortion of the temporal structure of the synthetic speech. In this paper, a shape-invariant technique is proposed which aligns the excitation phases at excitation points, whilst allowing for variations in the frequency of the sinusoidal components.

1. INTRODUCTION

Concatenative speech synthesis requires short segments of speech to be joined together with modifications to their pitch and time-scales. The sinusoidal speech model [1] is now recognised as a reliable basis for both pitch and time-scale modification of speech. Several techniques which use the sinusoidal framework are well documented including the so-called "Shape-Invariant" techniques [3]. A synthetic speech waveform is said to be shape-invariant if, after time-scaling, its individual pitch cycles resemble those of the original speech. Attempts to modify the pitch or time-scale of speech without preserving the shape, have been found to produce speech which has a reverberant quality [2]. The temporal structure of a speech waveform is largely influenced by the periodic closure of the glottis. This, it may be assumed, forces the glottal excitation into phase once every pitch cycle at times known as excitation points. By "into phase" we mean that the instantaneous phase of each harmonically related sinusoid is an integer multiple of 2π. In each pitch cycle, the glottal excitation is therefore made impulse-like during voiced speech. Achieving this phase relationship, when the glottal excitation points have been redefined by the time or pitch scaling requirements and therefore no longer coincide with the excitation points of the original speech, is more difficult than may first appear. The main difficulty arises from the fact that the instantaneous phases of the sinusoids modelling the excitation will not be directly known at the synthesis frame boundaries and must instead be deduced from a knowledge of the waveform at some other point or points in time i.e. at the excitation points. The solution to the problem proposed by McAulay and Quatieri [3] is to calculate the phase of each sinusoid at the end of each synthesis frame by assuming a constant frequency between the frame boundary and the nearest time and/or pitch scaled excitation point where the phase is taken to be $2\pi M$ for some integer, M. A value of M is taken which minimises the variation of instantaneous frequency over the frame, when cubic interpolation of instantaneous phase is applied between frame boundaries. In general, this estimate does not guarantee maximum achievable phase coherence at each excitation point because the instantaneous frequency, being represented as a quadratic function of time, will not remain constant between the frame boundary and the excitation point.

This paper is concerned with attempts to better preserve phase coherence in the glottal excitation without making contradictory assumptions about the variation of frequency. The criteria for shape-invariant voiced speech are described in the following section, an analysis/synthesis method designed to meet the criteria is presented in sections 3 and 4 and a comparison with the method described in [3] is given in section 5.

2. THE SHAPE-INVARIANT MODEL OF SPEECH

Commonly used models of speech production [5] assume that stationary segments of voiced speech may be produced by passing a train of scaled impulses $e(t)$ through a filter modelling the effect of the glottis, vocal tract and lip-radiation. The excitation $e(t)$ may be written as

$$e(t) = a + 2a\sum_{n=1}^{\infty}\cos\left[n\omega_0(t-\tau)\right] \qquad (1)$$

Pitch pulse locations occur at $t=\tau$, $t=\tau \pm 2\pi/\omega_0$, $t=\tau \pm 4\pi/\omega_0$, etc. i.e. where all the excitation phases of the harmonics are integer multiples of 2π. Since, in practice, voiced speech is quasi-stationary and band-limited, $e(t)$ may be better approximated as the sum of a finite number of amplitude and frequency modulated sinusoids

$$e(t) = \sum_{l=0}^{L-1} a_l(t)\cos\left[\Omega_l(t)\right] \qquad (2)$$

where $a_l(t)$ and $\Omega_l(t)$ are the instantaneous amplitude and phase respectively for frequency component l. To preserve an impulse-like shape for the excitation signal, even when the instantaneous frequencies of the pitch frequency harmonics become variable, the excitation phases must also be made to be integer multiples of 2π once every pitch cycle. The speech signal $s(t)$ can then be

produced by the introduction of the vocal system model (i.e. glottis, vocal tract and lip-radiation) parameters, i.e.

$$s(t) = \sum_{l=0}^{L-1} a_l(t).\, M_l(t) \cos\left[\Omega_l(t) + \psi_l(t)\right] \tag{3}$$

where $\psi_l(t)$ and $M_l(t)$ are the slowly evolving vocal system phases and magnitudes respectively at the harmonic frequencies.

3. ANALYSIS

The following description is concerned with voiced speech only. Unvoiced speech is dealt with by a simple extension to the technique. For the application of this research, the sinusoidal model parameters are extracted pitch-synchronously from a knowledge of excitation points. During voiced speech, an analysis window with width equal to two and a half times the average pitch period is placed symmetrically around each excitation point and Hamming weighted. A zero-padded 1024 point FFT is computed and a set of peaks at harmonically related frequencies are chosen from the magnitude spectrum according to [4]. The magnitude of each peak is calculated and the corresponding phase is obtained from the FFT spectrum. Since the analysis window is centred around an excitation point where all the excitation phases are assumed to be integer multiples of 2π, the wrapped phase measured at each peak frequency is, in principle, the phase of the system component, $\psi(\omega; t)$. Finally, vocal system magnitude and phase envelopes are computed by fitting cubic spline interpolation functions to the measurements at and adjacent to each peak. These parameters thus extracted are a characterisation of the speech signal at an excitation point.

4. SYNTHESIS

To synthesise pitch and time-scale modified speech, the first step is to compute modified excitation points which represent glottal closure times according to the new time-scale. This is most simply achieved by accumulating estimates of excitation points whose spacing equals the modified pitch-period [3].

We define a synthesis frame to be the segment lying between a time-scaled pair of analysis update points. It is convenient to define its boundaries to occur at relative times $t=0$ and $t=T$. It is assumed that all pitch dependent parameters have been pitch modified using the method described in [3]. The resulting sets of frequencies, system phases and amplitudes at time $t=0$ are then associated with the sets of frequencies, system phases and amplitudes respectively at time $t=T$ according to the frequency-matching technique described in [1]. The synthesis problem is therefore to generate for each frame a set of sinusoids whose instantaneous amplitudes, frequencies and system phases vary gradually from the values specified at $t=0$ to those specified at $t=T$. To maintain continuity at frame boundaries, the excitation phase component of each sinusoid at $t=0$, i.e. the phase offset, must be equal to that achieved by the corresponding sinewave at $t=T$ in the previous frame. The phase of each sinewave at $t=T$ can be made arbitrary (quadratic interpolation) but greater naturalness is achieved if an attempt is made to choose these phases in such a way that the phases of all sinusoids coalesce (i.e. are integer multiples of 2π) at each excitation point. A description of the interpolation scheme for one of the sinusoids, i.e. l, is described below.

An excitation point is selected from the modified set of excitation points which lies closest to the end of the synthesis frame, T. The relative position from $t=0$ of the excitation point is denoted by Z. The excitation phase of sinusoid l at $t=Z$, i.e. Ω_l^Z, is known to be equal to $2\pi M_l$ where M_l is some integer. The instantaneous frequency at $t=0$ is ω_l^0 and the corresponding frequency at $t=T$ is ω_l^T. The excitation phase, Ω_l^T, at time T could therefore be deduced from

$$\Omega_l^T = 2\pi M_l - \int_T^Z \omega_l(t) \cdot dt \tag{4}$$

where $\omega_l(t)$ is the slowly evolving instantaneous frequency for component l. Unfortunately the exact nature of $\omega_l(t)$ is unknown so we have no way of truly estimating Ω_l^T. Using the method described in [3], Ω_l^T can be estimated by assuming that the frequency is constant between T and Z such that

$$\Omega_l^T = 2\pi M_l - \omega_l^T(Z - T) \tag{5}$$

and cubic interpolation is performed exactly as in the baseline sinusoidal system [1]. For the purposes of this paper, this process shall be referred to as 'interpolation-by-linear-estimation'.

To get an exact solution to the problem, it is necessary to satisfy the conditions of the phase at Z and the instantaneous frequency at T simultaneously. That is, to find a function $\Omega_l(t)$ which has a value equal to $2\pi M_l$ at time Z whilst having a slope of ω_l^T at T. An exaggerated example of the problem is shown in figure 1 for the case where Z lies outside the synthesis frame 0 to T.

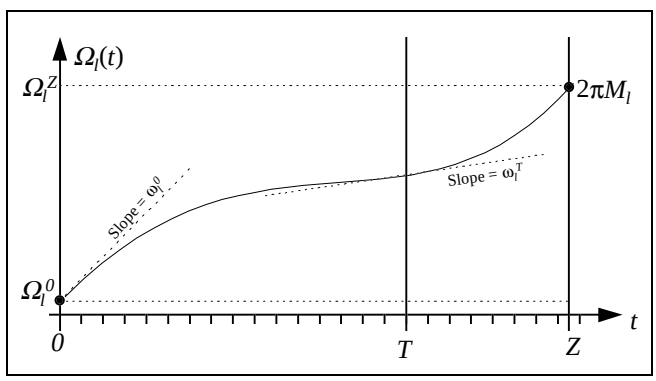

Figure 1: Interpolation of excitation phase for Z greater than T.

For the case where Z is greater than T, the main purpose of the interpolation function is to set a trajectory for the excitation phase such that the phase is 'aimed' at $2\pi M_l$ at the start of the next synthesis frame. As with the baseline model a cubic phase interpolation function is proposed which is of the form

$$\Omega_l(t) = \Omega_l^0 + \omega_l^0 t + \alpha_l t^2 + \beta_l t^3 \tag{6}$$

where Ω_l^0 and ω_l^0 are the initial instantaneous phase and frequency respectively. The excitation phase at Z is given by:

$$\Omega_l(Z) = \Omega_l^Z = \Omega_l^0 + \omega_l^0 Z + \alpha_l Z^2 + \beta_l Z^3 = 2\pi M_l \tag{7}$$

and the instantaneous frequency at T is

$$\Omega_l'(T) = \omega_l^0 + 2\alpha_l T + 3\beta_l T^2 = \omega_l^T \tag{8}$$

Rearranging into matrix form gives the matrix equation:

$$A.\underline{x}_l = \underline{y}_l \tag{9}$$

where

$$A = \begin{bmatrix} Z^2 & Z^3 \\ 2T & 3T^2 \end{bmatrix}, \underline{x}_l = \begin{bmatrix} \alpha_l \\ \beta_l \end{bmatrix} \text{and } \underline{y}_l = \begin{bmatrix} 2\pi M_l - \Omega_l^0 - \omega_l^0 Z \\ \omega_l^T - \omega_l^0 \end{bmatrix}$$

which may be solved for non-singular matrices A to obtain:

$$\begin{bmatrix} \alpha_l \\ \beta_l \end{bmatrix} = \frac{1}{\left(3T^2Z^2 - 2TZ^3\right)} \begin{bmatrix} 3T^2 & -Z^3 \\ -2T & Z^2 \end{bmatrix} \begin{bmatrix} 2\pi M_l - \Omega_l^0 - \omega_l^0 Z \\ \omega_l^T - \omega_l^0 \end{bmatrix} \tag{10}$$

The value of M_l is chosen according to the criterion described in [1] which minimises the total squared change in slope over the time segment from $t=0$ to $t=T$ when $0<Z<T$ or from $t=0$ to $t=Z$ when $Z>T$. It may be shown that the required value of M_l is equal to the following expression rounded to the nearest integer:

$$u_l = \frac{1}{2\pi}\left[\Omega_l^0 + \omega_l^0 Z + \left(\omega_l^T - \omega_l^T\right)\frac{Y^2}{2T}\right] \tag{11}$$

where Y is taken to be the largest of the two variables Z and T. Since the interpolation function must be defined at all three times, 0, Z and T, we shall refer to this method as '*0ZT*' interpolation.

Figure 2 shows a graph of the determinant of A, for various ratios of Z and T. It can be seen that a solution to (9) will exist provided that the selected value of Z is not equal to 0 or $^3/_2T$. When $Z = {}^3/_2T$, the rank of A becomes 1, indicating that the slope and value of any cubic polynomial at times T and $^3/_2T$ respectively are non-separable conditions. i.e. if one condition is met, the other cannot be assigned arbitrarily. In practice, the solutions for Z between $1.4T$ and $1.6T$ and around 0 are also not useful due to the degree by which the polynomial must contort in order to satisfy the conditions given by the vector, $\underline{y}_l$. In such cases an alternative interpolation strategy is required. One approach may be to use a higher order polynomial. However, in the current work, an improved form of interpolation-by-estimation has been adopted when the matrix A becomes ill-conditioned. This technique assumes the instantaneous frequency, rather than remaining constant between Z and T, is a linear function of t over the range 0 to Z. Therefore we assume that

$$\omega_l(t) = \omega_l^0 + \frac{\left(\omega_l^T - \omega_l^0\right)}{T} t \tag{12}$$

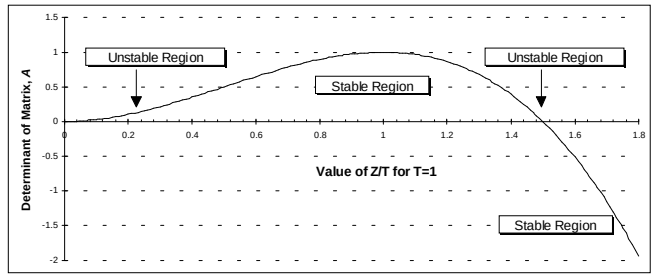

Figure 2: A graph of the normalised determinant of matrix A for ratios of Z and T from 0 to 1.8.

which, when substituted into (4) gives an estimate for Ω_l^T

$$\Omega_l^T = 2\pi\, M_l - \omega_l^0(Z - T) - \left(\omega_l^T - \omega_l^0\right)\frac{\left(Z^2 - T^2\right)}{2T} \tag{13}$$

Since we already know ω_l^0, Ω_l^0, and ω_l^T , cubic interpolation from 0 to T can be performed as described in [1]. This type of interpolation-by-estimation is valid for any values of Z and T and has been found to be more accurate than interpolation-by-linear-estimation. However, it is not as effective as the *0ZT* method in achieving maximum phase coherence for periods where reasonable *0ZT* solutions can be found.

The vocal system phase for sinewave l is now interpolated from 0 to T. Again cubic interpolation is used, since linear interpolation will impart an instantaneous frequency at the frame boundaries. The solution can be written as

$$\psi_l(t) = \psi_l^0 + \gamma_l t^2 + \delta_l t^3 \tag{14}$$

where γ_l and δ_l are chosen such that the instantaneous frequency, i.e. the slope of $\psi_l(t)$, is taken to be zero at times 0 and T.

Finally, the amplitude of the sinusoidal component l, $A_l(t)$ is found by linear interpolation of the measured peak amplitudes from 0 to T [1]. The modified synthetic speech signal is therefore given by

$$s(t) = \sum_{l=0}^{L-1} A_l(t)\cos\left[\Omega_l(t) + \psi_l(t)\right] \quad 0 \le t < T \tag{15}$$

5. RESULTS

As discussed in section 1, a useful method for measuring the degree to which the temporal structure of speech is preserved is by observing the glottal excitation $e(t)$. An approximation to $e(t)$ can be synthesised by removing the vocal system phase $\psi_l(t)$ from $s(t)$ and by assigning each harmonic a constant amplitude, C. i.e.

$$e(t) = \sum_{l=0}^{L-1} C\cos\left[\Omega_l(t)\right] \quad 0 \le t < T \tag{16}$$

Examples of time-scale modified waveforms of $e(t)$ for a female speaker are shown in figure 3 for both the *0ZT* interpolation method and the interpolation-by-linear-estimation method. It is

clear from this and other examples that greater preservation of the impulse-like nature of $e(t)$ occurs when the *0ZT* model is used, particularly when the variation of frequency of the sinewaves over a synthesis frame is large and/or different for each sinewave. For the cases where the chosen excitation point, Z, lies within the boundaries of the synthesis frame, maximum phase coherence is guaranteed, resulting in very high, symmetrical impulses. Although for the case where Z lies outside the synthesis frame, maximum phase coherence is not guaranteed, the resulting excitation signal is generally more impulse-like than that produced by interpolation-by-linear-estimation. An example of the speech produced by the two methods is shown in figure 4 a) and b). These are the excitation signals in figure 3, but with the amplitude and vocal system phase re-introduced. The original speech is shown in figure 4c) and the difference between the two methods is shown in figure 4d). Clearly there is a significant difference between the two methods, most noticeable is the difference due to the phase delay attributed to the misalignment of the sinusoids in b).

6. CONCLUSIONS

A model for shape-invariant speech modification has been presented which in general does not require an implicit estimate for excitation phase at synthesis frame boundaries. Although, under some circumstances, an estimate for phase is required, the technique by which this is achieved does not significantly degrade the performance of the model. The temporal structure of modified speech can be improved over that achieved by interpolation-by-linear-estimation due to its ability to better preserve the impulse-like nature of the glottal excitation. The result is highly shape-invariant pitch and time-scale modified voiced speech.

7. ACKNOWLEDGMENT

The authors wish to acknowledge the support of BT Laboratories and EPSRC in this work.

8. REFERENCES

[1] R. J. McAulay and T.F. Quatieri, “Speech analysis-synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 4, pp 744-754, Aug 1986.

[2] T. F. Quatieri and R. J. McAulay, “Speech transformations based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 6, pp 1449-1464, Aug 1986.

[3] R. J. McAulay and T.F. Quatieri, ”Shape invariant time-scale and pitch modification of speech,” *IEEE Trans. Signal Processing*, vol. 40, no. 3, pp 497-510, March 1992.

[4] D. B. Paul, “The spectral envelope estimation vocoder,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 4, pp 786-794, Aug 1981.

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech*, Englewood Cliffs: NJ: Prentice-Hall, 1978.

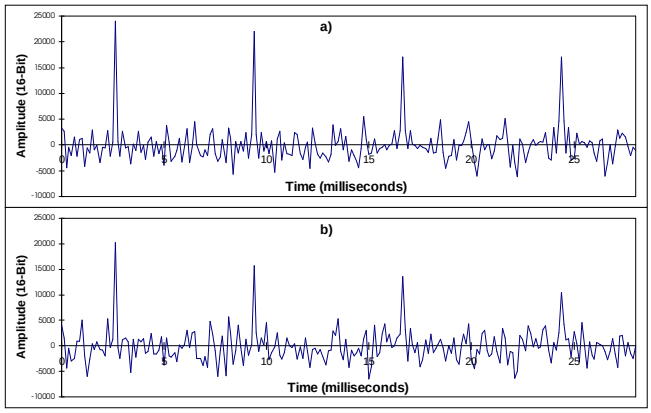

Figure 3: A comparison of the synthetic excitation signals $e(t)$ for time-scale modified female speech produced by a) The *0ZT* method and b) The interpolation-by-estimation method.

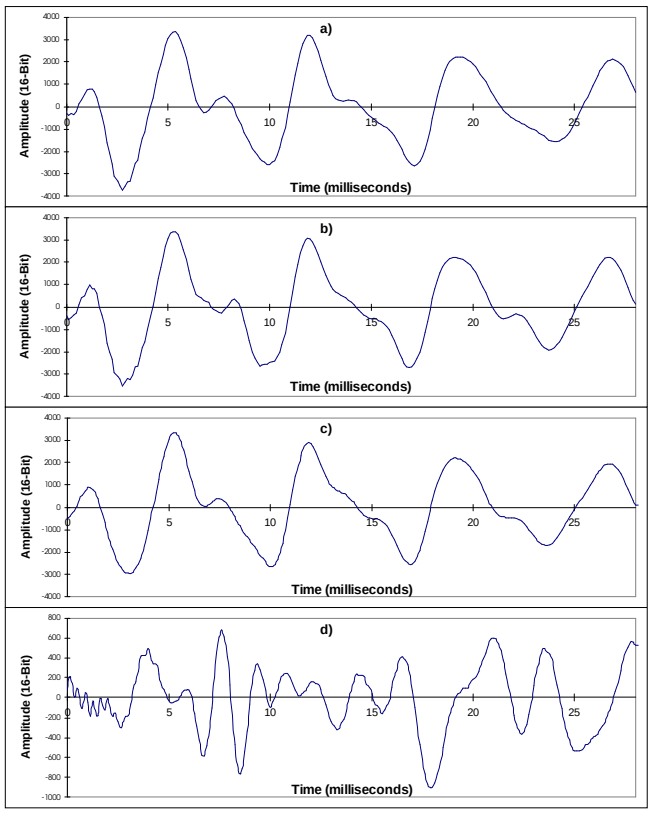

Figure 4: A comparison of the synthetic speech waveforms of time-scale modified female speech produced by a) The *0ZT* method and b) The interpolation-by-linear-estimation method. Also shown are c) The original speech signal and d) The difference between a) and b).