SHAPE-INVARIANT PITCH AND TIME-SCALE MODIFICATION OF SPEECH BY VARIABLE ORDER PHASE INTERPOLATION

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

[#]Department of Electrical Engineering and Electronics, The University of Liverpool, LIVERPOOL. L69 3BX. U.K.
matp@liv.ac.uk

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

ABSTRACT

To preserve the waveform shape and perceived quality of pitch and time-scale modified sinusoidally modelled voiced speech, the phases of the sinusoids used to model the glottal excitation are made to add coherently at estimated pitch pulse locations. The glottal excitation is therefore made to resemble a pseudo-periodic impulse train, a quality essential for shape-invariance. Conventional methods attempt to maintain the coherence once per synthesis frame by interpolating the phase through a single modified pitch pulse location, a time where all excitation phases are assumed to be integer multiples of 2π. Whilst this is adequate for small degrees of modification, the coherence is lost when the required amount of modification is increased. This paper presents a technique which is capable of better preserving the impulse-like nature of the glottal excitation whilst allowing its phases to evolve slowly through time.

1. INTRODUCTION

Many techniques currently exist for pitch and time-scale modification of speech, several of which are based on the sinusoidal framework [1] including the so-called "shape-invariant" techniques [2][3][4]. Attempts to modify the pitch or time-scale without preserving the waveform shape, have been found to produce speech which has a reverberent quality [5]. The temporal structure of voiced speech 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. The glottal excitation is therefore made impulse-like during voiced speech, an impulse occurring at each glottal closure time or excitation point. Achieving this phase relationship in synthetic speech is more difficult than may at 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 update points 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 [2] is, for each synthesis frame, to choose a suitable pitch and time-scale modified excitation point and then, for each sinusoid, to deduce a third order phase polynomial which is likely to have a value of $2\pi M$, for some integer M, at the time corresponding to the chosen excitation point. This technique involves estimating the phase at synthesis update points by linear interpolation from the chosen excitation point. Such an estimation, however, does not guarantee phase coherence at the excitation point and large errors in the estimate of excitation phase, which occur when the pitch is changing rapidly, can perceptually distort the synthetic speech. Moreover, for higher modification factors where the number of excitation points in a synthesis frame increases, attempting phase coherence at one single point may not be sufficient to preserve the shape. This is because the excitation phases are allowed to wander though the points at which phase coherence would normally be expected. If the synthesis frame is sufficiently long, the impulse-like nature of the excitation signal can be lost and perceptual reverberation returns.

In this paper a technique is presented which is capable of achieving maximum phase coherence at every excitation point in the synthesis frame by allowing the order of the interpolation polynomial to dynamically adapt to the specified pitch and time-scale modification requirements.

2. THE SHAPE-INVARIANT MODEL OF SPEECH

Commonly used models of speech production [6] 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 vocal system (i.e. 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

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

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

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 parameters, i.e.

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

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

3. ANALYSIS / SYNTHESIS

To synthesise pitch or time-scale modified voiced speech, the speech is first analysed and characterised in terms of the frequencies, amplitudes and vocal system phases of a set of slowly evolving sinewaves. These parameters are obtained by selecting a quasi-harmonically related set of peaks from an FFT magnitude spectrum [2][3] and computing the phase and magnitude components at each peak frequency. For the purposes of this work, analysis update points are made to coincide with excitation points [3][4], where all excitation phases are assumed to be multiples of 2π, so that in principle, the measured phase at each peak frequency is the vocal system phase. Representations of the magnitude and system phase envelopes are then derived at each analysis update point so that after pitch modification, a new set of vocal system magnitudes and phases may be obtained.

A synthesis frame is defined to be the region between a pair of time-scaled analysis update points. It is convenient to think of the frame boundaries to occur at times t=*0* and t=*T* relative to the start of the synthesis frame. All pitch-dependent parameters are pitch modified and the resulting sets of frequencies, amplitudes and vocal system phases at *t=0* are matched to their respective parameters at time *t=T* according to the "birth/death" process described in [1]. Assuming the position of the first excitation point in synthesis frame *k* is known, subsequent times up to and including the first excitation point in frame *k*+1 can be located by accumulating estimates of the modified pitch period *P'*(*t'*). Figure 1a) demonstrates this procedure. The next step is to determine the phases of each sinusoid at each of the *N* excitation points. By assuming the sinusoids are harmonically related, the number of 2π rotations of each sinewave between adjacent excitation points can be calculated. For example, the difference in phase between Z_1 and Z_2 for the fundamental is $1\times2\pi$. For the second harmonic, the difference is $2\times2\pi$ and so on. Hence for each sinewave, *l*, each excitation point Z_n can be associated with an integer value, $M_l(n)$ $(1 \le n \le N)$ which, when multiplied by 2π gives the phase for the sinewave at that excitation point.

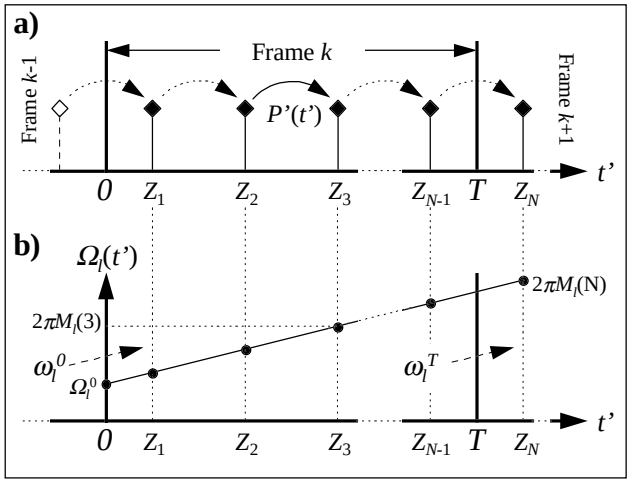

Figure 1. Excitation point selection and phase assignment for synthesis frame, *k*.

For continuity, the excitation phases of each sinewave at $t=0$, i.e. Ω_l^0, are known and their instantaneous frequencies at times $t=0$ and $t=T$ i.e. ω_l^0 and ω_l^T are the measured frequencies multiplied by the pitch modification factor, $q(t')$. Therefore, the problem is, for each sinewave l, to fit a smooth function $\Omega_l(t)$ such that the initial phase $\Omega_l(0)$ is Ω_l^0, the initial slope $\Omega_l'(0)$ is ω_l^0, the final slope $\Omega_l'(T)$ is ω_l^T, and the phases at each excitation point Z_1 to Z_N i.e. $\Omega_l(Z_1)$ to $\Omega_l(Z_N)$ equal $2\pi M_l(n)$. The problem is illustrated in figure 1b).

Although a solution for each sinewave may be found by solving a set of $N+3$ linear equations for an $N+2$ order system, it has been shown that specifying the slope of a polynomial and its value at independent times leads to ill-conditioned or even singular solutions [3]. Ill-conditioned solutions generate highly contorted phase polynomials and thus highly distorted speech. To avoid such problems, the order of the polynomial is increased by one. Therefore for N excitation points a polynomial of order $N+3$ is used. The following describes the interpolation procedure for an arbitrary sinewave l.

Having obtained the number of excitation points and the phases at each excitation point, an interpolation polynomial can be written as:

$$\Omega_l(t) = \Omega_l^o + \omega_l^o t + x_{l1}t^2 + x_{l2}t^3 + \ldots\ldots + x_{lN+2}t^{N+3} \qquad (4)$$

A matrix equation describing the required phases and instantaneous frequencies is derived as follows:

$$\begin{bmatrix} Z_1^2 & Z_1^3 & \cdots & Z_1^{N+3} \\ Z_2^2 & Z_2^3 & \cdots & Z_2^{N+3} \\ \vdots & \vdots & \ddots & \vdots \\ Z_N^2 & Z_N^3 & \cdots & Z_N^{N+3} \\ 2T & 3T^2 & \cdots & (N+3)T^{N+2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \\ x_{N+1} \\ x_{N+2} \end{bmatrix} = \begin{bmatrix} 2\pi M(1) - \Omega^0 - \omega^0 Z_1 \\ 2\pi M(2) - \Omega^0 - \omega^0 Z_2 \\ \vdots \\ 2\pi M(N) - \Omega^0 - \omega^0 Z_N \\ \omega^T - \omega^0 \end{bmatrix} \qquad (5)$$

or

$$A.\underline{x} = \underline{y} \qquad (6)$$

where the sinewave subscript "*l*" is omitted for convenience. For the polynomial to be realised, a solution for the vector $\underline{x}$ is required. Because the order is one greater than usual, the unknowns x_1 to x_{N+2} outnumber the independent equations. A consequence of this is that for every vector $\underline{y}$, there are an infinite number of vectors $\underline{x}$, which satisfy the conditions in matrix *A*. Given this extra degree of freedom, we therefore wish to choose the unique vector $\underline{x}$ which provides the smoothest solution across the synthesis frame. Using the smoothness criterion described in [1] we must therefore minimise

$$F(\underline{x}) = \int_0^{Z_N} \left[\ddot{\Omega}(t;\underline{x})\right]^2 dt \tag{7}$$

where $F(\underline{x})$ is the total squared change in slope from *0* to Z_N. $F(\underline{x})$ may be written as the quadratic form:

$$F(\underline{x}) = \underline{x}^T F \underline{x} \tag{8}$$

Using the Lagrange method of reduction [7], $F(\underline{x})$ may be expressed as a sum of squares

$$F(\underline{x}) = \sum_{i=1}^{N+2} X_i^2 \tag{9}$$

where

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_{N+2} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N+2} \\ 0 & p_{22} & \cdots & p_{2N+2} \\ 0 & 0 & \cdots & p_{3N+2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{N+2N+2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{N+2} \end{bmatrix} \tag{10}$$

or

$$\underline{X} = P.\underline{x} \tag{11}$$

Equation (6) may now be expressed in terms of the transformed vector $\underline{X}$.

$$G.\underline{X} = \underline{y} \tag{12}$$

where *G* is obtained by transforming of the matrix *A* as follows:

$$G = A.P^{-1} \tag{13}$$

The advantage of this transform is that a solution of (12) which minimises the sum of squares expression

$$\underline{X}^T \underline{X} = \sum_{i=1}^{N+2} X_i^2 = F(\underline{x}) \tag{14}$$

is given by

$$\underline{X} = G^{\#} \underline{y} \tag{15}$$

where $G^{\#}$ is the pseudo-inverse [7] of the non-square matrix *G*. The pseudo-inverse can be computed robustly using singular value decomposition [7]. The solution for $\underline{x}$ is then found from the inverse transformation:

$$\underline{x} = P^{-1} G^{\#} \underline{y} \tag{16}$$

This technique will be referred to as Variable Order Phase Interpolation (VOPI). By applying suitable interpolation schemes to the amplitude and vocal system phases [1][3] for each sinewave, the composite synthetic speech may be generated as follows:

$$s(t) = \sum_{l=0}^{L-1} A_l(t) \cos\left[\Omega_l(t) + \psi_l(t)\right] \quad 0 \le t < T \tag{17}$$

Where $\psi_l(t)$ and $A_l(t)$ are the instantaneous vocal system phase and amplitude respectively for sinewave *l*.

4. RESULTS

To demonstrate the effectiveness of this technique, the section of male speech shown in figure 2 was analysed and re-synthesised with its time-scale expanded by a factor of six for both VOPI and an adaptation of the shape-invariant technique described in [2]. The results from this experiment are shown in figure 3. A comparison with the original speech demonstrates the ability for VOPI to preserve the temporal structure of the original speech whereas the conventional technique tends to disperse the waveform resulting in reverberent speech. The reason for this can be seen by removing the vocal-system phases and assigning each sinewave a constant amplitude *C*. The result is an approximation of the excitation signal $e(t)$ given by:

$$e(t) = \sum_{l=0}^{L-1} C \cos\left[\Omega_l(t)\right] \tag{18}$$

This is shown in figure 4 for both methods. Since the conventional method attempts phase coherence only once every synthesis frame, much of the phase coherence is lost at the other excitation points in the frame. A large impulse does occur once every synthesis frame (six excitation points), but the potentially large distance between the chosen excitation point and the frame boundary prohibits complete alignment of the excitation phases, reducing the height of the impulse. Even under smaller modification factors, this mis-alignment can perceptually distort the speech. The VOPI technique, however, guarantees phase coherence at every excitation point resulting in a highly impulse-like excitation signal and natural sounding speech which is free from reverberation.

Although preserving phase coherence at every excitation point may be computationally intensive, experiments have shown that the order of the polynomial may be reduced to account for just one in every three excitation points without a noticeable degradation of shape or perceptual quality. For low to medium pitch and time-scale modification requirements, Quartic *OZT* interpolation [4], which uses a fourth order polynomial, has proven to be successful with a relatively low computational load.

5. CONCLUSIONS

An excitation phase interpolation technique has been presented which can ensure maximum phase coherence at any number of excitation points in a synthesis frame. No implicit estimation of

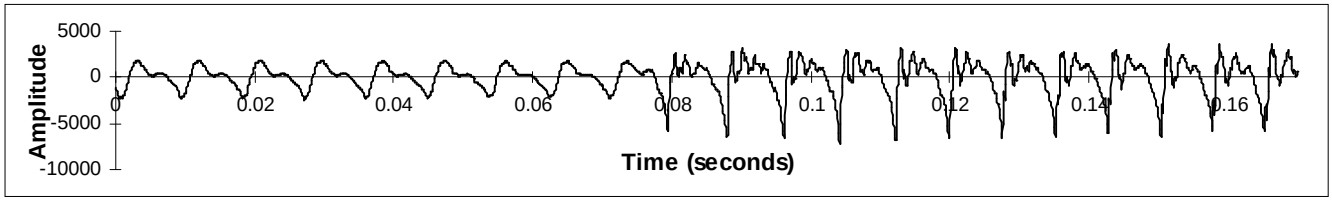

Figure 2. Original utterance of the transition /ɛ *i*:/ for a male speaker

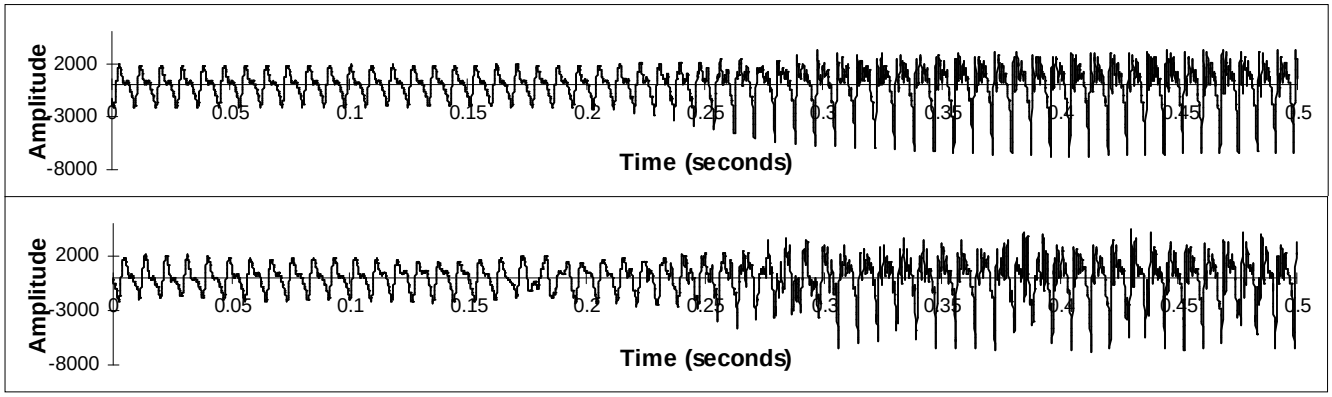

Figure 3. Reconstructed speech. Top: VOPI. Bottom: Conventional technique

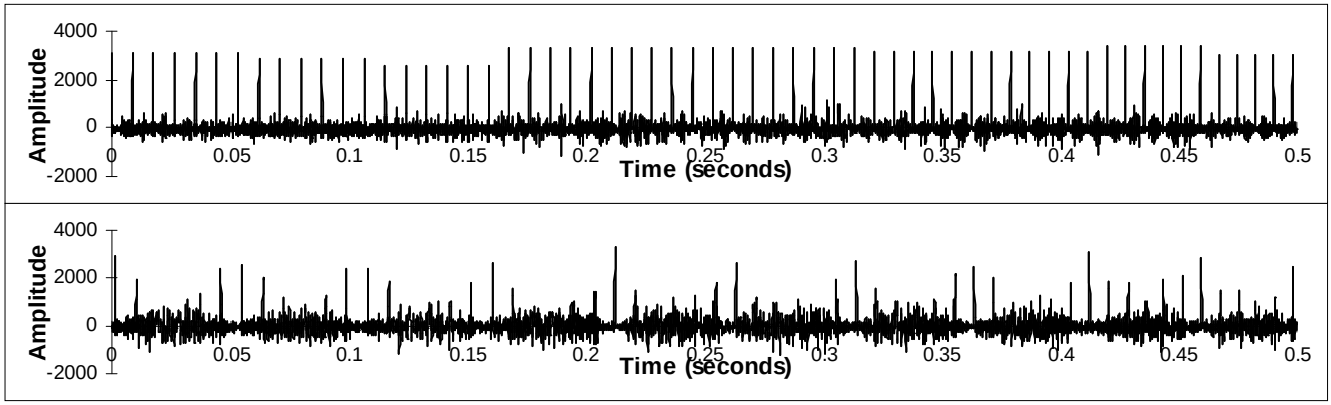

Figure 4. Reconstructed excitation signals Top: VOPI. Bottom: Conventional technique.

phase at synthesis frame boundaries is necessary. Whilst for large modification factors the computational burden may be high, the load may be reduced considerably without any degradation in quality by ignoring two out of every three excitation points. Synthetic speech produced using this method is natural sounding and of high perceptual quality. Investigations are under way to determine the benefits that this technique may give to other sinusoidally modelled applications.

6. 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] R. J. McAulay and T. F. Quatieri, "Shape Invariant Time-Scale and Pitch Modification of Speech" IEEE Trans. Acoust., Speech, Signal Processing, vol, ASSP-40, no. 3, pp 497-510, March. 1992.

[3] M. P. Pollard et al. "Enhanced Shape-Invariant Pitch and Time-Scale Modification for Concatenative Speech Synthesis" International Conference for Spoken Language Processing. Vol 3, pp 1429-1433. Oct. 1996.

[4] M. P. Pollard et al. "Phase Interpolation Methods for Pitch and Time-Scale Modification of Voiced Speech", The Institute of Acoustics (Speech & Hearing). Vol 18, pp 91-98. Nov. 1996.

[5] 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.

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

[7] P. Lancaster, Theory of Matrices, Academic Press, New York and London, 1969.

[8] G. W. Stewart, Introduction to Matrix Computations, Academic Press, New York and London, 1973.