Low Jitter Speech Synthesis Excitation Waveforms

Paul Milenkovic
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
1415 Johnson Drive
Madison, Wisconsin 53706 U.S.A.

December 8, 1986
Revised May 10, 1987

Abstract:

A glottal pulse waveform to be used as the excitation input to an LPC digital speech synthesizer is described by a parametric model. The model consists of a formula for a continuous time pulse waveform which is computed at equally spaced intervals to give the samples of a discrete time pulse waveform. This sampling operation will corrupt the resulting synthesized speech waveform with extraneous jitter, fundamental frequency perturbations, as well as shimmer, amplitude fluctuations, unless the original continuous time pulse waveform is bandlimited to prevent the aliasing that causes this jitter and shimmer. The pulse waveform can be bandlimited by applying a low pass filter. For a class of filters with continuous time impulse response $h(t)$, a closed form expression for the filtered glottal pulse waveform is obtained which allows for efficient computation of sample values.

Submitted for publication to IEEE Transactions on Acoustics, Speech, and Signal Processing

Contains Proprietary Information: Prepublication draft which may not be disclosed or distributed apart from permission of the author.
GLOSSARY (In order of appearance)

\[ p(t) \quad \text{Pulse waveform that starts at } t = 0 \text{ and ends at } t = T \]
\[ t \quad \text{Time variable} \]
\[ T \quad \text{Duration of pulse waveform } p(t) \]
\[ \tau \quad \text{Normalized time } t/T \]
\[ p_i(\tau) \quad \text{Polynomial basis functions on the interval } 0 \leq \tau \leq 1 \]
\[ n \quad \text{Polynomial order} \]
\[ m \quad \text{Number of component basis functions} \]
\[ b \quad \text{Vector of basis function weights} \]
\[ C_0 \quad \text{Matrix of coefficients used to express } p_i(\tau) \text{ in terms of } \tau \]
\[ C_1 \quad \text{Matrix of coefficients used to express } p_i(\tau) \text{ in terms of } \tau - 1 \]
\[ T_n \quad \text{Matrix which changes time scale from } [0,1] \text{ to } [0,T] \]
\[ u_t \quad \text{Vector of successive integrations of } \delta(t) \]
\[ u_{t-T} \quad \text{Vector of successive integrations of } \delta(t-T) \]
\[ h(t) \quad \text{Impulse response of the low pass filter applied to } p(t) \]
\[ h_t \quad \text{Vector of successive integrations of } h(t) \]
\[ h_{t-T} \quad \text{Vector of successive integrations of } h(t-T) \]
\[ q(t) \quad \text{Filtered pulse } h(t) * p(t) \]
\[ q_t(t) \quad \text{Part of } q(t) \text{ derived from first endpoint of } p(t) \]
\[ q_{t-T}(t) \quad \text{Part of } q(t) \text{ derived from second endpoint of } p(t) \]
\[ L() \quad \text{Integration operator} \]
\[ t_h \quad \text{Half sided duration of } h(t) \]
\[ k \quad \text{Fourier series order} \]
\[ x_t, y_t \quad \text{Vectors of elemental components of } h_t \]
\[ x_{t-T}, y_{t-T} \quad \text{Vectors of elemental components of } h_{t-T} \]
\[ H_x \quad \text{Matrix mapping } x_t, x_{t-T} \text{ into } h_t, h_{t-T} \]
\[ H_y \quad \text{Matrix mapping } y_t, y_{t-T} \text{ into } h_t, h_{t-T} \]
I. INTRODUCTION

A factor that needs to be taken into consideration in the design of a digital speech synthesizer is the pattern of fluctuations in fundamental period and waveform amplitude in voiced speech productions. Period-to-period fluctuations in the fundamental period are referred to as jitter while fluctuations in the amplitude of successive pitch periods are referred to as shimmer.

The levels of jitter and shimmer in the acoustic speech waveform are believed to influence the perceived voice quality. Support for this hypothesis is found in perceptual experiments which indicate that synthesized sounds with large amounts of jitter or shimmer are perceived as being rough or hoarse [1], [2]. These results lead us to seek methods of controlling jitter and shimmer in synthesized speech in order to reproduce the presence or absence of roughness and hoarseness as it occurs in natural speech.

To evaluate the ability of a digital speech synthesizer to reproduce the patterns of jitter and shimmer that occur in a natural utterance, we need to consider the synthesis method. A commonly used design for a digital speech synthesizer is one based on the source-filter model of vocal tract acoustics. The sound generation in the synthesizer is provided by either a noise source to model the sound of frication or by a source of periodic pulses to model the acoustic excitation signal produced by the larynx during voicing. The sound spectrum shaping in the synthesizer is provided by a filter which is patterned after the resonance structure of the vocal tract. The LPC synthesizer is an example of a source-filter synthesizer where the filter is the all-pole filter derived from the LPC coefficients [3].
The simplest structure we can use for the source waveform in a digital source-filter synthesizer is a train of discrete time impulses spaced at the voice fundamental period. With this type of excitation signal, however, we do not have full control over the level of jitter. For example, if the fundamental period is an integer multiple of the waveform sampling interval, it is possible to synthesize a source waveform with constant fundamental frequency and zero jitter. On the other hand, if the fundamental period is midway between two integer multiples of the sampling interval, the best we can do is to maintain the average of the desired fundamental period. The magnitude of the jitter, defined as the difference in fundamental period between two neighboring periods, will be equal to the sampling interval.

In some circumstances it may be reasonable to reduce jitter by rounding the pitch period value to integer multiples of the sampling interval. A synthesizer operating at a 10 kHz sampling rate can produce pitch increments of 1 Hz in the neighborhood of a 100 Hz fundamental frequency. While experiments with constant pitch sounds [13], [14] indicate a pitch discrimination threshold of .3 Hz, experiments conducted by Klatt on sounds with varying pitch [14] indicate a threshold in the vicinity of 2 Hz. The pitch difference caused by rounding the pitch period may not be noticeable in conversational speech.

In other circumstances it may be necessary to maintain an average value of the pitch period which is a fractional multiple of the sampling interval. In music synthesis of the singing voice, accurate control over pitch is required in order to maintain harmonies without introducing beat frequencies which are characteristic of mistuning. In the multi-pulse scheme for speech synthesis [15], where the objective is to control the synthesizer to track the acoustic waveform of individual pitch period
cycles of a natural speech production, the average pitch period of a sustained pitch portion of an utterance needs to be maintained. If the pitch period of a constant pitch segment is rounded to a constant integer number of samples, the synthesizer pitch periods will drift out of phase with the target utterance pitch periods.

While rounding the pitch period can result in zero jitter synthesis with a shift in overall pitch that is below audible thresholds, we may want to employ small but not zero levels of jitter in our synthesis. Measurements by Horii [4] indicate that the mean jitter level of normal male subjects can be under 50 microseconds, and a sampling rate of 40 kHz or more is required in order to measure these levels of jitter. A more recent study finds instances of jitter levels as low as 10 microseconds [5]. If we operate a synthesizer at a 10 kHz rate, however, the sample interval will be 100 microseconds, which is large compared to the jitter we are trying to reproduce.

The perceptual effects of changes in the pattern of jitter is a matter that has not been thoroughly investigated. For example, it is possible to reproduce a low average level of jitter even though the step changes in the pitch period are large. This is accomplished by separating step changes in the pitch period with a large number of waveform cycles with zero jitter. The peak jitter, however, will remain large. It is not known whether two sound productions with the same average jitter but different peak jitter are perceptually distinguishable. Finding an answer to this question will require a series of perceptual experiments concerning different vowel, speaking rate, and word contexts. In this paper, however, we seek to address the issue of synthesis techniques which are appropriate for conducting these studies.
It is theoretically possible to synthesize discrete time source waveforms with pitch periods that are fractional multiples of the sampling interval. If we have a means for computing the sample values of bandlimited continuous time pulses, it is possible to have continuous control of the pitch period and control of jitter down to the zero level. To see why this is so, one must consider that the discrete time output waveform of the synthesizer, if we are ever going to listen to it, is converted to a continuous time waveform by use of a digital-to-analog converter and a low pass signal reconstruction filter. In the ideal bandlimited case, the final continuous time waveform output to the synthesizer will be identical to that of an equivalent analog synthesizer which is excited directly by the continuous time pulse waveforms. Because we can exercise continuous control over the time spacing of pulse waveforms expressed in continuous time, we have complete control over the jitter pattern of the synthesizer output.

An approximate implementation of this concept is described by Hartwell and Prezas [6]. A prototype bandlimited excitation pulse is computed at closely spaced time samples and stored in a lookup table. The sampling rate of the stored waveform can be made much higher than the sampling rate at which the synthesizer operates. This saves the expense of operating the synthesizer at very high rates (> 40 kHz). Samples of the synthesizer excitation waveform are determined by retrieving entries from the lookup table. The minimum level of jitter is related to the shorter sampling interval of the lookup table rather than the longer sampling interval used by the synthesizer.
The pulse shape employed in this type of synthesizer is perceptually important. A study performed by Rosenberg [7] established that different pulse shapes can be perceptually discriminated and that listeners prefer a pulse shape patterned after the pulse shapes observed in inverse filter measurements of the glottal airflow signal. Titze [8] has synthesized the singing voice using a laryngeal model to produce pulse shapes typical of human glottal airflow patterns, and informal listening indicates a high quality of synthesis. Sambur, et al. [16] report that a pulse shape that differs from a pure impulse helps reduce the "buzzy" quality of LPC speech synthesis. They offer the explanation that the pulse they use introduces "irregularity" to the pitch period of synthesized waveform. By irregularity, they mean that the vowel waveform shape of a single pitch period is less peaked rather than any fluctuation in the waveform in the form of jitter or shimmer.

The Hartwell and Prezas synthesizer has the advantage of requiring few computations. It has the disadvantage of using a fixed pulse shape, which cannot be adjusted to reflect changes in the glottal pulse shape over the course of an utterance. The laryngeal model synthesizer has the advantage of being able to produce glottal pulses of varying shape in response to changes in the laryngeal parameters. It has the disadvantage of computational complexity. Another drawback of the laryngeal model is that it while it computes samples of a continuous time pulse waveform, the underlying continuous time waveform is not bandlimited on account of the discontinuities in glottal airflow that occur at the opening and the closing of the glottis. Titze [8] has reported that even at a 20 kHz sampling rate, a laryngeal model with abrupt opening and closing of the glottis resulted in a noticeably adverse perceptual quality. Superior synthesis quality was obtained using a laryngeal model with tapered glottal
opening and closing events.

Applying a taper to the glottal opening and closing discontinuities in the laryngeal model serves to reduce the amount of aliasing resulting from sampling. The taper may also change the spectral slope of the excitation signal. It is difficult to determine whether the improved voice quality is the result of eliminating aliasing or is the result of the change in the source spectral slope.

This situation has motivated consideration of a technique for bandlimiting glottal pulse waveforms of variable shape using a low pass filter of controllable properties rather than relying on an ad hoc procedure such as tapering the ends of the pulse. This technique is applicable to a source-filter model synthesizer structure such as the LPC synthesizer. This paper presents an analytical derivation of the filtered glottal pulse where the filtering is done in continuous time. An expression is found which allows us to compute sample values of the filtered pulse for pitch periods that are a fractional number of waveform samples.
II. METHODS

We first consider a description of the glottal waveform over the open glottis interval by a polynomial function of time. It has been shown experimentally that polynomial functions can give a good fit to the glottal airflow pulses estimated by inverse filtering the speech wave and that the polynomial pulses that are fitted to the inverse filter wave can be used to synthesize an approximation to the acoustic speech waveform [9] in a manner akin to the multi-pulse technique [15]. In addition, our physical intuition indicates that the glottal wave as well as its derivatives should be continuous over the open glottis interval. A waveshape of this type lends itself to a power series expansion. If we truncate the expansion to the first several terms, we obtain a polynomial.

Next, we derive a closed form result for the convolution of a polynomial pulse with a continuous time filter obtained from a frequency sampling design [10]-[12]. Frequency sampling designs lend themselves to low pass filters with stop band attenuation in excess of 80 dB [12], which should give an adequate reduction of aliasing in most applications.

Pulse Waveform

Consider a pulse waveform $p(t)$ which starts at time $t = 0$ and ends at time $t = T$. The excitation waveform for a digital speech synthesizer will be obtained for each pitch period by computing $p(t-t_d)$ at discrete values of $t$, where $t_d$ marks the beginning of the opening phase of the glottal pulse for that pitch period. Expressing $p(t)$ as a function of continuous time means that values for the pulse origin $t_d$ and pulse length
T of \( p(t-t_d) \) can be varied continuously.

In order to exercise independent control over the shape and duration of the pulse waveform \( p(t) \), we express it as the linear combination

\[
p(t) = b_1 p_1(t/T) + \ldots + b_m p_m(t/T)
\]

where the components \( p_i(t/T) \) are pulse waveforms of the form

\[
p_i(\tau) = C_0[i,0] + C_0[i,1] \tau + \ldots + C_0[i,n] \tau^n
\]

\[
= C_1[i,0] + C_1[i,1] (\tau - 1) + \ldots + C_1[i,n] (\tau - 1)^n
\]

\[0 \leq \tau \leq 1\]

for \( \tau = t/T \) where \( C_0[i,j] \) and \( C_1[i,j] \) are constant coefficients. That the pulse \( p(t) \) is zero outside the interval \([0,T]\) results in the identity

\[
p(t) = p(t)u(t) - p(t)u(t-T)
\]

where \( u(t) \) is the unit step function. Combining the previous three expressions gives

\[
p(t) = b_0 T_n u_{t+T} - b_1 T_n u_{t-T}
\]

where \( C_0 \) is a matrix with elements \( C_0[i,j] \), \( C_1 \) is a matrix with elements \( C_1[i,j] \), \( b \) is the vector \([b_1, \ldots, b_m]\), \( u_t \) is the vector \([u(t), u(t), \ldots, t^0 u(t)]\), \( u_{t+T} \) is the vector \([u(t-T), (t-T) u(t-T), \ldots, (t-T)^{T_0} u(t-T)]\), and \( T_n \) is a diagonal matrix with main diagonal elements \( 1/T_i \) for \( 0 \leq i \leq n \).
In an LPC speech synthesizer, it is often convenient to model the vocal tract transfer characteristic by the LPC forward filter and to model the radiation load transfer characteristic as a differentiator [3]. We can save the step of explicitly implementing this differentiator if the pulse input to the LPC forward filter is the time derivative of the glottal airflow pulse.

If the pulse p(t) which is sampled to form the input to the LPC forward filter is the first time derivative of the glottal pulse, the pulse p(t) will have a number of properties which can be derived from the properties of the glottal pulse. We will assume that the glottal flow starts at zero at the leading edge of the glottal pulse and ends at zero at the trailing edge. In the case where the glottis does not close completely on account of a glottal chink, for purposes of this model we will assume that glottal flow starts from and returns to the same constant value. We will also assume that the glottal waveform and its derivatives are continuous except at the points of glottal opening and glottal closing, where the derivatives of glottal flow can be discontinuous.

We will assume that the dominant aspects of the excitation signal which influence the speech spectrum are the pulse duration T as well as the discontinuities in the first two derivatives of the glottal wave at the glottal pulse endpoints. In the case of a glottal pulse which has a continuous first derivative between its endpoints, the asymptotic frequency spacing of the Laplace transform zeroes is controlled by the pulse duration T while the asymptotic damped of these zeroes is controlled by the ratio of the closing edge to the opening edge slopes [18], [20]. In our model we seek to control two pulse derivatives, the pulse endpoint slopes as well as the pulse endpoint curvatures, in order to have additional control over the pulse shape. As a consequence of pulse p(t) being the first derivative of
the glottal airflow pulse, we seek to control the magnitude of the discontinuities of \( p(t) \) and \( p'(t) \) at \( t = 0 \) and \( t = T \) by proper choice of the basis functions \( p_1(t/T) \).

We seek four basis functions, \( p_1(\tau) \) through \( p_4(\tau) \), of minimum polynomial order with which to exercise independent control of the boundary values \( p(0), p'(0), p(T), p'(T) \) by way of the weight vector \( b \). In order for the glottal pulse to begin and end at the same flow for all possible weight vectors \( b \), we will require that the DC value of \( p_1(\tau) \) through \( p_4(\tau) \) be zero. The minimum order basis functions under these conditions are polynomials of order \( n = 4 \).

In order to determine the coefficients of polynomials \( p_1(\tau) \) with desired boundary values, we note that the coefficients \( C_0 \) are linearly related to the boundary values of \( p_1(\tau) \) according to

\[
\begin{align*}
p_1(1) &= C_0[i,0] + C_0[i,1] + C_0[i,2] + C_0[i,3] + C_0[i,4] \\
p_1'(1) &= C_0[i,1] + 2C_0[i,2] + 3C_0[i,3] + 4C_0[i,4] \\
p_1(0) &= C_0[i,0] \\
p_1'(0) &= C_0[i,1] \\
DC &= C_0[i,0] + 1/2 C_0[i,1] + 1/3 C_0[i,2] + \frac{1}{4} C_0[i,3] + \frac{1}{5} C_0[i,4]
\end{align*}
\]

If we specify the DC value of \( p_1(\tau) \) to be zero, coefficients of the basis functions can be specified according to the setting of the boundary values \( p_1(0), p_1'(0), p_1(1), p_1'(1) \). Setting all but one of the four boundary values of each polynomial to zero results in the basis functions
\[ p_1(\tau) = 1 - 18\tau^2 + 32\tau^3 - 15\tau^4 \quad p_1(0) = 1 \]
\[ p_2(\tau) = \tau - 4.5\tau^2 + 6\tau^3 - 2.5\tau^4 \quad p_2'(0) = 1 \]
\[ p_3(\tau) = -12\tau^2 + 28\tau^3 - 15\tau^4 \quad p_3(1) = 1 \]
\[ p_4(\tau) = -1.5\tau^2 + 4\tau^3 - 2.5\tau^4 \quad p_4'(1) = -1 \]

It follows that in the formula \( p(t) = b C_0 T_n u_t^T - b C_1 T_n u_{t-T}^T \) that \( n = 4 \) and that the matrix \( C_0 \) is given by

\[
C_0 = \begin{bmatrix}
1 & 0 & -18 & 32 & -15 \\
0 & 1 & -4.5 & 6 & -2.5 \\
0 & 0 & -12 & 28 & -15 \\
0 & 0 & -1.5 & 4 & -2.5
\end{bmatrix}
\]

The four basis polynomials have the symmetry property \( p_1(-\tau) = p_3(\tau-1) \) and \( p_2(-\tau) = p_4(\tau-1) \). From this symmetry and the coefficients for the \( C_0 \) matrix it follows that the \( C_1 \) matrix is given by

\[
C_1 = \begin{bmatrix}
0 & 0 & -12 & -28 & -15 \\
0 & 0 & -1.5 & -4 & -2.5 \\
1 & 0 & -18 & -32 & -15 \\
0 & -1 & -4.5 & -6 & -2.5
\end{bmatrix}
\]
Filtered Pulse Waveform

We seek to filter \( p(t) \) by convolving it with the filter impulse response \( h(t) \). If \( h(t) \) is the response of a suitably chosen low pass filter, this operation serves to prevent aliasing when we sample \( p(t) \) to form an excitation signal for a digital speech synthesizer. The waveform \( q(t) = h(t) * p(t) \) is expressed as

\[
q(t) = q_t(t) - q_{t-T}(t) = b c_0 T_n h^T_t - b c_1 T_n h^T_{t-T}
\]

where \( h_t \) is the vector \([h(t)*u(t), h(t)*t u(t), \ldots, h(t)*t^nu(t)]\) and \( h_{t-T} \) is the vector \([h(t)*u(t-T), h(t)*u(t-T), \ldots, h(t)*(t-T)^nu(t-T)]\).

Suppose the low pass filter applied to the pulse \( p(t) \) is restricted to a functional form \( h(t) \) that is \( n+1 \) times integrable in closed form. A linear operator \( L() \) is defined according to

\[
L(g(t)) = \int_{-\infty}^{t} g(u)du
\]

Applying this operator to the unit delta function gives

\[
\begin{align*}
  u(t) &= L(\delta(t)) \\
t u(t) &= L(L(\delta(t))) \\
t^n u(t) &= n! L^{n+1}(\delta(t))
\end{align*}
\]

Because \( h(t) * \delta(t) = h(t) \) and because the convolution operator can be factored inside the linear operator \( L() \), it follows that

\[
 h(t)*t^i u(t) = i! L^{i+1}(h(t))
\]
where \( L^{i+1}(h(t)) \) can be expressed in closed form under the requirement that \( h(t) \) be \( n + 1 \) times integrable. This means that the vector \( h_t \) has the closed form expression \([L(h(t)), \ldots, L^{i+1}(h(t))]\) and the vector \( h_{t-T} \) has the closed form expression \([L(h(t-T)), \ldots, L^{i+1}(h(t-T))]\), which implies a closed form expression for \( q(t) = h(t)p(t) \). This closed form expression allows us to compute values of \( q(t) \) at discrete values of \( t \) for use in the digital synthesizer.

One way to specify a filter with an integrable impulse response is to employ a frequency sampling design [10]. The impulse response for such a filter is given by

\[
h(t) = h_0 + h_1 \cos(\pi t / t_h) + \ldots + h_k \cos(k \pi t / t_h)
\]

\[-t_h < t \leq t_h\]

In order to evaluate \( L^{i+1}(h(t)) \), we have to consider the regions \( t \leq -t_h \), \(-t_h < t \leq t_h \), and \( t_h < t \) separately. For \( t \leq -t_h \), it is obvious that \( L^{i+1}(h(t)) \) is zero. For \(-t_h < t \leq t_h \), integrating \( h(t) \) \( i \) times will produce terms involving powers of \( t \) up to \( i+1 \) as well as terms involving sines and cosines. For the range \( t > t_h \), \( L^{i+1}(h(t)) \) will only have terms involving powers of \( t \) up to the power \( i \). These terms result from repeated integration of \( h(t) \) over an interval where \( h(t) = 0 \) with non-zero initial conditions. For example, \( L(h(t)) \) will have a constant value \( L(h(t_h)) \) for \( t > t_h \) while \( L2(h(t)) \) would be given by \( L(h(t_h))t + L^2(h(t_h))t_h \) and so on for successive integrations. The results of performing the integrations can be stated formally as

\[
h^T_t = H_x x^T_t \quad -t_h < t \leq t_h
\]

\[
h^T_t = H_y y^T_t \quad t_h < t
\]
where \( H_x \) is a matrix of elements \( H_x[i,j] \) where \( 0 \leq i \leq n, 0 \leq j \leq n+1+2k \), \( x_t \) is the vector \([1, t, \ldots, t^{n+1}, \cos(\pi t/t_h), \ldots, \cos(k \pi t/t_h), \sin(\pi t/t_h), \ldots, \sin(k \pi t/t_h)]\), \( H_y \) is a matrix of elements \( H_y[i,j] \) where \( 0 \leq i, j \leq n \), and \( y_t \) is the vector \([1, t, \ldots, t^n] \). Likewise

\[
\begin{align*}
h_{t-T}^T &= H_x x_{t-T}^T & T-t_h < t \leq T+t_h \\
&= H_y y_{t-T}^T & T+t_h < t
\end{align*}
\]

where the vectors \( x_{t-T} \) and \( y_{t-T} \) are formed by substituting \( t-T \) for \( t \) into the expressions for \( x_t \) and \( y_t \).

At this point we consider means of computing the coefficients of \( H_x \) and \( H_y \). Because they are constants, these coefficients need only be computed once and then stored in a table. The coefficients of \( H_x \) and \( H_y \) can be computed according to the rules of symbolic integration. We start by setting \( H_x[-1,j] \) and \( H_y[-1,j] \) to

\[
\begin{align*}
H_x[-1,0] &= h_0 \\
H_x[-1,j] &= 0 & 1 \leq j \leq n + 1 \\
H_x[-1,n+1+j] &= h_j & 1 \leq j \leq k \\
H_x[-1,n+1+k+j] &= 0 & 1 \leq j \leq k \\
H_y[-1,j] &= 0 & 0 \leq j \leq n
\end{align*}
\]

where \( n \) is the polynomial order of \( p(t) \), \( k \) is the number of cosine terms in \( h(t) \) and \( h_j \) are the coefficients of the cosine terms in \( h(t) \). These terms are not themselves elements of the matrices \( H_x \) and \( H_y \) but are rather initial conditions of a recursion used to generate \( H_x \) and \( H_y \). The fact that \( d(t^j/j)/dt = t^{j-1} \) leads to the recursion
\[ H_x[i,j] = (i_1/j)H_x[i-1,j-1] \quad 1 \leq j \leq n+1 \quad 0 \leq i \leq n \]
\[ H_y[i,j] = (i_1/j)H_y[i-1,j-1] \quad 1 \leq j \leq n \quad 0 \leq i \leq n \]

where \( i_1 - i \) for \( i > 0 \) and \( i_1 - 1 \) for \( i = 0 \). That \( \frac{d((t_i/j_1) \sin(k \pi t_i/t_h))}{dt} = \cos(k \pi t_i/t_h) \) and \( \frac{d(-(t_i/j_1) \cos(k \pi t_i/t_h))}{dt} = \sin(k \pi t_i/t_h) \) for \( 0 \leq i \leq n \) leads to

\[ H_x[i,n+1+k+j] = (i_1t_h/j_1)H_x[i-1,n+1+j] \quad 1 \leq j \leq k \]
\[ H_x[i,n+l+j] = -(i_1t_h/j_1)H_x[i-1,n+l+k+j] \]

What remain to be specified are \( H_x[1,0] \) and \( H_y[1,0] \). These are the DC terms, and their values are determined by requiring continuity of the elements of the vector \( h_i \) at \( t = -t_h \) and at \( t = t_h \). From this it follows for \( 0 \leq i \leq n \) that

\[ -H_x[i,0] = (-t_h)H_x[i,1] + \ldots + (-t_h)^{n+1}H_x[i,n+1] + \]
\[ (-1)H_x[i,n+2] + \ldots + (-1)^kH_x[i,n+l+k] \]

\[ H_y[i,0] = H_x[i,0] + t_hH_x[i,1] + \ldots + t_h^{n+1}H_x[i,n+1] + \]
\[ (-1)H_x[i,n+2] + \ldots + (-1)^kH_x[i,n+l+k] - \]
\[ (t_hH_y[i,1] + \ldots + t_h^{n}H_y[i,n]) \]

We finally arrive at the point where \( q(t) \) can be evaluated in terms of its component parts \( q_t(t) \) and \( q_{t-T}(t) \) where

\[ q_t(t) = b C_0 T_n H_x x_t^T \quad -t_h < t \leq t_h \]
\[ = b C_0 T_n H_y y_t^T \quad t_h < t \]
\[ q_{t-T}(t) = b C_1 T_n H_x x_{t-T}^T \quad T-t_h < t \leq T+t_h \]
We do not need to consider \( q_{t-T}(t) \) for the interval \( T+t_h < t \) because \( p(t) = 0 \) for \( T < t \) implies \( q(t) = 0 \) for \( T+t_h < t \) on account of the finite duration of impulse response \( h(t) \), which obviates the need of explicitly computing \( q(t) \) or its components.

The terms \( b \ C_0 \ T_n H_x \), \( b \ C_0 \ T_n H_y \), and \( b \ C_1 \ T_n H_x \) are vectors which need to be computed once for each new pulse waveform \( p(t) \) which requires a new duration \( T \) and new pulse shape coefficients \( b \). For each time sample within the duration of the pulse, the evaluation of \( q(t) \) requires computing either one or two vector dot products, depending on the overlap between \( q_t(t) \) and \( q_{t-T}(t) \). The dot products are of length \( n + 2k + 2 \) or of length \( n + 1 \), depending on position within the pulse waveform.

The evaluation of \( x_t \) for each time sample requires \( n - 1 \) multiplications to compute the powers of time \( t \). The evaluation of \( y_t \) requires \( n \) multiplications to compute the powers of \( t \), plus it needs the evaluation of \( 2k \) sine and cosine terms. Because the digital synthesizer requires input samples at equally spaced time intervals, the \( 2k \) sine and cosine terms need only be computed for each pulse \( p(t-t_d) \) once for the initial sample of \( x_t \) as well as once for the initial sample of \( x_{t-T} \). Subsequent time samples of the sine and cosine terms can be obtained by performing \( k \) complex multiplies using the identities

\[
  e^{jk\omega t} = \cos(k\omega t) + j \sin(k\omega t) \\
  e^{jk\omega(t+1)} = e^{jk\omega t} e^{jk\omega}
\]

where \( j \) in this expression is the square root of \(-1\). Within each pulse, the cost of evaluating samples of \( q(t) \) is bounded by a constant multiple of \( 2k+n \). For each new pulse, we incur the cost of computing initial conditions requiring evaluation of \( 4k \) sine and cosine terms plus computation of the matrix multiplies.
III. RESULTS

Examples of a voice source pulse waveform computed using the proposed model are shown in Fig. 1. The top pulse is computed without the use of a low pass filter \( h(t) \). The middle pulse is computed with a low pass filter \( h(t) \) that is a \( k = 4 \) low pass frequency sampling design employing a single transition sample [10]. The bottom pulse in the figure is computed with a filter \( h(t) \) chosen to be a long duration raised cosine pulse. The waveform overshoot and ringing observed in the middle pulse is characteristic of the output of a low pass filter with a steep frequency cutoff while the smooth transitions in the bottom pulse are characteristic of a filter with a broad frequency cutoff.

The voice source pulses can be integrated to produce the glottal airflow pulse waveforms depicted in Fig. 2. The glottal pulses produced by the waveform model have slope discontinuities at the start and finish of the open cycle of the pitch period. The waveform increases monotonically, reaches a peak, and then decreases monotonically. The pulse waveform is skewed towards the closing edge to represent the effect of inductive loading of the vocal tract on the glottal pulse discussed by Rothenberg [19]. The effect of the different \( h(t) \) filters on the glottal pulse is less apparent than the effect on the voice source pulse though the bottom pulse shows considerable rounding of the corners.

The effect of the \( h(t) \) filter on the Fourier spectra of the voice source pulses is shown in Fig. 3. These spectrum plots are for the case of a 10 kHz waveform sampling rate. The \( k = 4 \) frequency sampling filter gives a good match to the spectrum of the unfiltered pulse up to about 3 kHz, where the filter begins to roll off. As is the case with an analog anti-alias filter, this corner frequency can be increased while keeping the same
level of anti-alias protection by using a more complex filter with a larger
value of k.

In addition to serving as an anti-alias filter, the h(t) filter can be
used to control the spectral slope of the voice source. As shown by the
bottom waveform plot in Fig. 2 and the bottom spectrum plot in Fig. 3, the
raised cosine h(t) filter which rounds the pulse corners has the effect of
increasing the spectral rolloff of the voice source waveform. This change
in the spectrum slope is accomplished without changing the glottal pulse
shape appreciably apart from the corners. This supports the conclusion
that the pulse opening and closing transitions are major contributors to
the source spectrum.

We used the voice source pulse model to synthesize constant pitch
synthetic vowels for purposes of measuring jitter and shimmer. These
results are summarized in Table I. We employed the same pulse shape and
h(t) filter as shown in the middle plots of Figs. 1-3. The open portion of
the glottal cycle was kept at a constant proportion of .55 of the pitch
period. The pulse coefficients were \( b_1 = 50, b_2 = 0, b_3 = .300, \) and \( b_4 =
1000. \) The h(t) filter coefficients were \( h_0 = .5, h_1 = 1, h_2 = 1, h_3 = 1, \)
and \( h_4 = .3904, \) which is a single transition sample frequency sampling
design. A waveform sampling rate of 10 kHz was used, and the source
waveform was fed into a 12 coefficient LPC synthesizer. The LPC
coefficients were obtained by the covariance method analysis without

Three synthesis conditions were used. The first condition is
synthesis without the use of the h(t) filter where the voice source pulse
were computed pitch synchronously. In the pitch synchronous condition, the
pitch period is rounded to an integer number of samples. The pitch period

18
is then adjusted from period to period to maintain the average pitch period value. The second condition is synthesis without the h(t) filter where the voice source pulse is computed asynchronously with a pitch period that is not an integer number of sample intervals. The voice source pulse is treated as a continuous time waveform which is sampled without regard for the resulting aliasing. In the third condition, the voice source pulse is computed asynchronously, but this time with the benefit of the anti-alias filter h(t).

Synthesis pitch periods were selected from two ranges. The range around 8 ms is representative of a male voice while the range around 4 ms is representative of a female voice. For each range, pitch period values were selected that were non-integer multiples of the sampling interval. The values 8.05 ms and 4.05 ms are midway between integer multiples of the sampling interval, and we expect synthesis jitter to be greatest at these values.

The values of jitter and shimmer are averages computed from a 250 ms sustained vowel portion of the synthesizer output. Jitter and shimmer are obtained by computing the short term autocorrelation function and using parabolic interpolation to find the peak [5]. The interpolated peak of the short term autocorrelation function provides an accurate measure of pitch period from which to compute jitter and the amplitude of the peak is used to compute shimmer. In the synthesis conditions without an h(t) filter we observe considerable shimmer, the value of which varies with the pitch range as well as the synthesis method. The variation of shimmer with pitch period is attributable to the overlap of the vocal tract impulse response between successive pitch periods [5]. Considerable reduction in both jitter and shimmer is obtained by using the h(t) filter to compute the voice source pulses.
IV. DISCUSSION

We have described a method for operating on a pulse waveform, expressed in terms of polynomial basis functions, with a filter specified in terms of a frequency sampling finite impulse response design. The filtering is conducted in continuous time, resulting in analytic expressions for the filtered pulse as a function of time.

This technique allows us to express values of a discrete time speech synthesizer excitation signal as samples of a low pass filtered continuous time glottal waveform model. The continuous time waveform model allows us to exercise continuous control over pulse duration and pulse location. By being able to specify a low pass filter independently from the waveform model, we can suppress aliasing in the conversion of this continuous time waveform to a discrete time waveform and back to continuous time.

The new pulse synthesis technique is applicable to the source-filter type of synthesizer where the voice source waveform is computed independently from a vocal tract filter used to shape the spectrum of the voice source waveform to produce the acoustic speech waveform. In the human voice, however, the glottal airflow waveform is influenced by the interaction of the glottal area function waveshape with the acoustic loading of subglottal and supraglottal portions of the airway [19], [21]-[23].

The source-tract interaction is responsible for two features of the glottal airflow pulse. The skewing of the glottal airflow pulse relative to a more symmetric glottal opening area pulse as observed in simultaneous inverse filtering and laryngeal imaging experiments [22] can be accounted for by the inductance of the subglottal and supraglottal air column [19]. The proposed glottal pulse model is able to produce skewed pulse shapes to
represent this effect. The second glottal pulse feature derived from source-tract interaction is formant ripple superimposed on the glottal airflow pulse [23]. This formant ripple is not accounted for by a low polynomial order \((m = 4)\) voice source pulse model.

The true glottal pulse with formant ripple can be replaced by a smooth glottal pulse where the vocal tract formant bandwidths are regarded as changing over the duration of the pitch period [18], [23]. This suggests that the acoustic effect of glottal formant ripple could be modelled in a source-filter synthesizer by changing parameters of the vocal tract filter synchronously with the glottal pulse. The low polynomial order source pulse model may be adequate to account for the effect of source-tract interaction on the speech waveform if the LPC coefficients of the vocal tract filter are adjusted to reflect time varying formant bandwidths.

Using a fixed vocal tract filter in a source-filter synthesizer, Holmes [17] reported good speech quality where the voice source pulse was a periodically replicated copy of a pulse exhibiting formant ripple derived from an inverse filter measurement of the glottal pulse. Applying the polynomial pulse synthesis method to this second approach, the polynomial order of the source model would need to be greatly increased in order to match the formant ripple observed in the inverse filter derived glottal wave.

The best synthesis quality may still require use of a laryngeal model synthesizer in order to fully account for source-tract interactions [8]. This synthesizer would need to be operated at a sampling rate many times in excess of the required signal bandwidth in order to suppress the aliasing effects inherent in computing a continuous time waveform at discrete sample points. The source-filter type of synthesizer that we are proposing offers
a considerable saving in computational complexity over the laryngeal model, both from the standpoint of reduced sample rate as well as the inherent simplicity of the source-filter model. Further work is required to develop methods of including source-tract interaction effects in the synthesis, either by increasing the polynomial order of the pulse model or by adjusting the vocal tract filter.
ACKNOWLEDGEMENT

This work was supported by NIH grant NS 21516-2. The author acknowledges input of the reviewers regarding the literature review.
FIGURE CAPTIONS

Fig. 1.
First derivative of the glottal airflow pulse computed using the polynomial weight coefficients $b_1 = 50$, $b_2 = 0$, $b_3 = -300$, $b_4 = 1000$. (a) Pulse computed without a low pass filter. (b) Pulse computed using the frequency sampling filter $h_0 = .5$, $h_1 = 1$, $h_2 = 1$, $h_3 = 1$, $h_4 = .3904$ with one sided impulse response duration $t_h = 5 t_s$ where $t_s$ is the sampling interval. (c) Pulse computed using the frequency sampling filter $h_0 = .5$, $h_1 = .5$ with one sided impulse response duration $t_h = 5 t_s$.

Fig. 2.
Glottal airflow pulse computed by integrating the pulse $p(t)$ computed from the polynomial weight coefficients $b_1 = 50$, $b_2 = 0$, $b_3 = -300$, $b_4 = 1000$. (a) Pulse computed without a low pass filter. (b) Pulse computed using the frequency sampling filter $h_0 = .5$, $h_1 = 1$, $h_2 = 1$, $h_3 = 1$, $h_4 = .3904$ with one sided impulse response duration $t_h = 5 t_s$ where $t_s$ is the sampling interval. (c) Pulse computed using the frequency sampling filter $h_0 = .5$, $h_1 = .5$ with one sided impulse response duration $t_h = 5 t_s$.

Fig. 3.
Fourier spectrum of the first derivative of the glottal airflow pulse computed using the polynomial weight coefficients $b_1 = 50$, $b_2 = 0$, $b_3 = -300$, $b_4 = 1000$. (a) Pulse computed without a low pass filter. (b) Pulse computed using the frequency sampling filter $h_0 = .5$, $h_1 = 1$, $h_2 = 1$, $h_3 = 1$, $h_4 = .3904$ with one sided impulse response duration $t_h = 5 t_s$ where $t_s$ is the sampling interval. (c) Pulse computed using the frequency sampling filter $h_0 = .5$, $h_1 = .5$ with one sided impulse response duration $t_h = 5 t_s$. 
<table>
<thead>
<tr>
<th>Pitch Period (ms)</th>
<th>Pitch Synchronous No Low Pass Filter</th>
<th>Pitch Asynchronous No Low Pass Filter</th>
<th>Pitch Asynchronous Low Pass Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jitter (µs)</td>
<td>Shimmer (percent)</td>
<td>Jitter (µs)</td>
</tr>
<tr>
<td>8.02</td>
<td>35</td>
<td>.31</td>
<td>34</td>
</tr>
<tr>
<td>8.05</td>
<td>70</td>
<td>.60</td>
<td>69</td>
</tr>
<tr>
<td>4.02</td>
<td>42</td>
<td>3.28</td>
<td>40</td>
</tr>
<tr>
<td>4.05</td>
<td>87</td>
<td>8.23</td>
<td>86</td>
</tr>
</tbody>
</table>
REFERENCES


