Glottal source modeling for singing voice synthesis

Hui-Ling Lu

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University, Stanford, CA94305, USA
vickylu@ccrma.stanford.edu

Julius O. Smith III

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University, Stanford, CA94305, USA
jos@ccrma.stanford.edu

ABSTRACT

Naturalness of sound quality is essential for singing-voice synthesis. Since 95% of singing is voiced sound (Cook, 1990), the focus of this paper is to improve the naturalness of the vowel tone quality via glottal excitation modeling. We propose to use the LF-model (Fant et al., 1985) for the glottal wave shape in conjunction with pitch-synchronous, amplitude-modulated Gaussian noise, which adds an aspiration component to the glottal excitation. The associated analysis and synthesis procedures are also provided in this paper. By analyzing baritone recordings, we have found simple rules to change voice qualities from "laryngealized" (or "pressed"), to normal, to "breathy" phonation.

1. INTRODUCTION

Glottal source modeling has been shown to be an important factor for improving the naturalness of speech synthesis (Childers and Hu, 1994). Since naturalness of the sound quality is essential for singing voice synthesis, and since roughly 95% of singing is voiced, we focus on improving the naturalness of the vowel tone quality via the glottal excitation model described in this paper. The motivation is to support variation of glottal excitation model parameters based on estimation results from recordings of singing.

To trade off between the complexity of the model and the corresponding analysis procedure, we propose to use a source-filter type synthesis model based on a simplified human voice production system. The source-filter model (Fant, 1970), shown in Fig. 1, decomposes the human voice production system into three elements: glottal source, vocal tract, and radiation impedance. The radiation impedance is approximated by a differencing filter. The vocal tract filter is assumed all-pole, since we will only deal with non-nasal voiced sound in this study. Since both the vocal tract filter and the radiation filter are linear and time-invariant (over short time frames), they can be commuted. The glottal source and the radiation are then combined to form the "derivative glottal wave" as shown in Fig. 2. Figure 2 illustrates the concept of the source-filter modeling clearly: the human voice is modeled as the output of a linear all-pole filter excited by a glottal excitation.

In addition to providing flexible pitch and volume controls, the desired excitation model is expected to be capable of changing the voice quality. "Voice quality" has a wide range of possible meanings. In this study, the voice quality dimension considered ranges from laryngealized (pressed), to normal, to "breathy" phonation (Klatt and Klatt, 1990). To enhance the versatility of voice synthesizers with limited available storage of voice data, conversion of voice data from the speaker stored to a different target speaker is widely studied. In contrast, intra-speaker voice quality variations have apparently not been explored until recently (d'Alessandro and Doval, 1998).

Klatt (1990) has summarized the important acoustic features for different voice qualities. Two acoustic parameters are considered perceptually important indicators of voice quality: (1) degree of aspiration noise intruding at high frequencies in vowels, and (2) the relative strength of the fundamental component of the glottal source wave. Therefore, our glottal excitation model consists of two parts, as depicted in Fig. 2: (1) high-pass glottal noise (turbulence noise); and (2) a smooth, quasi-periodic, derivative glottal wave. The derivative glottal wave shape can control the relative strength of the fundamental component of the glottal source.

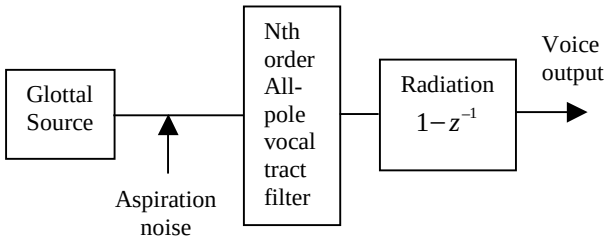

Figure 1. *Source-filter speech production model*

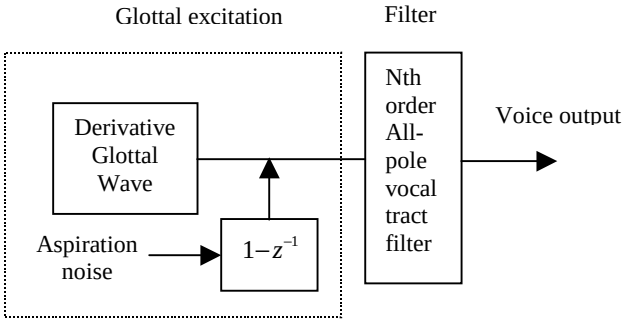

Figure 2. *A simplified source-filter model*

In the speech literature, there exist plenty of models for the smoothed derivative glottal wave. Cummings and Clement (1995) did an extensive literature survey on various models of the derivative glottal wave. The glottal models are divided into three rough categories: parametric non-interactive glottal volume velocity models, interactive parametric and mechanical glottal models, and three-dimensional physiological and numerical glottal models. The models differ as to how closely they approximate the physiology of voice production, and as to how much interaction is assumed among the glottal source, subglottal area, and supraglottal area.

Although a physical model is attractive for control flexibility in terms of physical parameters, a parametric, non-coupled glottal waveform model was chosen for this study, as such a model can provide good synthesis quality with the necessary controls. Among such glottal models, the LF-model and the KLGLOTT88 model appear to be most widely discussed.

The LF-model is used in this study to model the derivative glottal wave. Moreover, the aspiration noise is modeled as pitch-synchronous, amplitude-modulated Gaussian noise. We will describe both the LF-model and the noise model in the next section.

Once the synthesis model is constructed, the next step is to explore the space of model parameters such that one can generate desired voices by manipulating model parameters. The advantage of source-filter models over articulatory (Bavegard, 1996) or formant synthesis (Lin, 1990) models is that synthesis based on analysis is relatively simple. Hence, we can obtain the synthesis parameters via analysis of sound recordings.

For our model, the effort is then to estimate (1) the vocal tract filter parameters and (2) a glottal excitation waveform to mimic the desired singing vowels. We have developed a de-convolution algorithm using convex optimization techniques (Lu, 1999). Through this de-convolution, one can obtain the vocal tract filter parameters and the inverse-filtered glottal excitation waveform. The glottal excitation waveform consists of two parts: the derivative glottal wave and a residual (high-passed aspiration noise). These two components are separated by wavelet packet analysis (Coifman, 1992). We then use constrained nonlinear optimization to fit the derivative glottal wave to the LF-model. For breathy phonation, the noise residual is analyzed in terms of its average strength, location of maximum noise level, and the duty cycle of the amplitude modulation envelope.

In the remainder of this paper, we will describe the proposed synthesis model in Section 2. Section 3 will illustrate the overall analysis procedures for retrieving the model parameters from sound recordings. Using the proposed analysis procedure, the glottal excitation model parameters are studied for the case of a baritone singing sustained vowels with varying sound quality. Section 4 will show the results of the analysis and give a statistical summary of the model parameters.

2. Synthesis Model

The overall synthesis model is shown in Fig. 2. In this section, we will describe the LF-model and the noise residual model for glottal excitation modeling.

2.1 LF-model

The smoothed derivative glottal wave is modeled via the LF-model (Fant et al., 1985), which is a parametrized time-domain model of one cycle of the derivative glottal wave. This time-domain model characterizes the wave-shape of the derivative glottal wave in the open and the closed phases via only four parameters. The LF model is chosen because its properties appear to have been studied most extensively. It has been shown that the model can accommodate a wide range of natural variations. Its parameters can be estimated via inverse filtering of recorded samples, and the parameters vary with diverse voice qualities such as loudness, fundamental frequency and tenseness (Karlsson, 1995) (Childers, 1995). Moreover, synthesized voice quality can easily be altered by a single parameter in the extended transformed LF-model (Fant, 1995).

The LF-model models the differentiated glottal flow. The model consists of two segments. The first segment characterizes the differentiated glottal flow from the glottal opening to the maximum negative peak. The second segment characterizes the closure of the glottis. The model can be parametrized as follows:

$$g(t) = E_0 e^{\alpha t} \sin(\omega_g t), 0 \le t \le T_e \qquad (1)$$

$$= -\frac{E_e}{\varepsilon T_a}\left[e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)}\right], T_e \le t \le T_c \le T_0 \qquad (2)$$

Where the parameters are described below.

Figure 3 plots two periods of the glottal wave (top) and the derivative glottal wave generated from the LF-model using a typical set of normal phonation parameters (bottom).

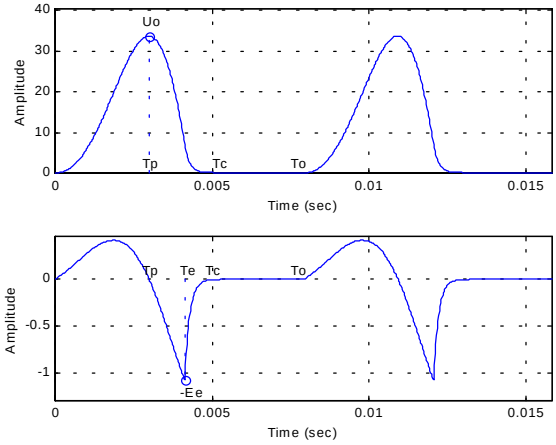

Figure 3. *Glottal wave/Derivative glottal wave from LF-model*

Along with the magnitude of the glottal closure excitation, E_e, the modeled waveform can be specified by two independent sets of parameters: the direct synthesis parameters ($E_0, \alpha, \omega_g, \varepsilon$), and the timing parameters (T_p, T_e, T_a, T_c). The parameter T_p denotes the instant of the maximum glottal flow. The parameter T_o is the fundamental period. The parameter T_c denotes the ending of the return phase. The parameter T_a is the effective duration of the return phase. The exponent ε in Eq. (2) is related to T_a and can be uniquely determined from T_a. Since the timing parameters can be easily identified from the estimated derivative glottal wave, one usually obtains the timing parameters and then derives the direct synthesis parameters from the timing parameters with the following constraints:

$$\int_0^{T_0} g(t)dt = 0, \quad \omega_g = \frac{\pi}{T_p}, \quad \varepsilon T_a = 1 - e^{-\varepsilon(T_c - T_e)}, \quad \text{and}$$

$$E_0 = -\frac{E_e}{e^{\alpha T_e}\sin(\omega_g T_e)}.$$

Lin (1990) gives further details regarding the implementation of the LF-model.

The return phase is a major constituent of the LF-model. The effective duration of the return phase, T_a, is perceptually the most important parameter for the LF-model. It has been shown to be inversely proportional to the frequency, $F_a = 1/(2\pi T_a)$, at which the spectrum of the derivative glottal

wave attains an extra –6 dB/oct slope. Hence, increasing T_a will lower the cut-off frequency F_a of an equivalent low-pass filter. Therefore, T_a determines the spectral tilt of the glottal source. For example, the closure of each glottis cycle is less abrupt for the breathy phonation. The gradual closure of the glottis, resulting in a larger T_a, introduces fewer higher harmonics and therefore greater spectral tilt.

In additional to the direct synthesis parameters and the timing parameters, one could also describe the model via a set of normalized timing parameters ($R_{a,} R_g, R_k$), defined as follows:

$$R_a = T_a / T_0 \quad (3)$$

$$R_g = T_o /(2 \cdot T_p) \quad (4)$$

$$R_k = (T_e - T_p)/T_p \quad (5)$$

In the transformed LF-model (Fant et al., 1994), a new wave-shape parameter, R_d, is introduced. This parameter is proportional to (U_o / E_e), where U_o is the peak value of the glottal flow, exclusive of a superimposed constant leakage flow, and E_e is the excitation amplitude at the glottal closure instants. R_d can also be estimated from the measured set of LF-parameters (Fant, 1997) as

$$R_d = (1/0.11)(0.5 + 1.2 \cdot R_k)(R_k / 4R_g + R_a) \quad (6)$$

The accuracy of this formula is within 0.5dB for $R_d < 1.4$. This formula tends to over-estimate R_d with a maximum error of 1.5dB at R_d = 2.7.

It has been shown that R_d is one of the most effective parameters for quantifying the quality of the voice source with a single numerical value. Another advantage of using R_d is that we can predict R_a, R_g, R_k from it. The predicted values are denoted R_{ap}, R_{gp}, R_{kp}. The following prediction equations are derived by Fant (1995) via linear regression:

$$R_{ap} = (-1 + 4.8R_d)/100 \quad (7)$$

$$R_{kp} = (22.4 + 11.8R_d)/100 \quad (8)$$

R_{gp} is then obtained by substituting (7) and (8) into (6).

These predicted values may deviate from the desired value; hence, three deviation constants are defined here:

$$K_a = R_a / R_{ap}, K_g = R_g / R_{gp}, K_k = R_k / R_{kp} \quad (9)$$

In Section 1.3, the analysis results are summarized in terms of R_d. In the synthesis stage, R_a, R_g, R_k are first predicted from R_d Timing parameters can then be obtained via equations (3), (4) and (5). Direct synthesis parameters are retrieved thereby.

In addition to R_d, the "open quotient" parameter is another good indicator of different phonations. The open quotient parameter, *OQ*, is defined as T_e / T_o. Breathy voices tend to have a larger open quotient. The open quotient can be related to the relative amplitude of the voice fundamental, H_1, and the second harmonic, H_2. By regression analysis of data generated by the LF-model for different phonation types, Fant (1997) found the following relationships.

$$H_1 - H_2 = -6 + 0.27\exp(5.5OQ) \quad (10)$$

$$H_1 - H_2 = -7.6 + 11.1R_d \quad (11)$$

These are good approximations for R_d up to 2.7.

2.2 Noise residual model

For the singing voice to sound "breathy", aspiration noise is perceptually most important (Klatt and Klatt, 1990). The increase of the relative amplitude of the fundamental component is secondary. Without the presence of aspiration noise, the increase of the fundamental component may induce the sensation of nasality in a high-pitch voice. Therefore, despite the fact that breathiness can be simulated to some extent by using a sophisticated glottal-source model, a more natural simulation of breathiness requires the addition of aspiration noise.

The aspiration noise (turbulence noise) is pitch synchronous with the smoothed quasi-periodic derivative glottal wave since the likelihood of the turbulence noise is proportional to the flow and inversely proportional to the radius of the aperture. Cook (1990) has calculated the likelihood of the existence of the turbulence noise. He concluded that the likelihood of the turbulence exists for the entire open phase and achieves maximum sound radiation power right after the point where the vocal folds begin to close. A high power burst of noise is also likely at the glottal opening instant, corresponding to highly pressurized air rushing through a small slit. Hence, it is expected that there are two pulses of noise per cycle in some cases. One occurs when the vocal folds are closing and the other one occurs when the vocal folds are opening. Pitch-synchronous noise is also found in other musical instruments such as bowed strings (Chafe, 1990).

From the psychoacoustics experiments, Hermes (1991) did a study on the perception of the synthetic breathy vowels. In his experiments, the synthetic vowels are generated based on the source-filter theory, and the derivative glottal wave is simplified as a low-pass filtered pulse train. He indicated that the noise segregates from the speech signal and perceived as a separate stream when stationary noise is used. Hence, adding stationary noise hardly contributes to the breathy timbre of the vowel. Hermes further concluded that using pitch-synchronous amplitude modulated noise, i.e., the noise has a temporal envelope of the same periodicity as the pulse train, could solve this problem.

Combined with the low-pass filtered pulse train, a de-emphasized, high-pass filtered, pitch-synchronous, amplitude-modulated noise is used as the source to excite the vocal tract filter. In this way, the noise integrates with the pulse train in the sense that a reduction of the loudness of the noise stream and a timbre change in the breathy vowel are perceived. It is found that the reduction of the noise loudness is maximized when the interval between the noise bursts and the pulses is less than 1 ms in a glottal period of 8 ms. The cut-off frequency of the high-pass filter is chosen from 1200 Hz to 2k Hz. A lower cut-off frequency results in a greater degree of breathiness. The de-emphasis is implemented by the low-pass filter $H(z) = 1/(1 - p \cdot z^{-1})$, where the pole p is set to 0.9. Childers (1995) also proposed a similar noise model, but without the de-emphasis.

As a first approximation, the nature of the double pulses of the aspiration noise is neglected. The noise residual before spectral shaping is roughly modeled as a pitch synchronous amplitude modulated Gaussian noise with larger power around the glottal closure instants.

Figure 4 shows the noise synthesis model for the high-passed aspiration noise. The first block is the Gaussian noise unit

generator, which generates a zero mean, unit variance, Gaussian white noise sequence. Second, a scaled Hanning window centered around the glottal closure instants modulates the amplitude of the Gaussian noise. The scale of the Hanning window for each glottal period is specified in the A_n parameter sequence. L in the above figure indicates the lag of the center of the Hanning window with respect to the glottal closure instants. The lag is specified as a percentage relative to the length of the glottal period. The locations of the Hanning window centers are then calculated from the glottal closure instants and the desired lag. After the amplitude modulation, a spectral shaping filter is used to account for the average spectral density of the aspiration noise. The shaping filter also includes the high-pass filtering introduced by commuting the radiation filter with the vocal tract filter as shown in Figure 2.

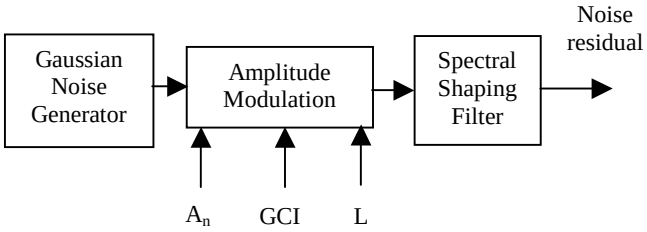

Figure 4. *Noise residual synthesis model*

The current spectral shaping filter used is a combination of a de-emphasis filter and a high-pass filter. We use the same de-emphasis factor 0.9 as Hermes did. The cut-off frequency of the high-pass filter is 4 kHz, which is chosen by observing the spectrum magnitude of the noise residual extracted from real recordings. Figure 5 shows several frequency spectra of noise residuals extracted from the baritone sustained singing vowels. Vowels /a/, /e/, and /i/ are included in this figure. We can see that the spectral shaping is quite consistent except for the residual vocal tract filter formant structure. Therefore, a better spectral shaping filter could be obtained by matching an ARMA filter to the frequency spectrum of the noise residuals extracted from singing recordings. This idea is still under investigation.

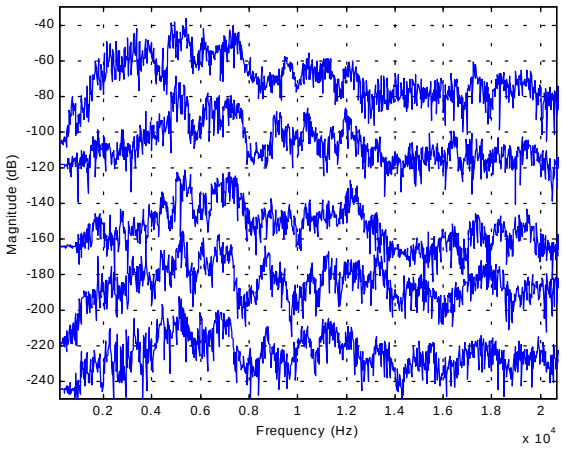

Figure 5. *Frequency spectra of the noise residuals*

3. Analysis Algorithms

In order to mimic the original recording, the analysis procedure involves three steps. The first step is to obtain the glottal excitation from sound recordings via a recently proposed inverse filtering method. The second step is to decompose the inverse filtered glottal excitation into a smoothed derivative glottal wave and the noise residual. Finally, the smoothed derivative glottal wave is fitted to the LF model via constrained nonlinear optimization. The magnitude of the noise residual is measured to determine the variance of the Gaussian noise around the glottal closure instants and the duty cycle of the amplitude modulation. These three steps will be illustrated in the following.

3.1 Source-filter de-convolution

A novelty of this newly proposed source-filter de-composition algorithm is that it provides a derivative glottal wave constraint when estimating the vocal tract filter. Hence, the resulting inverse-filtered derivative glottal waves are closer to the true glottal excitation, at least when the source-filter interaction is negligible. A brief discussion of this algorithm is provided in the following. A detailed analysis of the performance of this algorithm is published elsewhere (Lu, 1999).

The derivative glottal wave is constrained to the KLGLOTT88 model (Klatt, 1990). The KLGLOTT88 model consists of the Rosenberg model describing the basic wave-shape of the derivative glottal wave and a spectral tilt implemented by low-pass filtering. Figure 6 shows the synthesis model assumed under this algorithm, where the KLGLOTT88 model replaces the glottal excitation in Fig. 2.

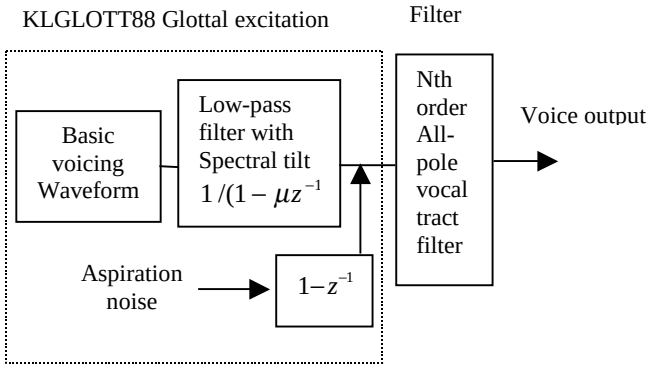

Figure 6. *The synthesis model at the analysis phase*

In the KLGLOTT88 model, the basic voicing waveform describes the wave-shape of the derivative glottal wave without the return phase. In the open phase, the derivative glottal wave is modeled by a simple second order polynomial,

$$g(n) = \begin{cases} 2an/F_s - 3b(n/F_s)^2, & 0 \le n \le T_0 \cdot OQ \cdot F_s \\ 0, & T_0 \cdot OQ \cdot F_s < n \le T_0 \cdot F_s \end{cases} \quad (12)$$

$$a = \frac{27 \cdot AV}{4 \cdot (OQ^2 \cdot T_0)} \quad (13)$$

$$b = \frac{27 \cdot AV}{4 \cdot (OQ^3 \cdot T_0^2)} \quad (14)$$

Where T_0 is the fundamental period of the voice, F_s is the sampling frequency, AV is the amplitude parameter, and OQ is the "open quotient" of the glottal source (0 for always closed and 1 for always open).

Although the basic waveform always has abrupt changes at the glottal closure instants, the low-pass spectral tilt filter gives some smoothing of the return phase. This first-order low-pass filter and the vocal tract filter are further combined to

form an order N+1 all-pole filter. The glottal spectral tilt can be associated with a single real root of the order N+1 filter.

Due to the simplicity of the basic waveform, we may formulate the parameter estimation as a convex optimization problem by estimating the parameters of the all-pole vocal tract filter and the glottal source waveform pitch synchronously. More explicitly, if we know the fundamental period T_0 and the open quotient *OQ*, the task is then to estimate the filter coefficients of the N+1 order all-pole filter and two shaping parameters (*a* and *b*) of the basic waveform using one or more periods of the voice pressure data.

To mimic the original speech pressure wave input, we would like to minimize the error between the estimated speech pressure wave and the true speech pressure wave. However, for the convex optimization formulation, a so-called "equation error" (Ljung and Soderstrom, 1987) is used for measuring the model fit. Using the equation-error method, we try to minimize the error between the estimated glottal wave and the true glottal wave signal.

Denote the filter coefficients as $A=[\hat{a}_1' \quad ... \quad \hat{a}_{N+1}']^T$, the known speech signal as *y(n)*, and the estimated derivative glottal waveform as $\hat{g}(n)$. We now form the error between the true glottal wave (that belongs to the KLGLOTT88 model by assumption) and the estimated glottal wave. For $i = 1,\ldots,m$, where *m* is the number of sampling points in one glottal period, we have

$$\begin{aligned} g(i)-\hat{g}(i) &= 2a\cdot i - 3b\cdot i^2 - y(i) + \hat{a}_1' y(i-1) + \ldots + \hat{a}_{N+1}' y(i-N-1) \\ &= \begin{bmatrix} y(i-1) & y(i-2) & \ldots & y(i-N-1) & 2i & -3i^2 \end{bmatrix} X - y(i) \\ &= a_i^T X - b_i \end{aligned} \tag{17}$$

$X = [\hat{a}_1' \quad ... \quad \hat{a}_{N+1}' \quad a \quad b]^T$ is the parameter vector we wish to estimate. Note that the chosen error is linear in the parameters, as is characteristic of equation error formulations.

If *OQ* and T_0 are assumed known, minimization of the equation error for a glottal period is then equivalent to the following convex optimization problem:

$$\begin{aligned} &\text{Minimize } \|AX-B\| \\ &\text{Subject to } a>0,\ b>0 \text{ and } a = T_0 \cdot OQ \cdot b \end{aligned} \tag{18}$$

where $A = [a_1 \quad ... \quad a_m]^T$, $B = [b_1{}^T \quad ... \quad b_2{}^T]^T$.

To trade off between robustness and computational efficiency, the L_2 norm minimization is chosen for our application. The above problem can then be solved via the sequential unconstrained minimization technique (SUMT) (Boyd, 1999). By uniformly sampling the *OQ* values and solving the problem at each sample, the best estimate can be obtained as the one having minimum error.

In addition to the convex constraints of the glottal wave-shape parameters (*a* and *b*), the value of the last coefficient of the filter, $\hat{a}_{N+1}'$, is also constrained to be within a predefined range. By observing that this coefficient is the product of all the poles and the spectral tilt μ, the poles of the all-pole filter are loosely regularized by the constraints on $\hat{a}_{N+1}'$.

The overall estimation procedure is illustrated in Fig. 7. Before SUMT estimation, we need to estimate the fundamental period T_0, and retrieve pitch-synchronized speech pressure data that start at the glottal closed phase. Several methods for GCI (glottal closure instant) detection from the speech signal have been discussed in the literature (Strube, 1974) (Ma, 1994). Most depend on either the short-time energy of the signal or the linear prediction residual signal. These methods are based on block data processing; since voice is rarely precisely stationary, there is some ambiguity in the locations of the detected glottal closure instants. To overcome this problem, Smits (1995) developed another type of method to extract the instants of significant excitation (glottal closure) for speech signals. This method assumes that the excitation signal within a pitch period, starting from the significant excitation, is minimum phase. The glottal closure instants are then estimated as the frame start time plus the average group-delay of the LP residual within an analysis frame.

The group-delay method has been shown to be robust against noise and distortion, since the average phase characteristics of the signal are determined mainly by the strength of the excitation (Murthy, 1999). Hence, this algorithm was chosen as the main algorithm for determining the GCIs. Since the estimates from this method tend to have a lag bias, the Frobenius norm approach (Ma, 1994) is also used in conjunction with the group-delay method to determine the final GCI locations.

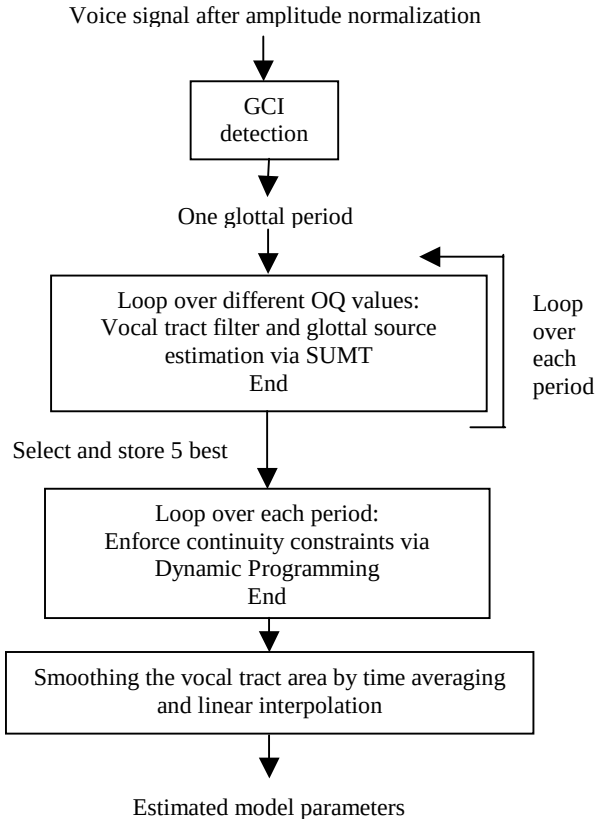

Figure 7. *The analysis procedure for source-filter decomposition*

By studying the synthetic data, we have found that the inverse filtered glottal excitation is still acceptable even when the detected GCIs are off 15% from the correct locations. Hence, a byproduct of the SUMT estimation is GCI detection, where the GCIs are detected by locating the local minima of the inverse-filtered glottal excitation. In our experiments for sustained vowels, the GCIs are actually further corrected using this method.

3.2 Noise residual extraction

From the above source-filter de-convolution procedure, one could obtain the inverse-filtered glottal excitation. We have modeled the glottal excitation signal as the sum of the smoothed derivative glottal wave and the high-passed noise residual. In this section, we want to separate the noise residual from the derivative glottal wave signal.

Short-time Fourier analysis methods consider the noise residual as the aperiodic component. The periodic part is considered as the derivative glottal wave. Since the derivative glottal wave, especially for pressed and normal phonation, has a sharp discontinuity around the glottal closures, traditional short-time Fourier analysis cannot represent it well around the glottal closure instants. A certain amount of averaging around the glottal closure instants is inevitable. A strength of wavelet analysis is that it can remove the noise component without compromising the sharp detail of the original signal. Wavelet Packet Analysis (WPA) (Coiffman, 1992) is a generalization of the wavelet decomposition in that it offers a larger range of signal representations. Therefore, WPA is used to extract the noise residual.

The following three figures compare the effectiveness of three noise extraction methods by de-noising a synthetic glottal excitation signal. Figure 8 is the de-noising result obtained using the SMS (sinusoidal modeling) decomposition (Serra, 1997). This is a short-time Fourier analysis type method. The first plot on Fig. 8 shows the desired derivative glottal wave and the estimated derivative glottal wave. The second plot shows the original noise component, and the third plot is the extracted noise component. In this example, we use 60 harmonics to present the derivative glottal wave in order to capture the sharp corner of the glottal closure instants. Since the number of harmonics is high, it tends to over-fit the input data, so that some noise is included in the estimated derivative glottal wave.

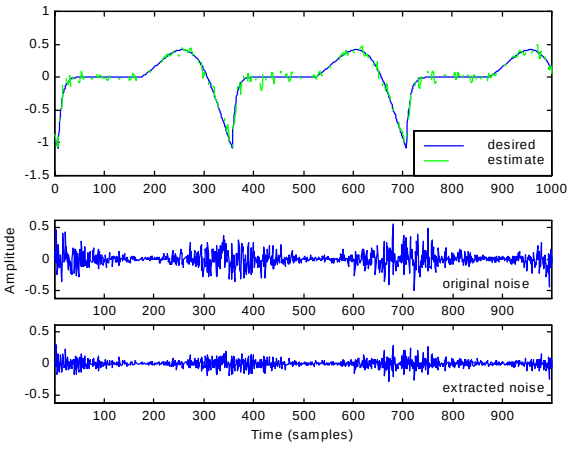

Figure 8 *De-noising results using SMS*

Fig. 9 shows the de-noising results obtained using a basic wavelet decomposition. The wavelet decomposition is performed at level 3 with the 3rd order Daubechies wavelet. A heuristic variant of the principle of Stein's Unbiased Risk Estimate (SURE) thresholding is used to truncate the noise components (Donoho, 1995). The noise residual is obviously under-estimated in this example.

Figure 10 shows the result for the Wavelet Packet Analysis (WPA) method. The WPA is performed at level 4 with the 2nd order Daubechies wavelet. The variance of the original noise and the extracted noise are comparable. The amplitude envelope of the extracted noise is very close to the original one.

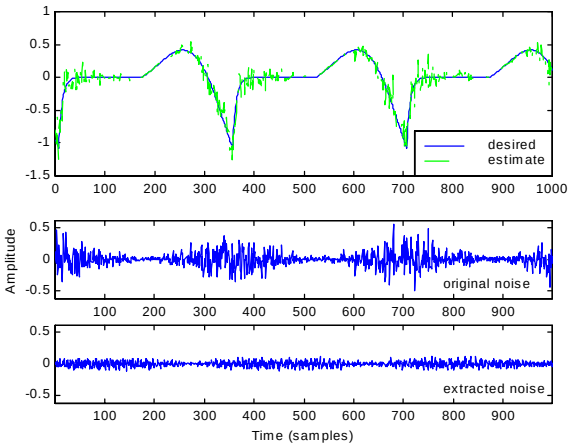

Figure 9. *De-noising results using a wavelet decomposition*

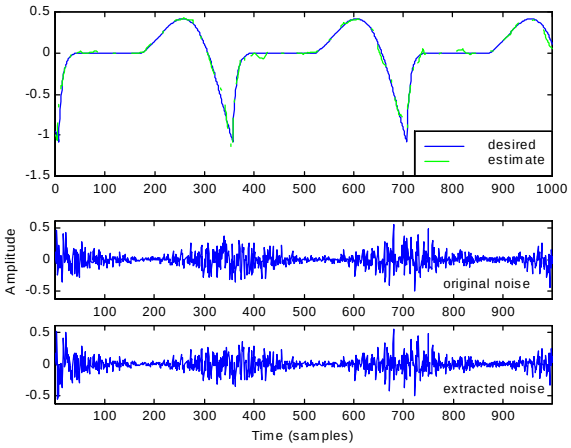

Figure 10. *De-noising results using Wavelet Packet Analysis*

3.3 LF-model fitting

Once the estimated derivative glottal wave is obtained, the LF-model is used to parametrize the derivative glottal wave period-by-period. Two steps are involved in LF-model fitting.

For each period of the derivative glottal wave, the LF-model timing parameters and glottal excitation, E_e, are first retrieved by direct estimation methods (Strik, 1998). Since the estimated derivative glottal wave is usually noisy due to the model mismatch, the direct estimation method may yield unreliable results. The LF-model parameters are further refined via constrained nonlinear optimization using the Sequential Quadratic Programming method. The timing parameters are constrained such that the open quotient *OQ* will be bounded around the estimated value from Equation (10).

4. Results and Discussion

Fig. 11 illustrates the effectiveness of the source-filter de-convolution algorithm. These results are estimated from the baritone sung vowel /a/ at pitch B2. The first plot overlays the inverse filtered glottal excitation, the KLGLOTT88 and the LF

synthetic derivative glottal waves. The parameters for the KLGLOTT88 model are obtained at the source-filter decomposition step. The LF-model parameters are obtained via fitting the inverse filtered signal to the LF-model. We can observe some formant ripples during the closed phase. These inevitable ripples appear to result from the source-tract interaction that is beyond the assumptions of the source-filter model. One simple remedy for this problem is to introduce two different vocal tract filters, one at the closed phase and another one at the open phase. The perceptual improvements due to using two such filters are still under evaluation.

The second plot shows a snapshot of the spectra of the inverse-filtered glottal excitation and two fitted synthetic derivative glottal waves. The third plot shows the spectra of the original recording and the synthetic sound generated by exciting the estimated vocal tract filter with the LF synthetic derivative glottal wave. Both of the frequency responses match quite well. We have separated these spectra by scaling their magnitude for better viewing. The synthetic vowel also sounds almost like the original singing from informal subjective listening.

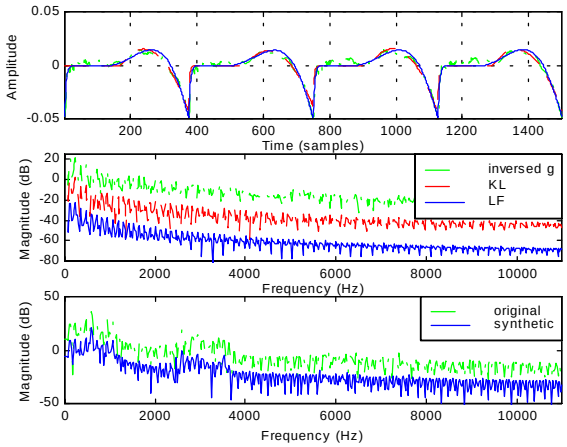

Figure 11. *Source-filter de-convolution results for a low-pitched baritone sung vowel /a/*

Fig. 12 shows the results of noise residual extraction using Wavelet Packet Analysis. The original recording is the breathy sung vowel /a/ at pitch B2. The upper plot shows the inverse-filtered glottal excitation and the de-noised signal. The bottom plot is the extracted noise residual. The estimated glottal open instants are also indicated in the plots as solid stem lines. The estimated glottal closure instants are shown by dashed stem lines. From this figure, we can clearly see that the noise bursts occur right after the glottal closure instants and the glottal opening instants. This result is consistent with the study by Cook (Cook 1990).

Table 1 summarizes the LF-model parameters from the analysis of the baritone sung vowels. The results are averaged across different singing phonations, pitch scales, and vowel types. The low pitch is B2, the medium pitch is G3 and the high pitch is B3. Comparing to Karlsson's results for speech data (Karlsson, 1995), we have a larger R_d parameter and a lower cut-off frequency F_a, i.e., a larger spectral tilt. From the summary, we also see that the deviation constants are close to 1 in most of the cases; hence, one could predict the LF parameters quite well from a single R_d parameter. Furthermore, we see that increasing the degree of breathiness or the phonation frequency will result in a larger R_d .

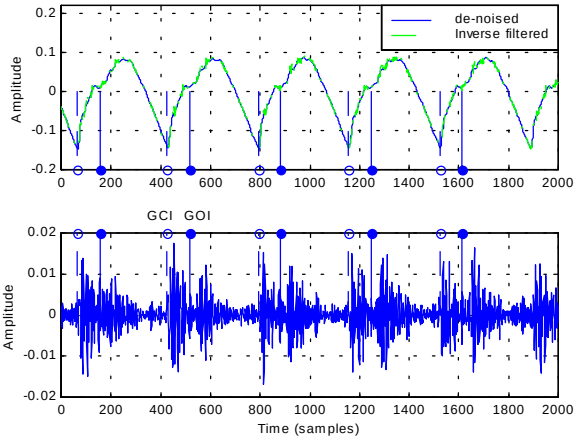

Figure 12. *Extracted noise residual for a breathy sung /a/*

Condition	OQ_i	F_a	K_k	K_a	R_d
Pressed	0.49	675	1.07	1.23	0.84
Normal	0.64	515.3	1.04	0.95	1.19
Breathy	0.78	161	0.89	1.21	2.90
Low F_0	0.61	349	1.05	0.97	1.30
Medium F_0	0.64	326	1.02	1.05	1.63
High F_0	0.69	300.7	0.90	1.35	2.26
Vowel /a/	0.63	349	0.99	1.10	1.57
Vowel /e/	0.66	318.9	1.01	1.10	1.79
Vowel /i/	0.65	298.3	1.00	1.12	1.81

Table 1. *The summary of the LF-model parameterization*

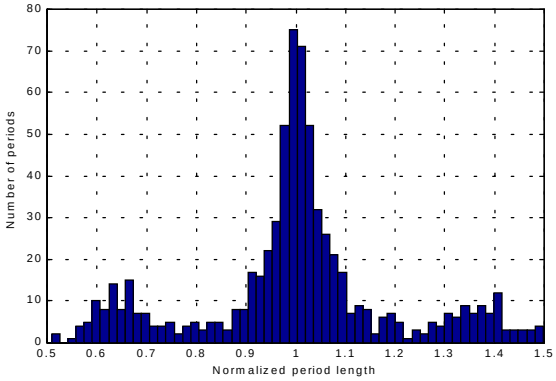

Figure 13. *Histogram of normalized noise period length*

A statistical analysis of the noise residual for breathy vowels was carried out. First, the degree of pitch synchrony was evaluated by measuring the intervals between adjacent peaks in the smoothed noise time-domain magnitude envelope. The intervals were normalized such that an interval of 1 is one period. A histogram of these normalized amplitude-envelope peak intervals (see Fig. 13) shows that the noise residual is highly pitch synchronous. We also found that the average noise peaks occurs at a lag of 10% of the period after the glottal closure instants. The corresponding interval between the major excitation and the noise bursts is less than 1ms. This is consistent with D.J. Hermes's experiments with synthetic noise.

We also evaluate the strength and the duty cycle of the noise bursts. The strength of the noise burst is measured as the normalized amplitude-envelope peak for each glottal period. The amplitude is normalized by the strength of the glottal excitation, E_e. We found that the normalized strength could be approximated as a constant except at the onset and the offset of the vowel. The average strength was 0.04 in our experiments. The duty cycle is measured as the normalized time duration for the noise bursts decay to 10% above the noise floor. The duty cycle is less consistent across different sung tones. The average duty cycle measured was 42%.

Fig. 14 shows the normalized noise floor. The normalized noise floor is defined as the minimum noise amplitude-envelope magnitude relative to the normalized strength of the noise burst in the period. The average noise floor measured was 0.02. The normalized noise floor was found to be quite consistent and has a tendency to increase at the offset where the noise is less pitched.

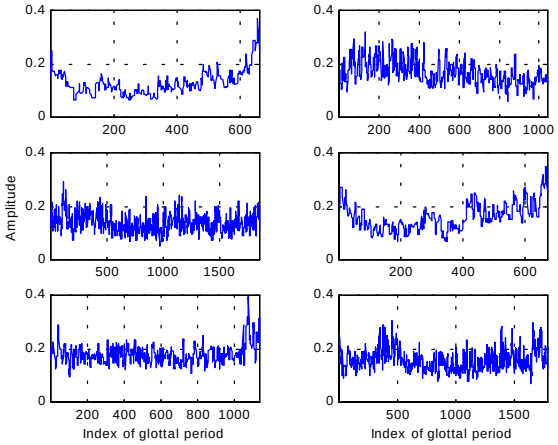

Figure 14. *Normalized noise floor each period*

5. SUMMARY AND CONCLUSIONS

In this paper, we proposed a model for glottal excitation which consists of the sum of a parametric glottal waveform (modeled using the LF method) and a pitch-synchronous amplitude-modulated Gaussian noise (the aspiration component). We described the associated analysis and resynthesis procedures. By analyzing baritone recordings with different voice qualities, parameters of the glottal excitation model were computed and summarized. We conclude that the proposed model is capable of achieving a wide variety of synthetic voice "textures" by varying the wave-shape parameter R_d and the strength of the noise component.

6. REFERENCES

[1] Bavegard, Mats 1996. *Towards an articulatory speech synthesize: model development and simulation*. Ph.D. thesis, KTH.

[2] Boyd, S. and Vandenberghe, L. 1999, *Course Reader for EE364: Introduction to Convex Optimization with Engineering Applications*. Stanford University.

[3] Childers, D. G. and Hu, H. T. 1994. "Speech synthesis by glottal excited linear prediction." *Journal of the Acoustical Society of America* 96(4):2026-2036.

[4] Childers, D. G. and Ahn, C. 1995. "Modeling the glottal volume-velocity waveform for three voice types." *Journal of the Acoustical Society of America* 97(1):505-519.

[5] Chafe, C. 1990. "Pulsed Noise in Self-Sustained Oscillations of Musical Instruments." *ICASSP* (2):1157-1160.

[6] Coifman, R. R. and Wickerhauser, M. V. 1992. "Entropy-based algorithms for best basis selection." *IEEE Trans on Inf. Theory* 38(2):713-718.

[7] Cook, P.R. 1990. *Identification of control parameters in an articulatory vocal tract model with applications to the synthesis of singing.* Ph.D. thesis, Report No. STAN-M-68.

[8] Cummings, K. E. and Clements, M. A. 1995. "Glottal models for digital speech processing: a historical survey and new results." *Digital signal processing* (5):21-42.

[9] Donoho, D. L. 1995. "De-noising by soft thresholding." *IEEE Trans on Inf. Theory* 41(3):613-627.

[10] Fant, G. 1970. *Acoustic theory of speech production*. Mouton, The Hague.

[11] Fant, G., Liljencrants, J., and Lin, Q. 1985. "A four-parameter model of glottal flow." *STL-QPSR* (4):1-13.

[12] Fant, G. 1995. "The LF-model revisited. Transformations and frequency domain analysis." *STL-QPSR* (2-3):119-156.

[13] Fant, G. 1997. "The voice source in connected speech." *Speech communication* (22):125-139.

[14] Hermes, D. J. 1991. "Synthesis of breathy vowels: some research methods." *Speech communication* (10):497-502.

[15] Karlsson, I. and Liljencrants, J. 1996. "Diverse voice qualities: models and data." *STL QPSR* (2):143-146.

[16] Klatt, D. H. and Klatt, L. C. 1990. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *Journal of the Acoustical Society of America* 87(2):820-857.

[17] Lin, Qiguang 1990. *Speech Production Theory and Articulatory Speech Synthesis*. Ph.D. dissertation, Dept. of Speech Communication & Music Acoustics, Royal Inst. Of Technology (KTH), Stockholm.

[18] Ljung, L., and Soderstrom, T. 1987 *Theory and Practice of Recursive Identification*, the MIT Press.

[19] Lu, Hui-Ling. 1999. "Joint estimation of vocal tract filter and glottal source waveform via convex optimization," *Proc. IEEE Workshop on application of signal processing to audio and acoustics*, pp. 79-82.

[20] Ma, C., Kamp, Y., and Willems, L.F. 1994 "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. on Acoustics, Speech, and Signal Processing*. 2(2):258-264.

[21] Murthy, P. S. and Yegnanarayana, B. 1999. "Robustness of group-delay-based method for excitation of significant instants of excitation from speech signals." *IEEE Trans. Speech, Audio Processing* 7(6):609-619.

[22] Serra, Xavier. 1997. "Musical sound modeling with sinusoids plus noise." *Musical Signal Processing*, Swets & Zeitlinger Publishers.

[23] Smits, R. and Yegnanayarana, B. 1995. "Determination of instants of significant excitation in speech using group delay function." *IEEE Trans. Speech Audio Processing* (3):325-333.

[24] Strik, H. 1998. "Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses" *Journal of the Acoustical Society of America* 103(5):2659-2669.

[25] Strube, H. W. 1974. "Determination of the instant of glottal closure." *Journal of the Acoustical Society of America* 56:1625-1629.