APPENDIX 7

GLOTTAL EXCITATION MODELS

A7.1 INTRODUCTION

Chapters 3 through 7 often refer to the need for models of the excitation waveform, especially for speech synthesis. In this appendix, we introduce two glottal models that are used in the software that accompanies this book: the Liljencrants-Fant (LF) model (Fant et al., 1985) and the polynomial model (Childers and Hu, 1994; Milenkovic, 1993). Also covered is a noise model for generating aspiration and fricative noise for unvoiced sounds. This noise model is also used for enhancing the naturalness of voiced sounds for certain voice types. This appendix presents some typical values for these models that have been found useful for generating several voice types, such as modal, breathy, vocal fry, whisper, harsh, and falsetto.

A typical glottal excitation volume–velocity waveform (glottal flow) obtained by inverse filtering is shown in Figure A7.1 along with its spectrum. As discussed in Chapters 3 through 6, some time domain acoustic features of the glottal source are the pitch period, pitch perturbation (jitter), amplitude perturbation (shimmer), the glottal flow pulse width, the glottal flow skewness, the abruptness of closure of the glottal flow, and aspiration noise. While some important frequency domain features are the spectral tilt (slope), harmonic richness factor, and harmonic-to-noise ratio. The models in this appendix generally can account for these features via the manipulation of the model parameters.

The relationship between the volume–velocity (glottal flow), the differentiated glottal flow, and the speech waveforms is illustrated in Figure A7.2. Note that the main excitation occurs at the initiation of the glottal closed phase. The speech waveform decays beyond this point until the next excitation occurs.

A7.2 THE LF MODEL

Instead of modeling the glottal flow, Liljencrants and Fant (Fant et al., 1985) chose to model the differentiated glottal flow. This model has become known as the LF model and is illustrated in Figure A7.3. The glottal waveform parameters are given as

$$\begin{aligned} g(t) &= E_0 e^{\alpha t} \sin(\omega_g t) \quad \text{for } 0 \le t \le t_e \\ &= -\frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{\varepsilon(t_c-t_e)}\right] \quad \text{for } t_e \le t \le t_c \le T_0 \end{aligned} \tag{A7.2.1}$$

where T_0 is the pitch period, which is typically larger than t_c. The following conditions hold.

$$\int_0^T g(t)\,dt = 0, \quad \omega_g = \frac{\pi}{t_p}, \quad \varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)}, \quad \text{and} \quad E_0 = -\frac{E_e}{e^{\alpha t_e}\sin(\omega_g t_e)} \tag{A7.2.2}$$

The modeled waveforms can be specified by either the direct synthesis parameters (E_0, α, ω_g, and ε) or the timing parameters (t_p, t_e, t_a, and t_c). The parameter t_p denotes the instant of the maximum glottal flow model waveform. The parameter t_e is the instant of the maximum negative differentiated glottal flow model. The parameter t_a is the time constant of the exponential curve of the second segment of the LF model. The parameter t_c is the instant at which complete glottal closure is reached for the model. One reason this model is important and useful is that it is related to actual glottal waveforms via the

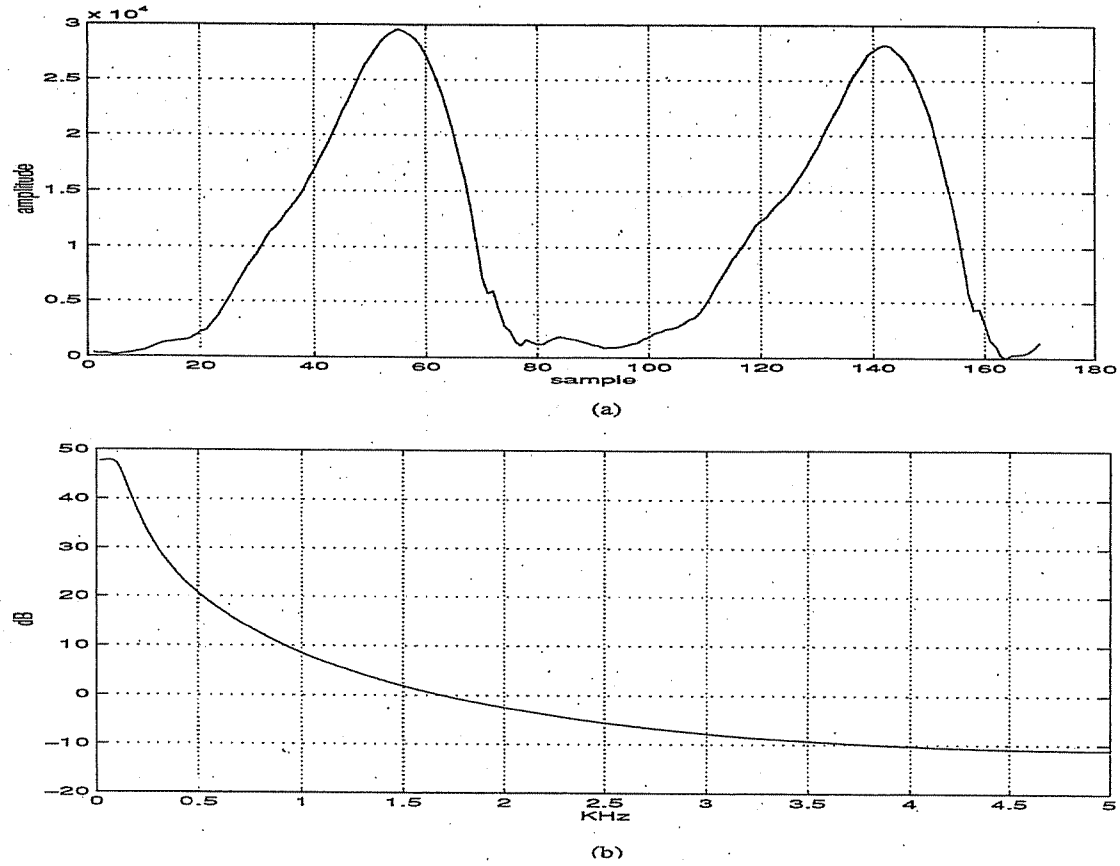


FIGURE A7.1 A typical glottal volume–velocity waveform obtained by inverse filtering and its spectrum.

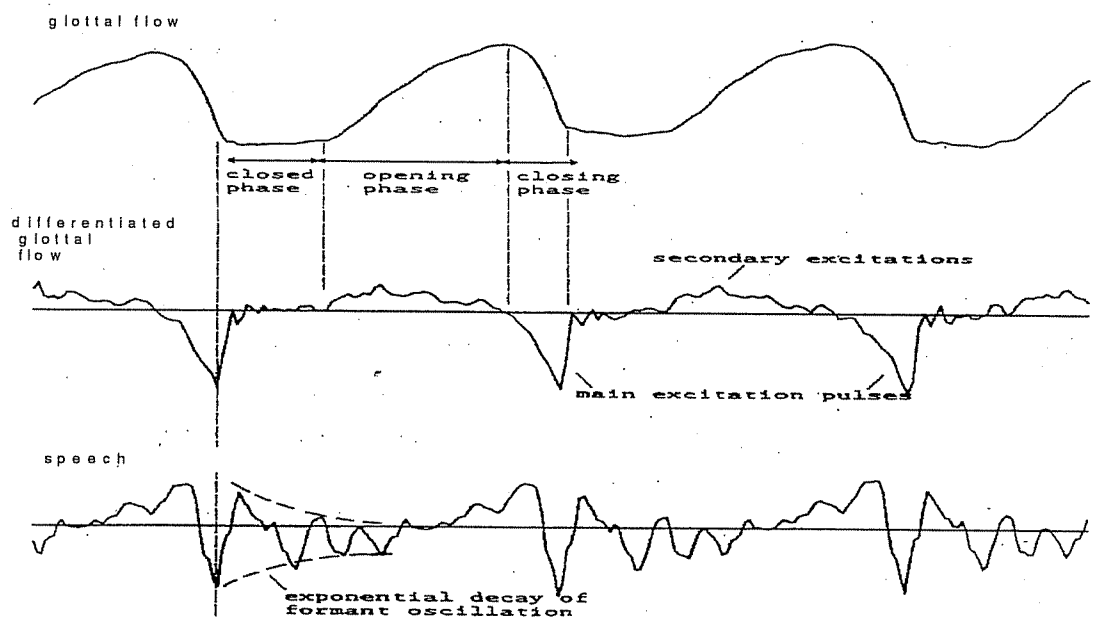


FIGURE A7.2 Illustration of the relationship between the glottal flow, differentiated glottal flow, and speech waveforms.

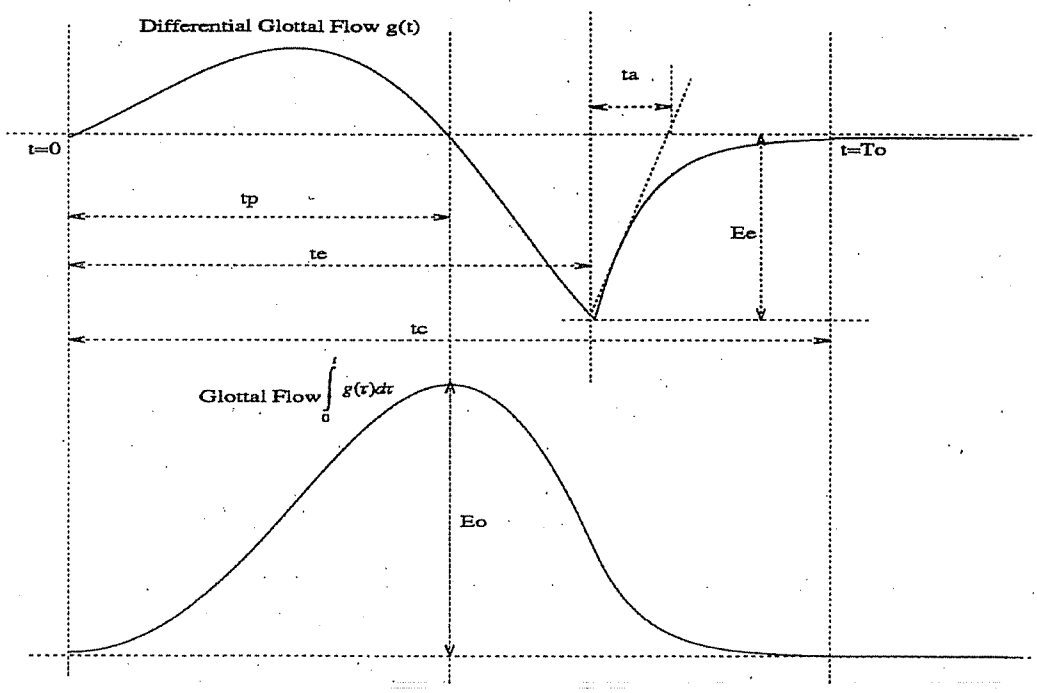


FIGURE A7.3 The LF model waveforms.

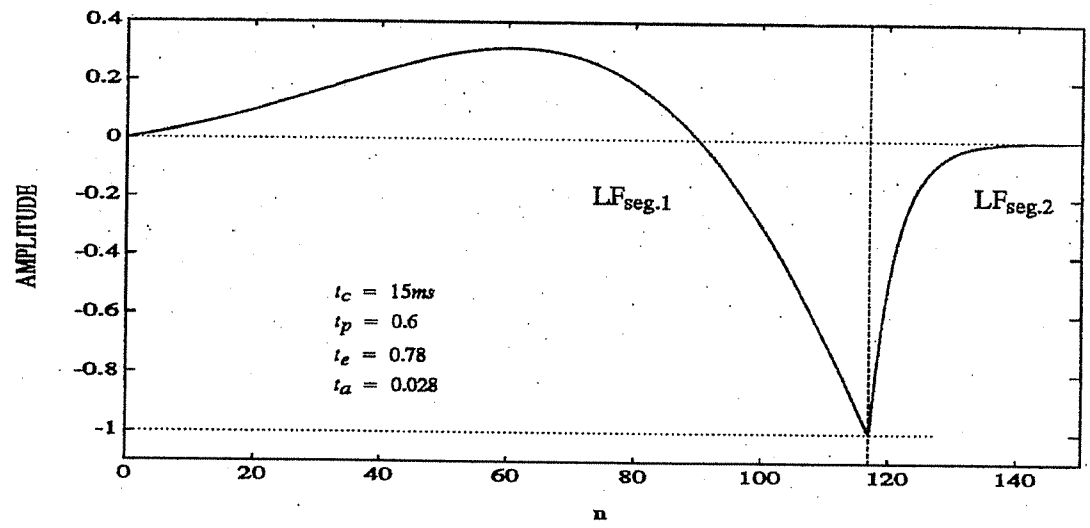
timing parameters, which have been shown to be important in psychoacoustic studies (Childers and Ahn, 1995). Lalwani and Childers (1991a,b) added modulated aspiration noise, jitter, and shimmer to the LF model to form a more complete glottal model. The LF model parameters are such that $0 \leq t_p \leq t_e \leq t_c$, and $t_a \leq 0$.

The first segment of the LF model characterizes the differentiated glottal flow over the interval from the glottal opening to the maximum negative excursion of the waveform. The second segment represents a residual glottal flow that comes after the maximum negative excursion. It can be shown that the spectrum of the first segment is dominated by the exponential component, $e^{\alpha t}$, of which the "negative bandwidth" equals $\frac{\alpha}{\pi}$. Likewise, the frequency response of the second segment can be approximated by a first-order low-pass filter with a cutoff frequency $F_a = 1/(2\pi t_a)$ (Fant and Lin, 1988). As a result, the bandwidths of the first and second segments are, respectively

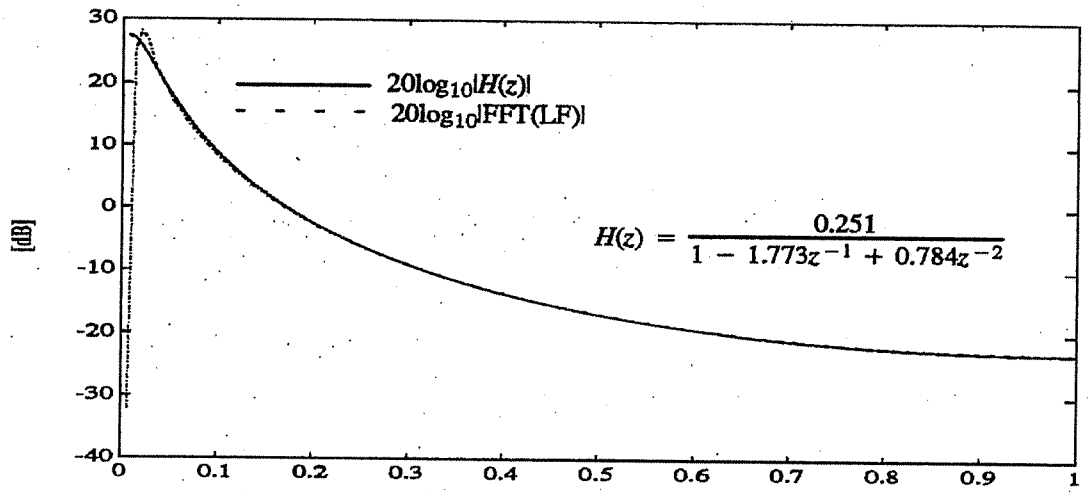
$$B_1 \cong \frac{\alpha}{\pi} \quad \text{and} \quad B_2 \cong \frac{1}{2\pi t_a}$$

Thus, the combination of the two segments of the differentiated glottal flow LF model can be approximated with a two pole filter. The center frequency of the poles and ω_g are nearly the same. Thus, the bandwidths control the filter. The bandwidth of inverse filtered data is in general close to B_1, making it difficult to estimate the waveshape of the first segment. However, B_2 is much greater than the bandwidth of the inverse filtered data. Thus, the second segment retains its waveshape after inverse filtering. However, the phase may be different. A typical example is given in Figure A7.4, which shows the spectra of the first and second segments of the LF model as well as the spectrum of the two pole model. The residue derived from the LF model has a flat spectral envelope and the waveform exhibits a sharp pulse at the conjunction between the two segments, where glottal closure occurs in the LF model. A knowledge of the relationship between the LF model and the residue serves as an aid for retrieving one from the other. Although the residue does not appear to be informative, its integral tends to favor the shape of the LF model waveform, as seen in Figure A7.5, where a known excitation waveform is used to synthesize the vowel /IY/. The synthesized speech is inverse filtered to obtain the residue, which in turn is integrated. This latter waveform is quite similar to the original excitation, as it should be.

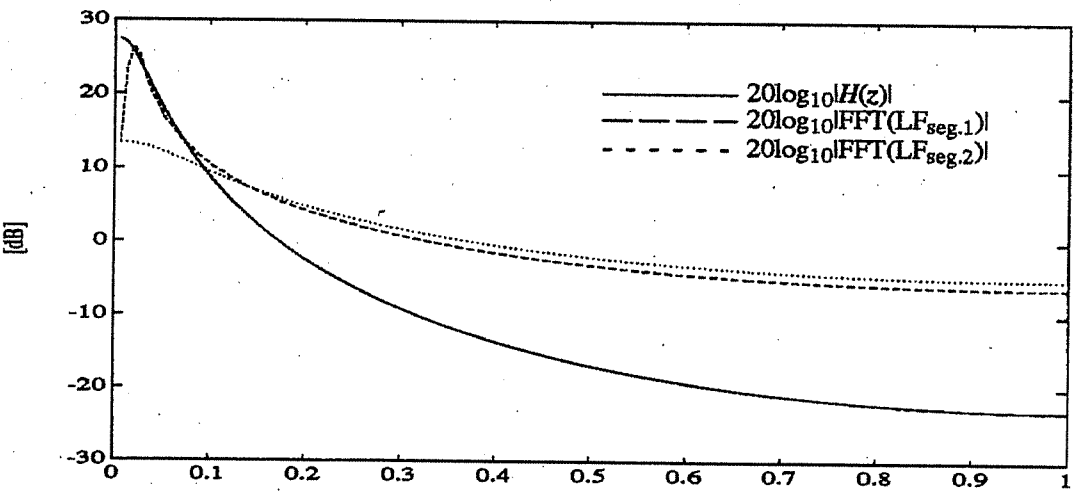
Three differentiated glottal waveforms obtained by inverse filtering are compared in Figure A7.6 with their respective LF model waveforms for three voice types: modal, vocal fry, and breathy. The



(a)



(b)



(c)

FIGURE A7.4 (*a*) The LF model waveform. (*b*) The FFT spectra of the LF model waveform and two pole model, H(z). (*c*) The FFT spectra of the segments of the LF model.

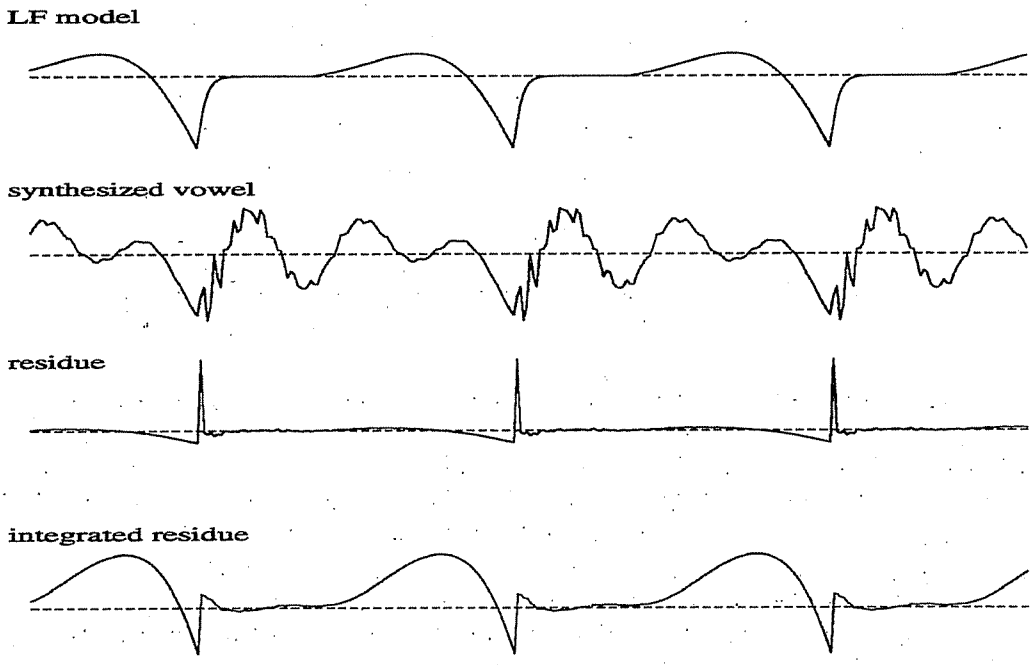


FIGURE A7.5 Illustration of the similarity between the LF model waveform and the integrated residue.

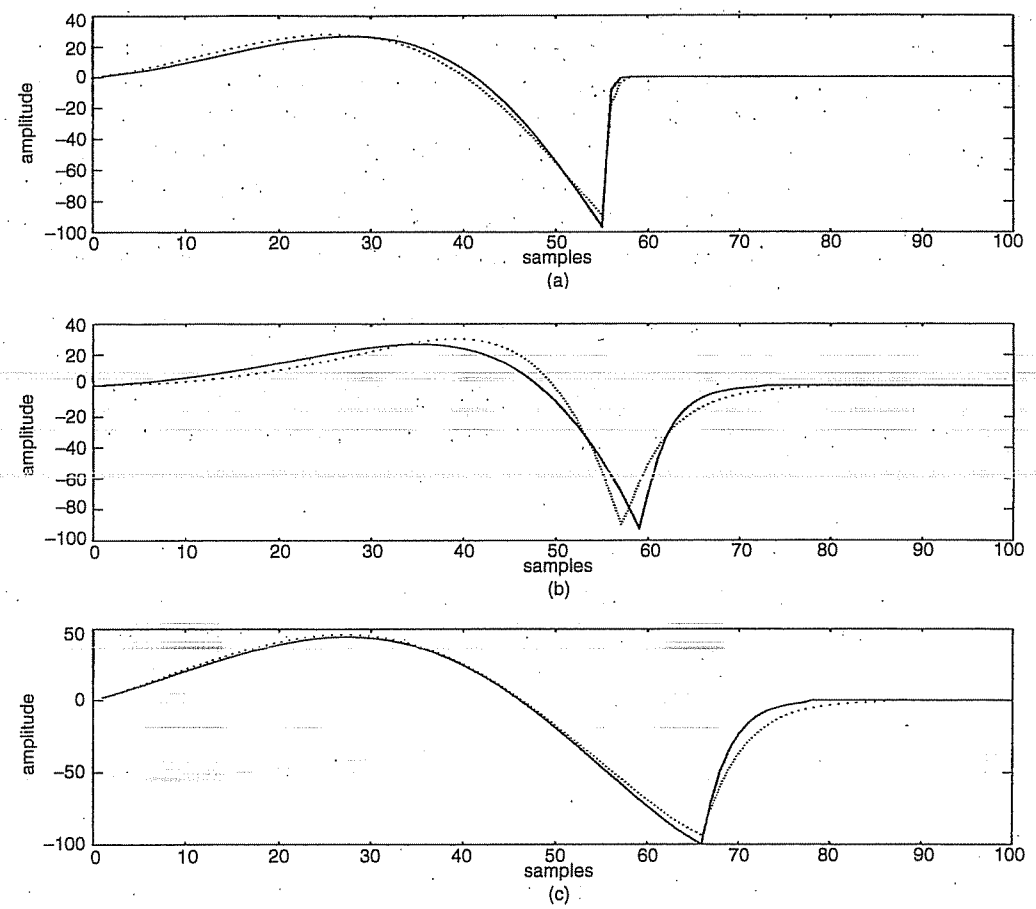


FIGURE A7.6 Comparison of differentiated glottal waveforms with their respective LF model waveforms for three voice types. The solid line represents the data and the dotted line is the model.

LF model parameters are: (*a*) $t_p = 41.3$, $t_e = 55.4$, $t_a = 0.4$, $t_c = 58.2$ (modal); (*b*) $t_p = 48.1$, $t_e = 59.6$, $t_a = 2.7$, $t_c = 72.0$ (vocal fry); and (*c*) $t_p = 46.2$, $t_e = 66.0$, $t_a = 2.7$, $t_c = 77.1$ (breathy).

From experiments by Childers and Lee (1991), Childers and Hu (1994), and Childers and Ahn (1995), some general characteristics of the time domain and frequency domain properties of the glottal waveform have been determined. Before these characteristics are tabulated, we define some terms that often occur in the literature. The **open quotient** of the glottal vibratory cycle is

$$\text{OQ} = \frac{\text{open glottal interval}}{\text{pitch period}} \tag{A7.2.3}$$

The open quotient is primarily determined by the **glottal pulse width**. The OQ can be expressed as follows using the LF model parameters.

$$\text{OQ}_{\text{LF}} = \frac{(t_e + kt_a)}{T_0} \tag{A7.2.4}$$

where the value of k is a function of the parameter t_a. It has been found that k has values in the range 2.0 to 3.0 when $0\% < t_a < 10\%$, where t_a is represented by a percentage of the pitch period. Note that $k = 0$ when $t_a = 0$. In using actual data to make calculations, a definition of various events is needed. For example, for the LF model calculations, the instant of glottal closure is defined as the instant at which the glottal flow waveform falls below 1% of the peak value of the waveform.

The **speed quotient** is

$$\text{SQ} = \frac{\text{opening interval}}{\text{closing interval}} = \text{SQ}_{\text{LF}} = \frac{t_p}{t_e + kt_a - t_p} \tag{A7.2.5}$$

The SQ is a measure of **glottal pulse skewness**. Another definition often used for the speed quotient is

$$\text{SQ} = \text{SQ}_{\text{LF}} = \frac{t_p}{t_c - t_p} \tag{A7.2.6}$$

However, this definition is such that often $t_e + kt_a$ is nearly equal to t_c, so the two definitions reduce to the same.

The **abruptness of closure** of the glottal pulse is measured by the value of t_a. If t_a is small, then the abruptness of closure is fast.

The **aspiration noise** is determined by the **signal-to-noise ratio** (SNR).

The **spectral tilt** is defined as the slope of the spectrum in dB of the glottal waveform. This is typically −12 dB/octave, but can vary from −6 dB to −18 dB. Low spectral tilt is defined as −6 dB, medium as −12 dB, and high as −18 dB.

The **normalized noise energy** (NNE) is 10 log [NE/SE], where SE is the total signal (speech plus noise) energy. The NNE is often measured by calculating the energy in the noise spectrum and the energy in the total signal spectrum. The **harmonics-to-noise ratio** (HNR) is 10 log [HE/NE], where NE is the noise energy and HE is the harmonic energy. The HE is calculated by subtracting NE from SE. The **harmonics richness factor** (HRF) is 10 log of the ratio of the energy of the sum of all the harmonics in the speech signal (less the fundamental) to the energy of the fundamental frequency. The NNE, HNR, and HRF are all measured in dB.

Table A7.1 summarizes the features of the time domain glottal factors, while Table A7.2 summarizes the frequency domain factors for several voice types. Some typical numerical values

TABLE A7.1. Time Domain Glottal Factors

Voice	Pitch period	Pulse width	Pulse skewness	Abruptness of closure	Aspiration noise	Jitter	Shimmer
Modal	Medium	Medium	Medium	Medium	High	Low	Low
Vocal fry	Long	Short	High	Fast	Medium	High	High
Breathy	Medium	Long	Low	Slow	Low	Low	Medium
Rough	Medium	Medium	Medium	Medium	Medium	High	High
Hoarse	Medium	Medium	Medium	Medium	Low	High	High

TABLE A7.2. Frequency Domain Glottal Factors

Voice	Fundamental frequency	Spectral tilt	Harmonic richness factor	Harmonic to noise ratio
Modal	Medium	Medium	Medium	High
Vocal fry	Low	Low	High	Medium
Breathy	Medium	High	Low	Medium
Rough	Medium	Medium	Medium	Low
Hoarse	Medium	Medium	Medium	Low

for some the LF and noise model parameters are summarized later after we introduce the noise model.

Another form of the LF model is called the transformed LF model (Fant, 1995), which is a simplified version of the LF model described previously. However, the features of the model are the same.

An alternative to the LF model was recently introduced by Velhuis (1998), which is computationally more efficient than the LF model.

In summary, in some research applications it is desired to fit a LF model to actual data. In this case, the LF model parameters are determined by a least square error fit of the data to the model. For the speech synthesizers in Chapters 6 and 7, the user is provided with a means to adjust the model parameters to obtain a desired LF excitation model waveform.

A7.3 NOISE MODEL

Both the LF model and the polynomial glottal waveform models tend to account for the low-frequency component of the glottal waveform. By subtracting the low-frequency component (as a model) from the differentiated glottal waveform (actual data), a noise-like waveform is obtained. This noise-like signal is attributed to turbulent noise and is important for the naturalness of both real and synthetic speech (Holmes, 1976; Kasuya et al., 1986; Klatt, 1980; Lalwani and Childers, 1991a,b). The parameters that are important for this noise are intensity, spectral shape, and timing (Childers and Lee, 1991; Klatt and Klatt, 1990; Lalwani and Childers, 1991a,b). The turbulent noise model used here is shown in Figure A7.7.

The noise model uses Gaussian white noise along with six parameters, defined as follows.

- snr: The power ratio of the low-frequency component to the aspiration noise.
- ampl: The amplitude modulation index 1 such that $0 \leq \text{ampl} \leq 1.0$.

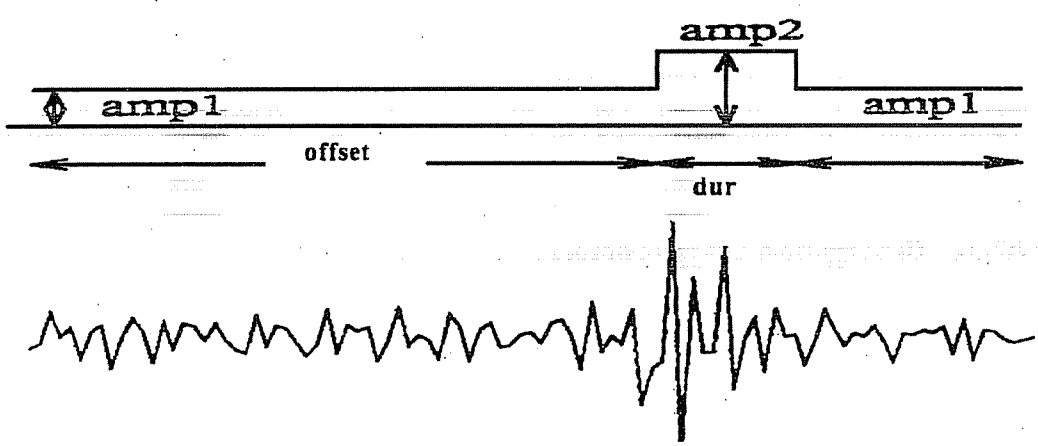


FIGURE A7.7 The turbulent noise model.

TABLE A7.3. Typical LF and Noise Model Parameter Values for Several Voice Types

Voice	t_p (%)	t_e (%)	t_a (%)	t_c (%)	Jitter (%)	snr (dB)	amp1 (%)	amp2 (%)	offset (%)	dur (%)
Modal	41.0	55.0	0.4	58.0	2.0	40.0	0.0	100.0	50.0	50.0
Vocal fry	48.0	59.0	2.7	72.0	10.0	20.0	0.0	100.0	20.0	20.0
Breathy	46.0	66.0	2.7	77.0	5.0	20.0	100.0	100.0	50.0	50.0
Whisper	50.0	80.0	8.0	100.0	2.0	−20.0	100.0	100.0	50.0	50.0
Falsetto	50.0	80.0	8.0	100.0	2.0	50.0	0.0	0.0	50.0	50.0
Harsh	25.0	30.0	1.0	50.0	10.0	10.0	100.0	100.0	50.0	50.0

- amp2: The amplitude modulation index 2 such that $0 \leq \text{amp2} \leq 1.0$.
- offset: The duration of the noise with amp1. This starts from the instant of the opening of the glottis.
- dur: The duration of amp2. The dur starts from the end of offset. The sum of the offset and dur must be less than or equal to the pitch period.

The term offset is used for the duration of amp1 for historical reasons in the author's research and has no other special significance.

The values used for the noise parameters are speaker dependent and speech dependent. Generally, for modal voices the snr is small, being about 0.25%. Some examples for these parameters are summarized in Table A7.3. This table is a summary of values obtained from the analysis of data (Childers and Ahn, 1995) as well as used in speech synthesis experiments in voice conversion (Childers et al., 1989; Childers and Hu, 1994; Childers and Lee, 1991; Childers and Wu, 1990; Shue, 1995).

The values for modal, vocal fry, and breathy voice were determined experimentally by analyzing data (Childers and Ahn, 1995), while the values for the three other voice types were determined via listening tests with synthesized speech. The 100% value for amp1 and amp2 means that the sliders on the noise settings in the source specification option are to be set to the maximum value of 1.0, or 100%. Jitter is a parameter that is not available in the formant synthesizer in Chapter 6.

A7.4 SOME EXAMPLES

Figure A7.8 shows some example excitation LF model waveforms with noise added for six voice types. These waveforms were created using the formant synthesizer. Figure A7.9 shows the synthesized vowel /IY/ waveforms using the excitation waveforms in Figure A7.8 as the input to the formant synthesizer. The model parameters are given in Table A7.4, which is similar to Table A7.3.

TABLE A7.4. LF and Noise Model Parameters for the Synthesized Data Shown in Figures A7.8 and A7.9

Voice	t_p (%)	t_e (%)	t_a (%)	t_c (%)	Jitter (%)	snr (dB)	amp1 (%)	amp2 (%)	offset (%)	dur (%)
Modal	45.0	60.0	0.5	65.0	2.0	40.0	0.0	100.0	50.0	50.0
Vocal fry	20.0	25.0	0.2	35.0	10.0	20.0	0.0	100.0	20.0	20.0
Breathy	50.0	80.0	8.0	100.0	5.0	20.0	100.0	100.0	50.0	50.0
Whisper	50.0	80.0	8.0	100.0	2.0	−20.0	100.0	100.0	50.0	50.0
Falsetto	50.0	80.0	8.0	100.0	2.0	50.0	0.0	0.0	50.0	50.0
Harsh	25.0	30.0	1.0	50.0	10.0	10.0	100.0	100.0	50.0	50.0

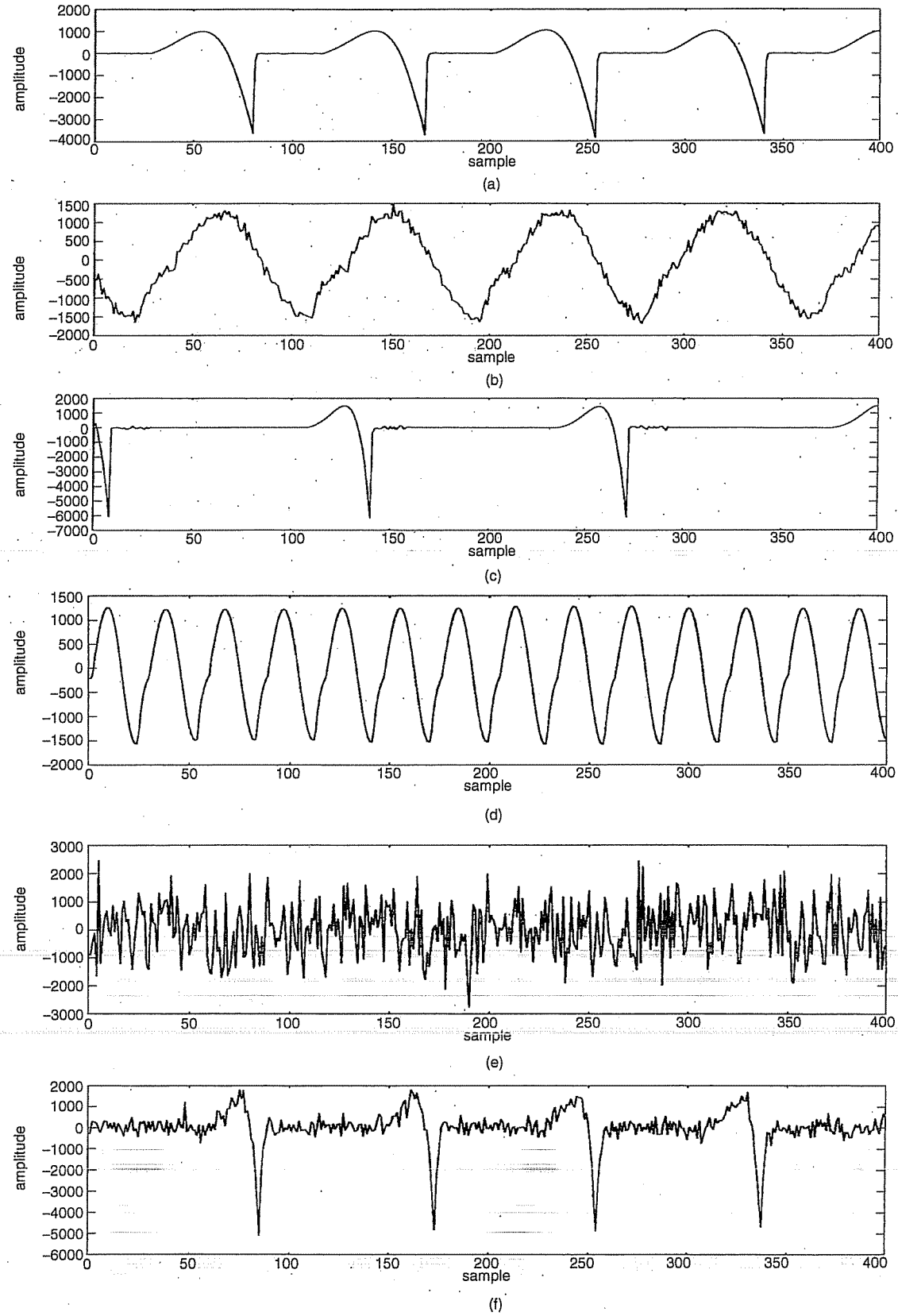


FIGURE A7.8 Examples of LF model excitation waveforms with noise added for six voice types: (*a*) modal, (*b*) breathy, (*c*) vocal fry, (*d*) falsetto, (*e*) whisper, and (*f*) harsh.

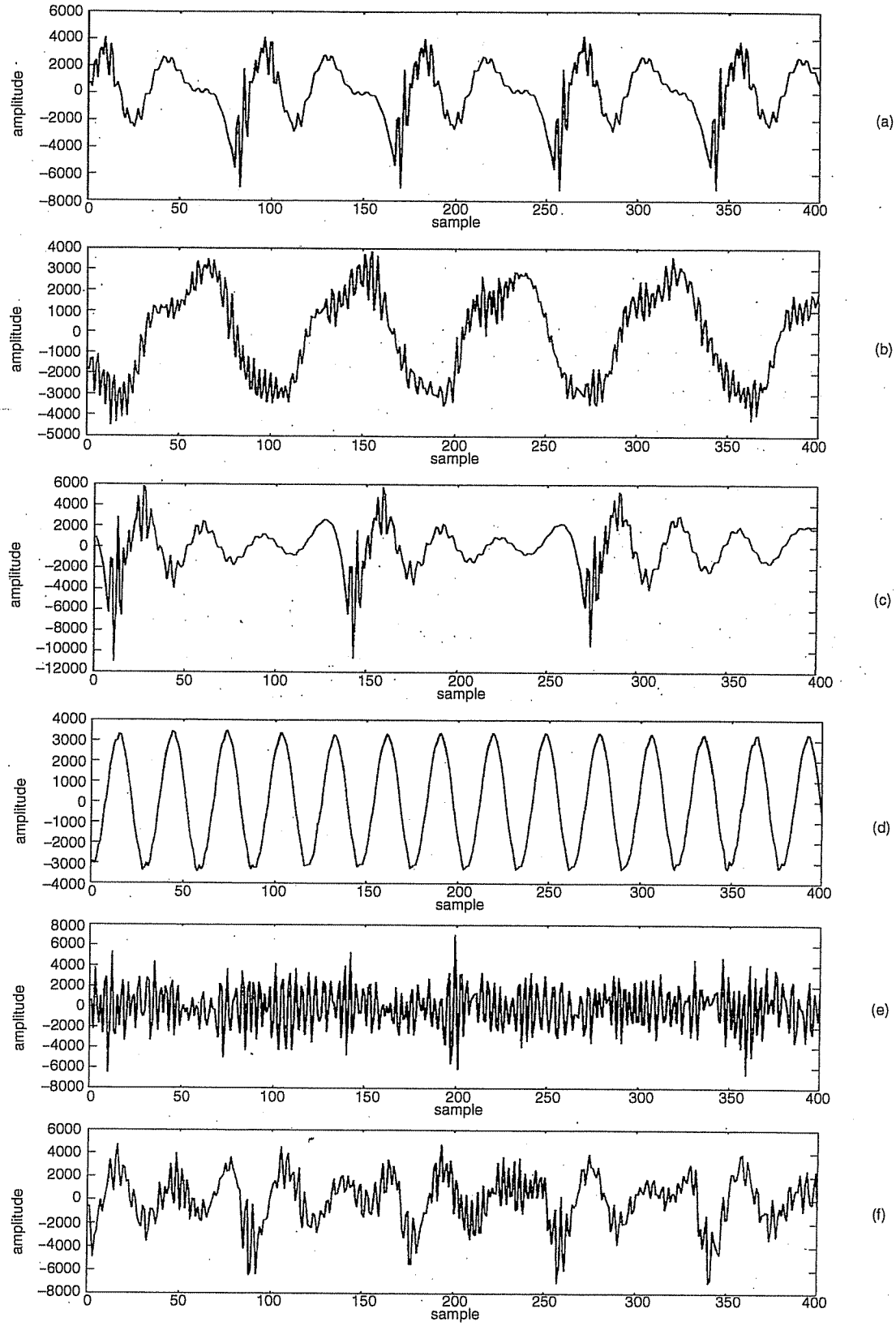


FIGURE A7.9 The synthesized vowel /IY/ using the excitation waveforms and the formant synthesizer: (*a*) modal, (*b*) breathy, (*c*) vocal fry, (*d*) falsetto, (*e*) whisper, and (*f*) harsh.

A7.5 THE POLYNOMIAL MODEL

The polynomial model described here appears in Childers and Hu (1994) and is similar to the Milenkovic (1993) polynomial model. The model is

$$p(t) = c_0 + c_1\tau + c_2\tau^2 + c_3\tau^3 + c_4\tau^4 + c_5\tau^5 + c_6\tau^6 \qquad \text{(A7.5.1.)}$$

where $\tau = \frac{t}{T}$, t is the independent variable and T is the period of the pulse waveform (or the pitch period). If the model is being used to fit data, then the coefficients are determined by fitting the polynomial function, p(t), to the estimated differentiated glottal waveform in a least squares sense, which is similar to that described in Milenkovic (1993). However, for speech synthesis using the software in Chapters 6 and 7, the user can adjust the coefficients to obtain a desired excitation waveform. An example of a waveform obtained by this model is shown in Figure A7.10 for both the glottal flow (lower) and its derivative (upper). While the coefficients of the polynomial have no obvious physical or physiologic interpretation, they have been shown to be suitable for the synthesis of high-quality speech (Milenkovic, 1993).

An overview of a system that uses the features of the LF, polynomial, and noise models for a glottal excited LP speech synthesizer is given in Figure A7.11 (Childers and Hu, 1994). In such a system, the excitation and noise models are designed as codebooks. The features of this system include an all pole synthesis filter, voiced and unvoiced excitations and gains, a glottal pulse codebook for voiced sounds, a stochastic codebook for unvoiced sounds, and an analysis to determine the glottal closure instants (GCIs).

For additional details on the formant and LP synthesizers consult the following PhD dissertations: Hu (1993), Shue (1995), and Hsiao (1996).

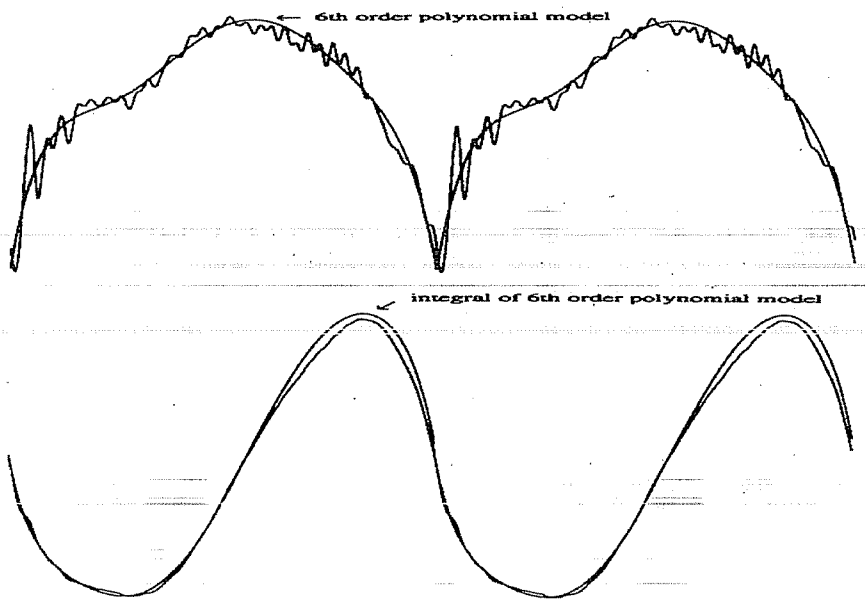


FIGURE A7.10 An example of the polynomial model waveforms: (upper) derivative of the glottal flow, (lower) the glottal flow.

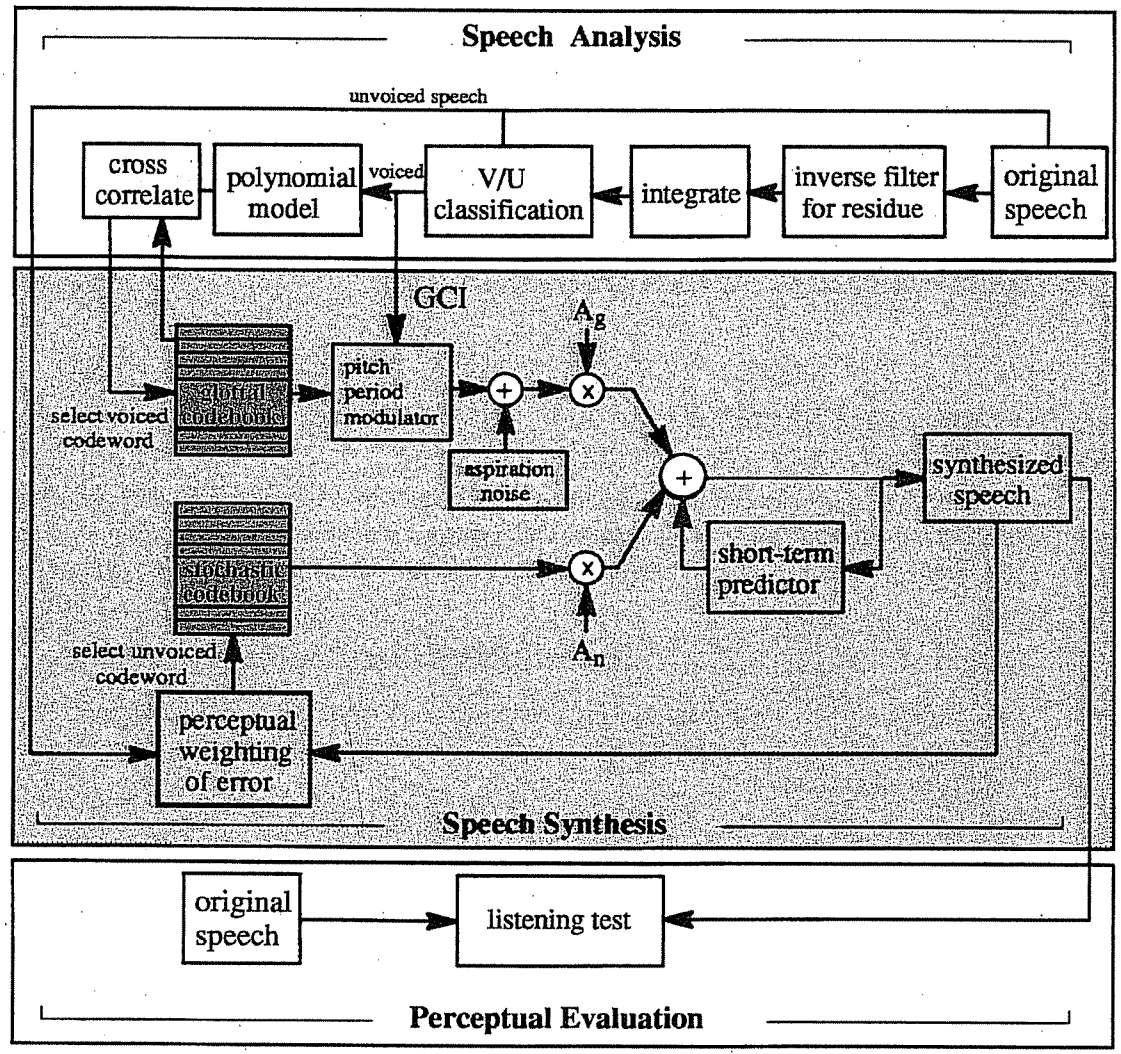


FIGURE A7.11 A glottal excited, linear prediction (GELP) speech synthesis system.