APPENDIX 11

ARTICULATORY SPEECH SYNTHESIS: THEORY

A11.1 INTRODUCTION

Both formant and LP synthesis models represent acoustic domain models of speech, with no interaction between the glottal flow excitation and the vocal-tract filter. Articulatory synthesis is the production of speech sounds using a model of the vocal tract, which directly or indirectly simulates the movements of the speech articulators. This method provides a means for gaining an understanding of speech production and for studying phonetics. In this model coarticulation effects arise naturally, and in principle it should be possible to deal with glottal source properties, including interaction between the vocal tract and the vocal folds, as well as the contribution of the subglottal system, and the effects of the nasal tract and sinus cavities.

Articulatory synthesis usually consists of two separate components as shown in Figure A11.1. In the articulatory model, the vocal tract is divided into numerous small sections and the corresponding cross-sectional areas are used as parameters to represent the vocal tract characteristics. In the acoustic model, each cross-sectional area is approximated by an electrical analog transmission line. To simulate the movement of the vocal tract, the area functions change with time. Each sound is represented by a target configuration of the vocal tract. The movement of the articulators of the vocal tract is specified by a change in the vocal tract configuration. This is done by specifying changes in the articulators from frame to frame over the speech file that is to be synthesized.

Presently, the complexity of articulatory synthesis is partially due to the analysis procedure, which usually requires an "articulatory-to-acoustic inverse transformation" from the speech signal, that is, speech inverse filtering. The complexity of the relationship between articulatory gestures and the acoustic signal makes it difficult to automatically generate the details of articulatory control needed to produce a synthetic copy of a given sample of human speech. Despite such drawbacks, articulatory speech synthesis has several advantages.

- The model has a direct relation to the human speech production process. Consequently, it is conjectured that articulatory synthesis may lead to a simpler and more elegant synthesis by rule; for example, text-to-speech applications (Parthasarathy and Coker, 1990, 1992) and articulation-based speech recognition systems (Erler and Deng, 1993).
- The articulatory parameters in the human voice production system vary slowly. Consequently, researchers have suggested that these parameters are potential candidates for efficient coding; for example, low bit-rate speech communication (Flanagan et al., 1980).
- To the extent that we can accurately obtain the speech gestures (articulatory movements or trajectories), articulatory synthesizers may be valuable for research scientists and physicians, since such synthesizers can be used to study linguistic theories, to provide a feedback mechanism for teaching speech production, and to explore the effects of vocal tract surgical techniques on speech production prior to surgical intervention (Childers, 1991); and they hold out the ultimate promise of high quality, natural-sounding speech with a simple control scheme (Klatt, 1987).

A properly constructed articulatory synthesizer is capable of reproducing all the naturally relevant effects for the generation of fricatives and plosives, modeling coarticulation transitions as well as source-tract interaction in a manner that resembles the physical process that occurs in real speech

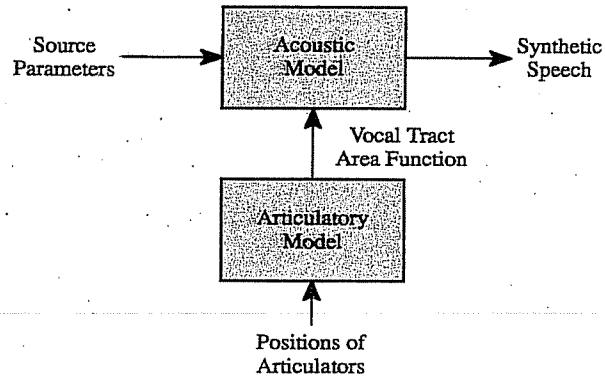


FIGURE A11.1 A model of articulatory speech synthesis.

production. Articulatory synthesizers will continue to be of great importance for research purposes, and to provide insights into various acoustic features of human speech. Thus, an articulatory synthesizer may provide both an efficient description of natural speech and a means for synthesizing natural-sounding speech. However, a major problem with the articulatory synthesizer is the lack of a means to derive articulatory configurations from the speech signal using speech inverse filtering. The software toolbox provided in Chapter 10 provides one means for speech inverse filtering. In addition, the toolbox displays the following characteristics:

- The articulator movements (gestures) for synthesizing words or sentences.
- The cross-sectional area and acoustic transfer function of the vocal tract.
- The pressure and volume–velocity waveforms at selected points in the vocal tract.
- The excitation source waveform and power spectral density.
- The synthesized speech waveform.

The optimization procedure is based on the simulated annealing (SA) algorithm. Using this method, the articulatory parameters are optimized to minimize the error distance between the natural (target) and the model-generated first four formants. The excitation source model for the articulatory synthesizer is a combination of a glottal source model (a modified LF-model), a subglottal, and a glottal area model. The toolbox can modify and calculate the vocal tract and the nasal tract models, which in turn are used to calculate the acoustic transfer function at selected points in the vocal tract. The toolbox is able to insert a noise source model at the center of a vocal tract constriction, or downstream from the constriction, or distributed along a specified spatial interval of the constriction. In addition, the toolbox can simulate viscous, heat conduction, and yielding wall losses in the vocal and nasal tracts. This model also includes the effects of the sinus cavities.

The speech synthesis process can vary the following parameters.

- The synthesis sampling frequency.
- The degree of nasalization, including the velopharyngeal port opening.
- The excitation source parameters, including subglottal coupling.
- The type and location of excitation.

A11.2 ARTICULATORY SPEECH SYNTHESIS

The articulatory speech synthesizer is based on a model of the physiology of the human speech production process. As shown in Figure A11.1, the articulatory synthesizer has two components. The articulatory model represents the articulatory positions, which are converted into vocal tract cross-sectional area functions. The acoustic model, which includes subglottal coupling, source-tract

interaction, the vocal tract, the nasal tract with sinus cavities, and acoustic radiation, simulates speech production and propagation as well as the physics of the physiological-to-acoustic transformation. This appendix presents the implementation of the articulatory model and the realization of the acoustic model. Our articulatory model is based on Mermelstein's model (1973), which is implemented in MATLAB. The time-domain approach is used to implement the acoustic model, since it offers the ability to simulate the dynamic properties of the vocal system as well as a method to improve the quality of the synthesized speech. Methods for estimating articulatory data from acoustic measurements are reviewed and described later in this appendix.

A11.2.1 Review of Articulatory Models

According to the acoustic theory of speech production, the human vocal tract can be modeled as an acoustic tube with nonuniform and time-varying cross-sections. This model modulates the excitation source to produce various sounds. The acoustic tube can be adjusted to various shapes by adjusting the articulatory parameters. These articulatory parameters are expressed in the form of a vector and specify the positions of the tongue body, tongue tip, jaw, lips, hyoid, and velum. Articulatory models are well known in the literature and can be classified into two major types: parametric area models and midsagittal distance models.

A11.2.1.1 Parametric Area Models The parametric area models do not represent articulatory positions directly but rather, concentrate on modeling the area function as a function of distance along the tract subject to certain constraints (Atal et al., 1978; Fant, 1960; Fant and Lin, 1991; Flanagan et al., 1980; Lin, 1990; Stevens, 1998; Stevens and House, 1955; Yu, 1993). A common feature of these models is a specification of the minimum constriction area and its axial location. The area of the vocal tract is usually represented by a continuous function such as a hyperbola, a parabola, or a sinusoid (Lin, 1990). Consonant articulations have generally not been implemented. Figure A11.2 shows one example of a parametric area model.

A11.2.1.2 Midsagittal Distance Models The midsagittal distance models are usually based on a representation of the midsagittal plane as seen from an x-ray image. They describe the speech articulator movements in a midsagittal plane and require the specification of the positions of the articulators (Levinson and Schmidt, 1983; Mermelstein, 1973; Prado, 1991; Sondhi and Schroeter, 1986) or a means to control the movements of the articulators by rules (Coker, 1976; Parthasarathy and Coker, 1990, 1992). The output is an estimate of the vocal tract cross-sectional area. Visualization and articulatory state interpretation are the major advantages of these models. Figure A11.3 shows an example of a midsagittal distance model.

A11.2.2 Implementation of the Articulatory Model

Articulatory models are used to transform articulatory parameters to a vector representation of the vocal tract cross-sectional area, which in turn are transformed to acoustic characteristics within the vocal tract. Our articulatory model is a modified version of the Mermelstein's model (1973),

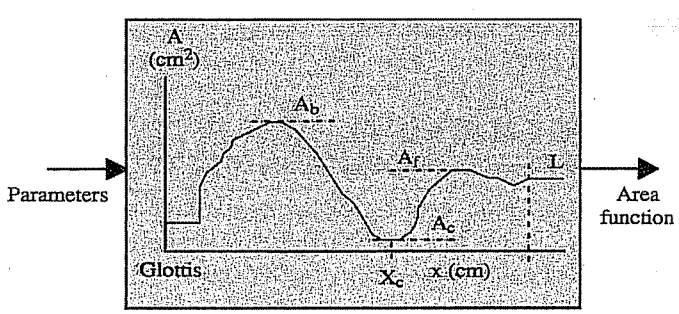


FIGURE A11.2 A parametric area model.

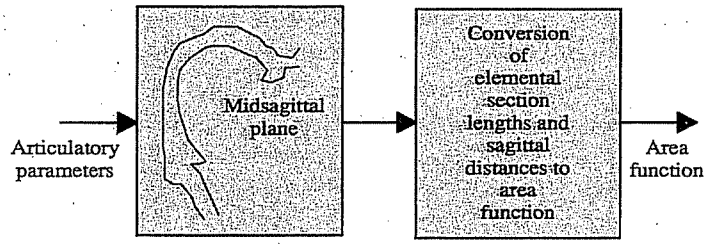


FIGURE A11.3 A midsagittal distance model.

which is implemented in MATLAB. Mermelstein's model (1973) achieves a match between x-ray tracings and a midsagittal vocal tract outline. However, there is insufficient information for a robust representation of the lower part of the pharynx and for the region between the tongue tip and the jaw. Our approach modifies the lower part of the pharynx and optimizes this region whenever necessary. This is discussed in more detail later in this appendix. We have also modified the hyoid and tongue-tip-to-jaw regions.

A11.2.2.1 Articulatory Parameters and Midsagittal Vocal Tract Outline In the articulatory model, a set of variables is used to specify the inferior outline of the vocal tract (Figure A11.4). These variables, referred to as articulatory parameters, are:

- Tongue body center: This is represented with an arc (DL-B) of a circle with a moving center and fixed radius. The tongue body center, denoted as tongc, has polar coordinates (sc, thetaj + thetab) with respect to the fixed point F. However, the rectangular coordinates (tbodyx, tbodyy) are used for display and optimization.
- Tongue tip: The tongue tip is represented by the rectangular coordinates (tipx, tipy) of point T. Arcs B-T and T-PF, specify the tongue blade outline. Since the location of point B varies with the tongue-body center (tongc) and the jaw angle (jaw), the tongue blade movements depend on the tongue body and jaw positions.
- Jaw: The point JAW with polar coordinates (sj, thetaj) are used to represent the jaw location. The distance sj is kept constant for most phonemes. The parameter jaw is used to denote the

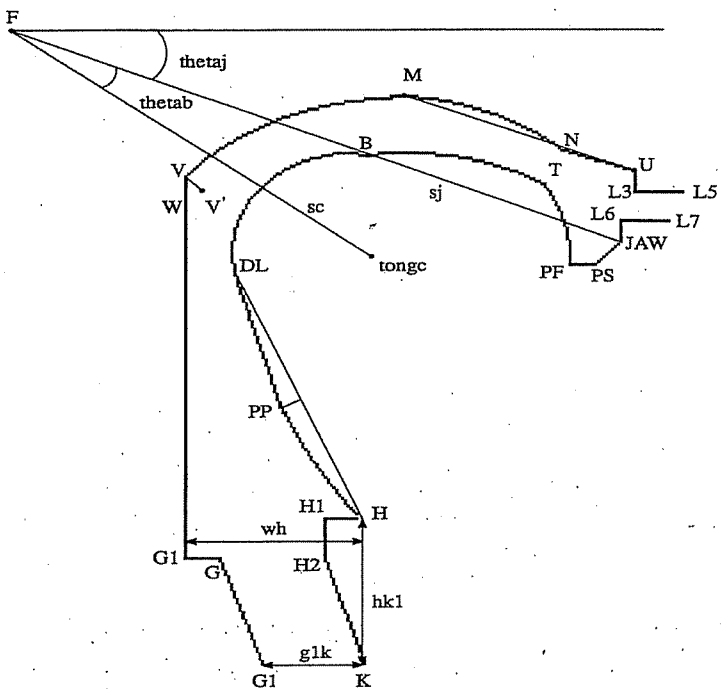


FIGURE A11.4 Articulatory model parameters.

angle thetaj. Note that the concave jaw is approximated by a polyline, a connected sequence of line segments (PF-PS-JAW-L6).

- Lips: The lips are represented by points L5 (upper) and L7 (lower). With respect to the point JAW, the coordinates of the lower lip are represented by (lipp, lipo), which specify the lip protrusion and lip opening, respectively. The use of lipp and lipo as separate variables allows lip closure, lip separation, or rounded lips. The upper lip L5 has the same coordinate values with respect to point U.
- Hyoid: The hyoid is specified by the parameter hyoid, the distance from point PP to the line segment H-DL. The point PP is on the normal bisector of the line segment H-DL, which is tangent to the tongue body arc outline at point DL. The line segment DL-PP and arc PP-H as well as the tongue body determine the anterior shape of the pharynx. The point H represents the intersection of the anterior edge of the epiglottis with the top edge of the hyoid bone. The point K represents an estimate of the anterior extremity of the larynx.
- The superior outline of the vocal tract is represented by the position of the upper teeth, U, the hard palate curve U-N-M, the highest point on the maxilla M, the soft palate arc M-V, the velum position V, the back wall of pharynx position W, and the highest point of the periarytenoid G. On the hard palate curve, the point N is located on the line segment M-U such that the distance M-N is twice the distance N-U. Circular arcs M-V and M-N are drawn with centers on a vertical line through M. The posterior–superior outline is generally considered fixed except for the soft palate curve near the velum point V. To specify the opening area of the velopharyngeal port, we treat the velum as an articulatory parameter.
- Velum: The state of the velum is represented by the position V of the tip of the uvula moving along a line segment (V-V′). The velar opening area is assumed proportional to the distance between the point V and the most elevated point of the velum. This distance is specified by the variable velum.

A11.2.2.2 Determination of the Vocal Tract Section Lengths and Cross-Sectional Areas The vocal tract cross-sectional area function is determined by the areas of the sections with projections on the X-Y plane that form the sagittal grids of the vocal tract, as shown in Figure A11.5. These grid lines vary with the positions of the articulators (they are fixed in Mermelstein's model); that is, the interval between two adjacent parallel sagittal grid lines

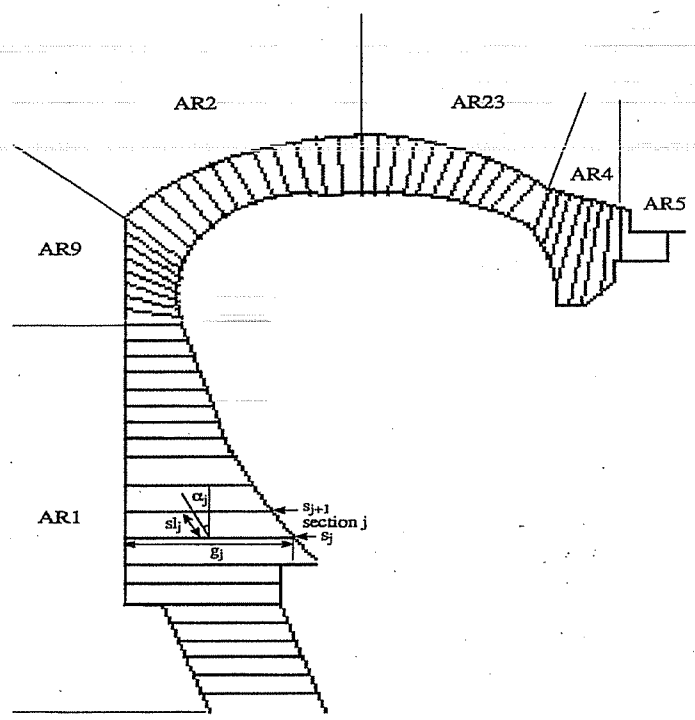


FIGURE A11.5 Midsagittal grids and areas for the articulatory model.

(regions AR1 and AR5) or the angle between two adjacent radial sagittal grid lines (rest regions) is not fixed. A total of 60 sections, 59 sections for the vocal tract plus one section (fixed length and area) for the outlet of the glottis, are used in our model. This feature provides more reliable estimates of the sagittal distances and cross-sectional areas.

The distance between the midpoints of two consecutive sagittal lines, s_j and s_{j+1}, represents the length of section j, sl_j (see Figure A11.5). The sagittal distance g_j of section j is defined as the grid line segment length between posterior–superior and anterior–inferior outlines. The sagittal distances are converted to cross-sectional areas by an empiric function based on previously published data (Mermelstein, 1973). In general, the cross-sectional area function is formulated as

$$A_j = F(j, g_j)\cos(\alpha_j) \qquad \text{(A11.2.2.2.1)}$$

where $j = 2$ (vocal tract inlet), ..., 60 (lips). F is an empiric function and has a different formula for the pharyngeal region, oral region, and labial region, and α_j is the deviation angle of the direction of wave propagation from the normal to the jth grid line (Guo and Milenkovic, 1993; Mermelstein, 1973; Rubin et al., 1981).

In the pharyngeal region (AR1 and AR9 in Figure A11.5), the empiric function is

$$F(j, g_j) = \pi g_j b_j \qquad \text{(A11.2.2.2.2)}$$

where g_j is one axis and $b_j = g_j + \Delta g$, where Δg belongs to the interval [1.5 3], is another axis of the ellipse, since we approximate each pharyngeal section as an elliptic cylinder. The b_j is proportionally increased as the grid line moves toward the larynx. In the soft palate region (AR2 in Figure A11.5), the empiric function has the form

$$F(j, g_j) = (2.0)g_j^{1.5} \qquad \text{(A11.2.2.2.3)}$$

In the hard palate region (AR23 in Figure A11.5), the empiric function is given by

$$F(j, g_j) = (1.6)g_j^{1.5} \qquad \text{(A11.2.2.2.4)}$$

For the labial region (AR5 in Figure A11.5), the empiric function is

$$F(j, g_j) = g_j[2.0 + 1.5(\text{lipo} - \text{lipp})] \qquad \text{(A11.2.2.2.5)}$$

For the other region (AR4 in Figure A11.5), the empiric function is

$$\begin{aligned} F(j, g_j) &= (1.5)g_j, && \text{for } g_j < 0.5 \\ &= (0.75) + 3(g_j - 0.5), && \text{for } 0.5 < g_j < 2 \\ &= 5.25 + 5(g_j - 2), && \text{for } g_j > 2 \end{aligned} \qquad \text{(A11.2.2.2.6)}$$

A11.2.2.3 Calculation of Formant Frequencies from the Vocal Tract Cross-Sectional Areas The calculation of formant frequencies from a given vocal tract cross-sectional area function has been well established in the acoustic theory of speech production (Atal et al., 1978; Badin and Fant, 1984; Fant, 1960; Fant, 1985; Fant and Lin, 1991; Lin, 1990, 1992; Stevens, 1998; Wakita and Fant, 1978). By computing the acoustic transfer function of a given vocal tract configuration, we can decompose the formant frequencies from the denominator of the acoustic transfer function. Refer to Appendix A11-B for the detailed acoustic transfer function calculations.

Let an all-pole acoustic transfer function be

$$H(s) = \frac{1}{H_p(s)} \qquad \text{(A11.2.2.3.1)}$$

The denominator $H_p(s)$ is normally a complex number

$$H_p(s) = N_b(s) + jN_a(s) \qquad \text{(A11.2.2.3.2)}$$

For a lossless vocal tract $N_a(s)$ is zero. When the losses are small, $N_a(s)$ is small compared with $N_b(s)$. Consequently, the roots of the complex function $H_p(s)$ should be located in the neighborhood of the roots of $N_b(s)$. Based on this assumption, a two-step approach was proposed by Fant (1960) and was referred to as the N_b method (Lin 1990, 1992). Figure A11.6 illustrates the flow chart of the method.

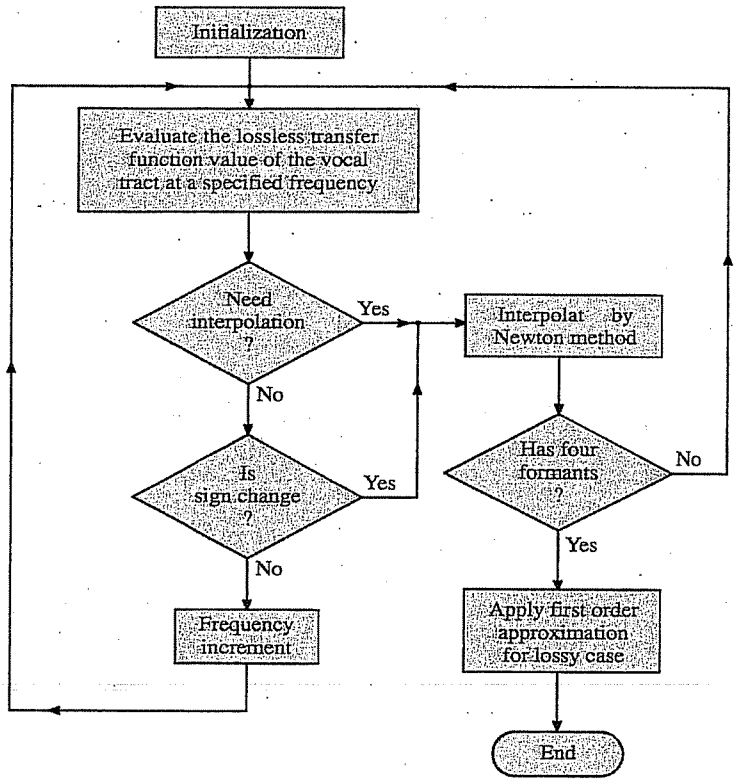


FIGURE A11.6 Flow chart of the N_b method for the decomposition of formants.

The first step of the N_b method is to search for the roots of $N_b(s) = 0$. At a given frequency f_n, the value $N_b(j2\pi f_n)$ is computed. The frequency is next increased (a few hundred hertz) and the value $N_b(j2\pi f_{n+1})$ is computed at the new frequency f_{n+1}. If the polarity changes, $(N_b(j2\pi f_n))(N_b(j2\pi f_{n+1})) < 0$, within this interval, then a root is present. Newton's approximation or other methods can be used to determine the root within this interval. Let f_0 be the estimated root frequency found by setting $N_b(s) = 0$. The second step is to account for the finite $N_a(s)$ by means of a first-order approximation term for $H_p(s)$ in the vicinity of $(j2\pi f_0)$.

$$H_p(s) \cong H_p(j2\pi f_0) + (s - j2\pi f_0)(H'_p(j2\pi f_0)) \qquad \text{(A11.2.2.3.3)}$$

where

$$H'_p(j2\pi f_0) = \left.\frac{d(H_p(s))}{ds}\right|_{s=j2\pi f_0} = N'_a(j2\pi f_0) - j(N'_b(j2\pi f_0)) \qquad \text{(A11.2.2.3.4)}$$

Set $H_p(s) = 0$ and let the roots be denoted as

$$s_n = \sigma_n + j(2\pi f_0 + \Delta\omega_n) \qquad \text{(A11.2.2.3.5)}$$

From the above equations, we have

$$\sigma_n = \frac{N_a N'_b}{(N'_a)^2 + (N'_b)^2} \quad \text{and} \quad \Delta\omega_n = -\sigma_n \frac{N'_a}{N'_b} \qquad \text{(A11.2.2.3.6)}$$

The final pole frequency and bandwidth is given by

$$f_n = f_0 + \frac{\Delta\omega_n}{2\pi} \quad \text{and} \quad B_n = -\frac{\sigma_n}{\pi} \qquad \text{(A11.2.2.3.7)}$$

By repeating the two-step procedure, one can terminate the search when the first four formants have been found or the incremental frequency is over 5 kHz.

In summary, the N_b method samples $N_b(s)$ at specific frequency increments, checking for changes in polarity of the function. Then a linear interpolation, such as Newton's method, is used to

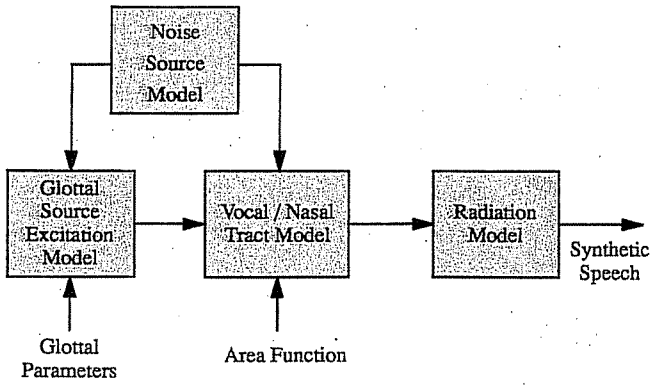


FIGURE A11.7 An acoustic model for the articulatory speech synthesizer.

obtain the frequency root. To determine the final frequencies of H(s), the derivatives N'_a and N'_b are approximated by finite differences.

A11.2.2.4 Estimate the Vocal Tract Cross-Sectional Area from the Formant Frequencies One of the functions of the articulatory model is to compute the articulatory information (in particular, the vocal tract cross-sectional area) from the acoustic information (the first four formant frequencies) that are obtained from the speech signal. In general, an optimization scheme is used to solve this speech inverse problem. The optimization scheme varies the articulatory parameters iteratively to achieve a match between the model-generated first four formants and the target (desired) first four formants. Later in this appendix we describe a simulated annealing optimization scheme for this purpose.

A11.2.3 Acoustic Models

Basically, the acoustic model of the human vocal system embodies several submodels, as shown in Figure A11.7. Both the vocal tract and nasal tract models simulate the sound propagation in these tracts. The excitation source model represents and generates the voiced excitation waveforms for the vocal tract. The turbulent air flow at a constriction for fricatives or plosives is generated by the noise source model. The radiation model simulates the acoustic energy radiating from the lips and the nostrils.

A11.2.3.1 Vocal Tract Models The vocal tract is a bent (or curved), three-dimensional acoustic tube with a slowly time-varying shape; it has soft wall vibrations, viscous friction and heat conduction losses, and varying boundaries at both the lips and glottis. There is a nasal side branch, beginning at the top of the pharynx, of fixed dimensions but variable coupling. The nasal tract is discussed later.

Research has demonstrated that the Navier–Stokes description of fluid flow has the feasibility for realistically characterizing the nonlinearities involved in voiced sound generation by the vocal cords, voiceless-fricative generation from turbulent flow at constrictions, and resonance and radiation effects conditioned by sound propagation in a nonuniform, lossy, soft wall human vocal tract (Hegerl and Höge, 1991; Iijima et al., 1992; Thomas, 1986). However, the results have been limited by the extensive computational requirements for solving the time-dependent, turbulent Navier–Stokes equations on a dense time-space grid for realistic geometric configurations. This limitation suggests the need for a simplified version of the acoustic model of the vocal tract.

For a bent (curved) vocal tract with variable cross-sectional area, the computation of its resonances (or its acoustic transfer function) is difficult. Fortunately, Sondhi (1986) has shown that the shift in the resonance frequencies below 4 kHz is in the range of 2% to 8% for typical dimensions of the vocal tract when it is straightened out. Thus, the vocal tract can be represented as a straight tube of varying cross-sectional area, as shown in Figure A11.8(*a*), with fixed shape (circular or elliptic) without a loss in accuracy. The next assumption is that sound propagation is planar along the axis of

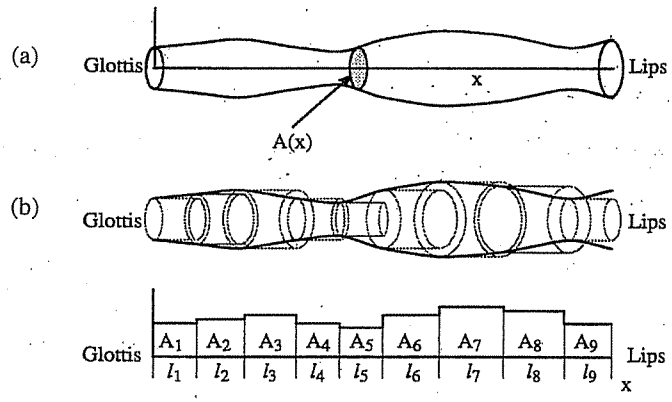


FIGURE A11.8 Vocal tract approximation. (*a*) Ideal vocal tract with variable cross-sectional area. (*b*) A step-wise approximation using concatenated acoustic tubes.

the tube. There are two reasons that make this assumption reasonable. First, the soft tissue along the vocal tract prevents radial propagation of the sound wave. Second, the average lateral (cross-sectional) dimension of the vocal tract is about 2 cm, which is much smaller than the wavelength of sound at 4 kHz, which is $\lambda = c/f = 34{,}300/4000 \cong 8.6$ cm. Strictly speaking, this assumption is valid only for frequencies below 4 kHz. But for speech, where 5 kHz is considered to be an appropriate bandwidth, the planar propagation assumption remains adequate. By neglecting the losses due to friction, heat conduction, and yielding wall vibration, a pair of equations characterizing the wave propagation in the vocal tract can be derived. In general, the solutions to such a pair of equations can only be obtained numerically. Thus, a further approximation is needed. A more tractable approach is to represent the vocal tract as a number of contiguous cylindrical sections, as depicted in Figure A11.8(*b*). If the number of concatenated sections is large, these short-length elemental sections provide a stepwise approximation of the continuous area function. We can expect that at resonant frequencies, the concatenated tubes are indistinguishable from the continuous ones. The uniform elementary cylindrical section is easy to treat. Once the lossless uniform tube has been analyzed, the effects of losses in the vocal tract can be accounted for.

A11.2.3.1.1 Sound Propagation. The linear plane wave propagation in the vocal tract is governed by the law of continuity and Newton's force law (Morse and Ingard, 1968)

$$\frac{1}{\rho c^2}\frac{\partial p(x,t)}{\partial t} = -\nabla \vec{v}(x,y,z,t) \qquad \text{(A11.2.3.1.1.1)}$$

$$\rho\frac{\partial \vec{v}(x,y,z,t)}{\partial t} = -\nabla \cdot p(x,t) \qquad \text{(A11.2.3.1.1.2)}$$

where ∇ represents the gradient, $\nabla\cdot$ is the divergence, $p(x, t)$ is the variation in sound pressure in the tube at position x and time t, $\vec{v}(x, y, z, t)$ is the particle velocity vector within the vocal tract, c is the velocity of sound, and ρ is the density of air in the tube.

If plane wave propagation is assumed, then all particles at a given displacement x and a specific time t will have the same velocity independent of location (y, z) within the cross-sectional area $A(x, t)$. Since the velocity vector points in a single direction, we can drop the vector notation in Equations (A11.2.3.1.1.1) and (A11.2.3.1.1.2). Define the volume velocity flow at position x and time t as

$$u(x,t) = A(x,t)v(x,t) \qquad \text{(A11.2.3.1.1.3)}$$

Applying the plane wave propagation assumption and substituting the volume velocity definition from Equation (A11.2.3.1.1.3) into Equations (A11.2.3.1.1.1) and (A11.2.3.1.1.2), we have the following equations (Deller et al., 1993; Rabiner and Schafer, 1978).

$$-\frac{\partial u(x,t)}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (p(x,t)A(x,t))}{\partial t} + \frac{\partial A(x,t)}{\partial t} \qquad \text{(A11.2.3.1.1.4)}$$

$$-\frac{\partial p(x,t)}{\partial x} = \rho\frac{\partial (u(x,t)/A(x,t))}{\partial t} \qquad \text{(A11.2.3.1.1.5)}$$

There is no closed form solution except for the simple configurations. If, however, the cross-sectional area A(x, t) and associated boundary conditions are specified, numerical solutions can be obtained. One method of simplifying the last pair of equations is to assume the vocal tract consists of concatenated uniform lossless sections, as depicted in Figure A11.8(*b*).

A11.2.3.1.2 Uniform Lossless Section. Assume that the vocal tract is composed of S_N uniform elemental sections and each section has cross-sectional area A_k and length l_k, where $1 \leq k \leq S_N$. This corresponds to spatial sampling, with l_k being the sampling interval for the kth section. Consider the ith section, of length l_i, with constant cross-sectional area A_i. Define $x_{i-1} = \sum_{k=1}^{i-1} \ell_k$, which represents the vocal tract length from the glottis to section i. Then the coupled differential Equations (A11.2.3.1.1.4) and (A11.2.3.1.1.5) for this elemental section become

$$-\frac{\partial u_i(x,t)}{\partial x} = \frac{A_i}{\rho c^2}\frac{\partial (p_i(x,t))}{\partial t} \tag{A11.2.3.1.2.1}$$

$$-\frac{\partial p_i(x,t)}{\partial x} = \frac{\rho}{A_i}\frac{\partial (u_i(x,t))}{\partial t} \tag{A11.2.3.1.2.2}$$

where $u_i(x, t)$ and $p_i(x, t)$ are the volume velocity and pressure, respectively, along the x axis with $x_{i-1} \leq x \leq x_i$. The solutions to Equations (A11.2.3.1.2.1) and (A11.2.3.1.2.2) have the form (Deller et al., 1993; Flanagan, 1972; Rabiner and Schafer, 1978)

$$u_i(x,t) = u_i^+\left(t - \frac{x}{c}\right) - u_i^-\left(t + \frac{x}{c}\right) \tag{A11.2.3.1.2.3}$$

$$p_i(x,t) = \frac{\rho c}{A_i}\left[u_i^+\left(t - \frac{x}{c}\right) + u_i^-\left(t + \frac{x}{c}\right)\right] \tag{A11.2.3.1.2.4}$$

where $u_i^+(t - \frac{x}{c})$ and $u_i^-(t + \frac{x}{c})$ indicate forward (transmitted) and backward (reflected) traveling waves, respectively. The boundary conditions at both ends of each section determine the relationship between the traveling waves in adjacent sections. They are derived from the physical principle that pressure and volume velocity must be continuous in both time and space everywhere within the tract.

A11.2.3.1.3 Approaches for Vocal Tract Simulation. Based on the above analysis, there are two approaches used for vocal tract simulation.

Wave propagation approach: This approach is based on the analytical solutions of Equations (A11.2.3.1.2.3) and (A11.2.3.1.2.4) for a lossless elemental uniform tube section. The pressure at any point within the section is considered to be made up of two components, a forward wave and a backward wave. At the junction of two cylindrical sections with different cross-sectional areas and lengths (see Figure A11.9), each wave has a forward propagation and backward reflection. Define

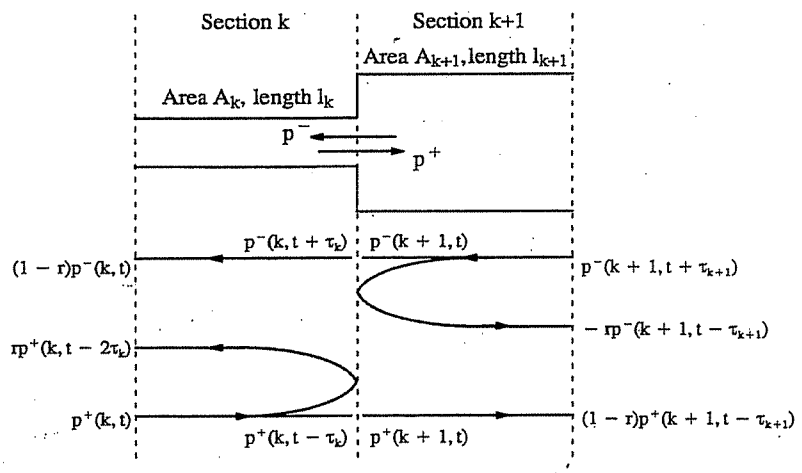


FIGURE A11.9 Reflection relationships at the junction of two lossless sections.

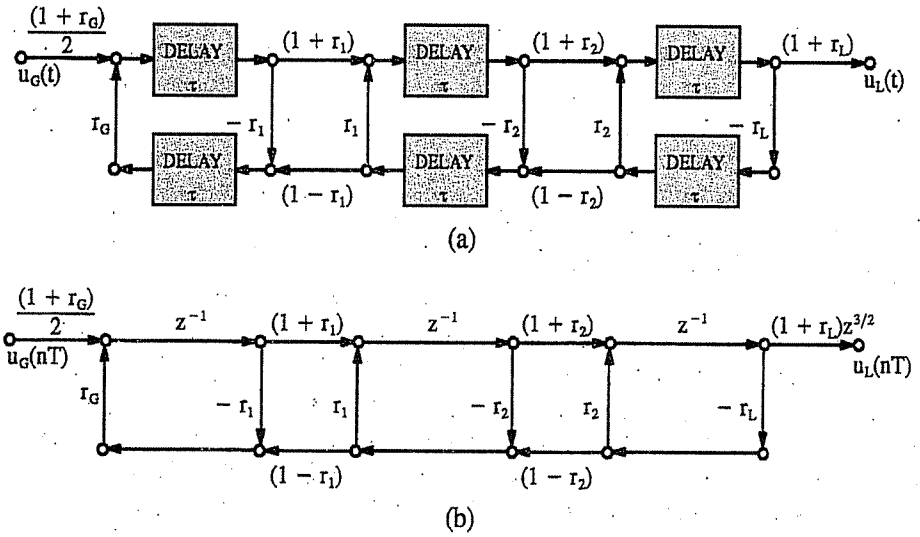


FIGURE A11.10 Discrete time lossless vocal tract. (*a*) Signal flow graph for lossless tube model. (*b*) Equivalent discrete time system.

the reflection coefficient r and the propagation delay τ, then apply the continuity conditions at each junction, and account for the losses at the glottis and lips as boundary conditions. This results in the signal flow graph and the equivalent discrete-time system, as depicted in Figure A11.10. See Rabiner and Schafer (1978) for detailed derivations.

This approach was first used by Kelly and Lochbaum (1962) for speech synthesis and has been called the Kelly–Lochbaum model or lattice structure. A more elegant realization is given by means of wave digital filters (WDF) (Fettweis and Meerkötter, 1975; Liljencrants, 1985; Meyer and Strube, 1984; Meyer et al., 1989; Meyer et al., 1991; Strube, 1982). The WDF has been implemented in special hardware for real-time synthesis (Meyer et al., 1989). Neglecting losses and using a fixed vocal tract length are the major drawbacks of this approach. By varying the sampling rate, the dynamic variation of vocal tract length can be simulated (Wright and Owens, 1993). Some progress has been made in incorporating losses (Liljencrants, 1985; Meyer et al., 1989).

Transmission-line approach: A transmission-line analog of the vocal tract (or equivalent electrical circuit model) is based on the similarity between the acoustic wave propagation in a cylindrical tube and the propagation of an electrical wave along a transmission line. The derivation from the basic equations of acoustic wave propagation to an equivalent electrical quadripole representation is well known (Fant, 1960; Flanagan, 1972; Linggard, 1985). The analogs are summarized in Chapter 5, Figure 5.1. For the circuit in the lower part of Figure 5.1, the series resistor R is used to represent the acoustic loss due to viscous drag, where the energy loss is proportional to the square of the volume velocity. The shunt conductance G represents the loss due to heat conduction, which is proportional to pressure squared. The shunt impedance Z_w is the acoustic equivalent mechanical impedance of the yielding wall. This wall impedance, which represents a mass–compliance–viscosity loss of the soft tissue, has three components, R_w, L_w, and C_w. Figure A10.27 in Appendix 10 provides some physical definitions for all the circuit components in Figure 5.1. Note that both R and G are a function of frequency. We describe the wall impedance in the next subsection.

A11.2.3.1.4 Wall Impedance. The pressure variation inside the vocal tract causes the cross-sectional area to change, since it exerts a varying force on the tract's elastic walls. Assume that the walls are locally reacting and the resulting cross-sectional area variation is small, that is,

$$A(x, t) = A_0(x, t) + \Delta A(x, t) = A_0(x, t) + y(x, t)S_0(x, t) \qquad \text{(A11.2.3.1.4.1)}$$

where $A_0(x, t)$ is the nominal area, $\Delta A(x, t)$ is a small variation, $y(x, t)$ is the yielding amount of the walls, and $S_0(x, t)$ is the circumference of the tract (Maeda, 1982a; Rabiner and Schafer, 1978). The wall vibration is modeled as a mass–compliance–viscosity mechanical model. The pressure variation

is governed by the following differential equation

$$S_0(x, t)p(x, t) = m\frac{\partial^2 y(x, t)}{\partial t^2} + b\frac{\partial y(x, t)}{\partial t} + ky(x, t) \qquad \text{(A11.2.3.1.4.2)}$$

where m, b, and k are the mass, viscosity, and compliance, respectively, of the wall per unit length of the tract (Maeda, 1982a). Define the volume velocity generated by the wall vibration as

$$u_w(x, t) = \frac{\partial(y(x, t)S_0(x, t)\ell)}{\partial t} \qquad \text{(A11.2.3.1.4.3)}$$

where ℓ is the length of tract. By substituting Equation (A11.2.3.1.4.3) into Equation (A11.2.3.1.4.2), the wall vibration can be rewritten as

$$p(x, t) = \frac{m}{S_0^2(x, t)\ell}\frac{\partial^2 u_w(x, t)}{\partial t^2} + \frac{b}{S_0^2(x, t)\ell}\frac{\partial u_w(x, t)}{\partial t} + \frac{k}{S_0^2(x, t)\ell}u_w(x, t) \qquad \text{(A11.2.3.1.4.4)}$$

For an elemental uniform section, Equation (A11.2.3.1.4.4) is simplified to

$$\begin{aligned} p(x, t) &= \frac{m}{S_0^2\ell}\frac{\partial^2 u_w(x, t)}{\partial t^2} + \frac{b}{S_0^2\ell}\frac{\partial u_w(x, t)}{\partial t} + \frac{k}{S_0^2\ell}u_w(x, t) \\ &= L_w\frac{\partial^2 u_w(x, t)}{\partial t^2} + R_w\frac{\partial u_w(x, t)}{\partial t} + \frac{1}{C_w}u_w(x, t) \end{aligned} \qquad \text{(A11.2.3.1.4.5)}$$

where the circuit components, L_w, R_w, and C_w representing the wall vibration impedance, have the apparent defintions.

The wall impedance can be either included in every elemental section of the vocal tract as a distributed element (Flanagan, 1972a; Flanagan and Ishizaka, 1976; Flanagan et al., 1975, 1980; Ishizaka et al., 1975; Maeda, 1982a) or inserted as a lumped shunt element, one in the pharynx and one at the level of the cheek (Badin and Fant, 1984; Lin, 1990; Wakita and Fant, 1978). As Wakita and Fant (1978) indicated, the lumped wall impedance, which is independent of the vocal tract configurations may not give satisfactory results. The distributed wall impedance is used in our software.

Figure A10.25 in Appendix 10 presents data concerning the wall mass, viscosity, and compliance found in the literature. In some cases, the compliance is not used since it has virtually no effect on the resonances of the model (Wakita and Fant, 1978). The data measured by Ishizaka and colleagues (1975) are used in our software. As Maeda (1982a) pointed out, the total mass of the walls may vary unrealistically if the yielding wall parameters are specified in terms of a unit surface area. Thus, the per unit length specification was used in Maeda's vocal tract simulation. We follow Maeda's specification.

A11.2.3.2 Nasal Tract and Sinus Cavities The nasal tract constitutes a side branch of the vocal tract. The velopharyngeal port controls the coupling between these two tracts for producing certain sounds. A general rule is that when the opening area is smaller than 20 mm^2 there is no apparent nasality. A wider opening produces nasal resonance, and speech is perceived as nasal when the area approaches 50 mm^2 (Borden and Harris, 1980). In our articulatory model, the opening area of the velopharyngeal port is simulated by lowering the velum along a line segment, as mentioned in Section A11.2.2.1.

The nasal tract has two channels at the nostrils. Usually, an acoustically approximated single tract is used owing to its quasi-symmetrical profile. The minor errors due to this approximation have been analyzed by Lin (1990). Figure A11.11 shows the area function of the nasal tract used by Maeda (1982b), where the nasal tract is assumed to be 11 cm long and consists of 11 elemental uniform sections. Generally, the nasal tract has a fixed structure except for the first few sections, indicated by a dashed line in Figure A11.11, where the area varies with the velopharyngeal port opening. Maeda (1982b) used linear interpolation to interpolate the areas (the second and third sections) between the coupling section (the first section) and the first fixed section (the fourth section).

From sweep frequency measurements of the acoustic transfer function (Lindqvist-Gauffin and Sundberg, 1976) and simulation studies (Fant, 1985; Lin, 1990; Maeda, 1982b) of the nasal tract, it has been found that the nasal sinuses should be considered as a part of the acoustic system. In a model of speech production, Lindqvist-Gauffin and Sundberg (1976) indicated that at least two

TABLE A11.1 Data on the Sinuses

	R_{sin} (dyne · sec/cm^5)	L_{sin} (g/cm^4)	C_{sin} (cm^4 · sec^2/g)	F (Hz)	B (Hz)
Maxillary (Lin, 1990)	1.1	3.42×10^{-3}	29.7×10^{-6}	500	51.2
Frontalis (Lin, 1990)	7.2	11.4×10^{-3}	1.13×10^{-6}	1,400	100.5
Maxillary (Sondhi and Schroeter, 1987)	1.0	5.94×10^{-3}	15.8×10^{-6}	519.5	26.8

shunting cavities, the sinus maxillares and the sinus frontales, must be added to improve the nasal quality. Since the opening area, which couples the sinus cavities to the nasal tract, is rather small, the sinus cavities can be regarded as Helmholtz resonators. According to the Lindqvist-Gauffin and Sundberg (1976) study, a reasonable estimate of the resonant frequencies would be 200 to 800 Hz for the maxillary sinuses and 500 to 2000 Hz for the frontal sinuses. The effect of the sinus resonance on the acoustic system is modeled as a shunt circuit element (Fant, 1985), as shown in Figure A11.12, and the resonance can be tuned to the required frequency. Fant (1985) inserted the sinus maxillares and frontales at positions 6 cm and 8 cm from the nostrils, respectively. The two sinuses are tuned to resonate at 500 Hz and at 1400 Hz respectively. However, Maeda (1982b) inserted only the sinus maxillares at a position 4 cm from the nostrils and showed that the quality of all nasalized vowels was satisfactory. Table A11.1 lists data for the shunt circuit components used in the literature (Lin, 1990; Sondhi and Schroeter, 1987).

To investigate the effects of the nasal tract and sinus cavities, our software system provides the user with a method to vary the nasal tract structure (area and length), to assign the number of coupled sinus cavities, and to change the circuit component values and the coupling position of each sinus.

A11.2.3.3 Radiation Models of Lips and Nostrils Acoustic energy escapes from the vocal tract via the lips. From the transmission-line analogs, the lips are treated as a radiation impedance that loads the vocal tract. The radiation impedance contains a resistive part that represents acoustic energy loss and a reactance part that represents the mass inertia of air at the lips (Fant, 1960; Stevens, 1998). Radiation from a spherical baffle is one model for the radiation impedance that is represented by nonlinear functions (Morse, 1948; Morse and Ingard, 1968). Stevens et al. (1953) made approximations and represented the radiation impedance by a resistive load and three other frequency-dependent components. Fant made another approximation and modeled the impedance by two frequency-dependent components, one being resistive and the other inductive (Fant, 1960; Wakita and Fant, 1978).

Another simplified radiation model is to assume that the radiating surface is set in a plane baffle of infinite extent. In this case, the radiation impedance is formed by a first order Bessel function and Struve function (Flanagan, 1972a; Rayleigh, 1945; Wakita and Fant, 1978). Flanagan (1972) provided a good approximation to this complicated representation by a parallel connection of a resistance and an

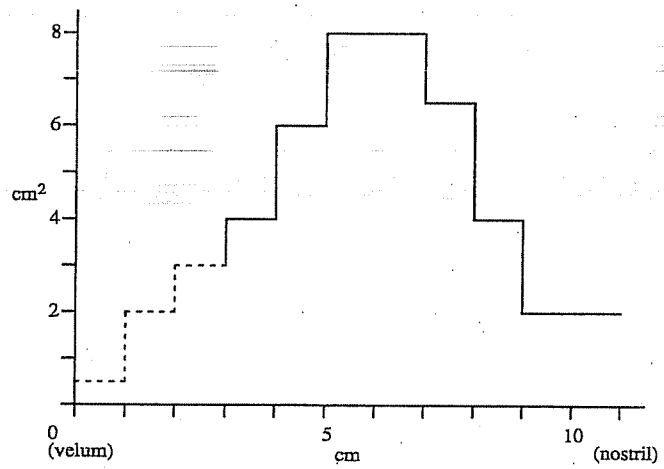


FIGURE A11.11 The area function for the nasal tract (after Maeda, 1982b).

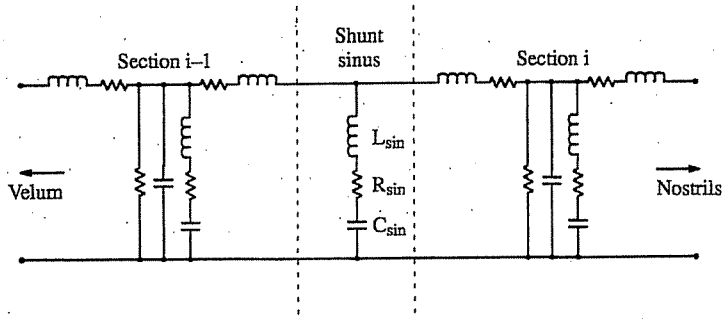


FIGURE A11.12 An extra shunt inserted between two nasal sections to model the sinus.

inductance. The most important feature of Flanagan's model (1972a) is that both circuit components are frequency independent. Figure A10.28 in Appendix 10 illustrates the Stevens et al. (1953) model and the Flanagan (1972a) model.

Comparisons between models have been made by researchers (Badin and Fant, 1984; Lin, 1990; Wakita and Fant, 1978). The Stevens et al. (1953) model yields the most accurate result. However, the Flanagan (1972a) model is usually preferred for time-domain synthesis (Flanagan and Ishizaka, 1976; Flanagan et al., 1975, 1980; Maeda, 1982a) and is used in our synthesis model. The same radiation model is used for the nostrils.

The relationship between the volume velocity at the lips or nostrils and the radiated pressure at a distance d cm from the lips or nostrils is given by (Fant, 1960)

$$\frac{p_r(\omega)}{u_r(\omega)} = \frac{\rho\omega}{4\pi d} K_T(\omega) \qquad \text{(A11.2.3.3.1)}$$

The factor $K_T(\omega)$ is a function to provide a smooth high frequency emphasis. Due to a lack of experimental verification, $K_T(\omega)$ is generally set to unity and the relationship is essentially a differentiation (Badin and Fant, 1984).

A11.2.3.4 Excitation Source Models Basically, there are two kinds of speech sounds. One is voiced, which involves quasi-periodic vibrations of the vocal folds. The other is unvoiced, which involves the generation of turbulence noise by the rapid flow of air past a narrow constriction. In the case of voiceless speech, the excitation waveform appears somewhat like a random noise source, which we will discuss in section A11.2.3.6. For voiced speech, the excitation source is a quasi-periodic pulse train located at the glottis.

A11.2.3.4.1 Excitation at the Glottis. In the case of voiced speech, the conventional LPC methods use only an impulse train as the excitation, which does not generate natural sounds (Childers and Wu, 1990). It is well-known that the "naturalness" of synthetic speech is closely related to the shape of the glottal pulse (Childers and Lee, 1991; Childers and Wu, 1990; Holmes, 1973; Klatt and Klatt, 1990; Rosenberg, 1971). We do not yet have a complete understanding of the phonatory behavior of the vocal folds. Thus, we lack an efficient model for the voice source. However, several models capable of describing the major characteristics of the glottal flow have been proposed. They can be classified into two major categories: interactive and noninteractive models (Fujisaki and Ljungqvist, 1986).

In the interactive models, there are two approaches to generate the glottal volume velocity. For the method known as the nonphysical approach, the glottal flow is calculated by modeling the glottal area (Allen and Strong, 1985; Ananthapadmanabha and Fant, 1982; Pinto et al., 1989; Titze, 1984) or conductance (Rothenberg, 1981) function and by incorporating the various impedances of the acoustic system into the model. For the method known as the physical approach, structural modeling of the mechanical vibration of the vocal cords (Flanagan and Landgraf, 1968; Ishisaka and Flanagan, 1972; Titze, 1973) and a kinematic model for the three-dimension glottis (Titze, 1989) have been attempted. The need to know the details of the physical characteristics of the various parts of the vocal cords is the major drawback of the interactive models. Furthermore, the computational burden for such models is high. We have, however, implemented one mechanical vibratory model in Chapter 9 with the theory presented in Appendix 10.

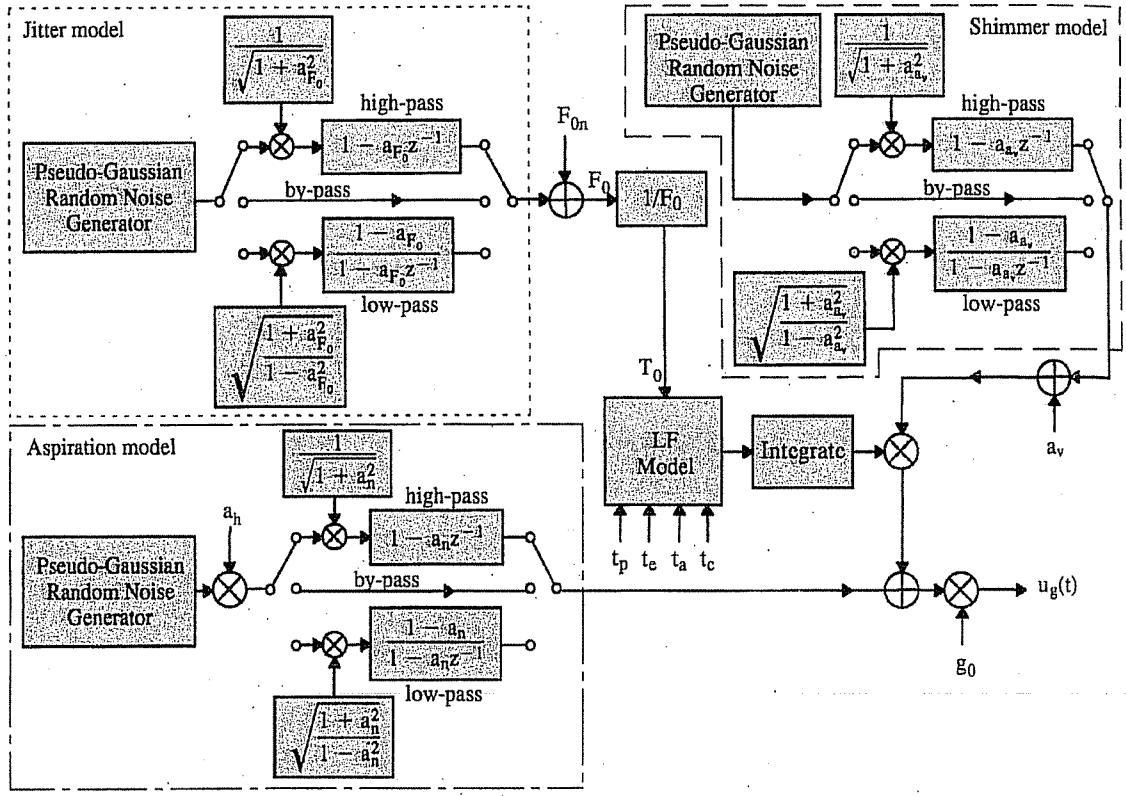


FIGURE A11.13 A simplified excitation model (Lalwani and Childers, 1991a).

In contrast, the noninteractive models directly parameterize the glottal flow or flow derivative function. If the parameters are sufficient to represent the glottal waveform, it may be possible to reconstruct the waveform from a given set of parameters. Therefore, the parameterization provides a method for generating, classifying, and storing a large number of glottal waveforms for various voicing conditions. A number of non–interactive models exist in the literature (Ananthapadmanabha, 1982; Fant, 1979; Fant et al., 1985; Fujisaki and Ljungqvist, 1986; Klatt and Klatt, 1990; Rosenberg, 1971). The Liljencrants–Fant (LF) model (Fant et al., 1985) is often used because: (1) it is preferred by listeners when they evaluate synthesized speech (Childers, 1991; Eggen, 1992); and (2) it has been shown to be superior to other models of the same complexity (Fant et al., 1985; Fujisaki and Ljungqvist, 1986). The LF model requires four parameters for modeling the differential glottal waveform (see Appendix 7). For additional details of the LF model refer to Fant (1986, 1988, 1993), Fant and Lin (1988, 1989), and Lin (1990). Another advantage of the LF model is that parameters of the model can be measured or estimated from the inverse filtered speech and the EGG signal or from the inverse filtered speech signal only (Childers and Lee, 1991; Lee, 1988).

It is well known that "jitter," the aperiodicity of the fundamental frequency of voicing (Horii, 1979), and "shimmer," the period-to-period random fluctuations in glottal-pulse amplitude (Horii, 1980), also contribute to a natural sounding voice. Klatt and Klatt (1990) included a slow quasi-random drift called "flutter" into the voicing source model to simulate jitter (pitch perturbation) but did not include a shimmer model. Lalwani and Childers (1991a) proposed a unified glottal excitation model that includes the pitch perturbation model with rate of perturbation control and the aspiration noise model with amplitude modulation into the LF model. This unified model has the capability to include the "shimmer," that is, the amplitude perturbation model. We use a simplified version of this unified model as a noninteractive glottal excitation model in our software. Figure A11.13 illustrates the block diagram of such a simplified excitation model.

A11.2.3.4.2 Excitation in the Vocal Tract. A speaker routinely phonates using the vocal folds. Unfortunately, over 1.5 million nonspeaking persons in the United States, excluding some deaf individuals (Klatt, 1987), cannot phonate using the vocal folds. Although aerodynamic-mechanical devices can aid the vocally handicapped, they produce no sound until pulmonary air is

diverted through them (Hilgers and Schouwenburg, 1990). In 1980, an electrically driven intraoral artificial larynx was invented (Lowry, 1981). This new speech prosthesis consists of a small speaker (with battery) and a resonator horn, which are joined to a dental plate and placed in the oral cavity of the subject (Myrick and Yantorno, 1993). With such a device the subject can produce intelligible speech, although the quality is still inferior. Such a device, called UltraVoice, is available through Health Concepts, Inc. (Malvern) in Pennsylvania.

Since this electrical-driven speech prosthesis must be placed in the vocal tract, it is reasonable to expect that the driving point acoustic transfer function is different from that seen by the glottis. Thus, the excitation signal must have a different waveform from the glottal pulse to produce the same speech sounds. Myrick and Yantorno (1993) presented the vocal tract frequency response when the excitation is located at the sixth section of a ten-section vocal tract by using the Kelly–Lochbaum model (1962) and the lossless transmission-line analog model (Flanagan, 1972a). C. Wu (1996) provides a detailed derivation of a new form of the discrete time acoustic equations with the excitation located within the vocal tract.

Our software system provides several advanced features for the user to investigate the properties when the excitation is located in the vocal tract. They are:

- The excitation can be located at any section of a sixty-section vocal tract with soft-wall vibration, thermal, and heat conduction losses.
- The nasal tract, with or without sinus cavities, can be coupled into the vocal tract by varying the opening area of the velopharyngeal port.
- The subglottal system can be coupled to the vocal tract when the glottis is opened.

A11.2.3.5 Glottal Impedance and Subglottal Models According to the classical formulation of the acoustic theory of speech production (Fant, 1960; Flanagan, 1972a), the voicing source is characterized as a current source. This assumes that the glottal waveform depends very little on the shape or impedance of the vocal tract. Similarly, the vocal tract is modeled by a time-invariant linear filter since the glottal impedance is assumed much higher than the vocal tract impedance. However, recent work by several researchers has shown that there does exist a certain degree of dependency of the glottal flow on the load of the vocal tract and the subglottal cavities. A number of major interaction consequences have been identified. They are:

- First formant, F_1, ripple in the source waveform: One may often observe a "hump" in the rising portion of the glottal volume velocity waveform obtained by using inverse filtering (Childers and Wu, 1990).
- Nonlinear interaction between F_0 and F_1: The pharyngeal pressure standing waves may cause a nonlinear increase in the glottal source strength whenever F_1 is near an integral multiple of F_0 (Ananthapadmanabha and Fant, 1982).
- Truncation of the F_1 damped sinusoid: The time-varying glottal impedance affects the vocal tract transfer function primarily by increasing the first-formant bandwidth, which leads to a truncation of the F_1 damped sinusoid when the glottis is open (Ananthapadmanabha and Fant, 1982).
- Pulse-skewing: The inertive loading by the subglottal and supraglottal acoustic systems results in a skewing to the right of the glottal pulse (Rothenberg, 1981).

Although speech can be synthesized by using a simple noninteractive excitation model, it seems to be essential that a glottal excitation model used in the articulatory speech synthesizer should reproduce the variations in the acoustic features of the excitation more naturally (Sondhi and Schroeter, 1987). Interactive physical models are hard to implement since one needs the physiologic characteristics of the vocal folds, and, furthermore, most models are computationally inefficient. On the other hand, interactive nonphysical models are attractive for researchers. Klatt and Klatt (1990) included the ability to change the first-formant bandwidth pitch synchronously to simulate the interaction between source and vocal tract in their formant synthesizer. For the articulatory synthesizer, a prescribed glottal area time function is usually used for source–tract interaction.

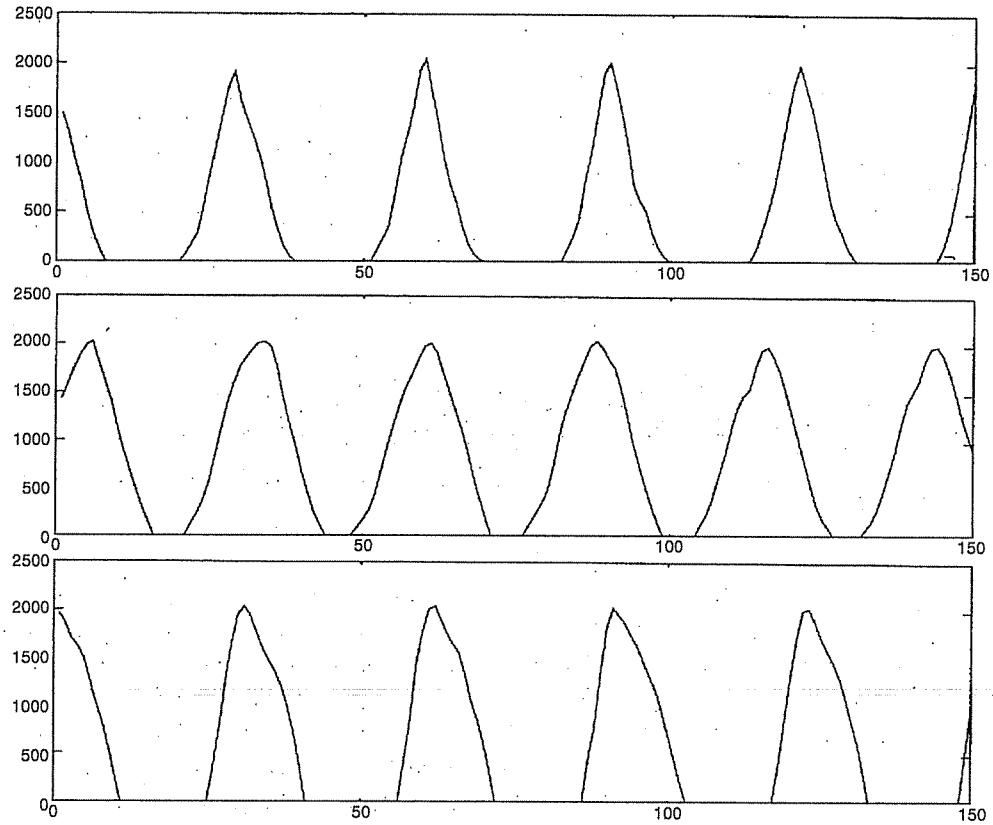


FIGURE A11.14 Three glottal area waveforms measured from ultra-high speed films.

The glottal area is defined as the opening between the vocal folds. It is time-varying during voiced phonation and quasi-steady for voiceless phonation. One possible method to obtain the time-varying glottal area is from ultra-high speed films (Childers and Krishnamurthy, 1985; Childers and Larar, 1984; Childers et al., 1984; Childers et al., 1990; Moore and Childers, 1983). The area function measured in this manner is the projected area, that is, the minimum area of the glottis. Figure A11.14 shows three measured glottal area waveforms from ultra-high speed films, where one can see that the projected glottal area tends to have a roughly triangular shape that is slightly skewed to the left. The sharp peak of the glottal area function usually results in the excitation at the apex being exaggerated (Lin, 1990). On the other hand, Cranen and Boves (1987) have derived the glottal area from a vertically uniform glottis by using the two-mass model (Ishizaka and Flanagan, 1972). The glottal area function derived in this manner is called the effective glottal area function and does not coincide with the projected glottal area. However, some simple functions such as a triangle, a sine, and a raised cosine are used to model the glottal area (Ananthapadmanabha and Fant, 1982). Figure A11.15 illustrates the glottal area waveforms modeled by triangular, sine, and raised-cosine functions. Our software program implementation provides these three functions as options for modeling the glottal area. In addition, for the triangular and raised-cosine functions, the opening and closing durations can be specified to model the glottal area skewing.

The time-varying glottal impedance is determined by the time-varying glottal area function. It contains a resistance and an inductance. Assume that the glottis is modeled as a rectangular slit with A_g, l_g, and d as the area, length, and thickness, respectively. Then the glottal inductance is given by

$$L_g = \frac{\rho d}{A_g} \qquad \text{(A11.2.3.5.1)}$$

The resistance of the glottis, according to an experiment by van den Berg and co-workers (1957), is formulated by

$$R_g = \frac{12\mu d \ell_g^2}{A_g^3} + k_g \frac{\rho u_g}{2A_g^2} \qquad \text{(A11.2.3.5.2)}$$

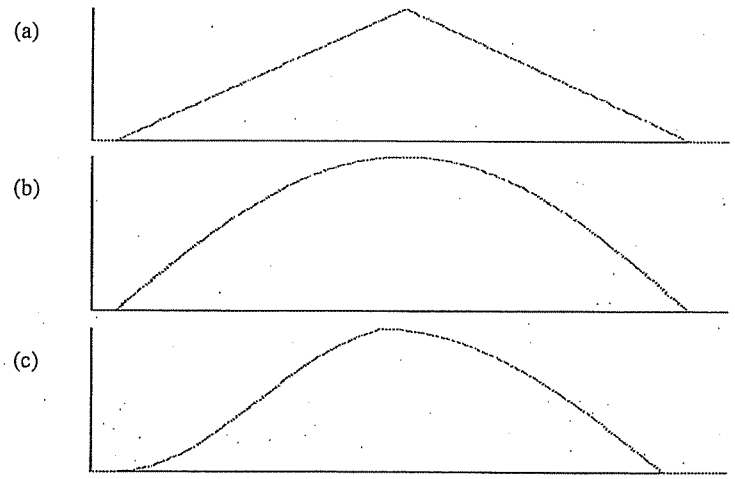


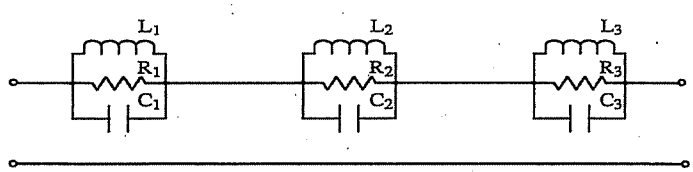
FIGURE A11.15 Glottal area models. (*a*) Triangular. (*b*) Sine. (*c*) Raised cosine.

where μ is the viscosity of air, k_g is a coefficient, ρ is the density of air in the tube, and u_g is glottal volume velocity. Some typical values of k_g used in the literature are 0.875 (van den Berg et al., 1957), 0.9 (Stevens, 1971), 1.1 (Ananthapadmanabha and Fant, 1982), and 1.38 (Maeda, 1982a).

The subglottal system, which includes the tracheal tube and lungs is usually omitted in vocal tract simulation, since its effect on speech spectra is assumed to be minor, except for unvoiced sounds, where the glottis is large (Ishizaka et al., 1976). However, when the glottal opening is large, which means the glottal impedance is no longer very high, the coupling between the subglottal system and the vocal tract is not negligible. From measurements of laryngectomized subjects, Ishizaka et al. (1976) measured the acoustic input impedance of the subglottal system. Ananthapadmanabha and Fant (1982) used Ishizaka et al. (1976) experimental data and represented the subglottal system as three cascaded RLC resonance modules called the Foster-chain circuit. Figure A11.16 shows the circuit and the corresponding component values. The subglottal formants were located at 640, 1335, and 2110 Hz, with the corresponding bandwidths of 246, 155, and 140 Hz. The effects of the Foster-chain subglottal model on the vocal tract formants and bandwidths have been analyzed (Ananthapadmanabha and Fant, 1982; Badin and Fant, 1984; Lin, 1990). The summary is that the acoustic effect of the subglottal system is small, except for unvoiced sounds, where the glottal opening is fairly large.

Combining the simplified Lalwani and Childers excitation source model (1991a), Ananthapadmanabha and Fant glottal area model (1982), and Foster-chain subglottal model, we use an interactive excitation model, shown in Figure A11.17.

A11.2.3.6 Noise Source Models When there is a flow of air through a constriction or past an obstruction, turbulence is created (Shadle, 1991; Stevens, 1971, 1993a, 1993b). The random velocity fluctuations in the flow can act as a source of sound called turbulence. Three types of consonants produced in this manner are fricatives, stops (plosives), and affricates. Fricatives are generated with the turbulent flow excitation located in the region of a constriction in the vocal tract. Plosives are produced by making a complete closure of the tract, building up pressure, and abruptly releasing it. The stop release is frequently followed by a turbulence noise excitation. Affricates are



$L_1 = 3.80$ mH	$L_2 = 0.72$ mH	$L_3 = 0.27$ mH
$R_1 = 36.7\ \Omega$	$R_2 = 53.6\ \Omega$	$R_3 = 53.9\ \Omega$
$C_1 = 17.6\ \mu$F	$C_2 = 19.2\ \mu$F	$C_3 = 21.1\ \mu$F

FIGURE A11.16 Foster-chain circuit model of the subglottal system.

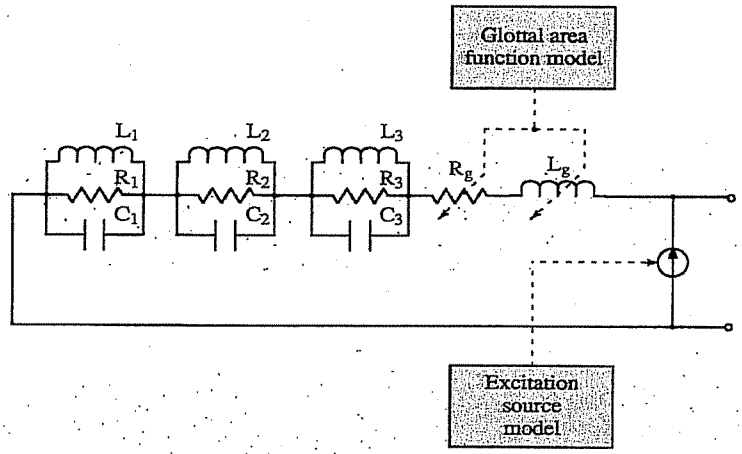


FIGURE A11.17 An interactive excitation source model used in our software.

dynamic sounds that can be modeled as the concatenation of a stop and a fricative. A special phoneme, denoted as HH or /h/, called aspirate, is produced with turbulent flow through the glottis. Refer to Broad (1977a), Borden and Harris (1980), and Stevens (1998) for more details on the generation of the unvoiced speech sounds.

Under the plane wave assumption, the sound pressure of turbulent flow can be taken as proportional to the square of the volume velocity of the airflow and inversely proportional to the constriction area, (Stevens, 1971). The turbulence noise source may be located at the center of, or immediately downstream from, the constriction region or possibly, at a combination of these places, or spatially distributed along the constriction region (Fant, 1960; Flanagan and Cherry, 1968; Lin, 1990; Stevens, 1971, 1993a, 1993b, 1998). The spectrum of the turbulence noise is broadly distributed over a wide range of frequencies (2 to 8 kHz), with some accentuation in the mid-audio range (Childers and Lee, 1991; Stevens, 1971, 1993a, 1993b, 1998).

Basically, the noise source model defines the characteristics of the noise source as a function of the airflow through the constriction and of the constriction cross-sectional area. Meyer-Eppler (1953) (Broad, 1977b) found that the rms sound pressure, P_{rms}, of the noise could be expressed as

$$P_{rms} = A_c \left(R_e^2 - R_{ec}^2\right) \qquad \text{(A11.2.3.6.1)}$$

where R_e is the Reynolds number and R_{ec} is the critical Reynolds number. Fant (1960) adopted a serial noise pressure source and reformulated the P_{rms} as a function of the pressure drop through the constriction and the effective width of the constriction. Lin (1990) extended Fant's model to include the frictional and turbulent losses inside the constriction. Both Fant and Lin tried to reconstruct the fricative spectra from area functions by using the acoustic transfer function. Some fricatives have been modeled quite successfully and some are unsatisfactory (Badin, 1989, 1991). Klatt (1980) used a random number generator, a spectrum-shaping filter, and an amplitude modulator to model the turbulent flow for the formant synthesizer. The spectrum-shaping filter was designed to simulate the spectral characteristics of the turbulent flow. A first order IIR filter was used to obtain the volume velocity due to a random pressure source. Childers and Lee (1991) have used a FIR filter to model highpass-filtered turbulence noise. Cook (1991, 1993) used a four-pole filter to model the spectral properties of the noise source.

By including a latent random pressure source, P_n, and an inherent constriction loss, R_n, in each elemental section of the vocal tract, Flanagan and Cherry (1968) could automatically introduce the turbulent flow excitation at any section. The P_n source was produced from Gaussian noise, which was bandpass-filtered from 500 to 4000 Hz, and the flow, U_n, was lowpass-filtered to 500 Hz before it modulated the P_n noise source. Figure A11.18 illustrates the schematic diagram. Such a turbulence noise model has been used in several studies (Flanagan and Ishizaka, 1976; Flanagan et al., 1975, 1980). However, as Sondhi and Schroeter (1987) pointed out, the Flanagan and Cherry (1968) model did not produce satisfactory unvoiced sounds due to the high "back" cavity impedance. Sondhi and Schroeter (1986, 1987) modified the above distributed and series pressure noise source model into

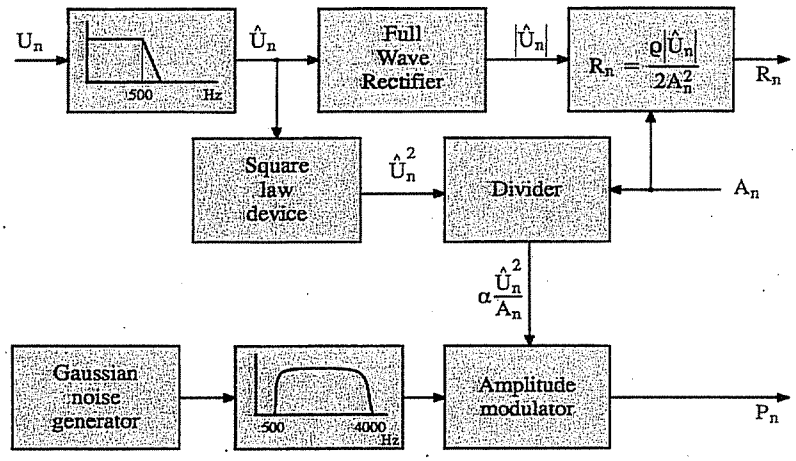


FIGURE A11.18 Schematic diagram of Flanagan and Cherry (1968) turbulence source generation model. A_n is the constriction area.

a parallel flow source $U_n = P_n/R_n$, which was located downstream from the constriction. The P_n is given by

$$\begin{aligned} P_n &= (\text{turbg})(\text{rand})\left(R_e^2 - R_{ec}^2\right), && \text{for } R_e > R_{ec} \\ &\doteq 0, && \text{for } R_e < R_{ec} \end{aligned} \qquad \text{(A11.2.3.6.2)}$$

where turbg is empirically determined as the turbulence gain, and rand is a random number uniformly distributed between -0.5 and 0.5. A first-order IIR filter with cutoff frequency 2000 Hz is used to lowpass the flow. Figure A11.19(*a*) and (*b*) shows the equivalent circuits of the serial and parallel turbulence sources, respectively.

We adopt the turbulence noise source model from Sondhi and Schroeter (1986, 1987). However, our model allows the user to place the turbulence noise source at the center of, or immediately downstream or upstream from, the constriction region, or spatially distributed along the constriction region. The turbulence gain and critical Reynolds number can also be specified.

We considered an acoustic model of the human vocal system. Next, we construct a transmission-line circuit model for the vocal system. Figure A11.20 illustrates the model structure of the vocal system for the proposed articulatory synthesizer. Based on this structure, the acoustic transfer function for different characteristics of the vocal system can be evaluated. The main purpose of this model structure is for deriving the acoustic equations for synthesizing speech. Refer to Appendices A11–B and A11–C for acoustic transfer function calculations and the derivation of the acoustic equations, respectively.

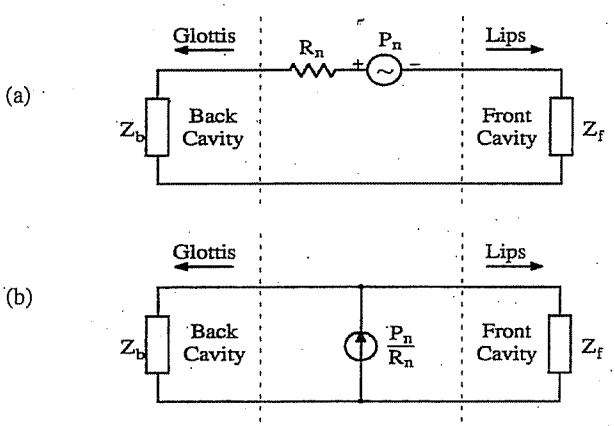


FIGURE A11.19 Equivalent circuits for the turbulence source. (*a*) Serial (after Flanagan and Cherry, 1968). (*b*) Parallel (after Sondhi and Schroeter, 1986).

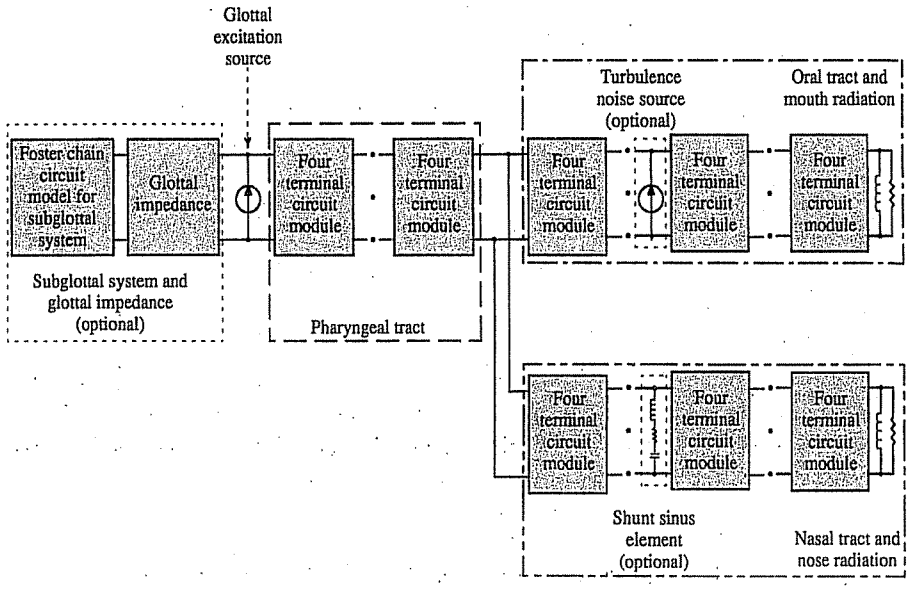


FIGURE A11.20 A model structure for the vocal system for the articulatory speech synthesizer.

A11.2.4 Analysis of Various Vocal System Characteristics

In this section, we analyze the effects of various vocal system characteristics. This analysis provides the basis for selecting the appropriate vocal system model structure and component values for the articulatory synthesizer. Five American vowels and diphthongs (/AA, IH, IY, OW, UW/) are investigated under different vocal characteristics. The vocal tract cross-sectional areas for these vowels and diphthongs are given in Appendix 5, while the methods for calculating the acoustic transfer function are given in Appendix A11-B.

A11.2.4.1 Frequency-Dependent Components As mentioned in Section A11.2.3.1.3, the series resistance R and the shunt conductance G in the equivalent circuit representation of a lossy section are frequency dependent. In the time-domain approach for the articulatory synthesizer, the frequency-dependent components have to be simulated at a fixed frequency. The effects of using a fixed frequency for these two components on the acoustic transfer function are not well documented. Wakita and Fant (1978) illustrated the effects on formant frequencies and bandwidths for the five Russian vowels given in Appendix 5, when the frequency was fixed at 1 kHz. They determined that the formant frequencies are scarcely affected, but that the bandwidths are affected rather appreciably. We have investigated the acoustic transfer functions for five vowels for three fixed frequencies: 1 kHz, 2.5 kHz, and 4 kHz. The acoustic transfer functions were calculated with the glottis closed and no nasal tract coupling. The formants were not affected appreciably, which agrees with Wakita and Fant's (1978) result. For formant frequencies below 1.5 kHz, the formant bandwidths for the fixed frequency cases were wider than the frequency-dependent case. For formant frequencies above 1.5 kHz, the 1 kHz case had the narrowest formant bandwidths, which is narrower than the frequency-dependent case. As a trade-off, we have set the frequency at 2.5 kHz for the frequency-dependent components.

A11.2.4.2 Number of Vocal Tract Sections As described in Section A11.2.3.1.2, the vocal tract can be approximated by a concatenation of uniform elemental sections. The number of elemental sections, S_N, should be large enough so that the acoustic characteristics of the concatenated tubes is indistinguishable from a continuous tube. In this section, we examine the influence of spatial sampling on the acoustic transfer function of the vocal tract. Figure A11.21 shows the acoustic transfer functions of different numbers of vocal tract sections for five vowels. The acoustic transfer functions are calculated with the glottis closed and no nasal tract coupling. It can be seen that the formant frequencies generally shift upward as the spatial sampling interval increases. The curves for 60 sections (solid line) are nearly identical to the results obtained for the last section with the frequency set at

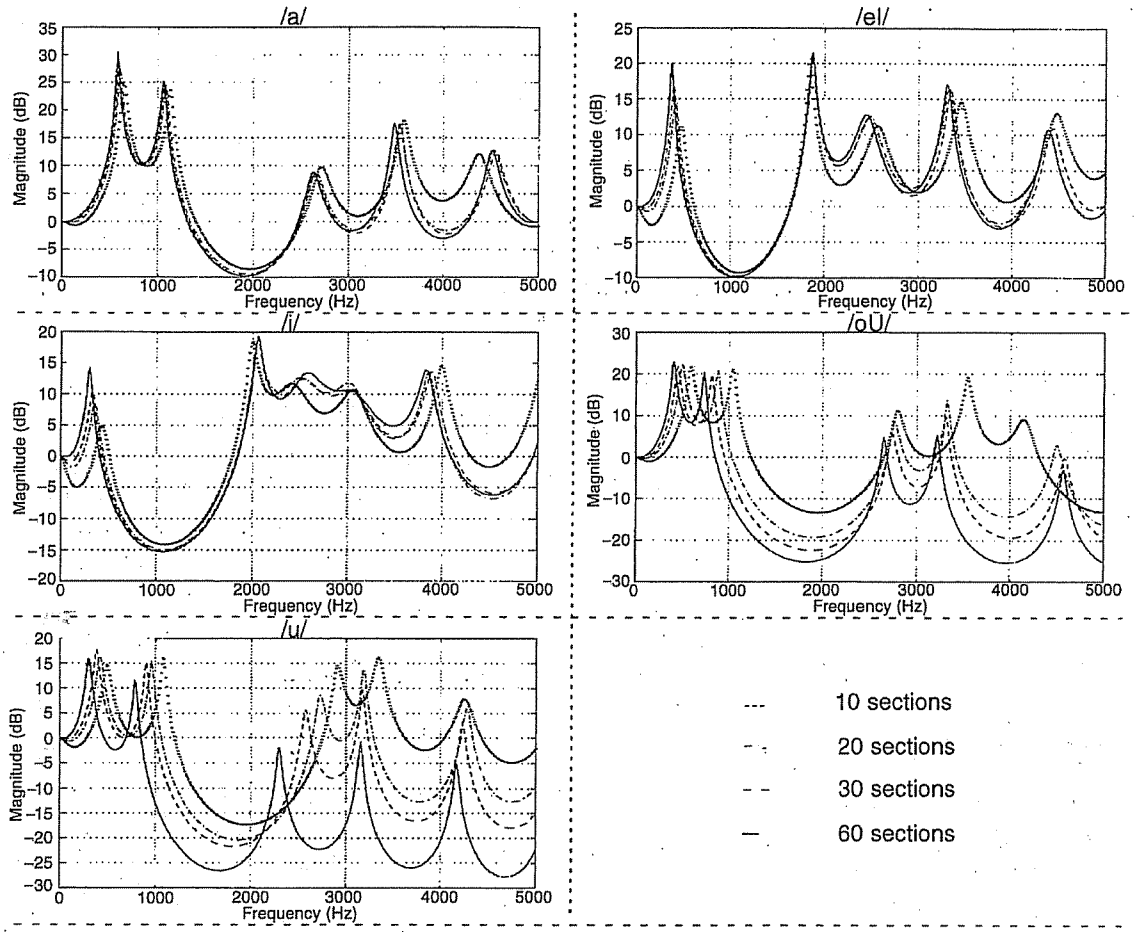


FIGURE A11.21 The effect of the spatial sampling interval on the acoustic transfer function for five vowels.

2.5 kHz for the frequency-dependent components. The exception is for the second and third formants for the vowel /IY/, where the formants shift downward. The vowels have been labeled using the IPA symbols. The corresponding upper case ARPAbet symbols are as follows: /a/ = /AA/ (as in Bach), /eI/ = /IH/ (as in bit), /i/ = /IY/ (as in beet), /oU/ = /OW/ (as in boat), /u/ = /UW/ (as in boot). From Figure A11.21, we can see that the spatial sampling interval, that is, the number of elemental sections, has a more significant effect on the acoustic transfer functions of vowels /u/ = /UW/ and /oU/ = /OW/ than others. However, a ten-section cross-sectional area function is not sufficient to represent the acoustic characteristics of the vocal tract.

A11.2.4.3 Nasal Tract System To study the acoustic properties of the nasal tract, the acoustic transfer function of the nasal tract for different opening areas of the velopharyngeal port are calculated. Figure A11.22 shows the resonant characteristics of the nasal tract for various velopharyngeal port opening areas. Basically, the nasal cavity has three resonant frequencies. The first resonance is not affected by the velopharyngeal port opening area. However, the second and third resonances shift downward as the velopharyngeal port opening area increases.

The effect of the extra sinus cavities on the acoustic transfer function of the nasal tract can be examined from Figure A11.23. The maxillary sinus and the frontal sinus are located at 4 cm and 8 cm, respectively, from the nostrils. The two sinus cavities are tuned to 500 Hz and 1400 Hz, respectively. The frontal sinus has a limited effect on the nasal tract acoustic transfer function. It has the effect of a zero-pole pair in the vicinity of its resonance frequency. For this reason, Maeda (1982b) ignored this sinus without losing the essentials of the nasal tract. The first resonance of the sinus nasal tract is shifted downward with a lower peak level as a result of the maxillary sinus coupling. The maxillary sinus also brings about a pole-zero pair in the vicinity of its resonance frequency.

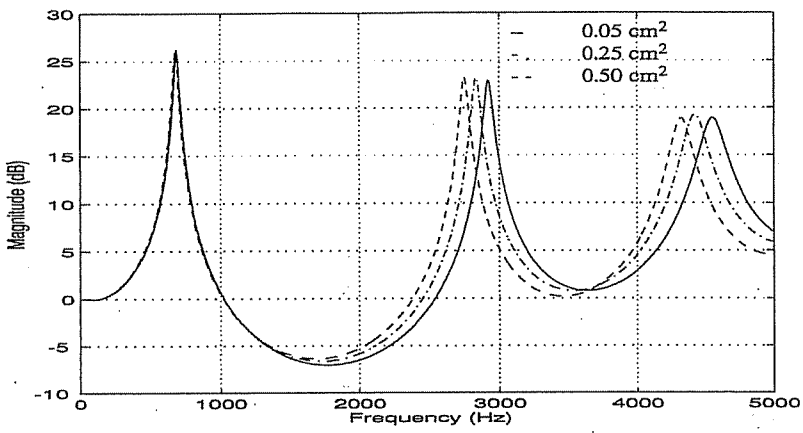


FIGURE A11.22 The effect of the velopharyngeal port opening on the nasal tract acoustic transfer function.

Lowering the velum creates a side passage for the air flow through the nasal cavity, giving rise to complex modifications of the acoustic characteristics of the sound. Figure A11.24 shows how this mechanism affects the acoustic transfer function of the vocal system. The velopharyngeal opening area is 0.5 cm^2. For the nasal tract with the sinus cavity, only the maxillary sinus is included and is tuned to 500 Hz. It is well known that the parallel branching of the nasal tract at the velum causes antiresonances in the vocal tract acoustic transfer function. The antiresonances are at the vicinities of 1 kHz and 3 kHz and can be seen clearly for some vowels. The effect of an extra sinus on the acoustic transfer functions for vowels /a/ = /AA/ and /eI/ = /IH/ is more significant than for vowels /i/ = /IY/, /u/ = /UW/, and /oU/ = /OW/. This result supports Maeda's statement (1982b) that the high vowels, such as /IY/ and /UW/, are more nasalized than the middle and low vowels, such as /AH/ and /AA/, even when the nasal sinus is not included.

A11.2.4.4 Glottal Impedance and Subglottal System The influence of the glottal impedance and the subglottal system on the acoustic transfer function of the vocal system can be examined from Figure A11.25. When the glottal area is small; that is, the glottal impedance is relatively high, the influence is insignificant. For a large glottis, the increased loading of the vocal tract causes an increase in the bandwidths and to some extent in the formant frequencies. It is obvious that the influence of the subglottal system depends on the glottal impedance. When the glottal area is small, the influence of the subglottal resonances is small, and vice versa.

A11.2.4.5 Excitation in the Vocal Tract Placing the excitation source in the vocal tract results in a very complicated system, as mentioned in Section A11.2.3.4.2. This section

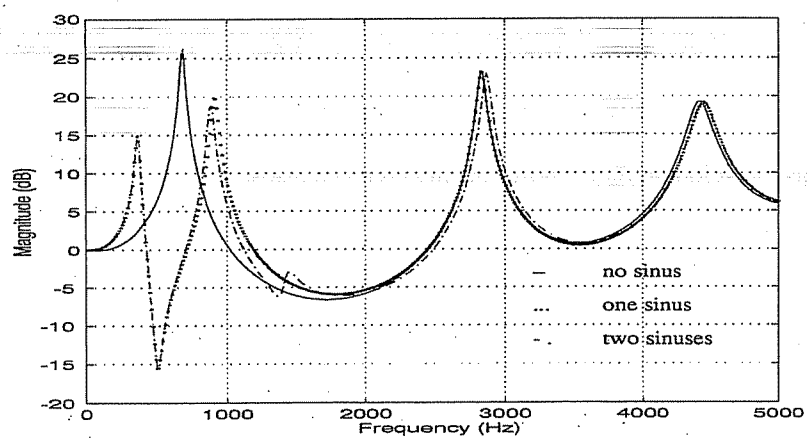


FIGURE A11.23 The effect of extra sinus cavities on the nasal tract acoustic transfer function.

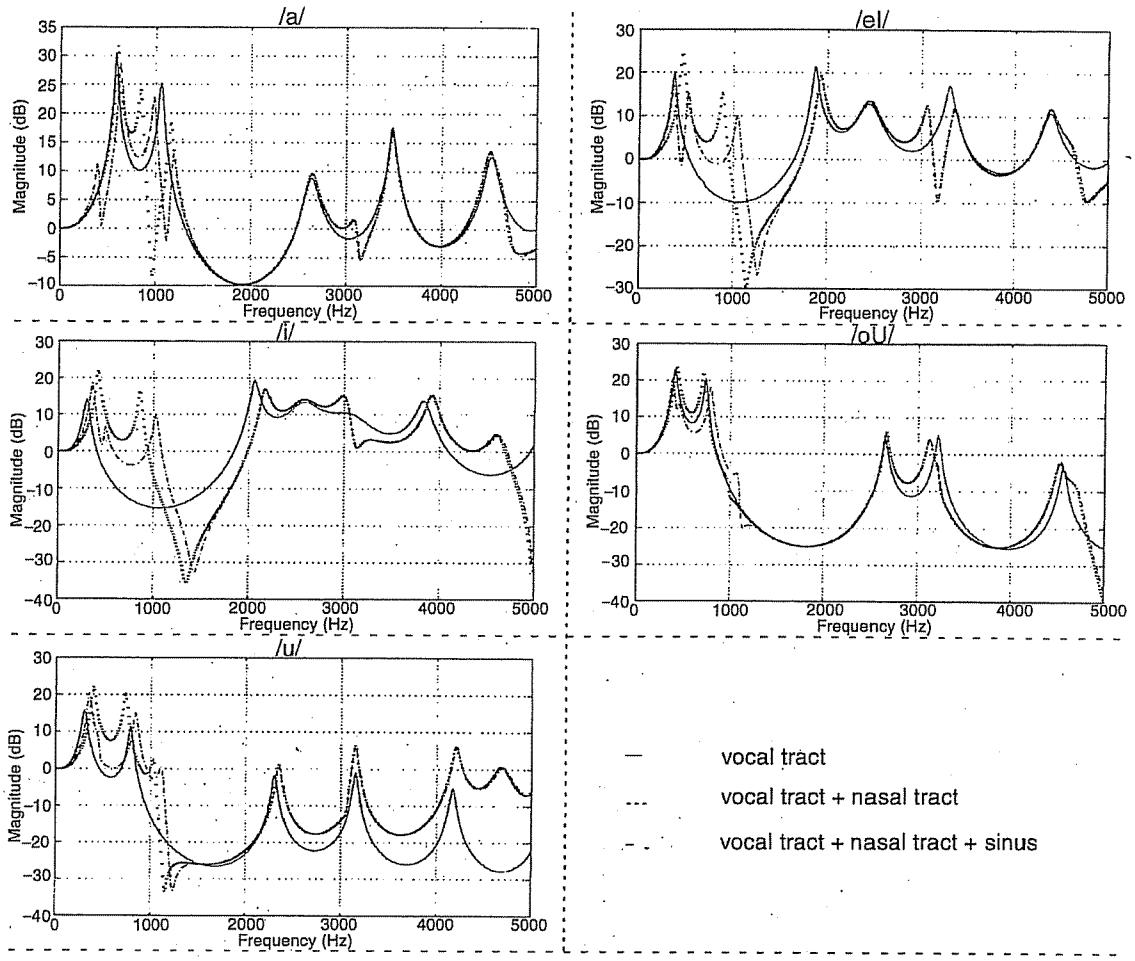


FIGURE A11.24 The effect of the nasal tract and an extra sinus on the acoustic transfer function for five vowels.

examines the effects of relocating the excitation on the acoustic transfer function (see Figure A11.26). It is found that the vocal tract resonant frequencies are relatively unaffected when the excitation is placed forward from the pharynx to the front oral cavity, even though this introduces antiresonances into the acoustic transfer function. Another feature is that the number of antiresonances increases when the excitation is placed forward from the pharynx to the front oral cavity. The excitation source waveform can be obtained by deconvolving the speech signal from the modified acoustic transfer function. It is expected that such an excitation waveform differs from the glottal waveform, which generally contains no zeros in its transfer function. Different vowels have different antiresonances, due to different vocal tract shapes. This makes the modeling of the excitation waveform inside the vocal tract difficult. One possible way to generate the excitation waveform inside the vocal tract is to prefilter the glottal pulse with the inverse filter of the modified acoustic transfer function.

A11.2.5 Articulatory Speech Synthesis

Basically, there are three approaches used in articulatory speech synthesis. The wave digital filter approach (Fettweis and Meerkötter, 1975; Lawson and Mirzai, 1990) extends the Kelly–Lochbaum model (1962). This approach is based on forward and backward traveling waves in a lossless acoustic tube (Meyer et al., 1989; Rubin et al., 1981; Strube, 1982; Titze, 1973) and can produce real-time synthesis (Meyer et al., 1989). It usually omits many acoustic effects, such as proper handling of existing losses in the tract, realistic modeling of the glottis, and appropriate modeling of source–tract interaction. Also the vocal tract length cannot be varied easily since the length of each section is fixed and related to the sampling frequency (Wakita, 1973). Recently, some progress has been made in

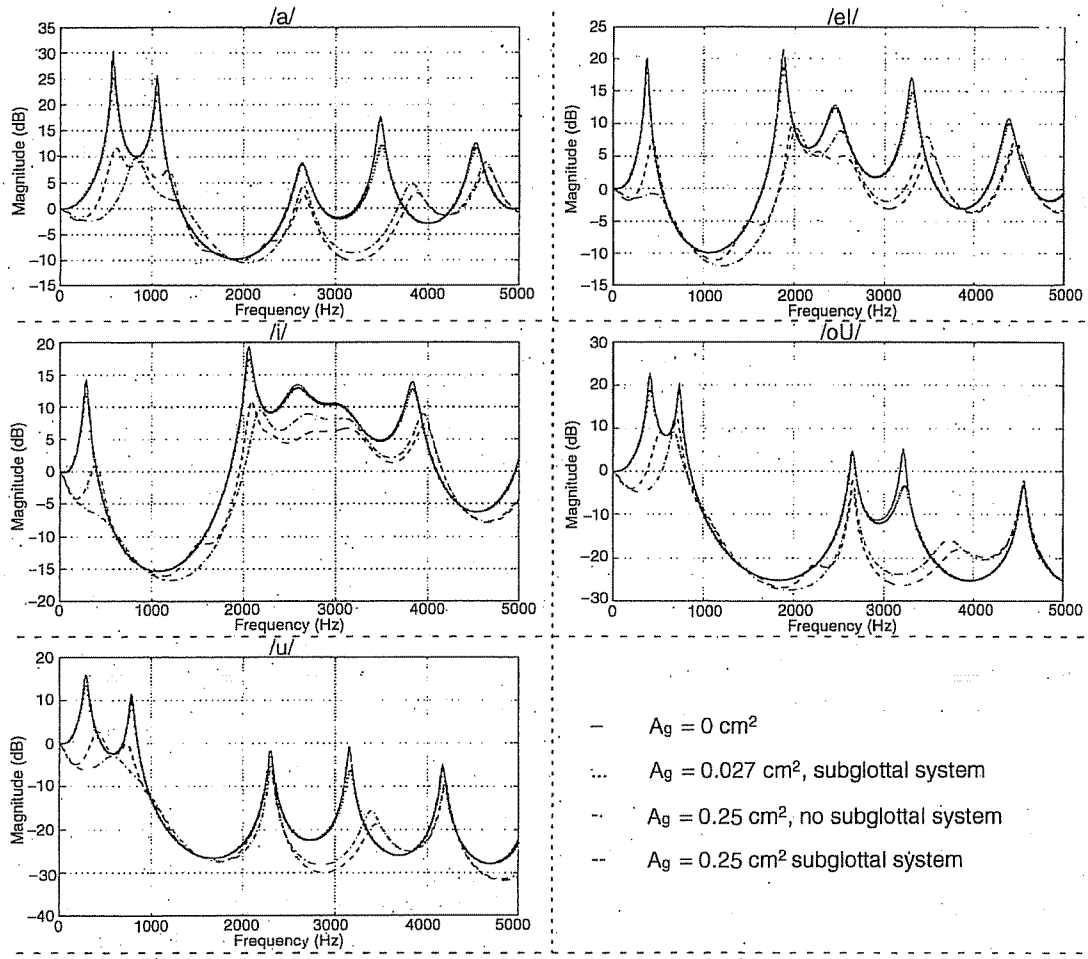


FIGURE A11.25 The effect of the glottal impedance and the subglottal system on the acoustic transfer function for five vowels.

modeling voiceless excitation, damping, and the glottal excitation (Meyer et al., 1989). The dynamic variation of vocal tract length can be simulated by varying the sampling rate (Wright and Owens, 1993).

The second approach uses a hybrid time–frequency domain method, which models the highly nonlinear glottal characteristics in the time domain and the linear tract with frequency-dependent losses and wall vibration characteristics in the frequency domain (Allen and Strong, 1985; Sondhi and Schroeter, 1986, 1987). The tract filter function and glottal source excitation function are interfaced by an inverse Fourier transformation and digital convolution. The problems with this approach are that it is incapable of producing the dynamic transitions of certain phonemes; for example, plosives, and it needs additional care to cope with the interaction between voiced and voiceless sources (Lin, 1990). In addition, it does not calculate the pressure and volume velocity.

The third approach is to model the human vocal system as a large set of linear or nonlinear difference equations to be solved in each sampling interval to give samples of the pressure and volume velocity at each point in the transmission-line circuit (Flanagan and Cherry, 1968; Flanagan and Ishizaka, 1976; Flanagan and Landgraf, 1968; Flanagan et al., 1975, 1980). The values of pressure and volume velocity at one time instant are used to determine the losses for the next time interval. This approach has been referred to as the time-domain approach (Sondhi and Schroeter, 1987). Figure A11.27 shows the schematic diagram of this approach.

In the time-domain approach, a very high sampling rate is usually required to avoid frequency-warping distortion (Wakita and Fant, 1978). In addition, the frequency-dependent components are simulated at a fixed frequency (see Section A11.2.4.1). Natural-sounding speech, however, can be generated. Several advantages have made the time-domain approach popular, although its computation

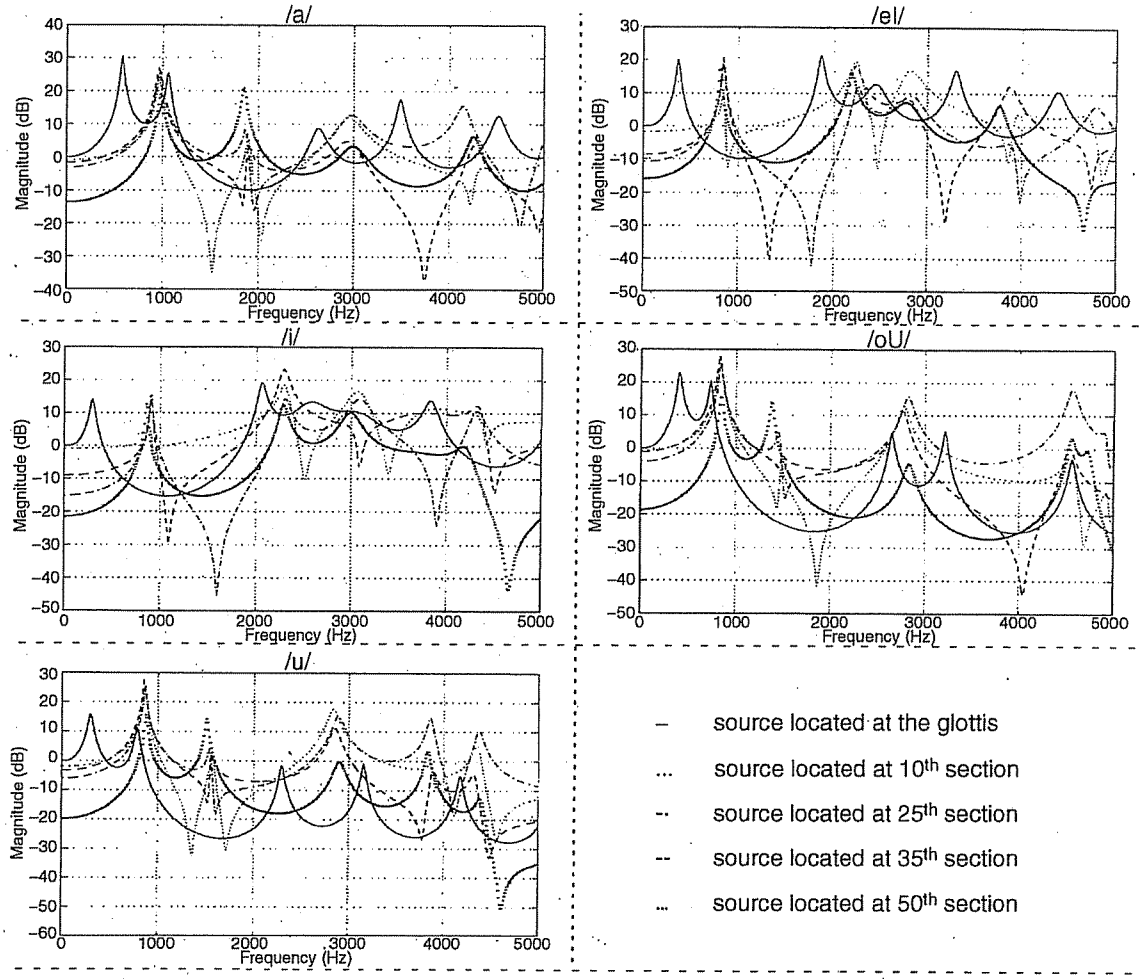


FIGURE A11.26 The acoustic transfer function for the five vowels when the source is located at different positions within the vocal tract.

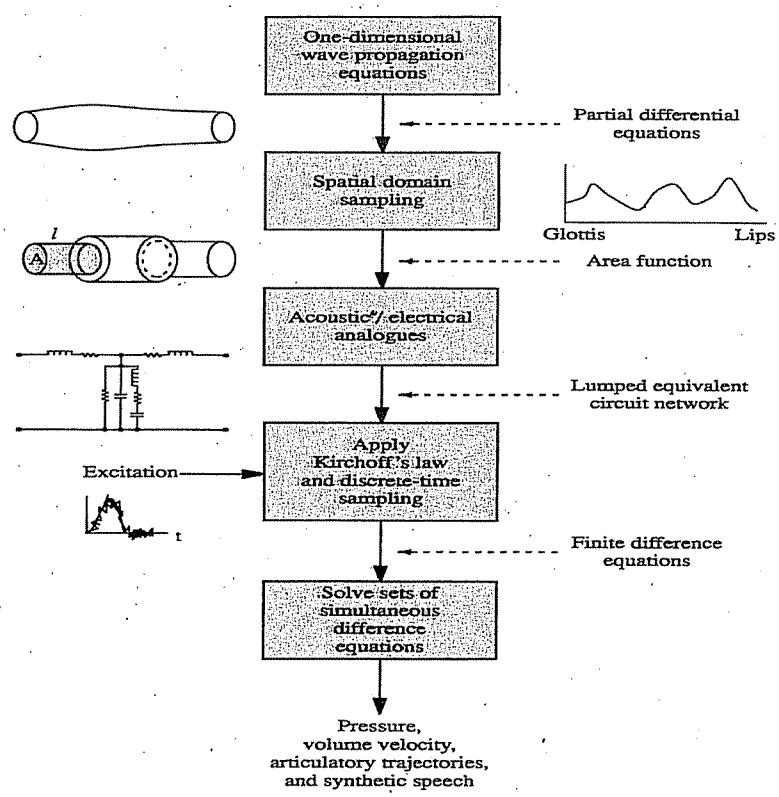


FIGURE A11.27 Time-Domain approach for articulatory speech synthesis.

is lengthy. These advantages are that the aerodynamic interaction is inherently included, the pressure and volume velocity at any point can be computed, and the dynamic articulatory gestures can be obtained when combined with the articulatory model. In our software, the time-domain approach is used for realization of the acoustic model.

Maeda (1982a) simplified Flanagan's model by replacing the mechanical vibratory model of the vocal cords with glottal area control parameters, discarding noise sources within the vocal tract, and omitting the effects of the nasal sinuses. These simplifications made the synthesis procedure much faster. Bocchieri (1983) and Bocchieri and Childers (1984) introduced other simplifications, such as reducing the number of noise sources and modeling the vocal tract variation by drawing a sequence of midsagittal vocal tract outlines on a graphic terminal. Based on Maeda's (1982a) work, Childers and Ding (1991) implemented an articulatory speech synthesizer by using a discrete circuit model that converts the acoustic equations into linear algebraic equations.

We rederived the acoustic equations (see Appendix A11–C) of the vocal system to include the subglottal system, the glottal impedance, the turbulence noise source, and the sinus cavities. Table A11.2 compares several articulatory synthesizers with our implementation (Figure A11.20).

A11.2.5.1 Realization Figure A11.28 shows a software block diagram of the model constructed for time-domain articulatory synthesis. Two options are provided for the interpolation of the vocal tract configuration: vocal tract cross-sectional area and articulatory parameters. If the articulatory parameters are interpolated, the articulatory model is used to transform the parameters to the vocal tract cross-sectional area. The number of vocal tract sections is 60. The vocal tract cross-sectional area is transformed to the equivalent RLC-network. On the other hand, the excitation parameters are interpolated and the excitation waveform is generated, according to the interpolated parameters, as the source input to the circuit network. The nasal sinus cavities and/or the subglottal system can be included optionally in the circuit network. By applying Kirchoff's and Ohm's laws and the trapezoidal algorithm, the discrete-time acoustic matrix equations are formed (see Appendix A11–C for the details). The pressure at the midpoint of each section and volume velocity at the junction of adjacent sections are calculated using the elimination procedure and a backward substitution. The

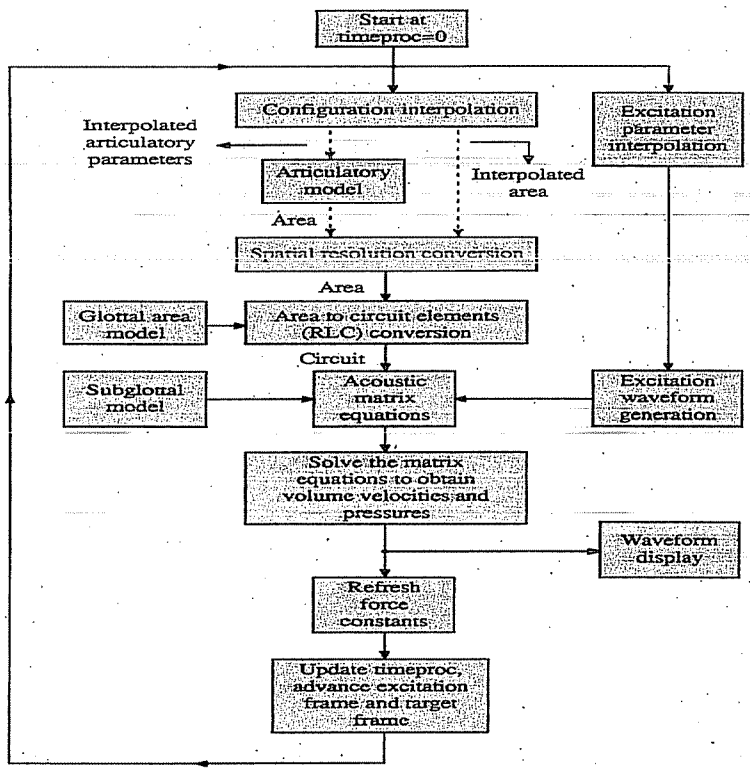


FIGURE A11.28 Block diagram of the articulatory speech synthesizer implemented in our software.

TABLE A11.2 Comparison of Several Articulatory Speech Synthesizers

	Flanagan	Maeda	Bocchieri and Childers	Childers and Ding	Prado	Proposed Synthesizer
Model of excitation	Self-oscillating two-mass	A slit	Self-oscillating two-mass	Passive one-mass	Parametric two-mass	LF
Jitter and shimmer models included	No	No	No	No	No	Yes
Noise source at the glottis	No	No	Yes	Yes	Yes	Yes
Noise source in the vocal tract	Every section	No	At constriction	At constriction	At constriction	Center of, or downstream or upstream from, or distributed along the constriction
Excitation in the vocal tract	No	No	No	No	No	Yes
Method for changing model parameters	Recompile source code	Recompile source code	Recompile source code	Through parameter files	Through parameter files	On screen by moving slider

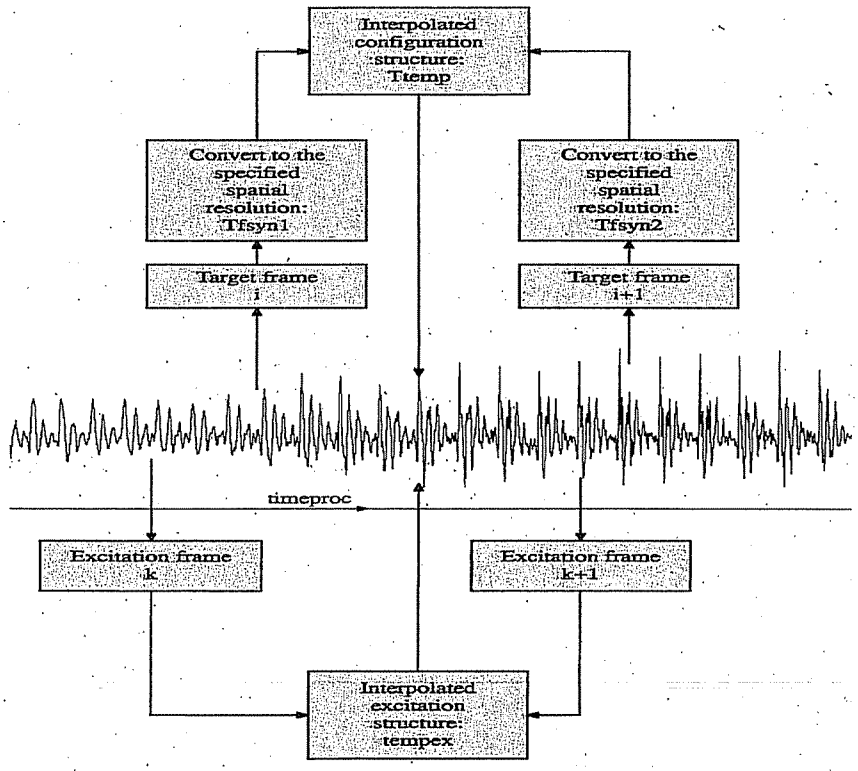


FIGURE A11.29 The timing sequence and interpolation of target frames and excitation frames.

synthetic speech is the backward difference between the sum of the volume velocities at the nostrils and lips at the current time instant and the sum of the volume velocities at the nostrils and lips at the previous time instant. The synthesis procedure is repeated by refreshing the force constants (see Appendix A11–C), updating the time instant, and advancing the target and/or excitation frames until the time epoch is reached.

A sketch of the configuration and excitation parameter interpolations is illustrated in Figure A11.29. Only two target frames and two excitation frames are shown in this figure. Either the vocal tract cross-sectional areas or the articulatory parameters are interpolated between the current target frame i and the next target frame i + 1 during the synthesis of speech. Assume that the target frame structure is pointed to by pointers Tfsyn1 and Tfsyn2, respectively. A temporary pointer, Ttemp, is used to point to the interpolated structure configuration. Similarly, a temporary pointer, tempex, is used to point to the interpolated excitation structure. Both data structures are used to generate speech.

A11.2.5.2 Interpolation Function Only one interpolation function, namely linear, is provided to interpolate the vocal tract configuration: (1) vocal tract cross-sectional area; or (2) articulatory parameters. For any one articulatory parameter or any one vocal tract cross-sectional area, linear interpolation can be described through the following function.

$$y = \alpha + \beta t \qquad \text{(A11.2.5.2.1)}$$

where y is the interpolated articulatory parameter or vocal tract cross-sectional area, t is time, and α and β are interpolation parameters. The two interpolation parameters are determined by the two target frames, i and i + 1.

A11.3 SPEECH INVERSE FILTERING

The recovery of articulatory movements from the speech signal, sometimes referred to as the speech inverse filtering problem, is difficult due to the nonuniqueness of the solution. This problem has been

the subject of research for several applications, including articulatory synthesis, speech recognition, low-bit-rate speech coding, and text-to-speech synthesis. Here, we offer a new solution using the simulated annealing algorithm, which is a "constrained multidimensional nonlinear optimization problem." The coordinates of the jaw, tongue body, tongue tip, lips, velum, and hyoid compose the multidimensional articulatory vector. A comparison between the model-derived and the target-frame first four formant frequencies forms the cost function. There are two constraints: (1) the articulatory-to-acoustic transformation function; and (2) the boundary conditions for the articulatory parameters. The optimum articulatory vector is obtained by finding the minimum cost function. Once the optimum articulatory vector is determined, the articulatory model determines the vocal tract cross-sectional area function, which in turn is used by the articulatory speech synthesizer.

A11.3.1 Review of the Derivations of the Vocal Tract Area Function

Geometric data concerning the vocal tract is essential to understanding articulation, and is a key factor in speech production. The acoustical theory of speech production (Fant, 1960) views the vocal tract as an acoustical tube with a varying cross-sectional area. The success of articulatory modeling depends to a large extent on the accuracy with which the vocal tract cross-sectional area function, A(x), can be specified for a particular utterance. Measurement of the vocal tract geometry is difficult. Basically, there are two methods for obtaining the vocal tract cross-sectional area function: (1) direct measurements from images such as x-rays; and (2) estimating the area function from acoustic data.

A11.3.1.1 Direct Measurements Direct measurements of the vocal tract have been made from lateral x-ray images (e.g., Chiba and Kajiyama, 1941; Fant, 1960; Johansson et al., 1983). Unfortunately, these direct measurements and their evaluations are laborious. In addition, the exposure to x-ray for utterances of long durations is a problem owing to dosage limitations. Magnetic resonance imaging (MRI) (Baer et al., 1991), which is free from the disadvantages associated with x-ray methods, might appear to be the best available method to collect the necessary data. The drawback, however, is that the subject may fatigue since the imaging process requires a long time. Additional drawbacks stem from the fact that the resolution of air–tissue boundaries may depend on the thickness of the tissue section, and the calcified structures contain little mobile hydrogen and, thus, may be indistinguishable from the airway.

A11.3.1.2 Estimation from Acoustic Data Several researchers have proposed analytical methods to derive the vocal tract cross-sectional area function, A(x), from acoustic data. Two approaches are based on LPC and the tube impulse response, respectively. The LPC approach is based on the fact that the filtering process of the lossless nonuniform acoustic tube model of the vocal tract is identical to that of the optimal inverse filter model with proper boundary conditions at the glottis and the lips (Atal and Hanauer, 1971; Wakita, 1973, 1979; Wakita and Gray, 1975; Gobl, 1988). The reflection coefficients are extracted by inverse filtering the speech signal. Then, the vocal tract cross-sectional area function is obtained from the set of reflection coefficients. The main problem with this approach is the articulatory compensation or the "ventriloquist effect;" that is, the fact that different vocal tract shapes can produce the same formant frequencies (Atal et al., 1978; Bonder, 1983; Charpentier, 1984; Mermelstein, 1967; Schroeder, 1967; Schroeder and Strube, 1979). For the tube impulse response approach, the basic concept is that if the transfer function of the vocal tract is known, then the A(x) can be derived uniquely (Gopinath and Sondhi, 1970; Milenkovic, 1984, 1987; Paige and Zue, 1970; Schroeder, 1967; Sondhi, 1979; Sondhi and Resnick, 1983). However, finding the transfer function of the vocal tract involving the use of impedance tubes with externally generated excitation does not allow the subject to phonate sounds.

To finesse some of the difficulties of the analytical methods, the "sorting" and "codebook" methods perform sampling of the articulatory parameters from the articulatory model and establish tables of vocal tract shapes and related acoustical representations. For the sorting method, reference tables are established by covering the articulatory space with a uniform or nonuniform grid and storing

the acoustic values computed at every vertex of the grid (Atal et al., 1978; Charpentier, 1984; Cook, 1991). These tables can be used to look up the effective vocal-tract geometric representations that have similar acoustic features. Some refinements, such as singular value decomposition and local region linearization (Atal et al., 1978), have been used to solve the ambiguous geometric subspace. On the other hand, the codebook method samples the articulatory space randomly and prunes it to retain only the reasonable shapes in the codebook. This method provides the basis for the vector quantization of the articulatory space (Larar et al., 1988). In 1990, Schroeter et al. (1990) made some improvements for the generation of codebooks by using a dynamic programming search. The codebooks are accessed through evaluating a weighted cepstral distortion measure as given by Meyer et al. (1991). There are several drawbacks with this numerical approach: cumbersome computations, sensitivity to the source excitation, mapping ambiguities, and acoustic modeling limitations (Schroeter and Sondhi, 1994).

A recent approach is to apply an artificial neural network (ANN) model to the speech inverse filtering, since it is a promising approach to implement codebooks (Shirai, 1993). The ANN model is trained with a large set of acoustic parameter patterns. Then, a test pattern of acoustic parameters is used to search the codebook to retrieve a corresponding articulatory pattern parameter set (Båvegård and Högberg, 1992, 1993; Papcun et al., 1992; Rahim et al., 1993; Xue et al., 1990). However, the learning of a large set of training patterns to span the articulatory space is still a challenge for the ANN model (Xue et al., 1990). In addition, as Schroeter and Sondhi (1994) pointed out, no clear advantage has so far been shown for ANN compared to other approaches.

The feedback methods try to optimize the articulatory parameters that are adjusted until the synthetic speech features differ minimally from the actual speech features. The selected speech features can be formants (Prado, 1991; Prado et al., 1992), spectral (Flanagan et al., 1980; Guo and Milenkovic, 1993; Gupta and Schroeter, 1993; Levinson and Schmidt, 1983; Parthasarathy and Coker, 1990, 1992; Riegelberger and Krishnamurthy, 1993; Heike, 1979), or others. The optimization can be done on a phoneme-by-phoneme basis (Parthasarathy and Coker, 1990, 1992) or on a frame-by-frame basis (Flanagan et al., 1980; Gupta and Schroeter, 1993; Levinson and Schmidt, 1983; Prado, 1991; Prado et al., 1992). Several search algorithms have been used, such as the Hooke and Jeeves algorithm (Flanagan et al., 1980; Gupta and Schroeter, 1993; Parthasarathy and Coker, 1990, 1992), the optimal gradient algorithm (Levinson and Schmidt, 1983), and combinations of the modified Fletcher–Reeves method and linear successive approximation (Prado, 1991; Prado et al., 1992). The problem of local minima related to the nonlinearity in the speech inverse filtering is a major impediment of this method.

Advances in computer technology have allowed the solution of optimization problems that require large numbers of complicated function evaluations to be computed on relatively inexpensive machines in a reasonable time. Thus, stochastic methods, such as genetic algorithms that serve as search procedures based on the mechanics of natural selection and natural genetics (Goldberg, 1989), can be applied to the speech inverse filtering problem. Some preliminary results have been obtained by McGowan (1994). Articulatory trajectories of an articulatory model were recovered by means of a genetic algorithm from the first three formant frequencies using a task-dynamic model (Saltzman, 1986; Saltzman and Kelso, 1987; Saltzman and Munhall, 1989) of speech articulation. Tests on synthesized utterances show that the method can recover the major aspects of an original trajectory, but it has trouble in obtaining the precise timing of events. An additional difficulty for the genetic algorithm, as Goffe et al. (1994) experienced, stems from the fact that it needs further development to become more usable for continuous function problems, since it has difficulty with a relatively flat surface.

In general, finding the global minimum value of a cost function with many degrees of freedom is difficult, since the cost function tends to have many local minima. A procedure for solving such optimization problems should sample values of the cost function in a manner that has a high probability of finding a near-optimal solution and should also lend itself to efficient implementation. Over the past few years, simulated annealing has emerged as a viable technique that meets these criteria. Simulated annealing is based on models found in nature; for example, some processes in thermodynamics (Kirkpatrick et al., 1983; Metropolis et al., 1953) can be modeled using a stochastic optimization method. Simulated annealing explores a function's entire surface and tries to optimize the function while moving uphill and downhill. Thus, this technique is largely independent of the starting values, which is often a critical factor in conventional optimization algorithms. Simulated annealing also makes less stringent assumptions regarding the function than do conventional algorithms. For example, the function need not be continuous since the method does not require the calculation of derivatives.

Because of these relaxed assumptions, it can more easily deal with functions that have ridges and plateaus. In addition, it can be applied to optimize a "black box" system for which one only needs to define the state (the parameter space) and to compute the corresponding energy (cost function value). Finally, functions that are not defined for some parameter values can also be optimized by the simulated annealing method (Bohachevsky et al., 1986; Corana et al., 1987; Goffe et al., 1992, 1994; Vanderbilt and Louie, 1984).

Based on the above reviews and discussions, we selected the simulated annealing algorithm for optimizing the nonlinear acoustic-to-articulatory transformation; that is, speech inverse filtering.

A11.3.2 Simulated Annealing Algorithms

Simulated annealing was first derived from statistical mechanics, where the thermodynamic properties of a large system in thermal equilibrium at a given temperature were studied (Metropolis et al., 1953). A description of the physical annealing process inspired this algorithm. In this situation, a solid metal is to be melted at a high temperature. After slow cooling (annealing), the molten metal arrives at a low energy state, since careful cooling brings the material to a highly ordered, crystalline state. Inherent random fluctuations in energy allow the annealing system to escape local energy minima to achieve the global minimum. However, if the material is cooled very quickly (or quenched), it might not escape local energy minima and when fully cooled it may contain more energy than annealed metal. Simulated annealing attempts to minimize an analogue of energy in an annealing process to find the global minimum. Kirkpatrick et al. (1983) were the first to propose and demonstrate the application of simulated annealing techniques to problems of combinatorial optimization, specifically to the problems of wire routing and component placement in VLSI design. Both Vanderbilt and Louie (1984) and Bohachevsky et al. (1986) modified simulated annealing for continuous variable problems. However, the Corana et al. (1987) implementation of simulated annealing for continuous variable problems appears to offer the best combination of ease of use and robustness, so it is used for our optimization process.

A11.3.2.1 Origin of the Algorithm As far back to 1953, Metropolis et al. (1953) proposed a method for computing the equilibrium distribution of a set of particles in a "heat bath" using a computer simulation method. For the system in thermal equilibrium at a given temperature T, they assumed that the probability $\pi_T(c)$ that the system is in a given configuration c depends upon the energy E(c) of the configuration and follows the Boltzmann distribution.

$$\pi_T(c) = \frac{e^{-\frac{E(c)}{kT}}}{\sum_{s \in C} e^{-\frac{E(s)}{kT}}} \qquad \text{(A11.3.2.1.1)}$$

where k is Boltzmann's constant and C is the set of all possible configurations. The configuration of the system is identified with the set of spatial positions of the particles. A stochastic relaxation technique was developed to simulate the behavior of the system. Suppose that the system is in configuration C_t at time t. A candidate configuration C_n for the system at time $t+1$ is generated randomly. The criterion for selecting or rejecting configuration C_n as a new configuration (state) depends on the difference of energies between configuration C_n and configuration C_t. Define p, the ratio of the probability of being in C_n to the probability of being in C_t, as

$$p = \frac{\pi_T(C_n)}{\pi_T(C_t)} = e^{-\frac{E(C_n)-E(C_t)}{kT}} \qquad \text{(A11.3.2.1.2)}$$

Then, apply a criterion, which has come to be known as the Metropolis criterion or algorithm, to decide the acceptance of C_n. The Metropolis criterion can be stated as follows. If $p > 1$, that is, the energy of C_n is strictly less than the energy of C_t, then configuration C_n is automatically accepted as the new configuration for time $t+1$. If $p \leq 1$, that is, the energy of C_n is greater than or equal to that of C_t, then configuration C_n is accepted as the new configuration with probability p. So a move to a state of higher energy is accepted in a limited way. By repeating this process for a large enough number of moves, that is, as $t \to \infty$, regardless of the starting configuration, it can be shown that the

distribution of configurations generated converges to the Boltzmann distribution (Geman and Geman, 1984).

A11.3.2.2 The Cooling Schedule A fundamental question arises in statistical mechanics concerning the system in the limit as it approaches a low temperature, for example, whether cooling produces crystalline or glassy solids in a metallurgic process. To achieve ground state (a low-energy crystalline configuration), simply lowering the temperature is not sufficient. Rather, a cooling schedule must be followed, where the temperature of the system is elevated, and then gradually lowered, spending enough time at each temperature to guarantee that thermodynamic equilibrium has been reached. If insufficient time is spent at each temperature, especially at a lower temperature, then the probability of achieving a low-energy crystalline state is greatly reduced.

The application of the annealing process to optimization problems involves several steps. First, one must identify the analogues of the physical concepts in the optimization problem. The energy function becomes the cost function. The configuration of particles becomes the combination of independent variable values. The rearrangement of particles becomes the iterative improvement of function values by changing variable values. Finding a low-energy configuration is a near-optimal solution, and the temperature becomes the control parameter for the process. Second, one must have a way of generating the candidate states. Usually, states are generated with a probability density function g(x) that has a gaussian-like peak. Third, one must have a way of selecting the new state. A state acceptance probability allows occasional hill-climbing as well as descents. The acceptance probability is based on the chances of obtaining a new state relative to a previous state. Two acceptance probability equations have been used successfully, the Boltzmann machine and the Metropolis algorithm, which are given as follows:

Boltzman machine

$$p(\Delta E) = \frac{1}{\left(1 + e^{-\frac{\Delta E}{T}}\right)}, \quad \text{for all } \Delta E \tag{A11.3.2.2.1}$$

Metropolis algorithm

$$\begin{aligned} p(\Delta E) &= 1.0, && \text{for } \Delta E \leq 0 \\ &= e^{-\frac{\Delta E}{T}}, && \text{for } \Delta E > 0 \end{aligned} \tag{A11.3.2.2.2}$$

where ΔE = E(new state) − E(current state) is the energy gap between the new state and the current state. The Boltzmann machine is known to better approximate the physical metaphor, but is more computationally expensive (Davis and Ritter, 1987). The Metropolis algorithm is used in our implementation. Fourth, one must specify a cooling schedule consisting of:

- An initial value of the control parameter; that is, the initial artificial temperature T.
- A decrement function for decreasing the value of the control parameter; that is, the cooling rate.
- A final value of the control parameter or a stop criterion.
- A finite number of moves for each downward control parameter value, that is, the amount of time spent at each temperature.

Such an analogy was first suggested by Kirkpatrick et al. (1983). They linked the algorithm with combinatorial optimization, specifically to the problems of wire routing and component placement in VLSI design. The rapid increase in inexpensive computing power has lead to several applications of the simulated annealing algorithm, including computer and circuit design (Vecchi and Kirkpatrick, 1983), image restoration and segmentation (Carnevali et al., 1985; Geman and Geman, 1984), the traveling salesman problem (Bonomi and Lutton, 1984), artificial intelligence (Hinton and Sejnowski, 1983), digital filter design (Benvenuto, et al., 1992; Pitas, 1994), and vector quantization (Rose et al., 1992). Because of the success of the simulated annealing in combinatorial optimization problems, its potential has been investigated for solving continuous function minimization problems. Vanderbilt and Louie

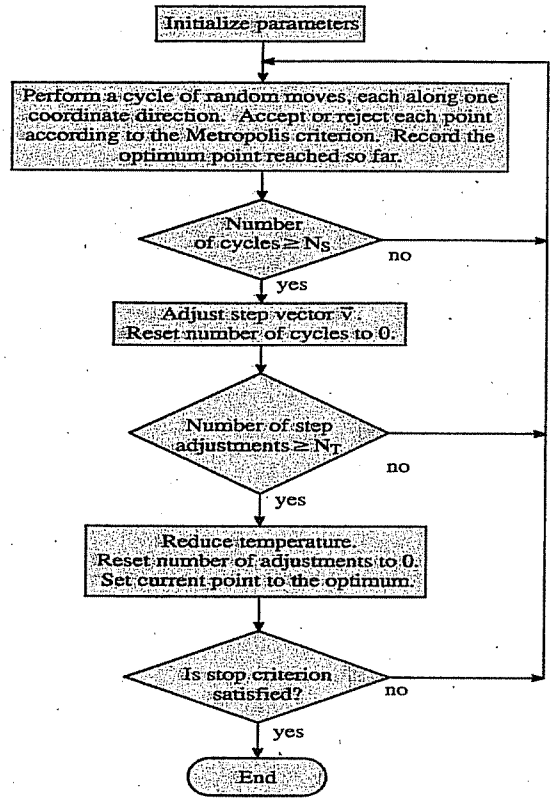


FIGURE A11.30 The simulated annealing algorithm (after Corana et al., 1987).

(1984) first modified simulated annealing by using a covariance matrix for controlling the transition probability. Bohachevsky et al. (1986) presented a simple method that was easy to implement, wherein the length of a generation step is constant. However, the Corana et al. (1987) implementation of simulated annealing for continuous variable problems appears to offer the best combination of ease of use and robustness, and has been used in econometric problems (Goffe et al., 1992, 1994).

A11.3.2.3 The Simulated Annealing Algorithm The Corana et al. (1987) algorithm is schematically shown in Figure A11.30. While a detailed description of the algorithm can be found there, we briefly describe it as follows. Let $\vec{x}$ be an M-dimensional vector with components $[x_1, x_2, \ldots, x_M]$. Let $\varepsilon(\vec{x})$ be the cost function and $lb_j \leq x_j \leq ub_j$, $j = 1, \ldots, M$; be the M variables with corresponding boundaries. The algorithm proceeds iteratively as follows. First, a cost-function evaluation is made at the initial point $\vec{x}$ and its value ε is recorded. Next, a new candidate point, $\vec{x}_n$, is generated by varying element i of $\vec{x}$, namely

$$x_{ni} = x_i + r(v_i) \qquad \text{(A11.3.2.3.1)}$$

The variable r is a uniformly distributed random number from $[-1, 1]$ and v_i is element i of $\vec{v}$, the step length vector of $\vec{x}$. The new function value ε_n is then computed. If ε_n is less than ε, $\vec{x}_n$ is accepted, $\vec{x}$ is set to $\vec{x}_n$, ε is set to ε_n, and the search path moves downhill. If this is the smallest ε at this point, it and $\vec{x}$ are recorded, since this is the best current value. If ε_n is greater than or equal to ε, the Metropolis criterion (Metropolis et al., 1953) determines acceptance. Compute the value p as follows.

$$p = e^{\frac{\varepsilon - \varepsilon_n}{T}} \qquad \text{(A11.3.2.3.2)}$$

In this equation, T represents the current temperature. Generate a uniformly distributed random number from [0, 1]. Decide the action based on the result of value comparison between p and p_u. If p_u is less than p, the new point is accepted, $\vec{x}$ is updated with $\vec{x}_n$, and the search path moves uphill. Otherwise, $\vec{x}_n$ is rejected; that is, no move is made. Thus, the process repeats from the new point (candidate point is accepted) or from the current position (candidate point is rejected). From Equation

(A11.3.2.3.2), we can see that the probability of an uphill move decreases when the temperature is lower and the difference in the function's value is larger.

After N_S steps through all elements of $\vec{x}$, the step length vector $\vec{v}$ is adjusted so that 50% of all moves are accepted. The goal is to make the algorithm follow the cost function (Corana et al., 1987). A greater percentage of accepted points means that the candidate points are too close to the current point. Thus, the step length vector $\vec{v}$ is enlarged. For a given temperature, this step adjustment increases the number of rejections and decreases the percentage of acceptances. On the other hand, a higher percentage of rejected points means that the candidate points are too far from the current point. A reduced step length decreases the rejection rate.

After N_T times through the above loops (corresponding to thermal equilibrium), the temperature, T, is reduced. The temperature is updated according to the following equation.

$$T_n = (r_T)T \tag{A11.3.2.3.3}$$

The reduction coefficient r_T has a value between 0 and 1. The starting point at the new temperature T_n is the optimum point obtained at the last temperature T. This makes the search path start at the most favorable point. Since the temperature characterizes the degree of "excitation" of the system, a lower temperature decreases the number of uphill moves, so the number of rejections increases and the step size reduces. The lower temperature (smaller step size) makes the search space shrink and focus on the most promising area, that is, concentrates most of the search in a smaller subset of low energy points.

The terminating criterion checks if there have been no significant moves for the last N_ε temperatures. Assume that the optimum value obtained at temperature T_k is ε_k^*. Let ε_{opt} be the current optimum value at the temperature T_{k+1}. If

$$\begin{aligned} \left|\varepsilon_k^* - \varepsilon_{k-m}^*\right| &\le \eta, \quad \text{where } m = 1, \ldots, N_\varepsilon \\ \left|\varepsilon_k^* - \varepsilon_{opt}\right| &\le \eta, \end{aligned} \tag{A11.3.2.3.4}$$

then stop the search. Note that η is a specified small constant. This check makes sure that the global or near-global minimum is reached. Another stop criterion is that the total number of cost-function evaluations exceeds a specified constant N_{tot}.

In summary, the simulated annealing algorithm starts at some high temperature specified by the user. A sequence of points is then generated until an equilibrium is approached. During this random walk process, the step length vector is periodically adjusted to better follow the cost function behavior. After thermal equilibrium, the temperature is reduced and a new sequence of moves is made starting from the current optimum point, until thermal equilibrium is reached again, and so forth. The process is terminated at a low temperature such that no more useful moves can be made, according to the stopping criterion.

A11.3.3 Speech Inverse Filtering Strategy and Procedure

In general, the relationship between the shape of the vocal tract and its acoustic output can be represented by a multidimensional function of a multidimensional argument

$$\vec{y} = f(\vec{x}) \tag{A11.3.3.1}$$

where $\vec{x}$ is a vector formed by the coordinates of the articulators, $\vec{y}$ is a vector formed by the corresponding acoustic features, and f is the function relating these vectors. Given an acoustic measurement $\vec{y}_d$, the problem is to find an articulatory state $\vec{x}_0$ such that $f(\vec{x}_0)$ is the best match to $\vec{y}_d$. In other words, with the optimization approach, $\vec{x}_0$ is the solution to the nonlinear optimization problem

$$\vec{x}_0 = \text{minimal arguement of } \|f(\vec{x}) - \vec{y}_d\| \tag{A11.3.3.2}$$

where $\|\cdot\|$ is a norm on the acoustic space.

A11.3.3.1 Strategy Speech inverse filtering is a "constrained multidimensional nonlinear optimization problem." As we defined in Section A11.2.2.1, the coordinates of the tongue body

(tbodyx, tbodyy), tongue tip (tipx, tipy), lips (lipp, lipo), jaw (jaw), and hyoid (hyoid) compose the multidimensional articulatory vector; that is

$$\vec{x} = [\text{thodyx, tbodyy, tipx, tipy, lipp, lipo, jaw, hyoid}] \qquad \text{(A11.3.3.1.1)}$$

Note that $\vec{x}$ is an eight-dimensional vector. Usually, the velum is set at different default positions for nasal, non-nasal, or nasalized phonemes, but it can be optimized for some phonemes. The dimensions of the lower pharynx are also allowed to be optimized whenever this is necessary.

We designate the articulatory vector as

$$\vec{x} = [x_1, x_2, \ldots, x_M] \qquad \text{(A11.3.3.1.2)}$$

where the value of M represents the number of dimensions of the articulatory domain to be optimized. As mentioned in the previous paragraph, M has a value of 8. For nasal and nasalized sounds, we may include the velum as an additional articulatory parameter; that is, M is set to 9. For middle vowels, some back vowels, and semivowels, three more parameters, which are anterior–posterior movements of K and H, g1k, and wh, and the height between K and H, hkl (refer to Figure A11.4), are included; that is, M is set to 11. Finally, one more parameter, the velum, can be included, that is, M = 12.

The acoustic vector is composed of the first four formant frequencies; that is, $\vec{y} = [F_1, F_2, F_3, F_4]$. The cost function (error distance) is derived from a comparison of the first four formant frequencies of the articulatory model and the first four formant frequencies determined from speech analysis (the target formants). A percentage of the weighted least-absolute-value (l_1-norm) error distance is defined as

$$\varepsilon(\vec{F}_m(\vec{x})) = \sum_{i=1}^{4} \frac{W_i |F_{mi}(\vec{x}) - F_{ti}|}{F_{ti}} \% \qquad \text{(A11.3.3.1.3)}$$

where F_{mi} is the ith model-derived formant, which is a function of the articulatory vector, F_{ti} is the ith target-frame formant estimated from the analysis of speech signal, and W_i is the assigned weight.

The constraints, which include the articulatory-to-acoustic transformation function f (Equation A11.3.3.1) and the boundary conditions of the articulatory parameters, are described as follows

$$\vec{y}_m = f(\vec{x}) = f([x_1, x_2, \ldots, x_M]) = \vec{F}_m(\vec{x}) = [F_{m1}(\vec{x}), F_{m2}(\vec{x}), F_{m3}(\vec{x}), F_{m4}(\vec{x})] \qquad \text{(A11.3.3.1.4)}$$

where $lb_j \leq x_j \leq ub_j, j = 1, \ldots, M$ are the lower and upper bounds of the articulatory parameters, and the subscript m represents the model-derived parameter values.

The object of the optimization process is to find the optimal articulatory vector that generates the acoustic vector (model-derived) as close to the desired (target-frame) as possible. The ideal minimum value of $\varepsilon(\vec{F}_m(\vec{x}))$ is 0%, but some approximations used in the articulatory model (see Section A11.2.2) make this value hard to obtain. The first approximation is related to the articulatory model. A non-robust representation of the lower part of the pharynx and the tongue tip-to-jaw region may cause some deviations of the midsagittal vocal tract outline. The second, and more significant deviation, is the uncertainty of the sagittal distance to cross-sectional area transformations. Different empirical transformation formulas can be found in the literature (Heinz and Stevens, 1964; Mermelstein, 1973; Sundberg et al., 1987). The final approximation is the area to formant frequency conversion. We have determined that an error criterion requiring the final value of error distance function to be less than 1% is adequate.

A11.3.3.2 Procedure To extract the articulatory trajectories from a speech sentence, the first step is to obtain a smoothed formant trajectory from the speech signal. A feature is available in our software for doing this. Then N target frames are selected. The target frame selection is based on the results of the speech analysis, which includes the formant trajectory, the location of the word endpoints, and the estimated phoneme boundaries of the speech signal. An example of selecting (marking) selected target frames of the formant tracks obtained from a speech file is shown in Chapter 10. Next, the speech inverse filtering procedure is applied to each target frame to obtain the optimum articulatory parameters.

Figure A11.31 shows the block diagram of the speech inverse filtering procedure, which is performed frame-by-frame. For each target frame, an initial value of the error distance function (cost function) is evaluated from the initial articulatory vector. The evaluation of the error distance function

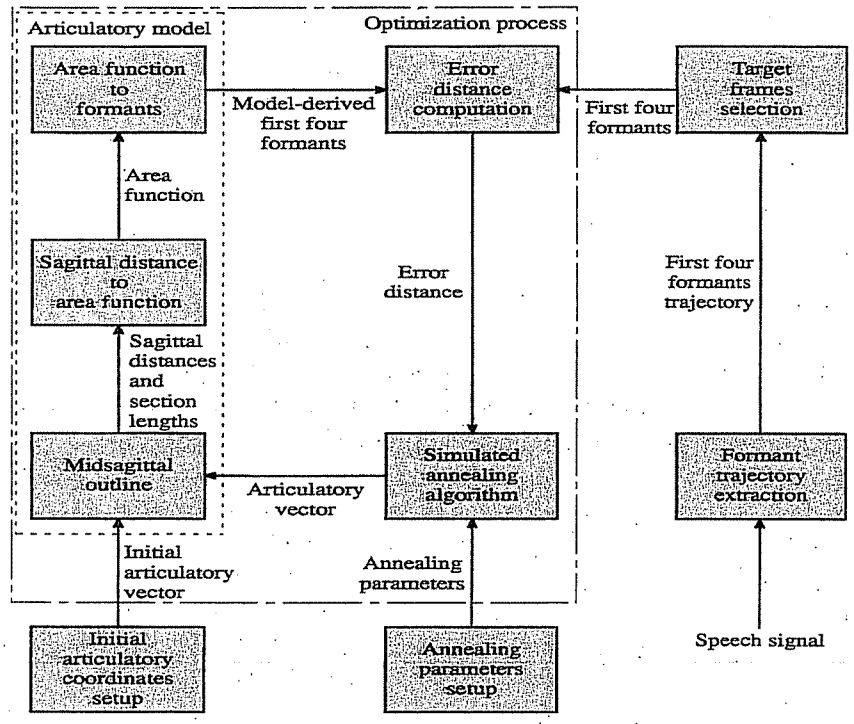


FIGURE A11.31 Block diagram of the speech inverse filtering procedure.

includes the computations of the sagittal distances and the section lengths, the calculations of the vocal tract cross-sectional area and the acoustic transfer function, the decomposition of the first four formants from the acoustic transfer function, and the calculation of the error distance. Then, the simulated annealing algorithm controls the movement of the search path. Each movement requires the generation of a next candidate point, the error distance function evaluation for the candidate point, and the decision to move. After a number of steps, the temperature is lowered and a new search begins. The process stops if the near-global minimum is reached or the maximum allowed number of function evaluations is exceeded. The speech inverse filtering procedure terminates when all target frames are optimized. The articulatory parameters and the vocal tract cross-sectional areas of all the optimized N target frames can be saved as a disk file for later use or can be directly passed to the articulatory synthesizer for synthesis.

Table A11.3 lists the components of the data structure. The first four formant frequencies, the frame starting time, and the frame duration are the initial components. After the minimum error distance is obtained, the first section area of the nasal tract, the optimal articulatory coordinates, the section lengths, and the cross-sectional areas are stored as the content of the target-frame data structure.

TABLE A11.3 Components of the Target Frame Structure

Data Type	Component	Description
Short	tfsetflag	Flag for optimization
Double	ttime	Frame starting time
Double	tdur	Frame duration
Float	ntla	Velopharyngeal port area
Structure pointer	areaf	Area function
Structure pointer	vtlen	Section lengths
Structure pointer	shape	Articulatory coordinates
Structure pointer	tfptr	Target formants
Structure pointer	next	
Structure pointer	previous	Both used for double-linked list

TABLE A11.4 Default Annealing Parameter Values

Annealing Parameters	Default Values
T = artificial temperature (as control parameter)	0.1–0.2 degrees
r_T = temperature reduction coefficient	0.85
N_S = number of steps to adjust the step length vector	20
N_T = number of adjustments at each temperature	5
N_ε = number of successive temperatures to test for stopping	4
η = termination criterion	0.005
N_{tot} = total number of function evaluations	5001
v_i = step length; where i = 1, 2, . . ., M	3.0

The annealing parameters that control the simulated annealing algorithm include the initial temperature T, the temperature reduction coefficient r_T, the number of steps to adjust the step length vector N_S, the number of step adjustments at each temperature N_T, the number of successive temperature reductions to test for termination N_ε, a small constant used for the termination criterion η, and the maximum number of function evaluations N_{tot}. The analogues between the annealing process and the articulatory problem can be identified as follows. First, the percentage of the weighted least-absolute-value (l_1-norm) error distance, Equation (A11.3.3.1.3), corresponds to the energy of the material. The articulatory vector, Equation (A11.3.3.1.1), corresponds to the configuration of particles. The change of articulatory parameters corresponds to the rearrangement of particles. Finding a near-optimal articulatory vector corresponds to finding a low-energy configuration. The temperature of the annealing process, T, becomes the control parameter for the speech inverse filtering process. Second, the Metropolis algorithm corresponds to the random fluctuations in energy. Third, the temperature reduction coefficient r_T corresponds to the cooling rate. Fourth, the finite number of moves at each downward control temperature value, $(N_S)(N_T)$, corresponds to the amount of time spent at each temperature.

Reasonable values of the parameters (Table A11.4) are used as defaults for the optimization process. However, a guideline of the optimization process is given in Appendix A11–D along with two examples.

A11.3.4 Results

A11.3.4.1 Optimization of Vowels Appendix A11–A presents the articulatory and acoustic characteristics for typical American vowels. The midsagittal vocal tract outline and the corresponding vocal tract cross-sectional area function are obtained from sustained vowel phonations by using the simulated annealing algorithm. Appendix A11–A, shows that the simulated annealing optimization algorithm works well, since most of the error distances are less than 0.5%. From these results, we arrive at the following observations.

- Different vowels are characterized by a different set of resonant frequencies (formants), thus, a different vocal tract shape. For example, front vowels (/IY, IH, EH, AE/) are characterized by a large difference between F_1 and F_2, which in turn corresponds to a large back cavity; middle vowels (/AA, ER, AH /) and back vowels (/UW, UH, OW, AO/) have a small difference between F_1 and F_2, which in turn corresponds to a narrow back cavity.
- The three middle vowels (/AA, ER, AH/) and the low, back vowel (/AO/) have similar vocal tract shapes, except for the retroflexed vowel, /ER/, which has a more significant tongue curl-up, which results in a distinctly low F_3.
- For all middle vowels (/AA, ER, AH/) and some other vowels (/AE, AO, OW/), the three articulatory parameters of the lower pharynx; that is, wh, g1k, and hk1, must be optimized.
- To investigate the complexity of the error distance function for each vowel, we used the same initial vocal tract configuration for the optimization process for all vowels. Experience with the optimization process indicates that the error distance function for the high front vowels

(/IY, IH/) and the high, back, rounded vowels (/UW, UH/) converges faster than for other vowels. The middle vowel (/AA/) and the low, back vowel (/AO/) have the slowest convergent rate because these two vowels have a more complex error distance function. For the middle vowel (/ER/), the curl of the tongue causes a distinctly low F_3, which may result in a slow convergence of the optimization process. To articulate the low, back vowel (/AO/), the tongue needs to be lowered and the jaw needs to be opened to widen the oral cavity. Also, the tongue must move back to narrow the pharyngeal cavity. These articulations may make the error distance function complex and slow the convergence of the optimization process. Notice that a more complex error distance function needs a higher initial temperature and more evaluations, since it can have more local minima to escape. See Appendix A11–D for a guideline of the optimization process.

A11.3.4.2 Optimization for a Sentence In Appendix A11–D, the simulated annealing algorithm is applied to perform the speech inverse filtering for the same sentence spoken by two male subjects. The following are some general observations regarding sentence optimization.

- There are two semivowels (/W, Y/) in the speech token analyzed. According to the phonetic classifications, these two semivowels have been categorized as glides. The formant tracks for the sentence, "We were away a year ago," show that the formants, especially the second and third formants, glide up or down to the next vowel. The /W/ glides (frames 1, 6, and 10 of subject A, and frames 1, 4, and 8 of subject B in Appendix A11–D) have three places of articulation in common: the protruded lips, high tongue tip, and high back tongue. However, some of these have a significant upward tongue tip curl. As frame 16 of subject A and frame 13 of subject B show, the tongue blade of the semivowel /Y/ approximates the palate and has been called a palatal glide. In summary, the vocal tract shape for glides "glide" to the next vowel with a fast movement of the tongue and lips.
- Both subjects have quite similar vocal tract shapes for vowel /IY/ and for vowel /ER/, respectively. The lip opening decreases during the transition from vowel /IY/ to semivowel /W/ (frame sequences 3-4-5-6 of subject A, and 2-3-4 of subject B) in order to have protruded lips for /W/.
- The diphthong /EI/ ending with vocal tract shape for /IH/ entails tongue movement forward up from the /EI/ (frame sequences 11-12-13-14 of subject A, and 9-10-11 of subject B). The diphthong /OU/ ending with the vocal tract shaped for /UH/ entails tongue movement back and up, concurrent with lip protrusion (frame sequences 24-25-26 of subject A, and 20-21-22 of subject B).
- The voiced stop /G/ is usually classified as a velar consonant. From frame 22 of subject A and frame 18 of subject B, a classification as the palatal-velar consonant is more correct (Borden and Harris, 1980, p. 117).
- The simulated annealing algorithm performs well. On the average, over 87% of the total frames have an error distance less than 0.1%.

A11.3.4.3 Remarks The above results illustrate the usefulness of the simulated annealing algorithm, which is efficient and flexible in dealing with the problems that are inherent to the acoustic-to-articulatory transformation. However, the selection of parameters for the annealing schedule is difficult, since we know little about the relation between the argument domain (articulatory vector) and the technology (the algorithm). The guideline in Appendix A11–D and the default annealing parameter values in Table A11.4 are considered a good procedure. The evaluation of the error distance function is the most computationally intensive part of the program. On the average, 2000 computations per minute are needed to obtain relatively quick results.

A11.4 SYNTHESIS OF UNVOICED SPEECH

The speech synthesis we have discussed so far has focused on voiced speech. In this section, we discuss the synthesis of unvoiced speech using the articulatory speech synthesizer. The acoustics and

aerodynamics involved in the production of unvoiced speech are far from being completely understood. It is not sufficient to determine the segments of unvoiced speech by estimating the vocal tract transfer function and the waveshape at the glottis. Fricatives are generated at a constriction within the vocal tract, while the glottis is the source for aspiration /HH/. Still, there are many unvoiced speech sounds that are generated by both types of constrictions: one at the glottis and another in the vocal tract (Stevens, 1971). Here we describe some factors relevant to the production of unvoiced speech.

A11.4.1 Introduction

When there is a flow of air through a constriction or past an obstruction, turbulence is created (Shadle, 1991; Stevens, 1971, 1993a, 1993b). Unvoiced speech can be divided into three groups, that is, unvoiced fricatives, unvoiced stops, and affricates. Aspiration noise can be regarded as unvoiced speech as well. An unvoiced fricative is generated by exciting the vocal tract with a steady air flow, which becomes turbulent near the constriction where the velocity of the airflow increases due to the reduced cross-sectional area of the constriction. The location of the constriction determines which sound is produced. The constriction separates the vocal tract into two cavities: the front cavity and the back cavity. The unvoiced speech sound is generated from the front cavity, and the back cavity traps energy and thereby introduces antiresonances into the vocal output. An unvoiced stop consonant is produced by forming a complete closure in the vocal tract and then releasing that closure. The closure is formed by a particular articulator; that is, the lips, the tongue blade, or the tongue body, and then released by moving the articulator rapidly. The initial rapid increase of the cross-sectional area at the constriction gives rise to a transient, and there is a brief burst of turbulence noise following the transient. An unvoiced affricate is a dynamic sound that can be modeled as a concatenation of the unvoiced stop and the fricative consonant. Therefore, this sound has the characteristics of both stop and fricative consonants. An aspiration is produced by turbulent air flow at the glottis with little or no vibration of the vocal folds. Because the vocal tract shape is in the position for the following vowel during the production of the aspiration noise, the characteristics of the aspiration /HH/ are similar and dependent on the vowel that follows the aspiration noise. Since the noise source for aspiration is located at the glottis, there is no front and back cavity in aspiration.

Among these four unvoiced sound groups, the fricative and aspiration are generated by a relatively steady flow of air. Therefore, it is much easier to analyze and synthesize fricatives and aspirations than other unvoiced consonant categories. To understand factors that control the unvoiced speech generation process, we describe the models for generation of unvoiced consonants, which include a turbulence noise generator and the vocal tract structure.

A11.4.2 Unvoiced Speech Production

A11.4.2.1 Models for the Generation of Unvoiced Consonants

A11.4.2.1.1 Fricative Consonants. A schematized model of the vocal tract for an unvoiced fricative consonant is shown in Figure A11.32. A_g and A_c are the cross-sectional areas of the glottis and the constriction, respectively. Likewise, L_g and L_c are the length of the glottis and the

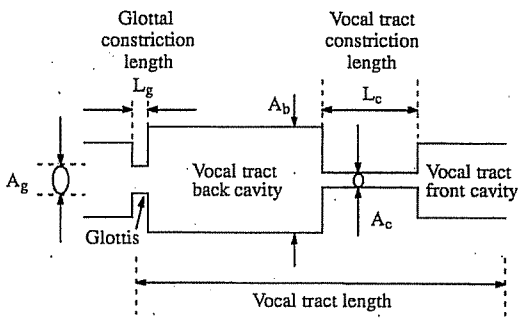


FIGURE A11.32 A model of the vocal tract for a fricative consonant.

vocal tract constriction, respectively. Unvoiced fricatives are produced by a turbulence noise source near a constriction either at the glottis or above the glottis depending on the cross-sectional areas, A_g and A_c. The location of the constriction, that is, the place of the articulation, determines the sound produced. If $A_g > A_c$, the supraglottal constriction plays a major role in the generation of fricative speech. The constriction is formed by the tongue body and/or tongue tip. The length of the constriction, L_c, is typically a few centimeters long. Aspiration noise is generated when $A_g < A_c$. The air flow is independent of the supraglottal constriction. The length of the constriction formed by the glottis, L_g, may be only a few millimeters, which is the average width of the glottis.

As shown in Figure A11.32, when the noise source is located in the vocal tract, the constriction separates the vocal tract into two cavities: the front cavity and the back cavity. Speech is radiated from the front cavity, while the back cavity serves to trap energy, and, thereby, introduces antiresonances into the vocal output. It is reported that the back cavity resonances may have negligible effect on the fricative spectrum if the constriction is sufficiently long and narrow (Heinz and Stevens, 1961). Therefore, the most prominent feature of the fricative spectrum is determined by the turbulence noise source at the constriction and by the resonances of the oral cavity in front of the constriction; that is, the front cavity resonance frequencies.

The constriction in the vocal tract may be abrupt or gradual depending on the articulatory position. It has been found experimentally that the pressure drop across the constriction is proportional to the square of the particle velocity, $(V_c)^2$, and the pressure drop across the constriction Δ_p is approximated by the Bernoulli equation

$$\Delta p = k\rho\frac{V_c^2}{2} = k\rho\frac{U^2}{2A^2} \qquad \text{(A11.4.2.1.1.1)}$$

where k is a constant, ρ is the density of the air, V_c is the flow velocity in the constriction, U is the volume velocity in the constriction, and A is the cross-sectional area of the constriction. The constant k is dependent on the ratio of the cross-sectional area A_b and A_c. It is also dependent on the rate of contraction and expansion of the constriction. Stevens found that a value of 0.9 was a reasonable value for constrictions for normal speakers (Stevens, 1971, 1998).

A11.4.2.1.2 Stop Consonants. For an unvoiced stop consonant, the vocal tract is also modeled as a tube with two constrictions; one at the glottis and one formed by an articulator within the vocal tract. The principal driving source for the flow is the subglottal pressure. At the release of an unvoiced stop consonant, the initial intraoral pressure is equal to the subglottal pressure. Both the glottal constriction area A_g and the supraglottal constriction area A_c change with time as the supraglottal constriction is formed and released. As pressure builds up above the glottis during closure for a consonant, outward forces are exerted on the vocal-fold surfaces, causing a passive increase in the glottal area. The rate of change of cross-sectional area of the supraglottal constriction near the consonantal closure and release can be estimated from cineradiographic, photographic, and airflow data (Childers 1977). It is reasonable to assume the glottal area to be constant and the supraglottal constriction area to increase with a trajectory of the form

$$A_c = (1 - e^{-\frac{t}{\tau}}) \qquad \text{(A11.4.2.1.2.1)}$$

The airflow through the constriction and through the glottis in the few tens of milliseconds following the release gives rise to a sequence of four types of sources: transient, fricative, aspiration, and voicing (Stevens, 1993b). Immediately following the release, there is a brief transient as the air that has been compressed in the vocal tract discharges through the opening constriction. Following this initial transient, the rapid airflow through the constriction gives rise to turbulence and hence to a sound source immediately downstream from the constriction. This source is identified as fricative noise. The turbulence noise source is usually represented as a sound-pressure source near an obstacle downstream from the constriction. Rapid airflow through the glottis also gives rise to turbulence noise, called aspiration noise, with source characteristics similar to those for fricative noise. Vocal-fold vibration begins simultaneously with or immediately following the aspiration noise.

A11.4.2.1.3 Affricate Consonants. The release mechanism for an affricate consonant is different from that for a simple stop consonant. The constriction that is formed by the major articulator to produce an affricate has two parts that can be manipulated independently, an anterior section

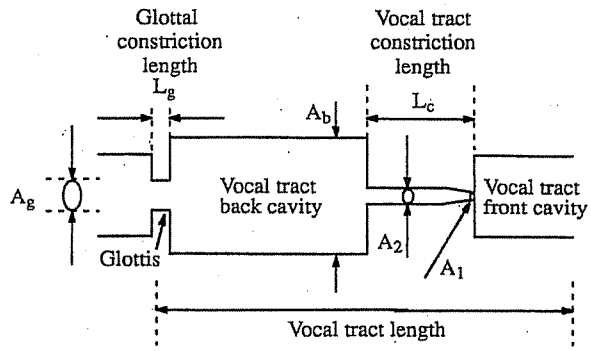


FIGURE A11.33 A model of the vocal tract for an affricate consonant.

and a posterior section. The configuration of the vocal tract can be schematized as in Figure A11.33, where the vocal tract posterior to the constriction is represented as a relatively wide uniform tube. The consonantal constriction has a narrowing near its anterior end, with area A_1 and a longer section with area A_2 behind this narrower section. The cross-sectional area of the glottal opening is A_g. The articulator used to produce an affricate can be the lower lip, the tongue blade, or the tongue body. The acoustic pattern includes an initial transient and an interval over which the fricative noise undergoes changes in amplitude and spectrum as the tongue-tip constriction increases in size.

There is a diverse sequence of acoustic and aerodynamic events at the release of an affricate consonant. During the closure interval, there is a pressure buildup behind the constriction and A_1 is set to zero. A consequence of this increased intraoral pressure is the creation of a force on the articulatory structure that forms the constriction, and it causes a passive downward displacement of the surface. The release occurs in two stages: first the area A_1 is increased, and this initial release is followed by a slower increase of the area A_2 of the longer constriction. At the initial release of the closure, the rate of movement of the articulatory structure that forms anterior area A_1, and hence the rate of increase of A_1, is expected to be rapid, since the movement is enhanced by forces due to pressure behind the constriction. This rate of increase of cross-sectional area is taken to be 50 cm^2/sec during the initial 1 to 2 msec (Stevens, 1993a, 1998).

Following the initial release at the anterior end of the constriction, the force on the surface of the articulator behind the constriction A_1 is reduced, and there may be an initial inward movement of the surface. After that, the posterior area A_2 is maintained for a few tens of milliseconds and then released. The constriction during the later part of the release sequence (after about 50 msec) is much like the constriction for a fricative consonant. Finally, as this longer constriction is released, there may be a time interval in which turbulence noise is generated at the glottal constriction A_g before glottal vibration commences.

For the transient source after the initial 0.1 to 0.2 msec, the initial rate of increase of the airflow can be estimated roughly by assuming that the flow U is proportional to the area A of the opening, and is given by

$$U \cong A\sqrt{\frac{2P_m}{\rho}} \qquad \text{(A11.4.2.1.3.1)}$$

where P_m is the intraoral pressure and ρ is the density of the air. After the initial transient source, turbulence noise will be generated as a consequence of airflow through the palatoalveolar constriction impinging on the lower incisors. The effective amplitude of the turbulence noise source is assumed to be proportional to $U^3A^{-2.5}$, where U is the volume velocity through the constriction and A is the cross-sectional area of the constriction (Stevens, 1971, 1998).

Figure A11.34 gives an estimate of the time course of the change in cross-sectional area for the front portion of the affricate constriction, represented by A_1 in the schematized configuration in Figure A11.33. Also shown in Figure A11.34 is the estimated cross-sectional area of the palatal constriction following the initial release. This is the equivalent of the area A_2 in the schematized configuration in Figure A11.33. The time course of this cross-sectional area over the time interval beyond 50 msec in Figure A11.34, together with the change in cross-sectional area of the glottal opening A_g, are similar to the trajectories that are normally assumed for the release of a fricative consonant (Stevens et al., 1992; Stevens, 1998). In summary, the requirement for an affricate is that the constriction formed by

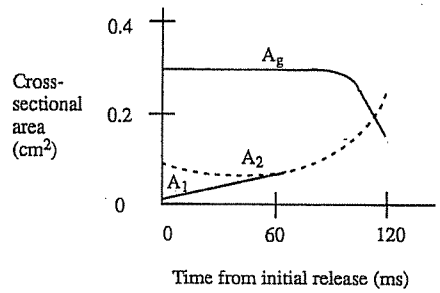


FIGURE A11.34 Time course of the change in cross-sectional area for an affricate consonant (after Stevens et al., 1993a).

the relevant articulator can be divided into an anterior portion, which executes the initial rapid release from closure, and a longer posterior portion, which is shaped to produce fricative noise corresponding to the fricative portion of the sequence.

A11.4.2.2 Model for Turbulence Noise Generation

A11.4.2.2.1 Characteristics of Equivalent Noise Source. The configuration of the vocal tract for the generation of an unvoiced turbulence noise source can be modeled as in Figure A11.35. This model was originally suggested by Fant (1960) who used it for calculations for various fricative and stop consonants. The source is assumed to be independent of the constriction area A_c. The impedance, Z_2, is seen looking upstream from the sound pressure source, P_s, and is frequency dependent. The impedance looking downstream from the pressure source is Z_1. According to this model, the front and back cavity resonance frequencies are determined by Z_1 and Z_2, respectively. The volume velocity U_0 at the mouth flows through the radiation impedance Z_r. The transfer function $\frac{U_0}{P_s}$, that is, the transfer function from the pressure source to the volume velocity at the lips can be calculated as

$$\frac{U_0}{P_s} = \frac{1}{Z_1 + Z_2} \qquad \text{(A11.4.2.2.1.1)}$$

Therefore, due to the frequency dependent impedance Z_2, the spectrum of the turbulence noise is characterized by poles at the natural frequencies for which $Z_1 + Z_2 = 0$, and by zeros at the frequencies for which $Z_2 = \infty$.

It has been found experimentally that the spectrum of the series pressure source in Figure A11.35 is relatively flat over a frequency range of two or three octaves centered on 0.2 V/D, where V is the velocity and D is a characteristic dimension (Stevens, 1971). More specifically, the center frequency can be represented as

$$f_c = 0.2\frac{U}{A^{\frac{3}{2}}} \qquad \text{(A11.4.2.2.1.2)}$$

where U is the volume velocity of the air flow near the constriction and A is the cross-sectional area of the constriction. For typical values of the volume velocities and constriction sizes encountered in turbulence noise generated speech, the center frequency is in the range of 500 to 3000 Hz. The lower end of this range is for aspiration noise /HH/ and the higher end is for fricative noise.

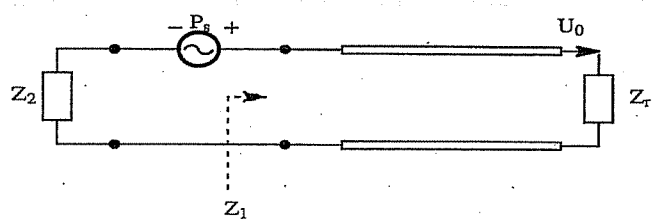


FIGURE A11.35 Equivalent circuit for turbulence noise generation.

In summary, the most important two components in this model are the turbulence noise source, P_s, at the constriction and the impedance Z_l looking downstream from the pressure source that determines the front cavity resonance frequencies. As the position of the constriction moves forward from the glottis to the lips, the pole of the transfer function moves higher in frequency, for example, from about 500 Hz for aspiration noise /HH/ to about 4000 Hz for fricative noise /S/ (Stevens, 1971; 1998).

A11.4.2.2.2 Noise Source Model. For convenience we repeat some of the results from Section A11.2.3.6. The airflow in a narrow tube generates turbulence noise, and the intensity and spectrum of the noise are decided by the Reynolds number R_e. If $R_e > R_{ec}$, where R_{ec} is the critical Reynolds number, then a noise with high intensity is generated, which is a turbulence noise. The Reynolds number is defined as

$$R_e = \frac{\rho \upsilon h}{\mu}$$

where ρ is air density, υ is linear velocity of air flow, h is characteristic dimension of the constriction, and μ is viscosity of air (Sorokin, 1994).

Basically, the noise source model defines the characteristics of the noise source as a function of the airflow through the constriction and of the constriction cross-sectional area. Broad (1977b) found that the rms sound pressure of the noise could be expressed as

$$P_{rms} = A_c \left(R_e^2 - R_{ec}^2\right) \qquad \text{(A11.4.2.2.2.1)}$$

where R_e is the Reynolds number and R_{ec} is the critical Reynolds number. Fant (1960) adopted a serial noise pressure source and reformulated the P_{rms} as a function of the pressure drop through the constriction and the effective width of the constriction. Under the plane wave assumption, the sound pressure of turbulent flow can be taken as proportional to the square of the volume velocity of the airflow and inversely proportional to the constriction area, A_c (Stevens, 1971). The location of the turbulence noise source may be located at the center of, or immediately downstream from, the constriction region, or possibly at a combination of these places, or spatially distributed along the constriction region (Fant, 1960; Flanagan and Cherry, 1968; Lin, 1990; Stevens, 1971, 1993a, 1993b).

The spectrum of the turbulence noise is broadly distributed over a wide range of frequencies (2 to 8 kHz) with some accentuation in the mid-audio range (Childers and Lee, 1991; Stevens, 1971, 1993a, 1993b). Klatt (1980) used a random number generator, a spectrum-shaping filter, and an amplitude modulator to model the turbulent flow for the formant synthesizer. The spectrum-shaping filter was designed to simulate the spectral characteristics of the turbulent flow. A first order IIR filter was used to obtain the volume velocity due to a random pressure source. Childers and Lee (1991) have used a FIR filter to model highpass-filtered turbulence noise. Cook (1991, 1993) used a four-pole filter to model the spectral properties of the noise source.

By including a latent random pressure source, P_n, and an inherent constriction loss, R_n, in each elemental section of the vocal tract, Flanagan and Cherry (1968) could introduce the turbulent flow excitation at any section. However, as Sondhi and Schroeter (1987) pointed out, the Flanagan and Cherry (1968) model did not produce satisfactory unvoiced sounds due to the high "back" cavity impedance. Sondhi and Schroeter (1986, 1987), thus, modified the model to a parallel flow source $U_n = \frac{P_n}{R_n}$, which was located downstream from the constriction. The P_n is given by

$$\begin{aligned} P_n &= (\text{turbg})(\text{rand})\left(R_e^2 - R_{ec}^2\right), && \text{for } R_e > R_{ec} \\ &= 0, && \text{for } R_e < R_{ec} \end{aligned} \qquad \text{(A11.4.2.2.2.2)}$$

where turbg is empirically determined as the turbulence gain and rand is a random number uniformly distributed between -0.5 and 0.5. The source resistance R_n is

$$R_n = \frac{\rho |\bar{U}|}{2A^2} \qquad \text{(A11.4.2.2.2.3)}$$

where $\bar{U}$ is a digitally low-pass filtered version of the volume velocity U at the constriction and is

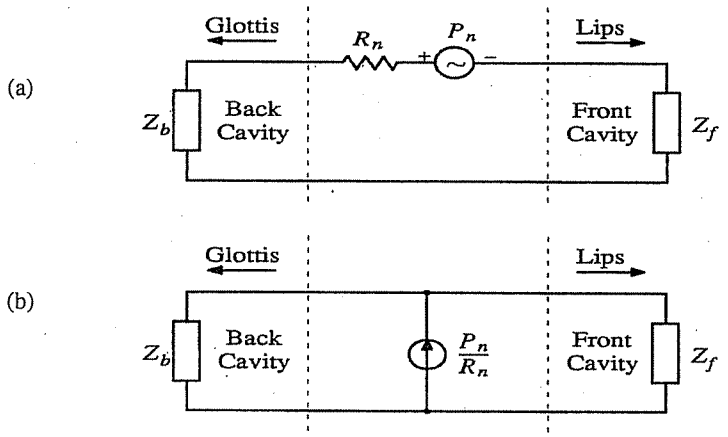


FIGURE A11.36 Equivalent circuits for the turbulence source. (*a*) Serial (after Flanagan and Cherry, 1968). (*b*) Parallel (after Sondhi and Schroeter, 1986).

defined as

$$\bar{U}(n) = \bar{U}(n-1) + [U(n) - \bar{U}(n-1)]2\pi f_g T \qquad \text{(A11.4.2.2.2.4)}$$

The value of the cutoff frequency f_g is not critical. Flanagan and co-workers (1975) used 500 Hz in order to ensure stability. A first-order IIR filter with a cutoff frequency of 2000 Hz was used to lowpass filter the flow. Figure A11.36(*a*) and (*b*) show the equivalent circuits of the serial and parallel turbulence sources, respectively.

Based on the previous discussion, we adopt the turbulence noise source model from Sondhi and Schroeter (1986, 1987) into the transmission-line circuit model of the vocal system. This model allows the user to place the turbulence noise source at the center of, or immediately downstream or upstream from, the constriction region, or spatially distributed along the constriction region. The turbulence gain and critical Reynolds number can also be specified.

A11.4.3 Synthesis Results

Examples of several speech tokens were synthesized using the articulatory speech synthesis toolbox given in Chapter 10. The speech tokens consisted of the affricate /CH/, fricative /S/, and the sentence, "Should we chase those cowboys?"

A11.4.3.1 Example 1—Synthesis of the Affricate /CH/ Affricates and fricatives are produced by the formation of a narrow supraglottal constriction and the generation of turbulence at the vicinity of this constriction. The experimental sections for fricatives are taken at the interval of maximum intensity, which occurs near the middle or towards the end of the sound. Obtaining articulatory data for unvoiced consonants from speech is only in a preliminary stage (Sorokin, 1994; Stevens, 1993a, 1993b, 1998). Thus, a vocal tract cross-sectional area (Figure A11.37) estimated from the x-ray photograph of Badin (1991) was adopted for the affricate /CH/ in this experiment. The actual width of a very narrow constriction cannot always be estimated accurately from sagittal x-ray photographs. In addition, the perpendicular dimensions are generally inaccessible to measurements. The simulations were carried out on the assumption that the minimum cross-sectional area was 0.16 cm^2. Errors in the estimation of these dimensions do not severely affect the calculations, since the length of a narrow constriction is more crucial than the cross-sectional area.

Modeling the complex turbulent phenomena is another challenging problem. In this experiment, the affricate /CH/ was synthesized with the turbulence noise source located at (1) the center of; (2) immediately downstream; (3) upstream from the constriction region; and (4) spatially distributed along the constriction region. The model presented in Figure A11.36(*b*) was used to simulate the turbulence noise source. The turbulence gain and critical Reynolds number were specified at 0.000002 and 2700, respectively. The glottal volume velocity was assumed to be a DC source and set at 1000 cm^3/sec.

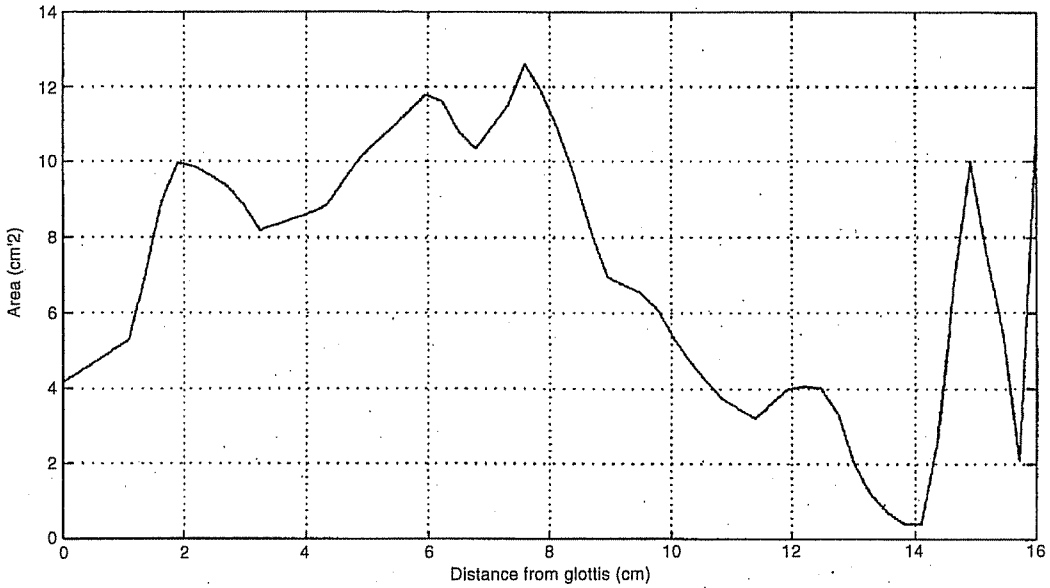


FIGURE A11.37 Vocal tract cross-sectional area for the affricate /CH/.

In Figure A11.38, the transfer function of /CH/ is displayed for various locations of the noise source. Placing the excitation source in the vocal tract introduces antiresonances into the acoustic transfer function. The vocal tract resonant frequencies are relatively unaffected when the excitation is placed forward from the pharynx toward the front oral cavity. Another feature is that the number of antiresonances increases as the excitation is moved forward from the pharynx to the front oral cavity. Figure A11.39 gives a comparison of the power spectral density of the synthesized and original speech for the affricate /CH/. A 6th-order LPC model was used to analyze a segment of the speech signal. Note that the spectral characteristics of the noise source located at the center of and upstream from the constriction region are similar but differ slightly from the other two locations. The resonant peak near 3700 Hz is due to the resonance of the small, short front cavity.

The resonance is less prominent when the turbulence noise source is distributed along the constriction region. The resonant frequency and amplitude for the downstream case (2000 Hz) are lower than those for original speech case (2200 Hz). The difference may be due to differences in the dimensions of the vocal tract for the location of the constriction and the source characteristics of the turbulence noise for the two speakers. The synthesized and original speech are similar in the low frequency region. This means that the back cavity has little effect on the synthesized speech.

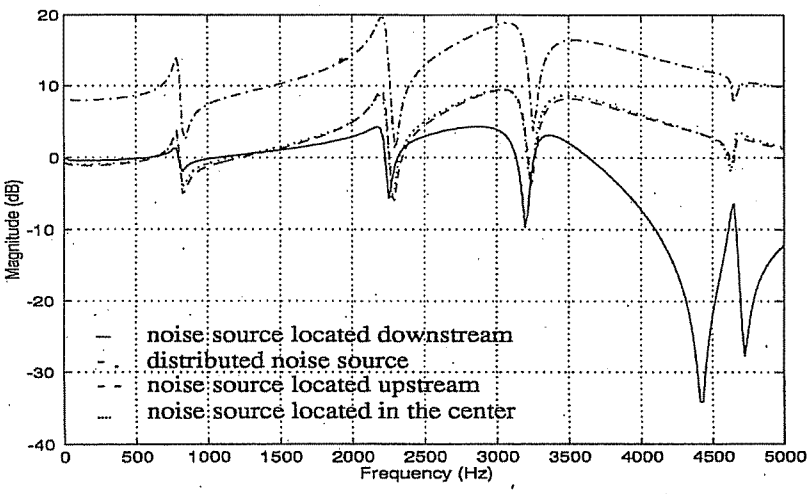


FIGURE A11.38 Transfer function for the affricate /CH/ for various locations of the noise source.

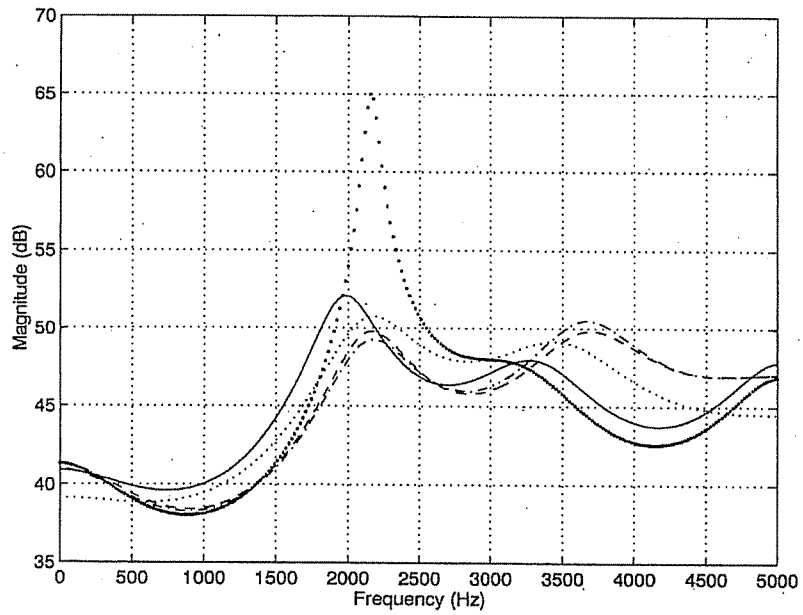


FIGURE A11.39 Power spectral density of the synthesized and original affricate /CH/.

A11.4.3.2 Example 2—Effect of the Back Cavity on the Synthesized Fricative /S/

For some fricative sounds generated with a supraglottal constriction, the effect of the back cavity resonance on the speech spectrum can be neglected if the constriction is narrow and long (Heinz and Stevens, 1961). Here, we confirm that the back cavity resonance does not greatly effect the synthetic fricative sound as long as the supraglottal constriction is sufficiently narrow and long. The shape of the area function corresponding to the back cavity is changed as specified in the left column of Figure A11.40. The front cavity and constriction of the three area functions remain the same; that is, the cross-sectional area of the supraglottal constriction is 0.16 cm^2, the front cavity length is 2.02 cm, and the constriction length is 2.02 cm.

The synthesized speech is similar to the original fricative /S/. Since the vocal tract configuration for /S/ has a distinct front cavity and back cavity, it is not difficult to observe the effect of the back cavity resonance on the speech spectrum. Synthesized fricatives generated with two different back cavity shapes are compared with a synthesized fricative using data for an area function measured from an x-ray (Fant, 1960) in Figure A11.40.

As expected, the formant structure of the three spectra are quite similar, since the back cavity resonance is decoupled from the constriction and the front cavity. Consequently, there is little influence on the overall speech spectrum. These results are in agreement with the fricative speech production theory; that is, the back cavity resonance does not play an important role in generating fricative sounds.

A11.4.3.3 Example 3—Synthesis of the Sentence, "Should We Chase Those Cowboys?"

The sentence, "Should we chase those cowboys?" consists of voiced and unvoiced sounds. For the voiced portion of the original speech, speech inverse filtering was performed with the simulated annealing algorithm to obtain the vocal tract cross-sectional area. Based on the pitch contour and LF parameters obtained in the analysis phase, the excitation waveform model was constructed. Unvoiced consonants have traditionally proven difficult to model and synthesize because of the complex nature of their production mechanisms and the lack of sufficient articulatory and aerodynamic data for these sounds. The dynamic, time-varying characteristics and the complex

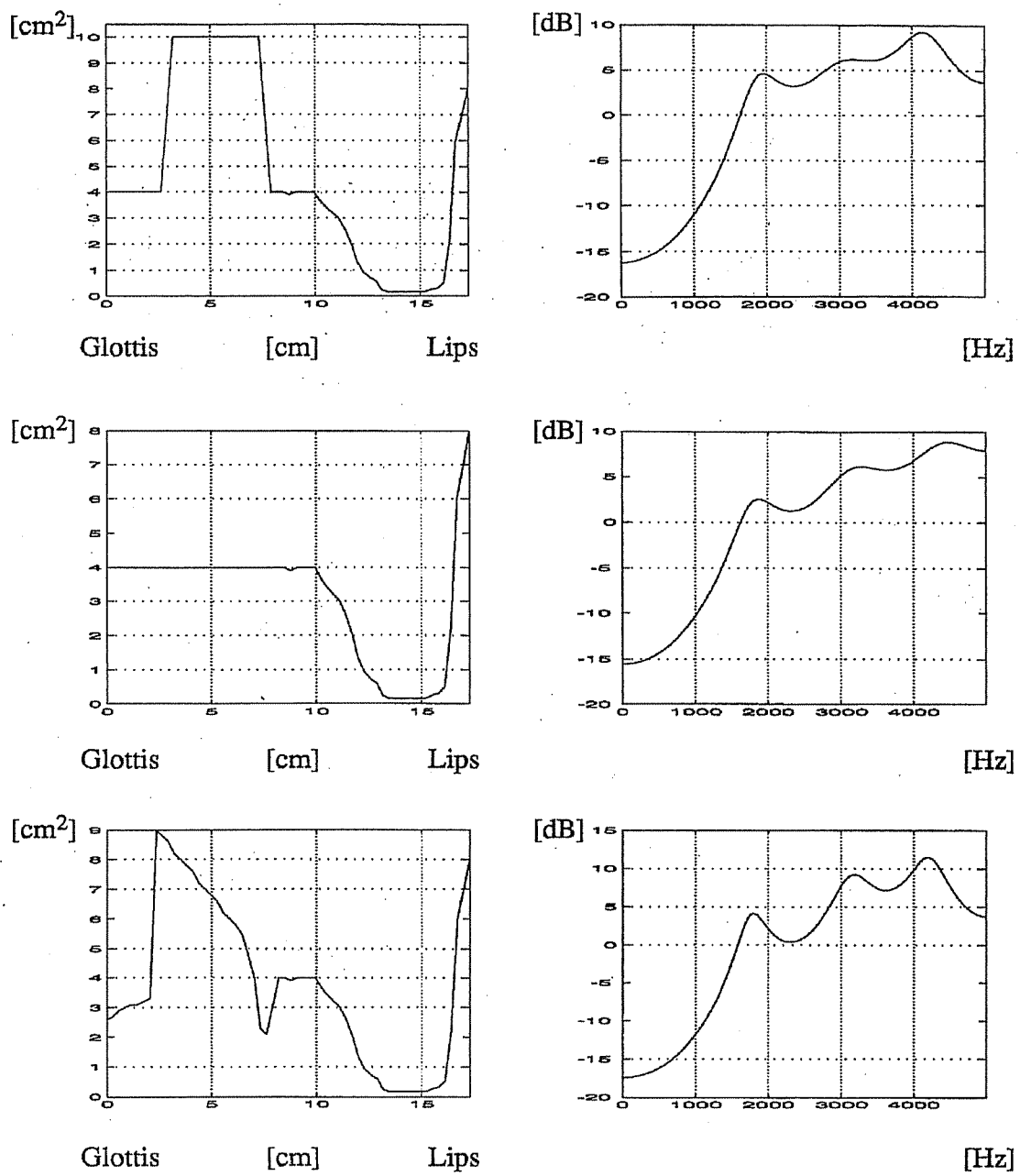


FIGURE A11.40 Area function and LPC spectrum for a synthesized /S/ for different back cavity shapes.

nature of their production have made plosives the most difficult phonemes to model and synthesize. Thus, only the fricative-like portion of unvoiced sounds and the corresponding cross-sectional areas are considered here.

The experimental sections for fricatives and affricates are taken at the interval of maximum intensity, which occurs near the middle or end of the sound. The experimental sections for stops are also taken at the interval of maximum intensity, which generally falls at the beginning of the unvoiced sound interval identified as the stop burst. The unvoiced simulations also assume a minimum cross-sectional area of 0.16 cm^2. The model presented in Figure A11.36(*b*) was used to simulate the turbulence noise source. The turbulence gain and critical Reynolds number were specified at 0.000002 and 2700, respectively. The glottal volume velocity was assumed to be a DC source and set at 1000 cm^3/sec. The original and synthetic speech signals and wideband spectrograms are shown in Figure A11.41,

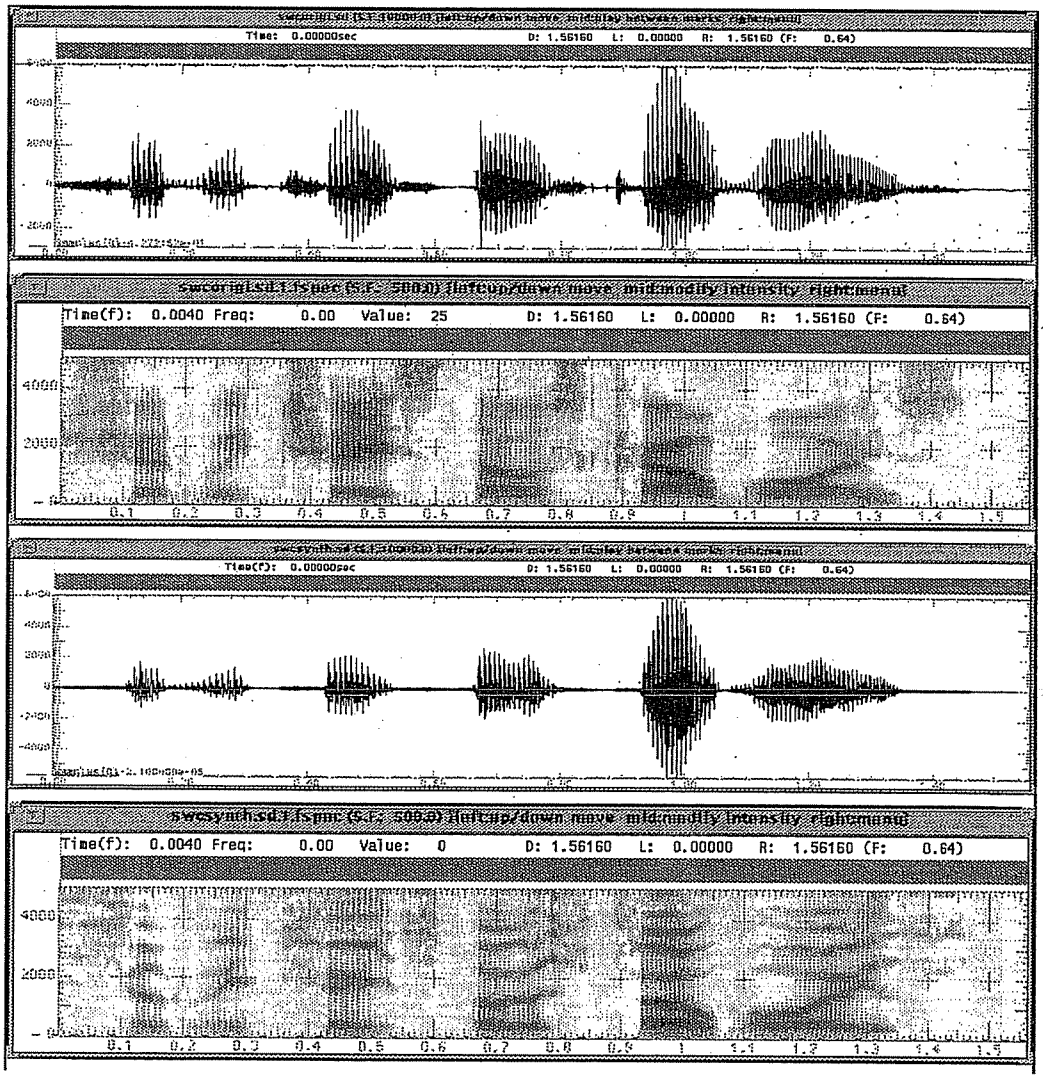


FIGURE A11.41 The original (upper) and synthesized (lower) speech and spectrograms for the sentence, "Should we chase those cowboys?"

and are reasonably similar. The synthesized speech is highly intelligible, but is not as natural as the original.

A11.4.4 Discussion

From the examples described previously (and others), it is concluded that the articulatory speech synthesis toolbox in Chapter 10 is able to produce a good quality synthetic speech. This indicates that the articulatory synthesis tool effectively identifies and simulates the human vocal system. From the results of these experiments, we summarize the production of unvoiced speech as follows.

- Schroeter and Sondhi (1994) concluded that the synthesis of fricatives in the articulatory synthesizer is not yet satisfactory. However, from our experiments, the turbulence noise source location was found to have important acoustic consequences. The downstream case is able to generate spectral characteristics close to the original speech. The major problem for synthesis of unvoiced sounds lies in the estimation of the relevant parameters from the acoustic speech signal, such as inferring articulatory and source information.
- Our results confirmed that the back cavity resonance does not have much effect on the synthetic fricative sound as long as the supraglottal constriction is sufficiently narrow and long. For the

fricative /S/, the supraglottal constrictions, which are configured by the tongue tip and tooth ridge, are narrow and long enough to decouple the back cavity resonance with the front cavity.

- Analysis and synthesis of voiced sounds provide us with an understanding of the production of phonetic information and vocal characteristics. Numerous algorithms have been proposed and many of them give successful results. For unvoiced consonants, the acoustics and aerodynamics involved in the production of unvoiced speech are far from being completely understood because of the complex nature of their production mechanisms and lack of sufficient articulatory and aerodynamic data. The experimental sections for unvoiced sounds in our analysis and synthesis are taken at their intervals of maximum intensity. Since many portions of speech are voiced sounds, the analysis and synthesis based on voiced sounds and experimental sections of unvoiced sounds are enough to generate an intelligible synthetic speech.

A11.5 SUMMARY

This appendix focuses on four main areas of the articulatory synthesizer model: the articulatory model, the acoustic model, the analysis of various vocal system characteristics, and the articulatory speech synthesizer. After a brief review of the articulatory model, we defined the articulatory parameters and described the articulatory model in some detail. Our articulatory model represents the vocal tract by 60 sections to provide more reliable estimates of the cross-sectional areas. We have made an attempt to cover the acoustic model of the entire vocal system, which includes the vocal tract, the nasal tract, the sinuses, the glottal impedance, the subglottal tract, the glottal excitation source, and the turbulence noise source. A transmission-line circuit model of the vocal system is provided. A model for unvoiced speech production is included along with several examples. Also included in this appendix is an analysis of several characteristics of the vocal system that is based on the calculation of the acoustic transfer function, which is described in Appendix A11–B. Such an analysis provides a basis for choosing appropriate parameters for the articulatory synthesizer. The effect of relocating the excitation source on the acoustic transfer function is also described. For more details see C. Wu (1996). Finally, the strategy of the implementation of the articulatory synthesizer is presented after a review of the articulatory synthesis approaches. The time-domain approach is implemented to provide the ability to investigate the dynamic properties of the vocal system. A derivation of the discrete time acoustic equations is given in Appendix A11–C. One interpolation function, linear, is used to interpolate the vocal tract cross-sectional area or articulatory parameters.

We described the simulated annealing optimization algorithm in detail after reviewing the derivations of the vocal tract cross-sectional area. The simulated annealing algorithm is based on the Corana et al. (1987) approach. The articulatory vector defines the set of parameters to be optimized. The cost function is a percentage of the weighted least-absolute-value error distance. It defines a comparison of the first four formant frequencies between the model-generated and the target-frame (from speech analysis). A 1% error criterion gives satisfactory results. Once the optimum articulatory vector is obtained, the articulatory model determines the vocal tract cross-sectional area function, which in turn is used by the articulatory speech synthesizer. Results and discussion of speech inverse filtering for twelve typical American vowels and two sentences are presented in Appendix A11–A and Appendix A11–D, respectively. Default annealing parameters that control the simulated annealing algorithm are also given in Table A11.4. The simulated annealing algorithm is efficient and flexible in dealing with the problems that are inherent to speech inverse filtering. The author is grateful for the results provided by Y. F. Hsieh (1994) and C. "John" Wu (1996).