Sondhi, M.M. & Schroeter, J. “Speech Production Models and Their Digital Implementations”
Digital Signal Processing Handbook
Ed. Vijay K. Madisetti and Douglas B. Williams
Boca Raton: CRC Press LLC, 1999

44

Speech Production Models and Their Digital Implementations

M. Mohan Sondhi
Bell Laboratories
Lucent Technologies

Juergen Schroeter
AT&T Labs — Research

44.1 Introduction

The characteristics of a speech signal that are exploited for various applications of speech signal processing to be discussed later in this section on speech processing (e.g., coding, recognition, etc.) arise from the properties and constraints of the human vocal apparatus. It is, therefore, useful in the design of such applications to have some familiarity with the process of speech generation by humans. In this chapter we will introduce the reader to (1) the basic physical phenomena involved in speech production, (2) the simplified models used to quantify these phenomena, and (3) the digital implementations of these models.

44.1.1 Speech Sounds

Speech is produced by acoustically exciting a time-varying cavity — the vocal tract, which is the region of the mouth cavity bounded by the vocal cords and the lips. The various speech sounds are produced by adjusting both the type of excitation as well as the shape of the vocal tract.

There are several ways of classifying speech sounds [1]. One way is to classify them on the basis of the type of excitation used in producing them:

- **Voiced** sounds are produced by exciting the tract by quasi-periodic puffs of air produced by the vibration of the vocal cords in the larynx. The vibrating cords modulate the air stream from the lungs at a rate which may be as low as 60 times per second for some males to as high as 400 or 500 times per second for children. All vowels are produced in this manner. So are laterals, of which **l** is the only exemplar in English.

- **Nasal** sounds such as **m, n, ng,** and nasalized vowels (as in the French word **bon**) are also voiced. However, part or all of the airflow is diverted into the nasal tract by opening the velum.
- **Plosive** sounds are produced by exciting the tract by a sudden release of pressure. The plosives **p, t, k** are voiceless, while **b, d, g** are voiced. The vocal cords start vibrating before the release for the voiced plosives.
- **Fricatives** are produced by exciting the tract by turbulent flow created by air flow through a narrow constriction. The sounds **f, s, sh** belong to this category.
- **Voiced fricatives** are produced by exciting the tract simultaneously by turbulence and by vocal cord vibration. Examples are **v, z,** and **zh** (as in **pleasure**).
- **Affricates** are sounds that begin as a stop and are released as a fricative. In English, **ch** as in **check** is a voiceless affricate and **j** as in **John** is a voiced affricate.

In addition to controlling the type of excitation, the shape of the vocal tract is also adjusted by manipulating the tongue, lips, and lower jaw. The shape determines the frequency response of the vocal tract. The frequency response at any given frequency is defined to be the amplitude and phase at the lips in response to a sinusoidal excitation of unit amplitude and zero phase at the source. The frequency response, in general, shows concentration of energy in the neighborhood of certain frequencies, called **formant frequencies**.

For vowel sounds, three or four resonances can usually be distinguished clearly in the frequency range 0 to 4 kHz. (On average, over 99% of the energy in a speech signal is in this frequency range.) The configuration of these resonance frequencies is what distinguishes different vowels from each other.

For fricatives and plosives, the resonances are not as prominent. However, there are characteristic broad frequency regions where the energy is concentrated.

For nasal sounds, besides formants there are anti-resonances, or zeros in the frequency response. These zeros are the result of the coupling of the wave motion in the vocal and nasal tracts. We will discuss how they arise in a later section.

44.1.2 Speech Displays

We close this section with a description of the various ways of displaying properties of a speech signal. The three common displays are (1) the **pressure waveform**, (2) the **spectrogram**, and (3) the **power spectrum.** These are illustrated for a typical speech signal in Figs. 44.1a–c.

Figure 44.1a shows about half a second of a speech signal produced by a male speaker. What is shown is the **pressure waveform** (i.e., pressure as a function of time) as picked up by a microphone placed a few centimeters from the lips. The sharp click produced at a plosive, the noise-like character of a fricative, and the quasi-periodic waveform of a vowel are all clearly discernible.

Figure 44.1b shows another useful display of the same speech signal. Such a display is known as a **spectrogram** [2]. Here the x-axis is time. But the y-axis is frequency and the darkness indicates the intensity at a given frequency at a given time. [The intensity at a time t and frequency f is just the power in the signal averaged over a small region of the time-frequency plane centered at the point (t, f)]. The dark bands seen in the vowel region are the **formants.** Note how the energy is much more diffusely spread out in frequency during a plosive or fricative.

Finally, Fig. 44.1c shows a third representation of the same signal. It is called the **power spectrum.** Here the power is plotted as a function of frequency, for a short segment of speech surrounding a specified time instant. A logarithmic scale is used for power and a linear scale for frequency. In

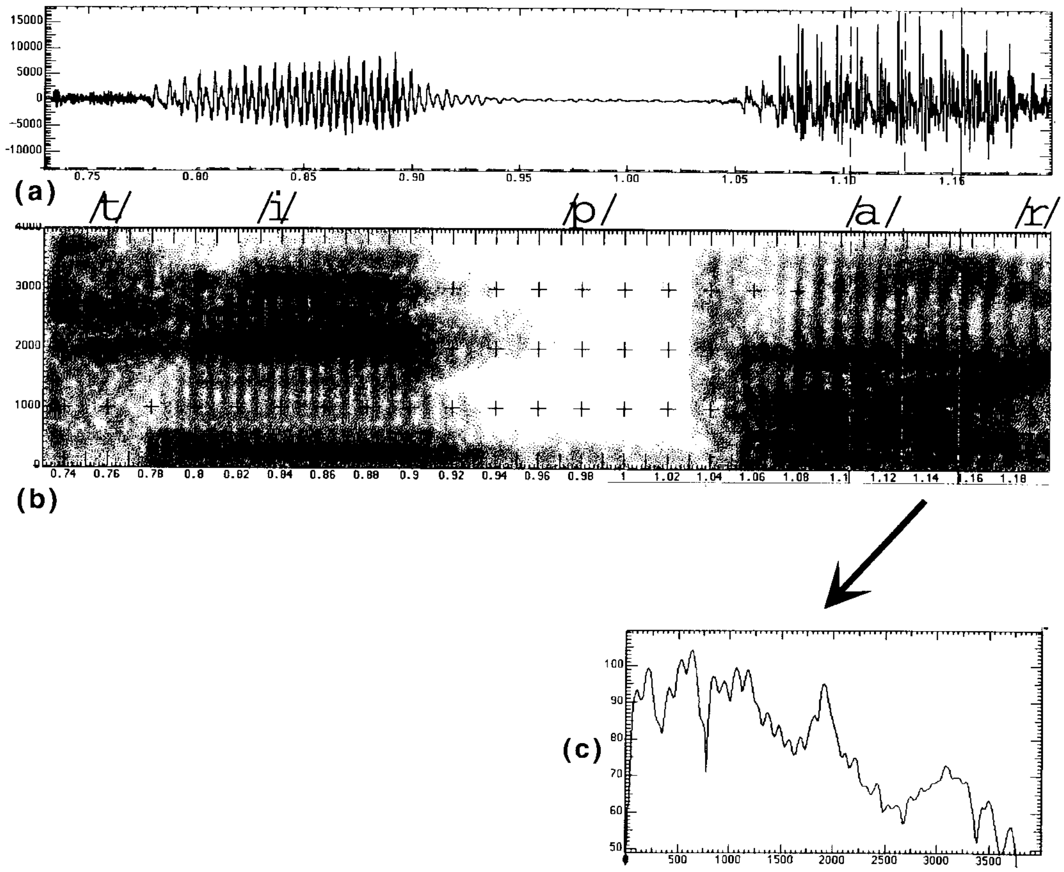

FIGURE 44.1: Display of speech signal: (a) waveform, (b) spectrogram, and (c) frequency response.

this particular plot, the power is computed as the average over a window of duration 20 msec. As indicated in the figure, this spectrum was computed in a voiced portion of the speech signal. The regularly spaced peaks — the fine structure — in the spectrum are the harmonics of the fundamental frequency. The spacing is seen to be about 100 Hz, which checks with the time period of the wave seen in the pressure waveform in Fig. 44.1a. The peaks in the envelope of the harmonic peaks are the formants. These occur at about 650, 1100, 1900, and 3200 Hz, which checks with the positions of the formants seen in the spectrogram of the same signal displayed in Fig. 44.1b.

44.2 Geometry of the Vocal and Nasal Tracts

Much of our knowledge of the dimensions and shapes of the vocal tract is derived from a study of x-ray photographs and x-ray movies of the vocal tract taken while subjects utter various specific speech sounds or connected speech [3]. In order to keep x-ray dosage to a minimum, only one view is photographed, and this is invariably the side view (a view of the mid-sagittal plane). Information about the cross-dimensions is inferred from static vocal tracts using frontal X rays, dental molds, etc.

More recently, Magnetic Resonance Imaging (MRI) [4] has also been used to image the vocal and nasal tracts. The images obtained by this technique are excellent and provide three-dimensional reconstructions of the vocal tract. However, at present MRI is not capable of providing images at a rate fast enough for studying vocal tracts in motion.

Other techniques have also been used to study vocal tract shapes. These include:

(1) ultrasound imaging [5]. This provides information concerning the shape of the tongue but not about the shape of the vocal cavity.

(2) Acoustical probing of the vocal tract [6]. In this technique, a known acoustic wave is applied at the lips. The shape of the time-varying vocal cavity can be inferred from the shape of the time-varying reflected wave. However, this technique has thus far not achieved sufficient accuracy. Also, it requires the vocal tract to be somewhat constrained while the measurements are made.

(3) Electropalatography [7]. In this technique, an artificial palate with an array of electrodes is placed against the hard palate of a subject. As the tongue makes contact with this palate during speech production, it closes an electrical connection to some of the electrodes. The pattern of closures gives an estimate of the shape of the contact between tongue and palate. This technique cannot provide details of the shape of the vocal cavity, although it yields important information on the production of consonants.

(4) Finally, the movement of the tongue and lips has also been studied by tracking the positions of tiny coils attached to them [8]. The motion of the coils is tracked by the currents induced in them as they move in externally applied electromagnetic fields. Again, this technique cannot provide a detailed shape of the vocal tract.

Figure 44.2 shows an x-ray photograph of a female vocal tract uttering the vowel sound /u/. It is seen that the vocal tract has a very complicated shape, and without some simplifications it would be very difficult to just specify the shape, let alone compute its acoustical properties. Several models have been proposed to specify the main features of the vocal tract shape. These models are based on studies of x-ray photographs of the type shown in Fig. 44.2, as well as on x-ray movies taken of subjects uttering various speech materials. Such models are called **articulatory models** because they specify the shape in terms of the positions of the **articulators** (i.e., the tongue, lips, jaw, and velum).

Figure 44.3 shows such an idealization, similar to one proposed by Coker [9], of the shape of the vocal tract in the mid-sagittal plane. In this model, a fixed shape is used for the palate, and the shape of the vocal cavity is adjusted by specifying the positions of the articulators. The coordinates used to describe the shape are labeled in the figure. They are the position of the tongue center, the radius of the tongue body, the position of the tongue tip, the jaw opening, the lip opening and protrusion, the position of the hyoid, and the opening of the velum. The cross-dimensions (i.e., perpendicular to the sagittal plane) are estimated from static vocal tracts. These dimensions are assumed fixed during speech production. In this manner, the three-dimensional shape of the vocal tract is modeled.

Whenever the velum is open, the nasal cavity is coupled to the vocal tract, and its dimensions must also be specified. The nasal cavity is assumed to have a fixed shape which is estimated from static measurements.

44.3 Acoustical Properties of the Vocal and Nasal Tracts

Exact computation of the acoustical properties of the vocal (and nasal) tract is difficult even for the idealized models described in the previous section. Fortunately, considerable further simplification can be made without affecting most of the salient properties of speech signals generated by such a model. Almost without exception, three assumptions are made to keep the problem tractable. These assumptions are justifiable for frequencies below about 4 kHz [10, 11].

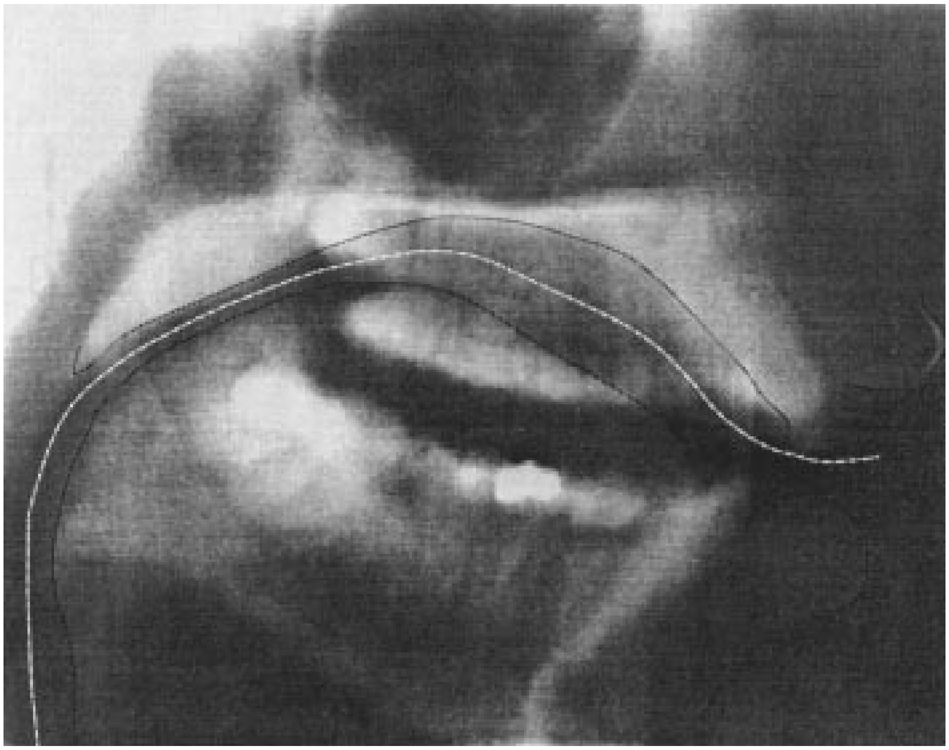

FIGURE 44.2: X-ray side view of a female vocal tract. The tongue, lips, and palate have been outlined to improve visibility. (*Source:* Modified from a single frame from “Laval Film 55,” Side 2 of Munhall, K.G., Vatikiotis-Bateson, E., Tohkura, Y., X-ray film data-base for speech research, ATR Technical Report Tr-H-116, 12/28/94, ATR Human Information Processing Research Laboratories, Kyoto, Japan. With permission from Dr. Claude Rochette, Departement de Radiologie de l’Hotel-Dieu de Quebec, Quebec, Canada.)

44.3.1 Simplifying Assumptions

1. It is assumed that the vocal tract can be **“straightened out”** in such a way that a center line drawn through the tract (shown dotted in Fig. 44.3) becomes a straight line. In this way, the tract is converted to a straight tube with a variable cross-section.
2. Wave propagation in the straightened tract is assumed to be **planar.** This means that if we consider any plane perpendicular to the axis of the tract, then every quantity associated with the acoustic wave (e.g., pressure, density, etc.) is independent of position in the plane.
3. The third assumption that is invariably made is that wave propagation in the vocal tract is **linear.** Nonlinear effects appear when the ratio of particle velocity to sound velocity (the **Mach number**) becomes large. For wave propagation in the vocal tract the Mach number is usually less than .02, so that nonlinearity of the wave is negligible. There are, however, two exceptions to this. The flow in the **glottis** (i.e., the space between the vocal folds), and that in the narrow constrictions used to produce fricative sounds, is nonlinear. We will show later how these special cases are handled in current speech production models.

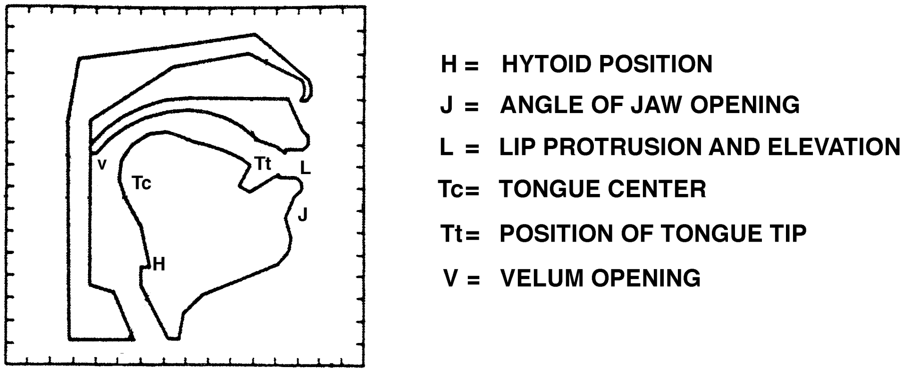

FIGURE 44.3: An idealized articulatory model similar to that of Coker [9].

We ought to point out that some computations have been made without the first two assumptions, and wave phenomena studied in two or three dimensions [12]. Recently there has been some interest in removing the third assumption as well [13]. This involves the solution of the so called **Navier-Stokes equation** in the complicated three-dimensional geometry of the vocal tract. Such analyses require very large amounts of high speed computations making it difficult to use them in speech production models. Computational cost and speed, however, are not the only limiting factors. An even more basic barrier is that it is difficult to specify accurately the complicated time-varying shape of the vocal tract. It is, therefore, unlikely that such computations can be used directly in a speech production model. These computations should, however, provide accurate data on the basis of which simpler, more tractable, approximations may be abstracted.

44.3.2 Wave Propagation in the Vocal Tract

In view of the assumptions discussed above, the propagation of waves in the vocal tract can be considered in the simplified setting depicted in Fig. 44.4. As shown there, the vocal tract is represented as a variable area tube of length L with its axis taken to be the $x-$axis. The glottis is located at $x = 0$ and the lips at $x = L$, and the tube has a cross-sectional area $A(x)$ which is a function of the distance x from the glottis. Strictly speaking, of course, the area is time-varying. However, in normal speech the temporal variation in the area is very slow in comparison with the propagation phenomena that we are considering. So, the cross-sectional area may be represented by a succession of stationary shapes.

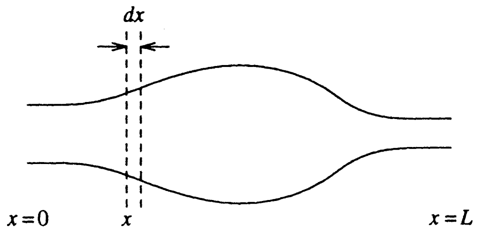

FIGURE 44.4: The vocal tract as a variable area tube.

We are interested in the spatial and temporal variation of two interrelated quantities in the acoustic wave: the pressure $p(x, t)$ and the volume velocity $u(x, t)$. The latter is $A(x)v(x, t)$, where v is the **particle** velocity. For the assumption of linearity to be valid, the pressure p in the acoustic wave is assumed to be small compared to the equilibrium pressure P_0, and the particle velocity v is assumed to be small compared to the velocity of sound, c. Two equations can be written down that relate $p(x, t)$ and $u(x, t)$: the equation of motion and the equation of continuity [14]. A combination of these equations will give us the basic equation of wave propagation in the variable area tube. Let us derive these equations first for the case when the walls of the tube are rigid and there are no losses due to viscous friction, thermal conduction, etc.

44.3.3 The Lossless Case

The **equation of motion** is just a statement of Newton's second law. Consider the thin slice of air between the planes at x and $x + dx$ shown in Fig. 44.4. By equating the net force acting on it due to the pressure gradient to the rate of change of momentum one gets

$$\frac{\partial p}{\partial x} = -\frac{\rho}{A}\frac{\partial u}{\partial t} \tag{44.1}$$

(To simplify notation, we will not always explicitly show the dependence of quantities on x and t.)

The **equation of continuity** expresses conservation of mass. Consider the slice of tube between x and $x + dx$ shown in Fig. 44.4. By balancing the net flow of air out of this region with a corresponding decrease in the density of air we get

$$\frac{\partial u}{\partial x} = -\frac{A}{\rho}\frac{\partial \delta}{\partial t} \, . \tag{44.2}$$

where $\delta(x, t)$ is the fluctuation in density superposed on the equilibrium density ρ. The density is related to pressure by the gas law. It can be shown that pressure fluctuations in an acoustic wave follow the adiabatic law, so that $p = (\gamma P/\rho)\delta$, where γ is the ratio of specific heats at constant pressure and constant volume. Also, $(\gamma P/\rho) = c^2$, where c is the velocity of sound. Substituting this into Eq. (44.2) gives

$$\frac{\partial u}{\partial x} = -\frac{A}{\rho c^2}\frac{\partial p}{\partial t} \tag{44.3}$$

Equations (44.1) and (44.3) are the two relations between p and u that we set out to derive. From these equations it is possible to eliminate u by subtracting $\frac{\partial}{\partial t}$ of Eq. (44.3) from $\frac{\partial}{\partial x}$ of Eq. (44.1). This gives

$$\frac{\partial}{\partial x} A \frac{\partial p}{\partial x} = \frac{A}{c^2}\frac{\partial^2 p}{\partial t^2} \, . \tag{44.4}$$

Equation (44.4) is known in the literature as **Webster's horn equation** [15]. It was first derived for computations of wave propagation in horns, hence the name. By eliminating p from Eqs. (44.1) and (44.3), one can also derive a single equation in u.

It is useful to write Eqs. (44.1), (44.3), and (44.4) in the frequency domain by taking Laplace transforms. Defining $P(x, s)$ and $U(x, s)$ as the Laplace transforms of $p(x, t)$ and $u(x, t)$, respectively, and remembering that $\frac{\partial}{\partial t} \to s$, we get:

$$\frac{dP}{dx} = -\frac{\rho s}{A} U \tag{44.1a}$$

$$\frac{dU}{dx} = -\frac{sA}{\rho c^2} P\psi \tag{44.3a}$$

and

$$\frac{d}{dx} A \frac{dP}{dx} = \frac{s^2}{c^2} AP\psi \tag{44.4a}$$

It is important to note that in deriving these equations we have retained only first order terms in the fluctuating quantities p and u. Inclusion of higher order terms gives rise to nonlinear equations of propagation. By and large these terms are quite negligible for wave propagation in the vocal tract. However, there is one second order term, neglected in Eq. (44.1), which becomes important in the description of flow through the narrow constriction of the glottis. In deriving Eq. (44.1) we neglected the fact that the slice of air to which the force is applied is moving away with the velocity v. When this effect is correctly taken into account, it turns out that there is an additional term $\rho v \frac{\partial v}{\partial x}$ appearing on the left hand side of that equation. The corrected form of Eq. (44.1) is

$$\frac{\partial}{\partial x}\left[p + \frac{\rho}{2}(u/A)^2\right] = -\rho \frac{d}{dt}\left[\frac{u}{A}\right] .\psi \tag{44.5}$$

The quantity $\frac{\rho}{2}(u/A)^2$ has the dimensions of pressure, and is known as the **Bernoulli pressure**. We will have occasion to use Eq. (44.5) when we discuss the motion of the vocal cords in the section on sources of excitation.

44.3.4 Inclusion of Losses

The equations derived in the previous section can be used to approximately derive the acoustical properties of the vocal tract. However, their accuracy can be considerably increased by including terms that approximately take account of the effect of viscous friction, thermal conduction, and yielding walls [16]. It is most convenient to introduce these effects in the frequency domain.

The effect of viscous friction can be approximated by modifying the equation of motion, Eq. (44.1a) as follows:

$$\frac{dP}{dx} = -\frac{\rho s}{A} U - R(x, s)U .\psi \tag{44.6}$$

Recall that Eq. (44.1a) states that the force applied per unit area equals the rate of change of momentum per unit area. The added term in Eq. (44.6) represents the viscous drag which reduces the force available to accelerate the air. The assumption that the drag is proportional to velocity can be approximately validated. The dependence of R on x and s can be modeled in various ways [16].

The effect of thermal conduction and yielding walls can be approximated by modifying the equation of continuity as follows:

$$\rho \frac{dU}{dx} = -\frac{A}{c^2} sP - Y(x, s)P\psi \tag{44.7}$$

Recall that the left hand side of Eq. (44.3a) represents net outflow of air in the longitudinal direction, which is balanced by an appropriate decrease in the density of air. The term added in Eq. (44.7) represents net outward volume velocity into the walls of the vocal tract. This velocity arises from (1) a temperature gradient perpendicular to the walls which is due to the thermal conduction by the walls, and (2) due to the yielding of the walls. Both these effects can be accounted for by appropriate choice of the function $Y(x, s)$, provided the walls can be assumed to be **locally reacting**. By that we mean that the motion of the wall at any point depends on the pressure at that point alone. Models for the function $Y(x, s)$ may be found in [16].

Finally, the lossy equivalent of Eq. (44.4a) is

$$\frac{d}{dx}\frac{A}{\rho s + AR}\frac{dP}{dx} = \left(\frac{As}{\rho c^2} + Y\right) P\ .\ \psi \tag{44.8}$$

44.3.5 Chain Matrices

All properties of linear wave propagation in the vocal tract can be derived from Eqs. (44.1a), (44.3a), (44.4a) or the corresponding Eqs. (44.6), (44.7), and (44.8) for the lossy tract. The most convenient way to derive these properties is in terms of **chain matrices**, which we now introduce.

Since Eq. (44.8) is a second order linear ordinary differential equation, its general solution can be written as a linear combination of two independent solutions, say $\phi(x, s)$ and $\Psi(x, s)$. Thus

$$P(x, s) = a\phi(x, s) + b\Psi(x, s)\psi \tag{44.9}$$

where a and b are, in general, functions of s. Hence, the pressure at the input of the tube ($x = 0$) and at the output ($x = L$) are linear combinations of a and b. The volume velocity corresponding to the pressure given in Eq. (44.9) is obtained from Eq. (44.6) to be

$$U(x, s) = -\frac{A}{\rho s + AR}\left[a d\phi/dx + b d\Psi/dx\right] .\psi \tag{44.10}$$

Thus, the input and output volume velocities are seen to be linear combinations of a and b. Eliminating the parameters a and b from these relationships shows that the input pressure and volume velocity are linear combinations of the corresponding output quantities. Thus, the relationship between the input and output quantities may be represented in terms of a 2×2 matrix as follows:

$$\begin{aligned} \begin{bmatrix} P_{\text{in}} \\ U_{\text{in}} \end{bmatrix} &= \begin{bmatrix} \mathbf{k}_{11} & \mathbf{k}_{12} \\ \mathbf{k}_{21} & \mathbf{k}_{22} \end{bmatrix} \begin{bmatrix} P_{\text{out}} \\ U_{\text{out}} \end{bmatrix} \\ &= \mathbf{K} \begin{bmatrix} P_{\text{out}} \\ U_{\text{out}} \end{bmatrix} . \end{aligned} \tag{44.11}$$

The matrix **K** is called a **chain matrix** or ABCD matrix [17]. Its entries depend on the values of ϕ and Ψ at $x = 0$ and $x = L$. For an arbitrarily specified area function $A(x)$ the functions ϕ and ψ are hard to find. However, for a **uniform** tube, i.e., a tube for which the area and the losses are independent of x, the solutions are very easy. For a uniform tube, Eq. (44.8) becomes

$$\frac{d^2P}{dx^2} = \sigma^2 P\psi \tag{44.12}$$

where σ is a function of s given by

$$\sigma^2 = (\rho s + AR)\left(\frac{s}{\rho c^2} + \frac{Y}{A}\right) .$$

Two independent solutions of Eq. (44.12) are well known to be $\cosh(\sigma x)$ and $\sinh(\sigma x)$, and a bit of algebra shows that the chain matrix for this case is

$$\mathbf{K} = \begin{bmatrix} \cosh(\sigma L)\psi & (1/\beta)\sinh(\sigma L) \\ \beta \sinh(\sigma L)\psi & \cosh(\sigma L) \end{bmatrix} \tag{44.13}$$

where

$$\beta = \sqrt{\left[Y + \frac{As}{\rho c^2}\right] / \left[R + \frac{\rho s}{A}\right]} .$$

For an arbitrary tract, one can utilize the simplicity of the chain matrix of a uniform tube by approximating the tract as a concatenation of N uniform sections of length $\Delta = L/N$. Now the output quantities of the ith section become the input quantities for the $i+1$st section. Therefore, if $\mathbf{K}_i$ is the chain matrix for the ith section, then the chain matrix for the variable-area tract is approximated by

$$\mathbf{K} = \mathbf{K}_1 \mathbf{K}_2 \cdots \mathbf{K}_N \ .\psi \tag{44.14}$$

This method can, of course, be used to relate the input-output quantities for any portion of the tract, not just the entire vocal tract. Later we shall need to find the input-output relations for various sections of the tract, for example, the tract from the glottis to the velum for nasal sounds, from the narrowest constriction to the lips for fricative sounds, etc.

As stated above, all linear properties of the vocal tract can be derived in terms of the entries of the chain matrix. Let us give several examples.

Let us associate the input with the glottal end, and the output with the lip end of the tract. Suppose the tract is terminated by the radiation impedance Z_R at the lips. Then, by definition, $P_{\text{out}} = Z_R U_{\text{out}}$. Substituting this in Eq. (44.11) gives

$$\begin{bmatrix} P_{\text{in}}/U_{\text{out}} \\ U_{\text{in}}/U_{\text{out}} \end{bmatrix} = \begin{bmatrix} \mathbf{k}_{11} & \mathbf{k}_{12} \\ \mathbf{k}_{21} & \mathbf{k}_{22} \end{bmatrix} \begin{bmatrix} Z_R \\ 1 \end{bmatrix} .\psi \tag{44.15}$$

From Eq. (44.15) it follows that

$$\frac{U_{\text{out}}}{U_{\text{in}}} = \frac{1}{\mathbf{k}_{21} Z_R + \mathbf{k}_{22}} .\psi \tag{44.16a}$$

Equation (44.16a) gives the **transfer function** relating the output volume velocity to the input volume velocity. Multiplying this by Z_R gives the transfer function relating output pressure to the input volume velocity. Other transfer functions relating output pressure or volume velocity to input pressure may be similarly derived.

Relationships between pressure and volume velocity at a single point may also be derived. For example,

$$\frac{P_{\text{in}}}{U_{\text{in}}} = \frac{\mathbf{k}_{11} Z_R + \mathbf{k}_{12}}{\mathbf{k}_{21} Z_R + \mathbf{k}_{22}} \tag{44.16b}$$

gives the **input impedance** of the vocal tract as seen at the glottis, when the lips are terminated by the radiation impedance.

Also, **formant frequencies**, which we mentioned in the Introduction, can be computed from the transfer function of Eq. (44.16a). They are just the values of s at which the denominator on the right-hand side becomes zero. For a lossy vocal tract, the zeros are complex and have the form $s_n = -\alpha_n + j\omega_n, n = 1, 2, \cdots$. Then ω_n is the frequency (in rad/s) of the nth formant, and α_n is its half bandwidth.

Finally, the chain matrix formulation also leads to **linear prediction coefficients** (LPC), which are the most commonly used representation of speech signals today. Strictly speaking, the representation is valid for speech signals for which the excitation source is at the glottis (i.e., voiced or aspirated speech sounds). Modifications are required when the source of excitation is at an interior point.

To derive the LPC formulation, we will assume the vocal tract to be lossless, and the radiation impedance at the lips to be zero. From Eq. (44.16a) we see that to compute the output volume velocity from the input volume velocity, we need only the $\mathbf{k}_{22}$ element of the chain matrix for the entire vocal tract. This chain matrix is obtained by a concatenation of matrices as shown in Eq. (44.14).

The individual matrices $\mathbf{K}_i$ are derived from Eq. (44.13), with $N = L/\Delta$. In the lossless case, R and Y are zero, so $\sigma = s/c$ and $\beta = A/\rho c$. Also, if we define $z = e^{2s\Delta/c}$, then the matrix $\mathbf{K}_i$ becomes

$$\mathbf{K}_i = z^{N/2} \begin{bmatrix} \frac{1}{2}\left(1 + z^{-1}\right) & \frac{A_i}{2\rho c}\left(1 - z^{-1}\right) \\ \\ \frac{\rho c}{2A_i}\left(1 - z^{-1}\right) & \frac{1}{2}\left(1 + z^{-1}\right) \end{bmatrix} .\psi \tag{44.17}$$

Clearly, therefore, $\mathbf{k}_{22}$ is $z^{N/2}$ times an Nth degree polynomial in z^{-1}. Hence, Eq. (44.16a) can be written as

$$\sum_{k=0}^{N} a_k z^{-k} U_{\text{out}} = z^{-N/2} U_{\text{in}} .\psi \tag{44.18}$$

where a_k are the coefficients of the polynomial. The frequency domain factor $z = e^{-2s\Delta/c}$ represents a delay of $2\Delta/c$s. Thus, the time domain equivalent of Eq. (44.18) is

$$\sum_{k=0}^{N} a_k u_{\text{out}}(t - 2k\Delta/c) = u_{\text{in}}(t - N\Delta/c) .\psi \tag{44.19}$$

Now $u_{\text{out}}(t)$ is the volume velocity in the speech signal, so we will call it $s(t)$ for brevity. Similarly, since $u_{\text{in}}(t)$ is the input signal at the glottis, we will call it $g(t)$. To get the time-sampled version of Eq. (44.19) we set $t = 2n\Delta/c$ and define $s(2n\Delta/c) = s_n$ and $g((2n - N)\Delta/c) = g_n$. Then Eq. (44.19) becomes

$$\sum_{k=0}^{N} a_k s_{n-k} = \varepsilon_n .\psi \tag{44.20}$$

Equation (44.20) is the LPC representation of a speech signal.

44.3.6 Nasal Coupling

Nasal sounds are produced by opening the velum and thereby coupling the nasal cavity to the vocal tract. In nasal consonants, the vocal tract itself is closed at some point between the velum and the lips, and all the airflow is diverted into the nostrils. In nasal vowels the vocal tract remains open. (Nasal vowels are common in French and several other languages. They are not nominally phonemes of English. However, some nasalization of vowels commonly occurs in English speech.)

In terms of chain matrices, the nasal coupling can be handled without too much additional effort. As far as its acoustical properties are concerned, the nasal cavity can be treated exactly like the vocal tract, with the added simplification that its shape may be regarded as fixed. The common assumption is that the nostrils are symmetric, in which case the cross-sectional areas of the two nostrils can be added and the nose replaced by a single, fixed, variable-area tube.

The description of the computations is easier to follow with the aid of the block diagram shown in Fig. 44.5. From a knowledge of the area functions and losses for the vocal and nasal tracts three chain matrices $\mathbf{K}_{gv}$, $\mathbf{K}_{vt}$, and $\mathbf{K}_{vn}$ are first computed. These represent, respectively, the matrices from glottis to velum, velum to tract closure (or velum to lips, in case of a nasal vowel), and velum to nostrils.

From $\mathbf{K}_{vn}$ with some assumed impedance termination at the nostrils, the input impedance of the nostrils at the velum may be computed as indicated in Eq. (44.16b). Similarly, $\mathbf{K}_{vt}$ gives the input impedance at the velum, of the vocal tract looking toward the lips. At the velum, these two impedances are combined in parallel to give a total impedance, say Z_v. With this as termination, the velocity to velocity transfer function, T_{gv}, from glottis to velum can be computed from $\mathbf{K}_{gv}$ as shown

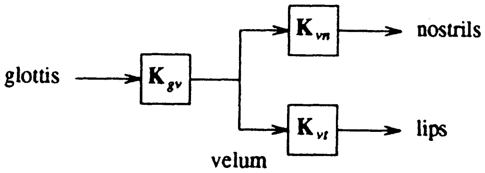

FIGURE 44.5: Chain matrices for synthesizing nasal sounds.

in Eq. (44.16b). For a given volume velocity at the glottis, U_g, the volume velocity at the velum is $U_v = T_{gv}U_g$, and the pressure at the velum is $P_v = Z_vU_v$. Once P_v and U_v are known, the volume velocity and/or pressure at the nostrils and lips can be computed by inverting the matrices $\mathbf{K}_{vn}$ and $\mathbf{K}_{vt}$.

44.4 Sources of Excitation

As mentioned earlier, speech sounds may be classified by type of excitation: periodic, turbulent, or transient. All of these types of excitation are created by converting the potential energy stored in the lungs due to excess pressure into sound energy in the audible frequency range of 20 Hz to 20 kHz.

The lungs of a young adult male may have a maximum usable volume (“vital capacity”) of about 5 l. While reading aloud the pressure in the lungs is typically in the range of 6 to 15 cm of water (6000 to 15000 Pa). Vocal cord vibrations can be sustained with a pressure as low as .2 cm of water. At the other extreme, a pressure as high as 195 cm of water has been recorded for a trumpet player. Typical average airflow for normal speech is about 0.1 l/s. It may peak as high as 5 l/s during rapid inhales in singing.

Periodic excitation originates mainly at the vibrating vocal folds, turbulent excitation originates primarily downstream of the narrowest constriction in the vocal tract, and transient excitations occur whenever a complete closure of the vocal pathway is suddenly released. In the following, we will explore these three types of excitation in some detail. The interested reader is referred to [18] for more information.

44.4.1 Periodic Excitation

Many of the acoustic and perceptual features of an individual’s voice are believed to be due to specific characteristics of the quasi-periodic excitation signal provided by the vocal folds. These, in turn, depend on the morphology of the voice organ, the **larynx**. The anatomy of the larynx is quite complicated, and descriptions of it may be found in the literature [19]. From an engineering point of view, however, it suffices to note that the larynx is the structure that houses the **vocal folds** whose vibration provides the periodic excitation. The space between the vocal folds, called the **glottis**, varies with the motion of the vocal folds, and thus modulates the flow of air through them. As late as 1950 Husson postulated that each movement of the folds is in fact induced by individual nerve signals sent from the brain (the Neurochronaxis hypothesis) [20]. We now know that the larynx is a self-oscillating acousto-mechanical oscillator. This oscillator is controlled by several groups of tiny muscles also housed in the larynx. Some of these muscles control the rest position of the folds, others control their tension, and still others control their shape. During breathing and production of fricatives, for example, the folds are pulled apart (abducted) to allow free flow of air. To produce voiced speech, the vocal folds are brought close together (adducted). When brought close enough together, they go into a spontaneous periodic oscillation. These oscillations are driven by Bernoulli pressure (the same mechanism that keeps airplanes aloft) created by the airflow through the glottis.

If the opening of the glottis is small enough, the Bernoulli pressure due to the rapid flow of air is large enough to pull the folds toward each other, eventually closing the glottis. This, of course, stops the flow and the laryngeal muscles pull the folds apart. This sequence repeats itself until the folds are pulled far enough away, or if the lung pressure becomes too low. We will discuss this oscillation in greater detail later in this section.

Besides the laryngeal muscles, the lung pressure and the acoustic load of the vocal tract also affect the oscillation of the vocal folds.

The larynx also houses many mechanoreceptors that signal to the brain the vibrational state of the vocal folds. These signals help control pitch, loudness, and voice timbre.

Figure 44.6 shows stylized snapshots taken from the side and above the vibrating folds. The view from above can be obtained on live subjects with high speed (or stroboscopic) photography, using a laryngeal mirror or a fiber optic bundle for illumination and viewing. The view from the side is the result of studies on excised (mostly animal) larynges. From studies such as these, we know that, during glottal vibration, the folds carry a mechanical wave that starts at the tracheal (lower) end of the folds and moves upwards to the pharyngeal (upper) end. Consequently, the edge of the folds that faces the vocal tract usually lags behind the edge of the folds that faces the lungs. This phenomenon is called **vertical phasing**. Higher eigenmodes of these mechanical waves have been observed and have been modeled.

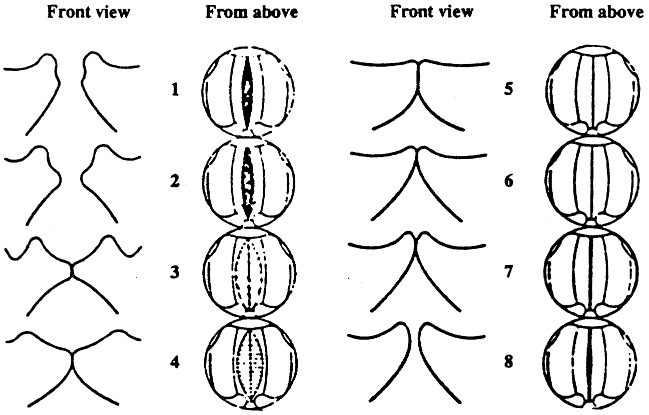

FIGURE 44.6: One cycle of vocal fold oscillation seen from the front and from above. (After Schönhärl, E., 1960 [25]. With permission of Georg Thieme Verlag, Stuttgart, Germany.)

Figure 44.7 shows typical acoustic flow waveforms, called **flow glottograms**, and their first time derivatives. In a normal glottogram, the **closed phase** of the glottal cycle is characterized by zero flow. Often, however, the closure is not complete. Also, in some cases, although the folds close completely, there is a parallel path — a chink — which stays open all the time.

In the **open phase** the flow gradually builds up, reaches a peak, and then falls sharply. The asymmetry is due to the inertia of the airflow in the vocal tract and the sub-glottal cavities. The amplitude of the fundamental frequency is governed mainly by the peak of the flow while the amplitudes of the higher harmonics is governed mainly by the (negative) peak rate of change of flow, which occurs just before closure.

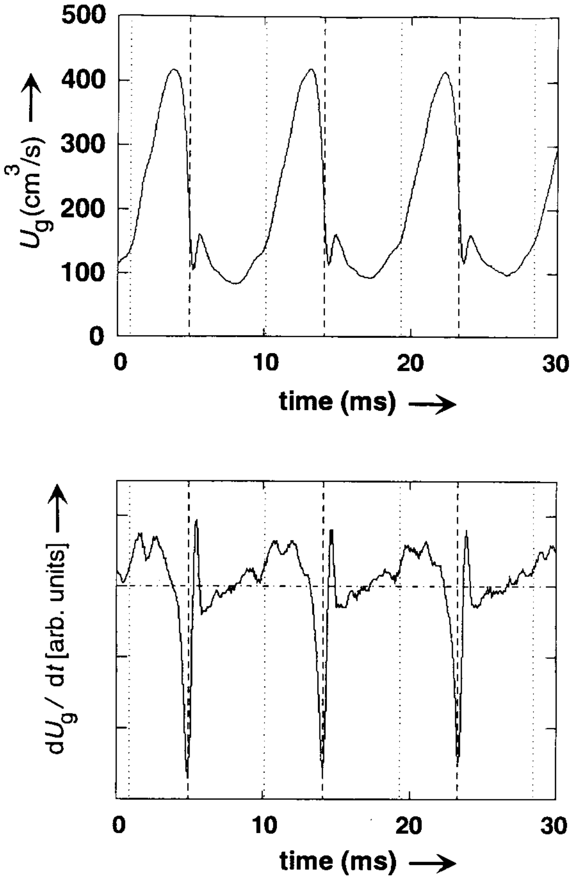

FIGURE 44.7: Example of glottal volume velocity and its time derivative.

Voice Qualities

Depending on the adjustment of the various parameters mentioned above, the glottis can produce a variety of phonations (i.e., excitations for voiced speech), resulting in different perceptual voice qualities. Some perceptual qualities vary continuously whereas others are essentially categorical (i.e., they change abruptly when some parameters cross a threshold).

Voice timbre is an important continuously variable quality which may be given various labels ranging from "mellow" to "pressed". The spectral slope of the glottal waveform is the main physical correlate of this perceptual quality. On the other hand, nasality and aspiration may be regarded as categorical qualities.

The physical properties that distinguish a "male" voice from a "female" voice are still not well understood, although many distinguishing features are known. Besides the obvious cue of fundamental frequency, the perceptual quality of "breathiness" seems to be important for producing a female-sounding voice. It occurs when the glottis does not close completely during the glottal cycle. This results in a more sinusoidal movement of the folds which makes the amplitude of the fundamental frequency much larger compared to those of the higher harmonics. The presence of leakage in the abducted glottis also increases the damping of the lower formants, thus increasing their bandwidths. Also, the continuous airflow through the leaking glottis gives rise to increased levels of glottal noise (aspiration noise) that masks the higher harmonics of the glottal spectrum. Finally, in glottograms of female voices, the open phase is a larger proportion of the glottal cycle (about 80%) than in glottograms of male voices (about 60%). The points of closure are also smoother for female voices, which results in lower high frequency energy relative to the fundamental.

Finally, the individuality of a voice (which allows us to recognize the speaker) appears to be dependent largely on the exact relationships between the amplitudes of the first few harmonics.

Models of the Glottis

A study of the mechanical and acoustical properties of the larynx is still an area of active interdisciplinary research. Modeling in the mechanical and acoustical domains requires making simplifying assumptions about the tissue movements and the fluid mechanics of the airflow. Depending on the degree to which the models incorporate physiological knowledge, one can distinguish three categories of glottal models:

Parametrization of glottal flow is the "black-box" approach to glottal modeling. The glottal flow wave or its first time derivative is parametrized in segments by analytical functions. It seems doubtful that any simple model of this kind can match all kinds of speakers and speaking styles. Examples of speech sounds that are difficult to parametrize in this way are nasal and mixed-excitation sounds (i.e., sounds with an added fricative component) and "simple" high-pitch female vowels.

Parametrization of glottal area is more realistic. In this model, the area of the glottal opening is parametrized in segments, but the airflow is computed from the propagation equations, and includes its interaction with the acoustic loads of the vocal tract and the subglottal structures. Such a model is capable of reproducing much more of the detail and individuality of the glottal wave than the black box approach. Problems are still to be expected for mixed glottal/fricative sounds unless the tract model includes an accurate mechanism for frication (see the section on turbulent excitation below).

In a complete, self-oscillating model of the glottis described below, the amplitude of the glottal opening as well as the instants of glottal closure are automatically derived, and depend in a complicated manner on the laryngeal parameters, lung pressure, and the past history of the flow. The area-driven model has the disadvantage that amplitude and instants of closure must be specified as side information. However, the ability to specify the points of glottal closure can, in fact, be an advantage in some applications; for example, when the model is used to mimic a given speech signal.

Self-oscillating physiological models of the glottis attempt to model the complete interaction of the airflow and the vocal folds which results in periodic excitation. The input to a model of this type is slowly varying physical parameters such as lung pressure, tension of the folds, pre-phonatory glottal shape, etc. Of the many models of this type that have been proposed, the one most often used is the 2-mass model of Ishizaka and Flanagan (I&F). In the following we will briefly review this model.

The I&F two-mass model is depicted in Fig. 44.8. As shown there, the thickness of the vocal folds that separates the trachea from the vocal tract is divided into two parts of length d_1 and d_2, respectively, where the subscript 1 refers to the part closest to the trachea and 2 refers to the part closest to the vocal tract. These portions of the vocal folds are represented by damped spring-mass systems coupled to each other. The division into two portions is a refinement of an earlier version that represented the folds by a single spring-mass system. By using two sections the model comes closer to reality and exhibits the phenomenon of vertical phasing mentioned earlier.

In order to simulate tissue, all the springs and dampers are chosen to be nonlinear. Before discussing the choice of these nonlinear elements, let us first consider the relationship between the airflow and the pressure variations from the lungs to the vocal tract.

Airflow in the Glottis

The dimensions d_1 and d_2 are very small — about 1.5 mm each. This is a very small fraction of the wavelength even at the highest frequencies of interest. (The wavelength of a sound wave in air at 100 kHz is about 3 mm!). Therefore we may assume the flow through the glottis to be incompressible.

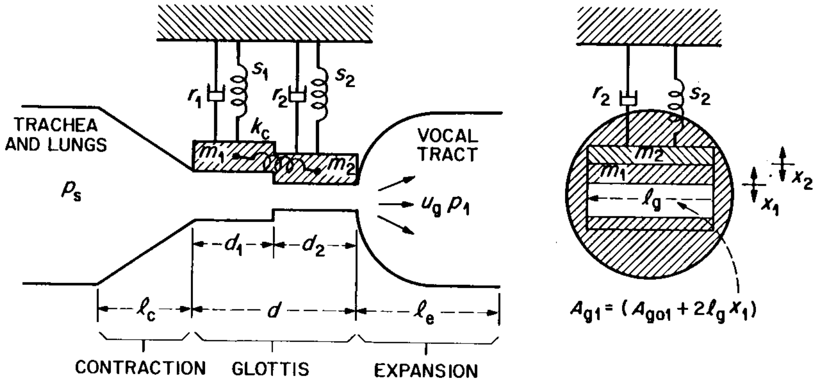

FIGURE 44.8: The two-mass model of Ishizaka and Flanagan [21].

With this assumption the equation of continuity, Eq. (44.2), merely states that the volume velocity is the same everywhere in the glottis. We will call this volume velocity u_g. The relationship of this velocity to the pressure is governed by the equation of motion. Since the particle velocity in the glottis can be very large, we need to consider the nonlinear version given in Eq. (44.5). Also, since the cross-section of the glottis is very small, viscous drag cannot be neglected. So we will include a term representing viscous drag proportional to the velocity. With this addition, Eq. (44.5) becomes:

$$\frac{\partial}{\partial x}\left[p + \frac{\rho}{2}\left(u_g/A\right)^2\right] = -\rho\frac{\partial}{\partial t}\left(\frac{u_g}{A}\right) - R_v\left(u_g/A\right) \ . \tag{44.21}$$

The drag coefficient R_v can be estimated for simple geometries. In the present application a rectangular aperture is appropriate. If the length of the aperture is l, its width (corresponding to the opening between the folds) is w and its depth in the direction of flow is d, then $R_v = \frac{12\mu d}{lw^3}$, where μ is the coefficient of shear viscosity. The pressure distribution is obtained by repeated use of Eq. (44.21), using the appropriate value of A (and hence of R_v) in the different parts of the glottis. In this manner, the pressure at any point in the glottis may be determined in terms of the volume velocity, u_g, the lung pressure, P_s, and the pressure at the input to the vocal tract, p_1.

The detailed derivation of the pressure distribution is given in [21]. The derivation shows that the total pressure drop across the glottis, $P_s - p_1$, is related to the glottal volume velocity, u_g, by an equation of the form

$$P_s - p_1 = Ru_g + \frac{d}{dt}(Lu_g) + \frac{\rho}{2}\left(u_g/\alpha\right)^2 \ . \tag{44.22}$$

With the analogy of pressure to voltage and volume velocity to current, the quantity R is analogous to resistance and L to inductance. The term in u_g^2 may be regarded as u_g times a current-dependent resistance. The quantity α has the dimensions of an area.

Models of Vocal Fold Tissue

When the pressure distribution derived above is coupled to the mechanical properties of the vocal folds, we get a self-oscillating system with properties quite similar to those of a real larynx. The mechanical properties of the vocal folds have been modeled in many ways with varying degrees of complexity ranging from a single spring-mass system to a distributed parameter flexible tube. In the following, by way of example, we will summarize only the original 1972 I&F model.

Returning to Fig. 44.8, we observe that the mechanical properties of the folds are represented by the masses m_1 and m_2, the (nonlinear) springs s_1 and s_2, the coupling spring k_c, and the nonlinear dampers r_1 and r_2. The opening in each section of the glottis is assumed to have a rectangular shape with length l_g. The widths of the two sections are $2x_j$, $j = 1, 2$. Assuming a symmetrical glottis, the cross-sectional areas of the two sections are

$$A_{gj} = A_{g0j} + 2l_g x_j , \quad j = 1, 2 , \tag{44.23}$$

where A_{g01} and A_{g02} are the areas at rest. From this equation, we compute the lateral displacements $x_{j\,\min}$, $j = 1, 2$ at which the two folds touch each other in each section to be $x_{j\,\min} = -A_{g0j}/(2l_g)$. Displacements more negative than these indicate a collision of the folds. The springs s_1 and s_2 are assumed to have restoring forces of the form $ax + bx^3$, where the constants a and b take on different values for the two sections and for the colliding and non-colliding conditions.

The dampers r_1 and r_2 are assumed to be linear, but with different values in the colliding and non-colliding cases. The coupling spring k_c is assumed to be linear. With these choices, the coupled equations of motion for the two masses are:

$$m_1 \frac{d^2 x_1}{dt^2} + r_1 \frac{dx_1}{dt} \quad + \quad f_{s1}(x_1) + k_c(x_1 - x_2) = F_1 , \tag{44.24a}$$

and

$$m_2 \frac{d^2 x_2}{dt^2} + r_2 \frac{dx_2}{dt} \quad + \quad f_{s2}(x_2) + k_c(x_2 - x_1) = F_2 . \tag{44.24b}$$

Here f_{s1} and f_{s2} are the cubic nonlinear springs. The parameters of these springs as well as the damping constants r_1 and r_2 change when the folds go from a colliding state to a non-colliding state and vice versa. The driving forces F_1 and F_2 are proportional to the average acoustic pressures in the two sections of the glottis. Whenever a section is closed (due to the collision of its sides) the corresponding driving force is zero. Note that it is these forces that provide the feedback of the acoustic pressures to the mechanical system. This feedback is ignored in the area-driven models of the glottis.

We close this section with an example of ongoing research in glottal modeling. In the introduction to this section we had stated that breathiness of a voice is considered important for producing a natural-sounding synthetic female voice. Breathiness results from incomplete closures of the folds. We had also stated that incomplete glottal closures due to abducted folds lead to a steep spectral roll-off of the glottal excitation and a strong fundamental. However, practical experience shows that many voices show clear evidence for breathiness but do not show a steep spectral roll-off, and have relatively weak fundamentals instead. How can this mystery be solved? It has been suggested that the glottal "chink" mentioned in the discussion of Fig. 44.7 might be the answer. Many high-speed videos of the vocal folds show evidence of a separate leakage path in the "posterior commissure" (where the folds join) which stays open all the time. Analysis of such a permanently open path produces the stated effect [22].

44.4.2 Turbulent Excitation

Turbulent airflow shows highly irregular fluctuations of particle velocity and pressure. These fluctuations are audible as broadband noise. Turbulent excitation occurs mainly at two locations in the vocal tract: near the glottis and at constriction(s) between the glottis and the lips. Turbulent excitation at a constriction downstream of the glottis produces fricative sounds or voiced fricatives depending on whether or not voicing is simultaneously present. Also, stressed versions of the vowel *i*, and liquids *l* and *r* are usually accompanied by turbulent flow. Measurements and models for turbulent excitation are even more difficult to establish than for the periodic excitation produced by the glottis because, usually, no vibrating surfaces are involved. Because of the lack of a comprehensive model, much confusion exists over the proper sub-classification of fricatives. The simplest model for turbulent excitation is a "nozzle" (narrow orifice) releasing air into free space. Experimental work has shown that half (or more) of the noise power generated by a jet of air originates within the so-called mixing region that starts at the nozzle outlet and extends as far as a distance four times the diameter of the orifice. The noise source is therefore distributed. Several scaling relations hold between the acoustic output and the nozzle geometry. One of these scaling properties is the so-called **Reynolds number**, Re, that characterizes the amount of turbulence generated as the air from the jet mixes with the ambient air downstream from the orifice:

$$Re = \frac{u}{A}\frac{x}{\nu} \ . \tag{44.25}$$

Here u is the volume velocity, A is the area of the orifice (hence, u/A is the particle velocity), x is a characteristic dimension of the orifice (the width for a rectangular orifice), and $\nu = \mu/\rho$ is the kinematic viscosity of air. Beyond a critical value of the Reynolds number, Re_{crit} (which is about 1200 for the case of a free jet), the flow becomes fully turbulent; below this value, the flow is partly turbulent and becomes fully laminar at very low velocities. Another scaling equation defines the so-called **Strouhal number**, S, that relates the frequency $F_{\max}$ of the (usually broad) peak in the power spectrum of the generated noise to the width of the orifice and the velocity:

$$S = F_{\max}\frac{x}{u/A} \ . \tag{44.26}$$

For the case of a free jet, the Strouhal number S is 0.15. Within the jet, higher frequencies are generated closer to the orifice and lower frequencies further away.

Distributed sources of turbulence can be modeled by expanding them in terms of monopoles (i.e., pulsating spheres), dipoles (two pulsating spheres in opposite phase), quadrupoles (two dipoles in opposite phase), and higher-order representations. The total power generated by a monopole source in free space is proportional to the fourth power of the particle velocity of the flow, that of a dipole source obeys a $(u/A)^6$ power law, and that of a quadrupole source obeys a $(u/A)^8$ power law. Thus, the low order sources are more important at low flow rates, while the reverse is the case at high flow rates. In a duct, however, the exponents of the power laws decrease by 2, that is, a dipole source's noise power is proportional to $(u/A)^4$, etc.

Thus far, we have summarized noise generation in a free jet or air. A much stronger noise source is created when a jet of air hits an obstacle. Depending on the angle between the surface of the obstacle and the direction of flow, the surface roughness, and the obstacle geometry, the noise generated can be up to 20 dB higher than that generated by the same jet in free space. Because of the spatially concentrated source, modeling obstacle noise is easier than modeling the noise in a free jet. Experiments reveal that obstacle noise can be approximated by a dipole source located at the obstacle.

The above theoretical findings qualitatively explain the observed phenomenon that the fricatives th and f (and the corresponding voiced dh and v) are weak compared to the fricatives s and sh. The teeth (upper for s and lower for sh) provide the obstacle on which the jet impinges to produce the higher noise levels. A fricative of intermediate strength results from a **distributed** obstacle (the "wall" case) when the jet is forced along the roof of the mouth as for the sound y.

In a synthesizer, dipole noise sources can be implemented as series pressure sources. One possible implementation is to make the source pressure proportional to $Re^2 - {Re_{\mathrm{crit}}}^2$ for $Re > Re_{\mathrm{crit}}$ and zero otherwise [11]. Another option [23] is to relate the noise source power to the Bernoulli pressure $B = .5\rho(u/A)^2$. Since the power of a dipole source located at the teeth (and radiating into free space) is $(u/A)^6$, it is also proportional to B^3, and the noise source pressure $p_n \propto B^{3/2}$. On the other hand, for wall sources located further away from the lips, we need multiple (distributed) dipole sources with source pressures proportional either to $Re^2 - {Re_{\mathrm{crit}}}^2$ or to B. In either case, the source should have a broadband spectrum with a peak at a frequency given by Eq. (44.26).

When a noise source is located at some point inside the tract, its effect on the acoustic output at the lips is computed in terms of two chain matrices — the matrix $\mathbf{K}_F$ from the glottis to the noise source, and the matrix $\mathbf{K}_L$ from the noise source to the lips. For fricative sounds, the glottis is wide open, so the termination impedance at the glottis end may be assumed to be zero. With this termination, the impedance at the noise source looking toward the glottis is computed from $\mathbf{K}_F$ as explained in the section on chain matrices. Call this impedance Z_1. Similarly, a knowledge of the radiation impedance at the lips and the matrix $\mathbf{K}_L$ allows us to compute the input impedance Z_2 looking toward the lips. The volume velocity at the source is then just $P_n/(Z_1 + Z_2)$ where P_n is the pressure generated by the noise source. The transfer function obtained from Eq. (44.16a) for the matrix $\mathbf{K}_L$ then gives the volume velocity at the lips.

It can be shown that the series noise source P_n excites all formants of the entire tract (i.e., the ones we would see if the source were at the glottis). However, the spectrum of fricative noise usually has a high pass character. This can be understood qualitatively by the following considerations.

When the tract has a very narrow constriction, the front and back cavities are essentially decoupled, and the formants of the tract are the formants of the back cavity plus those of the front cavity. If now the noise source is just downstream of the constriction, the formants of the back cavity are only slightly excited because the impedance Z_1 also has poles at those frequencies. Since the back cavity is usually much longer than the front cavity for fricatives, the lower formants are missing in the velocity at the lips. This gives it a high pass character.

44.4.3 Transient Excitation

Transient excitation of the vocal tract occurs whenever pressure is built up behind a total closure of the tract and suddenly released. This sudden release produces a step-function of input pressure at the point of release. The output velocity is therefore proportional to the integral of the impulse response of the tract from the point of release to the lips. In the frequency domain, this is just P_r/s times the transfer function, where P_r is the step change in pressure. Hence, the velocity at the lips may be computed in the same way as in the case of turbulent excitation, with P_n replaced by P_r/s. In practice, this step excitation is usually followed by the generation of fricative noise for a short period after release when the constriction is still narrow enough. Sometimes, if the glottis is also being constricted (e.g., to start voicing) some aspiration might also result.

44.5 Digital Implementations

The models of the various parts of the human speech production apparatus which we have described above can be assembled to produce fluent speech. Here we will consider how a digital implementation of this process may be carried out. Basically, the standard theory of sampling in the time and frequency domains is used to convert the continuous signals considered above to sampled signals, and the samples are represented digitally to the desired number of bits per sample.

44.5.1 Specification of Parameters

The parameters that drive the synthesizer need to be specified about every 20 ms. (The assumed quasi-stationarity is valid over durations of this size.)

Two sets of parameters are needed — the parameters that specify the shape of the vocal tract and those that control the glottis. The vocal tract parameters implicitly control nasality (by specifying the opening area of the velum) and also frication (by specifying the size of the narrowest constriction).

44.5.2 Synthesis

The vocal tract is approximated by a concatenation of about 20 uniform sections. The cross-sectional areas of these sections is either specified directly, or computed from a specification of articulatory parameters as shown in Fig. 44.3. The chain matrix for each section is computed at an adequate sampling rate in the frequency domain to avoid time-aliasing of the corresponding time functions. (Computation of the chain matrices requires a specification of the losses also. Several models exist which assign the losses in terms of the cross-sectional area [11, 16]).

The chain matrices for the individual sections are combined to derive the matrices for various portions of the tract, as appropriate for the particular speech sound being synthesized. For voiced sounds, the matrices for the sections from the glottis to the lips are sequentially multiplied to give the matrix from the glottis to the lips. From the $\mathbf{k}_{11}, \mathbf{k}_{12}, \mathbf{k}_{21}, \mathbf{k}_{22}$ components of this matrix, the transfer function $\frac{U_{\text{out}}}{U_{\text{in}}}$ and the input impedance are obtained as in Eqs. (44.16a) and (44.16b). Knowing the radiation impedance Z_R at the lips we can compute the transfer function for output pressure, $H = \frac{U_{\text{out}}}{U_{\text{in}}} Z_R$. The inverse FFT of the transfer function H and the input impedance Z_{in} give the corresponding time functions $h(n)$ and $z_{\text{in}}(n)$, respectively. These functions are computed every 20 ms, and the intermediate values are obtained by linear interpolation.

For the current time sampling instant n, the current pressure $p_1(n)$ at the input to the vocal tract is then computed by convolving z_{in} with the past values of the glottal volume velocity u_g. With p_1 known, the pressure difference $P_s - p_1$ on the left hand side of Eq. (44.22) is known. Equation (44.18) is discretized by using a backward difference for the time derivative. Thus, a new value of the glottal volume velocity is derived. This, together with the current values of the displacements of the vocal folds, gives us new values for the driving forces F_1 and F_2 for the coupled oscillator Eqs. (44.24a) and (44.24b). The coupled oscillator equations are also discretized by backward differences for time derivatives. Thus, the new values of the driving forces give new values for the displacements of the vocal folds. The new value of volume velocity also gives a new value for p_1, and the computational cycle repeats, to give successive samples of p_1, u_g, and the vocal fold displacements.

The glottal volume velocity obtained in this way, is convolved with the impulse response $h(n)$ to produce voiced speech.

If the speech sound calls for frication, the chain matrix of the tract is derived as the product of two matrices — from the glottis to the narrowest constriction and from the constriction to the lips, as discussed in the section on turbulent excitation. This enables us to compute the volume velocity at the constriction, and thus introduce a noise source on the basis of the Reynolds number.

Finally, to produce nasal sounds, the chain matrix for the nasal tract is also computed, and the output at the nostrils computed as discussed in the section on chain matrices. If the lips are open, the output from the lips is also computed and added to the output from the nostrils to give the total speech signal. Details of the synthesis procedure may be found in [24].

References

[1] Edwards, H.T., *Applied Phonetics: The Sounds of American English*, Singular Publishing Group, San Diego, 1992, Chap. 3.

[2] Olive, J.P., Greenwood, A., and Coleman, J., *Acoustics of American English Speech*, Springer Verlag, New York, 1993.

[3] Fant, G., *Acoustic Theory of Speech Production*, Mouton Book Co., Gravenhage, 1960, Chap. 2.1, 93-95.

[4] Baer, T., Gore, J.C., Gracco, L.C., and Nye, P.W., Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, *J. Acoust. Soc. Am.*, 90 (2),799-828, Aug 1991.

[5] Stone, M., A three-dimensional model of tongue movement based on ultrasound and microbeam data, *J. Acoust. Soc. Am.*, 87 (5), 2207-2217, May 1990.

[6] Sondhi, M.M. and Resnick, J.R., The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis, *J. Acoust. Soc. Am.*, 73 (3), 985-1002, March 1983.

[7] Hardcastle, W.J., Jones, W., Knight, C., Trudgeon, A., and Calder, G., New developments in electropalatography: A state of the art report, *Clinical Linguistics and Phonetics*, 3, 1-38, 1989.

[8] Perkell, J.S., Cohen, M.H., Svirsky, M.A., Mathies, M.L., Garabieta, I., and Jackson, M.T.T., Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements, *J. Acoust. Soc. Am.*, 92 (6), 3078-3096, Dec 1992.

[9] Coker, C.H., A model of articulatory dynamics and control, *Proc. IEEE*, 64 (4), 452-460, April 1976.

[10] Sondhi, M.M., Resonances of a bent vocal tract, *J. Acoust. Soc. Am.*, 79 (4), 1113-1116, April 1986.

[11] Flanagan, J.L., *Speech Analysis, Synthesis and Perception,* 2nd ed., Springer Verlag, New York, 1972, Chap. 3.

[12] Lu, C., Nakai, T., and Suzuki, H., Three-dimensional FEM simulation of the effects of the vocal tract shape on the transfer function, *Intl. Conf. on Spoken Lang. Processing,* Banff, Alberta, 1, 771-774, 1992.

[13] Richard, G., Liu, M., Sinder, D., Duncan, H., Lin, O., Flanagan, J.L., Levinson, S.E., Davis, D.W. and Slimon, S., Numerical simulations of fluid flow in the vocal tract, *Proc. Eurospeech '95, European Speech Comm. Assoc.*, Madrid, Spain, 18-21, Sept. 1995.

[14] Morse, P.M., *Vibration and Sound,* McGraw Hill, New York, 1948, Chap. 6.

[15] Pierce, A.D., *Acoustics,* 2nd ed., McGraw-Hill, 360, 1981.

[16] Sondhi, M.M., Model for wave propagation in a lossy vocal tract, *J. Acoust. Soc. Am.*, 55 (5), 1070-1075, May 1974.

[17] Siebert, W. McC., *Circuits, Signals and Systems,* MIT Press/McGraw-Hill, pp. 97, 1986.

[18] Sundberg, J., *The Science of the Singing Voice,* Northern Illinois University Press, DeKalb, IL, 1987.

[19] Zemlin, W.R., *Speech and Hearing Science, Anatomy, and Physiology,* Prentice-Hall, Englewood Cliffs, NJ, 1968.

[20] Husson, R., Etude des phénomenes physiologiques et acoustiques fondamentaux de la voix cantée, Disp edit Rev Scientifique, 1-91, 1950. For a discussion see Diehl, C.F., *Introduction to the anatomy and physiology of the speech mechanisms,* Charles C Thomas, Springfield, IL, 110-111, 1968.

[21] Ishizaka, K. and Flanagan, J.L., Synthesis of voiced sounds from a two-mass model of the vocal cords, *Bell System Tech. J.,* 51 (6), 1233-1268, July-Aug. 1972.

[22] Cranen, B. and Schroeter, J., Modeling a leaky glottis, *J. Phonetics*, 23, 165-177, 1995.

[23] Stevens, K.N., Airflow and turbulence noise for fricative and stop consonants: Static considerations, *J. Acoust. Soc. Am.,* 50 (4), 1180-1192, 1971.

[24] Sondhi, M.M. and Schroeter, J., A hybrid time-frequency domain articulatory speech synthesizer, *IEEE Trans. on Acous., Speech, and Sig. Proc.*, ASSP-35 (7), 955-967, July 1987.

[25] Schönhärl, E., *Die Stroboskopie in der praktischen Laryngologie,* Georg Thieme Verlag, Stuttgart, Germany, 1960.