# A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer

MAN MOHAN SONDHI AND JUERGEN SCHROETER, MEMBER, IEEE

*Abstract*—**High quality speech at low bit rates (e.g., 2400 bits/s) is one of the important objectives of current speech research. As part of long range activity on this problem, we have developed an efficient computer program that will serve as a tool for investigating whether articulatory speech synthesis may achieve this low bit rate. At a sampling frequency of 8 kHz, the most comprehensive version of the program, including nasality and frication, runs at about twice real time on a Cray-1 computer.**

## Introduction

Low bit rate coding of speech (Flanagan *et al.* [1]) is an important objective of current speech research. There are essentially three different methods used for speech synthesis from low bit rate data (e.g., Flanagan [2] and Linggard [3]): formant synthesis, synthesis from linear prediction coefficients (LPC), and articulatory speech synthesis. Formant synthesis models the spectrum of speech while linear prediction models the signal waveform using correlation techniques. For these methods there exist accompanying analysis procedures for obtaining the low bit rate data directly from the speech input.

Articulatory synthesis models the speech production mechanism directly. At present, however, there is no appropriate analysis procedure which satisfactorily solves the related "inverse" problem of obtaining articulatory parameters from spoken utterances (for preliminary attempts see, e.g., Atal and Hanauer [4], Wakita [5], Atal *et al.* [6], Flanagan *et al.* [7], Levinson and Schmidt [8], and Kuc *et al.* [9]). Despite this drawback, articulatory speech synthesis has several advantages as follows.

a) Articulatory speech synthesis has the potential for very natural speech output at bit rates below 4800 bits/s, provided that "good" articulatory parameters are available to control the synthesizer.

b) The control signals of articulatory speech synthesizers have a direct interpretation in terms of physiological and physical data. In the human voice production system, they vary slowly enough to be potential candidates for efficient coding.

c) The model parameters are easier to interpolate than those of more abstract waveform or spectrum synthesizers. This is because interpolated values for the control signals of an articulatory synthesizer are physically realizable. (This is not true in general. An LPC vector interpolated between two realizable vectors might correspond to an unstable filter; interpolation of a set of formants between two reasonable sets of formants might yield a set that corresponds to an unreasonable, if not impossible, vocal tract shape, etc.) For the same reason, slightly erroneous control signals usually do *not* result in "unnatural" speech.

In this paper we will describe recent efforts to develop a *hybrid* articulatory speech synthesizer. This approach takes advantage of both frequency and time domain techniques in order to obtain a fast and versatile realization.

In the synthesizer which we will describe, the wave propagation in the tract is assumed to be planar and linear. During the voiced portions of speech, however, the excitation of the tract is provided by a nonlinear model of the vocal cord oscillator (Ishizaka and Flanagan [10]) which is controlled by lung pressure, glottal rest area, intraglottal damping, pitch factor, and supraglottal pressure. It is the dependence on the supraglottal pressure which differentiates this type of synthesizer from simpler source-filter approaches. This dependence maintains a natural acoustic interaction between glottal source and the load provided by the vocal tract.

The unvoiced portions of speech, both aspiration and frication, are generated automatically by introducing noise sources at the glottis and downstream from the narrowest constriction of the vocal tract, respectively. The strengths of these sources depend on the local Reynolds number (e.g., [2, p. 55]).

Since the vocal tract is assumed to be linear, it is most efficiently modeled in the frequency domain by a product of $2 \times 2$ chain matrices (also called ABCD matrices). The elements of these matrices are specified to include wall vibration, viscous friction, etc. (Sondhi [11], [12]). The nasal tract, including the sinus cavities, is modeled similarly, only the parameters are modified in view of the fact that the ratio of its perimeter to its cross-sectional area is higher than the average value of this ratio for the vocal tract. The glottis, however, must be modeled in the time domain.

In order to combine these two descriptions, the vocal/nasal tract chain matrices are used to compute the input reflectance of the tract at the glottis end, and the transfer functions from glottal flow to the radiated sound pressure at the lips and nostrils. The corresponding impulse responses, calculated by inverse Fourier transforms, are convolved with the glottal flow to yield supraglottal pres-

Manuscript received March 20, 1986; revised July 8, 1986. The sequence of names was decided on by a coin toss.

The authors are with AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 8714399.

sure, flow at the narrowest constriction, and the voiced portion of the output speech. Whenever the flow at the constriction is large enough to generate frication, the unvoiced portion of the speech output is obtained by convolving the noise at the constriction with the appropriate transfer impulse response computed in a similar manner. We also have provision for adding to the speech output the sound radiated by the vibration of the vocal tract wall near the glottis. The effect of this component, however, is very small.

At present, the synthesizer is driven either by interactively generated vocal tract areas (from Bocchieri [13]), by measured area data (Sondhi and Resnick [14]), or by parameters generated by a text-to-speech transcription program (Coker [15]). Such data are adequate for text-to-speech synthesis. For low bit rate speech *transmission* applications, however, we must derive the articulatory parameters from the speech wave. One (not entirely satisfactory) method which we plan to try is to use LPC derived "pseudoareas" [4], [5]. Another possibility is the idea of adaptive estimation of these parameters by matching synthetic speech to natural speech [7]. We believe experimentation with our synthesizer will suggest better methods.

## I. Articulatory Speech Synthesis

An overall outline of an articulatory speech synthesizer is given in Fig. 1. Articulatory speech synthesizers differ from each other in the ways in which they model the two parts—the glottis and the vocal tract.

Besides implementing an electrical hardware network analog (e.g., Hecker [16]), there are mainly two different approaches used in articulatory speech synthesizers. The first approach is to model the glottal source *and* the vocal tract by finite difference equations. This results in a large system of linear or nonlinear equations to be solved for each sampling interval. Examples of this approach are to be found in [10], [13], [17], [18], and [19]. Unfortunately, this method is computationally very cumbersome. (For example, Bocchieri [13, p. 52], reports a computer time of 5 h for 1 s of speech on a Data General Eclipse S/130.)

Kelly and Lochbaum [20] presented a much faster method. This method was later theoretically substantiated by Fettweis [21], and is now called the wave digital filter method. It is based on forward and backward traveling waves in a digital line analog of the vocal tract, and can be conveniently realized in special hardware [22]. Recently, progress has been made in incorporating losses, in modeling of the glottis, and in taking into account the time-varying vocal tract area ([23]–[26]).

## II. The Hybrid Method

The hybrid method differs from both of these approaches. While the glottis is modeled in the time domain because of its highly nonlinear nature, the vocal and nasal tracts are modeled in the frequency domain, taking advantage of the more convenient and accurate modeling of losses and radiation. The two models are then interfaced by inverse Fourier transformation and digital convolution. At the time of this writing we have come across an article (Allen and Strong [27]) which reports on the generation of steady vowels by a method somewhat similar to ours. However, their approach does not include a model for self-oscillation of the vocal cords, ignores frication and nasality, and does not deal with the dynamic variations in the glottal parameters or in the shape of the vocal tract. All these effects are included in our model whose details are given in the following sections.

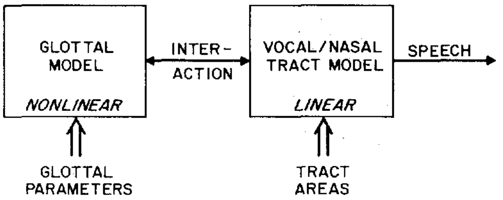


Fig. 1. General outline of an articulatory speech synthesizer.

### A. Model for the Glottal Source

The purpose of a glottal model (left box of Fig. 1) is to transform a few slowly varying control parameters into a relatively fast varying vocal tract excitation function. Although synthetic speech can be generated by using a simple excitation waveform specified uniformly over an utterance (Rosenberg [28]), it seems to be essential that a model of the glottal source used in articulatory speech synthesis should reproduce more naturally the variations in the essential acoustic features of the excitation. Such features are [10] the proper interaction with the acoustic load provided by the coupled tract input impedance $Z_{\text{in}}$, and realistic dependence of the glottal volume velocity waveform ($u_g$) and pitch ($F_0$) on lung pressure ($p_S$), pitch factor ($q$), and glottal rest area ($A_{g0}$). The $u_g$-waveform and $F_0$ can be considered as being output variables of the glottal model; $Z_{\text{in}}$, $p_S$, $q$, and $A_{g0}$ are the related quasi-stationary control variables.

Although at first glance it seems important to include the acoustic properties of the subglottal trachea and lungs, several authors have demonstrated that their influence on the performance of an articulatory speech synthesizer is only minor (e.g., [2], Wakita and Fant [29]). Guérin [30] found only slightly more skewed $u_g$-waveforms when the subglottal structures were also simulated. Therefore, at present, we have not included a simulation of subglottal structures in our synthesizer.

Acoustic interaction of source and tract is the subject of five recent papers (Fant [31], Guérin [30], Rothenberg [32], Fant *et al.* [33], Ananth *et al.* [34]). In the latter paper, a method was proposed for measuring the source–tract interaction from speech, expressing source/tract interaction as "variation in damping" of the speech waveform occurring as a result of open and closed glottal phases.

Acoustic interaction gives rise to several effects. One of these effects is skewing of the glottal waveform whenever moderately narrow constrictions of the vocal tract occur, such as in /a/ near the glottis, and in /i/ near the lips. In these cases, the negative slope of $u_g$ becomes steeper compared to the positive slope, thus giving rise to higher energy in the high-frequency part of the excitation spectrum. (Rothenberg [35] gives a good explanation of this effect.)

Another interaction effect consists of oscillatory ripples, mostly visible on the positive slope portion of the glottal waveform. The frequency of these ripples usually lies slightly below twice the frequency of the first formant ($F_1$) of the tract (Fant *et al.* [33]).

Other interaction effects are the dependence of duty cycle (ratio of open to closed phase) of the glottis on $F_1$ (e.g., [30, Fig. 38-8]) for high pitch voicing, and the dependence of pitch frequency $F_0$ on load characteristics ([30, p. 485]), resulting in a "pull effect" reported by Ishizaka and Flanagan ([10, Fig. 16]).

There are several different models of the glottal source which attempt to reproduce these interactions. Some models parameterize the glottal waveform (see, e.g., Rothenberg [35], Ananthapadmanabha and Fant [36], Fant [31]) in terms of lung pressure and glottal area. Others try to explicitly model the mechanical system of the glottis, by one or more *discrete* mechanical oscillators (Flanagan and Landgraf [37], Flanagan and Cherry [38], Ishizaka and Flanagan [10]). The most complex models solve coupled acoustomechanical boundary-value problems of the *distributed* and *layered* mechanical system (e.g., Guérin [30, p. 492], Titze and Talkin [39]). (More recently, Titze [40] described a method for parameterization of his model, thus decreasing the computational load.)

The model we selected for our synthesizer is the two-mass model of Ishizaka and Flanagan. Guérin [30] and Cranen and Boves [41], [42] have shown that this model has very realistic properties. An outline of the model is shown in Fig. 2. We will refer to that figure for an explanation of the variables appearing in the rest of this section. The glottal volume velocity $u_g(t)$ satisfies the differential equation

$$R_{\text{tot}} u_g + L_{\text{tot}} \frac{du_g}{dt} = p_S - p_1 - p_{ng}. \qquad (1)$$

Here $p_S$ is the subglottal (i.e., in our case the lung) pressure, $p_1$ is the pressure downstream of the glottal expansion (see Fig. 2), and $p_{ng}$ is a series noise pressure source located at the interface between expansion and first (variable) section of the vocal tract (see Section II-B-2). We discretize (1) using backward differences. For sampling instant $n$ this yields

$$u_g(n) = \frac{[p_S(n) - p_1(n) - p_{ng}(n)]\, t_s + L_{\text{tot}}(n)\, u_g(n-1)}{t_s R_{\text{tot}}(n) + L_{\text{tot}}(n)}. \qquad (2)$$

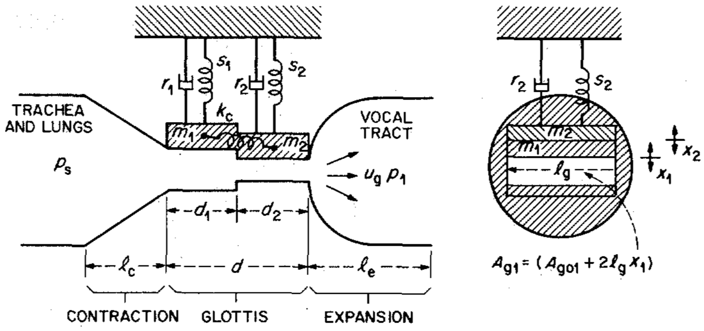


Fig. 2. The two-mass model of the vocal cords; after Ishizaka and Flanagan [10].

Here $t_s$ is the sampling interval, and $R_{\text{tot}}$ and $L_{\text{tot}}$ are the total quasi-stationary resistance and inductance representing the contraction, glottis, and expansion, as depicted in Fig. 2. These variables are given by

$$L_{\text{tot}} = \rho(d_1/A_{g1} + d_2/A_{g2}) \qquad (3)$$

and

$$R_{\text{tot}} = \frac{\rho}{2}\left[\frac{0.37}{A_{g2}^2} + \frac{1 - 2\dfrac{A_{g2}}{area_1}\left(1 - \dfrac{A_{g2}}{area_1}\right)}{A_{g2}^2}\right]|u_g| + 12\mu l_g^2(d_1/A_{g1}^3 + d_2/A_{g2}^3), \qquad (4)$$

where $area_1$ is the area of the first section of the vocal tract. In (3) and (4) the quantities $A_{g1}$, $A_{g2}$, $area_1$, and $u_g$ are all time varying. To simplify the expressions, their dependence on $n$ is not shown explicitly. The glottal areas $A_{g1}$ and $A_{g2}$ are related to the lateral displacements $x_1$ and $x_2$ by

$$A_{g1}(n) = A_{g01} + 2l_g x_1(n) \qquad (5)$$

and

$$A_{g2}(n) = A_{g02} + 2l_g x_2(n), \qquad (6)$$

where $A_{g01}$ and $A_{g02}$ are commonly set to the same glottal rest area $A_{g0}$.

Our realization differs from the original version of Ishizaka and Flanagan in three ways. The most important difference is that, due to the hybrid method, we do not have to set up a *large* system of equations. Instead, we compute the present value of $p_1$ in terms of the present value of $u_g$ and past values of $p_1$ and $u_g$, as shown in (18) of Section II-B-2. Equations (2) and (18) can thus be solved for $p_1(n)$ and $u_g(n)$. Note, however, that $R_{\text{tot}}$ and $L_{\text{tot}}$, appearing in (2), depend on the lateral deflections $x_1$ and $x_2$, as shown by (3)–(6). The deflections are obtained as the solutions of two second-order differential equations driven by $p_1$ and $p_S$. All four equations are thus coupled and must be solved simultaneously.

The second difference between our realization and that of Ishizaka and Flanagan is a more consistent discretization of these equations. The details of this discretization are given in the Appendix.

TABLE I
VARIABLES USED IN THE TWO-MASS MODEL OF ISHIZAKA AND FLANAGAN [10]

| Variable | Meaning | Value | Unit |
|---|---|---|---|
| $m_1$ | mass 1 of vocal cords | $0.125/q$ | g |
| $m_2$ | mass 2 of vocal cords | $0.025/q$ | g |
| $d_1$ | thickness of $m_1$ | $0.25/q$ | cm |
| $d_2$ | thickness of $m_2$ | $0.05/q$ | cm |
| $\eta_{k1}$ | nonlinear spring coeff. | 100 | — |
| $\eta_{k2}$ | nonlinear spring coeff. | 100 | — |
| $\eta_{h1}$ | nonlinear spring coeff. | 500 | — |
| $\eta_{h2}$ | nonlinear spring coeff. | 500 | — |
| $h_1$ | nonlinear spring coeff. | $3k_1$ | dyn/cm |
| $h_2$ | nonlinear spring coeff. | $3k_2$ | dyn/cm |
| $k_1$ | linear spring coeff. | $80000q$ | dyn/cm |
| $k_2$ | linear spring coeff. | $8000q$ | dyn/cm |
| $k_c$ | coupl. spring coeff. | $25000q^2$ | dyn/cm |
| $\mu$ | viscosity of air | $1.86 \times 10^{-4}$ | dyn s/cm$^2$ |
| $\rho$ | density of air | $1.14 \times 10^{-3}$ | g/cm$^3$ |
| $r_{1_{\text{open}}}$[a] | damping resistance | $2 \times 0.2\sqrt{k_1 m_1}$ | g/s |
| $r_{1_{\text{closed}}}$[a] | damping resistance | $2 \times 1.1\sqrt{k_2 m_2}$ | g/s |
| $r_{2_{\text{open}}}$[a] | damping resistance | $2 \times 0.6\sqrt{k_1 m_1}$ | g/s |
| $r_{2_{\text{closed}}}$[a] | damping resistance | $2 \times 1.9\sqrt{k_1 m_1}$ | g/s |

[a] Dynamically modified by the damping variable $g_s : r \rightarrow r/g_s^2$.

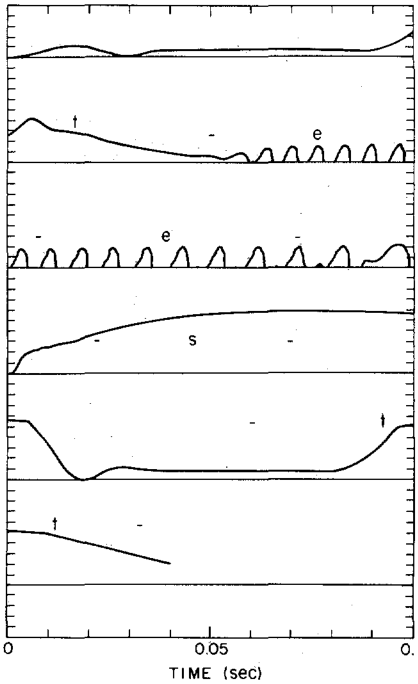


Fig. 3. Time waveform of the glottal flow $u_g$ for the word "test." The figure shows six consecutive intervals, each 0.1 s long. Articulatory parameters are from Coker's program [15].

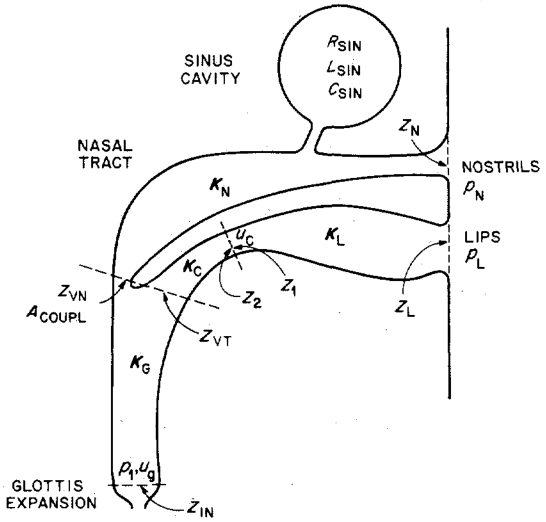


Fig. 4. Sketch of the vocal and nasal tracts. Chain matrices $K_G$, $K_N$, $K_C$, and $K_L$ cover the glottal region, the nasal tract, the tract between velum and constriction, and the tract between the constriction and the lips, respectively.

Third, following Coker, we added another time-varying parameter $g_S$ to modify the glottal damping resistances (i.e., $r_1 \rightarrow r_1/g_S^2$ and $r_2 \rightarrow r_2/g_S^2$), because it was found that the original glottal model did not stop voicing fast enough when $A_{g0}$ was suddenly increased in order to produce stops and fricatives after a vowel. The minimum value chosen for $g_S^2$ is typically about 0.16. There are several justifications for such a variable damping (e.g., Hirose [43] and Hirose *et al.* [44]). Further, Hutters [45] has recently shown that aspirated and unaspirated stops are generated by different types of glottal gesture rather than by a different timing of the glottal and supraglottal articulations. She produced evidence that devoicing is an active process carried out by the interarytenoid muscle and by the posterior cricoarytenoid muscle (see, e.g., Zemlin [46, pp. 150–155] for the anatomy of these muscles), rather than just by a diminished transglottal pressure due to closure of the vocal tract.

The values of all variables (except the ones controlled externally) used in our version of the two-mass model are given in Table I. As an example of its output we show the $u_g$-waveform of the synthesized word "test" in Fig. 3. The externally controlled variables in this example were obtained by using the method of Coker [15].

### B. Modeling of the Vocal and Nasal Tracts

First we will describe the method of analysis in the frequency domain. This is followed by a description of the method of synthesis in the time domain.

*1) Frequency Domain Analysis:* In Fig. 4 the vocal and nasal tracts are outlined. The velum is assumed to be at 8 cm downstream from the glottal expansion. We also detect the position of the narrowest constriction between velum and lips, which will be needed, if small enough, to introduce frication noise. We assume that the tract can never be constricted between glottis and velum. This is true for most western languages. Aspiration noise will be generated *at* the glottis, as will be described in detail later.

*a) Vocal Tract.* The different portions of the tract are described by four chain matrices, namely, $K_G$ for the laryngeal region between glottis and velum, $K_N$ for the nasal tract from the velum to the nostrils, $K_C$ from the velum to the constriction, and $K_L$ from the constriction to the lips.

A general chain matrix of a portion of the tract relates (planar) output pressure $P_{\text{out}}$ and volume velocity $U_{\text{out}}$ to the input pressure $P_{\text{in}}$ and volume velocity $U_{\text{in}}$ (capitalized in order to denote variables in the frequency domain). Thus,

$$\begin{pmatrix} P_{\text{out}} \\ U_{\text{out}} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \times \begin{pmatrix} P_{\text{in}} \\ U_{\text{in}} \end{pmatrix} = K \times \begin{pmatrix} P_{\text{in}} \\ U_{\text{in}} \end{pmatrix}, \quad (7)$$

where the input is the glottal side, the output side is toward the lips or nostrils. Assuming loss distribution as in [11], the chain matrix parameters for a homogeneous cylindrical tube of length $\Delta l$ and area *area* are given by

$$A = \cosh(\sigma\Delta l/c); \qquad B = -\frac{\rho c}{area}\gamma \sinh(\sigma\Delta l/c)$$

$$C = -\frac{area}{\rho c}\frac{\sinh(\sigma\Delta l/c)}{\gamma}; \quad D = \cosh(\sigma\Delta l/c). \quad (7a)$$

Here, the complex variables $\sigma$ and $\gamma$ are defined as

$$\gamma = \sqrt{\frac{a + j\omega}{\beta + j\omega}} \quad (7b)$$

and

$$\sigma = \gamma(\beta + j\omega), \quad (7c)$$

where

$$\alpha = \sqrt{j\omega c_1} \quad (7d)$$

and

$$\beta = \frac{j\omega\omega_0^2}{(j\omega + a)j\omega + b} + \alpha. \quad (7e)$$

The last equation differs from (29) in [11] because we also take wall compliance into account. The parameters $\omega_0^2$, $a$, and $b$ are defined by the equations

$$\omega_0^2 = \frac{\rho c^2}{area\ L'_w} \quad (7f)$$

$$a = R'_w/L'_w \quad (7g)$$

$$b = 1/(L'_w C'_w), \quad (7h)$$

where $L'_w$, $C'_w$, and $R'_w$ are the mass compliance, and resistance of the wall per unit length, respectively. Note that as in Sondhi [11], we assume the wall parameters to vary with *area* in such a way that $\omega_0^2$, $a$, and $b$ are independent of *area*. Also, $\omega_0$ is interpreted as the lowest angular resonance frequency of the tract when it is closed at both ends. Data for all variables of the vocal and nasal tract are given in Table II.

Each of the four matrices $K_G$, $K_N$, $K_C$, and $K_L$ is obtained by concatenating elementary matrices of the type given by (7) using the same $\Delta l$ for all elements.

At the velum two special coupling matrices are needed. For computing the chain matrix from glottis to lips, the nasal side branch is represented by the matrix $K_{cN}$ given by

$$K_{cN} = \begin{pmatrix} 1 & 0 \\ -1/Z_{VN} & 1 \end{pmatrix} \quad (8)$$

where $Z_{VN}$ is the input impedance of the nasal branch at the velum. Similarly, for computing the chain matrix from glottis to the nostrils, the vocal tract is represented by the coupling matrix $K_{cT}$ which is the same as $K_{cN}$ with $Z_{VN}$ replaced by the input impedance of the oral cavity $Z_{VT}$.

*b) Nasal Tract:* The nasal tract was modeled after geometrical data of Maeda [47]. It has a geometrical length of 11 cm. Modeling the nasal tract by only a variable-area tube of that length results in too high a frequency for the first pole/zero pair of the nasal tract transfer function (e.g., Fujimura and Lindqvist [48]). Lindqvist and Sundberg [49] got more realistic frequency locations for the first pole and zero when the sinus (side) cavities were also modeled. Maeda [47] used a single side cavity of 20.8 cm$^3$ coupled to the main nasal tract via a 0.5 cm long tube of 0.1 cm$^2$ cross section at a distance of 7 cm from the velum. This side cavity introduces an additional pole slightly below its resonance frequency and an additional zero slightly above its resonance frequency. In contrast to Maeda, who modeled the side cavity too by concatenating homogeneous line sections, we used a computationally more economical approach by modeling the sinus cavity by a discrete Helmholtz resonator, having an impedance

$$Z_{\text{sin}} = R_{\text{sin}} + j\omega L_{\text{sin}} + \frac{1}{j\omega C_{\text{sin}}}.$$

The coupling of the sinus cavity to the nasal tract is achieved by a special chain matrix analogous to (8). The model of Maeda also features a special method to reduce the area in the vocal tract according to the opening of the velum. This reduction results in a noticeable downshift of all higher formants of the vocal tract, a phenomenon observed also in natural nasalized speech. Good agreement with the nasal transfer function reported in Maeda [47] was obtained with the resonator data and the viscous loss parameter $c_1$ given in Table II.

*2) Time Domain Synthesis of Voiced Sounds:* Here we will describe how the glottal model and the vocal/nasal tract model are interfaced to each other.

Assuming that the previously defined vocal/nasal tract chain matrices $K_G$, $K_N$, $K_C$, $K_L$, $K_{cN}$, and $K_{cT}$ are known, we form global matrices as follows (see Fig. 4). The chain matrix from the glottis to the point of the narrowest constriction is

$$K_{\text{fric}} = K_C K_{cN} K_G. \quad (9)$$

The chain matrix from the glottis to the lips (computed only if the constriction is not completely closed) is given by

$$K_{\text{tract}} = K_L K_{\text{fric}}. \quad (10)$$

The chain matrix from the glottis to the nostrils is

$$K_{\text{nasal}} = K_N K_{cT} K_G. \quad (11)$$

The input impedance $Z_{\text{in}}$ is given by

$$Z_{\text{in}} = \frac{D_{\text{tract}} Z_L - B_{\text{tract}}}{A_{\text{tract}} - C_{\text{tract}} Z_L}, \quad (12)$$

TABLE II
VARIABLES USED IN THE FREQUENCY DOMAIN ANALYSIS OF THE VOCAL AND NASAL TRACTS

| Variable | Meaning | Value | Unit |
|---|---|---|---|
| | (Vocal Tract) | | |
| $\Delta l$ | length of elementary section | 0.85 | cm |
| $a$ | ratio of wall resistance to mass | $130\pi$ | rad/s |
| $b$ | squared angular freq. of mechan. resonance | $(30\pi)^2$ | $(\text{rad/s})^2$ |
| $c_1$ | correction for thermal conductivity and viscosity | 4 | rad/s |
| $\omega_0^2$ | lowest squ. ang. freq. of acoust. resonance | $(406\pi)^2$ | $(\text{rad/s})^2$ |
| | (Nasal Tract, All Other Variables as for Vocal Tract) | | |
| $c_1$ | correction for thermal conductivity and viscosity | 72 | rad/s |
| | (Nasal Tract, Sinus Cavity) | | |
| $R_{\text{sin}}$ | acoust. resistance of coupling section | 1 | dyn s/cm$^5$ |
| $L_{\text{sin}}$ | acoust. mass of coupling section | $5.94 \times 10^{-3}$ | g/cm$^4$ |
| $C_{\text{sin}}$ | acoust. compliance of 20.8 cm$^3$ | $15.8 \times 10^{-6}$ | cm$^4$ s$^2$/g |

where $Z_L$ is the radiation impedance at the lips and $A_{\text{tract}}$, $B_{\text{tract}}$, $C_{\text{tract}}$, and $D_{\text{tract}}$ are the elements of the matrix $K_{\text{tract}}$. The input impedances $Z_{VN}$ and $Z_{VT}$ have similar expressions. The radiation impedance is taken to be that of a pulsating sphere [2, p. 36] with a radius equal to that of the lip opening.

Denoting $U_g$ and $P_L$ for the Fourier transformed glottal flow $u_g$ and sound pressure $p_L$ radiated at the lips, respectively, the transfer function $H_L$ from $U_g$ to $P_L$ is

$$H_L = \frac{P_L}{U_g} = \frac{Z_L}{A_{\text{tract}} - C_{\text{tract}} Z_L}. \qquad (13)$$

$H_L$ is set to zero for a closed tract. The transfer function $H_N$ from $U_g$ to $P_N$ (the Fourier transform of the sound pressure radiated at the nostrils) is

$$H_N = \frac{P_N}{U_g} = \frac{Z_N}{A_{\text{nasal}} - C_{\text{nasal}} Z_N}. \qquad (14)$$

$H_N$ is set to zero if the nasal tract is not coupled to the vocal tract ($A_{\text{coupl}} = 0$, see Fig. 4). The (voiced) speech output is computed by convolving $u_g$ with the impulse response $h_{\text{out}}$[1] obtained by inverse Fourier transforming the transfer function $H_{\text{out}}$, given by

$$H_{\text{out}_{\text{voiced}}} = \frac{P_{\text{speech}}}{U_g} = H_L + H_N + H_{\text{vib}}, \qquad (15)$$

where $H_{\text{vib}}$ is the transfer function

$$H_{\text{vib}} = \frac{area_1}{c} \frac{j\omega a_{\text{vib}}}{c + j\omega a_{\text{vib}}} Z_{\text{in}} \beta \qquad (16)$$

representing the sound pressure radiated by a sphere of radius $a_{\text{vib}}$ vibrating with the particle velocity of the vocal tract wall at the glottis; $area_1$ is the first area data of the tract at the glottis end, and $Z_{\text{in}}$ is the input impedance of the tract in the same plane. (According to Fant *et al.* [50], the wall vibration is maximum at the glottis, followed by lip vibration which is about 4 dB down. We did not model lip vibration.) The definition of $\beta$ is given by (7e).

The interaction between the glottal model and the tract model takes place via the supraglottal pressure $p_1$. This pressure can be obtained by convolving the inverse Fourier transform of the impedance $Z_{\text{in}}$ with the glottal volume velocity $u_g$. Using the term $z_{\text{in}}$ for this impulse response, we obtain

$$p_1(t) = z_{\text{in}}(t) * u_g(t),$$

where * denotes convolution. In the discrete realization we obtain

$$p_1(n) = z_{\text{in}}(0)\, u_g(n) + \sum_{k=1}^{N-1} z_{\text{in}}(k)\, u_g(n-k), \qquad (17)$$

where $N$ is the length of the truncated impulse response.

In an early version of our synthesizer we computed $p_1$ from (17). It was found, however, that impractically long impulse responses had to be used in order to obtain a realistic buildup of supraglottal pressure during oral closure (see, e.g., Flanagan *et al.* [7, Fig. 6]). (It is ironic that this high computational load is required when in fact there is no speech output.) Schumacher [51] suggests a different approach which solves this problem. By utilizing the relationship between input impedance and input reflectance he shows that

$$p_1(t) = Z_0 u_g(t) + r_{\text{in}}(t) * [p_1(t) + Z_0 u_g(t)].$$

Here $Z_0$ is the characteristic impedance of the first section of the vocal tract model ($Z_0 = \rho c / area_1$) and $r_{\text{in}}(t)$ is the inverse Fourier transformed input reflectance $R_{\text{in}}$ of the tract ($R_{\text{in}}(\omega) = (Z_{\text{in}} - Z_0)/(Z_{\text{in}} + Z_0)$). After discretization we obtain

$$p_1(n) = \frac{1 + r_{\text{in}}(0)}{1 - r_{\text{in}}(0)} Z_0 u_g(n) + \frac{1}{1 - r_{\text{in}}(0)} \sum_{k=1}^{N-1} r_{\text{in}}(k) \cdot [p_1(n-k) + Z_0 u_g(n-k)]. \qquad (18)$$

[1] Note that because of the symmetry properties of the functions appearing in (7a-h), $h_{\text{out}}$ is a real function. So also is the function $z_{\text{in}}(t)$ derived from $Z_{\text{in}}$.

At time $n$, the term represented by the summation is known so that (18) and (2) are two equations for the two unknowns $u_g(n)$ and $p_1(n)$. The solution of these equations along with the coupled oscillator equations (see the Appendix) completes the computations for the time instant $n$.

The recursive form (18) is more economical compared to (17) even for the open tract since $r_{\text{in}}$ invariably decays much faster than $z_{\text{in}}$. This is because $r_{\text{in}}$ is measured with a matched (reflectionless) termination at the measurement point, while $z_{\text{in}}$ is measured with a *rigid* termination. The former obviously results in faster energy loss. (For a rigid-walled tube terminated in a lossless load, it is shown in Atal and Hanauer [4], that the inverse Fourier transforms of the chain matrix parameters are all of approximately 1 ms duration. A recursive formulation directly in terms of these parameters should therefore give the shortest convolutions. We tried this approach but abandoned it; due to yielding walls and nasal coupling, the durations turn out to be much longer, especially for the parameter $C$. The convolutions were also found to be not very stable numerically.)

*3) Time Domain Synthesis of Unvoiced Sounds:* In natural speech unvoiced sounds are either produced by aspiration (at the glottis) or by frication (at a constriction of the tract). Both cases are treated accordingly in our synthesizer; that is, noise of the correct amplitude is automatically injected at the correct place within the tract.

For aspiration we follow the suggestions of Fant [52, pp. 272-275], Flanagan [2, p. 251], and Flanagan *et al.* [18], and add a noise pressure source with amplitude proportional to the difference of the squared local Reynolds number $\text{Re}^2$ and a threshold (critical Reynolds number $\text{Re}^2_{\text{crit}}$). Thus, we set

$$\begin{aligned} p_{ng} &= g_{ng} \, random(\text{Re}^2 - \text{Re}^2_{\text{crit}}), \quad && \text{Re} > \text{Re}_{\text{crit}} \\ &= 0, && \text{Re} \leq \text{Re}_{\text{crit}} \end{aligned} \tag{19}$$

where $g_{ng}$ is an empirically determined gain (about $2 \times 10^{-6}$), *random* is a random number uniformly distributed between $-0.5$ and $0.5$, and $\text{Re}^2_{\text{crit}}$ is empirically found to be about $2700^2$. $p_{ng}$ is then substituted into (1).

For frication generated in the vocal tract we have made several modifications to the original suggestions. First, while Flanagan *et al.* [18] used a noise source in each $T$-section of their vocal tract network, we use only one noise source at the point of maximum constriction. In Flanagan's proposal the noise source is a pressure source and appears in the vocal tract network as sketched in Fig. 5(a). Here, the impedance $Z_1$ is the backward input impedance seen toward the glottis at the constriction. Assuming a wide open glottis and zero subglottal impedance, we obtain

$$Z_1 = -B_{\text{fric}}/D_{\text{fric}} \tag{20}$$

and

$$Z_2 = \frac{D_L Z_L - B_L}{A_L - C_L Z_L}. \tag{21}$$

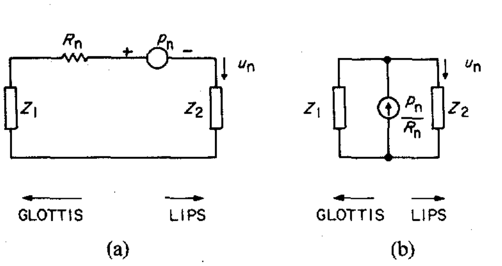


Fig. 5. Implementing constriction noise sources in the vocal tract model. Impedances $Z_1$ and $Z_2$ represent the residual vocal tract from the reference plane to the glottis and to the lips, respectively (see Fig. 4). (a) Series pressure source model; (b) parallel volume velocity source model.

Here $A_L$, $B$, $\cdots$, are the elements of $K_L$, etc. The squared Reynolds number $\text{Re}^2$ is

$$\text{Re}^2 = \frac{4\rho^2}{\pi\mu^2} \frac{\bar{u}_C^2}{area_C}, \tag{22}$$

where $\bar{u}_C$ is a digitally low-pass filtered version of the volume velocity $u_C$ at the constriction:

$$\bar{u}_C(n) = \bar{u}_C(n-1) + [u_C(n) - \bar{u}_C(n-1)] \, 2\pi f_g t_s. \tag{23}$$

The choice of cutoff frequency $f_g$ is not critical. Flanagan *et al.* [18] used 500 Hz in order to ensure stability. We used 2000 Hz. The source resistance of the noise pressure source is

$$R_n = \frac{\rho |\bar{u}_C|}{2 area_C^2}. \tag{24}$$

The current volume velocity at the constriction due to the glottal flow $u_g$ is

$$u_C(n) = \sum_{k=0}^{N-1} h_u(k) \, u_g(n-k), \tag{25}$$

where $h_u$ is the impulse response corresponding to

$$H_u(\omega) = \frac{U_c}{U_g} = \frac{1}{D_{\text{fric}} Z_2 - B_{\text{fric}}}. \tag{26}$$

The noise sound pressure radiated at the lips is then obtained by convolving the noise flow in $Z_2$ with the impulse response corresponding to the appropriate volume velocity to sound pressure transfer function [in analogy to (13)].

This model did not give satisfactory results because the "internal" impedance of the noise source was too high. (We do not report here how we incorporated the time dependent resistance $R_n$ in calculating $u_C$ in this case. Even with $R_n = 0$, for example, we could not get a good quality /t/, because $Z_1$ was much too high, thus preventing any large volume velocity $u_n$.) It turned out that good results could be obtained by introducing the noise source at a "reference plane" downstream of the constriction. Unfortunately, for good results, the position of the reference plane had to be different for different fricatives and stop consonants. This finding is consistent with the results of Shadle [53]. Because of the need to modify the position of the reference plane, this method is very inconvenient, and we abandoned it in favor of the following alternative.

Shadle [54] has shown that if the noise source is represented by a volume velocity source, then its position is not very critical. Following the procedure outlined by Liljencrants [25, pp. 5–4, eq. 5.109], we therefore used the short-circuit noise flow $u_n = p_n / R_n$ in a parallel network shown in Fig. 5(b). An acceptable position for the reference plane was found to be one section downstream of the outlet of the narrowest constriction (as long as the constriction was not at the lips). By informal listening tests, the optimum critical squared Reynolds number $\mathrm{Re}_{\mathrm{crit}}^2$ was found to be about $3500^2$, the appropriate gain for computing $p_n$ in analogy to (19) was found to be about 0.0001.

The transfer function between noise flow $U_n$ and radiated noise pressure at the lips can be derived by inspection of Fig. 5(b) and by using, for example, (13). We obtain

$$H_{\mathrm{outfric}} = \frac{P_{\mathrm{speech}}}{U_n} = \frac{Z_L}{A_L - C_L Z_L} \frac{Z_1}{Z_1 + Z_2}. \qquad (27)$$

We denote the impulse response corresponding to $H_{\mathrm{outfric}}$ with the symbol $h_n$ in order to avoid double indexes on $h_{\mathrm{out}}$. (In the description above, we introduced the noise source at the reference plane but the amplitude of the source was supposed to be based on the flow at the constriction. In the actual implementation, we calculate the amplitude from the flow at the reference plane rather than at the constriction. This is not a significant change, but eliminates the necessity to compute another transfer function.)

### C. Practical Realization

*1) Outline of the Program:* The synthesizer program is organized in three major parts. The main program SYNTHC does the I/O, both for documentation and articulatory parameter input (coded in ASCII), and for "speech data" output (coded in binary). There exist versions of the program to read input data from Sondhi and Resnick [14] and from Bocchieri [13]. Different variables can easily be selected for output in a scaled 16-bit integer format for convenient listening or plotting (see example in Fig. 3). As an option, SYNTHC also generates linearly interpolated frames of glottal and tract related input data (needed, for example, for the currently still too sparsely sampled tract area functions of [14]), and does all time-domain processing with the exception of the two-mass model, which is evaluated in the subroutine CORD.

An important feature of the program is the sample-by-sample linear interpolation of all impulse responses between two adjacent frames. This interpolation does not *exactly* correspond to a linear interpolation of the areas, as can be shown using the theoretical framework given in [14, eqs. (3–10)]. This method avoids, however, spurious signals otherwise generated due to the undersampling of the tract areas (also see, e.g., Liljencrants [25, pp. 0–5]).

SYNTHC also interpolates the glottal parameters with a first-order low-pass filter in analogy to (23). Here the cutoff frequency is 10 Hz.

All variables to be used for convolution (e.g., $u_g$, $u_n$, $p_1 + Z_0 u_g$) are stored twice in delay lines (arrays), using an offset of $N$ (the length of the impulse responses). This greatly simplifies the addressing for all convolutions.

The main program calls two major subprograms, namely, TRACT and CORD. TRACT calculates the target values of all impulse responses once per frame (of typical duration 10 ms). CORD is called for each sample to update the displacements $x_1$ and $x_2$ of the two masses of the vocal cord model.

TRACT analyzes the vocal tract in the frequency domain. For voiced speech and for a completely closed tract (e.g., for velar nasals [52, p. 150]), it computes the two impulse responses $h_{\mathrm{out}}$ and $r_{\mathrm{in}}$ according to (12), (15), and (19), when there is a small enough nonzero opening somewhere in the tract, TRACT also computes the impulse responses $h_n$ and $h_u$ corresponding to $H_{\mathrm{outfric}}$ [(27)] and $H_u$ [(25), (26)].

During intervals when there is no narrow constriction in the tract, the frication related impulse responses $h_u$ and $h_n$ are set to zero, thus yielding a "soft" onset of frication in the next frication interval. (Note: all impulse responses are interpolated on a sample-by-sample basis.) It was also found that these impulse responses should be computed (and used) if the previous frame had a constriction and the current frame does not. (Additionally, in order to get a good stop release, the area value at the point of closure is prevented from being larger than 0.2 cm$^2$ for the duration of one frame following closure.)

All transfer functions are low-passed by a zero-phase filter with a gradual rolloff starting at half-Nyquist frequency and are set to zero above three-quarters of the Nyquist frequency. This was found to be a reasonable compromise at a sampling frequency of 20 kHz. For a sampling frequency of 8 kHz, this filter has to be adjusted, otherwise the synthesized speech will sound muffled. The impulse responses are computed by inverse Fourier transformation and then windowed by applying the right half of a Hamming window of length $2N$, where $N$ is the length of the inverse FFT. For a sampling frequency of 20 kHz, a length of $N = 512$ points was found to be appropriate; for 8 kHz 256 points could be used.

*2) Synthesis Results:* In Fig. 6 we show the three glottal parameters $A_{g0}$, $p_S$, and $q$ for synthesizing the word "test." (The resulting $u_g$-waveform was already reported in Fig. 3.) Fig. 7 shows the speech output and the supraglottal pressure $p_1$ on the same time scale as Figs. 3 and 6.

Finally, we show some spectrograms where corresponding ones from natural and/or synthetic speech can be found in the literature. The spectrogram in Fig. 8 ("we were away a year ago") can be compared to the spectrogram reported by Atal and Hanauer [4, Fig. 10]. The spectrogram in Fig. 9 ("noon is the sleepy time of day") can be compared to the spectrogram shown by Flanagan [55, Fig. 5]. Both sentences were synthesized from data obtained by Coker's program. The spectrogram in Fig. 10 ("how are you?") can be compared to the spectrogram reported by Sondhi and Resnick [14, Fig. 26]. This sen-

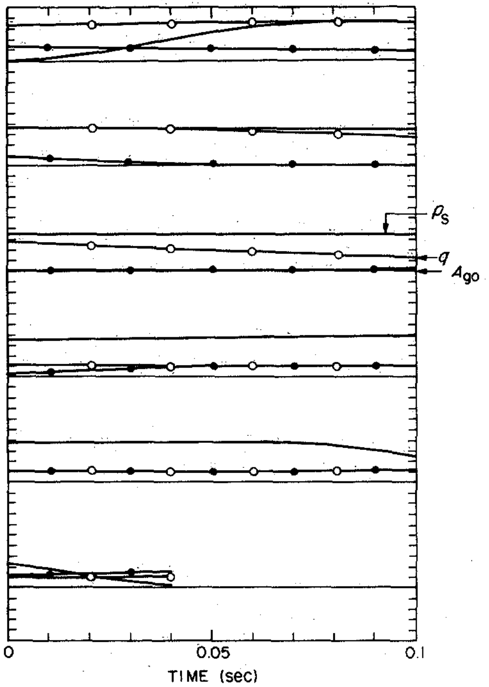


Fig. 6. Articulatory parameters $A_{g0}$, $p_S$, and $q$ used to synthesize "test." (The scaling is different for each of these parameters.)

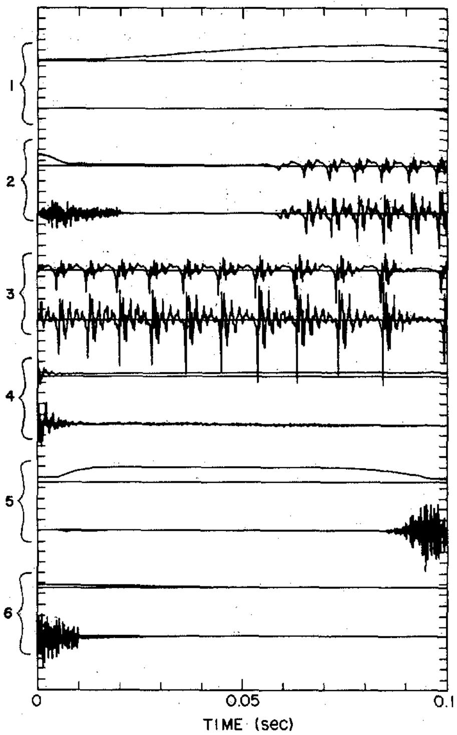


Fig. 7. Resulting supraglottal pressure $p_1$ and speech waveform (upper and lower curve, respectively). Note that no noise appears in the $p_1$-waveform due to the fact that we do not compute its feedback to the glottis. Note also that $p_1$ is attenuated by a factor of 20 relatively to the speech output.

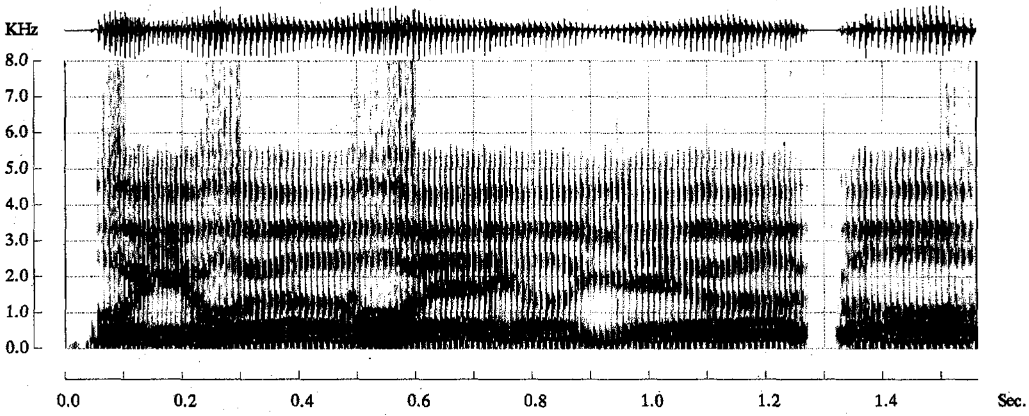


Fig. 8. Spectrogram for the sentence "we were away a year ago." Area data from a text-to-articulatory parameter program [15].

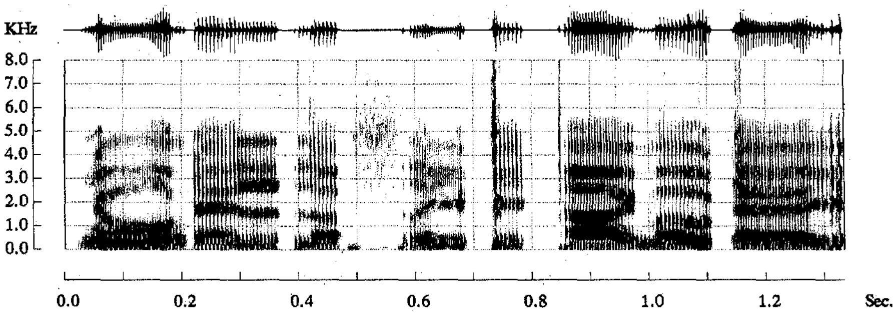


Fig. 9. Spectrogram for the sentence "noon is the sleepy time of day." Area data and glottal parameters from a text-to-articulatory parameter program [15].

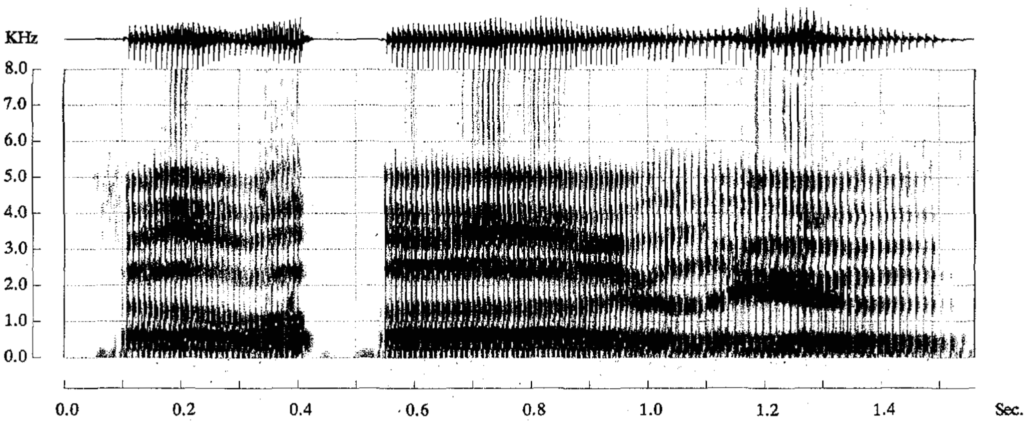


Fig. 10. Spectrogram for the sentence "how are you?" Area data as measured by Sondhi and Resnick [14].

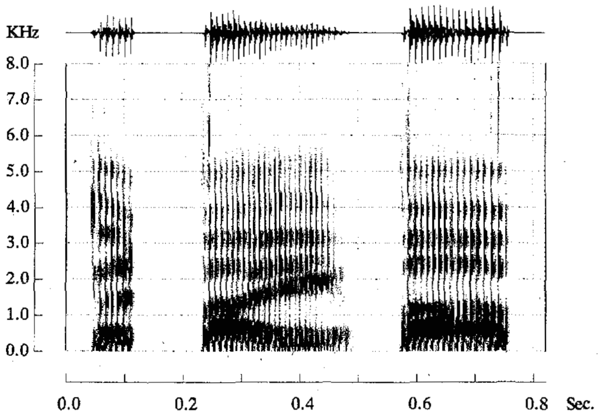


Fig. 11. Spectrogram for the sentence "goodbye Bob." Area data from Bocchieri [13].

tence was synthesized using the same measured areas as in Sondhi and Resnick [14], but manually generated glottal parameters. Finally, Fig. 11 ("goodbye Bob") shows the spectrogram obtained from Bocchieri's area data and manually generated glottal parameters.

## III. Discussion and Conclusions

In this paper we have described an articulatory speech synthesizer in which the properties of the oral and nasal tracts are computed in the frequency domain, converted to the time domain by inverse Fourier transformation, and interfaced with a time-domain nonlinear model of the vocal cords. The synthesizer allows for nasalization, frication, and aspiration. Many of the techniques used in the synthesizer have existed in the literature. What we have done is to bring them all together in a comprehensive synthesizer. We mention here some noteworthy features of our synthesizer, comparing them in particular to the synthesizer of Ishizaka and Flanagan [10].

1) Our chain matrix representation for the vocal tract is much more accurate and economical compared to the lumped parameter representation of [10]. To illustrate this point, note that for a hardwalled, lossless, uniform tract the chain matrix gives exact values for the formants with even one section. On the other hand, in the ten-section lumped parameter representation of [10] the second formant is in error by 1 percent and the fourth formant by 5 percent. Bocchieri [13] found that he needed 60 sections to get good spectral shapes with the lumped-parameter representation.

2) The hybrid simulation is much more versatile than the method of [10].

Note, for instance, that if one has a codebook of articulatory shapes, the impulse responses of those shapes can be precomputed and the computation reduced drastically. The lumped parameter method is incapable of utilizing such a codebook.

As another example, we might mention the ease with which losses such as viscous and thermal losses can be included in our simulation. Since such losses vary as the square root of the frequency, they do not have lumped parameter representations. In simulations with the method of [10], these losses can be incorporated only as convolutions with impulse responses proportional to $\sqrt{t}$.

Similar remarks apply to the wall impedance. At present we (as well as [10]) approximate this as a compliance-mass-resistance system. However, the wall may be better represented in terms of frequency-dependent elements. Including such elements would require a trivial modification of our simulation, but would present formidable difficulties for the [10] simulation.

3) Finally, we compare the computational complexity of our method to that of [10]. At a sampling rate of 8 kHz, the most comprehensive version of our synthesizer, including frication and nasality, takes about 2 s of Cray-1 CPU time per second of speech. It is difficult to directly compare this to the run times of other comparable synthesizers. Bocchieri [13] required 5 h of computation for 1 s of speech on the Data General Eclipse S/130. Flanagan *et al.* [7] quote the figure of 200 s per second of speech on the Eclipse S/200 computer, for a 10-section implementation of the method of [10]. In a private communication, K. Ishizaka has informed us that when the last mentioned program was transformed to the Cray-1 computer, the computation time did not decrease significantly; however, after modifying the method of time-dis-

cretization, the computation time was reduced to about 2 s per second of speech. This is still about 5 or 6 times slower than our method because, as mentioned above, comparable modeling accuracy requires about 60 sections rather than the 10 used by [10].

At present we have the following sources of input data for the synthesizer: a) data for one sentence from Bocchieri [13] (this was manually generated in an interactive trial and error procedure); b) data for one sentence and a few words from the impedance tube measurements of Sondhi and Resnick [14]; and c) data for a few sentences generated by the text-to-speech transcription program of Coker [15].

For text-to-speech synthesis, these types of data are useful but with the following limitations.

a) The interactive method is too time consuming to be useful for synthesis of large amounts of speech. Its main purpose appears to be experimental and educational.

b) The transcription program provides a complete set of control parameters. We believe that such a program coupled to our synthesizer can produce good quality synthesis. At present, however, the control parameters have to be manually altered by quite a significant amount before they produce reasonable speech. This is because the parameters have been optimized by trial and error for use with a very different synthesizer. Considerable effort must be invested to optimize them for our synthesizer.

The impedance tube method can provide large amounts of data in a relatively short time. The present measurements have inadequate spatial resolution and too slow a temporal sampling rate. Both these limitations will be reduced with a new experimental setup now being assembled. However, it must be remembered that this method will yield only vocal tract shapes. It does not provide control parameters for the vocal cords, nasalization, and frication. The main use of this method will be in providing statistical properties of area functions which would lead to improved quantization rules.

For applications to low bit rate speech transmission, we must find a way to derive the control parameters from the speech signal. One method, which is not entirely satisfactory, is to use LPC derived pseudoareas and a pitch detector. Another is to estimate the parameters adaptively by matching synthetic speech to natural speech. For each of these methods we could find the vocal tract shapes by exhaustive search of a code book of shapes derived from impedance tube measurements. We believe experimentation with the synthesizer will suggest other, better methods.

## Appendix

### Solution of the Coupled Oscillator Equations

For every sampling instant $n$, (2) and (18) are to be solved for $u_g(n)$ and $p_1(n)$. By rearranging terms, those two equations can be written as follows:

$$t_s p_1(n) + den\, u_g(n) = [p_S(n) - p_{ng}(n)]\, t_s + L_{\text{tot}}(n)\, u_g(n-1) \quad \text{(A-1)}$$

and

$$p_1(n) - R_1 u_g(n) = \frac{1}{1 - r_{\text{in}}(0)} \sum_{k=1}^{N-1} r_{\text{in}}(k) \cdot [p_1(n-k) + Z_0 u_g(n-k)], \quad \text{(A-2)}$$

where

$$den = t_s R_{\text{tot}}(n) + L_{\text{tot}}(n) \quad \text{(A-3)}$$

and

$$R_1 = \frac{1 + r_{\text{in}}(0)}{1 - r_{\text{in}}(0)} Z_0. \quad \text{(A-4)}$$

These two simultaneous equations can be solved trivially once $L_{\text{tot}}$ and $R_{\text{tot}}$ are known. As shown in (3)-(6) of the text, these are functions of the lateral displacements $x_1$ and $x_2$ of the two masses. The lateral displacements are solutions of the differential equations

$$m_1 \ddot{x}_1 + r_1 \dot{x}_1 + s_1 + k_c(x_1 - x_2) = f_1 \quad \text{(A-5)}$$

and

$$m_2 \ddot{x}_2 + r_2 \dot{x}_2 + s_2 + k_c(x_2 - x_1) = f_2, \quad \text{(A-6)}$$

where $m_1$, $k_1$, etc., are as given in Table I, and $f_1$ and $f_2$ are complicated functions of $p_1$ and $u_g$ (as given in Ishizaka and Flanagan [10, eq. (18)]. The spring restoring forces $s_1$ and $s_2$ have the form

$$s_1 = h_1 x_1 + s_1' \quad \text{(A-7)}$$

$$s_2 = h_2 x_2 + s_2' \quad \text{(A-8)}$$

where $s_1'$ and $s_2'$ represent nonlinear restoring forces given by (A-13) and (A-14). If (A-5) and (A-6) are also discretized by backward differences, then it is clear that we would have a coupled set of cubic equations. This is avoided by a discretization scheme in which the cubic terms are delayed by one sample. This results in the linear system

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad \text{(A-9)}$$

where, closely following the notation of Ishizaka and Flanagan, the matrix elements and the $b$'s are given by

$$a_{11} = (k_1 + h_1 + k_c) t_s^2 + r_1 t_s + m_1; \quad a_{12} = -k_c t_s^2$$
$$a_{21} = -k_c t_s^2; \quad a_{22} = (k_2 + h_2 + k_c) t_s^2 + r_2 t_s + m_2. \quad \text{(A-10)}$$

(Set $h_1 = 0$, if $x_1(n-1) > -A_{g01}/2l_g$, and set $h_2 = 0$, if $x_2(n-1) > -A_{g02}/2l_g$.)

$$b_1 = (2m_1 + r_1 t_s) x_1(n-1) - m_1 x_1(n-2) - s_1' t_s^2 + f_1(n-1) \quad \text{(A-11)}$$

$$b_2 = (2m_2 + r_2 t_s) x_2(n-1) - m_2 x_2(n-2) - s_2' t_s^2 + f_2(n-1). \quad \text{(A-12)}$$

The nonlinear portions, $s_1'$ and $s_2'$, of the spring forces are

$$s_1' = \begin{cases} k_1\eta_{k1}x_1^3(n-1), & x_1(n-1) > -\dfrac{A_{g01}}{2l_g} \\ k_1\eta_{k1}x_1^3(n-1) - h_1\left[\dfrac{A_{g01}}{2l_g} + \eta_{k1}\left(x_1(n-1) + \left(\dfrac{A_{g01}}{2l_g}\right)\right)^3\right], & x_1(n-1) \leqslant -\dfrac{A_{g01}}{2l_g} \end{cases} \qquad \text{(A-13)}$$

$$s_2' = \begin{cases} k_2\eta_{k2}x_2^3(n-1), & x_2(n-1) > -\dfrac{A_{g02}}{2l_g} \\ k_2\eta_{k2}x_2^3(n-1) - h_2\left[\dfrac{A_{g02}}{2l_g} + \eta_{k2}\left(x_2(n-1) + \left(\dfrac{A_{g02}}{2l_g}\right)\right)^3\right], & x_2(n-1) \leqslant -\dfrac{A_{g02}}{2l_g}. \end{cases} \qquad \text{(A-14)}$$

In Ishizaka and Flanagan's paper, too, the cubic terms were delayed by one sample. However, for further simplification, they also delayed the linear term involving $k_c$ in an inconsistent manner. This inconsistency has been eliminated in the above discretization. In practice we found that this slight modification considerably improved numerical stability.

## Acknowledgment

The authors would like to thank M. Liberman and M. Melchner for their help in producing the spectograms, and C. Coker for many hours of consultations and for some changes he made to his program. We acknowledge helpful discussions with O. Fujimura, J. Larar, and T. Thomas.

## References

[1] J. L. Flanagan, M. R. Schroeder, B. S. Atal, R. E. Crochiere, N. S. Jayant, and J. M. Tribolet, "Speech coding," *IEEE Trans. Commun.*, vol. COM-27, pp. 710-737, 1979.

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

[3] R. Linggard, *Electronic Synthesis of Speech.* New York: Cambridge University Press, 1985.

[4] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637-655, 1971.

[5] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of the acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417-427, 1973.

[6] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Amer.*, vol. 63, no. 5, pp. 1535-1555, 1978.

[7] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Signal models for low bit-rate coding of speech," *J. Acoust. Soc. Amer.*, vol. 68, no. 3, pp. 780-791, 1980.

[8] S. E. Levinson and C. E. Schmidt, "Adaptive computation of articulatory parameters from the speech signal," *J. Acoust. Soc. Amer.*, vol. 74, no. 4, pp. 1145-1154, 1983.

[9] R. Kuc, F. Tuteur, and J. R. Vaisnys, "Determining vocal tract shape by applying dynamic constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 1985, pp. 1101-1104.

[10] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1233-1268, 1972.

[11] M. M. Sondhi, "Model for wave propagation in a lossy vocal tract," *J. Acoust. Soc. Amer.*, vol. 55, no. 5, pp. 1070-1075, 1974.

[12] ——, "An improved vocal tract model," in *Proc. 11th Int. Congr. Acoust.*, Paris, France, vol. 4, 1983, pp. 167-170.

[13] E. L. Bocchieri, "An articulatory speech synthesizer," Doctoral dissertation, Univ. Florida, 1983.

[14] M. M. Sondhi and J. R. Resnick, "The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis," *J. Acoust. Soc. Amer.*, vol. 73, no. 3, pp. 985-1002, 1983.

[15] C. H. Coker, "A model of articulatory dynamics and control," *Proc. IEEE*, vol. 64, pp. 452-460, 1976.

[16] M. H. L. Hecker, "Studies of nasal consonants with an articulatory speech synthesizer," *J. Acoust. Soc. Amer.*, vol. 34, no. 2, pp. 179-188, 1962.

[17] M. R. Portnoff, "A quasi-one-dimensional digital simulation for the time-varying vocal tract," S.B./S.M. thesis, M.I.T., Cambridge, MA, 1973.

[18] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell Syst. Tech. J.*, vol. 45, no. 3, pp. 485-506, 1975.

[19] S. Maeda, "A digital simulation method of the vocal tract system," *Speech Commun.*, vol. 1, pp. 199-229, 1982.

[20] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. 4th Int. Congr. Acoust.*, Paper G-42, 1-4; reprinted in *Speech Synthesis*, J. L. Flanagan and L. R. Rabiner, Eds. Stroudsburg, PA: Dowden, Hutchinson & Ross, 1962, pp. 127-130.

[21] A. Fettweis, "Digital filter structures related to classical filter networks," *Archiv für Elektronik und Übertragungstechnik*, vol. 25, pp. 79-89, 1971.

[22] J. Braas, "Ein digitales Leitungsmodell als Hilfsmittel zur Sprachsynthese" ("A digital line analog as a tool for speech synthesis"), dissertation, Ruhr-Univ. Bochum, Federal Republic of Germany, 1981.

[23] Y. Kabasawa, K. Ishizaka, and Y. Arai, "Simplified digital model of the lossy vocal tract and vocal cords," in *Proc. 11th Int. Congr. Acoust.*, Paris, France, vol. 4, 1983, pp. 175-178.

[24] P. Meyer and H. W. Strube, "Calculations on the time varying vocal tract," *Speech Commun.*, vol. 3, pp. 109-122, 1984.

[25] J. Liljencrants, "Speech synthesis with a reflection-type line analog," dissertation, Roy. Inst. Technol., Stockholm, Sweden, 1985.

[26] ——, "Dynamic line analog for speech synthesis," Speech Transmission Lab., Quart. Progr. and Status Rep., STL-QPSR, vol. 1/1985, Roy. Inst. Technol., Stockholm, Sweden, pp. 1-14, 1985.

[27] D. R. Allen and W. J. Strong, "A model for the synthesis of natural sounding vowels," *J. Acoust. Soc. Amer.*, vol. 78, no. 1, pp. 58-69, 1985.

[28] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, part 2, pp. 583-590, 1971.

[29] H. Wakita and G. Fant, "Toward a better vocal tract model," Speech Transmission Lab., Quart. Progr. and Status Rep., STL-QPSR, vol. 1/1978, Roy. Inst. Technol., Stockholm, Sweden, pp. 9-29, 1978.

[30] B. Guérin, "Effects of the source tract interaction using vocal fold models," in *Vocal Fold Physiology*, J. R. Titze and R. C. Scherer, Eds. Denver, CO: The Denver Center for the Performing Arts, 1985, pp. 482-499.

[31] G. Fant, "The voice source theory and acoustic modeling," in *Vocal Fold Physiology*, J. R. Titze and R. C. Scherer, Eds. Denver, CO: The Denver Center for the Performing Arts, 1985, pp. 453-464.

[32] M. Rothenberg, "Source tract acoustic interaction in breathy voice," in *Vocal Fold Physiology*, J. R. Titze and R. C. Scherer, Eds. Denver, CO: The Denver Center for the Performing Arts, 1985, pp. 465-481.

[33] G. Fant, Q. Lin, and C. Gobl, "Notes on glottal flow interaction,"

Speech Transmission Lab., Quart. Progr. and Status Rep., STL-QPSR, vol. 2-3/1985, Roy. Inst. Technol., Stockholm, Sweden, pp. 21-45, 1985.

[34] A. S. Ananth, D. G. Childers, and B. Yegnanarayana, "Measuring source tract interaction from speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 1985, pp. 1093-1096.

[35] M. Rothenberg, "An interactive model for the voice source," Speech Transmission Lab., Quart. Progr. and Status Rep., STL-QPSR, vol. 1/1981, Roy. Inst. Technol., Stockholm, Sweden, pp. 1-17, 1981. Also in *Vocal Fold Physiology*, D. M. Bless and J. H. Abbs, Eds. San Diego, CA: College-Hill, 1981, pp. 155-165.

[36] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Commun.*, vol. 1, pp. 147-184, 1982.

[37] J. L. Flanagan and L. Landgraf, "Self-oscillating source for vocal tract synthesizers," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 57-64, 1968.

[38] J. L. Flanagan and L. Cherry, "Excitation of vocal tract synthesizers," *J. Acoust. Soc. Amer.*, vol. 45, no. 3, pp. 764-769, 1969.

[39] I. R. Titze and D. T. Talkin, "A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 60-74, 1979.

[40] I. R. Titze, "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *J. Acoust. Soc. Amer.*, vol. 75, no. 2, pp. 570-580, 1984.

[41] B. Cranen and L. Boves, "A set-up for testing the validity of the two mass model of the vocal folds," in *Vocal Fold Physiology*, J. R. Titze and R. C. Scherer, Eds. Denver, CO: The Denver Center for the Performing Arts, 1985, pp. 500-513.

[42] ——, "Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production," *J. Acoust. Soc. Amer.*, vol. 77, no. 4, pp. 1543-1551, 1985.

[43] H. Hirose, "Laryngeal adjustments in consonant production," *Phonetica*, vol. 34, pp. 289-294, 1977.

[44] H. Hirose, H. S. Park, and M. Sawashima, "Activity of the thyroarytenoid muscle in the production of Korean stops and fricatives," in *Vocal Fold Physiology*, J. R. Titze and R. C. Scherer, Eds. Denver, CO: The Denver Center for the Performing Arts, 1985, pp. 105-112.

[45] B. Hutters, "Vocal fold adjustments in aspirated and unaspirated stops in Danish," *Phonetica*, vol. 42, pp. 1-24, 1985.

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

[47] S. Maeda, "The role of the sinus cavities in the production of nasal vowels," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1982, pp. 911-914.

[48] O. Fujimura and J. Lindqvist, "Sweep-tone measurements of vocal tract characteristics," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, part 2, pp. 541-558, 1971.

[49] J. Lindqvist and J. Sundberg, "Acoustic properties of the nasal tract," Speech Transmission Lab., Quart. Progr. and Status Rep., STL-QPSR, vol. 1/1972, Roy. Inst. Technol., Stockholm, Sweden, pp. 13-17, 1972.

[50] G. Fant, L. Nord, and P. Branderud, "A note on the vocal tract wall impedance," Speech Transmission Lab., Quart. Progr. and Status Rep., STL-QPSR, vol. 4/1986, Roy. Inst. Technol., Stockholm, Sweden, pp. 13-20, 1986.

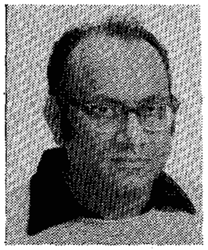
[51] R. T. Schumacher, "Ab initio calculations of the oscillations of a clarinet," *Acoustica*, vol. 48, no. 2, pp. 71-85, 1981.

[52] G. Fant, *Acoustic Theory of Speech Production*, 2nd ed. The Hague, The Netherlands: Mouton, 1970.

[53] C. Shadle, "The acoustics of fricative consonants," Doctoral dissertation, Mass. Inst. Technol., Cambridge, 1985.

[54] ——, "Turbulence noise in the vocal tract," in *Proc. 11th Int. Congr. Acoust.*, Paris, France, vol. 4, 1983, pp. 171-174.

[55] J. L. Flanagan, "Voices of men and machines," *J. Acoust. Soc. Amer.*, vol. 51, no. 5, part 1, pp. 1375-1387, 1972.



**Man Mohan Sondhi** received the B.S. degree in physics (Honours) in 1950 from Delhi University, Delhi, India; the D.I.I.Sc. degree in communications engineering in 1953 from the Indian Institute of Science, Bangalore, India; the M.S. degree in electrical engineering in 1955; and the Ph.D. degree in 1957 from the University of Wisconsin, Madison.

He has been with Bell Laboratories, Murray Hill, NJ, since 1962. Before joining Bell Laboratories, he worked for $1\frac{1}{2}$ years at the Avionics Division of John Oster Mfg. Co., Racine, WI; for 1 year at the Central Electronics Research Institute of Pilani, India; and taught for 1 year at Toronto University, Toronto, Ont., Canada. At Bell Laboratories his research has included work on speech signal processing, echo cancellation, adaptive filtering, modeling of auditory, speech, and visual processing by human beings, acoustical inverse problems, and speech recognition based on hidden Markov modeling of speech. From 1971 to 1972 he was a Guest Scientist at the Royal Institute of Technology, Stockholm, Sweden.

**Juergen Schroeter** (M'79) received the Dipl.-Ing. (EE) degree in 1976, and the Dr.-Ing. (EE) degree in 1983, Ruhr-Universitaet Bochum, West Germany.

He has been with AT&T Bell Laboratories, Murray Hill, NJ, since 1986. From 1976 to 1985 he was with "Lehrstuhl fuer Grundlagen der Elektrotechnik und Akustik," Prof. J. Blauert, Ruhr-Universitaet Bochum, where he taught courses in acoustics and fundamentals of electrical engineering, and did research in binaural hearing, hearing protection, and signal processing. At AT&T Bell Laboratories he currently is working on speech coding methods employing models of the vocal tract and vocal cords.

Mr. Schroeter is a member of ASA.