Techniques for Estimating Vocal-Tract Shapes from the Speech Signal

Juergen Schroeter, *Senior Member, IEEE,* and Man Mohan Sondhi

***Abstract*—This paper reviews methods for mapping from the acoustical properties of a speech signal to the geometry of the vocal tract that generated the signal. Such mapping techniques are studied for their potential application in speech synthesis, coding, and recognition. Mathematically, the estimation of the vocal tract shape from its output speech is a so-called inverse problem, where the direct problem is the synthesis of speech from a given time-varying geometry of the vocal tract and glottis. Different mappings are discussed: mapping via articulatory codebooks, mapping by nonlinear regression, mapping by basis functions, and mapping by neural networks. Besides being nonlinear, the acoustic-to-geometry mapping is also nonunique, i.e., more than one tract geometry might produce the same speech spectrum. We will show how this nonuniqueness can be alleviated by imposing continuity constraints.**

I. Introduction

ONE attractive approach for improving the naturalness of speech synthesizers is to employ models of the glottis (vocal cords) and the vocal tract which incorporate the physiological and physical constraints of the human speech production mechanism. Such models should also benefit low-bit-rate coders and speech recognizers. Unfortunately, so far no one has been able to demonstrate this potential advantage in a practical system. However, several steps have been taken towards achieving this elusive goal. One such step is the development of mapping techniques for estimating the shape of a talker's vocal tract from his or her speech signal. Two distinct groups of researchers are intererested in such mapping techniques. One group would like to estimate geometric properties of the tract, such as the shape of the tongue, lip rounding, etc., for the purpose of displaying these features, or for training aids for the deaf, etc. The main goal of the other group, by contrast, is to synthesize the best quality speech from the recovered shapes, for coding, text-to-speech synthesis, etc. The criteria of these two groups are not identical. In this paper we will summarize the various mapping methods—including neural networks—that have been proposed for such mappings. Our review will show that so far it has not been demonstrated that neural networks perform better than other methods for either application.

A model of human speech production synthesizes speech from slowly-varying *physiological* parameters such as lung pressure, glottal widths, shape of the tongue, coupling to the nasal cavity, and lip opening. Such a system is called an *articulatory synthesizer*. When an articulatory synthesizer is combined with methods for estimating its control parameters, we call the combined system an *articulatory speech mimic*. However, while the direct problem of synthesizing speech from a given set of time-varying articulatory parameters is well understood, the inverse problem of estimating these parameters from natural input speech is difficult because of the nonuniqueness of the acoustic-to-articulatory mapping.

The first attempt at creating an articulatory speech mimic was reported by Flanagan, Ishizaka, and Shipley in [1]. The authors closed an optimization loop around their articulatory speech synthesizer developed earlier [2] by comparing the spectra of the synthesized speech with given spectra of consecutive target speech frames of 19.2 ms duration. For each frame, an optimization procedure tried to minimize an acoustic distance between the two speech signals, thus, in effect, estimating articulatory parameters by an *analysis-by-synthesis procedure*. The authors of the current paper continued along these lines by creating a new articulatory synthesizer [3], and an articulatory speech mimic [4]. Elsewhere, similar approaches were taken (e.g., [5], [6]).

A major stumbling block in articulatory analysis-by-synthesis is the initialization of the optimization loop. Since most optimization algorithms will only find the local minimum of a given cost function that is near the initial parameters, one needs to choose good startup parameters. This can be achieved by employing an acoustic-to-articulatory mapping. One possible realization of such a map is a so-called *articulatory codebook*, that is a table of corresponding acoustic and geometric vectors [7]. The idea is to use a given acoustic representation as a key to look up (retrieve) the associated vocal-tract shape. Since articulatory codebooks can be pre-computed and searched exhaustively without computationally expensive speech synthesis, one can start off a follow-up optimization close to the global optimum. In fact, if the codebook-lookup were good enough, one might avoid the iterative optimization altogether. This last step would be essential for speech coding purposes. In this paper, we will review articulatory codebooks, as well as alternative mapping techniques.

This paper is structured as follows. Section II will discuss relevant portions of the physics of the vocal tract. In Section III we will review articulatory speech synthesis before focusing in Section IV on acoustic-to-articulatory mappings. In that section, we will discuss procedures for accessing articulatory codebooks, including dynamic programming methods of finding optimal parameter sequences to match sequences of speech spectra. Finally, we will discuss other mapping techniques,

Manuscript received April 7, 1993; revised September 15, 1993.

The authors are with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9214435.

such as regression techniques, basis functions, and neural networks. Section V concludes this paper.

II. Sound Propagation in the Vocal Tract

A. Wave Motion

Due to its complexity, it is not feasible to *exactly* model sound waves in the vocal tract. In order to make the problem at all tractable several simplifying assumptions are necessary. Almost all analyses make the following three assumptions: (a) the vocal tract can be straightened out and hence approximated as a variable-area tube; (b) the wave motion in the tract is planar, that is, pressure and velocity are constant in a plane perpendicular to the (straightened out) axis of the tract; and (c) the linear wave equation is valid.

None of these assumptions is strictly true. However, there is evidence that indicates that they are all reasonable. It can be shown, for instance, that the resonances of a straight tube of *uniform* cross-section remain almost unchanged if the tube is bent to a curvature approximating that of a vocal tract [8]. Also, the *cross*-dimension of the tract (i.e., perpendicular to the direction of wave propagation) is rarely larger than about 5 cm, which is a half wavelength at 3.5 kHz. The cross modes of sound waves in the tract are therefore negligible at frequencies lower than 3.5 kHz. Since most of the acoustic energy in a speech wave is concentrated in this range of frequencies, assumption (b) is reasonable. As for the linearity of the wave equation, calculation of Mach numbers shows that except at the vocal cords and at extremely narrow constrictions in the tract, this assumption is accurate as well [9].

In view of assumptions (a) and (b), as far as acoustical properties are concerned, the shape of a vocal tract is completely specified by the area function, $A(x)$, which specifies the cross-sectional area as a function of position along the tract. We will take $x = 0$ to be the glottis (i.e., the vocal cords) end of the tract.

Let us assume that the walls of the vocal tract are rigid, and $A(x)$ is a slowly varying function of x. Assuming that there are no viscous or thermal losses, then the pressure, $P(x, s)$, and the volume velocity, $U(x, s)$ in the tube satisfy the pair of first order differential equations

$$\frac{dP}{dx} = -\frac{\rho s}{A}U$$
$$\frac{dU}{dx} = -\frac{As}{\rho c^2}P. \tag{1}$$

In these equations, s is the complex frequency variable, ρ is the density of air, and c is the velocity of sound. The volume velocity can be eliminated from (1) to yield the second-order equation in pressure

$$\frac{d}{dx}A\frac{dP}{dx} - \frac{s^2}{c^2}AP = 0 \tag{2}$$

which is known as Webster's Horn equation [9]. Similarly, an equation in $U(x, s)$ alone can be derived by eliminating $P(x, s)$.

The effects of viscous friction, thermal conduction and yielding walls can all be approximately accounted for by appropriately modifying (1) or (2). In certain special cases, when the functions specifying the losses and wall impedance have a special form, the modification consists of just a transformation of the variable s. Such simple functional forms are adequate for many applications. In the general case, with arbitrarily distributed, frequency dependent losses and wall impedance, (2) gets modified to

$$\frac{d}{dx}M(x, s)\frac{dP}{dx} - N(x, s)P = 0. \tag{3}$$

Here the functions M, N can be computed in terms of $A(x)$, the viscous and thermal losses, and the wall impedance[10].

B. The Direct Problem

If the area function $A(x)$, the wall impedance, and the loss parameters of the vocal tract are specified, then (1) or (2) can be solved for any given boundary conditions at the lips and glottis. With a proper choice of boundary conditions, we can compute the speech signal for a variety of sounds. By way of illustration, consider the computation of nonnasalized vowel sounds.

For such sounds the boundary condition at the lips is that the tract is terminated in the radiation impedance, $Z_L(s)$. Let $H_P(x, s)$ be the solution for the pressure in the tract, which satisfies this boundary condition at the lips, and for which the volume velocity at the glottis is unity. Let $H_U(x, s)$ be the corresponding volume velocity. Then the volume velocity in the vocal tract due to any other input $U_g(s)$ at the glottis is just

$$U(x, s) = H_U(x, s)U_g(s). \tag{4}$$

In particular, the volume velocity at the lips is obtained by setting $x = L$, the length of the vocal tract. The function $H_U(L, s)$ is called the transfer function of the tract. Then $S(L, s) = U(L, s)Z_L(s)$ is just the speech signal in the frequency domain. The inverse Laplace transform of $S(L, s)$ (or the inverse Fourier transform of the function obtained by setting $s = j\omega; j = \sqrt{-1}$) gives the time domain speech signal.

A slight modification of this procedure allows one to generate fricative and nasal sounds, too. For fricative sounds, the excitation of the vocal tract is by a noise-like signal generated by turbulence at a narrow constriction somewhere inside the vocal tract. And for nasal sounds the nasal cavity gets coupled to the vocal tract. Finally, note that in natural speech, the function $A(x)$ changes continuously in time. However, these changes are, in general, slow, so that the motion of the tract can be approximated by a succession of stationary shapes.

It can be shown (see, e.g., [9]) that the transfer function $H_U(L, s)$, for vowel sounds, has no zeros (other than zeros of the radiation impedance[1]). On the other hand, it has an infinite

[1]Note that the acoustic impedance $Z(s)$ in a plane perpendicular to the direction of (1-D) wave propagation in a tube is the ratio of pressure $P(s)$ and volume velocity $U(s)$ in that plane. In the time domain, $z(t)$ is the pressure impulse response observed in that plane when we excite the tube with a volume velocity impulse in the same plane.

number of poles. In the speech literature, these poles are called *formants*. (Only the lowest 3–5 are of practical importance.) The imaginary part of a formant is called its frequency and twice the real part is called its bandwidth.

Although $H_U(L, s)$ has no zeros, the speech spectrum can have zeros even for vowel sounds. These zeros may arise for the following reasons: (a) the input volume velocity at the glottis has zeros, (b) the point of acoustic excitation is somewhere inside the vocal tract, or (c) the nasal cavity is coupled to the vocal tract.

C. The Inverse Problem

What we have described in the preceding paragraphs is the problem of computing the speech signal from a specification of articulatory information (i.e., $A(x)$, etc.). The problem of interest in the present paper is the *inverse* of this problem, that is, the problem of computing articulatory information (in particular, $A(x)$) from acoustic information that can be obtained from the speech signal. It turns out that this inverse problem does not have a unique solution. In order to show this nonuniqueness, let us first discuss the types of acoustic information that *are* sufficient to recover the tract shape.

1) Frequency-Domain Methods: Over 40 years ago Borg [11] considered an idealized form of our problem, and proved a remarkable result that allows computation of $A(x)$ from a knowledge of certain sets of eigenvalues of boundary value problems associated with (2). Stated in terms of the vocal tract problem, his result may be summarized as follows:

Consider an ideal, lossless, rigid-walled vocal tract, and impose some homogeneous lossless boundary condition at one end. For concreteness, assume the condition $U(0, s) = 0$, that is, complete closure at the glottis. Suppose further that $\lambda_1, \lambda_2, \cdots$ and $\mu_1, \mu_2, \cdots$ are the infinite sequences of eigenvalues for two independent homogeneous lossless boundary conditions at the other end. Again for concreteness, suppose these boundary conditions are (1) $U(L, s) = 0$ and (2) $P(L, s) = Z_L(s)U(L, s)$. The first of these conditions corresponds to a complete closure of the lips. The second corresponds to terminating the lips in the impedance $Z_L(s)$, which may be taken to be the radiation impedance discussed in the preceding section, provided its resistive part is negligible.

Then what Borg showed is that a knowledge of this doubly-infinite sequence of eigenvalues is sufficient to uniquely determine $A(x)$. This information is also *necessary*, in that if any one of these eigenvalues is changed, the corresponding area function changes. The two sets of eigenvalues may also be looked upon as defining the input impedance of the vocal tract looking in at the glottis. This function is given by

$$Z_{\text{in}}(s) = \frac{\prod_i (s - \lambda_i)}{\prod_i (s - \mu_i)}. \tag{5}$$

Thus Borg's theorem may be summarized by saying that $A(x)$ is uniquely determined by $Z_{\text{in}}(s)$ and vice versa. Clearly, the input impedance can be specified at either end of the tract. We used the impedance at the glottis as an illustration. *Measurement* of the input impedance is, of course, much easier at the lips.

For a lossy vocal tract we know of no frequency domain method. However, as pointed out by Atal [12] there is one very special configuration with loss, for which the inverse problem has a solution. That happens if the tract and the termination at one end are lossless, and the other end is terminated in an impedance which is lossy at all frequencies, with known loss. Clearly, this is not a useful method from a practical point of view. (See [13] for a discussion.)

There is a variant of Borg's result in which $A(x)$ is derived from the poles and *residues* of the input impedance [14]. Specification of the infinite set of poles and zeros is, of course, equivalent to specifying the infinite set of poles and residues. However, the reconstruction procedure and the approximations involved are quite different for the two approaches.

We close this section by mentioning the work of Schroeder and that of Mermelstein. Schroeder [15] showed that, to first order perturbation theory, the frequencies of the poles and zeros of the input impedance give an estimate of the odd and even coefficients, respectively, of the Fourier expansion of $\log A(x)$. Thus, if $\log A(x)$ is assumed to be antisymmetric around the midpoint of the vocal tract, then $A(x)$ can be approximately determined from just the formant frequencies. Of course, the assumption of antisymmetry is a drastic assumption which cannot be justified for a real vocal tract. Mermelstein [16] showed that measured values of the first few poles and zeros give a good approximation of the "band-limited" $A(x)$ that is, the $A(x)$ obtained by retaining only low-order Fourier coefficients of $\log A(x)$. Obviously, this method is not applicable to the problem of deriving $A(x)$ from the speech signal, because the input impedance must be measured.

2) Time-Domain Method: The methods of estimating $A(x)$ from the frequency domain data discussed in the preceding section have several disadvantages in practice. First of all, the length of the vocal tract and the boundary conditions at the two ends must be known. Each of these strongly affects the computed $A(x)$, and none of them is accurately known. Second, there is no known theory that can deal with distributed losses and yielding walls.

A very different type of method was proposed by Sondhi and Gopinath [17] which is based on the *time domain* specification of the input impedance, $z_{\text{in}}(t)$. This method alleviates both problems mentioned above. Sondhi and Gopinath showed that there is a unique one-to-one correspondence between $z_{\text{in}}(t)$ for $0 < t < T$ and $A(x)$ for $0 < x < cT/2$. Further, the method can be generalized to include the effect of losses and yielding walls [18], [19], provided that these losses are known. However, like the method of Mermelstein, this method is not useful for deriving $A(x)$ *from the speech signal*, because one needs to make a measurement of the input impedance.

3) The nonuniqueness: The poles of the input impedance at the glottis, are precisely the poles of the transfer function $H_U(L, s)$. The nonuniqueness that we mentioned earlier follows directly from this observation. For under the best of circumstances, the speech signal can at most give us the transfer function, and hence the poles of $Z_{\text{in}}(s)$. Without changing this transfer function, the *zeros* of the impedance can be specified arbitrarily, except for the constraint that they must

interlace with the poles, that is, $\lambda_1 < \mu_1 < \lambda_2 < \mu_2 < \cdots$. Each specification of zeros compatible with this constraint yields a new $A(x)$. Thus there is a nondenumerable infinity of area functions consistent with a given set of formant frequencies.

As we will see in Section IV-C, the only way to deal with this nonuniqueness is to use constraints of temporal continuity of the area function.

III. Articulatory Synthesis

Dennis Klatt wrote in [20] (p. 747): "An alternative solution to the problem of producing a natural female voice quality by a formant synthesizer might be to employ articulatory models of the trachea, vocal folds, and vocal tract, as well as their interactions, in a sophisticated articulatory synthesizer. Thus we now turn to efforts to produce speech by direct simulation of the mechanisms involved in speech generation." He continues (p. 749): "...an articulatory model" (i.e., synthesizer) "is likely to be the ultimate solution to the objective of natural intelligible speech synthesis by machine, but computational costs and lack of data upon which to base rules prevent immediate application of this approach." In the following we will discuss these issues.

Several approaches have been tried. Flanagan *et al.* [2], Portnoff [21], Maeda [5], and Bocchieri [22] solved discretized partial differential equations (PDE's). The advantage is that the implementation is simple, but the method is computationally costly. Kelly and Lochbaum [23], Kabasawa *et al.* [24], and Meyer *et al.* [6] used so-called wave digital filters (WDF's). Compared to the PDE-method, WDF's are faster, but lack flexibility in that they do not allow frequency-dependent losses and also warp the frequency axis due to the bilinear transform used to map the continuous to the discrete problem. They can accommodate time-varying area functions [25], but have problems with a variable tract length. Finally, our own synthesizer [3] uses the chain-matrix approach that analyses the tract in the frequency domain (thus allowing for realistic frequency-dependent losses) but assumes a static tract for each speech frame. Let us focus on this approach as an example for what issues need to be addressed in an articulatory synthesizer. Before we do this, however, let us define what we mean by "articulatory model."

A. Articulatory Models

Articulatory models are first in the chain of models used to transform articulatory parameters (coordinates) such as tongue center position, tongue tip position, jaw angle, velum opening, etc., to a vector of tract areas, and from there, to acoustic characteristics of the vocal tract. Based primarily on X-ray studies, articulatory models are aimed at capturing the inherent constraints of the vocal tract ("tongue cannot go through the roof of the mouth"). Hence, in creating such models, researchers are usually more concerned about geometric accuracy and less concerned about an optimal acoustic match between target and re-synthesized speech. Despite this potential drawback, such models are useful in articulatory speech mimicking since articulatory parameters usually have a lower dimensionality compared to area parameters. The lower dimensionality might reduce ambiguity in the acoustic-to-geometric mapping. On the other hand, articulatory parameters generally show a nonuniform parameter sensitivity in speech mimicking experiments and might be unable to provide good tract areas for certain sounds [26]. In this paper we will use the term "articulatory parameters" to include tract areas, since most of the techniques discussed, in principle, can be applied for articulatory model parameters, as well as (linear or log) tract areas.

Articulatory models can be static or *dynamic*, descriptive or *functional*. An example of a dynamic functional articulatory model is that of Henke [27]. It is controlled by gestures, or articulatory targets (Henke calls this "goal-driven"), that is, the model is governed by equations of motion for the articulators. It also contains a "look-ahead" text input that accounts for anticipatory coarticulation, that is, pre-adjustment of the articulators in preparation for an upcoming sound to be produced. In doing so, the speech production system takes advantage of the freedom it has in producing a target speech spectrum: for any given speech sound, there are "critical" and "noncritical" articulators [28], and different articulators can compensate for each other (e.g., [29]). For example, in the case of rounded vowels (e.g. /u/), the jaw position can be compensated for by the lip aperture (rounded lips lower all formant frequencies by lengthening the tract; raising the jaw has a similar effect [30]). Noncritical articulators are free to assume convenient positions which could mean that the motor control can anticipate future sounds where these articulators become critical. What is advantageous for speech production, of course, means ambiguities in parameter estimation. We will come back to this issue in the next section. Other examples of dynamic functional models are the one by Perkell [31] and the one developed at Haskins Laboratories (e.g., [32]). Muscular structures are simplified and modeled by springs and dampers. In these models, coproduction of speech sounds can be described in terms of the dynamics of the system.

Linear component articulatory models are based on data from X-ray studies and measurements of lip opening. Examples are the one by Kiritani *et al.* [33] and the one by Maeda [34], both of which are static, that is, do not incorporate task dynamics. Measured tongue and palate (roof of the mouth) mid-sagittal profiles are parametrized through curve fitting techniques. Curve parameters are then related to measured or inferred tract areas by a linear factor analysis. The result is a set of orthogonal articulatory parameters which are sometimes hard to interpret. It is interesting to note that, to the best of our knowledge, nobody has tried using neural networks for a nonlinear mapping between curve parameters and area data.

By far the easiest to understand are descriptive static articulatory models, such as the ones developed by Coker and Fujimura [35], by Mermelstein [36], and by Coker [37]. These models are based on X-ray mid-sagittal projections and on intuition. For example, they describe the tongue body as a "circle in a circle." Because these models are 2-D, they can compute tract areas only by incorporating certain assumptions to substitute for the missing information about the depth dimension. In our own work we primarily used Mermelstein's and Coker's articulatory models.

B. Vocal Tract Representation in an Articulatory Synthesizer

Since we assume planar and linear wave propagation, we can evaluate the vocal tract in either the time domain or in the frequency domain. We choose the frequency domain description because for frequencies above a few Hertz, we can easily incorporate effects of yielding walls (affecting frequencies and bandwidths of the lower formants), and viscous losses and radiation (affecting mainly high frequencies). For voiced speech, our frequency-domain model of the vocal tract is excited by a nonlinear model of the vocal cords. The tract and the glottal model are interfaced via Discrete Fourier Transform (DFT) and convolution. For the synthesis of aspirated or fricative sounds, the synthesizer generates noise at the glottis and at the narrowest constriction.

For a given vector of vocal tract areas that specify the tract shape from glottis to lips and the coupling to the nasal tract, we need an acoustic model to calculate the transfer function from glottal flow to pressure at lips, the transfer function from a noise source at some point in the tract to the pressure at the lips, and the input impedance of the tract as seen from the glottis. The latter is important for specifying the load to the glottal model. Acoustic characteristics of the vocal tract are computed using the chain-matrix approach.

A chain matrix relates the pressure and volume velocity at the output of a tube $[P_2, U_2]$ to those at the input $[P_1, U_1]$

$$\begin{bmatrix} P_2 \\ U_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P_1 \\ U_1 \end{bmatrix}. \tag{6}$$

Note that capital letters denote variables in the frequency domain. Computing the matrix for an arbitrary tube is complicated. However, for a tube of *uniform* area it can be computed quite easily. Therefore, we represent an arbitrary tube as a concatenation of elementary uniform sections. The overall vocal-tract chain matrix for nonnasal sounds is obtained as the product of the sequence of elementary 2×2 matrices.

Consider nonnasal sounds for simplicity, and let A, B, C, D denote the chain matrix elements of the overall tract from glottis to lips. Then the transfer function from glottal flow U_g to flow at the lips U_L is

$$H_U = \frac{U_L}{U_g} = \frac{1}{A - CZ_L}. \tag{7}$$

where Z_L is the radiation impedance at the lips. Note that the transfer function of the glottal flow U_g to sound pressure at the lips is $H_P = H_U Z_L$. In contrast to simpler source-filter synthesizers, articulatory synthesizers model source-tract interactions. These manifest themselves, for example, by formant ripples in the glottal flow waveform. The tract input impedance as seen from the glottis is given by:

$$Z_{\text{in}} = \frac{DZ_L - B}{A - CZ_L} \tag{8}$$

Since (7) and (8) have the same denominator, the tract transfer function H_U and the tract input impedance Z_{in} have identical poles.

Nasal sounds can be handled similarly, with some minor modifications. Details of the synthesizer, including the glottal model and noise excitation, can be found in [3].

IV. Acoustic-To-Articulatory Mappings

A. The Nonuniqueness of Such Mappings

In Section II we pointed out that acoustic-to-articulatory mappings generally are nonunique, that is, there is more than one tract shape that can produce a given tract transfer function. (Ventriloquists use this feature to their advantage.) This problem is compounded by another kind of nonuniqueness: changes in the source (glottis) can compensate for changes in certain features of the tract transfer function, such as spectral tilt, and/or bandwidth of the first (lowest) formant. Ignoring this second kind of nonuniqueness for the time being, one can explore conditions and characteristics of the first kind separately. Atal *et al.* [38] did fundamental work on this issue.

Atal *et al.* [38] established tables of vocal tract shapes and related acoustic representations. These tables can be used to look up pairs of vectors $(\mathbf{x}_l, \mathbf{y}_l), l = 1, \cdots, M$ of the mapping

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) \tag{9}$$

where $\mathbf{y}$ is the articulatory/geometric representation that is related to the acoustic representation $\mathbf{x}$. In one experiment the authors used the frequencies of the lowest three formants as components of $\mathbf{x}$ and four variables of a simple articulatory model as components of $\mathbf{y}$. In another experiment, they used the lowest three (log) formant frequencies and amplitudes of the vocal-tract transfer function at these formant frequencies for $\mathbf{x}$ and log-areas at twenty points along the tract for $\mathbf{y}$. Atal *et al.* explored the nonlinear mappings by linearization in small regions. For such a small region where a linear map was assumed valid, the authors used singular-value decomposition (SVD) [39] to determine the effective dimensionality of the *geometric* representation. Whenever this is larger than the effective dimensionality of the *acoustic* representation, the dimensionality of the "null space" of the mapping from acoustics to geometry, is simply the difference of the two dimensionalities. This null space is the ambiguous geometric subspace that relates to the same acoustics (the authors call it a "fiber"). In practice, the effective dimensionality of a fiber has to be determined by thresholding the eigenvalues of the system matrix and by discarding the lowest eigenvalues.

B. Articulatory Codebooks

Based on the ideas put forth in [38], we have used linked lists of vocal tract shape vectors $\mathbf{y}$ and related acoustic vectors $\mathbf{x}$ in our speech mimic [4], [40]. As mentioned in the introduction, searching such a codebook provides a set of good articulatory startup-vectors for further optimization. Generation of such a codebook consists of selecting a method of obtaining training vectors that adequately span both, the acoustic signal space, and the articulatory parameter space. We also need a method for clustering the training vectors and a definition of distance between acoustic vectors and between articulatory vectors.

There are several options for obtaining geometrical data. For example, one can employ direct geometric measurements via (microbeam) X-ray tracings or ultrasound imaging (e.g., [41]), electromagnetic "articulographs" [42], [43], or magnetic

resonance imaging (MRI) [44]. Also, vocal tract areas can be obtained by the acoustic impulse method mentioned in Section II [45]. All these methods are very cumbersome; some of them do not allow simultaneous acquisition of speech which is required for the training data. Therefore, for our purposes, we chose to use Mermelstein's and Coker's articulatory models (see Section III-A) coupled to our synthesizer (Section III-B). As will be pointed out in subsection D below, other authors actually chose to acquire, for example, X-ray microbeam data and, simultaneously, speech data for training their mappings.

Two procedures were used for obtaining geometric and acoustic training vectors from articulatory models: the root-shape interpolation method [7], and the random sampling method [46]. The reader is referred to these references for details. Suffice it to say that the root-shape interpolation method starts from predefined (in an articulatory sense) "extreme" root shapes and then fills the multidimensional articulatory space between these roots by interpolation and clustering. In contrast to this, the random sampling method explores the whole articulatory space as it is manifested in the articulatory model. Of course, while the root-shape method has the advantage of only generating "reasonable" (i.e., realistic) tract shapes, it has the disadvantage over the random sampling method that the roots are difficult to define. Choosing the wrong root shapes will create "holes" in the covered articulatory space with the consequence of bad matches for certain speech sounds. The random sampling method, on the other hand, creates lots of "unreasonable" tract shapes that have to be filtered out.

In [46], [47] we evaluated different acoustic distance measures for looking-up tract shapes in an articulatory codebook (i.e., accessing it), given a single (static) speech frame. A difficult problem with accessing such a vocal-tract codebook is the codebook's ignorance about glottal excitation. This problem is illustrated in Fig. 1. The speech spectrum (left) and the log-magnitude of the tract's transfer function (right) are very different. Note, for example, that the comb filter-like structure of the speech spectrum on the left was created by the (quasi-) periodicity of the glottal excitation. The peaks seen in the spectrum are the harmonics of the fundamental frequency (pitch) of glottal vibration. Another characteristic introduced by the glottis is spectral tilt (higher energy at low frequencies, lower energy at high frequencies) that can vary depending on speaking style (e.g., high tilt for a mellow voice and low tilt for loud shouting). Unfortunately, due to the changing radiation impedance at the lips, spectral tilt is also affected by small changes in a small lip aperture (e.g., in /u/ or /w/), without a significant effect on formant frequencies. Thus, a change in spectral tilt can be due to a certain glottal gesture or can be due to a specific tract gesture. This constitutes the second kind of nonuniqueness mentioned in Section IV-A. Another glottal effect is the apparent broadening of the first (and maybe second) formant which is due to shifting resonances and changing losses in the tract when the glottis opens and closes [48].

FFT-derived cepstral measures (see, e.g., Ch. 10 of [49]) using certain weights in the cepstral domain (so-called "lifters") combined with other weights in the frequency domain (filters) turned out to be optimal in dealing with glottal variability.

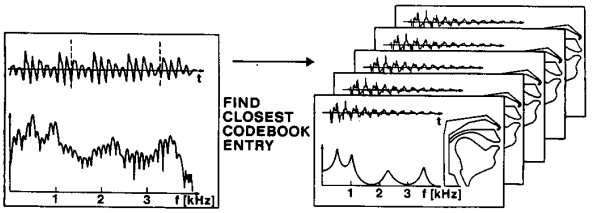


Fig. 1. Speech signal (left) and articulatory codebook (right), time (top) and frequency domain representation (bottom). Articulatory codebook entries also contain the vocal-tract shape.

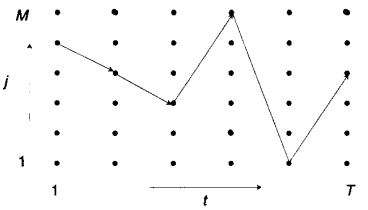


Fig. 2. Codebook access via dynamic programming. Trellis showing a possible path through a decision grid of M times T nodes. Here M is the size of the codebook, and T is the number of time steps.

Cepstral liftering discriminates between acoustic vectors based primarily on formant frequencies while largely ignoring differences in formant bandwidths and in spectral tilt. This gives the advantage of a reduced sensitivity to glottal effects. It is accompanied, however, by the disadvantage that a larger list of candidate tract shapes must be considered (with different spectral tilts, and possibly different wall-vibration losses; see Section III-B) for any given spectral vector. This means that cepstral liftering increases ambiguities of the first kind mentioned in Section IV-A. Therefore, it is important to use additional measures for ranking a list of candidate tract shapes. This is the topic of the following subsection which will also give details of the cepstral distance measures mentioned above.

C. A Dynamic Programming Method for Estimating Vocal-Tract Dynamics

Strube and colleagues [6], [50], [51] use linear and nonlinear Kalman filtering to model vocal-tract dynamics in an articulatory speech mimic. Saltzman and Munhall [32] devised a "task-dynamic" model of articulation that could be used for mimicking speech. The problem with both approaches is that one needs to have accurate information on the dynamics of the system in order to incorporate this knowledge in a speech mimic. So far, this information is largely lacking. A competing approach is to penalize large "articulatory efforts," that is, fast changes in the vocal tract, and look for smoothly evolving articulatory trajectories under the constraint of matching a given sequence of speech spectra. This is conveniently done with dynamic programming.

Consider Fig. 2 [52]. Given a codebook of M entries, and a sequence of spectral vectors $\mathbf{x}_t, t = 1, 2, \cdots, T$ for T successive frames of a speech signal, we wish to find the best sequence of shapes $\mathbf{y}_{j(t)}, t = 1, 2, \cdots, T$, where $j(t) \in [1, M]$ is the index of the codebook entry chosen for the t-th frame. Instead of exhaustively searching over all possible

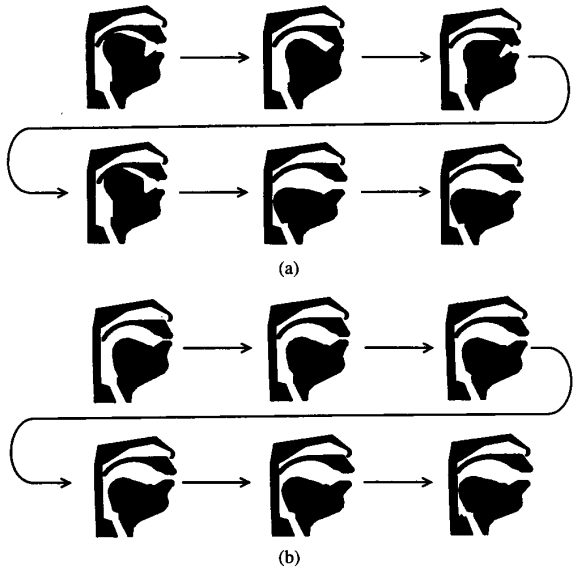


Fig. 3. (a) Sequence of tract shapes retrieved using spectral match only (independently for each frame). (b) Sequence of tract shapes retrieved using dynamic programming. (/wa/; i.e., the first 60 ms of the word "why").

M^T tract shape sequences, dynamic programming can be used to retrieve the best sequence of shapes with an effort of only $T \times M^2$ distance computations. Since M can be on the order of 100 000 or more, and (as we found experimentally) T can be 20 (the equivalent of 200 ms of speech assuming a frame shift of 10 ms) or larger, the problem is only tractable through dynamic programming. The decision space is a grid of $M \times T$ points. Let $D(\mathbf{y}_{j(t-1)}, \mathbf{y}_{j(t)})$ be the geometric cost of making the transition from shape $\mathbf{y}_{j(t-1)}$ at time $t-1$ to shape $\mathbf{y}_{j(t)}$ at time t, and $d(\mathbf{x}_t, \mathbf{x}_{j(t)})$ be the acoustic distance between the given acoustic vector $\mathbf{x}_t$ and the acoustic vector $\mathbf{x}_{j(t)}$ related to the candidate tract shape (both at time t). A practical definition of the "best" sequence of indices $j(t)$ is the sequence which minimizes the accumulated composite cost

$$C_T = d(\mathbf{x}_0, \mathbf{x}_{j(0)}) + \sum_{t=1}^{T} [d(\mathbf{x}_t, \mathbf{x}_{j(t)}) + D(\mathbf{y}_{j(t-1)}, \mathbf{y}_{j(t)})]. \tag{10}$$

Suppose that we have determined the optimal path to every node $(j, T-1), j = 1, 2, \cdots, M$ for time $T-1$, and have computed C_{T-1} for every node in the column, let's say $C(j, T-1)$. Then, it can be seen from (10) that the optimal path that ends at node $(k(T), T)$ is the one coming from the node $(j^*(k(T)), T-1)$, where j^* is given by

$$j^*(k(T)) = \arg\min_{j \in [1,M]} [C(j, T-1) + D(\mathbf{y}_j, \mathbf{y}_{k(T)})]. \tag{11}$$

Applying (11) for all k at time T, we see that the given optimal paths up to $T-1$ can be extended to T with M^2 computations. Note that at each value of T and $k(T)$, the previous node from which the optimal path arrives at node $(k(T), T)$ is known. Therefore, the complete path can be recovered by backtracking.

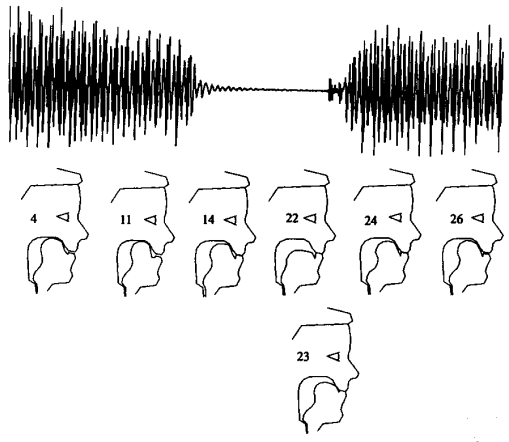


Fig. 4. Speech and retrieved tract shapes for (/ibi/). The shape at the bottom is for the burst release of the /b/. Numbers denote 10 ms frames. "Faces" are time-aligned with the speech signal.

As an example, consider Fig. 3. This figure shows tract shapes retrieved for the first 60 ms of the word "why." Here the codebook size was $M = 2523$. Each frame of speech was modeled by a vector of linear prediction coefficients (LPC's) which characterize the smoothed power spectrum [53]. In this example, the acoustic distance used was the symmetrized likelihood ratio distance [54]

$$d(\mathbf{a}, \hat{\mathbf{a}}) = 0.5 \times \left(\frac{\mathbf{a}'\hat{\mathbf{V}}\mathbf{a}}{\hat{\mathbf{a}}'\hat{\mathbf{V}}\mathbf{a}} + \frac{\hat{\mathbf{a}}'\mathbf{V}\hat{\mathbf{a}}}{\mathbf{a}'\mathbf{V}\mathbf{a}} \right) - 1 \tag{12}$$

where the $\mathbf{a} = (1, a_1, a_2, \cdots, a_p)$ are LPC vectors of order $p = 10$, the primes $'$ denote matrix/vector transposition, and the $\mathbf{V}$'s are the corresponding autocorrelation matrices. The geometric distance used was

$$D(\mathbf{y}, \hat{\mathbf{y}}) = \left[\frac{1}{n_y} \sum_{i=1}^{n_y} (y_i - \hat{y}_i)^2 \right]^{1/2} \tag{13}$$

where $n_y = 11$ is the number of components of tract-shape vectors $\mathbf{y}$ of the Mermelstein articulatory model (see Section III-A).

Tract shapes retrieved every 10 ms using a spectral match only (12) are shown in Fig. 3(a). Tract shapes retrieved for the same instants of time using the DP-procedure outlined above are shown in Fig. 3(b). While a highly irregular sequence of tract shapes is evident in the startup of the utterance when only the spectral distance is used (Fig. 3(a)), a much smoother sequence of shapes is realized by the DP-procedure (Fig. 3(b)). Note the smooth opening gesture of the lips. Results of "anticipatory articulation" (see Section III-A) can be seen in the fact that the tongue center (which is a "noncritical articulator" for the /w/) moves smoothly to the back *before* the actual start of the vowel /a/.

Other, more difficult, examples are depicted in Figs. 4 and 5. Fig. 4 shows the speech signal and retrieved tract shapes for the utterance /ibi/. A codebook of 30 000 nonnasal tract shapes, generated with the random sampling method (Section IV-B), was used. The problem here is with the silence gap in

the middle of the waveform and with the "explosion" (noise burst) at the /b/-release. Experiments indicated that the simple scheme used for obtaining the results shown in Fig. 3 fails in this case for two reasons. First, the DP-algorithm has no spectral information in the silence gap, and very inaccurate information in the noise burst (since we don't have a good frication model at the present time). Also, we conjecture that estimating tract shapes in a static manner for fricatives is more ambiguous than estimating them for voiced speech. Second, it was found that spectral information for the place of articulation for the /b/ (e.g., closure at the lips vs. closure at the tongue tip) is contained in just a few frames before and after the silence gap and the burst. Hence, the acoustic, as well as the geometric distance measures of (12) and (13), respectively, had to be modified. We did this in two ways. First, we ran two independent dynamic programs: one DP up to the onset of silence (defined as 25 dB below the maximum level of the overall energy in a frame for this utterance), and another DP that started when voicing set in after the burst. The tract shapes for the intermediate frames were obtained by extrapolation of the articulatory parameters of the Coker articulatory model from the left and right contexts. Note that the tongue center is unrealistically low for frame 22 (mid of silence gap). However, the closure at the lips is correctly extrapolated (forward) from speech frames 11 and 14, as well as extrapolated (backward) from frames 24 and 26. Second, to boost the importance of the transitionary frames before and after the silence, we introduced a time-varying weighting factor on the geometric distance that, in effect, normalized the geometric changes by corresponding spectral changes. That is, we tolerated large geometric changes whenever they were accompanied by large acoustic changes. Then the modified version of (10)

$$C_T = \alpha_0 d(\mathbf{x}_0, \mathbf{x}_{j(0)}) + \sum_{t=1}^{T} [\alpha_T d(\mathbf{x}_t, \mathbf{x}_{j(t)}) + D(\mathbf{y}_{j(t-1)}, \mathbf{y}_{j(t)})/\Delta(\mathbf{x}_{t-1}, \mathbf{x}_t)] \quad (14)$$

where the α_t's down-weighted the acoustic distance with decreasing log energy:

$$\alpha_t = \left(\frac{\max[L_t - (L_{\max} - 25 \text{ dB}), 0]}{25 \text{ dB}} \right)^{0.5}. \quad (15)$$

Here the variable L_t is the log energy in dB of frame t, and $L_{\max}$ is the maximum log energy of the utterance in dB. Consequently, $0 \leq \alpha_t \leq 1$. The acoustic distance chosen was the liftered cepstral distance (mentioned in Section IV-B)

$$d(\mathbf{c}, \hat{\mathbf{c}}) = \sum_{k=1}^{k_{\max}} w_k^2 (c_k - \hat{c}_k)^2$$

$$w_k = \begin{cases} (k/20)^{0.4}, & 0 < k \leq 20 \\ 0.5 + 0.5 \cos[\pi(k-21)/20], & 20 < k \leq k_{\max} \end{cases} \quad (16)$$

where the c-vectors were FFT-derived, frequency-domain weighted cepstral vectors of 20 ms speech frames (Hamming-windowed, 10 ms frame shift) and w is the Meyer lifter [47]. The highest cepstral order was $k_{\max} = 40$. Note that the log energy, c_0, is not used. The frame-to-frame acoustic distance $\Delta(x_{t-1}, x_t)$ is based on unmodified FFT-cepstra, equivalent to the first part of (16) with $k_{\max} = 9$ and $w_k = 1/9 \ \forall k = 1, \cdots 9$.

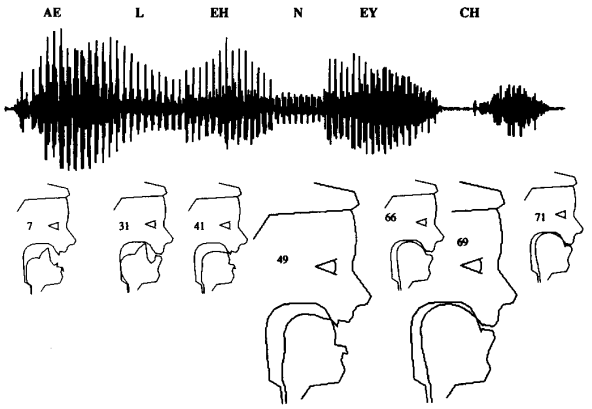


Fig. 5. Same as Fig. 4 for utterance "l-n-h" (spoken letters). The enlarged "face" for frame 49 shows the complete closure of the tract at the tongue tip (for /n/). Similarly, frame 69 stresses the correct closure at the tongue tip just before the /t/-release. Note also that the vocal tract is *not* closed at frame 71.

Fig. 5 shows results for a medial nasal /n/ and an utterance-final unvoiced dental stop /t/ in the affricate /t∫/ (as in "*ch*in") in the spelling of the letters "l", "n", and "h" (see, e.g., [55], for comprehensive information on the sounds of the American English language). To obtain the correct /n/-shape, a crude nasal detector (using just the lowest nine cepstral coefficients) was used for "declaring" the nasal "off limits" to the DP, in fact, treating it the same way as the silence gap in Fig. 4.[2] For the utterance-final affricate, only forward extrapolation could be used, correctly introducing a dental closure during the silence period. A simple *ad hoc* rule ("release the closure when nonsilent") is responsible for the open tract shape in frame 71 (which is open although the figure doesn't show it clearly).

In a more extensive "recognition" test using 204 spelled letters that were recorded from 4 male talkers over a local telephone line (using a carbon-button microphone), the resulting articulatory gestures were evaluated by visual inspection. The "error rate" of this crude "articulatory recognizer" was 34.8%, that is, 71 of the 204 letters contained at least one wrong gesture (e.g., closure at the wrong position). No (obvious) errors were made in the vowel portions.

D. Other Mapping Techniques

While the codebook idea is a simple way to obtain reasonable startup tract shapes for further optimization, it is clearly *not* the most efficient. Since the articulatory and the acoustic domain needs to be covered by points, articulatory codebooks can get very large (the largest we ever used contained 250 000 tract shapes). Exhaustively searching such a large codebook for every speech frame is very time consuming. Although sub-

[2] Our experience shows that although our synthesizer produces preceptually acceptable nasals, the spectral details of a specific nasal are very different between the synthesizer and any natural nasal. In fact, due to the large variability of the nasal cavity between talkers, different talkers also show highly different nasal spectra.

optimal techniques can be employed to make the DP-search more efficient (such as using just the top N acoustic choices; N is usually less than 1000), it still takes up to 10 minutes of CPU time on a 40 MFLOP (sustained, not peak) machine to generate startup shapes for a 2 second utterance from a 100 000 entry codebook. Clearly, a more efficient method is necessary. In the following, we will look at other methods that parametrize whole regions (instead of storing point-to-point information) of the acoustic-to-articulatory mapping.

In all the methods described below, the aim is to split the acoustic space into regions such that each region maps to a corresponding region in the articulatory domain, with a mapping that is unique in both directions. For each such region a map is defined in terms of a set of parameters, and the parameters optimized to best fit a set of examples. When a map is linear in the parameters, the optimal set of parameters is obtained by a simple matrix inversion. If the parameters enter the mapping nonlinearly, then optimization is accomplished by some iterative procedure.

1) Nonlinear Regression: Möller *et al.* [56] and Möller [57] derived vocal tract shapes from speech spectra by representing articulatory parameters as polynomials in a large number of spectral variables. This work was based on earlier work by Atal [58] who derived the map in the opposite direction and then found the inverse transformation. (It was later continued in Göttingen [59], also see Section IV-D-3.)

Suppose that $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Here $\mathbf{g}$ is a vector function, each component of which is, in general, a function of all the components of $\mathbf{x}$. The idea is to assume that over a small region around a point $\mathbf{x}^0$ each component of the articulatory vector $\mathbf{y}$ can be approximated as a polynomial function of the components of the acoustic vector $\mathbf{x}$. Then the coefficients of the polynomial may be estimated by minimizing the mean squared error over a training set of simultaneous measurements of $\mathbf{x}$ and $\mathbf{y}$. Let $x_i, i = 1, \cdots, n_x$ and $y_i, i = 1, \cdots, n_y$ and $g_i, i = 1, \cdots n_y$ be the components of $\mathbf{x}, \mathbf{y}$, and $\mathbf{g}$, respectively. Using a Taylor series to second order as the approximation, we may write

$$\hat{y}_i = y_i^0 + \sum_{k=1}^{n_x} \delta_{ik} x_k + \sum_{k=1}^{n_x} \sum_{j=1}^{n_x} \gamma_{ijk} x_j x_k \tag{17}$$

where y_i^0 is the (unknown) value of g_i at $\mathbf{x}^0$. By redefining parameters, (17) can be rewritten as

$$\hat{y}_i = y_i^0 + \sum_{k=1}^{n_x} b_{ik} \tilde{x}_k \tag{18}$$

where b_{ik} are just the coefficients δ_{ik} and γ_{ijk} in a rearranged order, and $\tilde{x}_k$ are linear or quadratic expressions in the components of $\mathbf{x}$. If all the linear and quadratic terms are included, then $N_a = n_x + N_b$, where $N_b = (n_x + 1)n_x/2$. (The process may, of course be trivially extended to cubic and higher order terms).

In matrix notation, (18) becomes

$$\hat{\mathbf{y}} = \mathbf{y}^0 + \mathbf{B}\tilde{\mathbf{x}} \tag{19}$$

and the problem is to estimate $\mathbf{y}^0$ and the matrix $\mathbf{B}$. If $(\mathbf{x}_i, \mathbf{y}_i), i = 1, \cdots, M$ is a set of simultaneous measurements of $\mathbf{x}$ and $\mathbf{y}$, then $\mathbf{y}^0$ and $\mathbf{B}$ may be derived by minimizing

$$E = \sum_{i=1}^{M} ||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2. \tag{20}$$

The solution is obtained by a matrix inversion as will be shown in Section IV-D-2 below (see (26)).

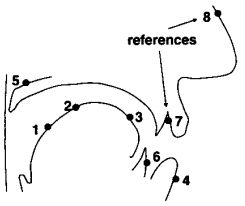


Fig. 6. Pellet positions used by Möller [57].

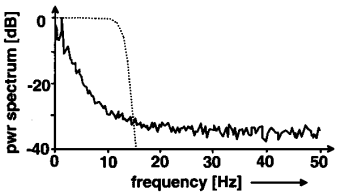


Fig. 7. Power spectrum of the horizontal coordinate of pellet 3 [57, p. 31]. The dashed line represents the low-pass filter used to suppress measurement noise.

As a first trial, Möller, Atal, and Schroeder [56] focused on the 1-D degree of coupling of the nasal cavity to the vocal tract ("velum height "). Later, Möller [57] also included 2-D (horizontal and vertical) coordinates of five additional (moving) pellets traced in 1976 by X-ray microbeam (see Section IV-B) at the Institute of Logopedics and Phoniatrics at the University of Tokyo, Japan [60]. Fig. 6 shows the location of all moving and reference pellets. Speech and pellet information was recorded simultaneously. Fig. 7 shows the measured power spectrum of the horizontal component measured for pellet 3 (tongue tip) together with the low-pass filter used. To describe the acoustic properties of speech, different spectral representations $\mathbf{x}$ were tried, including log and linear power-spectrum samples, arcsin-transformed reflection coefficients and log-area coefficients derived from LPC (for all three, see, e.g., [53]), and cepstral coefficients [49]. From the 5 pellet horizontal and vertical coordinates and the velum height, and linear combinations of these, orthogonal components were obtained by factor analysis.[3] The original pellet coordinates, as well as the orthogonal factor components were used as articulatory parameters $\mathbf{y}$.

Fig. 8 compares velum positions computed from single-word utterances containing nasals to the original X-ray measurements. Seventy-six generalized spectral components $\tilde{x}_i$ generated from 21 samples (250 Hz apart) of the smoothed log power spectrum were used. It is noteworthy that good results were obtained only after incorporating low-pass filtering into the regression (18) for estimating velum height.

[3]It can be shown that factor analysis is identical to singular-value decomposition (SVD) of the autocorrelation matrix of the given problem.

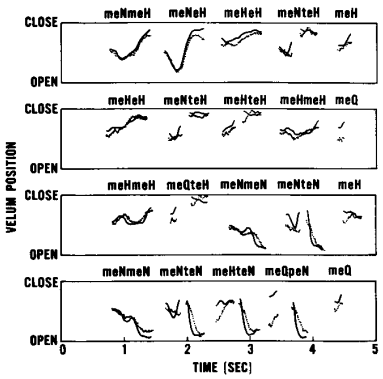


Fig. 8. Velum positions computed from speech [56]. Bold line: X-ray measurements, broken line: regression estimates.

(This finding did not carry over to other articulators in [57].) Also, it was found that about 40–50 generalized acoustic variables contributed very little to the regression, suggesting that these could be eliminated if an algorithm could be developed for their automatic identification. A sub-optimal method for eliminating inefficient generalized acoustic variables $\tilde{x}_i$ was introduced in [57].

The correlation of measured and estimated velum heights in [56] was between 0.93 and 0.98; corresponding mean-square errors were 21–37% of the measured variance of velum height. In [57], for each pellet coordinate, the minimum error varied between 40% and 87% of the measured variance. The worst pellet coordinates were the horizontal components of pellets 2 and 3 (tongue tip and middle); the best was pellet 5 (velum height). It was found that different acoustic representations $\mathbf{x}$ did not lead to significant differences in error (also found by Atal [58]). For the principal components, it was found that 80% of the pellet-coordinate variance could be explained by just two components. Errors (normalized by respective variance) for the first (most important) component were between 44% (velum height) and 59% (mid-tongue pellet). In summary, Möller found that components of tongue movement that are perpendicular to the tongue's surface (thus having a most direct influence on the tract area function) are the easiest to estimate. The most difficult problems were found with incomplete closures of the lips for /m/, too "open" constrictions/incomplete closures near and at velar consonants[4] (/n/ as in "lung", /k/, /g/), and tongue center positions that came out too high for the central vowel /ʌ/ (as in "above") and too low for the front vowel /ɛ/ (as in "get"). In respect to timing, it was found that while estimation was generally good at the center of vowels and liquids (incl. /l/ and /r/), word ends could constitute problems.

2) Basis Functions: Basis function networks can be viewed as a special class of neural networks [61] (p. 23). However, we note that what Hush and Horne call the "weights" in a (radial) basis function network are actually "centroids," that is, centers of gravity of the basis functions. As will be pointed out below, these centroids can be obtained by techniques of vector quantization or they can be set randomly.

[4]Note that only voiced speech frames were considered.

Presentation of input vectors to basis-function networks involves computing distances to these centroids (see Hush and Horne's (30) and (33) while multilayer perceptrons (MLP's), for example, use multiplicative "weighting" of the components of each input vector.

Due to the nonuniqueness of the mapping, a single global mapping for all speech is not well behaved. Locally, however, the mapping is continuous and unique almost everywhere. This observation was exploited by Parthasarathy and Sondhi [62]. In their work, mappings were learned from a large number of example pairs $(\mathbf{x}_l, \mathbf{y}_l), l = 1, 2, \cdots, M$, of acoustic vectors $\mathbf{x}_l$ and geometric vectors $\mathbf{y}_l$, which we will call the training data $\mathbf{T}$. The first step towards deriving the maps is to cluster $\mathbf{T}$ into a selected number N_x of acoustic clusters (i.e., on the basis of the similarity of the acoustic vectors). Using the iterative procedure of Linde, Buzo, and Gray [63], an optimal set of $\mathbf{N}_x$ centroids, $C \equiv [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{N_x}]$, may be identified. The set of centroids C is optimal in the sense that the average distance of a vector $\mathbf{x}$ from the nearest centroid is a minimum. Each centroid, $\mathbf{c}_k$, is associated with a cluster—the subset of training vectors $\mathbf{x}$ which are closer to $\mathbf{c}_k$ than to any other member of C.

The above clustering procedure partitions the training data into N_x clusters of pairs of vectors $(\mathbf{x}, \mathbf{y})$ which have the property that the $\mathbf{x}$ vectors within each cluster are close to each other. If the mapping of $\mathbf{x}$ to $\mathbf{y}$ were unique and continuous, then within each cluster the $\mathbf{y}$ vectors would also be close to each other. However, as mentioned earlier, the mapping from acoustic to articulatory data is not unique. Assume that each $\mathbf{x}$ maps to at most $N_y, \mathbf{y}$'s. Therefore, each cluster is partitioned into N_y subclusters, this time using the articulatory similarity (e.g., geometric distance) as the criterion for clustering.

At this stage, $\mathbf{T}$ is partitioned into a total of $N_x \times N_y$ subclusters, with the property that within each subcluster the $\mathbf{x}$ vectors are close to each other *and* the $\mathbf{y}$ vectors are also close to each other. For each of these subclusters we may define a map of the type

$$\mathbf{g}(\mathbf{x}) = \mathbf{y}^0 + \mathbf{A}\mathbf{x} + \sum_{i=1}^{N_b} \mathbf{a}_i \phi_i(\mathbf{x}). \tag{21}$$

Here $\mathbf{y}^0$ is a constant vector, $\mathbf{A}$ is a constant matrix, and the ϕ_i's are a set of N_b basis functions that span the region of the $\mathbf{x}$ domain occupied by the training data in that subcluster. A convenient choice for the ϕ_i's is a set of multidimensional gaussian functions. Thus

$$\phi_i(\mathbf{x}) = \exp[-(\mathbf{x} - \boldsymbol{\mu}_i)' \mathbf{D}_i (\mathbf{x} - \boldsymbol{\mu}_i)] \tag{22}$$

where $\mathbf{D}_i$ is a positive diagonal matrix. Since ϕ_i has ellipsoidal symmetry about $\boldsymbol{\mu}_i$, these functions may be termed "generalized radial basis functions." (The term radial basis function is used for a basis function with spherical, or radial symmetry.)

The parameters $\mathbf{y}^0, \mathbf{A}$, and $(\mathbf{D}_i, \boldsymbol{\mu}_i, \mathbf{a}_i), i = 1, 2, \cdots, N_b$ were determined for each subcluster so as to minimize the mean squared value of the mapping error for the training data. Thus if $(\mathbf{x}_i, \mathbf{y}_i), l = 1, 2, \cdots, m$ is the training data comprising a given subcluster, the parameters for that subcluster are

chosen to minimize the error e given by

$$e = \sum_{i=1}^{m} \| \mathbf{g}(\mathbf{x}_l) - \mathbf{y}_l \|^2 . \quad (23)$$

The error e may be minimized by means of a standard optimization procedure due to Hook and Jeeves [64]. In this manner, one map was derived for each of the $N_x \times N_y$ subclusters.[5]

The way in which we would use these maps for the dynamic programming (DP) search problem would be to search the centroids of the N_x acoustic clusters for the one closest to a given acoustic vector. Then the N_y maps for the selected acoustic cluster are used to compute N_y mapped articulatory vectors, which are declared as possible candidates.

To check on the feasibility of this method, Parthasarathy and Sondhi derived maps from training data $\mathbf{T}$ consisting of $M = 125\,000$ pairs taken from one of our articulatory codebooks. They chose $N_x = 64, N_y = 4, N_x \times N_y = 256, N_b = 16$. Euclidean distances were used, between acoustic vectors $\mathbf{x}$ of line spectral frequencies (LSF's, also called line spectral pairs, LSP's, derived from a tenth order LPC obtained from the tract's impulse response associated with (7); for a definition of LSF's see, e.g., [65]), and between (geometric) log area vectors $\mathbf{y}$ (20 areas from glottis to lips). To test the accuracy of their procedure, they generated 5000 test LSF vectors $\mathbf{x}_l, l = 1, \cdots, 5000$ outside the training set. These vectors were chosen to reasonably cover the space of LSF vectors. Each test vector $\mathbf{x}_l$ was mapped to $N_y = 4$ articulatory vectors $\hat{\mathbf{y}}_{kl} = \mathbf{g}_k(\mathbf{x}_l), k = 1, \cdots, N_y$, by the procedure outlined above. For each mapped articulatory vector $\hat{\mathbf{y}}_{kl}$, the LSF vector $\hat{\mathbf{x}}_{kl}$ was computed and compared to $\mathbf{x}_l$. For the closest of the $N_y = 4$ maps, the average spectral distortion was 0.33 dB.[6]. The average distance for the second best choice was 1.06 dB. Thus if the correct map is selected by the DP-algorithm then the mapping procedure outlined above can provide an accurate transformation from the spectral to the articulatory domain.

A modification of the basis-function approach was introduced by Atal and Rioul [66]. Instead of adapting the basis functions to the training set, those authors used a *large* number of *random* basis functions. This approach has the benefit of being very robust against changes in the training set. As will be detailed below, it also has the advantage of requiring only the solution of a *linear* system of equations for the weights in this network of basis functions. In the following, we will summarize Atal and Rioul's work.

Atal and Rioul redefined $\tilde{x}_k$ in (18) to be

$$\tilde{x}_k = \begin{cases} x_k \quad , & 1 \le k \le n_x \\ \phi_i\left(\sum_{j=1}^{n_x} r_{ij} x_j + \theta_i\right), & i = k - n_x, \quad n_x < k \le N_a \end{cases} . \quad (24)$$

Here, $N_a = n_x + N_b$, where N_b is the number of nonlinear functions ϕ_i. The coefficients $r_{ij}, 1 \le i \le N_b, 1 \le j \le n_x$ and θ_i, are random numbers distributed uniformly between -1 and $+1$. For simplicity, Atal and Rioul chose the nonlinear functions ϕ_i to be identical for all i

$$\phi_i(z) = \phi(z) = \frac{2}{1 + \exp(-z)} - 1 = \frac{1 - \exp(-z)}{1 + \exp(-z)} = \tan \mathrm{h}(z/2). \quad (25)$$

Note that (25) is a scaled and shifted version of the so-called sigmoid function $\check{\phi}(z) = [1 + \exp(-z)]^{-1}$ used in neural networks.[7] Note also that (19) can be rewritten as $\hat{\mathbf{y}} = \mathbf{W}\mathbf{u}$ with $\mathbf{u} = (1, x_1, \cdots, x_{n_x}, \tilde{x}_{n_{x+1}}, \cdots, \tilde{x}_{n_a})'$ an $(N_a + 1)$-dimensional known input vector and $\mathbf{W}$ an $n_y \times (N_a + 1)$ matrix of unknown coefficients. Given a large number of pairs of training vectors $(\mathbf{x}_l, \mathbf{y}_l), l = 1, \cdots, M$, an error analogous to that of (23) may be defined. Minimization of this error leads to a system of linear equations which always has a unique solution obtained directly by inverting a positive definite symmetric matrix $\mathbf{U}\mathbf{U}'$ in

$$\mathbf{W} = \mathbf{Y}\mathbf{U}'(\mathbf{U}\mathbf{U}')^{-1} \quad (26)$$

where the columns of the $(n_y \times M)$ matrix $\mathbf{Y}$ and those of the $(N_a + 1) \times M$ matrix $\mathbf{U}$ are, respectively, the $\mathbf{y}$ and $\mathbf{u}$ vectors of the training set $\mathbf{T}$ containing pairs of vectors $(\mathbf{x}_l, \mathbf{y}_l), l = 1, \cdots, M$.

For training and testing, Atal and Rioul used one of our articulatory codebooks (derived using the root-shape interpolation method and the Mermelstein articulatory model; see Section III-A) containing 10 182 nonnasal shapes. Half of the codebook was used for training, the other half was used for testing. As in [62], LSF's were used as acoustic vectors $\mathbf{x}$. LSF's and articulatory parameters were scaled to lie between -1 and $+1$. The number of random basis functions was chosen $500 \le N_b \le 1000$. In this "synthetic" test (no real speech used), different articulatory parameters showed different degrees of accuracy. The results are listed in Table I. Note that the vertical (y) coordinate of the tongue tip showed a maximum error of 15 mm. In the acoustic space, comparing the test set's acoustic results with the original, 83% of the estimated tract shapes had less than 4 dB spectral error. Note that this spectral distortion is much worse than the result reported above for Parthasarathy's and Sondhi's experiment.

In another experiment using LSF's obtained from 5000 speech frames for training and another 1000 for testing, results

TABLE I
ESTIMATION OF MERMELSTEIN ARTICULATORY PARAMETERS FROM LINE SPECTRAL FREQUENCIES USING RANDOM BASIS FUNCTIONS [66]

Parameter	Range	MaxError
Jaw angle	0.29–0.36 rad	0.035 rad (2 deg)
Tongue center x	6.0–8.5 cm	2.0 mm
Tongue center y	3.7–6.3 cm	3.0 mm
Tongue tip x	7.5–13.0 cm	10.0 mm
Tongue tip y	2.0–5.5 cm	15.0 mm
Lip position x	0.2–1.2 cm	3.0 mm
Lip position y	–0.05–0.4 cm	1.6 mm
Hyoid position x	6.1–6.4 cm	1.4 mm
Hyoid position y	8.45–9.0 cm	2.1 mm

[5]Note that if the shapes of the basis functions are fixed, that is, if $(\mathbf{D}_i, \boldsymbol{\mu}_i)$ are specified *a priori*, then A and the vector $\mathbf{a}_i$ can be obtained by just a matrix inversion analogous to the one in (26) below.

[6]The stated average distortion was computed from LPC-spectra (order $p = 10$). Note that a spectral distortion of less than about 1 dB is considered inaudible.

[7]Digital signal processing experts realize the similarity of this equation to the bilinear transform (e.g., [49], p. 207).

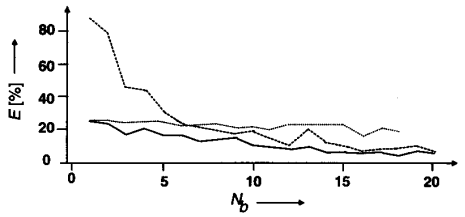


Fig. 9. Relative error E versus the number of nonlinear nodes N_b. Dotted: back propagation with linear part; solid: random basis functions including linear terms; dashed: random basis functions excluding linear terms.

from the random basis function method were compared to those of the "more traditional" back-propagation algorithm (started from a random set of weights; see next section of this paper) on networks of identical topology. In Fig. 9, the relative error $E\left(\sqrt{e/|\mathbf{y}|^2}\right)$ (E being the expectation operator; e from (23) with $m = M$) is plotted against the number of nonlinear nodes N_b (in the "hidden", intermediate layer). Not surprisingly, it was found that the back-propagation algorithm is slower by 2 to 3 orders of magnitude. Therefore, no comparison was done for $N_b > 20$. As can be seen from Fig. 9, the back-propagation performed worse than the random basis function approach. Finally, note that for N_b large enough, the linear part of the mapping can be approximated quite well even without including the linear term in (24).

Neural Networks Employing Multilayer Perceptrons (MLP's) Multilayer perceptrons (MLP's) have become the workhorse of the neural network community. This is also true of the application of neural networks to the vocal-tract inference problem. Therefore, this subsection will focus—with one exception (i.e., the work done at the 3rd Physical Institute in Göttingen, FRG)—on MLP's. Instead of rehashing the basic principles of this kind of network, we refer the reader to the excellent tutorials by Lippmann [67] and by Hush and Horne [61].

As an introduction, let us mention the work of Soquet *et al.* [68]. These authors used a 1-hidden-layer MLP (with ten units in the hidden layer) to estimate 30 tract areas given target values for the lowest three formant frequencies of 11 French vowels. Note that, for reasons stated in Sections II-C and IV-A, three formant frequencies are not enough to uniquely identify the tract area function. Hence, the vocal-tract was constrained to have a constant volume of 85 cm^3 and also to have specific geometrical symmetries. In addition, the objective function contained a term that penalized tract shapes that were spatially rough. Note that we, too, found this necessary when optimizing tract areas directly [26].

Shirai and Kobayashi [69] reported on using a 2-hidden-layer MLP to estimate articulatory parameters from voiced speech. Their MLP consisted of a 12-node input layer (for 12 LPC-derived cepstral coefficients), two 24-node hidden layers, and a four-node output layer providing parameters for their articulatory model (tongue center, tongue height, jaw opening, and lip rounding). The total number of weights in the network was 960. The network was trained on 250 speech segments from one male talker (50 segments for each of the five Japanese vowels /a/, /i/, /u/, /e/, and /o/ spoken without pauses in between; a total of 2778 speech frames). Then, using their existing speech mimic, they estimated articulatory parameter tracks including all inter-vocalic glides. These parameter tracks were finally used in training the network. The test data consisted of 5094 speech frames outside the training set. The normalized error[8] was 0.059 on the training set (averaged over the 5 vowels), and 0.051 on the testing set for /a/, /e/, and /o/. For the other two vowels, the averaged error was 0.163. Shirai and Kobayashi attributed this finding to a problem in their existing speech mimic rather than to the neural network. More interesting, however, is the finding that the neural network took less than 10% of the computer time for estimating articulatory parameters than did their (nonneural net) speech mimic.

Papcun *et al.* [28] avoided the use of a mimic by employing geometric data acquired by X-ray microbeam. Three male students provided this data together with their speech. Articulator positions were acquired from pellets at the middle of the lower lip, at the midline of each subject's tongue at distances of 10 and 60 mm from the tongue tip. Only the vertical movements of these pellets were reported in the paper. (Note Möller's conjecture about the higher difficulty of estimating horizontal pellet movements at the end of Section IV-D-1.) Utterances consisted of identical consonant-vowel utterances CəCəCəCəCə where ə is the first vowel in "above". Each utterance contained one consonant drawn from either /t,l,p/, /d,v,j/ , /g,b,ð/ (/ð/ as in "thy"), or /k,ʒ,s/ (/ʒ/ as in "measure"). Frames of 15.98 ms and a frame shift of 50% were used. Sixteen Bark-scale (i.e., auditory-filter) spectral components between 200 Hz and about 4 kHz represented the acoustics. Inputs were scaled to lie in the range 0 to 1. Pellet data were scaled to lie between 0.1 and 0.9. A separate network was trained for each of the three pellets. All three networks had the same structure. Different from Shirai and Kobayashi (and from Möller [57]), Papcun *et al.* combined data from 25 frames consecutive in time (covering a total time span of 207.74 ms) to form one input vector (the authors called this a "context frame"). Each network consisted of 400 input nodes (25 frames times 16 spectral values), two hidden layers of 8 units each, and one output unit. The standard sigmoid nonlinearity was used. In addition, output vectors in the testing phase were smoothed over a 10-frame window (20 frames in training). Only every fourth input/output vector pair was used for training which was done by standard backpropagation (see, e.g., [61], p. 13) with a momentum term weighted by 0.3.

For evaluation, Papcun *et al.* [28] used two traditional measures: rms-error and Pearson product-moment correlation. The rms-error reflects the overall distance between estimated trajectories and measured ones. Correlation compares the similarity of the shapes of these trajectories (i.e., whether both rise or fall in synchrony), discarding magnitude altogether. The highest correlations (>0.88) were found for the "critical" articulators (i.e., lower lip for the bilabials /p/ and /b/, the tongue tip for the alveolars /t/ and /d/, and the tongue dorsum for the velars /k/ and /g/), while the "noncritical" articulators showed correlations as low as 0.19. Surprisingly, the authors found that the rms-error was usually higher for the critical articulator than for the two others (with the exception of the

[8]Note that this error is *not* a recognition error for vowels but the average estimation error of the articulatory parameters.

/d/). After renormalizing the pellet coordinates individually for each dimension instead of normalizing across all pellets, this anomaly vanished. This means that the greater range of the critical articulator was the reason for this effect. In addition to rms-error and correlation, Papcun *et al.* used the "gesture recognition error rate" of a template-based recognizer as a more global measure of performance. Articulatory templates were extracted from the release of the /Cɔ/ syllables. Five repetitions and speech of three talkers each using two voicing conditions (thus producing a total of 30 syllables) were averaged for obtaining the nine templates (3 consonant categories times 3 pellets). In testing, the template that resulted in the lowest rms-error was selected as the best match. All but one out of 90 gestures in the test set were correctly recognized. Worse results were obtained using more stringent testing schemes.

In Göttingen [70]–[72], researchers devised competitive networks using the "counter-propagation" (CP) method of Hecht-Nielsen [73]. According to the authors, this method has the advantage of allowing tradeoffs between errors made in the geometric vs. errors made in the acoustic domain. In CP-networks, an intermediate layer of cells represents classes of input and output variables in the sense of (self-organized) vector quantization (VQ) of the one- or even bi-directional mapping. The network topology is symmetric, leading to two combined input/output (i/o) layers (one for each domain, acoustic and articulatory). The intermediate layer forms clusters of training pairs $(\mathbf{x}, \mathbf{y})$ using a winner-take-all strategy.

Let L_X be the i/o-layer for the acoustic domain and let $L_\mathbf{Y}$ be the i/o-layer for the geometric domain. Furthermore, let us denote the intermediate layer as L_{VQ}. Different from [73] where scalar products were maximized (requiring normalized training patterns), and similar to our basis function networks (Section IV-D-2), Strube and colleagues minimized Euclidean distances during training of their network as follows:

Step 1: Let $\mathbf{v}_i = \left(v_1^i, \cdots, v_{n_x}^i\right)$ be the weight vector connecting the i/o-layer L with L_{VQ}. Similarly, let $\mathbf{w}_i = \left(w_1^i, \cdots, w_{n_y}^i\right)$ be the weight vector connecting the i/o-layer $L_\mathbf{Y}$ with L_{VQ}. For every vector pair $(\mathbf{x}_l, \mathbf{y}_l)$ in the training set, select a winner node $i = i^*$ using

$$i^* = \arg\min_{i \in [1, n_{VQ}]} \left[r \parallel \mathbf{x}_l - \mathbf{v}_i \parallel^2 + (1 - r) \parallel \mathbf{y_l} - \mathbf{w_i} \parallel^2 \right], 0 \le r \le 1. \quad (27)$$

Here r is a parameter determining the relative influence of both domains, acoustic and articulatory, respectively, and n_{VQ} is the number of nodes in L_{VQ}

Step 2: Adapt the weights $\Delta \mathbf{v}_{i^*} = \alpha(t)(\mathbf{x}_i - \mathbf{v}_{i^*}), \Delta \mathbf{w}_{i^*} = \alpha(t)(\mathbf{y}_i - \mathbf{w}_{i^*})$. Learning parameter $\alpha(t)$ decreases with iteration number t, as is done in the classic stochastic approximation algorithm of Robbins and Monro [74].

Step 3: To generate articulatory output vectors, optional, so-called "outstar", weights $\mathbf{u}$ between L_{VQ} and L_Y can be adapted using $\Delta \mathbf{u}_i^* = \beta(t)(\mathbf{y}_l - \mathbf{u}_i^*)$. (Alternatively, the $\mathbf{w}$'s could be used.) Again, $\beta(t)$ is a learning parameter that decreases with time. The advantage of using outstar weights is to "decouple" the selection of the winner node (27) from the generation of output vectors. Note that to estimate articulatory parameters from input speech, no outstar weights are needed between L_{VQ} and L_X.

For training, steps 1–3 are repeated many times over the total set of acoustic/articulatory vector pairs. For testing, one presents only the acoustic "key" $\mathbf{x}$ to the network. The winner node i^* minimizing $||\mathbf{x} - \mathbf{v}_i||^2$ is identified and the corresponding "centroid" $\mathbf{u}_{i^*}$ is returned as the output. It is obvious that for estimating articulatory from given acoustic vectors in training step 1, r should be close to unity. For $r = 1$ we minimize the quantization error in the acoustic domain. The following idea, however, justifies $r < 1$. In cases where multiple y's exist for any given $\mathbf{x}$, a network trained with $r = 1$ would cluster all corresponding $\mathbf{y}$'s into one cluster producing, in testing, an estimate $\mathbf{u}_{i^*}$ that is close to the average $\bar{\mathbf{y}}$ of all $\mathbf{y}$'s seen in training. However, if $r < 1$ during training, a vector pair $(\mathbf{x}_l, \mathbf{y}_l)$ will update only the centroid $\mathbf{v}_i^*$., in effect, creating a bias towards the one that is closest to $\mathbf{y}_l$.

In their work, Strube and colleagues used an 11-parameter articulatory model [71] which is similar to the ones created by Mermelstein or by Coker (see Section III-A). The acoustic vectors were composed of 12 LPC-PARCOR coefficients [53] that were obtained from 30 ms long *synthetic* speech frames (Hamming windowed) with a frame shift of 2.5 ms. Networks with up to 800 nodes in L_{VQ} were tried. The parameter r was chosen to be 0.85. Output vector components were median-filtered with a window of 13 samples (corresponding to a time frame of 13 times 2.5 ms = 32.5 ms). Due to the VQ-nature of the network, output errors tend to manifest themselves by discontinuities (going from one cluster to another) instead of the more "noise-like" errors seen in the work of, for example, Möller [56][9]. Most recently [75], the group in Göttingen adapted our work on DP-based mappings [52,77] which led to much smoother results (in terms of articulatory parameter trajectories and also in terms of quality of the re-synthesized speech). Also, they replaced the PARCOR coefficients by Bark-scale power spectra raised to an exponent smaller than one.

Several alternatives were tried for training the network. The instantaneous adaptation described above leads, for large α, to what the authors call "dynamic" adaptation. This means that the same node in L_{VQ} gets adapted all the time because, for speech, the current input pair of vectors is likely to be close to the previous one. For small α, the convergence is very slow. The problem can be solved by the following procedure [75]. For each speech frame, sort the nodes in L_{VQ} with respect to how well they match the acoustics of the current speech frame. Then adapt all nodes in L_{VQ} according to their rank, that is, make $\alpha(t)$ and $\beta(t)$ (see step 2 above) for each node depend on a linear combination of acoustic and articulatory distances. This way, for each frame the best node is updated the most, followed by the second best node, etc. In addition, as time t (i.e., the number of iterations) increases, the algorithm is made to focus more on the best nodes. This leads to the advantage that, in the beginning, all nodes learn, but as time progresses the tendency is for only the best node to be adapted.

[9] This could be alleviated by using the interpolation mode of the CP-network as described in [73].

Consequently, all nodes of the network move quickly in the beginning to cover the relevant sub-space. Over time, however, more and more specialization of each node occurs. This idea is based on the "Neural-Gas" algorithm [76].

The end result of training a CP-network is a codebook of $(\mathbf{x}, \mathbf{y})$ pairs much like the ones described in Section IV-B. It remains to be seen what advantages result from using this particular training procedure.

Rahim *et al.* [77] used a wave-digital filter articulatory synthesizer (see introduction III) to explore the use of MLP's for speech mimicking. To overcome the fundamental nonuniqueness of the acoustic-to-articulatory mapping, the authors applied the concept of dynamic programming (Section IV-C) to training and accessing an assembly of competing mappings (an idea that came out of Parthasarathy's and Sondhi's work; see Section IV-D-2). This method has the advantage of maintaining the inherent nonuniqueness of the mapping while allowing for smooth transitions from one MLP to the next. Mermelstein's articulatory model was used with the tract length fixed at 17.5 cm.[10] The networks, initially trained on one of our Mermelstein codebooks of 75,238 tract shapes, were further trained ("bootstrapped") on 20 minutes of voiced speech from male talkers.[11] The acoustic vector $\mathbf{x}$ consisted of 18 FFT-derived cepstral coefficients, liftered by the Juang lifter [78]. Ten log areas represented the tract shapes $\mathbf{y}$.

Slightly different from Parthasarathy and Sondhi in Section IV-D-2, Rahim *et al.* used $N_x = 32$, (and again), resulting in a total number of clusters $N_x \cdot N_y = 128$. For comparison, a single MLP was trained with two hidden layers of 140 and 60 nodes, respectively. The 128 competing MLP's all had only a single hidden layer of 26 nodes.

Fig. 10 shows the concept of using dynamic programming for deciding which network to employ for any given speech frame. (This figure is analogous to Fig. 2.) DP-access to the assembly of MLP's can easily be done *after* the networks are trained. During training, however, the problem is more difficult. How do we select which network should be adjusted for a given speech frame? Rahim *et al.* tried several approaches, among them adjusting networks along the optimal DP-path, and adjusting the top N acoustic matches only. The latter algorithm turned out to be better, possibly due to the "dynamic allocation" problem noted by Strube and colleagues (the same mapping used for the previous speech frame tends to be chosen also for the current frame; see above) ; N was chosen to be 6. Consequently, it is noteworthy that re-training the networks on 20 minutes of natural speech improved the quality of the synthetic speech most dramatically for the single MLP (by about 1 dB in spectral distortion), and to a much lesser extent (about 0.4 dB) for the assembly of MLP's. The final result was a spectral distortion of 1.9 dB (averaged over three voiced test sentences). Note that this number cannot be compared, for example, to Parthasarathy's and Sondhi's results because those numbers were not obtained on real speech data.

Figs. 11–14 show, respectively, the original speech, synthetic speech from DP-codebook access, synthetic speech from the single MLP, and synthetic speech from the assembly of neural networks. The WDF-synthesizer was used for all synthetic speech. Single MLP and assembly of MLP's had been trained on (other) natural speech before seeing this sentence for the first time. Note that energy of synthetic and original speech was not matched, resulting in some obvious errors in amplitudes. Synthesis was done pitch-synchronously using a preset glottal excitation waveform. Output vectors $\mathbf{y}$ were median-filtered over five pitch epochs.

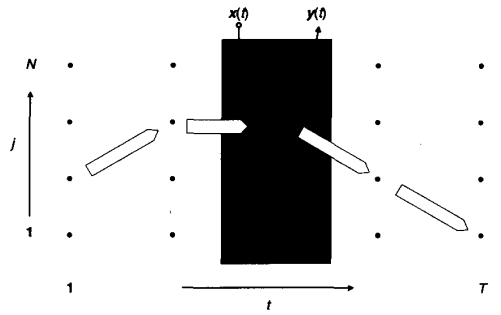


Fig. 10. Extension of the dynamic programming concept introduced in Fig. 2. Here we use multilayer perceptrons instead of articulatory codebook entries. Shown is a trellis outlining a possible path (large white arrows). N is the number of MLP's competing with each other.

Although the assembly of MLP's clearly outperformed the single MLP, and also showed a lower distortion compared to the codebook-lookup, the figures clearly show some weaknesses. First, the traces of the third and fourth formants are less smooth for the networks compared to the codebook, implying tract areas that were rougher in a spatial sense (glottis to lips). This could be due to the fact that log areas were adapted instead of parameters of the articulatory model and/or due to the fact that no spatial smoothness measure was applied (for such a measure see, e.g., [79]). Second, the WDF-synthesizer is lacking the capability of lengthening the vocal tract when necessary. This is the case for the /r/ (all formant frequencies low at about 1.4 s into the utterance). Also, the representation of losses in the vocal tract was inadequate. Finally, note the fact that the networks went through a phase of training on natural speech while the codebook did not. However, the assembly of networks is computationally more efficient than the codebook-lookup: it used only about 4% of the memory and 5% of computation time. It also could be bootstrapped from speech, while doing the same for a large articulatory codebook is much more laborious.

V. Summary and Conclusion

In this paper we have reviewed techniques available for inferring the shape of the vocal tract from the speech signal. The paper is aimed at readers interested in applying neural network methods to problems in the analysis and synthesis of speech.

After a discussion of the fundamentals of sound propagation in the vocal tract, we briefly mentioned the direct and inverse problems for the vocal tract, and indicated how to implement an articulatory synthesizer. The discussion led to the important topic of the nonuniqueness of the acoustic-to-articulatory

[10] Note that the fixed tract length is required by the WDF-synthesizer.

[11] Silent and unvoiced portions had been excluded from the training database.

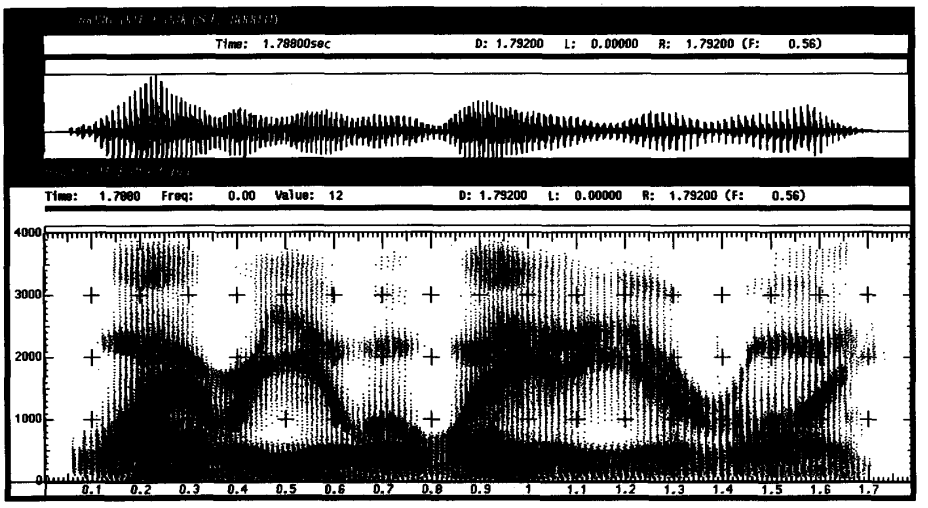


Fig. 11. Waveform and wideband spectrogram of the original sentence "Why were you away a year, Roy?" spoken by a male talker.

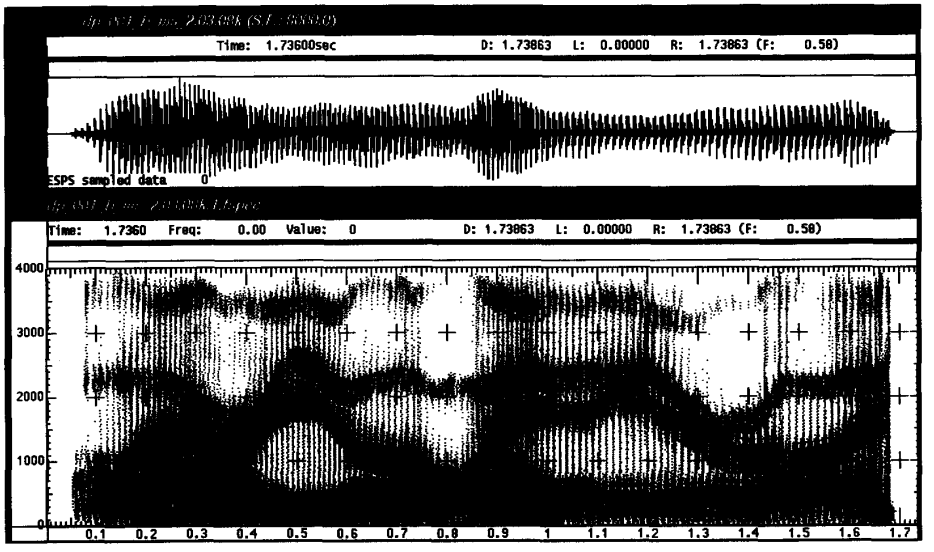


Fig. 12. Waveform and wideband spectrogram of same sentence synthesized from tract shapes retrieved from articulatory codebook using dynamic programming. Avg. distortion=2.03 dB.

domain mapping. We saw that the acoustic input impedance of the tract uniquely specifies the area function while the transfer function does not. We defined two kinds of nonuniqueness. The first kind is due to the fact that different tract shapes may have (almost) the same transfer function. The second kind arises from the fact that the same speech spectrum may be produced by two different tract shapes with appropriately selected inputs at the glottis (vocal cords). Both types of nonuniqueness have to be dealt with in an articulatory analysis/synthesis system.

The discussion of vocal tract fundamentals was followed by a review of work done over the past 20 years on acoustic-to-articulatory mappings. This includes work on articulatory codebooks, which are point-to-point mappings, as well as work on sets of parametric mappings, each covering a region of the acoustic space. The latter category includes mappings derived by nonlinear regression, expansions in terms of trained and randomly selected basis functions, multilayer perceptrons and counter propagation neural networks. In all these methods, the nonuniqueness is resolved by demanding temporal continuity. Optimal paths satisfying continuity constraints may be efficiently found by dynamic programming.

An important idea for alleviating the ambiguities in acoustic-to-articulatory mappings is the use of local (one for a subset of speech sounds), as opposed to global (one for all speech sounds) mappings. The acoustic-to-articulatory mapping is well-behaved only locally. Note, however, that the neighborhoods for the local mappings have to be determined carefully.

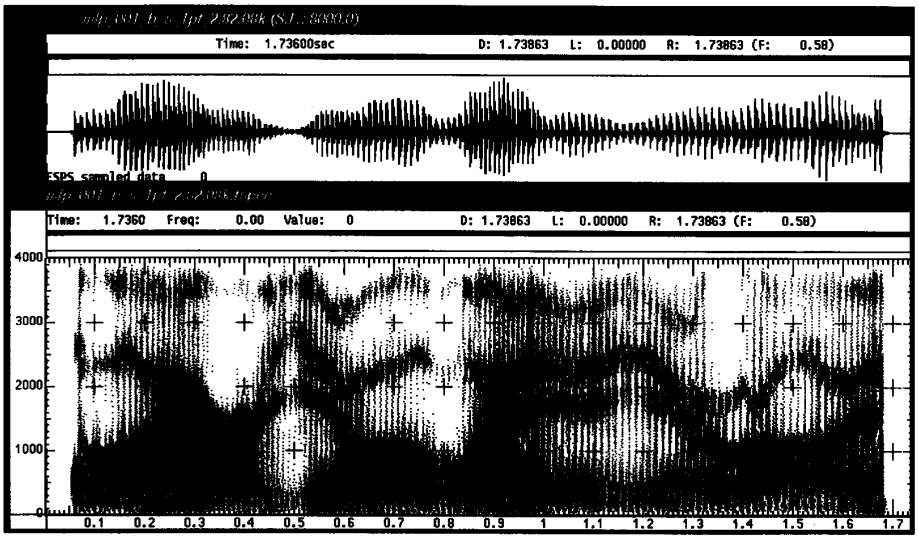


Fig. 13. Same for single MLP trained on natural speech. Avg. distortion=2.82 dB.

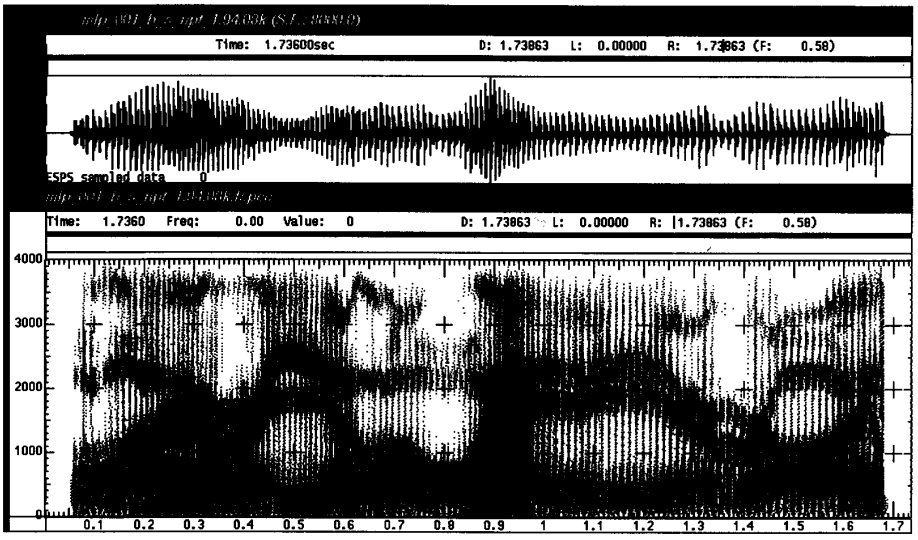


Fig. 14. Same for DP-accessed assembly of neural networks trained on natural speech. Avg. distortion=1.98 dB.

For the mapping problem the following challenges remain. So far, no one has successfully derived acoustic-to-articulatory mappings for *all* classes of speech sounds. Adequate mappings exist for small subsets, for example, for voiced speech only, or for "simple" consonant-vowel transitions. No good mappings exist at the present time for fricatives, stops, and nasals. For use of the articulatory approach in speech synthesis, recognition and coding, the mapping procedure must accommodate a variety of speaking styles, talkers and recording environments. It also has to include the effects of articulatory gestures related to the excitation (e.g., voiced, unvoiced, and transitions from one to the other). It seems to be highly desirable to develop approaches for estimating the parameters of a glottal model from the speech signal, in ways similar to the ones outlined in this paper for the vocal tract. For estimating the vocal tract *and* the control parameters of a nontrivial glottal model, we will encounter even more severe ambiguities than for estimating the vocal tract alone.

Concerning the applicability of neural networks to the mapping problem, it should be apparent from our discussion in Section IV-D that no clear advantage has so far been shown for them compared to other approaches. Vector quantization or optimization of expansions in terms of basis functions usually give better mappings than those derived by neural networks. However, neural networks and expansions in terms of basis functions might have an advantage in computational speed.

It remains to be seen if novel and improved applications of neural nets can actually provide significantly better mappings than the other approaches.

REFERENCES

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

[2] 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. 199–229, 1975.

[3] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 7, pp. 955–966, July 1987.

[4] J. Schroeter, J. N. Larar, and M. M. Sondhi, "Speech parameter estimation using a vocal tract/cord model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1987, pp. 308–311.

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

[6] P. Meyer, R. Wilhelms, and H. W. Strube, "A quasiarticulatory speech synthesizer for the German language running in real time," *J. Acoust. Soc. Am.*, vol. 82, no. 2, pp. 523–539, 1989.

[7] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 12, pp. 1812–1818, 1988.

[8] M. M. Sondhi, "Resonances of a bent vocal tract," *J. Acoust. Soc. Am.*, vol. 79, no. 4, pp. 1113–1116, 1986.

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

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

[11] G. Borg, "Eine Umkehrung der Sturm-Liouvilleschen Eigenwertaufgabe," ("An inversion of the Sturm-Liouville eigenvalue problem," in German) *Acta Mathematica*, vol. 78, pp. 1–96, 1946.

[12] B. S. Atal, "Determination of the vocal tract shape directly from the speech wave," *J. Acoust. Soc. Am.*, vol. 47, p. 65(A), 1970.

[13] M. M. Sondhi, "Estimation of vocal-tract areas: the need for acoustical measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 3, pp. 268–273, 1979.

[14] B. Gopinath and M. M. Sondhi, "Determination of the shape of the human vocal tract by acoustic measurements," *Bell Syst. Tech. J.*, vol. 49, no. 6, pp. 1195–1214, 1970.

[15] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.*, vol. 41, no. 4, pt. 2, pp. 1002–1010, 1967.

[16] P. Mermelstein, "Determination of vocal tract shapes from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41, no. 5, pp. 1283–1294, 1967.

[17] M. M. Sondhi and B. Gopinath, "Determination of vocal tract shape from impulse response at the lips," *J. Acoust. Soc. Am.*, vol. 49, no. 6, pt. 2, pp. 1867–1873, 1971.

[18] M. M. Sondhi and B. Gopinath, "Determination of the shape of a lossy vocal tract," in *Proc. Seventh Int. Congr. Acoust.* (Budapest, Hungary), 1971.

[19] J. R. Resnick, "Acoustic inverse scattering as a means for determining the area function of a lossy vocal tract: theoretical and experimental model studies," Ph.D. dissertation, Johns Hopkins Univ., Baltimore, MD, 1979.

[20] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987.

[21] 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.

[22] E. L. Bocchieri, "An articulatory speech synthesizer, Ph.D. dissertation, Univ. of Florida, Gainesville, 1983.

[23] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. Fourth Int. Congr. Acoust.*, 1962, pp. 1–4, reprinted in *Speech Synthesis*, J. L. Flanagan and L. R. Rabiner, Eds. Stroudsburg, PA: Dowden, Hutchinson & Ross, 1973, pp. 127–130.

[24] 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.

[25] H. W. Strube, "Time-varying digital filter and vocal tract models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1982, pp. 923–926.

[26] J. Schroeter, J. N. Larar, and S. Parthasarathy, "Vocal-Tract Areas versus Articulatory Parameters in Speech Production Modeling," *J. Acoust. Soc. Am.*, vol. 84, suppl. 1, S127.

[27] W. L. Henke, "Dynamic articulatory model of speech production using computer simulation," Ph.D. dissertation, M.I.T., Cambridge, MA, 1966.

[28] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *J. Acoust. Soc. Am.*, vol. 92, no. 2, pp. 688–700, 1992.

[29] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modeling* (NATO Advanced Study Institute series), W. J. Hardcastle and A. Marchal, Eds. Norwell, MA: Kluwer, 1990.

[30] B. E. F. Lindblom and E. F. Sundberg, "Acoustical consequences of lip, tongue, jaw and larynx movement," *J. Acoust. Soc. Am.*, vol. 50, no. 4 (pt. 2), pp. 1166–1179, 1971.

[31] J. S. Perkell, "A physiologically-oriented model of tongue activity in speech production," Ph.D. dissertation, M.I.T., Cambridge, MA, 1974.

[32] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychol.*, vol. 1, no. 4, pp. 333–382, 1989.

[33] S. Kiritani, S. Sekimoto, and H. Imagawa, "Parameter description of tongue point movements in vowel production," *Annu. Bull. RIHLP*, vol. 11, pp. 31–37, 1977.

[34] S. Maeda, "Un modele articulatoire de la langue avec des composantes lineaires," ("An articulatory model of the tongue with linear components," in French) in *10 è mes Journèes d' Étude sur la Parole* (Grenoble, France), May 30–June 1, 1979.

[35] C. H. Coker and O. Fujimura, "A model for specification of vocal tract area function," *J. Acoust. Soc. Am.*, vol. 40, p. 1271(A), 1966.

[36] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, no. 4, pp. 1070–1082, 1973.

[37] C. H. Coker, "A model of articulatory dynamics and control," *Proc. IEEE*, vol. 64, no. 4, pp. 452–460, 1976.

[38] 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. Am.*, vol. 63, no. 5, pp. 1535–1555, 1978.

[39] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: The Johns Hopkins University Press, 1989.

[40] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in: *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Dekker, 1992, pp. 231–267.

[41] M. Stone, "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2207–2217, 1990.

[42] B. Tuller, S. Shao, and J. A. S. Kelso, "An evaluation of an alternating magnetic field device for monitoring tongue movements," *J. Acoust. Soc. Am.*, vol. 88, no. 2, pp. 674–679, 1990.

[43] J. S. Perkell, *et al.*, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *J. Acoust. Soc. Am.*, vol. 92, no. 6, pp. 3078–3096, 1992.

[44] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels," *J. Acoust. Soc. Am.*, vol. 90, no. 2, pt. 1, pp. 799–828, 1991.

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

[46] J. Schroeter, P. Meyer, and S. Parthasarathy, "Evaluation of improved articulatory codebooks and codebook access distance measures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1990, pp. 393–396.

[47] P. Meyer, J. Schroeter, and M. M. Sondhi, "Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks," *IEEE Trans. Signal Processing*, vol. 39, no. 7, pp. 1493–1502, 1991.

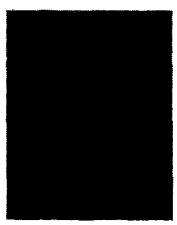
[48] J. N. Larar, Y. A. Alsaka, and D. G. Childers, "Variability in closed phase analysis of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 1985, pp. 1089–1092.

[49] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[50] R. Wilhelms, "Schätzung von artikulatorischen Bewegungen eines stilisierten Artikulationsmodells aus dem Sprachsignal" ("Estimation of articulatory movements of a stylized articulatory model from the speech signal," in German), doctoral dissertation, Univ. of Göttingen, FRG, 1987.

[51] E. Krüger, "Optimierung von akustisch—artikulatorischen Schtzungen mit Hilfe eines Systemmodells" ("Optimization of acousti-

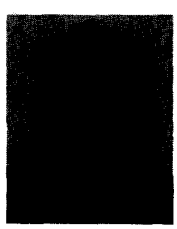
cal—articulatory estimations using a system model," in German), doctoral dissertation, Univ. of Göttingen, FRG, 1989.
[52] J. Schroeter and M. M. Sondhi, "Dynamic programming search of articulatory codebooks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1989, pp. 588–591.
[53] J. D. Markel and A. H. Gray, Jr., *Linear prediction of speech.* New York: Springer, 1976.
[54] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 1, pp. 67–72, 1975.
[55] H. T. Edwards, *Applied Phonetics, The Sounds of American English.* San Diego: Singular Publishing Group, 1992.
[56] J. W. Möller, B. S. Atal and M. R. Schroeder, "Determination of articulatory parameters of the human vocal tract from acoustic measurements," *J. Acoust. Soc. Am.*, vol. 60, S77(A), 1976.
[57] J. W. Möller, "Regressive Schötzung artikulatorischer Parameter aus dem Sprachsignal," ("Regressive estimation of articulatory parameters from the speech signal," in German), doctoral dissertation, Univ. of Göttingen, FRG, 1978.
[58] B. S. Atal, "Towards determining articulator positions from the speech signal," in *Proc. Preprints Speech Communication Seminar*, Stockholm, vol. 1, 1974, pp. 1–9.
[59] E. Krüger, G. Panagos, S. Jockusch, and H. W. Strube, "Artikulatorisch-akustische Abbildung auf grund gemessener artikulatorischer Parameter," ("Articulatory-acoustic mapping based on measured articulatory parameters," in German) in *Fortschritte der Akustik*, DAGA 88, Bad Honnef: DPG-GmbH, pp. 677–680.
[60] S. Kiritani, K. Itoh, and O. Fujimura, "Tongue pellet tracking by a computer-controlled X-ray microbeam system," *J. Acoust. Soc. Am.*, vol. 57, no. 6, pt. 2, pp. 1516–1520, 1975.
[61] D. R. Hush and B. G. Horne, "Progress in supervised neural networks—what's new since Lippmann," *IEEE Signal Processing Mag.*, vol. 10, no. 1, pp. 8–39, Jan. 1993.
[62] S. Parthasarathy and M. M. Sondhi, "Generalized radial basis functions for acoustic-to-articulatory mapping," unpublished.
[63] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communicat.*, vol. 28, pp. 84–95, 1980.
[64] J. L. Kuester and J. H. Mize, *Optimization techniques with Fortran.* New York: McGraw-Hill, 1973.
[65] F. Soong and B.-H. Juang, "Optimal quantization of LSP parameters," *IEEE Trans. Speech, Audio Processing*, vol. 1, no. 1, pp. 15–24, 1993.
[66] B. S. Atal and O. Rioul, "Neural networks for estimating articulatory positions from speech," *J. Acoust. Soc. Am.*, vol. 86, suppl. 1, S67, 1989.
[67] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE Signal Processing Mag.*, vol. 4, no. 2, pp. 4–22, Apr. 1987.
[68] A. Soquet, M. Saerens, and P. Jospa, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model: further results," in *Artificial Neural Networks* T. Kohonen, K. Mökisara, O. Simula, and J. Kangas, Eds. North Holland: Elsevier, 1991, pp. 371–376.
[69] K. Shirai and T. Kobayashi, "Estimation of articulatory motion using neural networks," *J. Phonetics*, vol. 19, pp. 379–385, 1991.
[70] S. Jokusch, G. Panagos, and H. W. Strube, "Anwendung neuronaler Netzwerke auf die Schätzung von Sprachparametern," ("Application of neural nets to the estimation of speech parameters," in German), in *Fortschritte der Akustik, DAGA'89*, Bad Honnef: DPG-GmbH, pp. 315–318, 1989.
[71] H. Wöhlbier and H. W. Strube, "Bestimmung der akustisch-artikulatorischen Abbildung an einem artikulatorischen Synthesemodell," ("Determination of the acoustic-articulatory mapping with an articulatory synthesis model," in German) in *Fortschritte der Akustik, DAGA'90*, Bad Honnef: DPG-GmbH, pp. 1079–1082, 1990.
[72] H. W. Strube and H. Wöhlbier, "Bestimmung artikulatorischer Steuerparameter zur Sprachsynthese mittels neuronaler Netzwerke," ("Determination of articulatory control parameters for speech synthesis by neural networks," in German) in: *Fortschritte der Akustik, DAGA'91*, Bad Honnef: DPG-GmbH, pp. 949–952, 1991.
[73] R. Hecht-Nielsen, "Counterpropagation Networks," in *Proc. IEEE 1st Int. Conf. Neural Networks*, vol. 2, 1987, pp. 19–32.
[74] H. Robbins and S. Monro, "A stochastic approximation method," *Annu. Math. Stat.*, vol. 22, pp. 400–407, 1951.
[75] H. Warneboldt and H. W. Strube, "Bestimmung der akustisch-artikulatorischen Abbildung für ein Vokaltraktmodell," ("Determination of the acoustic-articulatory mapping for a model of the vocal tract," in German), in *Fortschritte der Akustik, DAGA'93*, Bad Honnef: DPG-GmbH, 1993, in press.
[76] T. Martinetz and K. Schulten, "A 'Neural Gas' Network Learns Topologies," in *Artificial Neural Networks*, T. Kohonen, K. Mökisara, O. Simula, and J. Kangas, Eds. North Holland: Elsevier, 1991.
[77] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi, "On the use of neural networks in articulatory speech synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 2, pp. 1109–1121, 1993.
[78] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-35, no. 35, pp. 947–954, 1987.
[79] S. K. Gupta and J. Schroeter, "Pitch-synchronous frame-by-frame and segment-based articulatory analysis-by-synthesis," *J. Acoust. Soc. Am.*, in press.



Juergen Schroeter (M'79–SM'88) was born in Duisburg, Germany, on May 25, 1952. He received the Dipl.-Ing. (EE) and Dr.-Ing. (EE) degrees in 1976 and 1983, respectively, from Ruhr-Universitaet Bochum, Germany.

From 1976 to 1985, Dr. Schroeter 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. He is presently with AT&T Bell Laboratories, Murray Hill, NJ, where he is working on speech production models for speech synthesis, recognition, and coding.

Dr. Schroeter is a member of ASA.



Man Mohan Sondhi received the B.Sc. (Honours) degree in physics, 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, the M.S. degree in electrical engineering, in 1955, and the Ph.D. degree, in 1957, both from the University of Wisconsin, Madison.

Before joining AT&T Bell Laboratories, Murray Hill, NJ, he was with the Avionics division of John Oster Manufacturing Co., Racine, WI, the Central Electronics Research Institute, Pilani, India, and taught for one year at Toronto University, Toronto, Ontario, Canada. He was a Guest Scientist from 1971–1972 at the Royal Institute of Technology, Stockholm, Sweden, at CNET, Lannion, France (1982), and at NTT Human Interface Lab, Musashino, Japan (1990). He is a Supervisor and Distinguisher Member of Technical Staff in the Acoustics research department at AT&T Bell Labs. He holds nine patents, has authored or coauthored over 80 published articles, and co-edited *Advances in Speech Signal Processing*, and was Associate Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING for several years and has also been a Distinguished Lecturer of the ASSP society. His research interests have included speech signal processing; echo cancellation; adaptive filtering; modeling of auditory, speech, and visual processing by human beings; acoustical inverse problems; speech recognition; and analysis and synthesis of speech using articulatory models.