Waveform-Based Speech Recognition Using Hidden Filter Models: Parameter Selection and Sensitivity to Power Normalization

Hamid Sheikhzadeh and Li Deng, *Senior Member, IEEE*

***Abstract*—In this paper, we describe a novel approach to speech recognition by directly modeling the statistical characteristics of the speech waveforms. This approach allows us to remove the need for using speech preprocessors, which conventionally serve a role of converting speech waveforms into frame-based speech data subject to a subsequent modeling process. Central to our method is the representation of the speech waveforms as the output of a time-varying filter excited by a Gaussian source time-varying in its power. In order to formulate a speech recognition algorithm based on this representation, the time variation in the characteristics of the filter and of the excitation source is described in a compact and parametric form of the Markov chain. We analyze in detail the comparative roles played by the filter modeling and by the source modeling in speech recognition performance. Based on the result of the analysis, we propose and evaluate a normalization procedure intended to remove the sensitivity of speech recognition accuracy to often uncontrollable speech power variations. Effectiveness of the proposed speech-waveform modeling approach is demonstrated in a speaker-dependent, discrete-utterance speech recognition task involving 18 highly confusable stop consonant–vowel syllables. The high accuracy obtained shows promising potentials of the proposed time-domain waveform modeling technique for speech recognition.**

I. Introduction

In nearly all speech recognition systems developed to date, speech signals are first transformed away from the time domain, a process that is often referred to as feature extraction or preprocessing. Although preprocessing reduces the data rate, it also causes inevitable loss of possibly useful information for speech discrimination tasks. Speech waveforms in the time domain as raw speech data are generated directly by the speech production system and contain all the acoustic information necessary for recognition. Two principal components of the speech production system, the vocal tract and the excitation source, can be reasonably well parameterized by time-varying autoregressive filter models at the sample level [1]. In this modeling framework, which is known as linear predictive coding, the time-varying or nonstationary characteristics of the filters are handled in a nonparametric manner. That is, no constraints are imposed on the filters describing the speech signal at different time epochs.

For application of the above speech-sample modeling technique to speech recognition, however, one basic requirement is to represent the speech nonstationarity in a compact and parametric form. From the signal analysis viewpoint, the nonstationarity is associated with relatively slow variations in phonetic contents; from the speech production viewpoint, it is associated with the changes of the vocal tract configuration and of the excitation-source characteristics. This requirement is in contrast to the predictive coding scheme where the nonstationarity is treated on a nonparametric, frame-by-frame basis. One type of model that fulfills the above requirement is a stochastic speech production model with its parameters modulated by a Markov chain. The Markovian assumption, despite its simplicity, imposes a tight constraint on the possible variations in the vocal-tract filter parameters and allows a compact description of the time-varying speech signal before speech recognition algorithms can be formulated. In this study, a *hidden filter model* [2], [3], or the autoregressive hidden Markov model (AR–HMM), is used as a tool to explore the effectiveness of direct modeling of speech waveform for speech recognition without artificially breaking the recognizer into an independent preprocessing (which is usually based on spectral analysis nonparametric in handling the temporal dimension) and a subsequent modeling stage (which handles the temporal dimension parametrically).

The waveform-based speech recognition has a major advantage over the standard frame-based speech recognition in its avoidance of speech framing. Speech preprocessors in most existing recognizers involve blocking speech samples into fixed-length frames before the modeling stage. The framing carried out in preprocessing has no knowledge of and thus inevitably neglects the boundaries between any phonetic events. This often results in a relatively poor temporal resolution for fast varying speech sounds such as plosives. The method presented in this paper, on the other hand, is totally dispensable with the needs for framing and for the subsequent spectral analysis.

One basic issue arising from waveform-based speech recognition is the uncontrollable variations in the overall waveform power variance, even for the same word, due both to the variations in amplifier gain during recording and to the manner in which the words are uttered. A major contribution of this study is detailed analysis of the effects of power variations on

Manuscript received September 10, 1991; revised April 15, 1993. The associate editor coordinating the review of this paper and approving it for publication was D. Nahamoo. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1.

IEEE Log Number 9213393.

speech recognition accuracy and a simple but effective solution to remove such effects.

Poritz appears to be the first to propose the use of HMM's to represent the nonstationarity in an AR process [4]. In his model, a speech waveform of length $T \times M$ samples is decomposed into a sequence of T shorter segments, $\boldsymbol{Y}(t)$, each of length M : $\boldsymbol{Y}(t) = (y_{tM}, \cdots, y_{(t+1)M-1})$ for $t = 0, 1, \cdots, T-1$, and the consecutive vectors of time samples constitute an observation sequence. This AR–HMM was employed to solve the talker verification problem. In a more recent paper, Poritz discussed the *hidden filter* model [2], with no applications given. It is a special case of the model suggested in [4] with $M = 1$; that is, there is no grouping of speech samples in the model. Recently, Kenny *et al.* [3] used a version of the AR–HMM to describe sequences of cepstral vectors produced from a speech preprocessor, rather than modeling the aggregate speech waveform samples as a vector sequence as in [4]. In parallel with the above developments, Juang *et al.* [5] developed a more general type of AR–HMM, which associates a *mixture* of Gaussian autoregressive densities with each state in the HMM. Tishby [6] applied this mixture AR–HMM to a speaker recognition problem. In addition to the above various types of the nonstationary AR models targeted mainly for speech or speaker recognition applications, there has been a similar type of model, but with a different way to place constraints on the speech nonstationarity, developed for applications in speech coding [7].

In this paper, we use the simplest type of AR–HMM, the hidden filter model, for the applications of speech-waveform modeling and of ensuing speech recognition that have not been attempted in the past. The paper is organized as follows. In Section II, the theory of the hidden filter model is briefly reviewed. In Section III, we address the critical issue of power normalization of speech waveforms, which is inherently associated with any waveform-based speech recognition scheme. A detailed analysis is carried out on the effects of speech power variations on speech recognition accuracy. Based on the insight gained from the analysis, we propose and implement an effective solution to the power normalization problem. Finally, Section V presents experimental results on speech recognition for the evaluation of the proposed waveform-modeling technique.

II. The Hidden Filter Model

The formulation of the model is similar to [3]; the difference is that the observation vectors are single dimensional here. This is a standard AR model with its parameters associated with Markov states. The present value of a 1-D observation sequence O_t (digitized speech waveform in this study) is expressed as a linear combination of its past values plus a *driving sequence*, which is assumed to be a Gaussian i.i.d. process. Consider a Markov chain with N states and a state transition matrix $\underline{\underline{A}} = [a_{ij}], i, j = 1, 2, \cdots, N$. The AR model parameters are made conditional on the state of the Markov chain i. Thus, in the hidden filter model, the data-generation mechanism conditioned on state i is described by

$$O(t) = \underline{B}_i^* \underline{X}_t + e_i(t) \tag{1}$$

where $\underline{B}_i^* = \{B_i(l), l = 1, 2, \cdots, p\}$ is the vector of AR coefficients, $\underline{X}_t = \{O_{t-1}, O_{t-2}, \cdots, O_{t-p}\}^*$ is the sequence of the past p observations (* denotes transpose), and the driving sequence $e_i(t)$ (also called the *residual error*) is Gaussian and i.i.d. with a mean of μ_i and a variance of σ_i^2. In (1), if $p = 0$ (zeroth-order prediction), then the hidden filter model is reduced to the standard HMM [8]; that is, the present observation O_t is equal to $e_i(t)$ and the output distribution parameters (μ_i and σ_i) change according to the Markov state transition. On the other hand, when $N = 1$ the hidden filter model is reduced to the standard stationary AR model [1]. Since the driving sequence is assumed i.i.d., the likelihood of the observation sequence $O_1^T = \{O_1, O_2, \cdots, O_T\}$ given the state sequence $S_1^T = \{s_1, s_2, \cdots, s_T\}$ and p initial observations $\underline{X}_1$ under the model λ is calculated as

$$P(O_1^T | S_1^T, \underline{X}_1, \lambda) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_{s_t}} \exp[-\frac{1}{2\sigma_{s_t}^2}(O_t - \mu_{s_t} - \underline{B}_{s_t}^* \underline{X}_t)^2].$$

From this, the likelihood with which the model produces the observation sequence is obtained as

$$P(O_1^T | \underline{X}_1, \lambda) = \sum_{S_1^T} P(O_1^T | S_1^T, \underline{X}_1, \lambda) P(S_1^T). \tag{2}$$

We now discuss the Baum–Welch algorithm for parameter estimation of the hidden filter model. In the maximum likelihood estimation method that we pursue, it is required to maximize (2) with respect to $a_{ij}, B_i(k), \mu_i$, and σ_i^2 for each $i, i = 1, 2, \cdots, N$ and each $k, k = 1, 2, \cdots, p$. The estimation formulas, which is in an iterative form, can be derived by maximizing *Baum's auxiliary function* [9] for λ. Starting from an initial model λ_0, the objective function is given as

$$Q(\lambda, \lambda_0) = \sum_{S_1^T} P(S_1^T | O_1^T, \lambda_0) \log [P(O_1^T, S_1^T | \lambda)].$$

The objective function can be simplified to

$$Q(\lambda, \lambda_0) = \sum_{i,j=1}^{N} \sum_{t=1}^{T} \gamma_{ij}(t)[\ln a_{ij} + \ln D_i(O_t, \underline{X}_t)] \tag{3}$$

where

$$\ln D_i(O_t, \underline{X}_t) = \ln \frac{1}{\sqrt{2\pi}\sigma_i} - \frac{1}{2\sigma_i^2}(O_t - \mu_i - \underline{B}_i^* \underline{X}_t)^2$$

and $\gamma_{ij}(t)$ is the *a posteriori* probability of the transition from state i to state j given the observation sequence and the model λ_0: $\gamma_{ij}(t) = P(s_{t-1} = i, s_t = j | O_1^T, \lambda_0)$. The forward–backward algorithm [9] can be used to calculate $\gamma_{ij}(t)$ efficiently.

Equation (3) is optimized by differentiation with respect to each of $a_{ij}, \underline{B}_i, \mu_i$, and $\sigma_i^2 (i, j, = 1, 2, \cdots, N)$, respectively. As in standard HMM, a_{ij} is easily optimized

$$\hat{a}_{ij} = \sum_{t=1}^{T} \gamma_{ij}(t) \Big/ \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_{ij}(t). \tag{4}$$

The optimization of $\underline{B}_i$ and μ_i leads to the equations

$$\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)[O_t-\mu_i-\underline{B}_i^*\underline{X}_t]\underline{X}_t^*=\underline{0} \tag{5}$$

$$\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)[O_t-\mu_i-\underline{B}_i^*\underline{X}_t]=0. \tag{6}$$

If $\hat{\mu}_i$ and $\hat{\underline{B}}_i$ are the optimal values obtained by jointly solving (5) and (6), then

$$\hat{\sigma}_i^2=\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)[O_t-\hat{\mu}_i-\hat{\underline{B}}_i^*\underline{X}_t]^2\Big/\sum_{t=1}^{T}\gamma_{ij}(t). \tag{7}$$

In (5), it is noted that $(\underline{B}_i^*\underline{X}_t\underline{X}_t^*)^*=\underline{X}_t\underline{X}_t^*\underline{B}_i$ so that writing (5) in a transposed form, we have

$$\underbrace{\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)O_t\underline{X}_t}_{\triangleq\,\underline{S}_{OX}}=\underbrace{\left(\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)\underline{X}_t\right)}_{\triangleq\,\underline{S}_X}\mu_i+\underbrace{\left(\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)\underline{X}_t\underline{X}_t^*\right)}_{\triangleq\,\underline{\underline{S}}_{XX}}\underline{B}_i$$

Similarly, (6) gives

$$\underbrace{\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)O_t}_{\triangleq\,S_O}=\underbrace{\left(\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)\right)}_{\triangleq\,G_i}\mu_i+\underbrace{\left(\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)\underline{X}_t^*\right)}_{\triangleq\,\underline{S}_X}\underline{B}_i.$$

The above two equations constitute the system of equations

$$\begin{aligned}\underline{S}_{OX}&=\underline{S}_X\mu_i+\underline{\underline{S}}_{XX}\underline{B}_i\\ S_O&=G_i\mu_i+\underline{S}_X\underline{B}_i.\end{aligned} \tag{8}$$

Note that $\underline{S}_{OX},\underline{S}_X,\underline{\underline{S}}_{XX},S_O$, and G_i are independent of time. The system of equations (8) gives $(p+1)$ equations in $(P+1)$ unknowns for each state $1\le i\le N$, and hence may be solved for $\underline{B}_i$ and μ_i for each state i.

Equation (7) for the estimation of σ_i^2 can be simplified into

$$\hat{\sigma}_i^2=\frac{1}{G_i}\left[\sum_{j=1}^{N}\sum_{t=1}^{T}\gamma_{ij}(t)O_t^2+\hat{\underline{B}}_i^*\underline{\underline{S}}_{XX}\hat{\underline{B}}_i-2\underline{S}_{OX}^*\hat{\underline{B}}_i\right]-\hat{\mu}_i^2. \tag{9}$$

Throughout the above analysis, we have assumed that the model is trained using a single token. The multiple token training case analysis is similarly derived. Basically, in (3)–(9), this involves replacing each "t (time) summation" by a "k (token) summation and t (time) summation": $\Sigma_{t=1}^{T}\Rightarrow\Sigma_{k=1}^{K}\Sigma_{t=0}^{Tk}$. The subsequent analysis is exactly the same as discussed previously.

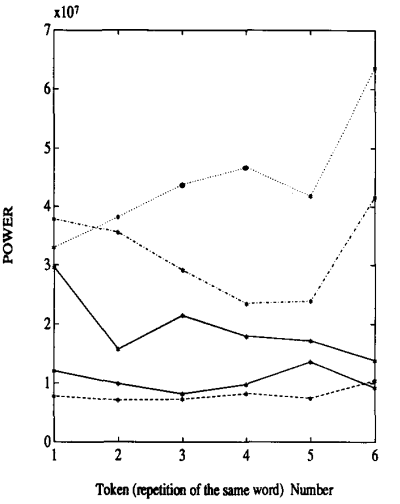


Fig. 1. Variation of average power for the same word uttered at different repetitions. The five curves are for five distinct words.

III. Effect of Power Variations on Recognition Accuracy

One basic problem encountered when applying the hidden filter model to speech recognition was uncontrollable variations of the signal power (closely related to the loudness) in spoken words having identical phonetic contents. Although the speakers would intend to maintain the same level of loudness, the actual level for different tokens of the same word often have loudness variations to a large degree. Fig. 1 shows such variations for five different one-syllable words, each having six tokens uttered during the same recording session. Although the average power of the utterance can be characteristic of the underlying word, different tokens of the same word often exhibit significant power variations. For the word illustrated by dotted lines, the variation is as high as 100% (comparing the first and the last tokens). In addition to this intrinsic cause, the power variation problem can be aggravated by environmental factors such as changes in the amplifier gain during recording. In this section, we furnish a detailed analysis on the effect of power variations on pattern discrimination in order to understand to what degree the variations can confound waveform-based speech recognition.

Separability of two sounds modeled by two different hidden filters is due to two different but related mechanisms: the sets of filter coefficients, $\underline{B}_i$ and the variances $\sigma_i^2, i=1,2,\cdots,N$. The former models the vocal tract transfer function, and the latter is related to the source excitation characteristics. Mathematically, these two sets of parameters determine the overall covariance matrix of the joint distribution of data sequence in a distinct manner: The AR coefficients determine the structure of the covariance matrix, while the variance is just a scaling factor of it (see appendix). The specific goal of the analysis described in this section is to determine the relative contribution of each of the above mechanisms to speech recognition performance.

A. Bhattacharyya Bound and Gaussian Approximation for the Classification Error

We employ the Bhattacharyya bound and the Gaussian approximation methods [10] to determine which parameter sets, the AR filter coefficients or the variances of the Gaussian driving sequences, are the dominant factor for the recognition error. Both methods give the class separability in terms of covariance matrices for two classes of hidden filters. Two different Gaussian AR models form a classifier, which can be used to determine to which class of the AR populations any unknown observation belongs. The Gaussian AR models are subsets of the more general classes of two multivariate Gaussian distributions. Throughout this section, we assume single-state hidden filters for simplicity purposes. For these special classes, the Bayes' recognition error can be calculated using numerical methods [10], [11]. To describe the methods, we first write the generic form of the *likelihood function* defined as

$$h(X) = +\ln p_1(X) = \ln p_2(X)$$

where $p_i(X)$ is the density function of the observation vector X, conditioned on class ω_i, and is assumed to be well trained. During recognition, the likelihood function is compared to a threshold, usually zero for two equiprobable classes. When $p_i(X)$'s are normal with expected vector M_i and covariance matrices Σ_i, the likelihood function can be written as

$$\begin{aligned} h(X) = {} & \tfrac{1}{2}(X - M_1)^T \Sigma_1^{-1} (X - M_1) \\ & -\tfrac{1}{2}(X - M_2)^T \Sigma_2^{-1} (X - M_2) + \tfrac{1}{2} \ln |\Sigma_1|/|\Sigma_2|. \end{aligned} \quad (10)$$

There are two type of errors: one (ϵ_1) results from misclassifying samples from class ω_1, and the other (ϵ_2) results from misclassifying samples from class ω_2. They can be calculated as

$$\epsilon_1 = \int_{ln P_1/P_2}^{+\infty} p_h(h|\omega_1) dh \quad (11)$$

and

$$\epsilon_2 = \int_{-\infty}^{ln P_1/P_2} p_h(h|\omega_2) dh \quad (12)$$

where P_i is the *a priori* probability for class ω_i and $p_h(h|\omega_i)$ is the conditional density of $h(X)$ for class ω_i. In later analysis, we assume $P_1 = P_2$ for simplicity.

Basically, there are three types of approaches to finding the recognition error: the exact methods, the approximation methods, and methods that bound the recognition error. We consider the Bhattacharyya bound and Gaussian approximation solutions that offer fairly simple implementations.

To simplify the analysis without loss of generality, we assume zero means in the two distributions characterizing the two classes. The Bhattacharyya bound [12] on the total recognition error for two classes of Gaussian distributions is given as

$$\epsilon_\mu = \sqrt{P_1 P_2} e^{-\mu} \quad (13)$$

where

$$\mu = \frac{1}{2} \ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{\sqrt{|\Sigma_1| \cdot |\Sigma_2|}}$$

and $|\Sigma_i|$ is the determinant of the covariance matrix of the class ω_i. It is well known that the Bhattacharyya bound is an important measure of the separability of two general distributions [10], [12].

The second method we use is based on Gaussian approximation of the likelihood function. Since the quadratic equation of (10) represents the summation of many terms, the central limit theorem suggests that the distribution of $h(X)$ is close to a Gaussian. This is particularly true when the eigenvalues λ_i of the equation $|\Sigma_2 - \lambda\Sigma_1| = 0$ are close to one [11]. Due to the near Gaussian distribution of $h(X)$, it is sufficient to characterize the distribution by the two moments [10]:

$$\begin{aligned} E[h(X)|\omega_1] &= -\tfrac{1}{2}\text{trace}[\Sigma_2^{-1}\Sigma_1 - I] + \tfrac{1}{2}\ln|\Sigma_1|/|\Sigma_2| \\ E[h(X)|\omega_2] &= \tfrac{1}{2}\text{trace}[\Sigma_1^{-1}\Sigma_2 - I] + \tfrac{1}{2}\ln|\Sigma_1|/|\Sigma_2| \\ \text{Var}[h(X)|\omega_1] &= \tfrac{1}{2}\text{trace}[\{\Sigma_2^{-1}\Sigma_1 - I\}^2] \\ \text{Var}[h(X)|\omega_2] &= \tfrac{1}{2}\text{trace}[\{\Sigma_1^{-1}\Sigma_2 - I\}^2]. \end{aligned} \quad (14)$$

After this, the approximation of the total recognition error ϵ can be found as [10]

$$\begin{aligned} \epsilon_i &= \text{erfc}\left[\frac{E[h(X)|\omega_i]}{\sqrt{\text{Var}[h(X)|\omega_i]}}\right], \qquad i = 1, 2 \\ \epsilon &= \frac{\epsilon_1 + \epsilon_2}{2} \end{aligned} \quad (15)$$

where $\text{erfc}(x)$ denotes the error function.

Both of the above methods approximate the recognition error of two multivariate Gaussian classes in terms of their covariance matrices. In the appendix, we describe a numerical method by which, given the AR parameter set, the covariance matrix of an AR (p) process can be obtained. The AR parameter set is obtained via the estimation procedure discussed in Section II.

B. Power Variation and Simulated Recognition Error

In this section, the Bhattacharyya bound and Gaussian approximation methods are used to estimate the recognition errors for two distinct hidden filter classes that are arbitrarily chosen and for those trained with speech data. To simplify the analysis, only the single-state hidden filters are considered. Given the AR coefficients and the variance of the driving sequence, the covariance matrix of the multivariate Gaussian distribution can be calculated numerically according to the procedure outlined in the appendix. The Bhattacharyya bound (13) and Gaussian approximation (15) methods can then be employed to estimate the total recognition error.

To simulate the effect of speech power variations on recognition accuracy, a carefully designed experiment is needed. The design we chose is based on the following assumption: the effect of speech power variations is equivalent to multiplying the covariance matrix of the modeled speech data for all different classes by a variable factor. (This assumption is valid provided the statistics of the speech waveform obey the hidden

filter model.) To make the design philosophy clear, we first prove that variances in the Gaussian distribution of the driving sequences σ_i^2's are the only set of parameters related to the power of the speech waveform. In other words, the power variation in the speech waveform does not affect estimates of the AR filter coefficients. The proof begins by multiplying the observation sequence (*training sequence*) O_1^T by an arbitrary constant L. Then, according to reestimation formulas (7) and (8), the only change in model parameters due to the variable observation scaling would be $\sigma'^2 = L^2 \cdot \sigma^2$, and the AR($p$) coefficients remain the same independently of the value of L.

The experiment that simulates the effect of power variations on recognition accuracy in speech recognition experiments is designed for classification of two AR(1) classes as follows. Let σ_1^2 and σ_2^2 be the variances of driving sequences for the two distinct AR processes representing the two classes of speech sounds. Let the AR parameter pairs for the two classes (B_1 and B_2) be chosen from combinations of values in the range $(-0.9, 0.9)$. Setting $\sigma_1^2 = 1$ and $\sigma_2^2 = k \cdot \sigma_1^2, k = 1, 1.2, \cdots, 3.6, 3.8$ with a fixed increment of 0.2, we find the total error $e_1(k)$ for each k according to (13). Since in actual speech recognition experiments speech power variations have equal effects on both speech classes, we continue the simulation by exchanging the role of σ_1^2 and σ_2^2. That is, set $\sigma_2^2 = 1$ and $\sigma_1^2 = k \cdot \sigma_2^2$, and find the total error $e_2(k)$, using (13) again. Assuming that the two classes have equal *a priori* probabilities, we then obtain the average recognition error $e_{\text{avg}}(k)$ by averaging $e_1(k)$ and $e_2(k)$.

A dimensionality of 20 is chosen for the multivariate Gaussian distribution. The simulated average recognition errors calculated according to the above paradigm are provided in Fig. 2 for seven different combinations of AR coefficient pairs B_1 and B_2 (marked by *a–g*). These results show the average recognition error $e_{\text{avg}}(k)$ as a function of the variance ratio of the two classes k. Two clear trends can be seen consistently from this figure. First, as the distance between the AR coefficients of the two classes ($|B_1 - B_2|$) increases, the maximum recognition error for each AR pair decreases (from curve *a* to curve *g*). Second, the sensitivity of the recognition error on the variance ratio k, measured as the average change in the recognition error resulting from a unit variance-ratio increase, decreases as evidenced by decreasing slopes in the curves from *a* to *g*. From these observations, it is inferred that when the AR coefficients of two classes are close to each other (curves *a, b, c* in Fig. 2), the dominant factor for recognition error is the variance parameters in the hidden filter models; otherwise (curves *d–g* in Fig. 2), the AR coefficient parameters in the hidden filter models are the principal determinants of the recognition error.

The same simulation experiment is carried out using the Gaussian approximation method. The average recognition errors calculated (using (15)) are shown in Fig. 3. The results are consistent with the previous conclusion drawn from the use of the Bhattacharyya method.

The next sets of simulation experiments are conducted using the hidden filter models with autoregression order $p = 8$. The AR coefficients in the four-state hidden filter models were determined from speech waveform samples according to

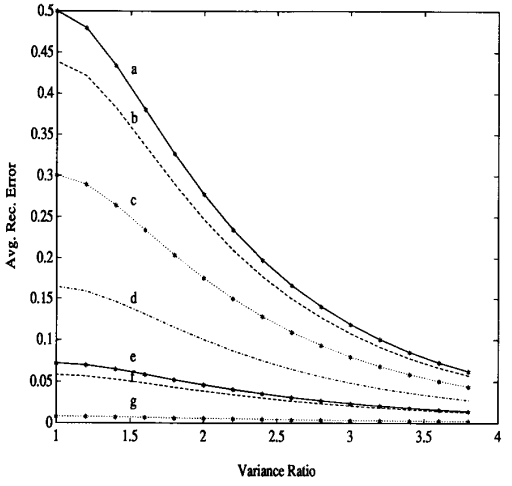


Fig. 2. Average recognition error versus variance ratio (k) using Bhattacharyya bound for seven different AR(1) pairs: (a) 0.5, 0.5, (b) 0.4, 0.6, (c) 0.3, 0.7, (d) 0.2, 0.8, (e) 0.1, 0.9, (f) −0.5, 0.5, (g) −0.5, 0.9.

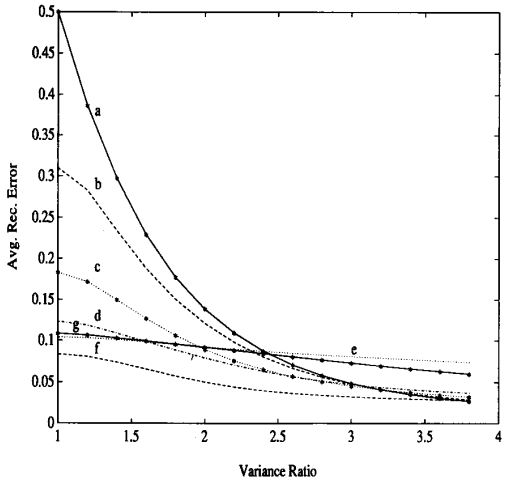


Fig. 3. Average recognition error versus variance ratio (k) using Gaussian approximation for seven different AR(1) pairs: (a) 0.5, 0.5, (b) 0.4, 0.6, (c) 0.3, 0.7, (d) 0.2, 0.8, (e) 0.1, 0.9, (f) −0.5, 0.5, (g) −0.5, 0.9.

the Baum–Welch reestimation formulas described in Section II. Shown in Fig. 4 are simulated recognition error curves, as a function of the variance ratio, for hidden filter models representing syllables /pu/ and /ku/, respectively. The four curves *a, b, c,* and *d* correspond to the four sequential states in the hidden filter models, respectively. They were obtained by using the Bhattacharyya solution through the same procedure as described earlier in this section for the AR(1) case. Shown in curves *a* and *b* are the results from the two starting states, which represent the burst/frication and aspiration segments in syllables /pu/ and /ku/. Their associated AR parameters should then correspond to two widely separated classes (i.e., very different covariance matrices in structure). Curves *c* and *d*, on the other hand, show the simulated errors for the two final states, which represent the vowel portions of the two

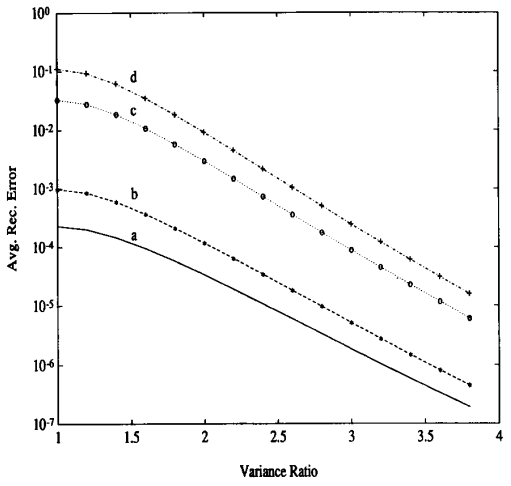


Fig. 4. Average recognition error as a function of variance ratio (k) using AR(8) parameters of the trained hidden filters for syllables /pu/ and /ku/. The pair of the AR(8) parameters is chosen from (a) first states, (b) second states, (c) third states, (d) fourth states of the hidden filters representing syllables /pu/ and /ku/, respectively. Bhattacharyya bound is used.

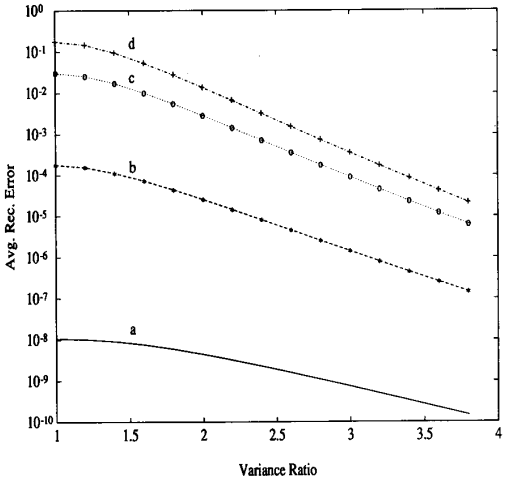


Fig. 5. Average recognition error as a function of variance ratio (k) using AR(8) parameters of the trained hidden filters for syllables /gu/ and /du/. The pair of the AR(8) parameters is chosen from (a) first states, (b) second states, (c) third states, (d) fourth states of the hidden filters representing syllables /gu/ and /du/, respectively. Bhattacharyya bound is used.

syllables (allophones /u/ in syllables /pu/ and /ku/), and hence are associated with two AR classes with more similar AR coefficients. Errors shown in c and d are very large due to the small separability of the AR coefficient sets (note that we are here dealing with separation of the same phoneme /u/ under two different contexts). They are also seen to be highly sensitive to the variance ratio k, with the maximum error of 0.11 at $k = 1$ and minimum error of $1.57e - 5$ at $k = 3.8$, in curve d for instance. In contrast, the errors shown in a and b are much smaller in magnitudes since we are dealing with separation of two different phonemes. (To be able to show them on the same plot as c and d, we use a log scale in the ordinate.) For all of the four states (curves a–d), the sensitivity of the error rate to the changing variance ratio is in close conformity with that predicted from Figs. 2 or 3. First, the error rate drops monotonically as the variance ratio increases. Second, and more importantly, the error rate, not in a log magnitude but in linear magnitude (as for Figs. 2 and 3) is much more sensitive to the variance ratio for the "vowel" region shown in curves c and d than for the "constant" region shown in curves a and b—note the absolute variation of the error rate is more than 10% (maximum error of 0.11 at $k = 1$ and minimum error of $1.57e-5$ at $k = 3.8$) for curve d, while the counterpart for curve b is only 0.1%.

Simulated recognition error curves for the four states of the hidden filter models representing syllables /gu/ and /du/ are shown in Fig. 5. No aspiration segments are present for the voiced stops, and the second states (one for each syllable) represent the transitional regions of the vocalic portions in the syllables. The error rates associated with the second states as depicted in curve b sit between the "consonant" region shown in curve a and the "vowel" region shown in curves c and d, in contrast to Fig. 4 where curve b was grouped into curve a. This can be accounted for by the earlier observations made from Figs. 2 or 3 since the AR coefficient sets representing two allophones of the same vowel should be closer to each other than the AR coefficient sets representing aspiration segments of two different stops.

As a final example, the simulated error curves for syllables /ba/ and /ga/ are plotted in Fig. 6. The behaviors of the errors are similar to those in Fig. 5, except that the errors in curve b, associated with the second states of the hidden filter models, are nearly three orders of magnitudes higher than their counterparts in Fig. 5 and are grouped totally into the "vowel" region shown in curves c and d. Interestingly, the speech recognition experiments, to be reported in Section V, demonstrated a substantially greater number of errors related to the /ba/–/ga/ confusion than any other syllable pairs.

All of the above simulation experiments were also performed using the Gaussian approximation method, and the results have been very consistent with those using the Bhattacharyya solution.

IV. A Prodedure for Power Normalization

The simulation experiments described in the last section suggest that when the AR coefficient sets for different speech classes are not sufficiently separated, power variations in the data will have a strong influence on the recognition accuracy. The existence of phonetic confusions in speech recognition is indicative of significant overlaps in the acoustic space between models representing different classes of speech sounds. Accordingly, a method for speech power normalization is necessary in order to remove the uncontrollable effects of power variations. The speech recognition experiments, to be reported in Section V, show that such effects tend to reduce the recognition accuracy. In addition, since different tokens from the same speech class are implicitly assumed to obey the statistical distribution described by the same model, they should maintain a relatively constant power level.

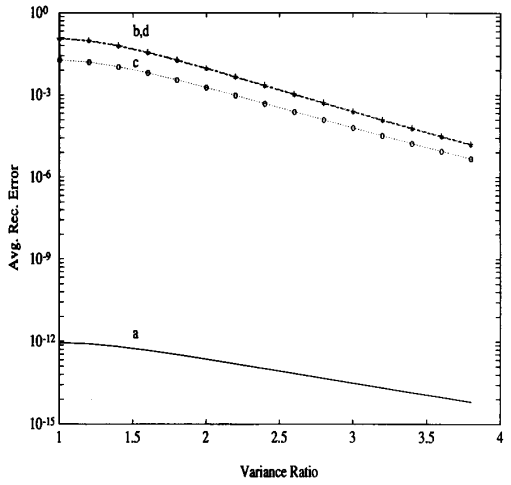


Fig. 6. Average recognition error as a function of variance ratio (k) using AR(8) parameters of the trained hidden filters for syllables /ba/ and /ga/. The pair of the AR(8) parameters is chosen from (a) first states, (b) second states, (c) third states, (d) fourth states of the hidden filters representing syllables /ba/ and /ga/, respectively. Bhattacharyya bound is used.

According to the above discussion, we propose a heuristic method for speech-waveform power normalization, and have implemented it in our hidden filter model-based speech recognizer. During the training phase for the hidden filter model, amplitudes of speech waveform corresponding to each word utterance are scaled such that the average power over the whole utterance for all tokens has a fixed, predetermined value. During recognition, the speech waveform for each test token is scaled so that its average power becomes identical to this predetermined value. This procedure normalizes the power of speech waveforms over the whole discrete utterance rather than over different segments within it. This is desirable because, in this way, the time course of power changes within the utterance is preserved.

The above method for speech waveform power normalization has been successfully applied to speech recognition experiments that we report below.

V. Speech Recognition Experiments

In this study, a left-to-right filter model was used for discrete-utterance speaker-dependent speech recognition. The database consisted of a total of 18 CV syllables where C stands for six English stop consonants /p/, /t/, /k/, /b/, /d/, /g/ and V stands for three cardinal vowels /i/, /a/, /u/. This task is relatively simple, allowing systematic experimentation on the model. All the syllables were uttered with a short pause in between by native English speakers in a normal office environment. Also, stop discrimination is known to be a difficult task, and is of special significance for general speech recognition problems [13].

Speech data were collected from four male speakers, each uttering 16 repetitions of each syllable. The Hyper-Signal software package running on an IBM-PS2 was used to collect time-domain data sampled at 16 kHz. Automatic end-point detection was carried out using the energy function calculated directly from speech waveforms based on methods described in [14].

TABLE I
Recognition Rates (%) and Average Ranks for Speaker 1 (Using the Power Normalization Procedure). N: Number of States in Hidden Filter Models, NTOK: Total Number of Training Tokens

Test No.	N	AR Order	NTOK	Rec. Rate	Avg. Rank
1	4	8	4	89.6	1.19
2	4	12	4	88.9	1.15
3	4	16	4	91.7	1.13
4	4	16	8	92.4	1.08
5	5	8	4	89.6	1.14
6	5	16	8	93.1	1.07
7	5	20	8	95.2	1.05
8	5	25	8	96.6	1.03

TABLE II
Recognition Rates (%) and Average Ranks for Speaker 2 (Using the Power Normalization Procedure)

Test No.	N	AR Order	NTOK	Rec. Rate	Avg. Rank
1	4	8	4	84.0	1.26
2	4	12	4	80.6	1.32
3	4	16	4	82.0	1.29
4	4	16	8	84.8	1.21
5	5	8	4	84.0	1.22
6	5	16	8	88.2	1.16
7	5	20	8	86.1	1.18
8	5	25	8	87.5	1.16

The range of parameters in the hidden filter models implemented is as follows. The number of states N in the models is four or five. The AR order varies from 8 to 25. Four or eight tokens (repetitions of the same word) were used for training, and eight disjoint tokens for each of the 18 CV syllables were used for testing.

The training and recognition algorithms were implemented according to the theory outlined in Section II. We have also utilized the segmental K-means algorithm [5] instead of the Baum–Welch algorithm to estimate the model parameters in order to save computation and memory storage. The power normalization procedure described in Section IV was used for the speech recognition experiments.

The performance of the hidden filter models for the above recognition task is measured by the percentage of test tokens correctly identified by the recognizer as the top one choice and by the average rank of the correct choice. Tables I and II contain these measures for two speakers as a function of the number of hidden filter states (N), the order of the AR filters, and the number of training tokens for each model. We have experimented with models with a large range of N values and AR orders. The optimal number of states is $N = 5$ for both speakers, and the optimal AR order is 25 for speaker 1 (Test No. 8) and for 16 for speaker 2 (Test No. 6). The optimum parameter set is quite consistent among all four speakers tested (results for two speakers only are shown).

The remaining recognition experiments were conducted for all four speakers using $N = 5$ and AR order of 25. These

TABLE III
RECOGNITION RATES (%) AND AVERAGE RANKS FOR FOUR SPEAKERS WITH AN OPTIMAL SET OF MODEL PARAMETERS (USING THE POWER NORMALIZATION PROCEDURE)

Speaker	Rec. Rate	Avg. Rank
1	96.6	1.03
2	88.2	1.16
3	85.5	1.23
4	86.8	1.15
Average	89.3	1.14

TABLE IV
COMPARISON BETWEEN RECOGNITION RATES (%) WITH AND WITHOUT THE USE OF SPEECH POWER NORMALIZATION PROCEDURE DESCRIBED IN SECTION IV

Speaker	Test No. (from Tables I and II)	Rec. Rate (Normalized Power)	Rec. Rate (Non-normalized Power)
1	4	92.4	88.2
1	6	93.1	91.3
1	8	96.6	93.8
2	4	84.8	82.7
2	6	88.2	82.0
2	8	87.5	84.0
1,2	Average	90.4	87.0

results are shown in Table III. The average recognition rate over the four speakers tested so far is 89.3% with an average rank of 1.14.

To examine the effect of speech waveform power normalization on recognition accuracy, we carried out a set of controlled experiments where the only experimental variable is the use of the normalization procedure or without its use. Table IV shows a comparison between the recognition accuracies affected by this variable only. We observed a 28.3% reduction of recognition errors, averaged over six experiments, after the power normalization procedure was implemented: the average recognition rate was improved from 87.0 to 90.4%.

In a further experiment for examining the sensitivity of recognition accuracy to power normalization, we artificially scaled speech waveforms of all the test tokens such that their overall powers varied systematically from one to as high as 30 times the predetermined power value set in the training phase. Speech recognition accuracies obtained by using various power ratios between test and training utterances are listed for a total of 144 test tokens in column two of Table V. Along with the recognition accuracies for the whole utterances, we also show in the remaining columns the constituent consonant and vowel recognition accuracies after analyzing the errors. It should be noted that the vowels used in our experiments are the three cardinal vowels /a/, /i/, and /u/, which are quite distinct in their formant frequencies and spectrums, and so their associated AR models are quite distinct too. Thus, as expected from the results of Section III-B, the principle determinants of the recognition error are the AR coefficients. As seen in Table V, the vowel recognition rate is not much affected by the power variations, which is consistent with the above conclusion. However, Table V shows that the recognition rate of the consonants is adversely affected by the

TABLE V
EFFECT OF POWER VARIATION ON SPEECH RECOGNITION ACCURACY (%). THE MODEL PARAMETERS ARE: NUMBER OF STATES = 5, AR MODEL ORDER = 25, NUMBER OF TRAINING TOKENS = 8 (SPEAKER 1)

Variance Ratio	Word Rec. Rate	Constant Rec. Rate	Vowel Rec. Rate
1.0	96.6	96.6	100.0
1.4	92.4	92.4	100.0
1.8	88.2	88.2	100.0
2.0	86.8	86.8	100.0
2.5	75.0	75.0	100.0
3.0	69.5	69.5	100.0
4.0	63.9	63.9	100.0
5.0	63.2	63.2	100.0
9.0	59.7	59.7	98.6
12	55.6	55.6	98.6
30	40.0	40.0	84.7

power variations. This is consistent with the fact that for the stop consonants, there is more overlap in the acoustic space, and thus the AR coefficients for some of them (e.g., /b/ and /g/) are close to each other. As a result, the power variations affect the recognition error much more.

The above experiments on speech power normalization suggest that power variations in speech waveforms have deteriorating effects on the performance of hidden filter model-based speech recognizers, and thus should be compensated for by power normalization schemes. This is in agreement with the conclusion we reached earlier in Section III-B after a series of simulation experiments.

VI. CONCLUSION

The principal motivation of this study is the desire to integrate the speech-preprocessor and the speech-modeling components of speech recognizers in order to achieve global optimization in the system design and be free from the time–frequency resolution dilemma. This is in contrast to nearly all the recognizers developed in the past that invariably break up the recognizer into the two serially connected, independent stages. Underlying the design philosophy of these conventional recognizers is the assumption that the output from the preprocessor preserves all the relevant information contained in the original speech waveform. One clear evidence against this assumption, however, is the smearing of acoustic information associated with fast varying speech sounds due to speech framing. Some speech events, such as stop bursts, frications, flaps, and formant transitions in a labial context, can have their time course as short as only a few milliseconds. This is smaller than the length of an average frame typically used in a speech preprocessor. Yet these short speech events often contain important cues for discriminating one speech sound from another [15].

In this study, we take an initial step towards direct modeling of speech waveforms for speech recognition without the use of speech preprocessors using fixed-length sample framing followed by subsequent spectral analysis. The method we used in this work is on the basis of a highly simplified assumption: speech waveforms are outputs of a time-varying

linear filter, which models the vocal tract resonator of the speech production system, excited by a Gaussian i.i.d. source with time-varying parameters in the distribution. The filter is assumed to be of an all-pole (i.e., autoregressive or IIR) type of a finite order. To formulate the speech recognition algorithm, we handle the time variation (nonstationarity) of the characteristics of the filter and the excitation source in a parametric form. Specifically, we assume that the filter coefficients and the variance parameters in the excitation-source distributions follow a Markov chain. The above assumptions give rise to the hidden filter model that we used to represent the statistics of speech waveforms and whose effectiveness in speech recognition is evaluated in this paper.

One major focus of this work is the comparative role played by the filter modeling and the source modeling in the speech recognition performance. In our current hidden filter modeling framework, the filter characteristics are described by a set of AR coefficients associated with Markov states, and the source characteristics by a set of variances (the mean parameters have been found to be near zero for all states). We study the relative significance of these separate sets of model parameters by conducting a series of controlled simulation experiments where the recognition errors were estimated (using the Bhattacharyya bound and Gaussian approximation method) as a function of the distance between the AR coefficient sets and that between the variance parameter sets associated with two AR classes. We concluded from these experiments that when the AR coefficients of the two AR classes are relatively close to each other (as in the task of discriminating two allophones of the same vowel), the recognition error rate is strongly influenced by the variance ratio of the two classes; AR coefficients, on the other hand, are the dominant factors of the recognition error if the AR classes are widely separated (as in the task of discriminating two stop consonants).

The implication of this conclusion is the need to devise appropriate power normalization schemes for speech waveforms in order to combat against the effects of speech power variations on speech recognition accuracy. Such variations are due to a number of uncontrollable factors. The normalization is necesary because overlaps in the acoustic space between the models describing different classes of speech sounds imply a dominant role of variance parameters in the models in determining theoretical speech recognition rates. The power normalization procedure we developed in this study (Section IV) was tested in a speech recognition task. Use of the normalization improved speech recognition accuracy ranged from 87.0 to 90.4%.

We conducted evaluation experiments aimed at testing effectiveness of the hidden filter modeling of speech waveforms in a speaker-dependent discrete-utterance speech recognition task involving 18 highly confusable stop consonant–vowel syllables. The highest recognition rate, obtained by setting the number of states in the models at five and setting the AR orders at 25, is 89.3% averaged over four speakers, each with 144 test syllable tokens. This is slightly inferior to but comparable with the recognition rate achievable with the standard HMM technique using mel frequency cepstral coefficients as the preprocessed speech representation [16].

The evaluation experiments conducted so far point to promising potentials in stochastic modeling of speech waveforms for speech recognition. One main contribution of this study is that it provides a clear understanding concerning the nature of the speech recognition sensitivity to the speech power variation as an uncontrollable variable in speech waveform generation. It is particularly encouraging that the power normalization procedure based on such an understanding leads to expected improvement is speech recognition performance. Our future work is directed towards overcoming two known deficiencies in the current formulation of the hidden filter model. First, state-conditioned i.i.d. Gaussian processes do not appear to be adequate for describing the statistical characteristics of the excitation sources in speech production. Second, while time-varying AR filters appear to be reasonable approximations to the vocal tract resonator configured to produce vowel-like utterances, they become poor ones for most consonants. In addition to providing better modeling schemes for both the source and the filter, we will also research into use of more general stochastic models developed in [17], [18] than the hidden filter model for speech waveform modeling. Once more faithful representations of speech waveforms are established, better speech recognition systems can be designed.

APPENDIX
NUMERICAL COMPUTATION OF THE COVARIANCE MATRIX OF AN AR PROCESS

The sequences generated by a (zero-mean) Gaussian autoregressive model will have a multivariate Gaussian distribution with covariance matrix Σ. For the special case of AR(1), the covariance matrix is known to have a closed form:

$$\Sigma = \frac{\sigma^2}{1-a^2}\begin{bmatrix} 1 & a & \cdots & a^{n-1} \\ a & 1 & & \vdots \\ \vdots & & \ddots & a \\ a^{n-1} & \cdots & a & 1 \end{bmatrix} \tag{16}$$

where σ^2 is the variance of the Gaussian distribution and a is the AR(1) coefficient. For higher order AR models, the covariance matrix does not have simple closed form, but can be calculated numerically. Consider the AR(p) process of

$$y_t = \sum_{k=1}^{p} a_k \cdot y_{t-1} + e_t \tag{17}$$

where $a_k, k = 1, 2, \cdots, p$ are the AR coefficients and e_t is a zero-mean Gaussian i.i.d. process with variance σ^2. The covariance matrix for the AR(p) process can be found via a recursive method proposed in [1]. It is a recursive implementation of the Yule–Walker equation pairs

$$r_i = \sum_{k=1}^{p} a_k \cdot r_{|i-k|}, \qquad i \geq 1 \tag{18}$$

$$r_0 = \sum_{k=1}^{p} a_k \cdot r_k + \sigma^2 \tag{19}$$

where r_i is the ith autocorrelation sample and a_k is the kth AR coefficient of an AR (p) process. Let $\boldsymbol{a}_m^{(n)}$ be an m-dimensional

vector whose kth component is $a_k^{(n)}$ and $\boldsymbol{r}_n$ an n-dimensional vector whose kth component is r_{n-k+1}. Let us define, for every vector, a reciprocal vector defined by the relationship

$$\boldsymbol{a}_m^{(n)*}|k = \boldsymbol{a}_m^{(n)}|_{m-k+1}.$$

Now, starting from $n = p$, we compute $\boldsymbol{a}_n^{(n)}$ for successively smaller values of n until $n = 1$ from

$$\boldsymbol{a}_{n-1}^{(n-1)} = \{\boldsymbol{a}_{n-1}^{(n)} + \boldsymbol{a}_n^{(n)}[\boldsymbol{a}_{n-1}^{(n)}]^*\}/\{1 - [a_n^{(n)}]^2\}. \quad (20)$$

Then, the autocorrelation function at the nth sampling instant is given by

$$r_n = \sum_{k=1}^{n} a_k^{(n)} r_{n-k}, \qquad \text{for } 1 \le n \le p. \quad (21)$$

The samples of the autocorrelation function are computed recursively for larger values of n starting from $n = 1$. Equation (21) is used to determine all the samples of the autocorrelation function with r_0 normalized to 1. The value of r_0 is finally determined from (19), and accordingly, all the values of the autocorrelation function are then normalized.

Finally, each element c_{ij} of the covariance matrix of the AR process (Σ) can be obtained by

$$c_{ij} = r(|i - j|), \qquad i, j = 1, 2, \cdots, T$$

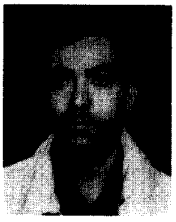
where T is the size of the covariance matrix. In our simulations (Section III-B), we chose $T \gg p$.

ACKNOWLEDGMENT

We wish to thank Dr. P. Kenny and Dr. A. Poritz for valuable discussions on various aspects of the hidden filter model, and we thank the reviewers for providing insightful comments that improve the quality of this paper.

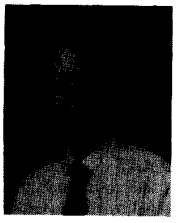
REFERENCES

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pt. 2, pp. 637–655, 1971.
[2] A. B. Poritz, "Hidden Markov models: A guided tour," in *Proc. ICASSP*, vol. 1, 1988, pp. 7–13.
[3] P. Kenny, M. Lennig, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 220–225, Feb. 1990.
[4] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. ICASSP*, May 1982, pp. 1291–1294.
[5] B. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals,"*IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 1404–1413, Dec. 1988.
[6] N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 563–570, Mar. 1991.
[7] J. R. Bellegarda and D. C. Farden, "Continuously adaptive linear predictive coding of speech," in *Proc. ICASSP*, May 1982, pp. 347–350.
[8] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734, Sept. 1982.
[9] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.
[10] K. Fukunaga, *Statistical Pattern Recognition*, 2nd ed. San Diego, CA: Academic, 1990.
[11] K. Fukunaga and T. F. Krile, "Calculation of Bayes' recognition error for two multivariate Gaussian distribution," *IEEE Trans. Comput.*, vol. C-18, pp. 220–229, Mar. 1969.
[12] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–110, 1943.
[13] L. Deng, M. Lennig, F. Seitz, and P. Mermelstein, "Modeling microsegments of stop consonants in a hidden Markov based word recognizer," *J. Acoust. Soc. Amer.*, vol. 87, pp. 2738–2747, 1990.
[14] L. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detection for isolated word recognition," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-29, pp. 777–785, Aug. 1981.
[15] V. W. Zue, *Notes on Speech Spectrogram Reading*. Cambridge, MA: M.I.T. Press, 1991.
[16] L. Deng and K. Erler, "Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech unit," *J. Acoust. Soc. Amer.*, vol. 92, pp. 3058–3067, Dec. 1992.
[17] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, vol. 27, pp. 65–78, Apr. 1992.
[18] ———, "A stochastic model of speech incorporating hierarchical nonstationarity," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 471–474, Oct. 1993.



Hamid Sheikhzadeh received the B.S. and M.S. degrees in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 1986 and 1989, respectively.

Since 1990 he has been a Research and Teaching Assistant in the Department of Electrical and Computer Engineering at the University of Waterloo, and has been pursuing the Ph.D. degree in electrical engineering. His research interests include signal processing and speech processing, with particular emphasis on speech recognition, speech enhancement, and auditory modeling.



Li Deng (S'83–M'86–SM'91) received the Ph.D. degree in electrical engineering from the University of Wisconsin, Madison, in 1986.

He worked on large vocabulary automatic speech recognition at INRS-Telecommunications, Montreal, Canada, from 1986 to 1989. Since 1989, he has been with Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada, where he is currently an Associate Professor. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge. His research interests include acoustic–phonetic modeling of speech, automatic speech recognition, statistical methods for signal analysis, computational phonology, auditory signal processing, and auditory neuroscience. In these areas, he has written more than 50 published papers.