A Minimax Classification Approach with Application to Robust Speech Recognition

Neri Merhav, *Member, IEEE*, and Chin-Hui Lee, *Member, IEEE*

***Abstract*— A minimax approach for robust classification of parametric information sources is studied and applied to isolated-word speech recognition based on hidden Markov modeling. The goal is to reduce the sensitivity of speech recognition systems to a possible mismatch between the training and testing conditions. To this end, a generalized likelihood ratio test is developed and shown to be optimal in the sense of achieving the highest asymptotic exponential rate of decay of the error probability for the worst-case mismatch situation. The proposed approach is compared to the standard approach, where no mismatch is assumed, in recognition of noisy speech and in other realistic mismatch situations.**

I. Introduction

A PROBLEM frequently encountered in speech recognition is mismatch between the underlying spectral characteristics associated with the training and testing conditions. This mismatch may arise from differences in recording conditions, e.g., background, time-varying channel characteristics, insufficient training data, speaker identity mismatch, additive noise, etc., or from variability in speech style, e.g., rate, intonation, accent, stress, etc. The issue of designing speech recognition systems that are robust to these types of spectral mismatch has been a long standing challenge for many researchers over the recent years (see, e.g., [1]–[36]). In particular, considerable effort has been devoted to the important special case of mismatch in the level of background noise between the training and recognition phases. Generally speaking, there are three major different approaches of handling wide-band noise in speech recognition systems.

In one approach the influence of noise is included in a statistical model of the speech signal (e.g., hidden Markov model (HMM)) and a classification rule that corresponds to the probability distribution (PD) of the noisy speech signal is applied. Theoretically, this is the most rigorous statistical approach because if the maximum *a posteriori* (MAP) decision rule is used, then the classification error rate is minimized [2], [27], [29]. The limitation of this method is that it requires complete knowledge of the unavailable underlying PD's and hence these have to be reliably estimated from the training data. The performance of this approach depends strongly on the availability of a faithful statistical model for the clean speech and the noise.

Manuscript received September 18, 1990; revised May 22, 1992.

N. Merhav was with AT&T Bell Laboratories, Murray Hill, NJ. He is now with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel.

C.-H. Lee is with AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9204546.

Another approach is associated with an estimator of the clean speech feature vectors or with a speech enhancement system at the receiving front end [9]–[14], [30]. The advantage of this approach is that it allows a high degree of flexibility as one can select any statistical model for the signal and for the noise and any estimation scheme. There are three major drawbacks, however, in this approach. First, the estimation error is usually not taken into account by the recognizer but the estimated vectors are treated as if they were clean. Second, some information that is useful for discrimination is lost in the estimation process. Finally, similarly to the first approach mentioned above, this method relies on a high degree of prior knowledge of the statistical characteristics of the noise which may not be available in practice.

The third approach is associated with robust front-end signal processing in the sense of reducing the noise sensitivity of the feature vectors [17]–[19], [31]–[33]. This includes spectral shaping methods [17], [33] which remove the influence of noise to some extent and methods based on modeling of the auditory system [18], [19], [31], [32]. In the latter, the underlying idea is to perform signal processing that simulates faithfully the operation of the human ear. It is believed that such processing is more robust to noise than that of the ordinary feature vectors commonly used. The disadvantages of this approach is that first, it is difficult to combine with other noise canceling schemes and second, very little is known and understood as to why noise sensitivity is reduced by this approach and what is the best model for the auditory system.

As mentioned earlier, the above approaches are all concerned with robustness of speech recognition systems against mismatch caused by additive noise. Additional techniques are usually aimed at other specific types of mismatch situations. These include stress compensation techniques [21], [34], [35], multistyle compensation methods [36], application of robust distortion measures [4], [17], [23]–[25], and novel representations of speech [7], [18]. These methods are summarized in [5] along with more details. Finally, another application of robust speech recognition is the case of channel mismatch due to variability of the transfer function. Some of the techniques used in combating additive noise can be used in this case as well (robust distortion measures [4], [17], [23]–[25], [33], auditory models [18], [19], [31], [32]). Other techniques are based on cepstral deconvolution [61], [12], [14], where the underlying idea is to subtract the average cepstrum of the transfer function.

A desirable objective is to develop a general robust speech recognition system that is capable of handling mismatch in

any of the above-mentioned adverse conditions rather than a specific condition. In a general mismatch situation the spectral distortion, even if small, might be of *any* form. A natural approach for improving robustness in the general case is to attempt to minimize the classification error rate for the *worst case* mismatch within some class. This is called the *minimax approach* and it was first introduced by Huber (see, e.g., [37]–[39]) for robust statistical inference. In [38] a robust version of the likelihood ratio test was developed for the two-class classification problem, where the probability measures P_0 and P_1, associated with the two classes, are not known accurately. The uncertainty about these measures was modeled in [38] *nonparametrically* by assuming that the true underlying measure lies in some neighborhood of either of the idealized (or nominal) measures P_0 and P_1. Specifically, this neighborhood was assumed to contain all measures of the form $Q = (1 - \varepsilon_i)P_i + \varepsilon_i H_i, i = 0, 1$, where $0 \leq \varepsilon_i < 1$ are fixed numbers that reflect the degree of uncertainty and H_i are arbitrary probability measures in some class $\boldsymbol{H}$. Assuming that the neighborhoods surrounding P_0 and P_1 do not overlap, Huber proposed optimal likelihood ratio tests for the worst-case (least favorable) pair of distributions from these neighborhoods. In [37, chap. VI] other nonparametric types of neighborhood, induced by various measures of distance between PD's, were considered as well. Later, the minimax approach has been widely investigated and studied with application to robust signal detection as well as in other application areas, e.g., parameter estimation, filtering, and coding. (See [40] and references therein.)

In this paper, we assume a general type of mismatch between the training and testing conditions and, therefore, adopt the minimax approach. Unlike [37]–[40], however, the neighborhoods of the nominal PD's are defined *parametrically*, yielding a formulation similar to that of the parametric composite hypotheses testing problem [41]. Specifically, the mismatch is modeled by allowing the underlying parameter vector, associated with the source of each word to be recognized, to be in a certain neighborhood of the parameter estimated from training utterances of the same word. The neighborhood is modeled in the parameter space rather than in the space of PD's for two reasons. First, if the underlying source is indeed a member of the parametric family and the only uncertainty is associated with the parameter value, then one expects that the parametric minimax approach will outperform the nonparametric approach as it incorporates more prior knowledge about the source. Second, the resulting test is relatively simple to implement.

A generalized likelihood ratio test (GLRT) is derived that attains the highest possible asymptotic rate of exponential decay of the error probability for the worst-case mismatch within the allowed neighborhood. This GLRT, which is similar to the test commonly used in parametric composite hypotheses testing problems (see, e.g., [41]), suggests that rather than the standard approach of comparing likelihood values associated with the testing signal for each trained model pointwise, one should first maximize the likelihood of this signal within the assumed mismatch neighborhood of each trained model, and then pick the word yielding the highest maximum. Clearly, the GLRT described above depends strongly on the specific parametric model being used as well as the topology of the neighborhood that models the mismatch, and the main problem in applying this methodology to speech recognition, is how to choose properly these two ingredients. The principles that guided us in choosing these key elements were based on theoretical results on one hand, and our wish to keep the recognition scheme as simple as possible, on the other.

Specifically, in [42] it has been proved analytically that if the underlying process is Gaussian, then under fairly mild regularity conditions, the sample cepstral coefficients are asymptotically uncorrelated and their asymptotic variances are independent of the spectrum of the underlying process. In view of this result, cepstral hidden Markov models (HMM's) with diagonal covariance matrices are adopted with a simple model of mismatch involving deviations in the cepstral means only.

As mentioned earlier, the resulting minimax test is easy to implement and it does not require any modification in the training procedure. Some degree of improvement in performance is usually attained as will be seen later. The minimax classification rule is also suitable for operating in fairly general mismatch situations because no particular assumptions are made as for the origin of the mismatch. Furthermore, it can be easily integrated with many other parametric robust recognition schemes, thereby combining the advantages of two or more methods. For instance, a speech recognition system that is designed specifically to handle a mismatch caused by additive white noise in the testing phase, is expected to perform better (in this particular mismatch situation) than the more general approach proposed here, which does not assume any prior knowledge on the particular type of mismatch. However, a small deviation from the additive white noise assumption might cause a "breakdown" in the performance of the former system while the latter still performs reasonably well. A combination of the two methodologies (see, e.g., [29]) is expected to result in an improvement in the additive white noise case while still preserving robustness to more general mismatch situations. A limitation of the proposed minimax approach is that it cannot be extended easily to connected speech recognition.

Experimental results based on isolated word, multispeaker recognition of noisy spoken versions of the English digits (with clean training data) show that the minimax approach gives significant improvement over the standard approach for a wide range of signal to noise ratio values. It also turns out that the error rate is fairly insensitive to design parameters that quantify the size and the shape of the mismatch neighborhood, and thus exact knowledge of these parameters is not crucial. A second experiment on the English digits, where training and testing were performed on two different databases that were recorded under completely different conditions (and hence mismatch is expected), also shows considerable improvement in the proposed approach. However, when the minimax approach has been tested on spoken versions of the English E-set letters, no significant improvement has been obtained. A simple possible reason for this fact is explained in the sequel and conclusions are drawn as for the situations in which the minimax approach is recommended.

The outline of the paper is as follows. In Section II the problem is formulated and the main result is stated and proved. Section III discusses the proposed classification rule and its relation to previous research. In Section IV, several guidelines are suggested as for the choice of the statistical model and the type of mismatch neighborhood to be used in speech recognition. In Section V, the main numerical procedures are described and the experimental results are presented. Finally, in Section VI some conclusions are summarized.

II. Problem Formulation and a Theoretical Result

Let $\{p_\lambda(\cdot), \lambda \in \Lambda\}$ be a parametric family of probability density functions (PDF's), where λ is a parameter vector, $\Lambda \subseteq \boldsymbol{R}^N$ is the parameter space, and $\boldsymbol{R}^N$ is the N-dimensional Euclidean space. For instance, $p_\lambda(\cdot)$ can be a PDF of a Gaussian cepstral HMM where λ denotes the set of parameters consisting of the initial state probabilities, the state transition probabilities, and the means and the variances associated with cepstral vectors in each state. Let $\lambda_i \in \Lambda$ be the parameter vector of the ith source, $1 \leq i \leq M$, which will be assumed known or given from a training procedure. In isolated-word speech recognition applications, λ_i denotes the parameter vector associated with the ith vocabulary word and M is the vocabulary size. Let $\Lambda_i, 1 \leq i \leq M$, denote nonoverlapping subsets of Λ, where $\lambda_i \in \Lambda_i$ for all i. The subset Λ_i will be henceforth referred to as the *mismatch neighborhood* or just the *neighborhood* of λ_i.

Comment: The assumption that λ_i are known accurately reflects our wish to focus on situations where the estimation error of the training phase is relatively small compared to deviations due to mismatch between the training and testing conditions. An alternative interpretation is that the estimation error is included in the mismatch neighborhood model.

Let $x = (x(1), x(2), \cdots, x(n)), x(t) \in \boldsymbol{X}, t = 1, 2, \cdots, n$, be a test sequence where $\boldsymbol{X}$ is the source alphabet which may be either a finite set, the real line $\boldsymbol{R}$, or the q-dimensional Euclidean space $\boldsymbol{R}^q$. For instance, in the cepstral HMM example mentioned above, $\boldsymbol{X} = \boldsymbol{R}^q$ and x is a sequence of q-dimensional cepstral vectors extracted from n successive frames of a spoken word to be recognized. It will be assumed that the test sequence x is generated by a PDF p_λ, where $\lambda \in \Lambda_m$ for some integer $1 \leq m \leq M$, where m is an unknown random variable. Given x and $\Lambda_i, 1 \leq i \leq M$, the classification problem is that of identifying m, the index of the neighborhood Λ_m that contains the underlying parameter vector λ. In other words, we wish to classify x into one out of M given sources (words) where the parameter λ of the source that governs x is allowed to depart from (mismatch) the nominal parameter value λ_m, associated with the true source m, but only within the neighborhood Λ_m. A decision rule Ω is a partition of the sample space of all possible test sequences $\boldsymbol{X}^n$, into M regions $\Omega_1, \Omega_2, \cdots, \Omega_M$ such that x is classified to the ith source if $x \in \Omega_i$. The worst-case probability of error $p_\Omega(e)$, associated with a decision rule Ω, is defined as

$$p_\Omega(e) \triangleq \sum_{i=1}^{M} p_i \max_{\lambda \in \Lambda_i} \int_{\Omega_i^c} p_\lambda(x)\, dx \tag{2.1}$$

where Ω_i^c is the complement set of Ω_i and p_i is the prior probability of the ith source, namely the probability that $m = i$. We seek a decision rule that minimizes the worst-case probability of error.

Unfortunately, the exact minimization of (2.1) is not trivial. It is desirable, however, to minimize (2.1) at least asymptotically as $n \to \infty$ for two reasons. First, although in practice n might not be large, it turns out that good decision rules in an asymptotic sense might yield fairly good performance even when n is small (see, e.g., [44]). Second, theoretical analysis of the error probability can be usually carried out only in the asymptotic limit.

The asymptotic approach that will be adopted here is supported by the fact that for many commonly used parametric families of PDF's $p_\lambda(\cdot)$ (e.g., independently identically distributed random processes, Markov processes, hidden Markov processes) the error probability (2.1) can be made exponentially small as a function of n. (See, e.g., [44]–[49] and references therein). Suppose that for the optimal classification rule, $p_\Omega(e)$ decays exponentially with n, say, $p_\Omega(e) \approx Ce^{-Bn}$ for some positive constants B and C. Then, clearly this optimal rule yields the highest possible exponential rate $B = B_{\max}$. We will demonstrate a suboptimal classification rule that is asymptotically optimal in the sense of yielding an error probability with the highest possible decay rate $B_{\max}$. In other words, for every arbitrarily small $\varepsilon > 0$ and all sufficiently large n, the error probability associated with our rule will be less than $C \exp(-(B_{\max} - \varepsilon)n)$. One might claim that this error probability might be $e^{\varepsilon n} \to \infty$ times larger than the minimum above. Note, however, that this factor is dominated by the decaying exponential $e^{-B_{\max} n}$ and hence exponentially insignificant. An alternative formulation of the above claim which will be used also in Theorem 1 below is that we minimize the limit of $n^{-1} \log p_\Omega(e)$ as $n \to \infty$.

For simplicity, assume that the Λ_i is a bounded subset of Λ and confine attention to a sequence of finite grids $\{\Lambda_i^n\}_{n \geq 1}, \Lambda_i^n \subset \Lambda_i, 1 \leq i \leq M$, of parameter values where as n grows, Λ_i^n becomes dense in Λ_i. We shall state the result of asymptotic minimization of $p_\Omega(e)$ in the above defined sense where the Λ_i in (2.1) are replaced by Λ_i^n. A similar result for the continuous subsets Λ_i can be shown from standard continuity arguments, however, the proof for this case is more involved and it requires stronger regularity conditions concerning uniform continuity of $p_\lambda(x)$ in λ, hence it is less general. In the discrete case considered here, on the other hand, the proof is fairly simple and general. It also exhibits a realistic situation since, in practice, λ can be represented with finite precision only.

We assume the following regularity conditions.

A1) $\Lambda_i, 1 \leq i \leq M$ are bounded sets.

A2) The grid $\Lambda_i^n \subset \Lambda_i$ becomes dense in Λ_i as n tends to infinity, i.e., for the Euclidean metric $d : \Lambda \times \Lambda \to \boldsymbol{R}^+$,

$$\lim_{n \to \infty} \min_{\lambda' \in \Lambda_i^n} d(\lambda, \lambda') = 0 \qquad \forall \lambda \in \Lambda_i. \tag{2.2}$$

A3) The cardinality $|\Lambda_i^n|$ of Λ_i^n is less than $e^{n\varepsilon_n}$ for some positive sequence $\{\varepsilon_n\}_{n \geq 1}$ that tends to zero.

The existence of a sequence of grids satisfying both conditions A2) and A3) can easily be implied from the boundedness of Λ_i and from the fact that it is a subset of a finite dimensional space. For example, if the number of grid points grows linearly with n in each dimension of $\boldsymbol{R}^N$, then (2.2) decays as fast as $1/n$ and at the same time $|\Lambda_i^n|$ is proportional to n^N, which is a polynomial resulting in $\varepsilon_n = n^{-1}N \log n$, where throughout the sequel, logarithms are taken to the base e unless otherwise specified.

Our main result is the following.

Theorem 1: Let $\Omega^* = \Omega^*(n)$ be a decision rule defined by

$$\Omega_i^*(n) \triangleq \{x \in \boldsymbol{X}^n : p_i \cdot \max_{\lambda \in \Lambda_i^n} p_\lambda(x) = \max_{1 \le j \le M} [p_j \cdot \max_{\lambda \in \Lambda_j^n} p_\lambda(x)]\}, \qquad 1 \le i \le M \tag{2.3}$$

where ties broken arbitrarily, and let Λ_i^n satisfy A1)–A3). Then, for any decision rule $\Omega = \Omega(n)$,

$$\frac{1}{n} \log p_{\Omega^*}(e) \le \frac{1}{n} \log p_\Omega(e) + \varepsilon_n \tag{2.4}$$

where $p_\Omega(e)$ and $p_{\Omega^*}(e)$ are defined as in (2.1) but with Λ_i replaced by Λ_i^n and ε_n is as in A3).

Proof of Theorem 1: For a given decision rule Ω let

$$\mu_\Omega(e) \triangleq \sum_{i=1}^M p_i \int_{\Omega_i^c} \max_{\lambda \in \Lambda_i^n} p_\lambda(x)\, dx. \tag{2.5}$$

Obviously, for every $\Omega, \mu_\Omega(e) \ge p_\Omega(e)$. However,

$$\begin{aligned} \mu_\Omega(e) &\le \sum_{i=1}^M p_i \int_{\Omega_i^c} \sum_{\lambda \in \Lambda_i^n} p_\lambda(x)\, dx \\ &= \sum_{i=1}^M p_i \sum_{\lambda \in \Lambda_i^n} \int_{\Omega_i^c} p_\lambda(x)\, dx \\ &\le \sum_{i=1}^M p_i |\Lambda_i^n| \max_{\lambda \in \Lambda_i^n} \int_{\Omega_i^c} p_\lambda(x)\, dx \\ &\le e^{n\varepsilon_n} p_\Omega(e). \end{aligned} \tag{2.6}$$

Next, observe from (2.5) that minimizing $\mu_\Omega(e)$ is equivalent to maximizing

$$\sum_{i=1}^M p_i \int_{\Omega_i} \max_{\lambda \in \Lambda_i^n} p_\lambda(x)\, dx$$

which is clearly attained for $\Omega = \Omega^*$ defined in (2.3). (This can be shown in a way similar to the proof of optimality of the classical maximum *a posteriori* decision rule). Hence for any decision rule Ω we have

$$\begin{aligned} \frac{1}{n} \log p_{\Omega^*}(e) &\le \frac{1}{n} \log \mu_{\Omega^*}(e) \\ &\le \frac{1}{n} \log \mu_\Omega(e) \\ &\le \frac{1}{n} \log e^{n\varepsilon_n} p_\Omega(e) \\ &= \frac{1}{n} \log p_\Omega(e) + \varepsilon_n. \end{aligned} \tag{2.7}$$

This completes the proof of Theorem 1.

A slightly weaker but considerably simpler version of the above result is that rather than minimizing $p_\Omega(e)$ of (2.1) which is a difficult task, we minimize an upper bound on $p_\Omega(e)$ given by

$$\tilde{\mu}_\Omega(e) \triangleq \sum_{i=1}^M p_i \int_{\Omega_i^c} \max_{\lambda \in \Lambda_i} p_\lambda(x)\, dx \tag{2.8}$$

which is similar to $\mu_\Omega(e)$ but with the finite grids Λ_i^n replaced by the continuous sets Λ_i. The minimization of $\tilde{\mu}_\Omega(e)$ is accomplished by the rule (2.3) where again, Λ_i^n are substituted by Λ_i. Note that here no regularity conditions are required except for integrability of $\max_{\lambda \in \Lambda_i} p_\lambda(x)$ as (2.8) is never smaller than (2.1). The result, however, is somewhat weaker as we are not minimizing the desired error probability but an upper bound on the error probability.

III. Discussion

The proposed decision rule Ω^* has an easy intuitive interpretation. Since the underlying parameter under each hypothesis is not known accurately, a two-step procedure is used. First, estimate the underlying parameter using the maximum likelihood (ML) approach within each neighborhood Λ_i (or Λ_i^n), i.e., compute $\lambda_i = \arg\max_{\lambda \in \Lambda_i} p_\lambda(x)$. Then, apply the maximum *a posteriori* (MAP) decision rule with λ_i replaced by $\hat{\lambda}_i, 1 \le i \le M$. Clearly, one can use other estimation approaches for the first step, for example, the minimum discrimination information (MDI) approach (see, e.g., [50]–[53]). The theorem tells us, however, that if the ML approach is chosen, then this two-step procedure is asymptotically optimal in a minimax sense.

The decision rule Ω^* is similar to the GLRT which is commonly used heuristically in two-class parametric composite hypothesis testing problems under the Neyman–Pearson formulation, where a uniformly most powerful test (UMPT) does not exist [41]. In these situations, if one wishes to test the hypothesis $H_0 : \lambda \in \Lambda_0$ against the alternative $H_1 : \lambda \in \Lambda_1$, a decision is made by comparing the generalized likelihood ratio, $\max_{\lambda \in \Lambda_1} p_\lambda(x) / \max_{\lambda \in \Lambda_0} p_\lambda(x)$, to a prescribed threshold. More recently, the GLRT was proved an asymptotically optimal solution in the Neyman–Pearson sense for various more specific classification problems: in [54] and [55] for testing whether two sequences were drawn from the same source, in [56] for testing for randomness and for independence, and in [57] and [58] for estimating the order of a statistical model. In contrast to our assumption, however, [54]–[58] are concerned with situations where the regions of uncertainty Λ_i are *nested* rather than disjoint.

In [29] a test similar to Ω^* has been applied to recognition of clean as well as noisy speech for the special case of gain mismatch using hidden Markov models (HMM's), based on the intuition described in the first paragraph of this section. Specifically, in the case of clean speech, it was assumed in [29] that the test sequence x is a sequence of vectors $x_t \in \boldsymbol{R}^q, t = 1, 2, \cdots, n$, drawn from a source p_λ which may differ from the nominal (training) source p_{λ_i} but only in a sequence of scaling parameters (gain factors) $\{g_t\}_{t=1}^n$ corresponding to

the vectors $\{x_t\}_{t=1}^n$. Namely, the neighborhood Λ_i of λ_i in this case consists of all parameter vectors λ which agree with λ_i in all components except for the gain factors. The proposed recognition test [29] was to pick the index i that maximizes the quantity $\max_G |G|^{-1} p_{\lambda_i'}(G^{-1}x)$, where x is represented as a qn-dimensional column vector, $G \triangleq \text{diag}(g_1 I, \cdots, g_n I)$ is a $(qn) \times (qn)$ diagonal matrix of the gain factors, I is the $q \times q$ identity matrix, and λ_i' is the vector of all components of λ_i except for the gain factors. Note, however, that in this case the dimension of Λ grows with n as λ contains n gain factors as components. Since Theorem 1 assumes the dimension N of Λ to be fixed and finite, it does not hold here. The weaker version of the Theorem, described at the end of Section II, however, holds for [29].

Another interesting special case of Ω^* arises when the subsets Λ_i form a *partition* of the parameter space Λ. In this case, Ω^* suggests that upon observing x, we first calculate the unconstrained ML parameter estimate $\hat{\lambda}_x \triangleq \arg\max_{\lambda \in \Lambda} p_\lambda(x)$, provided that it is unique, and then choose the source i for which $\hat{\lambda}_x \in \Lambda_i$. More specifically, if Λ_i is defined as the set of all parameter vectors λ for which $d(\lambda, \lambda_i) = \min_j d(\lambda, \lambda_j)$ for some distance measure $d : \Lambda \times \Lambda \to \boldsymbol{R}^+$, then Ω^*, in this case, picks the source i that minimizes $d(\hat{\lambda}_x, \lambda_i)$, namely, the nearest neighbor to $\hat{\lambda}_x$. In this case, since the neighborhoods are not isolated from each other, an exponential decay of $p_\Omega(e)$ cannot be expected, though Theorem 1 still holds.

IV. Choice of Parametric Family and Mismatch Neighborhoods

So far we have discussed the minimax classification approach for a general parametric family of sources $\{p_\lambda, \lambda \in \Lambda\}$ with an arbitrary type of mismatch neighborhood Λ_i. A key issue in applying the minimax decision rule proposed here to speech recognition, is an appropriate choice of these ingredients. In this section, an attempt is made to draw some guidelines which may suggest a reasonable choice of the parametric family and neighborhood type, and at the same time, keep the recognition scheme as simple as possible.

Consider first a stationary zero-mean Gaussian process $\{y_t\}$ with power spectrum density (PSD) $S(\omega)$, and the smoothed periodogram [59] for estimating the PSD, i.e., $\hat{S}_L(\omega) = \sum_{\tau=-L}^{L} \hat{R}(\tau) e^{j\omega\tau}$, where L is a fixed positive integer and $\hat{R}(\tau)$ is the empirical autocorrelation given by

$$\hat{R}(\tau) = T^{-1} \sum_{t=1}^{T-\tau} y_t y_{t+\tau}$$

T being the number of observations available. Next define the empirical cepstrum $\{\hat{c}_\tau^L\}, \tau = 0, 1, \cdots$, as the inverse Fourier transform of $\log \hat{S}_L(\omega)$. It is shown in [42] (along with more details) that under certain regularity conditions on the PSD $S(\omega)$, for every two fixed positive integers k and l,

$$\lim_{L\to\infty} \lim_{T\to\infty} T \cdot \text{cov}(\hat{c}_k^L, \hat{c}_l^L) = \delta_{kl} \tag{4.1}$$

where $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$, otherwise. Equation (4.1) tells us that if L is large and $T \gg L$, then the empirical cepstral coefficients are essentially uncorrelated and have same variance, which is approximately $1/T$, *independently* of the underlying PSD. More commonly [61], the estimated cepstrum coefficients $\{\tilde{c}_\tau^q\}$ are based on a qth-order AR spectrum estimator rather than the smoothed periodogram, i.e.,

$$\tilde{S}_q(\omega) = \hat{\sigma}^2 \left| 1 + \sum_{k=1}^{q} \hat{a}_k e^{-j\omega k} \right|^{-2} \tag{4.2}$$

where $\hat{\sigma}^2$ and $(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_q)$ are estimates of the gain and the AR coefficients, respectively, derived from Yule–Walker equations [62]. Here a result somewhat weaker than (4.1) holds (see [42]): If the underlying process is AR of fixed order p, then the asymptotic covariance matrix (as $T \to \infty$) of the first q empirical cepstrum coefficients tends (as $q \to \infty$) to the identity matrix in the Hilbert–Schmidt sense [63]. In other words, let ρ_{kl} be the asymptotic covariance $\lim_{T\to\infty} T \cdot \text{cov}(\tilde{c}_k^q, \tilde{c}_l^q)$, then $q^{-1} \sum_{k,l=1}^{q} |\rho_{kl}^q - \delta_{kl}|^2 \to 0$ as $q \to \infty$.

Suppose temporarily, that the nominal underlying sources, $1 \le i \le M$, are Gaussian and in the training phase, where the parameter λ_i, of each source is extracted, we observe for each i vectors of empirical cepstral coefficients of the form $\hat{c}^L = (\hat{c}_1^L, \hat{c}_2^L, \cdots, \hat{c}_q^L)$ (or $\tilde{c}^q = (\tilde{c}_1^q, \tilde{c}_2^q, \cdots, \tilde{c}_q^q)$). In light of (4.1) and the central limit theorem (similar to [60]), we wish to model this random vector as a Gaussian vector with a parameter λ_i consisting of a vector of means (which depends strongly on the underlying PSD), and the elements of a covariance matrix (which is roughly diagonal and depends weakly on the PSD). Suppose further that in the testing session, one of the Gaussian sources, say m, generates a sequence but its underlying PSD somewhat departs from that of the corresponding training source (e.g., due to linear distortions and noise) and again, we have access to the empirical cepstral coefficients. Since the source is still assumed Gaussian, then in view of (4.1), the deviation (mismatch) in the PSD is reflected mostly in the cepstral means, while the covariance matrix is essentially unaffected. This observation is further supported by several experimental studies of mismatch caused by noise [4], stress [34], [35], multistyle speech [36], and simulations on artificially generated data [64], and it suggests that the model of mismatch neighborhood Λ_i will be confined to the means only. An additional motivation for not including the cepstral variances in the mismatch model is our wish to keep the recognition scheme as simple as possible. We will allow, however, different variances for the different cepstral components at different states.

The next step is to define the shape of the mismatch neighborhood in the space of cepstral mean vectors. First observe that for large L and T the expected value of $\hat{c}_\tau^L$ (or of $\tilde{c}_\tau^q$ for large enough q) can be well approximated (see [42]) by the inverse Fourier transform of the true log spectrum, i.e.,

$$c_\tau \triangleq \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \cdot e^{j\omega\tau} \log S(\omega). \tag{4.3}$$

Hence, we shall refer to (4.3) as an approximation of the cepstral mean. Assume further, that the PSD is given by a

rational function of $e^{j\omega}$, i.e.,

$$S(\omega) = G \cdot \prod_{i=1}^{k} \frac{(1-\beta_i e^{-j\omega})(1-\beta_i e^{j\omega})}{(1-\alpha_i e^{-j\omega})(1-\alpha_i e^{j\omega})} \tag{4.4}$$

where $|\alpha_i|$ and $|\beta_i|, 1 \le i \le k$, are all less than 1, and G is a gain factor. It is easy to show, in this case, that for $\tau \ge 1$,

$$c_\tau = \frac{1}{\tau}\sum_{i=1}^{k}(\alpha_i^\tau - \beta_i^\tau). \tag{4.5}$$

Now let $S_1(\omega)$ and $S_2(\omega)$ be two distinct rational PSD's of the form (4.4) with parameters $\{(\alpha_i, \beta_i)\}_{i=1}^k$ and $\{(\gamma_i, \delta_i)\}_{i=1}^k$, respectively, and let $0 \le \rho < 1$ denote the maximum modulus:

$$\rho = \max_{1\le i\le k} \max\{|\alpha_i|, |\beta_i|, |\gamma_i|, |\delta_i|\}. \tag{4.6}$$

The difference between the cepstra $c_\tau^{(1)}$ and $c_\tau^{(2)}$ associated with $S_1(\omega)$ and $S_2(\omega)$, respectively, is bound above as follows:

$$\begin{aligned} |c_\tau^{(1)} - c_\tau^{(2)}| &= \frac{1}{\tau}\left|\sum_{i=1}^{k}(\alpha_i^\tau - \beta_i^\tau + \gamma_i^\tau - \delta_i^\tau)\right| \\ &\le \frac{1}{\tau}\sum_{i=1}^{k}(|\alpha_i^\tau| + |\beta_i^\tau| + |\gamma_i^\tau| + |\delta_i^\tau|) \\ &\le \frac{4k\rho^\tau}{\tau}. \end{aligned} \tag{4.7}$$

This suggests that the mismatch neighborhood in the space of cepstral means will be the set of all vectors whose τth component differs from the corresponding nominal value by a quantity proportional to $\tau^{-1}\rho^\tau$ for some $0 \le \rho < 1$. The mismatch model (4.7) agrees qualitatively with both analytical and experimental results of [4], where the influence of noise on cepstral mismatch has been observed. Specifically, it has been shown in [4] that the higher order cepstral coefficients are less affected by noise than the lower order coefficients. In particular, the mismatch decays exponentially with the index of the coefficient (see, e.g., [4, eq. (2.19)] which quite resembles the model (4.7)).

In statistical modeling of speech signals, it is common to use a set of estimated cepstrum coefficients as a spectrum related feature vector. Theoretically, in view of (4.1), this is convenient as it allows us to assume a diagonal covariance matrix. Experimentally, this is supported by previously reported studies (e.g., [61], [65]) which show that among different kinds of feature vectors: AR parameter set, impulse response, autocorrelation sequence, area function, and cepstrum function, the latter provides the highest recognition accuracy. To handle the nonstationarity of speech signals, however, the above-described Gaussian cepstrum model is combined with the HMM [66], [67]. We next provide a mathematical description of the cepstral HMM and define the form of Λ_i in this combined model.

Let $x = \{x(t), t = 1, \cdots, n\}$ be an observation sequence with $x(t) = \hat{c}^q(t) \in \boldsymbol{R}^q$ being an empirical cepstral vector based on the AR spectrum estimate (e.g., (4.2)) calculated from the tth frame (of length T) of the waveform. (We use $\check{c}^q(t)$ rather than $\hat{c}^L(t)$ as the former can be calculated recursively from the AR parameter estimates [61].) Consider cepstral HMM's with S states, where the state set is denoted by $\boldsymbol{S} = \{1, 2, \cdots, S\}$. Let $s = \{s_t, t = 1, \cdots, n\}, s_t \in \boldsymbol{S}$, be an unobserved sequence of states corresponding to x. The PDF of x is given by

$$p_\lambda(x) = \sum_s p_\lambda(x, s) = \sum_s p_\lambda(s) p_\lambda(x|s) \tag{4.8}$$

where $p_\lambda(s)$ is the probability of the sequence of states s, and $p_\lambda(x|s)$ is the probability of the given output sequence x given s. For first-order HMM's we have

$$p_\lambda(s) = \prod_{t=1}^{n} a_{s_{t-1}s_t} \tag{4.9}$$

where $a_{s_{t-1}s_t}$ denotes the transition probability from state s_{t-1} at time $(t-1)$ to state s_t at time t, and $a_{s_0 s_1} \triangleq \pi_{s_1}$ is the initial state probability. For $p_\lambda(x|s)$ we assume that

$$p_\lambda(x|s) = \prod_{t=1}^{n} b(x_t|s_t) \tag{4.10}$$

where $b(x_t|s_t)$ is a probability density function (PDF) of a q-dimensional Gaussian vector with uncorrelated components, $\{x_l(t)\}$, given by

$$b(x_t|s_t) = \prod_{l=1}^{q}\left\{[2\pi\sigma_l^2(s_t)]^{-1/2}\exp\left[-\frac{(x_l(t)-\mu_l(s_t))^2}{2\sigma_l^2(s_t)}\right]\right\}. \tag{4.11}$$

The parameter set of the cepstral HMM is $\lambda \triangleq (\pi, A, \mu, \Sigma)$, where $\pi = \{\pi_\alpha\}, A = \{a_{\alpha\beta}\}, \mu = \{\mu_l(\alpha)\}$, and $\Sigma = \{\sigma_l^2(\alpha)\}, \alpha, \beta \in \boldsymbol{S}, l = 1, 2, \cdots, q$, and $\lambda \in \Lambda$. We shall assume a *left-to-right* HMM [66, p. 266] for which $a_{\alpha\beta}$ vanishes for all $\beta < \alpha$ and $\beta > \alpha + 1$, namely, only the self transition and a transition to the next state are allowed. Clearly, once the last state $\alpha = S$ is visited, the process will remain in this state.

Following (4.1) and (4.7), the neighborhood Λ_i of the parameter $\lambda_i \triangleq \{\pi^{(i)}, A^{(i)}, \mu^{(i)}, \Sigma^{(i)}\}$, associated with the ith source, will be defined as

$$\begin{aligned} \Lambda_i = \{&(\pi, A, \mu, \Sigma) : \pi = \pi^{(i)}, A = A^{(i)}, \Sigma = \Sigma^{(i)}, \\ &|\mu_l(\alpha) - \mu_l^{(i)}(\alpha)| \le Cl^{-1}\rho^l, \alpha \in S, l = 1, 2, \cdots, q\} \end{aligned} \tag{4.12}$$

for some constants $C > 0$ and $0 \le \rho < 1$. We assume $\pi = \pi^{(i)}$ and $A = A^{(i)}$ for two reasons. First, the likelihood function $p_\lambda(x)$ is relatively insensitive to these components, and second, to simplify the computation. As we shall see in the next section, the constrained maximization $\max_{\lambda\in\Lambda_i} p_\lambda(x)$, needed to implement the decision rule Ω^*, is relatively easy to perform in this case.

V. Numerical Procedures and Experimental Results

In the first part of this section, the main numerical procedures of preprocessing, training and testing are described. The preprocessing and training procedures are conventional and are brought here for the sake of completeness only. A reader familiar with these procedures may skip directly to

the description of the testing procedure, which includes the particular modification to the minimax classification rule. In the second part, speech recognition results are provided.

A. Numerical Procedures

Pre-emphasis, windowing, AR parameter estimation, and cepstrum analysis of the speech signal are applied to obtain the observation sequences associated with the training and testing utterances. Specifically, the speech signal is first pre-emphasized using a filter of the form $P(z) = 1 - \alpha z^{-1}$, and then, successive waveform frames of length T, shifted D samples from each other, are multiplied by the generalized Hanning window:

$$w(\tau) = \beta - (1-\beta)\cos\left(\frac{2\pi\tau}{T}\right), \qquad \tau = 0, 1, \cdots, T-1. \tag{5.1}$$

For each frame $1 \le t \le n$, first, a vector of empirical autocorrelation coefficients $\{\hat{R}_t(\tau)\}_{\tau=0}^{q}$ is calculated as described in Section IV. Then, using the Yule–Walker equations [62], a vector of AR parameter estimates $\hat{a}(t) = (\hat{a}_1(t), \hat{a}_2(t), \cdots, \hat{a}_q(t))$ is derived. Finally, the AR cepstrum $x(t) = \hat{c}^q(t)$ is calculated from $\hat{a}(t)$, as described earlier using the well-known recursive formula [61]:

$$\begin{aligned} x_1(t) &= \hat{a}_1(t) \\ x_l(t) &= \sum_{r=1}^{l-1}\left(1 - \frac{r}{l}\right)\hat{a}_r(t)x_{l-r}(t) + \hat{a}_l(t), \qquad 1 < l \le q. \end{aligned} \tag{5.2}$$

In the training phase, the parameter λ_i associated with each source is estimated from a given training sequence, $y_i = (y_i(1), \cdots, y_i(n_i)), y_i(t) \in \boldsymbol{R}^q, t = 1, \cdots, n_i$, (or a set of such training sequences), generated from the ith source, namely, a training utterance of the ith vocabulary word. To this end, the maximum-likelihood approach is adopted, that is, we wish to find a parameter value λ_i that maximizes $p_\lambda(y_i)$. In practice, the maximization of the likelihood function is approximated by the segmental K-means algorithm [68], [69]. For a given observation sequence, say, $y = (y(1), \cdots, y(n))$, this algorithm performs local joint maximization of $p_\lambda(y, s)$ over the state sequence s and the parameter λ. This maximization of $p_\lambda(y, s)$ by the segmental K-means algorithm, rather than of $p_\lambda(y)$ by the Baum algorithm [70]–[72] is numerically easier and results in similar λ estimates provided that the empirical cepstrum is calculated from relatively long waveform frames [73]. The segmental K-means algorithm performs alternate maximization of $p_\lambda(y, s)$, once over the state sequences s for a given $\lambda \in \Lambda$ using the Viterbi algorithm, and then over λ for the resulting most likely state sequence s^* using reestimation formulas similar to those of the Baum algorithm. Specifically, following (4.8)–(4.11), we wish to minimize for a given $s = s^*$ the quantity, (5.3), shown at the bottom of the page, where $K = 0.5nq\log 2\pi$ and $n(\alpha, \beta)$ is the number of transitions from state α to state β that appear in s. It is easy to see that (5.3) is minimized if we choose $\pi_{s_1} = 1, a_{\alpha,\beta} = n(\alpha, \beta)/n(\alpha), n(\alpha)$ being $\sum_\beta n(\alpha, \beta)$, and,

$$\mu_l(\alpha) = \frac{1}{n(\alpha)}\sum_{t:s_t=\alpha} y_l(t) \tag{5.4a}$$

$$\sigma_l^2(\alpha) = \frac{1}{n(\alpha)}\sum_{t:s_t=\alpha}[y_l(t) - \mu_l(\alpha)]^2. \tag{5.4b}$$

This iterative algorithm was shown [69] to converge to a local maximum under certain regularity conditions. For left-to-right HMM's considered here, initial model estimates are obtained from uniform segmentation of y into S intervals and estimating

$$\mu(\alpha) = \{\mu_l(\alpha)\}_{l=1}^{q} \quad \text{and} \quad \Sigma(\alpha) = \{\sigma_l^2(\alpha)\}_{l=1}^{q}$$

from the αth segment, $\alpha \in \boldsymbol{S}$. Here π and A are initially estimated from the state sequence defined by this segmentation. The initial matrix A is, therefore, left–right and this structure is preserved in each iteration of the segmental K-means algorithm (see [66, eq. (44)]).

In the testing phase, a similar iterative algorithm is used. To approximate $\hat{\lambda}_i = \arg\max_{\lambda\in\Lambda_i} p_\lambda(x)$, we initialize with λ_i which was obtained in the training procedure (consider this as the nominal parameter value), and in each iteration, we first decode s^* using the Viterbi algorithm, and then minimize (5.3) (with y replaced by x) over the $\{\mu_l(\alpha)\}$, subject to the constraints $|\mu_l(\alpha) - \mu_l^{(i)}(\alpha)| \le Cl^{-1}\rho^l, \alpha \in \boldsymbol{S}, 1 \le l \le q$. Since the $x_l(t)$ are assumed uncorrelated, this constrained minimization is carried out for each component $\mu_l(\alpha)$ individually. Furthermore, since (5.3) is a convex function of each $\mu_l(\alpha)$, the minimization over this variable can be performed by first, computing the unconstrained minimizing mean component $\tilde{\mu}_l(\alpha)$ similarly to (5.4a), and then checking whether its value falls in the interval $I = [\mu_l^{(i)}(\alpha) - Cl^{-1}\rho^l, \mu_l^{(i)}(\alpha) + Cl^{-1}\rho^l]$: If this condition is met, then the cepstral mean component $\hat{\mu}_l^{(i)}(\alpha)$, associated with $\hat{\lambda}_i$ (as defined in Section III), agrees with $\tilde{\mu}_l(\alpha)$. Otherwise, $\hat{\mu}_l^{(i)}(\alpha)$ is the endpoint of I which is closest to $\tilde{\mu}_l(\alpha)$. The remaining components forming $\hat{\lambda}_i$ are left unchanged, following (4.12).

From the computational point of view, the complexity of the testing algorithm is essentially equivalent to that of the standard algorithm (which just computes $p_{\lambda_i}(x)$), multiplied by the number of iterations required for convergence of the

$$\begin{aligned} -\log p_\lambda(y, s) &= K - \sum_{t=1}^{n}\log a_{s_{t-1}s_t} + \frac{1}{2}\sum_{t=1}^{n}\sum_{l=1}^{q}\left[\frac{[y_l(t) - \mu_l(s_t)]^2}{\sigma_l^2(s_t)} + \log\sigma_l^2(s_t)\right] \\ &= K - \log\pi_{s_1} - \sum_{\alpha,\beta\in S} n(\alpha, \beta)\log a_{\alpha\beta} + \frac{1}{2}\sum_{\alpha\in S}\sum_{l=1}^{q}\sum_{t:s_t=\alpha}\left[\frac{[y_l(t) - \mu_l(\alpha)]^2}{\sigma_l^2(\alpha)} + \log\sigma_l^2(\alpha)\right] \end{aligned} \tag{5.3}$$

TABLE I
RECOGNITION RESULTS FOR NOISY SPEECH

SNR	Standard	Minimax	C	ρ
5	64.50 ± 13.25	88.25 ± 9.50	5	0.8
10	78.75 ± 8.50	93.50 ± 4.75	4	0.8
15	87.25 ± 6.25	95.50 ± 2.50	4	0.8
20	95.25 ± 2.75	97.25 ± 2.00	3	0.8
30	99.00 ± 2.00	99.50 ± 1.50	1	0.8
∞	99.75 ± 0.25	100.00 ± 0.00	1	0.8

above described procedure. Fortunately, as we shall see in the next part of this section, the convergence is reasonably fast in most situations.

B. Experimental Results

A simple special case of a mismatch situation is encountered when the testing signal is corrupted by additive white Gaussian noise, while the training data are clean. To examine the performance of the minimax decision rule Ω^*, it was first compared to the standard rule which assumes no mismatch (namely, where Λ_i includes λ_i only), in multi-speaker isolated-word recognition of spoken versions of the $M = 10$ English digits, recorded from 4 different speakers: 2 males, 2 females, through a telephone handset and sampled at 6.667 kHz. For each speaker and each digit, 5 training utterances and 10 testing utterances were used. While the training procedure was performed on clean data, in the testing phase, computer-generated Gaussian white noise, with various levels of intensity, was added to the original waveform prior to the preprocessing. The signal-to-noise ratio (SNR) was defined in a segmental manner, that is, if the clean signal $s(t)$ contains n frames of length T, and the noise power is σ_w^2 then,

$$\text{SNR} \triangleq \frac{1}{n} \sum_{t=1}^{n} 10 \log_{10} \frac{E_t}{T\sigma_w^2} \tag{5.5}$$

where E_t is the sum of squares of the T waveform samples in the tth frame.

Several design parameters were first experimentally optimized in order to obtain the best recognition accuracy in the standard classification approach without noise: the frame length T, the frame shift D, the cepstrum vector dimension q, the number of states S, the pre-emphasis filter coefficient α, and the window shape parameter β. The best values for this experiment were found to be $T = D = 200, q = 12, S = 8, \alpha = 0.95$, and $\beta = 0.5$.

Table I compares, for several SNR values in decibels (first column), the recognition accuracy in % of the standard decision rule (second column) to that of the minimax approach (third column) for the best mismatch neighborhood parameter values: C in the range [1, 8] (fourth column), and ρ in the range [0, 1] (fifth column). Each average score is given along with an uncertainty term (that follows the "±" sign) which is the empirical 90% confidence interval of the average scores associated with the individual digits, i.e., the smallest symmetric interval around the average score that includes 9 out of the 10 scores for each one of the digits.

As can be seen, the minimax decision rule introduces considerable improvement, especially at low SNR values, where the degree of mismatch between the noisy testing data and the clean training data is relatively high. Note that the mismatch shape parameter ρ in Table I is insensitive to the SNR, while the mismatch size parameter C decreases with the SNR, as high SNR values correspond to a small amount of mismatch.

Strictly speaking, the performance of the proposed rule depends on the appropriate choice of ρ and C, which in turn depends on the unknown amount of mismatch. It turns out, however, that the performance is fairly insensitive to these parameters in a reasonably wide range. This means that exact knowledge of ρ and C is not crucial, as considerable improvement (though not optimal) can be accomplished for any pair (ρ, C) in a fairly large domain. This is demonstrated in Table II which presents average recognition accuracy of the minimax rule as a function of ρ and C for SNR = 10 dB, which corresponds to the second line in Table I. A similar behavior was observed for other values of the SNR as well. Note also that the best values of ρ and C are not highly sensitive to the SNR. However, these can be always tuned online to best adjust to the environmental conditions. This exhibits an alternative to a dithering technique [28] which has been used to train models in moderate noise levels and worked fairly well in a wide range of SNR values. Of course, when there is no mismatch between the training and testing conditions performance usually improves. However, while in [28] is suitable for the special case of additive noise, here the technique is more general.

As for the computational aspect of the choice of ρ and C, it turns out as one might expect, that the number of iterations needed for convergence of the iterative algorithm for approximating $\max_{\lambda \in \Lambda_i} p_\lambda(x)$ in the testing procedure (see Section V-A), grows with ρ and C. The number of iterations ranged between 1 and 5 in the domain of ρ and C considered here.

Not surprisingly, the minimax classification rule is not comparable in performance to recognition schemes that are designed specifically to deal with noisy speech and utilize prior knowledge about the nature of the noise. To demonstrate the price paid for not using this prior knowledge, a comparison with the gain-adapted scheme of [29] showed a loss of about 5 dB in the minimax approach. This means that with the scheme of [29] similar recognition scores as in Table I are obtained when the SNR is 5 dB lower. It should be kept in mind, however, that the minimax rule is meant to be applicable in more general mismatch situations where the mismatch characteristics are unknown. It is interesting to note, however, that as explained in Section III, the gain-adapted recognizer [29] combines a noise-robust approach with the minimax methodology and hence, outperforms both systems for mismatch situations caused by additive noise. This supports our belief that the minimax approach has even higher potential when integrated with existing schemes.

Another situation where mismatch might be expected, occurs when the training data and the testing data are from different databases. In our second experiment, we examined

TABLE II
RECOGNITION ACCURACY AS A FUNCTION OF ρ AND C AT SNR = 10 dB

	C							
ρ	1	2	3	4	5	6	7	8
0.1	81.25	81.75	82.50	82.75	83.50	83.50	83.50	84.50
0.2	81.75	83.00	83.50	84.25	86.00	87.00	88.50	89.75
0.3	82.25	83.50	84.75	86.50	88.75	90.00	90.75	91.25
0.4	83.00	84.25	86.25	89.25	91.00	91.50	92.00	92.25
0.5	83.25	85.50	88.75	90.50	91.75	92.25	92.50	92.25
0.6	83.75	86.75	89.75	91.75	92.50	93.00	92.75	92.50
0.7	83.50	88.50	91.00	92.50	93.25	93.00	92.75	91.75
0.8	84.25	89.75	92.00	93.50	92.25	92.00	91.25	90.75
0.9	85.50	90.25	92.25	92.25	91.50	91.00	88.50	88.25
1.0	86.75	90.50	92.00	89.00	85.75	81.00	76.75	72.50

TABLE III
RECOGNITION ACCURACY FOR TWO DIFFERENT DATABASES

	C					
	A-Training, B-Testing			B-Training, A-Testing		
ρ	1	2	3	1	2	3
0.0	88.90	88.90	88.90	97.00	97.00	97.00
0.1	89.40	89.70	89.90	97.50	97.75	98.00
0.2	89.80	89.80	90.40	97.75	98.00	98.25
0.3	90.00	90.50	91.20	98.00	98.25	98.00
0.4	90.00	91.50	91.30	98.00	98.00	98.00
0.5	90.30	91.20	90.80	98.25	98.50	98.50
0.6	91.20	91.20	90.40	98.50	98.50	98.50
0.7	91.30	91.10	90.70	98.50	98.50	98.75
0.8	91.70	91.30	90.70	98.50	98.75	98.00
0.9	91.70	91.80	91.00	98.50	98.25	98.00
1.0	92.70	92.20	89.70	99.00	98.00	97.00

the minimax decision rule and compared to the standard approach in recognition of the English digits, using one database for training and another database, recorded under completely different conditions, for testing. Database A was the same as in the first experiment described above. Database B included one utterance of each digit from 100 different speakers, 50 males and 50 females, recorded from a telephone line and sampled at 6.667 kHz. The design parameters for both databases in this case were $T = 300, D = 100, q = 12, S = 8, \alpha = 0.95$, and $\beta = 0.54$. Table III presents the performance of the minimax rule as a function of ρ and C, when the training part of database A serves for training and the testing part of database B is used for testing (left part), and vice versa (right part).

Note that in both situations the highest improvement introduced by the minimax approach is attained for $\rho = C = 1$. When A is the training database and B is the testing database, the highest recognition accuracy of the minimax rule is 92.7% while the standard scheme ($\rho = 0$) attains 88.9%. When the roles of the databases are interchanged, the maximum improvement is from 97% to 99%. The ranges of recognition scores in these two situations is different, apparently because database B is larger, and hence when serves for training it yields more reliable parameter estimates $\{\lambda_i\}$. Again, observe that the sensitivity of the error rate to ρ and C is fairly weak and for most choices of these parameters some degree of improvement is obtained. The confidence intervals of the scores ranged between 1.10 to 1.50 in the left part and between 1.75 to 2.50 in the right part. The reason for smaller confidence intervals in the former case is that the testing database is relatively large.

An attempt has been made to examine the minimax scheme in recognition of spoken versions of the English E-set letters, /b/,/c/,/d/,/e/,/g/,/p/,/t/,/v/, and /z/, using an experiment similar to the latter. Unfortunately, in this case, no significant improvement over the standard scheme has been observed. A possible explanation for this is that since the E-set words are highly confusable and their discrimination is weak even without mismatch, the associated parameter vectors $\{\lambda_i\}$ lie very close to each other. Consequently, even for very small values of ρ and C some of the neighborhoods Λ_i may overlap, in which case the minimax rule breaks down. In view of this, it is recommended to apply the minimax scheme in situations of reasonably strong discrimination among the nominal sources.

VI. CONCLUSION

A parametric minimax approach for robust speech recognition has been developed which attains the best asymptotic performance for the worst case mismatch between the training and testing conditions. An attempt has been made to cover a class of mismatch situations as wide and general as possible. Experimentally, the proposed scheme introduces considerable reduction of the error rate over the standard nonrobust scheme. The algorithm is relatively simple to implement, it does not involve any modifications in the training procedure, and its performance is quite insensitive to design parameters that characterize the shape and the size of the uncertainty neighborhoods. It can be also easily integrated with many existing parametric isolated-word speech recognition schemes. A limitation of this approach appears to be its inability to reduce significantly the error rate, when the discrimination among the nominal sources is weak.

A possible direction of further research is improving the topology of the mismatch neighborhoods, which appears to be crucial for successful operation of the parametric minimax approach proposed here. In particular, a comprehensive comparative study of various shapes and topologies of the

mismatch neighborhood model might be interesting. Another interesting direction is to include the cepstral variances in the adaptation due to mismatch. This will make the scheme slightly more complicated but it is useful to examine whether it buys any significant improvement in performance. Finally, it should be examined whether performance can be gained by a developing a robust version of the training algorithm. These issues are currently under investigation.

ACKNOWLEDGMENT

The authors are grateful to Yariv Ephraim for providing us recognition scores from his gain adapted speech recognition system. Useful comments made by the anonymous referees are greatly appreciated.

REFERENCES

[1] G. A. Neben, R. J. McAulay, and C. J. Weinstein, "Experiments in isolated word recognition using noisy speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1156–1159, Apr. 1983.

[2] A. Nadas, D. Nahamoo, and M. A. Piceny, "Speech recognition using noise-adaptive prototypes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 517–520, Apr. 1988.

[3] D. Van Compernolle, "Increased noise immunity in large vocabulary speech recognition with the aid of spectral subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1143–1146, Apr. 1987.

[4] D. Mansour and B.-H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1659–1671, Nov. 1989.

[5] B.-H. Juang, "Recent developments in speech recognition under adverse conditions," in *Proc. Int. Conf. on Spoken Language Processing*, pp. 1113–1116, Nov. 1990.

[6] B.-H. Juang and R. L. Rabiner, "Signal restoration by spectral mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 2368–2371, Apr. 1987.

[7] D. Mansour and B.-H. Juang, "The short-time modified coherence representation and its application for noisy speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 795–804, June 1989.

[8] Y. Ephraim, J. G. Wilpon, and R. L. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1324–1327, Apr. 1987.

[9] D. Van Compernolle, "Spectral estimation using a log-distance error criterion applied to speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 258–261, May 1989.

[10] ——, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language*, vol. 3, no. 2, pp. 151–167, Apr. 1989.

[11] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 18A2.1–18A2.4, May 1984.

[12] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 849–852, Apr. 1990.

[13] A. Erell and M. Weintraub, "Estimation using log spectral distance criterion and Markov models for recognition of noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, to be published.

[14] ——, "Estimation using log spectral distance criterion for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 853–855, Apr. 1990.

[15] J. N. Holmes and N. C. Sedgwick, "Noise compensation for speech recognition using probabilistic models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 741–744, Apr. 1986.

[16] A. Noll, H. H. Hamer, H. Piotrowski, H. W. Ruehl, S. Dobler, and J. Weith, "Real-time connected-word recognition in a noisy environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 679–682, May 1989.

[17] B. H. Hanson and W. H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 968–973, 1987.

[18] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and Language*, vol. 1, pp. 109–130, 1986.

[19] ——, "Robustness against noise: The role of timing synchrony measurement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 2372–2375, May 1987.

[20] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 561–564, Apr. 1988.

[21] ——, "Stress compensation and noise reduction algorithms for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 266-269, May 1989.

[22] H. Gish, Y.-L. Chow, and R. Rohlicek, "Probabilistic vector mapping of noisy speech parameters for HMM word spotting," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 117–120, Apr. 1990.

[23] B. A. Hanson and T. H. Applebaum, "Robust speaker independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 857–860, Apr. 1990.

[24] H. Matsumoto and H. Imai, "Comparative study of various spectrum matching measures on noise robustness," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 769–772, Apr. 1986.

[25] F. Soong and M. M. Sondhi, "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 795–804, June 1989.

[26] D. B. Roe, "Speech recognition with a noise-adapting codebook," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1139–1142, Apr. 1987.

[27] A. P. Varga, R. K. Moore, J. Bridle, K. Ponting, and M. Russel, "Noise compensation algorithms for use with hidden Markov model based speech recognition," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 481–484, Apr. 1988.

[28] C.-H. Lee and K. Ganesan, "Speech recognition under additive noise," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 35.7.1–35.7.4, Mar. 1984.

[29] Y. Ephraim, "Gain adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 40, June 1992. (Also summarized in *IEEE Proc. Int. Symp. Inform. Theory*, p. 123, Jan. 1990.)

[30] A. Nadas, D. Nahamoo, and M. Picheny, "Adaptive labeling: Normalization of speech by adaptive transformations based on vector quantization," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 521–524, Apr. 1988.

[31] M. Hunt, "Speaker dependent and independent speech recognition experiments with an auditory model," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 215–218, Apr. 1988.

[32] W. Byrne, J. Robinson, and S. Shamma, "The auditory processing and recognition of speech," in *Proc. DARPA Workshop on Speech and Natural Language*, 1989.

[33] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 947–953, July 1987.

[34] Y. Chen, "Cepstral domain stress compensation for robust speech recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 717–720, Apr. 1987.

[35] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 713–716, Apr. 1987.

[36] R. P. Lippman, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 705–708, Apr. 1987.

[37] P. J. Huber, *Robust Statistical Procedures*. Society for Ind. and Appl. Math., no. 27, 1977.

[38] ——, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, no. 4, pp. 1753–1758, 1965.

[39] ——, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73–101, 1964.

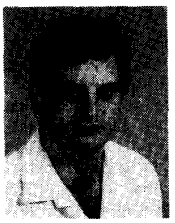
[40] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proc. IEEE*, vol. 73, no. 3, pp. 433-481, 1985.

[41] H. Van Trees, *Detection, Estimation, and Modulation Theory. part 1*. New York: Wiley, 1968, pp. 86–96.

[42] N. Merhav and C.-H. Lee, "On the asymptotic statistical behavior of empirical cepstral coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, to be published.

[43] R. S. Ellis, *Entropy, Large Deviations and Statistical Mechanics*. New York: Springer-Verlag, 1985.

[44] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 2157–2166, Oct. 1991.
[45] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–401, 1965.
[46] R. R. Bahadur, "Rates of convergence of estimates and test statistics," *Ann. Math. Statist.*, vol. 38, pp. 303– 324, 1967.
[47] A. D. M. Kester and W. C. M. Kallenberg, "Large deviations of estimators," *Ann. Statist.*, vol. 14, no. 2, pp. 648–664, 1986.
[48] S. Natarajan, "Large deviations, hypothesis testing, and source coding for finite Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 360–365, May 1985.
[49] I. Csiszár, T. M. Cover, and B.-S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 788–801, May 1987.
[50] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.
[51] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 26–37, 1980.
[52] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inverse problems," in *Proc. Int. Symp. on Information Theory*, Jan. 1990, p. 112.
[53] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1001–1013, Sept. 1989.
[54] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, pp. 278–286, Mar. 1988.
[55] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, pp. 401–408, Mar. 1989.
[56] ——, "On tests for randomness, tests for independence, and universal data compression," submitted for publication.
[57] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1014–1019, Sept. 1989.
[58] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 1109–1114, Sept. 1989.
[59] T. W. Anderson, *An Introduction to the Statistical Analysis of Data*. New York: Wiley, 1971.
[60] S. Kay and J. Makoul, "On the statistics of the estimated reflection coefficients of an autoregressive process," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 1447–1455, Dec. 1983.
[61] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, June 1974.
[62] J. Makoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[63] R. M. Gray, "*Toeplitz and Circulant Matrices: II*," Tech. Rep. 6504-1, Information Systems Lab., Stanford Univ., CA, Apr. 1977.
[64] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Some properties of continuous hidden Markov model representations," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1251–1270, July–Aug. 1985.
[65] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.
[66] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.
[67] L. R. Rabiner and J. G. Wilpon, "Some performance benchmarks for isolated word speech recognition systems," *Computer Speech and Language*, vol. 2, pp. 343–357, 1987.
[68] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1846–1856, Dec. 1989.
[69] B.-H. Juang and L. R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1639–1641, Sept. 1990.
[70] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, 1970.
[71] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.
[72] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data," *J. Roy. Statist. Soc.*, vol. B-39, pp. 1–38, 1977.
[73] N. Merhav and Y. Ephraim, "Hidden Markov modeling using a dominant sequence of states with application to speech recognition," *Computer, Speech and Language*, vol. 5, pp. 327–339, Oct. 1991.



Neri Merhav (S'86–M'87) was born in Haifa, Israel, on Mar. 16, 1957. He received the B.Sc., M.Sc., and D.Sc. degrees from the Technion, Israel Institute of Technology, Haifa, Israel, in 1982, 1985 and 1988, respectively, all in electrical engineering.

From 1982 to 1985 he was a research associate with the Israel IBM Scientific Center in Haifa, where he developed algorithms for speech coding, speech synthesis and adaptive filtering of speech signals in array sensors. From 1988 to 1990 he was with the Speech Research Department of AT&T Bell Laboratories, Murray Hill, New Jersey, where he investigated and developed algorithms for speech recognition. He is currently with the Electrical Engineering Department of the Technion and his research interests are statistical signal processing, information theory and statistical communications.

Chin-Hui Lee (S'79–M'81) received the B.S. degree from National Taiwan University, Taipei, in 1973, the M.S. degree from Yale University, New Haven, in 1977 and the Ph.D. degree from University of Washington, Seattle, in 1981, all in electrical engineering.

In 1981, he joined Verbex Corporation, Bedford, MA, and was involved in research work on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, where he engaged in research work in speech coding, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with the AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include speech modeling, speech recognition, speaker recognition, and signal processing. Since 1991, he has been an associated editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.