Minimum Classification Error Rate Methods for Speech Recognition

Biing-Hwang Juang, *Fellow, IEEE*, Wu Chou, *Member, IEEE*, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*— A critical component in the pattern matching approach to speech recognition is the training algorithm, which aims at producing typical (reference) patterns or models for accurate pattern comparison. In this paper, we discuss the issue of speech recognizer training from a broad perspective with root in the classical Bayes decision theory. We differentiate the method of classifier design by way of distribution estimation and the discriminative method of minimizing classification error rate based on the fact that in many realistic applications, such as speech recognition, the real signal distribution form is rarely known precisely. We argue that traditional methods relying on distribution estimation are suboptimal when the assumed distribution form is not the true one, and that "optimality" in distribution estimation does not automatically translate into "optimality" in classifier design. We compare the two different methods in the context of hidden Markov modeling for speech recognition. We show the superiority of the minimum classification error (MCE) method over the distribution estimation method by providing the results of several key speech recognition experiments. In general, the MCE method provides a significant reduction of recognition error rate.**

I. Introduction

THE METHOD of hidden Markov modeling has become prevalent in speech recognition applications recently [1]. The hidden Markov model (HMM) method is statistically based, and its success has triggered a renewed urge for a better understanding of the traditional statistical pattern recognition approach to speech recognition problems. This paper is thus intended to provide a revisit to the statistical formulation of the recognition problem, take a critical view of the approach, and hopefully inspire other innovations that would potentially lead to better solutions in the context of automatic speech recognition.

The statistical formulation has its root in the classical Bayes decision theory, which links a classification/recognition task to the problem of distribution estimation. This statistical formulation is the basis of various pattern recognition techniques developed in the past several decades. However, if we carefully reexamine the fundamental assumptions and limitations of the approach, we can find that there exist differences between the problem of optimal distribution estimation and the problem of optimal recognizer design. This is, as will be elaborated, due to the facts that we lack complete knowledge of the form of the data distribution and that training data are always inadequate, particularly in dealing with speech problems. Understanding of these differences would give us a better perspective in answering the question of optimal speech recognizer design.

The performance of a recognizer is normally defined by its expected recognition error rate, and we define an optimal recognizer to be one that achieves the least expected recognition error rate. The difference between the statistical and the proposed minimum classification error (MCE) approach lies in the way the recognition error is expressed and in the computational steps that would lead to the minimization of such error functions. A key to the development of the MCE method is a new error function that incorporates the recognition operation and performance in a functional form, which can be directly evaluated and optimized.

This paper begins in the next section with a brief review of the Bayes decision theory and its application to the formulation of statistical pattern recognition problems. We then discuss in Section III key considerations in choosing the distribution form for speech signals. The discussion is intended to cast the problem of automatic speech recognition in the framework of statistical pattern recognition, unlike other approaches such as the acoustic-phonetic approach or the artificial intelligence approach [2]. Based on the empirical observation, we explain why the HMM is a natural, simple choice for a speech signal distribution. We then discuss the estimation problem in HMM in Section IV. We point out, however, that despite its prevalence, an HMM is not the *true* distribution form for speech signals and a new approach based on the concept of discrimination for speech recognizer design becomes appropriate. In Section V, we introduce an MCE training method that aims at minimizing either the empirical error rate or the expected error rate, given an arbitrary choice of the distribution (discriminant) function. We elaborate the implementation of the new training method, again for the particular case of a hidden Markov model. We report several experimental results comparing the traditional maximum likelihood (ML) method (based on the distribution estimation formulation) and the new MCE training method in Section VI. We summarize discussions finally in Section VII.

II. Bayes Decision Theory

Let X be a random observation from an information source, consisting of M classes of events. A classifier's job is to correctly classify each X into one of the M classes. We denote these classes by $C_i, i = 1, 2 \cdots M$. Let $P(X, C_i)$ be the joint probability distribution of X and C_i, a quantity that is assumed to be known to the designer of the classifier. In other

Manuscript received August 22, 1995; revised September 18, 1996. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

The authors are with the Speech Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: wuchou@research.bell-labs.com; bhj@research.bell-labs.com).

Publisher Item Identifier S 1063-6676(97)03593-7.

words, the designer has full knowledge of the random nature of the source. From the set of joint probability distributions, the marginal and the conditional probability distributions can be easily calculated.

To measure the performance of the classifier, we further define for every class pair (j,i) a cost or loss function e_{ji} that signifies the cost of classifying (or recognizing) a class i observation into a class j event. The loss function is generally nonnegative with $e_{ii}=0$ representing correct classification.

Given an arbitrary observation X, a conditional loss for classifying X into a class i event can be defined as [3]

$$R(C_i|X)=\sum_{j=1}^{M} e_{ji}P(C_j|X) \tag{1}$$

where $P(C_j|X)$ is the *a posteriori* probability. This leads to a reasonable performance measure for the classifier, i.e. the expected loss, defined as

$$\mathcal{L}=\int R(C(X)|X)p(X)\,dX \tag{2}$$

where $C(X)$ represents the classifier's decision, assuming one of the M "values," $C_1,C_2\cdots C_M$ based on a random observation X drawn from a probability distribution $P(X)$. The decision function, $C(X)$, depends on the classifier design. Obviously, if the classifier is so designed that for every X

$$R(C(X)|X)=\min_i R(C_i|X) \tag{3}$$

the expected loss in (2) will be minimized.

For speech recognition, the loss function e_{ij} is usually chosen to be the zero-one loss function defined by

$$e_{ij}=\begin{cases}0, & i=j\\ 1, & i\neq j\end{cases}\qquad i,j=1,2\cdots M \tag{4}$$

which assigns no loss to correct classification and a unit loss to any error, regardless of the class. With this type of loss function, the expected loss $\mathcal{L}$ is thus the error probability of classification or recognition. The conditional loss becomes

$$\begin{aligned}R(C_i|X)&=\sum_{i\neq j}P(C_j|X)\\ &=1-P(C_i|X).\end{aligned} \tag{5}$$

The optimal classifier that achieves minimum $\mathcal{L}$ is thus the one that implements the following:

$$C(X)=C_i \quad \text{if} \quad P(C_i|X)=\max_j P(C_j|X). \tag{6}$$

In other words, for minimum error rate classification, the classifier employs the decision rule of (6), which is called the *maximum a posteriori* (MAP) decision. The minimum error rate achieved by MAP decision is called *Bayes risk*.

The required knowledge for an optimal classification decision is, thus, the *a posteriori* probabilities for the implementation of the MAP rule. These probabilities, however, are not given in practice and have to be estimated from a training data set with known class labels. The Bayes decision theory thus effectively transforms the classifier design problem into a distribution estimation problem. This is the basis of the statistical approach to pattern recognition, which can be stated: Given (or collect) a set of training data (observations) $\{X_1,X_2\cdots X_I\}$ with known class labels, estimate the a posteriori probabilities $P(C_i|X), i=1,2\cdots M$ for any X to implement the maximum a posteriori decision for minimum Bayes risk. The *a posteriori* probability $P(C_i|X)$ can be rewritten as

$$P(C_i|X)=P(X|C_i)P(C_i)/P(X). \tag{7}$$

Since $P(X)$ is not a function of the class index and, thus, has no effect in the MAP decision, the needed probabilistic knowledge can be represented by the class prior $P(C_i)$ and the conditional probability $P(X|C_i)$. For the simple case of isolated word speech recognition, the observations are the word utterances and the class labels are the word identities. The class prior $P(C_i)$ often appears as part of the language model [4] and in our present discussion is assumed to be uniform without loss of generality.

There are several issues associated with this classical approach. First, the distributions usually have to be parameterized in order for them to be practically useful for the implementation of the MAP rule of (6). The classifier designer therefore has to determine the right parametric form of the distributions. For most of the real-world problems, this is a difficult task. Our choice of the distribution form is often limited by the mathematical tractability of the particular distribution functions and is very likely to be inconsistent with the actual distribution. This means the true MAP decision can rarely be implemented, and the minimum Bayes risk generally remains an unachievable lower bound. Second, given a parameterized distribution form, the unknown parameters defining the distribution have to be estimated from the training data. A good parameter estimation method is therefore necessary. The estimation method has to be able to produce consistent parameter values. Third, the approach requires a training data set of known examples. In order to have a reliable parameter estimate, the training data set needs to be of sufficient size. Usually, the more the training data that is provided, the better the parameter estimate is. The difficulty, nevertheless, is that data collection and labeling is a labor-intensive and resource-demanding process, particularly for speech recognition applications. When the amount of the training data is limited, the quality of the distribution parameter estimates can not be guaranteed. These three basic issues point out a fundamental fact in the statistical pattern recognition approach; that is, despite the conceptual optimality of the Bayes decision theory and its applications to pattern recognition, it cannot be accomplished because practical "MAP" decisions in speech recognition are not true MAP decisions. This understanding is critical in our discussions below.

III. Probability Distributions for Speech

The statistical method, as discussed in the previous section, requires that a proper, usually parametric, distribution form for the observations be chosen in order to implement the MAP decision. Using the task of isolated word speech recognition as an example, we have to determine the distribution form for the speech utterances of each word before we apply an estimation method to find the values of the parameters. What is the right distribution form for speech utterances? This question involves

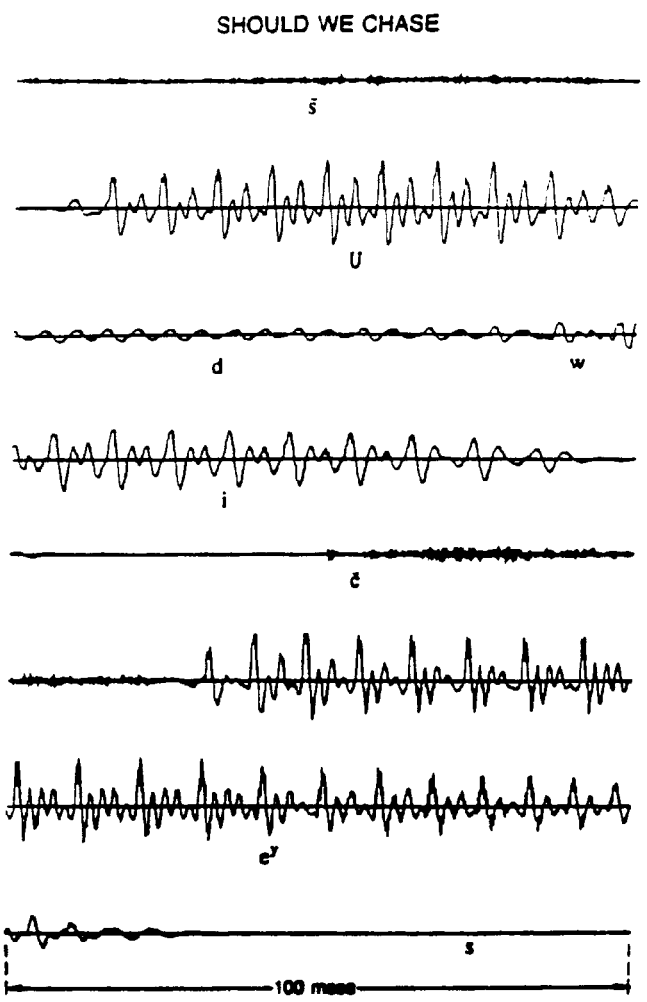

Fig. 1. Speech waveform and a segmentation and labeling of the constituent sounds of the phrase "Should we chase?"

two essential aspects: i) finding the speech dimensions that carry the most pertinent linguistic information, and ii) deciding how to statistically characterize the information along the chosen dimensions. We discuss these issues in this section.

A. Speech Characteristics

Speech is a time-varying signal. When we speak, our articulatory apparatus (the lips, jaw, tongue, and velum) modulates air pressure and flow to produce an audible sequence of sounds. Although the spectral content of any particular sound in speech may include frequencies up to several thousand hertz, our articulatory configuration (vocal tract shape, tongue movement, etc.) often does not undergo dramatic changes more than ten times per second. During the short interval where the articulatory configuration stays somewhat constant, a region of "quasistationarity" in the produced speech signal can often be observed. This is the first characteristic of speech that distinguishes it from other random, nonstationary signals. As an example, Fig. 1 shows the waveform of the speech segment "should we chase?" with the corresponding phoneme labels.

Furthermore, speech is not a memoryless process due to articulatory and phonotactic constraints. According to the phonological rule of a language, there is a certain dependency between sound pairs that occur in sequence; some occur more often than others, while some are simply nonexistent in the language. The speech model or distribution needs to have provisions to permit characterization of this sequential structure, ideally in a manner consistent with the slowly varying nature ("quasistationarity") of the speech signal. The HMM provides a simple means to characterize speech signals according to the above discussion.

B. Hidden Markov Model

Let X be a speech utterance, $X = (\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_T)$ where $\boldsymbol{x}_t$ denotes a short-time vector measurement. It has been found that short-time cepstral analysis [5] produces effective speech observations, in the form of low-frequency (10–16) cepstral coefficients, for recognition purposes. Thus, $\boldsymbol{x}_t$ conventionally is a cepstral vector.

Further consider a first-order N-state Markov chain governed by a state transition probability matrix $A = [a_{ij}]$, where a_{ij} is the probability of making a transition from state i to state j. Assume that at $t = 0$ the state of the system q_0 is specified by an initial state probability $\pi_i = P(q_0 = i)$. Then, for any state sequence $\boldsymbol{q} = (q_0, q_1 \cdots q_T)$, the probability of $\boldsymbol{q}$ being generated by the Markov chain is

$$P(\boldsymbol{q}|A, \pi) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}. \tag{8}$$

Suppose the system, when at state q_t, puts out an observation $\boldsymbol{x}_t$ according to a distribution $b_{q_t}(\boldsymbol{x}_t) = P(\boldsymbol{x}_t|q_t), q_t = 1, 2 \cdots N$. The HMM used as a distribution for the speech utterance X is then defined as

$$\begin{aligned} P(X|\pi, A, \{b_j\}_{j=1}^N) = P(X|\lambda) &= \sum_{\boldsymbol{q}} P(X, \boldsymbol{q}|\lambda) \\ &= \sum_{\boldsymbol{q}} P(X|\boldsymbol{q}, \lambda) P(\boldsymbol{q}|\lambda) \\ &= \sum_{\boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(\boldsymbol{x}_t) \end{aligned} \tag{9}$$

where $\lambda = (\pi, A, \{b_j\}_{j=1}^N)$ is the parameter set for the model.

As can be seen in (9), $\{b_{q_t}\}$ defines the distribution for short-time observations and A characterizes the behavior and interrelationship between different states of the speech-generation process. In other words, the structure of an HMM provides a reasonable means for characterizing the distribution of a speech signal. Normally, N, the total number of states, is much smaller than T, the time duration of the speech utterance. The state sequence $\boldsymbol{q}$ displays a certain degree of stability among adjacent q_t's due to the above-mentioned "quasistationarity." The use of the HMM as speech distributions has been shown to be practically effective.

Two points deserve further attention. First, in the above, the choice of state observation distributions $b_{q_t}(\boldsymbol{x}_t)$ is not specified. Different choices of speech dimensions for the observation space may require different forms of the state observation distribution. For cepstral vectors, a mixture Gaussian density is commonly employed. Second, regardless of the practical effectiveness of the HMM in speech recognition, it should not be taken as the true distribution form of speech and, therefore, any recognition system or decision rule that operates based on the HMM is not going to achieve the minimum error

rate as implied in the true MAP decision. We shall come back to this later in Section V.

IV. Hidden Markov Modeling Framework

The statistical method of hidden Markov modeling for speech recognition encompasses several interesting problems, namely, the evaluation problem, the decoding problem, and the estimation problem [1], [9]. In this paper, we discuss only the estimation problem, in light of the above discussions of Bayes decision theory approach, which transforms the recognizer design problem into a distribution estimation problem.

Given an observation sequence (or a set of sequences) X, the estimation problem involves finding the "right" model parameter values that specify a source model (distribution) most likely to produce the given sequence of observations. In solving the estimation problem, we usually use the method of ML; that is, we choose λ such that $P(X|\lambda)$ as defined in (9) is maximized for the given "training" sequence X. Note that in most simple cases, X is a speech utterance of a known word identity. The estimated model parameter set λ is then associated with each individual word class. For an M-word vocabulary, M such parameter sets are to be estimated for use in the recognizer.

The Baum–Welch algorithm [10] accomplishes likelihood maximization in a two-step procedure, known as "reestimation." Based on an existing model λ', the first step of the algorithm transforms the objective function $P(X|\lambda)$ into a new function $Q(\lambda', \lambda)$ that essentially measures a divergence between the initial model λ' and an updated model λ. The Q function is defined, for the simplest case, as

$$Q(\lambda', \lambda) = \sum_{\boldsymbol{q}} P(X, \boldsymbol{q}|\lambda') \log P(X, \boldsymbol{q}|\lambda) \tag{10}$$

where $P(X, \boldsymbol{q}|\lambda)$ can be found in (9). It can be shown that $Q(\lambda', \lambda) \geq Q(\lambda', \lambda')$ implies $P(X|\lambda) \geq P(X|\lambda')$. Therefore, the second step of the algorithm involves maximizing $Q(\lambda', \lambda)$ as a function of λ to obtain a higher, improved likelihood. These two steps iterate interleavingly until the likelihood reaches a fixed point.

The ML method is, however, not the only possible choice for solving the estimation problem. An in-depth discussion of various estimation criteria can be found in [9]. It should be pointed out that the ML method does not necessarily lead to a minimum error rate performance for the recognizer. As discussed above, this is due to i) the likely mismatch between the chosen distribution form (HMM in the present case), and the actual speech data distribution that is typically not available; and ii) the finite training (known) data set that is often inadequate.

V. MCE Training

As discussed, classifier design by distribution estimation often does not lead to an optimal performance. The problem is that in most situations, the estimated probabilities deviate from the true probabilities and the exact MAP rule cannot be implemented. In addition, when the assumed form of the distributions is different from the true one, the optimality of the estimated distribution has little to do with the optimality of the classifier, particularly in terms of recognition error rate. This leads to various criteria for estimating the classifier parameters in classifier design. In particular, criteria of maximum mutual information (MMI) and minimum discriminative information (MDI) are used in many applications [6]–[8]. Although these methods demonstrate significant performance advantages over the traditional ML approach, they are not based on a direct minimization of a loss function, which links to the classification error rate. Over the past few years, an attempt has developed to overcome the fundamental limitations of the traditional approach and to directly formulate the classifier design problem as a classification error rate minimization problem. This approach is called the *minimum classification error* (MCE) method, in which the goal of MCE training is to be able to correctly discriminate the observations for best recognition/classification results rather than to fit the distributions to the data.

Consider a set of *class conditional likelihood functions* $g_i(X;\Lambda), i = 1, 2 \cdots M$ defined by the parameter set Λ. In its simplest form for our present discussion of the HMM techniques, $g_i(X;\Lambda)$ can take essentially the same form as (9), i.e.

$$g_i(X;\Lambda) = P(X|\lambda^{(i)}) = P(X|\pi^{(i)}, A^{(i)}, \{b_j^{(i)}\}_{j=1}^{N}) \tag{11}$$

where the superscript i denotes the parameter set identity associated with word (class) i in the vocabulary. The entire parameter set of the classifier Λ is thus $\Lambda = \{\lambda^{(i)}\}_{i=1}^{M}$[1]. The choice of HMM of (9) is a reasonable one as discussed in Section III. The classifier/recognizer is operating under the following *decision rule:*

$$C(X) = C_i \quad \text{if} \quad g_i(X;\Lambda) = \max_j g_j(X;\Lambda). \tag{12}$$

The goal of classifier design is again to achieve the minimum error probability based on the loss function defined in (4).

The difficulty associated with the MCE training approach lies in the derivation of an objective function that has to be consistent with the performance measure (i.e., the error rate) and also suitable for optimization. The error rate based on a finite data set is a piecewise constant function of the classifier parameter Λ and, thus, a poor candidate for optimization by a simple numerical search method. We propose the following *embedded smoothing* for a loss function, which is a reasonable estimate of the error probability.

A. Optimization Criterion

The smoothed optimization criterion is a function of the *class conditional likelihood functions* $g_i(X, \Lambda), i = 1, 2 \cdots M$. Again, the classifier makes its decision for each input X by choosing the largest of the class conditional likelihood function evaluated on X. The key to the new error criterion is to express the operational decision rule of (12) in a functional form. There exist in this regard many possibilities, one of which is a *class misclassification measure* taking the following form:

$$d_i(X) = -g_i(X;\Lambda) + \log \left[\frac{1}{M-1} \sum_{j, j \neq i} \exp[g_j(X;\Lambda)\eta] \right]^{1/\eta} \tag{13}$$

[1] Note that $g_i(X, \Lambda)$ can be other reasonable functions, which are consistent with the error rate minimization.

where η is a positive number [11], [12]. This misclassification measure is a continuous function of the classifier parameters Λ and attempts to emulate the decision rule. For an ith class utterance $X, d_i(X) > 0$ implies misclassification and $d_i(X) \leq 0$ means correct decision. When η approaches ∞, the term in the bracket becomes $\max_{j,j\neq i} g_j(X;\Lambda)$. By varying the value of η and M, one can take all the competing classes into consideration, according to the individual significance, when searching for the classifier parameter Λ. In HMM-based speech recognition systems, each input feature vector $X(t)$ in a speech utterance X $(X = X(1), \cdots X(t_l))$ is often assigned to certain state of some HMM according to the best hidden-state sequence obtained through a Viterbi alignment process.

To complete the definition of the objective criterion, the misclassification measure of (13) is embedded in a smoothed zero-one function, for which any member of the sigmoid function family is an obvious candidate. A general form of the *loss function* can then be defined as

$$\ell_i(X;\Lambda) = \ell(d_i(X)) \tag{14}$$

where ℓ is a sigmoid function, one example of which is

$$\ell(d) = \frac{1}{1+\exp(-\gamma d+\theta)} \tag{15}$$

with θ normally set to zero and γ set to $\geq$ one. Clearly, when $d_i(X)$ is much smaller than zero, which implies correct classification, virtually no loss is incurred. When $d_i(X)$ is positive, it leads to a penalty which becomes essentially a classification/recognition error count. Finally, for any unknown X, the classifier performance is measured by

$$\ell(X;\Lambda) = \sum_{i=1}^{M} \ell_i(X;\Lambda) 1(X \in C_i) \tag{16}$$

where $1(\cdot)$ is the indicator function.

This three-step definition emulates the classification operation as well as the performance evaluation in a smooth functional form, suitable for classifier parameter optimization. Based on the criterion of (16), we can choose to minimize one of two quantities for the classifier parameter search; one is the expected loss and the other the empirical loss.

B. Optimization Methods

The purpose of the training process in the MCE approach is to find a set of parameters Λ so that a prescribed loss is minimized. As mentioned previously, the two kinds of loss we focus on are the expected loss and the empirical loss.

1) Expected Loss: For a classification problem involving M different classes, the expected loss is defined as

$$L(\Lambda) = E_X\{\ell(X;\Lambda)\} = \sum_{i=1}^{M} \int_{X\in C_i} \ell_i(X;\Lambda) p(X)\, dX. \tag{17}$$

Various minimization algorithms can be used to minimize the expected loss. The generalized probabilistic descent (GPD) algorithm is a powerful algorithm that can be used to accomplish this task [11]. In the GPD-based minimization algorithm, the target function $L(\Lambda)$ is minimized according to an iterative procedure. The following generalized probabilistic descent theorem [17] establishes the algorithmic convergence property of the algorithm.

Theorem 1: Suppose the following conditions are satisfied:[2]

$C1$: $\sum_{t=1}^{\infty} \epsilon_t = \infty, \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty \quad \epsilon_t \geq 0;$

$C2$: $\exists 0 \leq V < \infty,$ such that for all t,
$R_t(\epsilon_t, \theta_t) = \langle \nabla \ell(X,\Lambda_n), H(X,\Lambda_n + \epsilon_n\theta_n \nabla \ell(X,\Lambda_n)) \nabla \ell(X,\Lambda_n)\rangle \leq V,$
where H is the Hessian matrix of second-order partial derivatives;

$C3$: $\Lambda^* = \arg\min_{\Lambda} E_X \ell(X,\Lambda)$
is the unique Λ such that
$\nabla L(\Lambda)|_{\Lambda=\Lambda^*} = \nabla E_X \ell(X,\Lambda)|_{\Lambda=\Lambda^*} = 0.$

Then, Λ_t given by

$$\Lambda_{t+1} = \Lambda_t - \epsilon_t \nabla \ell(X_t,\Lambda)|_{\Lambda=\Lambda_t} \tag{18}$$

will converge to Λ^* almost surely (i.e., with probability one).

Condition $C3$ can be considerably weakened. Even without condition $C3$ the following is still true:

$$E_X \nabla \ell(X,\Lambda_{t_k}) \to 0 \tag{19}$$

where Λ_{t_k} is a subsequence of Λ_t. In this case, Λ_{t_k} will converge to a local minimum point Λ^* where $\nabla L(\Lambda)|_{\Lambda=\Lambda^*} = 0$.

The algorithm defined by equation (18) can also be generalized to the following form:

$$\Lambda_{t+1} = \Lambda_t - \epsilon_t U_t \nabla \ell(X_t,\Lambda)|_{\Lambda=\Lambda_t} \tag{20}$$

where U_t is a positive definite matrix [17].

Other theoretical properties of the GPD algorithm have been studied in the literature, often under the name of stochastic approximation [20]–[22]. However, in order to apply this algorithm to speech recognition, such as a speech recognition system using HMM's, the GPD algorithm has to accommodate various constraints imposed on the HMM structures. In particular, the GPD algorithm is an unconstrained minimization scheme that needs modification for solving minimization problems with constraints. As will be shown shortly, one can utilize parameter space transformation to resolve this issue. In this method, the original parameters are updated through the inverse transform from the transformed parameter space to the original parameter space. This is done in such a way that constraints on the original parameters are always maintained. More detailed illustrations of this approach are given in later sections.

It should be noted that the underlying probability distributions involved in minimizing (17) are often unknown to the designer. One of the advantages of a GPD-based minimization algorithm is that it does not make any explicit assumption on these unknown probabilities. This feature is important for recognition and adaptive learning problems.

[2] The proof of this theorem is based on the Martingale convergence theorem, which is out of the scope of this paper.

2) Empirical Loss: For a given training data set consisting of I samples $\{X_1, \cdots X_I\}$, the empirical probability measure P_I defined on the training data set is a discrete probability measure that assigns equal mass at each sample. The empirical loss, on the other hand, is thus expressed as

$$L_0(\Lambda) = \frac{1}{I} \sum_{j=1}^{I} \sum_{i=1}^{M} \ell_i(X_j;\Lambda) 1(X_j \in C_i) = \int \ell(X;\Lambda)\, dP_I \tag{21}$$

where j denotes the index of the training utterance X_j in the training set of size I, and P_I is the empirical measure defined on the training set. If the training samples are obtained by an independent sampling from a space with a fixed probability distribution P, the empirical probability distribution P_I will converge to P in distribution as $I \to \infty$. In other words, for any measurable function f

$$\int f\, dP_I \to \int f\, dP. \tag{22}$$

The empirical loss defined on the I independent training samples will converge to the expected loss, as the sample size I increases. With sufficient training samples, the empirical loss is an estimate of the expected loss. The goodness of this estimate is determined by the training sample size I and the convergence rate of the empirical probability measure P_I to the limit distribution P. Various upper bounds on the convergence rate of the empirical probability measure can be found in [18].

C. HMM as a Discriminant Function

As we argued previously, an HMM is a reasonable model/distribution form for speech observations, although we cannot explicitly prove that it is the true distribution form for speech. In this case, the MCE method is particularly appropriate for the training of the model parameters.

Following (9), we have several ways of using an HMM as the discriminant function. A basic component in (9) is the joint observation-state probability

$$P^{(i)}(X, \boldsymbol{q};\Lambda) = \pi^{(i)}_{q_0} \prod_{t=1}^{T} a^{(i)}_{q_{t-1}q_t} b^{(i)}_{q_t}(x_t) \triangleq g_i(X, \boldsymbol{q};\Lambda) \tag{23}$$

which is now defined as a component function $g_i(X, \boldsymbol{q};\Lambda)$ for class i as well. The discriminant function for class i can take several possible forms, as follows, based on $g_i(X, \boldsymbol{q};\Lambda)$:

1) $g_i(X;\Lambda) = \sum_{\boldsymbol{q}} g_i(X, \boldsymbol{q};\Lambda).$ (24)

2) $g_i(X;\Lambda) = \max_{\boldsymbol{q}} g_i(X, \boldsymbol{q};\Lambda).$ (25)

3) $g_i(X;\Lambda) = \left\{ \frac{1}{Q} \sum_{\boldsymbol{q}}^{Q} g_i(X, \boldsymbol{q};\Lambda)^{\alpha} \right\}^{1/\alpha}$ (26)

where Q is the total number of possible state sequences and α is a positive number.

4) functions of the above.

Note that (24) is equivalent to the likelihood function, (25) is equivalent to the maximum joint observation-state probability, and (26) is a generalized mixture model which approaches (25) when $\alpha \to \infty$. We use the logarithm of (25) as an example in our derivation below. The algorithm based on (25) is often called *segmental GPD* [12].

We define, for

$$X = (\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_T) \quad \text{and} \quad \boldsymbol{x}_t = [x_{t1}, x_{t2} \cdots x_{tD}]'$$

with D being the dimension of $\boldsymbol{x}_t$

$$g_i(X;\Lambda) = \log\{\max_{\boldsymbol{q}} g_i(X, \boldsymbol{q};\Lambda)\} = \log\{g_i(X, \overline{\boldsymbol{q}};\Lambda)\} = \sum_{t=1}^{T} [\log a^{(i)}_{\overline{q}_{t-1}\overline{q}_t} + \log b^{(i)}_{\overline{q}_t}(\boldsymbol{x}_t)] + \log \pi^{(i)}_{\overline{q}_0} \tag{27}$$

where $\overline{\boldsymbol{q}} = (\overline{q}_0, \overline{q}_1 \cdots \overline{q}_T)$ is the optimal state sequence that achieves $\max_{\boldsymbol{q}} g_i(X, \boldsymbol{q};\Lambda)$. We also assume that

$$b^{(i)}_j(\boldsymbol{x}_t) = \sum_{k=1}^{K} c^{(i)}_{jk} \mathcal{N}[\boldsymbol{x}_t; \mu^{(i)}_{jk}, R^{(i)}_{jk}] \tag{28}$$

where $\mathcal{N}[\cdot]$ denotes a normal distribution, $c^{(i)}_{jk}$ is the mixture weights, $\mu^{(i)}_{jk} = [\mu_{jk\ell}]^D_{\ell=1}$ the mean vector and $R^{(i)}_{jk}$ the covariance matrix which, for simplicity, is assumed to be diagonal, i.e. $R^{(i)}_{jk} = [\sigma^2_{jk\ell}]^D_{\ell=1}$.

It is desirable to maintain the original constraints in HMM as probability measure, such as: i) the function being non-negative, ii) $\Sigma_j\ a_{ij} = 1$ for all i, 3) $\Sigma_k\ c_{jk} = 1$ for all j, etc. Also, we assume $\sigma_{jk\ell} > 0$. The following parameter transformations allow us to maintain the following constraints during parameter adaptation:

1) $a_{ij} \to \tilde{a}_{ij}$ where $a_{ij} = \dfrac{e^{\tilde{a}_{ij}}}{\sum_k e^{\tilde{a}_{ik}}}$ (29)

2) $c_{jk} \to \tilde{c}_{jk}$ where $c_{jk} = \dfrac{e^{\tilde{c}_{jk}}}{\sum_k e^{\tilde{c}_{ik}}}$ (30)

3) $\mu_{jk\ell} \to \tilde{\mu}_{jk\ell} = \dfrac{\mu_{jk\ell}}{\sigma_{jk\ell}}$ (31)

4) $\sigma_{jk\ell} \to \tilde{\sigma}_{jk\ell} = \log \sigma_{jk\ell}.$ (32)

It can be shown that for $X_n \in C_i$ in the training set, discriminative adjustment of the mean vector follows

$$\tilde{\mu}^{(i)}_{jk\ell}(n+1) = \tilde{\mu}^{(i)}_{jk\ell}(n) - \epsilon \frac{\partial \ell_i(X_n;\Lambda)}{\partial \tilde{\mu}^{(i)}_{jk\ell}}\bigg|_{\Lambda=\Lambda_n} \tag{33}$$

where

$$\frac{\partial \ell_i(X;\Lambda)}{\partial \tilde{\mu}^{(i)}_{jk\ell}} = \frac{\partial \ell_i}{\partial d_i} \frac{\partial d_i}{\partial \tilde{\mu}^{(i)}_{jk\ell}} \tag{34}$$

$$\frac{\partial \ell_i}{\partial d_i} = \gamma \ell_i(d_i)(1 - \ell_i(d_i)) \tag{35}$$

$$\frac{\partial d_i(X;\Lambda)}{\partial \tilde{\mu}^{(i)}_{jk\ell}} = -\sum_{t=1}^{T} \delta(\overline{q}_t - j) \frac{\partial \log b^{(i)}_j(\boldsymbol{x}_t)}{\partial \tilde{\mu}^{(i)}_{jk\ell}} \tag{36}$$

and

$$\frac{\partial}{\partial \tilde{\mu}_{jk\ell}^{(i)}} \log b_j^{(i)}(\boldsymbol{x}_t) = c_{jk}^{(i)}(2\pi)^{-d/2}|R_{jk}^{(i)}|^{-1/2}(b_j^{(i)}(\boldsymbol{x}_t))^{-1} \cdot \left(\frac{x_{t\ell}}{\sigma_{jk\ell}^{(i)}} - \tilde{\mu}_{jk\ell}^{(i)}\right) \exp\left\{-\frac{1}{2}\sum_{\ell=1}^{D}\left(\frac{x_{t\ell}}{\sigma_{jk\ell}^{(i)}} - \tilde{\mu}_{jk\ell}^{(i)}\right)^2\right\} \tag{37}$$

where γ is the center slope of the exponential sigmoid function for ℓ_i as defined in (15) and $\delta(\cdot)$ denotes the Kronecker delta function. Finally

$$\mu_{jk\ell}^{(i)}(n+1) = \sigma_{jk\ell(n)}^{(i)}\tilde{\mu}_{jk\ell}(n+1). \tag{38}$$

Similarly, for the variance $\sigma_{jk\ell}^{(i)}$

$$\tilde{\sigma}_{jk\ell}^{(i)}(n+1) = \tilde{\sigma}_{jk\ell}^{(i)}(n) - \epsilon \frac{\partial \ell_i(X_n;\Lambda)}{\partial \tilde{\sigma}_{jk\ell}^{(i)}}\bigg|_{\Lambda=\Lambda_n} \tag{39}$$

where

$$\frac{\partial \ell_i}{\partial \tilde{\sigma}_{jk\ell}^{(i)}} = -\gamma \ell_i(d_i)[1-\ell_i(d_i)] \cdot \sum_{t=1}^{T} \delta(\overline{q}_t - j)\, \frac{\partial \log b_j^{(i)}(\boldsymbol{x}_t)}{\partial \tilde{\sigma}_{jk\ell}^{(i)}} \tag{40}$$

$$\frac{\partial \log b_j^{(i)}(\boldsymbol{x}_t)}{\partial \tilde{\sigma}_{jk\ell}^{(i)}} = c_{jk}^{(i)}(2\pi)^{-d/2}|R_{jk}^{(i)}|^{-1/2} \cdot \exp\left\{-\frac{1}{2}\sum_{\ell=1}^{D}\left(\frac{x_{t\ell}-\mu_{jk\ell}}{\sigma_{jk\ell}}\right)^2\right\} \cdot \left[\left(\frac{x_{t\ell}-\mu_{jk\ell}}{\sigma_{jk\ell}}\right)^2 - 1\right] \cdot (b_j^{(i)}(\boldsymbol{x}_t))^{-1}. \tag{41}$$

Finally

$$\sigma_{jk\ell}^{(i)}(n+1) = \exp\{\tilde{\sigma}_{jk\ell}(n+1)\}. \tag{42}$$

Similar derivations for the transition probabilities and the mixture weights can be easily accomplished.

D. Embedded MCE Training of HMM

In the above development of the MCE training formalism, the utterance observation X is assumed to be one of the M classes. For recognition of continuous speech or for speech recognition using subword model units, what usually happens is that X is a concatenated string of observations belonging to different classes. For example, a sentence is a sequence of words, each of which is to be modeled by a distribution. In this situation, one possible training criterion is to minimize the string error rate while the string model is constructed from concatenating a set of word (or substring) models.

Let $W = (w_1, w_2 \cdots w_S)$ be a word sequence that constitutes a sentence. We define for an observation sequence X

$$g(X, W_r;\Lambda) = \log P(X, \boldsymbol{q}_{W_r}, W_r|\Lambda) \tag{43}$$

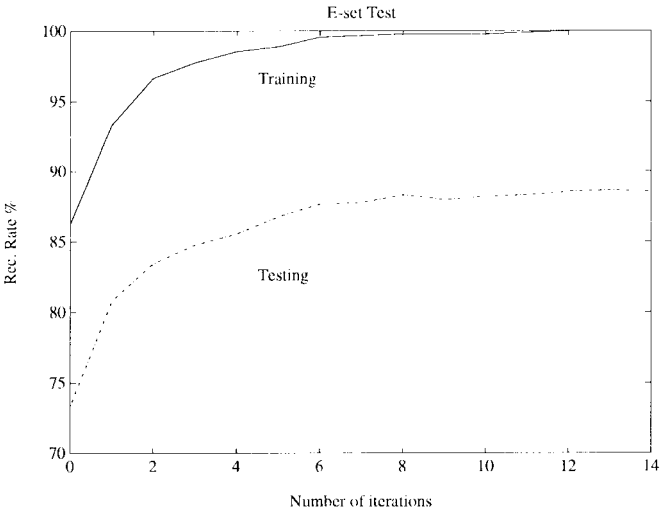

Fig. 2. Recognition curve of segmental GPD training.

where

$$W_r = \underset{W \neq W_1, \cdots, W_{r-1}}{\arg\max} P(X, \boldsymbol{q}_W, W|\Lambda)$$
$$= r\text{th best word sequence},$$
$$\boldsymbol{q}_{W_r} = \text{best state sequence corresponding to } W_r \tag{44}$$

and $P(X, \boldsymbol{q}_{W_r}, W_r|\Lambda)$ is the joint state-word sequence likelihood. Also, let W_0 be the known word sequence label for a training sentence X. Following the minimum error rate formulation, we define

$$d(X;\Lambda) = -g(X, W_0;\Lambda) + \log\left\{\frac{1}{r_m}\sum_{r=1}^{r_m}\exp[g(X, W_r;\Lambda)\cdot\eta]\right\}^{1/\eta} \tag{45}$$

where r_m is the total number of the competing word sequences, different from W_0, that will be taken into consideration in training. Again, the misclassification measure of (45) is embedded in a zero-one or sigmoid loss function to create a string error count. The rest of the procedure follows the above case for isolated utterances straightforwardly. For detailed discussion of this technique, consult [13].

VI. Experimental Results

For brevity, we cite one set of experimental results for the isolated class-utterance case [12] and another set for the connected word case [13].

The isolated class-utterance case involves recognition of utterances of the English E-set vocabulary, consisting of {b, c, d, e, g, p, t, v, z}. The data base was recorded over local dialed-up telephone lines from 100 American native speakers, 50 male and 50 female. Two utterances from each talker were recorded, one used for training and the other for testing. An HMM recognizer with ten-state, five-mixture/state models trained with the traditional ML method achieved an accuracy of 89% for the training data set and 76% for the test set.

Fig, 2 plots the recognition accuracy for both the training and the testing sets as a function of the number of iterations of the MCE training procedure. After ten iterations, the new recognizer achieved 99% accuracy for the training data set

TABLE I
PERFORMANCE COMPARISON IN CONNECTED DIGIT RECOGNITION

System	String Error Rate	# of String Errors	Error Reduction
Baseline	1.4%	120	N/A
Minimum String Error GPD	0.95%	82	31.6%

and 88% for the test set, representing a 50% reduction in recognition error.

For the connected word case, the experiment used TI connected-digit data base, which contains 8565 connected digit strings for training and 8578 strings for testing. The digit strings are of unknown length with a maximum of seven digits. The HMM recognizer configuration used ten-state, 64-mixture/state digit-based models. The MCE-based segmental GPD training method [13] was applied in the model training stage. Table I lists a comparison in string error performance for a baseline system trained with the conventional ML method and a new system trained with the segmental GPD method. The string error rate was reduced from 1.4% to 0.95%, representing a 32% reduction in recognition error.

The minimum-error-rate-based training approach was also applied to training context-dependent subword models in continuous speech recognition [13], [14]. For connected digit recognition on the TI data base, we used a set of acoustic-based context-dependent subword model units, in which 95% of the model units were interword context-dependent units. Each digit in the vocabulary was modeled by a context-dependent network with 12 fan-in heads and 12 fan-out tails. Fig. 3 illustrates the topology of each word model described by these acoustic-based context-dependent interword model units. The performance comparisons between different acoustic modeling approaches are given in Table II. Under the unknown length decoding condition, the model obtained from the minimum error rate training achieves a string error rate of 0.72% and a word error rate of 0.24% on the test set [14]. These are the best results reported so far for connected digit recognition on TI connected digit database. A similar approach was also applied to speaker recognition, details of which has been described in [24].

VII. SUMMARY

We have reexamined the classical Bayes decision theory approach to the problem of speech recognition and discussed the implied assumptions and issues that have been often left unresolved in the past. The classical decision-theoretic approach transforms the recognizer design problem into a problem in probability distribution estimation. The limitation of the approach, however, comes from the fact that the form of probability distributions for various dimensions of the speech signal is realistically unknown and virtually any assumed form will deviate from the true one and lead to suboptimal parameter estimates, thereby making the minimum error probability, as suggested by the Bayes approach, unattainable.

In light of this limitation, a new MCE approach based on learning for discrimination was discussed in this paper. The MCE approach to recognizer design aims at optimizing

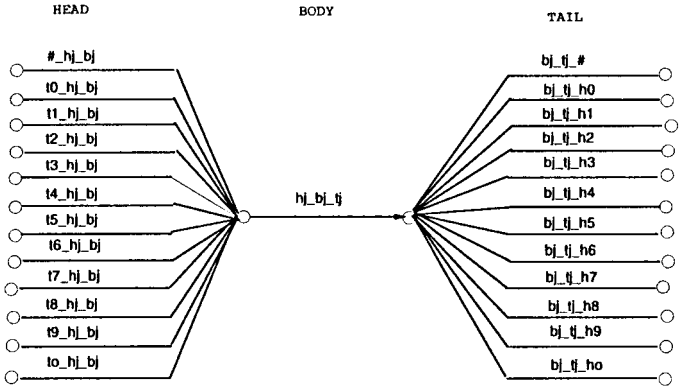

Fig. 3. Diagram of digit model with acoustic interword context-dependent model units.

TABLE II
PERFORMANCE COMPARISON OF ML AND MINIMUM ERROR RATE TRAINING

Training Method 8700 sents.	Word (Error_rate)	String (Error_rate)
Baseline (ML)	0.33%	0.97%
Str_Err_GPD	0.24%	0.72%
Err_Reduction	26%	25%

the recognizer parameters to minimize the error rate. We elaborate the issues and solutions associated with the new MCE approach in this paper in the context of HMM-based recognizer designs. The main issue or difficulty in this new approach concerns with means to formulate the error rate estimate as a smooth loss function for optimization. We show that a three-step smooth embedding leads to an error function, which is a close approximation to the error count and can be easily optimized. The development led to an algorithm called the generalized probabilistic descent (GPD) algorithm, an implementation of which, in terms of hidden Markov modeling, is discussed in detail in this paper. We further show that the new MCE approach indeed achieves better performance than the traditional probability distribution estimation approach in a number of speech recognition experiments. In general, the MCE method provides 30–50% reduction in error rate, compared to the traditional recognizer design.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.
[2] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[4] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE*, vol. 73, pp. 1616–1624, Nov. 1985.
[5] B.-H. Juang, L. R. Rabiner and J. G. Wilpon, "On the use of bandpass littering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 7, 947–954, 1987.
[6] L. R. Bahl, P. F. Brown, P. V. deSouza and R. L. Mercer, "Maximum mutual information estimation of HMM parameters for speech recognition," in *Proc. ICASSP-86*, pp. 49–52.
[7] B. Merialdo, "Phonetic recognition using hidden Markov models and maximum mutual information training," in *Proc. ICASSP-88*, pp. 111–114.
[8] Y. Normandin, "Optimal splitting of HMM Gaussian mixture components with MMIE training," in *Proc. ICASSP-95*, pp. 449–452.

[9] B.-H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, pp. 251–272, Aug. 1991.
[10] 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. Stat.*, vol. 41, pp. 164–171, 1970.
[11] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.
[12] W. Chou, C.-H. Lee and B.-H. Juang, "Segmental GPD training of an hidden Markov model based speech recognizer," *Proc. ICASSP-92*, pp. 473–476.
[13] ——, "Minimum error rate training based on N-best string models," in *IEEE Proc. ICASSP-93*, pp. II-652–II-655.
[14] ——, "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition," in *Proc. ICSLP-94* pp. 439–442, Yokohama, 1994.
[15] W. Chou, T. Matsuoka, B.-H. Juang and C.-H. Lee, "A high resolution N-best search algorithm using inter-word context dependent models for continuous speech recognition," in *Proc. ICASSP-94*, pp. II-1140–II-1143.
[16] W. Chou, C.-H. Lee, B.-H. Juang and F. K. Soong, "A minimum error rate pattern recognition approach to speech recognition," *Int. J. Pattern Recog. Artif. Intell.* vol. 8 no. 1, pp. 5–31, 1994.
[17] W. Chou and B.-H. Juang, "Adaptive discriminative learning in pattern recognition," Tech. Rep., AT&T Bell Labs, Murray Hill, NJ.
[18] D. Pollard, *Convergence of Stochastic Process, Springer Series in Statistics*. New York: Springer-Verlag, 1984.
[19] A. Benveniste, M. Metivier and P. Priouet, *Adaptive Algorithms and Stochastic Approximations*, New York: Springer-Verlag.
[20] J. R. Blum, "Multidimensional stochastic approximation methods," *Ann. Math. Stat.*, vol 25, pp. 737–744, 1954.
[21] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol 22, pp. 400–407, 1951.
[22] J. L. Doob, *Stochastic Process*. New York: Wiley, 1953.
[23] C.-H. Lee *et al.*, "Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition," *Comput. Speech Lang.*, vol. 4, no. 2, pp. 103–127, 1992.
[24] C.-S. Liu *et al.*, "A study on minimum error discriminative training for speaker recognition," *J. Acoust. Soc. Amer.* vol. 97, no. 1, pp. 637–648, 1995.

Biing-Hwang Juang (S'79–M'81–SM'87–F'92) received the B.Sc. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1973, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, in 1979 and 1981, respectively.

In 1978, he conducted research on vocal tract modeling at the Speech Communications Research Laboratory (SCRL). He joined Signal Technology, Inc., in 1979 as Research Scientist, working on signal and speech related topics. Since 1982, he has been with AT&T Bell Laboratories, Murray Hill, NJ, where he is engaged in a wide range of speech-related research activities. He has published extensively in the area of speech communication. He is co-author of *Fundamentals of Speech Recognition* (Englewood Cliffs, NJ: Prentice-Hall).

Dr. Juang was an editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH,AND SIGNAL PROCESSING (1986–1988), the IEEE TRANSACTIONS ON NEURAL NETWORKS (1992–1993), and the *Journal of Speech Communication* (1992–1994). He has served on the Digital Signal Processing and the Speech Technical Committees as well as the Conference Board of the IEEE Signal Processing Society, and was (1991–1993) Chairman of the Technical Committee on Neutral Networks for Signal Processing. He is currently Editor-in-Chief of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He also serves on international advisory boards outside the United States. He received the 1993 Senior Paper Award, the 1994 Senior Paper Award, and the 1994 Best Signal Processing Magazine Paper Award, all from the IEEE Signal Processing Society. He holds a dozen patents.

Wu Chou (SM'87–M'91) received the M.S. degree in electrical engineering in 1987, the M.S. degree in statistics in 1988, and the Ph.D. degree in electrical engineering in June 1990, all from Stanford University, Stanford, CA.

Since 1990, he has been with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ. His research interests include speech recognition, speaker recognition and decoding methods in large vocabulary speech recognition, wireless and multimedia communication, acoustic assisted image coding and animation, high-fidelity audio signal processing, mixed-signal DSP, and oversampled sigma-delta conversion.

Dr. Chou is a member of Sigma Xi.

Chin-Hui Lee (S'79–M'81–SM'90–F'96) received the B.S. degree from National Taiwan University, Taipei, Taiwan, R.O.C., in 1973, the M.S. degree from Yale University, New Haven, CT, 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, CA, where he engaged in research in speech coding, speech recognition, and signal processing for the development of the DSC-2000 Voice Server. Since 1986, he has been with Bell Laboratories, Murray Hill, NJ, where he is now a Distinguished Member of Technical Staff. His current research interests include signal processing, speech modeling, adaptive and discriminative modeling, speech recognition, speaker recognition, and spoken dialogue processing. His research scope is reflected in a recent edited book, *Automatic Speech and Speaker Recognition: Advanced Topics* (Boston: Kluwer, 1996).

From 1991 to 1995, he was an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was a member of the ARPA Spoken Language Coordination Committee between 1991 and 1995. He has also been a member of the Speech Technical Committee of the IEEE Signal Processing Society (SPS) since 1995. In 1996, he helped promote the newly formed SPS Multimedia Signal Processing (MMSP) Technical Committee, and is a member of the committee. Dr. Lee is a recipient of the 1994 SPS Senior Award, and currently serves as the Chairman of the SPS Speech Technical Committee.