On the Relations Between Modeling Approaches for Speech Recognition

YARIV EPHRAIM, MEMBER, IEEE, AND LAWRENCE R. RABINER, FELLOW, IEEE

Abstract —Some relations among approaches that have been applied to estimating models for acoustic signals in speech recognition systems are examined. In particular, the maximum likelihood (ML), maximum mutual information (MMI), and minimum discrimination information (MDI) modeling approaches are studied. It is shown that all three approaches can be formulated uniformly as MDI modeling approaches for simultaneous estimation of the acoustic models for all words in the vocabulary and that none of the approaches requires any model correctness assumption. The three approaches differ in the effective source being modeled and in the probability distribution attributed to this source.

I. Introduction

SPEECH RECOGNITION could be performed optimally if the probability of any word[1] in the recognizer's vocabulary and the probability distribution (PD) of the acoustic signal corresponding to that word were known. In this case the recognizer that is optimal in the sense of minimizing the probability of error is a maximum *a posteriori* (MAP) decoder that chooses from all possible words in the vocabulary that word that yields the highest conditional probability given the acoustic input signal. In practice, the word probability and the PD's of the acoustic signals are not known, so only suboptimal recognizers can be implemented.

The most commonly used speech recognition approach is first to estimate the unknown word probability and the unknown PD's of the acoustic signals from long training sequences of words and acoustic signals, respectively. Then, the optimal MAP decision rule is applied to the estimated word probability and PD's as if they were the true probability measures. Such an approach is referred to as the "plug-in" method in the statistical literature. In estimating both the word probability and the PD's of the acoustic signals, parametric models are first assumed for these probability measures, and then the parameter sets of the models are estimated from the given training sequences. The model assumed for the acoustic signal is referred to as the acoustic model. The model assumed for the probability of word occurrence is referred to as the word model. Thus the estimation of the unknown statistics of the speech signal becomes a parameter estimation problem. However, it is not a standard parameter estimation problem, since the statistics of the sources (here the acoustic signal and the word) generating the training sequences are not necessarily those of the models. This estimation problem is therefore better described in terms of source modeling by parametric models.

The acoustic model for a given word is usually chosen to be a Markov source, or a hidden Markov model (HMM) [1]–[3]. Similarly, the word model is also chosen to be Markovian [4]. The estimation of the parameter sets of the HMM's for the acoustic signal is usually performed by the maximum likelihood (ML) estimation approach [5]–[7]. An ML estimate results from local maximization of the likelihood function of the HMM for a given training sequence of speech signal. This statistical inference approach is chosen for two major reasons. First, there exists an efficient algorithm, the Baum algorithm [5]–[7], for performing the modeling. Second, under a model correctness assumption which implies that the acoustic signal is a Markov source, the ML estimator of the parameter set of a finite alphabet HMM is consistent [8, Theorem 3.4], provided that the ML estimate globally maximizes the likelihood function, and the initial state probabilities, the state transition probabilities, and the alphabet letter probabilities conditioned on the state are all strictly positive and time invariant. Hence one can argue intuitively that using the ML estimates of the acoustic models and the MAP decision rule can lead to a speech recognition system that is asymptotically optimal [9].

Since no fundamental achievable recognition bounds (similar to Shannon bounds in coding theory) are known, and since the optimal recognizer cannot be implemented, it is not clear how good (or bad) the performance of current state of the art recognizers is as compared to the ultimate achievable performance. Hence the assumptions upon which the suboptimal recognizer is based have been repeatedly challenged in an attempt to improve recognition accuracy. In particular, the optimality of the ML procedure for estimating the parameter sets of the HMM's for the acoustic signals has been questioned for two reasons. First, acoustic signals are not necessarily Markov sources. Second, the amount of training data available for

Manuscript received January 26, 1988; revised February 1989. This work was partially presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, April 1988.

The authors are with AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974.

IEEE Log Number 8933849.

[1] The term "word" is referred here to any language unit being modeled acoustically, i.e., subword units, physical words, phrases, etc.

modeling is usually limited. Hence consistency of the ML estimator is not well-defined, and the argument given earlier for the asymptotic optimality of the recognizer whose acoustic models are estimated by the ML estimation approach may no longer be valid.

Recently, two new approaches for estimating the parameter sets of HMM's for given acoustic signals were proposed. The first is the maximum mutual information (MMI) approach [9], [10]. This approach assumes that a word model is given and attempts to find the set of HMM's for the acoustic signals for which the sample average of the mutual information with respect to the given word model is maximum. The second approach is the minimum discrimination information (MDI) estimation approach [11]. The discrimination information, also known as the cross entropy, relative entropy, directed divergence, *I*-divergence, and Kullback–Leibler number, is a measure between two PD's, one is related to the source being modeled and the other to the model being used. The modeling is performed by joint minimization of the discrimination information measure over all PD's of the source that satisfy a given set of moment constraints and all PD's of the model from the given parametric family. The consistency of the MDI estimation approach was discussed by Shore and Johnson [19]. They showed that, if the source is characterized by moment constraints, then the MDI modeling approach is the only correct inference approach, in the sense of satisfying a set of consistency axioms. Any other inference approach will either provide the same estimate as the MDI approach or will lead to inconsistency. The important case of MDI hidden Markov modeling of acoustic signals for which covariance constraints are available was considered in [11].

Both MMI and MDI modeling approaches aim indirectly at reducing the error rate of the recognizer. The MMI approach was motivated by the idea of estimating the set of acoustic models that minimizes the average code length needed for correct decoding of each spoken word from its acoustic signal. The MDI approach capitalizes on modeling using reliable aspects of the training data. This results from estimating the model for each word that yields MDI with respect to *all* acoustic utterances of the word satisfying the given set of moment constraints. In either case, however, it is difficult to show theoretically that the modeling approach results in a recognition scheme that minimizes the probability of error. Hence the extent to which recognition accuracy is improved, as opposed to coding efficiency or modeling accuracy, is experimentally demonstrated and therefore is highly task dependent.

The primary purpose of this paper is to establish some relations among the ML, MMI, and MDI modeling approaches. We demonstrate that all three approaches can be formulated uniformly as being MDI modeling approaches that differ in the effective source being modeled and in the PD attributed to that source. This formulation is important since it provides a common basis for comparing the three modeling approaches. Furthermore, it clearly shows the difference among the approaches in terms of the assumptions being made about the true PD of the source to be modeled and the model itself. Although the relations developed here among the three modeling approaches are given in the context of speech recognition, they are generally correct for other pattern classification problems, since we do not use any particular property of either the speech signal or the Markovian models.

We conclude the paper by proposing a new approach for estimating the acoustic and word models which is more directly related to our main objective than the other approaches discussed here, i.e., the minimization of the recognition error rate. In this approach a set of acoustic and word models is designed by minimizing the empirical classification error rate for the MAP decision rule. The purpose here is to show that such modeling is possible with complexity not greater than that of the MMI or the MDI approaches. It should be mentioned that classification approaches, other than the two-step modeling–recognition approaches discussed here, were developed recently for sources whose statistics are not explicitly known [12], [22], [23].

II. ML, MMI, and MDI Modeling Approaches

Let Y be a random variable defined on the sample space, say $\boldsymbol{Y}$, of all acoustic signals corresponding to the words in the vocabulary. Let $y \triangleq \{y_t,\ t = 0,\cdots,T-1\}$ be a realization of Y, where $y_t \in R^K$, the K-dimensional Euclidean space. Let $M \in \{1,\cdots,L\}$ be a discrete random variable representing the words in a vocabulary of size L. Let $Q_{Y|M}$ and Q_M be, respectively, the PD's attributed to the acoustic signal from a given word and to the word itself. Let $P_{Y|M}$ and P_M be, respectively, the PD's of parametric models for the acoustic signal from a given word and for the word. The parameter set of $P_{Y|M=m}$, here the HMM for the mth word, will be denoted by λ_m. The parameter set of the word model P_M will be denoted by μ. Let $Q_{Y,M}$ be the joint PD attributed to the acoustic signal and the word. Similarly, let $P_{Y,M}$ be the joint PD of the model for the acoustic signal and the word. Since the vocabulary has L words, L acoustic models and a single word model have to be designed. The parameter sets of the acoustic models $\{\lambda_m,\ m = 1,\cdots,L\}$ are estimated from appropriate acoustic training sequences. Let $y_T(m) = \{y_t(m),\ t = 0,\cdots,T-1\}$, $y_t(m) \in R^K$, be the given training sequence of acoustic signals corresponding to the mth word. For simplicity of notation, we assume that all training sequences have the same length.

To simplify the discussion, we shall assume in all subsections but Section II-C that the space $\boldsymbol{Y}$ is finite. This assumption is always met in practice because our models and training sequences are always stored in a digital computer. From the theoretical point of view, this assumption will allow us to present the main ideas in this paper in a simple way without using measure theoretic arguments. The extension to the case where $\boldsymbol{Y}$ is infinite can be achieved in a manner similar to the approach used in Csiszár and Tusnady [13]. We shall use lower case letters to denote the probability mass functions (pmf's) corresponding to the PD's defined before. Thus $q(y|m)$, $q(m)$, $p(y|m)$, $p(m)$, $q(y,m)$, and $p(y,m)$, denote the pmf's corresponding to $Q_{Y|M}$, Q_M, $P_{Y|M}$, P_M, $Q_{Y,M}$, and $P_{Y,M}$, respectively.

The discrimination information $D(Q_{Y,M}\|P_{Y,M})$ between the two PD's $Q_{Y,M}$ and $P_{Y,M}$ will play a central role in this section. It is defined [14] by

$$D(Q_{Y,M}\|P_{Y,M}) = \begin{cases} \sum_{m=1}^{L}\sum_{y\in Y} q(y,m)\ln\frac{q(y,m)}{p(y,m)}, & \text{if } Q_{Y,M} \ll P_{Y,M} \\ +\infty, & \text{otherwise} \end{cases} \tag{1}$$

where $Q_{Y,M} \ll P_{Y,M}$ denotes that $Q_{Y,M}$ is absolutely continuous with respect to $P_{Y,M}$. For $Q_{Y,M} \ll P_{Y,M}$, we have that [14, p. 13]

$$D(Q_{Y,M}\|P_{Y,M}) = D(Q_M\|P_M) + \sum_{m=1}^{L} q(m)D(Q_{Y|M=m}\|P_{Y|M=m}) \tag{2}$$

where

$$D(Q_M\|P_M) \triangleq \sum_{m=1}^{L} q(m)\ln\frac{q(m)}{p(m)} \tag{3}$$

is the discrimination information between the PD attributed to the word and the parametric model for the word, and

$$D(Q_{Y|M=m}\|P_{Y|M=m}) \triangleq \sum_{y\in Y} q(y|m)\ln\frac{q(y|m)}{p(y|m)} \tag{4}$$

is the discrimination information between the PD attributed to the acoustic signal from the mth word and the acoustic parametric model for that word. Equation (2) suggests that

$$\min_{\{\mu,\lambda_m,\, m=1,\cdots,L\}} D(Q_{Y,M}\|P_{Y,M}) = \min_{\mu}\left\{D(Q_M\|P_M) + \min_{\{\lambda_m,\, m=1,\cdots,L\}} \sum_{m=1}^{L} q(m)D(Q_{Y|M=m}\|P_{Y|M=m})\right\}. \tag{5}$$

This means that if $\{P_{Y|M=m}, m=1,\cdots,L\}$ does not depend on μ, the word model parameter set, then jointly optimal word and acoustic modeling, in the sense of minimizing the discrimination information $D(Q_{Y,M}\|P_{Y,M})$, can be independently performed by minimizing $D(Q_M\|P_M)$ and $\sum_{m=1}^{L} q(m)D(Q_{Y|M=m}\|P_{Y|M=m})$, respectively.

In this section we shall be concerned only with the estimation of the acoustic models for the different words in the vocabulary, $\{P_{Y|M=m}, m=1,\cdots,L\}$. Whenever necessary, we shall assume *a priori* knowledge of the word model P_M. We now examine the three approaches for acoustic modeling, namely, the ML, MMI, and MDI, and formulate them as MDI modeling approaches. The MDI interpretation of the ML approach is based on the original work of Csiszár and Tusnady [13].

A. ML Estimation Approach

In ML estimation the parameter set λ_m of the acoustic model $P_{Y|M=m}$ is estimated from the given training sequence $y_T(m)$ by

$$\max_{\lambda_m} \ln p(y_T(m)|m). \tag{6}$$

Let $Q_{Y|M=m}$ be the empirical distribution of the mth training sequence, i.e., the pmf $q(y|m)$ is given by

$$q(y|m) = \delta(y - y_T(m)) \triangleq \begin{cases} 1, & y = y_T(m) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\delta(\cdot)$ is a probability measure which is concentrated on $y_T(m)$. On substituting (7) into (2) we obtain

$$D(Q_{Y,M}\|P_{Y,M}) = D(Q_M\|P_M) - \sum_{m=1}^{L} q(m)\ln p(y_T(m)|m) \tag{8}$$

where we have used

$$\sum_{y\in Y} q(y|m)\ln q(y|m) = \ln q(y_T(m)|m) = \ln 1 = 0. \tag{9}$$

From (8), and the fact that $D(Q_M\|P_M)$ is independent of $\{\lambda_m, m=1,\cdots,L\}$, we see that

$$\min_{\{\lambda_m,\, m=1,\cdots,L\}} D(Q_{Y,M}\|P_{Y,M}) = D(Q_M\|P_M) - \sum_{m=1}^{L} q(m)\max_{\lambda_m}\ln p(y_T(m)|m). \tag{10}$$

Equation (10) shows that the standard ML estimation approach (6) is an MDI modeling approach, for estimating all acoustic models simultaneously, when the PD attributed to the acoustic signal from each word is concentrated in the acoustic training sequence from that word. Note that this PD is concentrated in the entire training sequence from each word, rather than in the individual vectors of that training sequence, i.e., we use $q(y|m) = \delta(y - y_T(m))$ rather than

$$q'(y|m) = \frac{1}{T}\sum_{t=0}^{T-1}\delta(y_t - y_t(m)).$$

Using $q'(y|m)$ would have led to an MDI measure which is insensitive to permutations of the vectors in the acoustic training sequence from each word. Thus the temporal information about the acoustic signal carried by the properly estimated HMM (using (6)) for that signal could have been lost. The pmf $q'(y|m)$ was considered in [13], where the MDI interpretation of the ML approach was originally given, and applied to independent identically distributed (i.i.d.) vector sources. In our case, however, neither the speech signal nor the acoustic HMM's are i.i.d. vector sources.

The MDI derivation of the ML estimation approach clearly shows that this approach does not require any model correctness assumption, since the pmf's $\{q(y|m), m = 1,\cdots,L\}$ attributed to the acoustic signals and the pmf's $\{p(y|m), m = 1,\cdots,L\}$ of the acoustic models were independently chosen. Such an assumption was previously attributed to the ML estimation approach, for example, in [15]. The significance of the MDI interpretation of the ML estimation approach is that ML estimation of the parameter set of a model for a given source is equivalent to approximating the empirical distribution of the source by the PD of the model in the MDI sense. Thus a goodness criterion for the ML estimate is introduced. This interpretation also has operational consequences—it shows that the ML estimate may be sensitive to the specific training sequence from the source used for modeling. This is a familiar situation in modeling speech signals for recognition applications. The MDI equivalence of the ML estimation approach, however, does not show that the estimated models have any additional desirable properties like consistency (see, e.g., [16], [17]). Such properties may exist only under a model correctness assumption, i.e., when the statistics of the source are identical to those of the model.

B. MMI Estimation Approach

The MMI modeling approach was first proposed in [18, p. 262] and was first applied to acoustic modeling of speech signals in [9], [10]. Let $I(Y;M)$ be the mutual information between the two random variables Y and M:

$$I(Y;M) = \sum_{m=1}^{L} \sum_{y \in Y} q^*(y|m) q^*(m) \ln \frac{q^*(y|m)}{\sum_{l=1}^{L} q^*(y|l) q^*(l)} \tag{11}$$

where $q^*(y|m)$ and $q^*(m)$ are the true pmf's of Y given M and of M, respectively. Since these pmf's are not known, the modeling approach proposed in [9], [10] is to replace the pmf's in the argument of the information measure (i.e., the argument of the logarithm function) by the pmf's of the parametric models and to calculate the expected value involved in (11) with respect to estimates $q(y|m)$ and $q(m)$ of $q^*(y|m)$ and $q^*(m)$, respectively. This results in

$$\hat{I}(Y;M) = \sum_{m=1}^{L} \sum_{y \in Y} q(y|m) q(m) \ln \frac{p(y|m)}{\sum_{l=1}^{L} p(y|l) p(l)}. \tag{12}$$

The estimate of $q(y|m)$ suggested in [9] is the empirical distribution (7). Substituting (7) into (12) gives

$$\hat{I}(Y;M) = \sum_{m=1}^{L} q(m) \ln \frac{p(y_T(m)|m)}{\sum_{l=1}^{L} p(y_T(m)|l) p(l)}, \tag{13}$$

which should be maximized over all $\{\lambda_m, m = 1,\cdots,L\}$.

Note from (12) that

$$\hat{I}(Y;M) = D(Q_{Y,M} \| P_Y P_M) - D(Q_{Y,M} \| P_{Y,M}) \tag{14}$$

where P_Y is the PD of the model for all acoustic signals, with pmf $p(y) = \sum_{m=1}^{L} p(y|m)p(m)$. Since the estimated average mutual information $\hat{I}(Y;M)$ in (12) (and similarly in (13)) comprises the difference of two nonnegative discrimination information measures, this estimate can be negative; thus it is not a valid measure for mutual information. This happens if the two random variables Y and M are strongly dependent under the P measure, e.g., when $Q_{Y,M} = P_Y P_M \neq P_{Y,M}$. For a less trivial example, let Y be a scalar random variable that can take only L values, $Y = 1,\cdots,L$, let $p(y,m) = (1-\epsilon)/L\,\delta(y-m) + \epsilon/L^2$, $0 < \epsilon < 1$, and let $q(y,m) = 0$ if $y = m$ and $q(y,m) = 1/(L^2 - L)$ otherwise. In this case we have that $\hat{I}(Y;M) = \ln(\epsilon/L) < 0$. Nevertheless, acoustic models designed by maximizing $\hat{I}(Y;M)$ have been shown to be useful in speech recognition applications [9], [10]. Hence an interpretation must be found that explains the merit of this approach other than from its being an MMI estimation approach.

Let $Q_{M|Y}$ and $P_{M|Y}$ be, respectively, the *a posteriori* PD attributed to the word given the acoustic signal and the *a posteriori* PD of the model for the word given the acoustic signal. Let $q(m|y)$ and $p(m|y)$ be, respectively, the corresponding pmf's of $Q_{M|Y}$ and $P_{M|Y}$. From (7) we have

$$q(l|y_T(m)) = \frac{q(y_T(m)|l) q(l)}{\sum_{k=1}^{L} q(y_T(m)|k) q(k)} = \begin{cases} 1, & l = m \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

which shows that the *a posteriori* probability of the mth word is one if the mth training sequence is observed and zero otherwise. For $y = y_T(m)$, the pmf $p(l|y)$. $l = 1,\cdots,L$, is given by

$$p(l|y_T(m)) = \frac{p(y_T(m)|l) p(l)}{\sum_{k=1}^{L} p(y_T(m)|k) p(k)}, \tag{16}$$

and it depends on all parameter sets of the acoustic models, $\{\lambda_m, m = 1,\cdots,L\}$, as well as on the parameter set of the word model μ. Using (15) and (16), we can rewrite (13) as

$$\hat{I}(Y;M) = -\sum_{m=1}^{L} q(m) D(Q_{M|Y=y_T(m)} \| P_{M|Y=y_T(m)}) - \sum_{m=1}^{L} q(m) \ln p(m) \quad (17)$$

where

$$D(Q_{M|Y=y_T(m)} \| P_{M|Y=y_T(m)}) = \sum_{l=1}^{L} q(l|y_T(m)) \ln \frac{q(l|y_T(m))}{p(l|y_T(m))}. \quad (18)$$

Hence the MMI estimate of the set of acoustic models is obtained from

$$\min_{\{\lambda_m,\, m=1,\cdots,L\}} \sum_{m=1}^{L} q(m) D(Q_{M|Y=y_T(m)} \| P_{M|Y=y_T(m)}). \quad (19)$$

The MDI derivation of the MMI estimation approach given in (19) shows that this approach tries to estimate a set of acoustic models that minimizes the average discrimination information measure between the *a posteriori* empirical PD of the word given the acoustic signal and the *a posteriori* PD of the model for the word given the acoustic signal. The average is taken over all words in the vocabulary. This criterion is well defined since the average discrimination information measure is nonnegative, and it constitutes a similarity measure between $P_{M|Y}$ and $Q_{M|Y}$. In the extreme case where $P_{M|Y} = Q_{M|Y}$, the average discrimination information equals zero. This is, of course, a desirable property of any estimation procedure. Similarly to the ML case, the MMI approach does not use any model correctness assumption.

It is interesting at this point to compare the ML and the MMI estimation approaches based upon their MDI interpretations. Using (4)–(7), the ML estimate can be viewed as resulting from

$$\min_{\{\lambda_m,\, m=1,\cdots,L\}} \sum_{m=1}^{L} q(m) D(Q_{Y|M=m} \| P_{Y|M=m}). \quad (20)$$

As is clearly shown by (19) and (20), the two modeling approaches minimize average discrimination information; however, they do so between different pairs of PD's. In the ML approach, the empirical PD $Q_{Y|M=m}$ is approximated by the model $P_{Y|M=m}$, while the MMI approach approximates the empirical PD $Q_{M|Y=y_T(m)}$ by the model $P_{M|Y=y_T(m)}$. This means that the ML approach tries to estimate the set of acoustic models that best approximate probability measures concentrated in the acoustic training sequences. On the other hand, the MMI approach tries to estimate the set of acoustic models for which the probability of each word given a training sequence is as close as possible to unity if the training sequence comes from the same word and is as close as possible to zero otherwise.

In terms of the classification problem considered here, the MMI modeling approach seems reasonable, since recognition is performed on the basis of the *a posteriori* PD $P_{M|Y}$, by choosing the word m that maximizes $p(m|y = z)$ for the given test sequence z. Hence it is intuitively appealing to estimate the set of acoustic models which maximizes $P_{M|Y=y_T(m)}$ for $M = m$ (by making it as close as possible to one), and at the same time minimizes $P_{M|Y=y_T(m)}$ for $M \neq m$ (by making it as close as possible to zero).

C. MDI Estimation Approach

The MDI modeling approach proposed in [11] is for estimating an individual acoustic model from a given training sequence of the speech signal. We shall first review the principles of this approach, and then generalize it for multiple model estimation. The major differences between this approach and those considered previously is in the way the PD attributed to the acoustic signal is estimated from the given training data. Rather than assuming that the PD of the acoustic signal is concentrated in the given training sequence, the MDI approach of [11] considers all PD's that satisfy a set of moment constraints from the acoustic training sequence. The PD that yields MDI with respect to the acoustic model is chosen. This PD for the acoustic signal, called the MDI PD with respect to the model, depends on the set of moment constraints and on the parameter set of the model. The modeling is performed by minimizing the discrimination information measure between the MDI PD and the PD of the model, over all parameter sets of the model. In practice, the moment constraints have to be estimated from the acoustic signal. In this case the estimated moments are considered as if they were the true moments. If the signal is stationary and ergodic, and the frame length from which each moment is estimated is large enough, then good estimates result. Alternatively, the estimated moments of the acoustic signal can be considered as being a characterization of the acoustic signal, regardless of how well they approximate the true moments.

The special important case of modeling sources from which second-order moments are available, using HMM's with Gaussian autoregressive (AR) output PD's, was considered in [11]. Specifically, it was assumed that, for each time t, we are given a $K \times K$ covariance matrix $R_t(m)$, which agrees with the covariance matrix of $y_t(m) \in R^K$ within some symmetric band, say B. The elements of $R_t(m)$ that are outside B are assumed unknown. Such a matrix is referred to as the partial covariance of the source at time t. Let $R_T(m) \triangleq \{R_t(m), t = 0,\cdots,T-1\}$ be the sequence of given partial covariance matrices from the acoustic signal for the mth word. Let $\Omega(R_T(m))$ be the set of all PD's $Q_{Y|M=m}$ that satisfy the given set of partial covariance matrices. The MDI estimate of the parameter set of the model for the mth word λ_m is obtained from

$$\min_{\lambda_m} \min_{Q_{Y|M=m} \in \Omega(R_T(m))} D(Q_{Y|M=m} \| P_{Y|M=m}) \quad (21)$$

where now

$$D(Q_{Y|M=m}\|P_{Y|M=m}) = \int q(y|m)\ln\frac{q(y|m)}{p(y|m)}dy \qquad (22)$$

provided that $Q_{Y|M=m}$ and $P_{Y|M=m}$ are absolutely continuous with respect to Lebesgue measure. The convention that $\ln 0 = -\infty$, $\ln(c/0) = \infty$ for any positive number c, $0\cdot(+\infty) = 0$, and $0\cdot(-\infty) = 0$, is assumed in evaluating (22).

The double minimization in (21) is implemented by alternate minimization of $D(Q_{Y|M=m}\|P_{Y|M=m})$, once over all $Q_{Y|M=m} \in \Omega(R_T(m))$ assuming $P_{Y|M=m}$ is known, and then over all $P_{Y|M=m}$ for the resulting MDI PD with respect to the old model, as originally proposed in [13]. The algorithm starts from a given model, e.g., the model estimated by the ML approach, and generates a sequence of HMM's with nonincreasing discrimination information values. The algorithm is stopped if the difference in discrimination information in two consecutive iterations is smaller than or equal to a given threshold. For a given HMM $P_{Y|M=m}$ and partial covariance matrices $\{R_t(m)\}$ such that each $R_t(m)$ has any positive definite extension, a unique PD $Q_{Y|M=m}$ exists that minimizes $D(Q_{Y|M=m}\|P_{Y|M=m})$ over $\Omega(R_T(m))$. The probability density function (pdf) of the MDI PD is given by

$$q(y|m) = C_m p(y|m)\exp\left\{-\frac{1}{2}\sum_{t=0}^{T-1} y_t^{\#}\Lambda_t(m)y_t\right\} \qquad (23)$$

where C_m is a normalization constant that makes $\int dy q(y|m) = 1$, the pound sign denotes vector transpose, and $\{\Lambda_t(m)\}$ is a set of symmetric matrices of Lagrange multipliers that vanish outside B and are chosen so that

$$\int dy_t\, q(y_t|m)y_t y_t^{\#} = R_t(m) \qquad (24)$$

within the given band B. Note that C_m is a highly nonlinear function of λ_m and $\{\Lambda_t(m)\}$. The discrimination information between the MDI PD with respect to the given model and the model itself, called the MDI measure with respect to the given model, equals

$$\min_{Q_{Y|M=m}\in\Omega(R_T(m))} D(Q_{Y|M=m}\|P_{Y|M=m}) = -\ln C_m - \frac{1}{2}\operatorname{tr}\sum_{t=0}^{T-1} R_t(m)\Lambda_t(m). \qquad (25)$$

The Lagrange multipliers can be estimated from maximization of the right side of (25) over all $\{\Lambda_t(m)\}$ for which $q(y|m)$ is a nonsingular pdf, say the set δ_{λ_m}. This function is unimodal on δ_{λ_m}, and hence the maximization can be performed using any standard constrained optimization procedure. Given the MDI PD with respect to the old model, a new HMM that decreases the MDI measure (25) can be estimated by maximizing an appropriately chosen auxiliary function, using the "forward-backward" procedure. This results in a reestimation algorithm similar to the Baum algorithm [5]–[7].

The extension of the MDI approach to multiple model design can be done as follows. Let $\bar{Y} \triangleq \{Y(1),\cdots,Y(N)\}$ be a sequence of N random variables representing acoustic signals, where each random variable is defined on $\boldsymbol{Y}$. Let $\bar{M} \triangleq \{M_1,\cdots,M_N\}$ be a sequence of N discrete random variables representing words, where each can take L values. Let $\bar{y} \triangleq \{y(1),\cdots,y(N)\}$ be a realization of $\bar{Y}$, where $y(n) \triangleq \{y_0(n),\cdots,y_{T-1}(n)\}$ and $y_t(n) \in R^K$. Similarly, let $\bar{m} \triangleq \{m_1,\cdots,m_N\}$ be a realization of $\bar{M}$. Let $Q_{\bar{Y},\bar{M}}$, $P_{\bar{Y},\bar{M}}$, $Q_{\bar{Y}|\bar{M}}$, $P_{\bar{Y}|\bar{M}}$, $Q_{\bar{M}}$, and $P_{\bar{M}}$, respectively, be defined similarly to $Q_{Y,M}$, $P_{Y,M}$, $Q_{Y|M}$, $P_{Y|M}$, Q_M, and P_M. Let $\bar{R} \triangleq \{R_T(n),\ n = 1,\cdots,N\}$ be a given set of sequences of partial covariance matrices corresponding to the acoustic signals $\bar{Y}$. Let $\Omega(\bar{R}) \triangleq \{Q_{\bar{Y}|\bar{M}=\bar{m}}$ that satisfies $\bar{R}\}$. Similar to (2) we have that

$$D(Q_{\bar{Y},\bar{M}}\|P_{\bar{Y},\bar{M}}) = D(Q_{\bar{M}}\|P_{\bar{M}}) + \sum_{\bar{l}} q(\bar{l})D(Q_{\bar{Y}|\bar{M}=\bar{l}}\|P_{\bar{Y}|\bar{M}=\bar{l}}) \qquad (26)$$

where $\bar{l}$ is defined similarly to $\bar{m}$. When modeling is performed using partial covariance matrices $\bar{R}$ corresponding to a sequence of words, say $\bar{m}$, we implicitly assume that $q(\bar{l}) = 1$ for $\bar{l} = \bar{m}$ and $q(\bar{l}) = 0$, otherwise. Hence simultaneous MDI estimation of all acoustic models can be achieved by

$$\min_{\{\lambda_m\}_{m=1}^{L}}\ \min_{Q_{\bar{Y}|\bar{M}=\bar{m}}\in\Omega(\bar{R})} D(Q_{\bar{Y}|\bar{M}=\bar{m}}\|P_{\bar{Y}|\bar{M}=\bar{m}}). \qquad (27)$$

We shall assume here that $N \geq L$, and that $\bar{R}$ contains at least one sequence of partial covariance matrices for each word in the vocabulary. We shall now show that if the acoustic signals are assumed statistically independent under the P measure, then the minimization in (27) can be independently performed for each acoustic model using the approach developed in [11]. Hence this approach can simply be viewed as an MDI approach for simultaneous estimation of all acoustic models.

Assume that

$$p(\bar{y}|\bar{m}) = \prod_{n=1}^{N} p(y(n)|m_n). \qquad (28)$$

Following the proof of [11, theorem 1], it can be shown that if each partial covariance matrix $R_t(n)$, $t = 0,\cdots,T-1$, $n = 1,\cdots,N$, has any positive definite extension, then there exists a unique PD $Q_{\bar{Y}|\bar{M}=\bar{m}}$ that minimizes $D(Q_{\bar{Y}|\bar{M}=m}\|P_{\bar{Y}|\bar{M}=m})$ over $\Omega(\bar{R})$, with pdf given by

$$\begin{aligned} q(\bar{y}|\bar{m}) &= Cp(\bar{y}|\bar{m})\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{t=0}^{T-1} y_t(n)^{\#}\Lambda_t(n)y_t(n)\right\} \\ &= \prod_{n=1}^{N} C_n p(y(n)|m_n) \\ &\quad \cdot\exp\left\{-\frac{1}{2}\sum_{t=0}^{T-1} y_t(n)^{\#}\Lambda_t(n)y_t(n)\right\} \\ &= \prod_{n=1}^{N} q(y(n)|m_n), \end{aligned} \qquad (29)$$

TABLE I
SUMMARY COMPARISON OF ML, MMI, AND MDI MODELING APPROACHES

	Measure	Source	Model
ML	$\Sigma_{m=1}^{L} q(m) D(Q_{Y\|M=m} \| P_{Y\|M=m})$	$q(y\|m) = \delta(y - y_T(m))$	$p(y\|m)$
MMI	$\Sigma_{m=1}^{L} q(m) D(Q_{M\|Y=y_T(m)} \| P_{M\|Y=y_T(m)})$	$q(m\|y) = \delta(y - y_T(m))$	$p(m\|y) = \dfrac{p(y\|m)p(m)}{\Sigma_{l=1}^{L} p(y\|l)p(l)}$
MDI	$\Sigma_{m=1}^{L} D(Q_{Y\|M=m} \| P_{Y\|M=m})$	$q(y\|m) = C_m p(y\|m) \exp\left\{-\frac{1}{2}\sum_{t=0}^{T-1} y_t^{\#} \Lambda_t(m) y_t\right\}$	$p(y\|m)$

where $q(y(n)|m_n)$ is the MDI PD with respect to $p(y(n)|m_n)$, given $R_T(n)$, defined similarly to (23). In this case we have that [14, p. 12]

$$\min_{Q_{\bar{Y}|\bar{M}=\bar{m}} \in \Omega(\bar{R})} D(Q_{\bar{Y}|\bar{M}=\bar{m}} \| P_{\bar{Y}|\bar{M}=\bar{m}})$$
$$= \sum_{n=1}^{N} \min_{Q_{Y_n|M_n=m_n} \in \Omega(R_T(n))} D(Q_{Y_n|M_n=m_n} \| P_{Y_n|M_n=m_n}). \quad (30)$$

Now, if $N = L$, then we have a single acoustic utterance from each word, say $y_T(l)$ for $m_e = l$, and estimation of the parameter sets of the models is performed independently, since

$$\min_{\{\lambda_l\}_{l=1}^{L}} \min_{Q_{\bar{Y}|\bar{M}=\bar{m}} \in \Omega(\bar{R})} D(Q_{\bar{Y}|\bar{M}=\bar{m}} \| P_{\bar{Y}|\bar{M}=\bar{m}})$$
$$= \sum_{l=1}^{L} \min_{\lambda_l} \min_{Q_{Y_l|M_l=l} \in \Omega(R_T(l))} D(Q_{Y_l|M_l=l} \| P_{Y_l|M_l=l}). \quad (31)$$

If $N > L$, we may have more than one acoustic utterance for each word in the vocabulary. In this case modeling is performed by

$$\min_{\lambda_l} \sum_{n:m_n=l} \min_{Q_{Y_n|M_n=l} \in \Omega(R_T(n))} D(Q_{Y_n|M_n=l} \| P_{Y_n|M_n=l}),$$
$$l = 1, \cdots, L. \quad (32)$$

While the minimization over $Q_{Y_n|M_n=l} \in \Omega(R_T(n))$ is independently applied for each utterance of acoustic signal, the minimization over λ_l is performed using all acoustic utterances from the same word. This minimization can be accomplished similarly to the case of a single acoustic utterance per word through maximization of an auxiliary function that comprises the sum of individual auxiliary functions for the different utterances from the same word (see [11, (14)]), [21].

D. Discussion

We have seen that the three modeling approaches considered here—ML, MMI, and MDI—are optimal approaches for simultaneous estimation of all acoustic models, in the minimum average discrimination information sense. The ML and MDI aproaches minimize the average discrimination information measure between the PD attributed to the acoustic signal from a given word, and the PD of the acoustic model for that word (see (20) and (31)). The MMI modeling approach minimizes the average discrimination information measure between the PD attributed to the word given an acoustic signal, and the model for the word given the acoustic signal (see (19)). The ML and the MMI attribute to the acoustic signal a PD that is concentrated in the individual training sequences for the different words. Thus these two approaches make precisely the same assumptions about the acoustic signal being modeled. The MDI approach, however, attributes to the acoustic signal from each word a more robust PD. This PD is obtained by considering all PD's for the acoustic signal for a given word, which satisfy a given set of moments from this signal. Since the three modeling approaches are variants of the MDI modeling approach, and in MDI modeling the PD attributed to the source need not be the same as the PD of the model, none of the three approaches assumes model correctness.

Table I summarizes the three modeling approaches in terms of the average MDI measure they minimize, the PD attributed to the source being modeled, and the PD of the model itself.

The estimation of the acoustic models by the ML and MDI approaches can be done independently of each other, while in the MMI approach all acoustic models must be simultaneously estimated. Hence, if the size of the vocabulary is increased, all acoustic models must be redesigned in the MMI approach, whereas in the ML and MDI approaches only models for the new words must be designed. From the three estimation approaches, the ML approach is the easiest to implement using the Baum algorithm. The MMI approach is usually implemented using general constrained optimization procedures, e.g., a variant of the steepest descent method [20]. The implementation of the MDI algorithm requires constrained maximization in a high dimensional Euclidean space for estimating the Lagrange multipliers corresponding to the given moments. The estimation of the parameter set of the model itself, however, can be done efficiently by a procedure that generalizes the Baum algorithm.

III. MODEL DESIGN FOR MINIMUM EMPIRICAL ERROR RATE

Next, we propose a new approach for designing the set of acoustic and word models in which the models are optimized for the decoder used during recognition in the minimum empirical error rate sense. We focus on the

MAP decoder assumed throughout this paper which is implemented using the acoustic and word models. The acoustic and word models are designed by minimizing the empirical error rate function over the parameter sets of the models. The minimization can be performed using general optimization procedures similar to those used in implementing the MMI and MDI approaches.

The probability of classification error is given by,

$$P_e(\lambda,\mu) = 1 - \sum_{m=1}^{L} \sum_{y \in \omega_{\lambda,\mu}(m)} q^*(y,m) \qquad (33)$$

where $q^*(y,m)$ is the true joint pmf of the acoustic signal and word, $\lambda \triangleq \{\lambda_m,\ m = 1,\cdots,L\}$, and $\omega_{\lambda,\mu}(m)$ is the set of all acoustic signals $y \in Y$ that will be decoded as the mth word. The set $\{\omega_{\lambda,\mu}(m),\ m = 1,\cdots,L\}$ constitutes a partition of Y. For the MAP decoder we have

$$\omega_{\lambda,\mu}(m) = \left\{ y: \ln \frac{p(y|m)p(m)}{p(y|l)p(l)} \geq 0, \quad \begin{matrix} l = 1,\cdots,L \\ l \neq m \end{matrix} \right\}, \qquad m = 1,\cdots,L \qquad (34)$$

where we arbitrarily assign y to the set of the lowest index when ties occur. Let $1_{\omega_{\lambda,\mu}(m)}(y)$ be the characteristic function associated with $\omega_{\lambda,\mu}(m)$,

$$1_{\omega_{\lambda,\mu}(m)}(y) = \begin{cases} 1, & y \in \omega_{\lambda,\mu}(m) \\ 0, & \text{otherwise.} \end{cases} \qquad (35)$$

Assume that we have labeled training data of N pairs of words and acoustic signals $\{(w_n, y_T(n)),\ n = 1,\cdots,N\}$, where $w_n \in \{1,\cdots,L\}$ and $y_T(n) \in Y$ for all $n = 1,\cdots,N$. Furthermore, assume that $N \gg L$ to give meaningful estimation. The training data are used for estimating the unknown pmf $q^*(y,m)$. Using the empirical distribution estimate

$$q(y,m) = \frac{1}{N} \sum_{n=1}^{N} \delta(y - y_T(n), m - w_n) \qquad (36)$$

and (35), we get from (33) the following empirical classification error rate function

$$\hat{P}_e(\lambda,\mu) = 1 - \frac{1}{N} \sum_{m=1}^{L} \sum_{n:\ w_n = m} 1_{\omega_{\lambda,\mu}(m)}(y_T(n)). \qquad (37)$$

The acoustic and word models are estimated from minimization of $\hat{P}_e(\lambda,\mu)$ over the domain of $\{\lambda,\mu\}$. This error rate function is well-defined and attains its minimum for some set of models. The minimization of this function could be done using general optimization procedures, e.g., the steepest descent method, if $\hat{P}_e(\lambda,\mu)$ were differentiable. Since this is not the case here, we approximate the characteristic functions $\{1_{\omega_{\lambda,\mu}(m)}(y)\}$ by differentiable functions, e.g., sigmoid functions, and minimize the resulting approximated error rate function.

The major advantage of this approach, as compared to the other modeling approaches discussed in Section II, is that here the model design procedure is optimal in the minimum empirical classification error rate sense for the given decision rule. Thus the ultimate criterion of speech recognition is used in estimating the unknown statistics of speech, and the models are optimized for the decision rule used during recognition. It should be understood, however, that if the amount of training data is insufficient, minimizing the empirical error rate for the training set does not guarantee minimum error rate on the test data. Conditions for convergence of the empirical classification error rate to the probability of classification error can be found in [24].

Acknowledgment

The authors wish to thank Amir Dembo for many interesting discussions during this work. The valuable comments of Salim Roucos, Neri Merhav, and the anonymous referees are greatly appreciated.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[2] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[3] A. B. Poritz, "Hidden Markov models: A guided tour," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 7–13.

[4] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 2, pp. 179–190, Mar. 1983.

[5] 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.

[6] 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.

[7] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov Sources," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 5, pp. 729–734, Sept. 1982.

[8] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, pp. 1554–1563, 1966.

[9] P. F. Brown, "The acoustic-modeling problem in automatic speech recognition," Ph.D. dissertation, Dept. Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, PA, 1987.

[10] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1986, pp. 49–52.

[11] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. Inform. Theory*, vol. IT-35, no. 5, pp. 1001–1013, Sept. 1989.

[12] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 2, pp. 278–286, Mar. 1988.

[13] I. Csiszár and G. Tusnady, "Information geometry and alternating minimization procedures," Math. Inst. Hungarian Academy of Sciences, Budapest, Preprint 35, 1983.

[14] S. Kullback, *Information Theory and Statistics.* New York: Dover, 1968.

[15] J. E. Shore, "On a relation between maximum likelihood classification and minimum relative-entropy classification," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 6, pp. 851–854, Nov. 1984.

[16] J. C. Kiefer, *Introduction to Statistical Inference*. New York: Springer-Verlag, 1987.

[17] S. S. Wilks, *Mathematical Statistics*. New York: Wiley, 1961.

[18] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London: Prentice-Hall International, 1982.

[19] 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, no. 1, pp. 26–37, Jan. 1980 (cf. comments and corrections, *IEEE Trans. Inform. Theory*, vol. IT-29, no. 6, pp. 942–943, Nov. 1983).

[20] D. G. Luenberger, *Linear and Non-Linear Programming*. Reading, MA: Addison-Wesley, 1984.

[21] Y. Ephraim, A. Dembo, and L. R. Rabiner, "Extensions of the minimum discrimination information approach for hidden Markov modeling," unpublished paper.

[22] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. IT-35, no. 2, pp. 401–408, Mar. 1989.

[23] N. Merhav and J. Ziv, "Parameter estimation for Markov sources with empirically observed statistics," submitted for publication.

[24] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*. New York: Springer-Verlag, 1982.