S13.8

ESTIMATION OF HIDDEN MARKOV MODEL PARAMETERS BY MINIMIZING EMPIRICAL ERROR RATE

A. Ljolje
Y. Ephraim
L.R. Rabiner

Speech Research Department
AT&T Bell Laboratories
Murray Hill. New Jersey 07974

ABSTRACT

A new approach for designing a set of acoustic models for speech recognition applications which results in minimum empirical error rate for a given decoder and training data is studied. In an evaluation of the system for an isolated word recognition task, hidden Markov models (HMM's) are used to characterise the probability density functions of the acoustic signals from the different words in the vocabulary. Decoding is performed by applying the maximum a-posteriori decision rule to the acoustic models. The HMM's are estimated by minimizing a differentiable cost function, which approximates the empirical error rate function, using the steepest descent method. The HMM's designed by the minimum emprical error rate approach were used in multispeaker recognition of the English E-set words and compared to models designed by the standard maximum likelihood estimation approach. The proposed approach increased recognition accuracy from 68.2% to 76.2% on the training set and from 53.4% to 56.4% on an independent set of test data.

I. INTRODUCTION

Isolated word speech recognition could optimally be performed if the true probability of all words in the vocabulary were known, as well as the corresponding probability distributions (PD's) of the acoustic signal. In such a case the recognizer is a maximum a-posteriori (MAP) classifier which is optimal in the sense of minimizing the probability of error. The recognized word, selected from all possible words in the vocabulary, is chosen as the word with the highest joint probability with the acoustic input signal. In practice, however, word probabilities and acoustic signal PD's are not known and we can therefore only design suboptimal classifiers.

The traditional approach to this problem in speech recognition is to estimate the unknown word probabilities and the PD's of the acoustic signal from training data. Word data sets and acoustic signals are used respectively. These word probabilities and the PD's of the acoustic signal are consequently assumed to be the true word probabilities and PD's of the acoustic signal. The optimal MAP decoder is then applied to perform classification. The estimation of both word probabilities and PD's of the acoustic signal is most often based on some parametric model. The parameter sets of these models are estimated from the word and acoustic signal training sequences. The model assumed for the probability of the word occurrence is referred to as the word model and the model assumed for the acoustic signal is called the acoustic model. Unfortunately this creates the problem of choosing a model with the same statistics as the sources that generated the training sequences used to estimate the model parameters. This is generally not the case and the parameter estimation problem becomes a problem in 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 or language model is also chosen to be Markovian [4]. The estimation of the parameter sets of the HMM's for the acoustic signals 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, and some other mild assumptions, the ML estimator of the parameter set of the model is asymptotically efficient [8, Theorem 3.4]. Hence, one can intuitively argue that using the ML estimates of the acoustic models and the MAP decision rule can lead to a speech recognition system which is asymptotically optimal [9].

Recently a new approach for estimating the parameters of HMM's by maximizing the empirical mutual information between the acoustic signal and the corresponding word was proposed [10]-[11]. The maximization is performed using general optimization procedures, e.g., the steepest-descent method. HMM's designed by this approach were experimentally demonstrated to perform better than models designed by the ML estimation approach in speech recognition applications [11]. Nevertheless, since neither the likelihood nor the mutual information is directly related to the probability of classification error, HMM's designed by these approaches are not guaranteed to yield minimum error rate even for the given training data.

In this paper we focus on the problem of estimation of the parameters of the HMM's for the acoustic signals by minimizing the empirical error rate of the recognizer for the MAP decoder and a given set of training data [12]. We develop an algorithm for performing the estimation and study its performance in multi-speaker recognition of isolated versions of the English E-set words. The acoustic models are simultaneously estimated by minimizing a differentiable function which approximates the empirical error rate function. The minimization is performed using the steepest descent method. The cost function used in this approach is shown to be fundamentally different from the cost functions employed by the ML and the MMI approaches for hidden Markov modeling.

In section 2 we derive the algorithm for the minimum empirical error rate design of HMM parameters and compare this approach with the MMI approach. In Section 3 we discuss experimental recognition results and compare the proposed approach with the ML estimation approach. Comments are given in Section 4.

II. HMM DESIGN FOR MINIMUM EMPIRICAL ERROR RATE

Let Y be a random variable defined on the sample space, say Y, of all acoustic signals corresponding to the words in the vocabulary. Let $y=\{y_t,\ t=1,\cdots,\ T\}$, where $y_t \in R^K$, the K-dimensional Euclidean space, be a realization of Y. For simplicity of notation, we assume throughout this paper that all acoustic signals have the same duration T. Let $M \in \{1,\cdots,\ L\}$ be a discrete random variable representing the words in a vocabulary of size L. Let $P_{Y|M}$ and P_M be, respectively, the PD's of parametric models for the acoustic signal for a given word, and for the word. The parameter set of $P_{Y|M=m}$, here the HMM for the m-th word, will be denoted by λ_m. The parameter set of the word model P_M will be denoted by μ. We shall assume that Y is a discrete space and use lower case letters to denote probability density functions (pmf's). Thus, $p(y \mid m)$ and $p(m)$ will denote the pmf's corresponding to $P_{Y|M}$ and P_M, respectively.

The problem of designing a speech recognition system for an L word vocabulary is that of estimating the L acoustic models and the word model from a given set of training data. In this paper we shall mainly be concerned with the estimation of the acoustic models. Whenever necessary, we shall assume a-priori knowledge of the word model P_M. We assume that training data consisting of a labeled set of N pairs of words and acoustic signals is available. In particular, this training data is denoted by $\{(w_n, y_T(n)),\ n=1,\cdots,\ N\}$, where $w_n \in \{1,\cdots,\ L\}$ and $y_T(n) \in Y$ for all $n=1,\cdots,\ N$. Furthermore, for meaningful estimation of the models, we assume that $N \gg L$.

We now derive the proposed approach for hidden Markov modeling and compare it with the MMI and ML modeling approaches. Let $\omega_{\lambda,\mu}(m)$ denote the set of all acoustic signals $y \in Y$ to be decoded as the m-th

CH2847-2/90/0000-0709 $1.00 © 1990 IEEE

word, where decoding is performed using the acoustic and word models, λ and μ, respectively. For MAP decoding we have

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

where we arbitrarily assign y to the set of the lowest index when ties occur. The set $\{\omega_{\lambda,\mu}(m), m=1,\cdots,L\}$ constitutes a partition of Y. Let $q^*(y,m)$ denote the true joint pmf of the acoustic signal and word. The probability of classification error associated with the partition $\{\omega_{\lambda,\mu}(m), m=1,\cdots,L\}$ is given by

$$\begin{aligned} P_e(\lambda,\mu) &= 1 - \sum_{m=1}^{L} \sum_{y \in \omega_{\lambda,\mu}(m)} q^*(y,m) \\ &= 1 - \sum_{m=1}^{L} \sum_{y \in Y} 1_{\omega_{\lambda,\mu}(m)}(y) q^*(y,m), \end{aligned} \quad (2)$$

where $1_{\omega_{\lambda,\mu}(m)}(y)$ denotes the characteristic function of the set $\omega_{\lambda,\mu}(m)$,

$$1_{\omega_{\lambda,\mu}(m)}(y) = \begin{cases} 1 & y \in \omega_{\lambda,\mu}(m) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The empirical error rate is obtained by replacing the unknown pmf $q^*(y,m)$ in (2) by the sample distribution estimate $q(y,m)$ obtained from the given training data $\{(w_n, y_T(n)), n=1,\cdots,N\}$. This estimate is given by

$$q(y,m) = \frac{1}{N} \sum_{n=1}^{N} \delta(y - y_T(n), m - w_n). \quad (4)$$

From (2) and (4) we obtain the following empirical classification error rate function,

$$\hat{P}_e(\lambda,\mu) = 1 - \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} 1_{\omega_{\lambda,\mu}(m)}(y_T(n)). \quad (5)$$

This function constitutes the average number of utterances from the training data which were misclassified by the MAP decoder for a given set of acoustic and word models. Minimum error rate on the training data can therefore be achieved by minimizing (5) over $\{\lambda,\mu\}$. In practice, a differentiable approximation of (5) is minimized using steepest descent methods. We use the following approximation of $1_{\omega_{\lambda,\mu}(m)}(y)$,

$$1_{\omega_{\lambda,\mu}(m)}(y) \approx \frac{p(y \mid m)p(m)}{\sum_{l=1}^{L} p(y \mid l)p(l)}, \quad (6)$$

and perform the minimization of (5) over λ. For initial parameter set λ estimated by the ML approach, $p(y \mid m) \gg p(y \mid l)$, $l \neq m$ if y is an acoustic signal from the m-th word, and $p(y \mid l) \gg p(y \mid m)$ for some $l \neq m$, otherwise. Hence, the right hand side of (6) approximates $1_{\omega_{\lambda,\mu}(m)}(y)$ very well. As is easy to see, this approximation is a differentiable function provided that $\{p(y \mid l), l=1,\cdots,L\}$ are differentiable. On substituting (6) into (5) we obtain the cost function used for estimating the acoustic models in the minimum empirical error rate HMM design approach.

$$\hat{P}_e(\lambda,\mu) \approx 1 - \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \frac{p(y_T(n) \mid m)p(m)}{\sum_{l=1}^{L} p(y_T(n) \mid l)p(l)}. \quad (7)$$

For equally likely words, i.e., $p(m)=1/L$, we have

$$\hat{P}_e(\lambda,\mu) \approx 1 - \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \frac{p(y_T(n) \mid m)}{\sum_{l=1}^{L} p(y_T(n) \mid l)}. \quad (8)$$

It is interesting at this point to compare the approximate empirical error rate function (8) with the cost function associated with the MMI modeling approach [10]-[11]. Let $I(Y;M)$ be the mutual information between the two random variables Y and M.

$$I(Y;M) = \sum_{m=1}^{L} \sum_{y \in Y} q^*(y,m) \ln \frac{q^*(y \mid m)}{\sum_{l=1}^{L} q^*(y \mid l) q^*(l)}, \quad (9)$$

where $q^*(y \mid m)$ and $q^*(m)$ are the true pmf's of Y given M and of M, respectively. In the MMI modeling approach, the pmf's in the argument of the information measure (i.e., the argument of the logarithm function) are replaced by the pmf's of the parametric models, and the expected value involved in (9) is calculated with respect to the empirical distribution estimate (4) of $q^*(y,m)$. This results in the following cost function

$$\begin{aligned} \hat{I}(Y;M) &= \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \ln \frac{p(y_T(n) \mid m)}{\sum_{l=1}^{L} p(y_T(n) \mid l)p(l)} \\ &= \frac{1}{N} \ln \prod_{m=1}^{L} \prod_{n:w_n=m} \frac{p(y_T(n) \mid m)}{\sum_{l=1}^{L} p(y_T(n) \mid l)p(l)}, \end{aligned} \quad (10)$$

which is maximized over λ using the steepest-descent approach. For equally likely words, we have that

$$\hat{I}(Y;M) = \ln N + \frac{1}{N} \ln \prod_{m=1}^{L} \prod_{n:w_n=m} \frac{p(y_T(n) \mid m)}{\sum_{l=1}^{L} p(y_T(n) \mid l)}, \quad (11)$$

Comparing (8) and (11) shows that the two approaches use exactly the same statistics, $p(y_T(n) \mid m) / \sum_{l=1}^{L} p(y_T(n) \mid l)$, but in a completely different manner. In the minimum empirical error rate design approach, the sum of these statistics is maximized over λ, while in the MMI approach the product of these statistics is maximized. From (11), the MMI cost function is dominanted by the smallest term in $\{p(y_T(n) \mid m) / \sum_{l=1}^{L} p(y_T(n) \mid l), n=1,\cdots,N\}$, or by the least favorable utterance of acoustic signals over all models! Hence, the MMI cost function is likely to be insensitive to the parameters of the models λ, and hence an inappropriate cost function.

In the next section we study the performance of the proposed approach for model design by minimizing the empirical error rate and compare it with that of the ML approach. In the latter approach, the parameter set of the model for the acoustic signal from each word is designed from the training data corresponding to that word. The estimation is performed by

$$\max_{\lambda_m} \sum_{n:w_n=m} \ln p(y_T(n) \mid m), \quad (12)$$

where local maximization is performed using the Baum algorithm [5]-[7]. The minimization of (8), or equivalently, the maximization of the empirical correct decision rate given by

$$\hat{P}_c(\lambda) \approx \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \frac{p(y_T(n) \mid m)}{\sum_{l=1}^{L} p(y_T(n) \mid l)}, \quad (13)$$

was performed using the steepest-descent method. The approximate gradient of $P_c(\lambda)$ with respect to λ was calculated as follows

$$\begin{aligned} \nabla_\lambda \hat{P}_c(\lambda) \approx & \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \frac{\nabla_\lambda p(Y_T(n) \mid m)}{\left\{ \sum_{l=1}^{L} p(Y_T(n) \mid l) \right\}^2} \\ & - \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \frac{p(Y_T(n) \mid m) \nabla_\lambda \sum_{l=1}^{L} p(Y_T(n) \mid l)}{\left\{ \sum_{l=1}^{L} p(Y_T(n) \mid l) \right\}^2} \end{aligned} \quad (14)$$

which can further be reduced to the following approximation,

$$\begin{aligned} \nabla_\lambda \hat{P}_c(\lambda) \approx & \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \nabla_{\lambda_m} p(Y_T(n) \mid m) \\ & \frac{\sum_{l=1}^{L} p(Y_T(n) \mid l) - p(Y_T(n) \mid m)}{\left\{ \sum_{l=1}^{L} p(Y_T(n) \mid l) \right\}^2} \\ & - \frac{1}{N} \sum_{m=1}^{L} \sum_{n:w_n=m} \sum_{\substack{l=1 \\ l \neq m}}^{L} \frac{p(y_T(n) \mid m) \nabla_{\lambda_l} p(y_T(n) \mid l)}{\left\{ \sum_{l=1}^{L} p(Y_T(n) \mid l) \right\}^2}. \end{aligned} \quad (15)$$

Both likelihoods and their gradients with respect to model parameters can efficiently be estimated using forward and backward probabilities, $\alpha_t(i)$ and $\beta_t(j)$, respectively. They can be calculated recursively, after Baum [6], for $1 \leq t \leq T$,

$$\alpha_t(j) = \left[\sum_{i=1}^{n} \alpha_{t-1}(i) a_{ij} \right] b_j(y_t) \quad (16)$$

and for $T-1 \geq t \geq 1$,

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij} b_j(y_{t+1}) \beta_{t+1}(j). \quad (17)$$

We can then efficiently evaluate the likelihood function,

$$p(y \mid m) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_t(i) a_{ij} b_j(y_{t+1}) \beta_{t+1}(j). \quad (18)$$

The models selected for the experiments are HMM's with Gaussian mixtures of the form

$$b_j(y_t) = \sum_{k=1}^{m} c_{jk} N(y_t, M_{jk}, U_{jk}). \quad (19)$$

HMM's consist of a set of n states and they are defined by state transition probabilities, a_{ij}, $1 \le i,j \le n$, mixture weights, c_{jk}, observation means, μ_{jkl}, and observation variances, u_{jkl}, $1 \le j \le n$, $1 \le k \le m$, $1 \le l \le d$, where each observation vector is a d-dimensional vector of spectral components.

The experiments described below only considered a subset of the HMM parameters in order to most efficiently demonstrate the properties of the MEE approach. Only the state transition matrix, mixture weights and the mean values have been reestimated in the gradient search. That was achieved by estimating the gradients of the likelihood function with respect to these parameters, which can be performed efficiently using the forward and backward probabilities,

$$\frac{\partial}{\partial a_{ij}}(P) = \sum_{t=1}^{T-1} \alpha_t(i) b_j(O_{t+1}) \beta_{t+1}(j), \quad (20)$$

$$\frac{\partial}{\partial c_{jk}}(P) = \left[\delta_{1j} \frac{\partial b_j}{\partial c_{jk}} \bigg|_{o_1} \beta_1(j) + \sum_{t=1}^{T-1} \sum_{i=1}^{n} \alpha_t(i) a_{ij} \frac{\partial b_j}{\partial c_{jk}} \bigg|_{o_{t+1}} \beta_{t+1}(j) \right], \quad (21)$$

and

$$\frac{\partial}{\partial m_{jkr}}(P) = \left[\delta_{1j} \frac{\partial b_j}{\partial m_{jkr}} \bigg|_{o_1} \beta_1(j) + \sum_{t=1}^{T-1} \sum_{i=1}^{n} \alpha_t(i) a_{ij} \frac{\partial b_j}{\partial m_{jkr}} \bigg|_{o_{t+1}} \beta_{t+1}(j) \right]. \quad (22)$$

We can thus obtain the gradients of the objective function used in the gradient search, which is the approximation of the probability of correct classification on the training data.

III. APPLICATION OF THE MEE APPROACH

The MEE approach was applied to recognition of isolated spoken utterances of the English E-set words (ie. B, C, D, E, G, P, T, V, Z) recorded using a conventional telephone handset over local telephone lines. The database used in the first two experiments consisted of utterances spoken by two male and two female speakers. In the first experiment, training of a set of multi-speaker models was performed using five utterances spoken by each male speaker and seven utterances spoken by the female speakers, for each word in the vocabulary. Testing was performed using ten utterances from each speaker (ie. 360 test tokens). In the second experiment, the testing and training data were reversed, allowing us to evaluate the effects of increasing the amount of training data on the performance of the MEE approach.

In the third experiment, a considerably larger database of speakers was used. The training data consisted of one repetition of each vocabulary word by fifty male and fifty female speakers. The testing data consisted of the same number of utterances, spoken by the same speakers at a different time. Again a set of multi-speaker word HMM's were created for each of the nine words in the E-set.

The speech signal in these experiments was first sampled at a 6667 Hz rate and analyzed using a sliding window. In the first two experiments a 30ms analysis window and a 10ms window shift were used. In the third experiment a 45ms analysis window with a shift of 15ms was used. For each vector y_t, corresponding to the analysis window centered at time t, the resulting spectral vector, $S(\omega, t)$, was represented using $d = 12$ cepstral coefficients, in the first two experiments, and $d = 10$ coefficients in the third experiment. The cepstral coefficients were computed from the linear prediction parameters of y_t and they were liftered using a standard bandpass lifter [13]. A set of d additional parameters were obtained by evaluating the differential cepstral coefficients (called the delta cepstral coefficients), ΔC_m which contain important information about the temporal rate of change of the cepstrum [14]. The combined cepstral and delta cepstral vectors form a set of observation vectors, O_t, which were used in the first two experiments described below. (Each observation vector consisted of $d = 24$ parameters in the first two experiments and $d = 20$ parameters in the third experiment.)

In all three experiments the HMM's had identical structure. For each word in the vocabulary a five state left-to-right HMM with two Gaussian mixture components per state was used. The covariance matrix of all Gaussian pdf's was assumed diagonal.

The steepest descent method was initialized using a set of HMM's estimated by the Baum algorithm [6]. The Baum algorithm was initialized by splitting each isolated utterance in the training data into five segments of equal duration, whose sample means and variances were used as a coarse initial estimate for the model parameters. The Baum reestimation was then iterated until the likelihood function reached a local maximum. The acoustic models obtained by the Baum algorithm were also used for defining the baseline performance (of an ML system) for comparison with the performance of the models obtained by the MEE estimation procedure.

The gradient search optimization of the linearly constrained nonlinear function, $P_c(\lambda)$, was performed using the optimization package, Modular In-Core Nonlinear Optimization System (MINOS), developed at Stanford University [15]. It solves this class of problems using a reduced-gradient algorithm [16] in conjunction with a quasi-Newton algorithm [17]. The implementation follows that described in Murtagh and Sunders [18].

The gradient search procedure provided intermediate acoustic models each iteration. These intermediate models were used in a series of recognition experiments to see how the performance on both the test data and the training data changed with each iteration.

The results obtained in Experiment 1 for the smallest training set are shown in Table 1.

EXPERIMENT 1.				
	ML		MEE	
	training data	test data	training data	test data
log likelihood	207847	315454	202892	312173
$P_c(\lambda)$	0.945		0.991	
accuracy (%)	94.4	73.6	99.1	70.0

Table 1. Results based on training on the small data set and testing on the medium size data set.

The recognition performance on the training data is very good, but the recognition performance on the test data is relatively poor. This result is probably due to insufficient training data. This type of behavior (ie. excellent performance on training data, poor performance on test data) occurred for all iterations of the gradient search estimation procedure. Thus, in spite of the improved accuracy on the training data obtained using the MEE approach, the resulting models did not improve the accuracy on the test data, but instead slightly reduced it.

In the second experiment the training was performed using the larger data set which was previously used as test data. This provided a more balanced result, where the differences in accuracy on the training and the testing data were reduced somewhat. The increase in the amount of the training data provided better word models enabling the MEE approach to improve the accuracy on the test data as well as on the training data, as shown in Table 2.

EXPERIMENT 2.				
	ML		MEE	
	training data	test data	training data	test data
log likelihood	343798	192135	338330	184899
$P_c(\lambda)$	0.925		0.975	
accuracy (%)	92.5	75.0	97.0	76.9

Table 2. Results based on training on the medium size data set and testing on the small data set, obtained after 85 iterations.

The highest accuracy on the test data was not achieved with the same set of models that had highest accuracy on the training data.

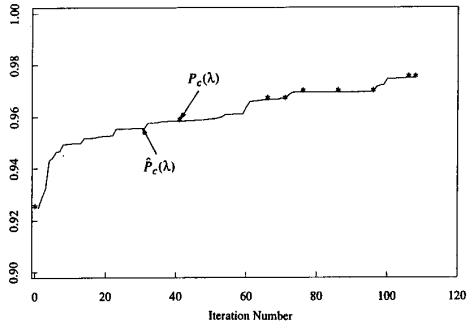

Figure 1. Objective function and recognition accuracy on the training data in Experiment 2.

To illustrate this point, Figure 1 shows the value of the objective function, and the resulting training set recognition accuracy as a function of the iteration number in the model design procedure. An excellent match between the value of the objective function, $P_c(\lambda)$, and the true probability of correct classification can be seen in the figure. A plot of the recognition accuracy on the test data, as a function of the iteration number in the model design procedure, can be seen in Figure 2.

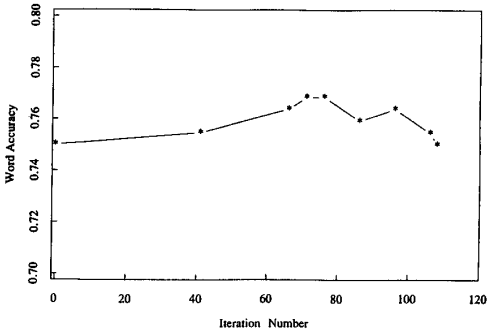

Figure 2. Recognition accuracy on the test data at different stages during the gradient search estimation in Experiment 2.

The recognition accuracy first increases from the ML score until a peak is reached at 76.9%. Then the accuracy starts dropping until it reaches the same value as obtained with the ML estimation set of models. (The gradient search estimation was then terminated.) It can also be seen from Tables 1 and 2 that MEE estimation results in a lower likelihood score since it attempts to maximize recognition accuracy without any explicit concern for the change in the likelihood function.

The results obtained from the third experiment performed on the much larger set of talkers are shown in Table 3.

EXPERIMENT 3.				
	ML		MEE	
	training data	test data	training data	test data
log likelihood	906786	903410	902169	896058
$P_c(\lambda)$	0.681		0.755	
accuracy (%)	68.2	53.4	75.0	56.4

Table 3. Results based on training and testing on the largest data set, obtained after 90 iterations.

The differences in performance on the training and the test data are somewhat smaller than in the previous experiments with the reduction in the error rate on the test data after performing MEE reestimation (3%) being about half the reduction in the error rate on the training data (7%). The differences between the value of the objective function and the actual word recognition accuracy on the training data, as shown in Figure 3, are more pronounced here because of the lower accuracy of the recognizer.

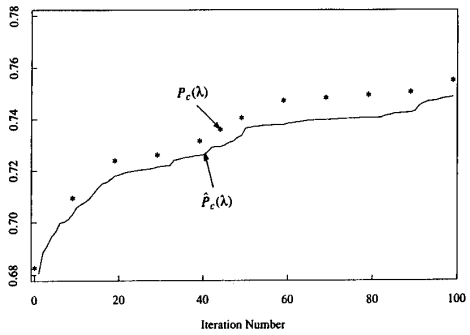

Figure 3. Objective function and recognition accuracy on the training data in Experiment 3.

Figure 4 shows the resulting accuracy on the test data as a function of the iteration number during the gradient search.

The accuracy on the test set is highly non-monotonic with the iteration number; hence in order to achieve the best recognition results, careful monitoring of performance during the steepest descent estimation is required. (This may not always be practical in actual implementations.)

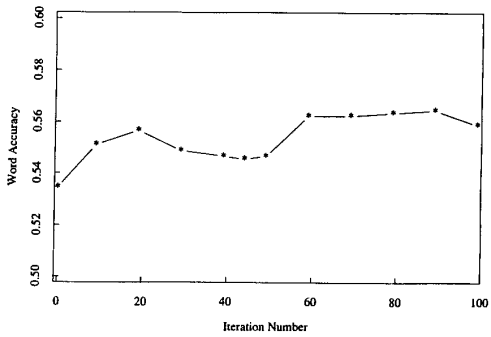

Figure 4. Recognition accuracy on the test data at different stages during the gradient search estimation in Experiment 3.

IV. CONCLUSIONS

A new technique for estimating HMM parameters based on minimizing empirical error rate has been proposed and studied. The MEE approach is more directly related to the goal of reducing classification error rate than other approaches currently used for estimating model parameters. It is implemented using steepest descent estimation by minimizing a differentiable approximation of the error function and thus maximizing recognition accuracy on the training data. An increase in accuracy on the training data *does not guarantee* an improvement in recognition rate on an independent set of test data. The experiments described here show that it is necessary to have a large amount of training data to adequately characterize the source and thus make the MEE approach usable. When sufficient training data is available, the MEE approach was shown to be capable of reducing the error rate on both training and test data as compared to the error rate of the ML training approach.

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. on ASSP*, p. 7-13, April 1988.

[4] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. on PAMI*, vol. 5, no. 2, pp. 179-190, March 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. Stat.*, 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. IT*, vol. 28, 1982.

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

[9] P.F. Brown, *The Acoustic-Modeling Problem in Automatic Speech Recognition*. Ph.D Thesis, Carnegie Mellon Univ., 1987.

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

[11] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. IEEE Int. Conf. ASSP*, pp. 49-52, 1986.

[12] Y. Ephraim, and L.R. Rabiner, "On the Relations Between Modeling Approaches for Speech Recognition," to appear in *IEEE Trans. on IT*.

[13] B.-H. Juang, L.R. Rabiner, and J.G Wilpon, "On the Use of Band-pass Liftering in Speech Recognition," IEEE Trans. on ASSP, Vol. 35, pp. 947-954, 1987.

[14] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.* 80(4), pp. 1016-1025, 1986.

[15] B.A. Murtagh and M.A. Saunders, "MINOS 5.1 User's Guide," Technical Report SOL 83-20R, Stanford University, 1987.

[16] P. Wolfe, "The reduced-gradient method," unpublished manuscript, Rand Corporation, 1962.

[17] W.C. Davidon, "Variable metric methods for minimization," A.E.C. R&D Report ANL-5990, Argonne National Laboratory, 1959.

[18] B.A. Murtagh and M.A. Saunders, "Large-scale linearly constrained optimization," *Math. Prog.* 14, pp. 41-72, 1978.