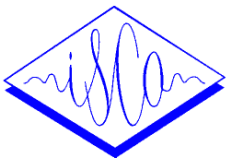

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

DISCRIMINATIVE TRAINING OF HMM BASED SPEECH RECOGNIZER WITH GRADIENT PROJECTION METHOD

Qiang Huo[†‡] and Chorkin Chan[†]

[†]Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
[‡]ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai, Seika-cho Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

In this paper, in order to examine the viability and characteristics of our previously proposed *gradient projection method* for HMM training, the method is applied to discriminatively train the DHMM parameters with the objective of minimizing the recognition error rate and a series of experiments have been conducted. The experiments involve speaker independent recognition of the highly confusable E-set of the English alphabet. Three kinds of training schemes, namely, *batch training*, *sequential block training* and *corrective training* are studied. The experimental results show that the minimum recognition error objective function is a viable formulation that can be optimized by the gradient projection method.

1. INTRODUCTION

For hidden Markov model (HMM) based speech recognizer, generally speaking, the purpose of training HMM λ is to yield a decoder of the lowest possible recognition error rate. This objective is achieved by maximizing an objective function $R(\lambda)$. There are thus two problems to consider. The first is to determine a meaningful objective function such that, if $R(\bar{\lambda}) > R(\lambda)$, then $\bar{\lambda}$ produces a better decoder than that by λ. Once a function $R(\lambda)$ has been chosen, the second problem (the estimation problem) is to find the parameter set $\bar{\lambda}$ which maximizes it. Traditionally, maximum likelihood (ML) estimation and its variants have long been the preferred HMM training methods due to the existence of corresponding training procedures which allow one to efficiently find the locally optimal values of the model parameters according to the adopted training criteria. The optimality of these methods is conditioned on the correct choice of the model formulation. Nevertheless, inaccuracies in modeling the speech signals in reality may lead to ML models that do not maximize the recognition accuracy. This concern motivates many recent efforts to consider alternative design criteria.

Recently, alternatives to ML training such as, "maximum mutual information (MMI) training" [2], "minimum discrimination information (MDI) training" [3], "corrective training" [1] and other discriminative training methods [7, 6] with the objective of lower recognition error rate have been proposed. All these methods involve objective functions that do not satisfy the conditions assumed by the conventional Baum-Welch formulation. Gopalakrishnan *et al.* [4] derived a Baum-like reestimation formula for discrete HMMs (DHMM) which applies to rational objective functions. Later, Normandin *et al.* [8] extended it to suit continuous density HMMs. However, there are still many cases of HMM training for speech recognition where the conditions required by the algorithm are apparently not satisfied.

By looking at the training of HMMs as a general constrained optimization problem with linear constraints, we have previously proposed a *gradient projection method* (GPM) to solve for the "optimal" values of the HMM parameters and this formulation is applicable to any analytic objective function [5]. Recently, Katagiri *et al.* [7, 6] proposed a new smooth objective function which approximates the empirical error rate for the training sample set. No existing Baum-like algorithm can be applied for the optimization of such a function. They used the so called *generalized probabilistic descent* (GPD) method to solve their problem. In this paper, this new discriminative criterion is adopted for DHMM training to verify the viability of our proposed GPM to solve this particular constrained optimization problem.

2. THE GRADIENT PROJECTION METHOD

The main idea of the GPM is to search along the projection of the gradient of the objective function on the constraint space for a local maximum. So, the GPM is essentially a steepest ascent method in the subspace defined by "the active constraints" of HMM parameters. Readers are referred to [5] for details of the GPM algorithm.

3. THE MEC OBJECTIVE FORMULATION

Let's consider a collection of M DHMMs, $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_M)$, where λ_m denotes the set of parameters of the m-th DHMM. Let $\mathbf{x}^{(m,n)}$ denote the nth training observation sequence of length $T^{(m,n)}$ associated with model m, and each model has W_m such observation sequences. The objective function for discriminative training is derived according to the following *minimum error classification* (MEC) formulation which is a three-step procedure [7, 6]. The three-step definition emulates the classification/recognition operation as well as the performance evaluation, particularly in terms of classification errors, in a smooth function form.

The first step of the formulation is to prescribe an appropriate discriminant function $g_i(\mathbf{x};\Lambda)$ which is used by the classifier to make its decision for each input $\mathbf{x}$ by choosing the largest of the discriminants evaluated on $\mathbf{x}$ which is often generically stated as

$$C(\mathbf{x}) = C_i, \quad \text{for } g_i(\mathbf{x};\Lambda) = \max_j g_j(\mathbf{x};\Lambda), \qquad (1)$$

where $C(\cdot)$ denotes a classification operation. The ith

discriminant function is defined as:

$$g_i(\mathbf{x};\Lambda) = \frac{1}{T}\ln P(\mathbf{x}|\lambda_i), \quad (2)$$

where $P(\mathbf{x}|\lambda_i)$ is the probability of the input utterance $\mathbf{x}$ given model λ_i and T is the length of $\mathbf{x}$. In order to balance the contributions of training utterances with different length in the final objective function, the normalized log-likelihood are adopted here as the discriminant function.

A misclassification measure is then introduced in the second step to embed the decision process in a function form. Among many possibilities, the misclassification measure for each class i is defined by

$$d_i(\mathbf{x};\Lambda) = -g_i(\mathbf{x};\Lambda) + \ln\left[\frac{1}{M-1}\sum_{j\neq i} e^{g_j(\mathbf{x};\Lambda)\eta}\right]^{\frac{1}{\eta}}, \quad (3)$$

where η is a positive value.

The third step is to define the smoothed loss function $l_i(\mathbf{x};\Lambda)$ of the misclassification measure for each class i. One possibility is to choose

$$l_i(\mathbf{x};\Lambda) = l_i(d_i(\mathbf{x};\Lambda)) = \frac{1}{1+e^{-\xi d_i(\mathbf{x};\Lambda)}}, \quad (4)$$

where ξ is a positive value. Thus, for any unknown $\mathbf{x}$, the classifier performance is measured by

$$l(\mathbf{x};\Lambda) = \sum_{i=1}^{M} l_i(\mathbf{x};\Lambda)1(\mathbf{x}\in C_i), \quad (5)$$

where $1(\cdot)$ is an indicator function:

$$1(\hbar) = \begin{cases} 1 & \text{if } \hbar \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

and C_i is used to denote both the class and the data set.

At this point, we define the objective function of discriminative training as the following *empirical average cost* for the entire training data set:

$$L(\Lambda) = \frac{1}{W}\sum_{m=1}^{M}\sum_{n=1}^{W_m} l_m(\mathbf{x}^{(m,n)};\Lambda), \quad (7)$$

where $W = \sum_{m=1}^{M} W_m$ is the total number of training tokens. By controlling parameters η, ξ and minimizing this *empirical average cost*, one can have an accurate approximation to the minimization of the classification error probability. Due to the fact that our GPM formulation is for maximization, the actual objective function adopted is $R(\Lambda) = -L(\Lambda)$. Given the above objective function, one can now apply the GPM to discriminatively adjust the DHMM parameters Λ which effectively minimize the cost function.

The training scheme presented above represents one extreme that the HMM parameters are adjusted after the presentation of the entire training data set using the objective function defined over the whole training set. It will be referred to as the *batch training* scheme. In fact, the updating schedule can be defined arbitrarily. Another extreme like the GPD method in [7, 6] is that the HMM parameters are updated upon presentation of each training token just like what the many so called stochastic descent (or ascent) methods do in neural network training.

With these considerations in mind, the so-called *sequential block training* scheme has been experimented with. Each time, a block of training data is presented. The HMM parameters are then updated after one iteration of the GPM with the objective function defined on this data block. If a complete pass through the training data set is called an *epoch*, then, the HMM parameters are updated several times over an *epoch*. Furthermore, a so-called *corrective training* method has also been attempted. Let Ξ represent the subset of the training tokens which are incorrectly recognized based on the recognizer with the current HMM parameters, the following objective function

$$R_c(\Lambda) = -\frac{1}{W_c}\sum_{\mathbf{x}_i\in\Xi}\sum_{m=1}^{M} l_m(\mathbf{x}_i;\Lambda)1(\mathbf{x}_i\in C_m) \quad (8)$$

is defined, where W_c denotes the number of training tokens in Ξ. Once an error set of training tokens is given and the objective function defined, several (particularly 5 in our experiments) iterations of the GPM are conducted and the HMM parameters Λ are updated accordingly. With the new HMM parameters obtained, a new error set Ξ can be identified and the algorithm iterates accordingly.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

A series of experiments have been conducted to examine the characteristics of the GPM. The experiments involve the recognition of the highly confusable E-set of the English alphabet, namely, b, c, d, e, g, p, t, v, and z. The E-set utterances from the OGI ISOLET database produced by 150 talkers (75 females and 75 males) are used. Each talker utters each of the letters twice. Among them, speech tokens from 120 talkers are used for training and the remaining tokens from the other 30 talkers for testing. To imitate the effect of a telephone bandwidth, the speech data originally sampled at 16 KHz are lowpass-filtered at 3.3KHz and down-sampled to 8 KHz. The feature vectors used in this study consist of 12 bandpass-liftered LPC-derived cepstral coefficients with a 30ms frame length and a 10ms frame shift. Throughout the experiments, each letter in the vocabulary is modeled by a single left-to-right 5-state DHMM with arbitrary state skipping. A 256-vector codebook is generated from the training tokens of 120 talkers by using the LBG algorithm with an Euclidean distortion measure and used in all experiments. For consideration of future experiments, the training tokens of all 26 English alphabet in ISOLET corpus are actually used to form the codebook. This will reduce the recognition rate somewhat compared to the case when the codebook is generated strictly from the E-set training tokens. In recognition, the decision rule assigns an unknown letter according to the highest forward-backward probability.

Although there are many alternatives to perform a crude line search in a GPM, the following *ad hoc* method is adopted in all experiments of this study. For a DHMM, the step length $\bar{r}$ in Step 4 of the basic algorithm of the GPM presented in [5] can always be computed and has a finite value. If the objective function computed at the maximum step length $\bar{r}$ along the current search direction has a better value than the one at the current point, then a step of $\bar{r}$ is taken without a thorough line

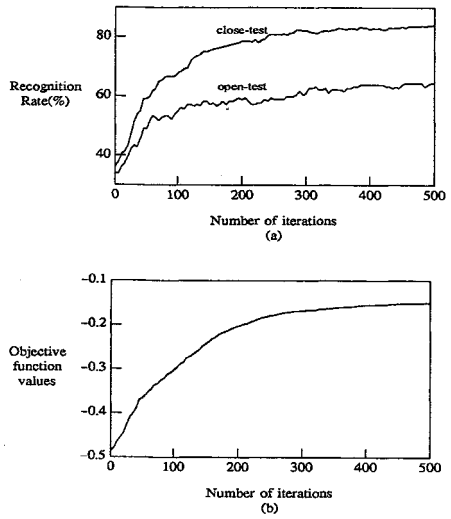

Figure 1: Learning curves of batch training with random initial models in "b, d, g" task. (a) close-test and open-test recognition rate. (b) objective function values.

search. Otherwise, the objective function is evaluated at 5 equally spaced locations (except the current one) along the search direction between the current position and that at the maximum step size $\bar{\tau}$. The location with the best objective function value is selected to approximate a line search. This *ad hoc* method works well in the particular experimental setups here.

4.2. Experiments on the "b, d, g" Task

To examine the characteristics of the GPM under different complexity (in terms of the number of HMM parameters and training tokens) of the recognition task, a relatively small task, *viz.*, speaker independent recognition of a subset of the E-set, *viz.*, the English letters "b, d, g", is considered first. Three sets of experiments are conducted. The first set of experiments is to discriminatively train the DHMMs with the GPM in a *batch training* mode. The parameters η and ξ used in equations (3) and (4) are respectively set at 2.0 and 4.0. To examine the effects of different initial models, two experiments are conducted. One is to start the training process from the well-trained initial models which themselves are trained from random initial models with the conventional Baum-Welch algorithm. With these initial models, the recognizer has a recognition rate of 66.11% on the testing set (83.33% on the training set). After 150 iterations of *batch training*, the recognizer achieves a recognition rate of 74.44% on the testing set (86.39% on the training set). About 25% error rate reduction is achieved by discriminative *batch training*. The second experiment is to perform discriminative *batch training* directly from random initial models. We plot in Figure 1 the objective function values, and the close and open test recognition rates at certain iterations to illustrate the performance improvement as a function of the training process. Figure 1 clearly shows that the performance improves very slowly due to the intrinsic linear convergence property of the GPM. For example, only after more than 240 iterations of *batch training*, an objective function value similar to that obtained by a recognizer of traditionally well-trained models is achieved. These traditional models are trained after tens of iterations with the Baum-Welch algorithm. Here the belief that gradient-based methods are not as efficient as the Baum-Welch one [4] is once again confirmed. After 455 iterations, the recognition rate rises to 64.44% on the testing set (83.19% on the training set) which is even worse than that obtained by the well-trained conventional Baum-Welch models, although the objective function value keeps on increasing. If the training process continues, the close-test recognition rate further improves and it can reach a rate similar to that of batch training from well-trained initial models. However, the open-test results have no further improvement. This reflects the fact that from different initial models, the algorithm converges to different local maximum points, so that good initial models are essential to good system training with this kind of parameter training algorithms. This also suggests that the proposed algorithm is most attractive for final "tune-up" and will usually be bootstrapped from well-trained initial models trained with other methods such as the traditional Baum-Welch algorithm.

Table 1: Performance comparison (% correct) of several training methods for speaker independent recognition of the English letters "b, d, g" (BW: Baum-Welch training; BT: batch training; SBT1: sequential block training with block size 60; SBT2: sequential block training with block size 120; CT: corrective training)

Schemes	BW	BT	SBT1	SBT2	CT
close-test	83.33	86.39	86.81	86.53	88.89
open-test	66.11	74.44	76.11	75.56	76.11

The second set of experiments is to discriminatively train the DHMMs with the GPM in a *sequential block training* mode starting from the traditionally well-trained initial models. The parameters η and ξ are the same as that used in the *batch training* mode. Two arbitrary block size schemes are studied. In the first experiment, each data block consists of 60 training tokens in total with 20 tokens per letter. Thus the HMM parameters are updated 12 times in each *epoch*. In the second experiment, the block size is doubled, i.e., each letter has 40 training tokens. The HMM parameters are updated 6 times in each *epoch*. When the block size is 60 tokens, after 4 *epoch* of *sequential block training* (equivalently 48 times of updating of parameters), the recognizer achieves a recognition rate of 76.11% on the testing set (86.81% on the training set). When the block size is 120 tokens, after 5 *epoch* of *sequential block* training (equivalently 30 times of updating of parameters), the recognizer achieves a recognition rate of 75.56% on the testing set (86.53% on the training set). In terms of recognition rate improvement, there is not much difference between the two block sizes experimented. In terms of computation involved to obtain similar performances, the case of smaller block seems to be superior. The optimal choice of block size is completely problem and data dependent and it can only be determined by extensive experiments. Experimental results here also show that the *sequential block training* scheme is more efficient than the *batch training* one in terms of both performance improvement and computation requirement involved.

The third set of experiments is to train the DHMMs with the GPM in a *corrective training* mode. The parameter η remains the same and ξ changes to 1.0. Corrective training is started with the well-trained initial models used in the *batch training* mode discussed above. After 85 iterations of *corrective training*, the recognizer achieves

Table 2: Performance comparison (% correct) of several training methods for speaker independent recognition of the English E-set vocabulary (BW: Baum-Welch training; BT: batch training; SBT1: sequential block training with block size 180; SBT2: sequential block training with block size 360; CT: corrective training)

Schemes	BW	BT	SBT1	SBT2	CT
close-test	70.97	73.98	74.31	73.94	71.67
open-test	51.48	53.89	53.70	53.15	52.22

a recognition rate of 76.11% on the testing set (88.89% on the training set). In other words, about 30% error rate reduction is achieved by *corrective training*. In this relatively small recognition task, the *corrective training* scheme seems efficient in performance improvement also.

In summary, all three training schemes, namely, *batch training, sequential block training* and *corrective training* work well in the "b, d, g" task studied here. The *sequential block training* method seems to be a good compromise between the two extremes of *batch training* and *on-line updating*. The close and open test recognition rates with these three training methods and the traditional Baum-Welch algorithm are summarized in Table 1 for comparison.

4.3. Experiments on the "E-set" Task

After an initial success of the GPM with the "b, d, g" task, the proposed method is then applied to the E-set task. The three sets of experiments discussed above are repeated with the E-set. Similar facts observed in the "b, d, g" task are repeated here. The close and open test recognition rates with these three training methods and the traditional Baum-Welch algorithm are summarized in Table 2 for comparison. Compared with the performance improvement in the "b, d, g" task, relatively little improvement is achieved here. It is also noted that in the E-set task studied here, the *corrective training* scheme seems less efficient than the *batch training* and *sequential block training* schemes. This is due to the fact that in *corrective training* only those incorrectly recognized training tokens are used in the training process. The more difficult and complex the task, the more likely fluctuations will appear. One should therefore be cautious in adopting the *corrective training* scheme in real applications.

5. SUMMARY

In this paper, in order to examine the viability and characteristics of our previously proposed *gradient projection method* for HMM training, the method is applied to discriminatively train the DHMM parameters with the objective of minimizing the recognition error rate and a series of experiments have been conducted. The experiments involve speaker independent recognition of the highly confusable English letters "b, d, g" and the complete E-set of the Latin alphabet. Three kinds of training schemes, namely, *batch training, sequential block training* and *corrective training* are studied. In the "b, d, g" task, all three training schemes work well which in the E-set case, the *corrective training* scheme is less efficient than the other two methods. The *sequential block training* method is shown to be a good compromise between the two extremes of *batch training* and *on-line updating*. The experimental results also show that good initial models are critical to system performance due to the nature of the algorithm converging only to a local maximum of the objective function. This fact suggests that the GPM is most attractive for final "tune-up" and will usually be bootstrapped from initial models trained with other methods such as the traditional Baum-Welch reestimation algorithm. Furthermore, experimental results verify that the minimum recognition error objective is a viable formulation that can also be optimized by the GPM. This general optimization technique is not only a viable alternative to the classical Baum-Welch algorithm, it can also serve as a preferable method in general HMM training when the objective function and constraints fail to satisfy the conditions demanded by the Baum-Welch reestimation formulas. Due to the existence of this kind of gradient based optimization methods, more flexible modeling of speech signals and more sophisticated training of model parameters for speech recognition become viable.

ACKNOWLEDGEMENT

The authors would like to thank Mr. Z.-Q. Ma for helping to conduct the related experiments. The first author would also like to thank Dr. C.-H. Lee for his continuous discussions on the experimental aspects, and Drs. Y. Yamazaki & Y. Sagisaka of ATR-ITL for their support of his preparing this manuscript.

REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. De Souza & R. L. Mercer. *A new algorithm for the estimation of hidden Markov model parameters.* Procs. ICASSP, 1988, pp.493-496.

[2] P. F. Brown. *The Acoustic Modeling Problem in Automatic Speech Recognition.* PhD thesis, Department of Computer Science, Carnegie Mellon University, 1987.

[3] Y. Ephraim, A. Dembo & L. R. Rabiner. "A minimum discrimination information approach for hidden Markov modeling". *IEEE Trans. on Information Theory*, Vol. 35, pp.1001-1013, 1989.

[4] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas & D. Nahamoo. "An inequality for rational functions with applications to some statistical estimation problems". *IEEE Trans. on Information Theory*, Vol. 37, pp.107-113, 1991.

[5] Q. Huo and C. Chan. "The gradient projection method for the training of hidden Markov models". *Speech Communication*, Vol. 13, pp.307-313, 1993.

[6] B.-H. Juang & S. Katagiri. "Discriminative learning for minimum error classification". *IEEE Trans. on Signal Processing*, Vol. 40, pp.3043-3054, 1992.

[7] S. Katagiri, C.-H. Lee & B.-H. Juang. *New discriminative training algorithms based on the generalized probabilistic descent method.* Proc. IEEE Workshop Neural Networks for Signal Processing, 1991, pp.299-308.

[8] Y. Normandin, R. Cardin & R. D. Mori. "High-performance connected digit recognition using maximum mutual information estimation". *IEEE Trans. on Speech and Audio Processing*, Vol. 2, pp.299-311, 1994.