DETERMINISTICALLY ANNEALED DESIGN OF SPEECH RECOGNIZERS AND ITS PERFORMANCE ON ISOLATED LETTERS

Ajit Rao, Kenneth Rose, and Allen Gersho

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106.

ABSTRACT

We attack the general problem of HMM-based speech recognizer design, and in particular, the problem of isolated letter recognition in the presence of background noise. The standard design method based on maximum likelihood (ML) is known to perform poorly when applied to isolated letter recognition. The more recent minimum classification error (MCE) approach directly targets the ultimate design criterion and offers substantial improvements over the ML method. However, the standard MCE method relies on gradient descent optimization which is susceptible to shallow local minima traps. In this paper, we propose to overcome this difficulty with a powerful optimization method based on deterministic annealing (DA). The DA method minimizes a randomized MCE cost subject to a constraint on the level of entropy which is gradually relaxed. It may be derived based on information-theoretic or statistical physics principles. DA has a low implementation complexity and outperforms both standard ML and the gradient descent based MCE algorithm by a factor of 1.5 to 2.0 on the benchmark CSLU spoken letter database. Further, the gains are maintained under a variety of background noise conditions.

1. INTRODUCTION

The recognition of spoken letters is an important subproblem in the design of spelled name recognition systems [1] which are used in applications such as automatic car navigation, automated directory assistance, voice activated call forwarding etc. Accurate name recognition in spelled name recognition systems is highly conditioned on the accurate recognition of individual letters. However, the task of english letter recognition is known to be challenging due to the high confusability of the alphabet. In particular, utterances of letters in the subsets $\{b, c, d, e, g, p, t, v, z\}$ (E-set), $\{a, k, j\}$, $\{f, s, x\}$, $\{i, r, y\}$ and $\{m, n\}$ are often confused with each other. The difficult nature of the recognition task is further aggravated in real-world situations by the presence of background noise and in the case of telephone-based systems, by the additional presence of channel distortion.

Here, we consider a standard recognition system based on hidden Markov models (HMM). The main contribution of this paper is a powerful new HMM design method that improves the recognizer's classification rate compared to standard design methods. The fundamental approach is however, not restricted to HMM systems and can be applied to the design of other pattern classifiers including neural network based speech recognition systems.

1.1. HMM-based speech recognition

The HMM paradigm is widely used in conventional speech recognition systems. The design of HMM-based recognizers has traditionally been based on maximum likelihood (ML) modeling of speech. However, the ultimate objective is not to model but rather to minimize the error rate of the classifier. In the standard ML approach, a labelled training set of classified speech patterns is divided into subsets of identically labelled patterns and the HMM corresponding to each label is designed *independently* from the corresponding subset via maximum likelihood estimation of the model parameters. Note that the maximum likelihood design objective does not minimize classification errors unless the HMM structure is the precise model for the speech. Further, the ML objective differs significantly from the optimal classification objective when the training set is short.

This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Inc., Advanced Computer Communications, Cisco Systems, Inc., DSP Group, Inc., DSP Software Engineering, Inc., Fujitsu Laboratories of America, Inc., General Electric Company, Hughes Electronics Corp., Intel Corp., Nokia Mobile Phones, Qualcomm, Inc., Rockwell International Corp., and Texas Instruments, Inc.,

The inadequacies of the ML design approach have been pointed out by several researchers [2, 3, 4]. who noted that significant gains in accuracy and robustness are possible by directly targeting the minimum classification error (MCE) objective. However, MCE optimization is difficult for two main reasons. The first and most widely recognized, is that unlike the ML cost, the classifier's error rate is a piecewise constant function of the HMM parameters. This implies that gradients with respect to the parameters vanish almost everywhere (an infinitesimal change in parameter values will not change the classification of any training pattern). Consequently, one cannot directly use gradient-based optimization of MCE. To address this problem the Generalized Probabilistic Descent (GPD) method [5] has been suggested. GPD replaces the classification error cost surface with a smooth approximation, thus allowing the application of gradient-based optimization. However, as we will demonstrate later, even if the cost surface is smoothed, it might still be highly complex and riddled with shallow local minima which tend to trap gradient descent algorithms. Another difficulty in MCE design is that it entails joint optimization of all the HMM parameters, which is computationally complex, even for an off-line design.

To overcome the optimization difficulties of the standard MCE approach, we propose here, an alternative method based on the technique of deterministic annealing (DA). DA was first proposed in the context of clustering [6], later extended to solve structurally constrained clustering problems such as the design of pattern classifiers [7] and regression functions [8], and recently applied to of time series classification [9]. We will show here that the DA method for HMM classifier design offers substantial gains by combining the right criterion of MCE with the optimization power of DA.

2. THE HMM DESIGN PROBLEM

In a typical isolated-word speech recognition system, we are given a training set

$$\mathcal{T} \equiv \{(\mathbf{y}_1, c_1), (\mathbf{y}_2, c_2), ..(\mathbf{y}_N, c_N)\} \tag{1}$$

of labelled *patterns*. The pattern $\mathbf{y}_i$ corresponds to an utterance of the word, c_i from the dictionary, $\mathcal{C} \equiv \{1, 2, ..M\}$. The vector $\mathbf{y}_i$ consists of a sequence of l_i observations. Our method allows for both discrete and continuous valued observations.

The HMM classifier system is specified by a set of HMMs $\{H_j,\ j = 1, 2, \cdots, M\}$, one per word in the dictionary. HMM H_j is specified by the parameter set Λ_j which is composed of the state transition probabilities, state-dependent output distributions and the initial probabilities of the states. In this paper, we consider an HMM system based on the "best path" discriminant [1]. Given a training sequence $\mathbf{y}_i$, for each HMM H_j, we define the discriminant $d_j(\mathbf{y}_i)$ as the log likelihood (based on the HMM model) of the most likely state sequence,

$$d_j(\mathbf{y}_i) = \max_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} l(\mathbf{y}_i, \mathbf{s}, H_j). \tag{2}$$

Here, $\mathcal{S}_{l_i}(H_j)$ is the set of all l_i length state sequences in H_j and $l(\mathbf{y}_i, \mathbf{s}, H_j)$ is the log likelihood of a particular state sequence $\mathbf{s}$. The output of the classifier is the word corresponding to the HMM with the highest discriminant:

$$C(\mathbf{y}_i) = arg \max_j d_j(\mathbf{y}_i). \tag{3}$$

This classification system can be viewed as a competition between paths. The observation is ultimately labeled by the class index of the HMM to which the winning path belongs. The classifier design problem can be stated as the joint optimization of the the HMM parameters $\{\Lambda_j\}$ to minimize the empirical misclassification rate measured over the training set,

$$\min_{\{\Lambda_j\}} P_e = 1 - \frac{1}{N} \sum_{i=1}^{N} \delta(C(\mathbf{y}_i), c_i). \tag{4}$$

Here δ is the error indication function: $\delta(u, v) = 1$ if $u = v$ and 0 otherwise.

The cost function of (4) is a piecewise constant function of the HMM parameter set thus making naive gradient-based optimization impossible. Although the GPD method replaces this cost function by a continuously differentiable function which is amenable to gradient descent, in practice GPD may easily get trapped in shallow local minima of the cost surface.

3. DETERMINISTIC ANNEALING

The deterministic annealing algorithm is based on the concept of a *random classification*. The idea is to replace the "best path" classifier by a random classifier during the design. Given an observation $\mathbf{y}_i$, the random classifier chooses from the set of all state sequences in all the HMMs, a random winning state sequence, $\mathbf{s}$

[1] Our design method can be easily modified to the case where the discriminant is obtained by appropriate averaging of the likelihood over all paths.

in HMM H_j with a probability obeying the Gibbs law:

$$P(\mathbf{s}, H_j|\mathbf{y}_i) = \frac{e^{\gamma l(\mathbf{y}_i, \mathbf{s}, H_j)}}{\sum_{j'} \sum_{\mathbf{s}' \in \mathcal{S}_{l_i}(H_{j'})} e^{\gamma l(\mathbf{y}_i, \mathbf{s}', H_{j'})}}. \quad (5)$$

The parameter, γ controls the "fuzziness" of the distribution. For $\gamma = 0$, the distribution over paths is uniform. For finite, positive values of γ, the Gibbs distribution assigns higher probabilities of winning to state sequences with higher log likelihood scores, $l(\mathbf{y}_i, \mathbf{s}, H_j)$. In the limiting case of $\gamma \rightarrow \infty$, the random classification rule reverts to the non-random "best path" classifier of (2), which selects with probability one, the path with the highest log likelihood. The Gibbs parametric form is not arbitrary, and can be derived in a systematic manner from information-theoretic principles [9]. At this point, we re-emphasize that the random classifier paradigm is adopted only during design - in the limit, the DA algorithm is designing a regular non-random HMM-based classifier.

The expected error rate (over the training set) of the random classifier is given by:

$$< P_e >= 1 - \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_{c_i})} P(\mathbf{s}, H_j|\mathbf{y}_i). \quad (6)$$

Simple minimization of the expected error rate of (6) with respect to all the HMM parameters and the scale parameter γ is possible although such a method (somewhat like GPD) is susceptible to shallow local minima traps on the cost surface. We adopt instead, an "annealing" strategy to overcome the poor local minima problem. It is based on an entropy-constrained formulation: Rather than minimize the misclassification cost ($< P_e >$) of the random classifier as is, we minimize it while enforcing a constraint on the Shannon entropy (a measure of randomness),

$$H = -\frac{1}{N} \sum_i \sum_j \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} P(\mathbf{s}, H_j|\mathbf{y}_i) \log P(\mathbf{s}, H_j|\mathbf{y}_i). \quad (7)$$

We thus optimize the HMM parameters so as to minimize the expected error rate $< P_e >$ while constraining the randomness to a prescribed entropy level, $H = \hat{H}$. We then gradually lower the entropy level while repeating the optimization process. The constrained optimization problem of minimizing $< P_e >$ at a given entropy level is equivalent to the unconstrained Lagrangian minimization:

$$\min_{\{\Lambda_j\}, \gamma} L =< P_e > -TH \quad (8)$$

where T is the corresponding Lagrange parameter. The parameter, T, is gradually reduced from a high value to zero, while tracking the minimum of L. As $T \rightarrow 0$, the optimization reduces to the unconstrained minimization of $< P_e >$ which forces $\gamma \rightarrow \infty$ leading to the optimal non-random maximum discriminant classifier. The gradual reduction of T is central to the ability of the algorithm to avoid shallow local minima on the cost surface. We refer to the Lagrange parameter T as the *temperature* because of interesting connections to statistical physics. The process of reducing T to zero is similar in principle to the phenomenon of annealing in physical systems. For more insights into the physical analogy, see [6, 7, 8].

The minimization of the Lagrangian cost function L is achieved by a series of gradient descent steps at each temperature. An important aspect of the proposed method is the discovery of an efficient forward-backward algorithm to determine the gradient parameters for the optimization. The complexity of DA scales similarly to the maximum likelihood method with respect to the number of states and training vectors. Details of the algorithm are omitted for brevity, but will be presented at the conference.

4. EXPERIMENTAL RESULTS

In this section, we report the results of our experiments on the recognition of isolated english letters. In these experiments, we fixed the front-end processing of a discrete observation HMM system (both feature set and feature quantization codebook) and compared design methods to optimize the HMM classifier. Three methods were considered: (i) standard maximum likelihood (ML) (ii) Generalized Descent (GD) [2] and (iii) deterministic annealing (DA).

The dataset, which is a part of the ISOLET database from CSLU [3], consists of english letters spoken by 60 speakers (30 male and 30 female). Every speaker uttered each letter once and for each utterance, four noise-corrupted versions were obtained by adding synthetic white noise, recorded car noise, computer fan noise and air-conditioner noise to the clean speech. The speech which was sampled at 16 KHz was divided into frames of 512 samples. Consecutive frames overlap by 256 samples. In each frame, a 28 dimensional feature

[2] The GD approach is in principle, similar to GPD as it smooths the misclassification cost and then applies gradient descent. As it is also a degenerate case of DA, we have chosen to use it for comparison, while maintaining all auxillary parameters identical.

[3] Information on the ISOLET and how to obtain it is available at http://www.cse.ogi.edu/CSLU/corpora/

consisting of 14 Mel-scale FFT-based cepstral coefficients (MFCC) and their first-order time derivatives (ΔMFCC coefficients) was extracted. The MFCC coefficents have the advantage of robustness to noise and ease of computation over other features. This 28 dimensional feature was quantized using a codebook of 64 vectors. The HMMs were each configured in a four-state left-to-right architecture.

The recognizer was designed as follows. First, the major confusable sets were identified: $\{b, c, d, e, g, p, t, v, z\}$, $\{a, k, j\}$, $\{f, s, x\}$, $\{i, r, y\}$ and $\{m, n\}$. Next, for each confusable set, an HMM classifier was designed using each of the GD and DA approaches with the objective of minimizing errors for patterns within that set. Also, an HMM classifier was designed for the entire alphabet using the ML approach. Since standard ML works quite well for letters outside the confusable sets, the full-alphabet ML-designed HMM classifier can be used as a first pass to identify the utterance. If the first pass maps the input pattern to a letter in one of the confusable sets, then a second pass is initiated to identify the utterance more carefully using the (GD or DA-designed) HMM classifier designed for that set. Otherwise, the letter with the highest discriminant in the first pass is identified as the recognized class. Clearly, the second pass that uses either DA or GD designed HMMs can improve on the performance of the first pass. In the table below, we compare the error rates obtained under three different setups: (a) only the ML-designed first pass was used (ML) (b) ML-designed first-pass was followed by a GD-designed second pass (GD) and (c) ML-designed first-pass was followed by a DA-designed second pass (DA). The error rates obtained in the three setups are presented for each of the background conditions. Clearly, the DA approach performs consistently and substantially better than both ML and GD approaches.

Background	ML	GD	DA
Clean	14.4%	13.9 %	7.0%
Car Noise	27.6%	25.4%	21.3%
White Noise	17.1%	15.4%	11.3%
Computer Fan noise	30.4%	28.5%	24.1%
Air-conditioner noise	23.6%	21.8%	13.8%

Table 1: Comparison of misclassification rates (P_e) obtained by the maximum likelihood (ML), gradient descent (GD), and deterministic annealing (DA) method for isolated letter recognition under different background conditions.

It is important to note that the results in the table which are for the training set are given to illustrate the improved optimization offered by the DA approach. Given the size of the training set, all three methods in fact overtrain the classifier and the performance of all three methods are greatly degraded outside the training set. The relative gains of DA are also similarly reduced. Modified variants of these algorithms which are designed to eliminate overtraining are currently under investigation. These results may be available for presentation at the conference.

5. REFERENCES

[1] J.-C. Junqua, "SmarTspelL: a multipass recognition system for name retrieval over the telephone", *IEEE Trans. Speech and Audio Processing*, March 1997, vol.5,(no.2):173-82.

[2] L. R. Bahl, P. F. Brown, P. V. DeSouza, R. L. Mercer, "Maximum Mutual Information estimation of hidden Markov model parameters", in *Proc. ICASSP-86*, pp. 49-52, Tokyo, Japan.

[3] B. H. Juang, W. Chou, C. H. Lee, " Minimum classification error rate methods for speech recognition ", *IEEE Trans. Speech and Audio Processing* , vol. 5, no.3, pp 257-65, 1997.

[4] H. Watanabe, S. Katagiri, "HMM speech recognition based on discriminative metric design", in *Proc. ICASSP-97*, pp 3237-3240, 1997.

[5] B. H. Juang, S. Katagiri, "Discriminative learning for minimum error classification", *IEEE Trans. Sig. Proc.* , vol. 40, pp 3043-3054, 1992.

[6] K. Rose, E. Gurewitz, G.C. Fox, "Vector quantization by deterministic annealing", *IEEE Trans. Information theory*, vol.38, p.1249-1258, 1992.

[7] D. Miller, A. Rao, K. Rose, A. Gersho, " A global optimization method for statistical classifier design ", *IEEE Trans. Signal Processing*, vol.44, no.12, p:3108-22, Dec. 1996.

[8] A. Rao, D. Miller, K. Rose, A. Gersho, " Mixture of Experts Regression Modeling by Deterministic Annealing ", *IEEE Trans. Signal Processing* , Nov. 1997.

[9] A. Rao, K. Rose, A. Gersho, "A Deterministic Annealing Approach to Discriminative Hidden Markov Model Design", *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 1997, pp. 266-275.