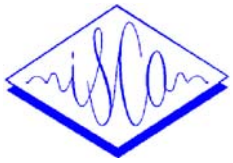

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

THE DETERMINISTIC ANNEALING APPROACH FOR DISCRIMINATIVE CONTINUOUS HMM DESIGN

Cecile Gelin-Huet, Kenneth Rose and Ajit Rao
Signal Compression Lab,
Department of Electrical and Computer Engineering,
University of California, Santa Barbara , CA 93106, USA
{huet, rose, ajit}@laurel.ece.ucsb.edu

ABSTRACT

We propose a deterministic annealing (DA) algorithm to design classifiers based on continuous observation hidden Markov models. The algorithm belongs to the class of minimum classification error (MCE) techniques that are known to outperform maximum likelihood (ML) design. Most MCE methods smooth the piecewise constant classification error cost to facilitate the use of local descent optimization methods, but are susceptible to the numerous shallow local minimum traps that riddle the cost surface. The DA approach employs *randomization* of the classification rule followed by minimization of the corresponding *expected* misclassification rate, while controlling the level of randomness via a constraint on the Shannon entropy. The effective cost function is smooth and converges to the MCE cost at the limit of zero entropy. The proposed algorithm significantly outperforms both standard ML and standard MCE design methods on the E-set database.
Keywords: Speech recognition, Discriminative training, Deterministic annealing, continuous HMM.

1. INTRODUCTION

Many practical speech recognition systems employ hidden Markov models (HMMs) to model speech units. Typically, the system associates an HMM with each speech unit, and recognition is performed via competition between these HMMs.

During the design (or training) phase, the HMM parameters are learned from a training set of speech utterances. Optimal training is a major challenge since the natural choice for design cost is the classification error rate (defined as the fraction of training patterns that is misclassified), which is a piecewise constant function of the HMM model parameters, and does not lend itself to gradient-based optimization. The popular Maximum Likelihood (ML) approach circumvents this difficulty by discarding the classification error cost and replacing it with the ML objective, which although mismatched is easier to optimize. Recently, there has been renewed interest in direct design approaches whose objective is minimum classification error (MCE). MCE techniques smooth the classification error cost function and jointly optimize all HMM parameters of the classifier via gradient descent. Of particular importance within the MCE family is the generalized probabilistic descent (GPD) [1,2,3,5]. Although MCE targets the true design cost and thereby offers significant performance gains over ML, it suffers from a significant drawback. The MCE cost surface is riddled with shallow local minima that easily trap local descent methods, and may substantially compromise performance.

The above provides the motivation for the deterministic annealing (DA) method proposed here, which is an MCE technique in that it targets the true cost (classification error rate), but does so while employing a powerful optimization tool. DA was first proposed for clustering and related problems [12,13] and later extended to solve problems which require structural constraints on the clustering rule [6], and applied to certain source coding systems [7], regression functions [8,10], pattern classifiers [6], etc. For a tutorial on DA see [14]. Most recently, DA was successfully applied in the design of discrete observation HMM classifiers [9,11], and was shown to substantially outperform both ML and GPD. In this paper, we propose a generalization of the DA method to the design of continuous observation HMMs.

2. PROBLEM STATEMENT

The isolated-word speech recognition problem is specified by a training set, $T=\{(\boldsymbol{x}_1,\boldsymbol{c}_1),(\boldsymbol{x}_2,\boldsymbol{c}_2),..,(\boldsymbol{x}_N,\boldsymbol{c}_N)\}$ of labeled training patterns. The pattern $\boldsymbol{x}_i$ corresponds to an utterance of the word $\boldsymbol{c}_i$ which belongs to a finite-sized dictionary, $C=\{1,2,...,M\}$. Pattern $\boldsymbol{x}_i$ is a sequence of T_i *feature vectors* extracted from the speech utterance. Each N_f–dimensionnal feature vector typically contains the cepstral or LPC coefficients and

This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, Cisco Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Intel Corp., Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., Rockwell International Corp., Panasonic Technologies, Inc., and Texas Instruments, Inc.

their derivatives. The recognition system consists of a set of HMMs $\{H_j; j=1,2..N_H\}$, usually, one per word in the dictionary. Model H_j is fully specified by the parameter set Λ_j, which includes the state transition probabilities $A_j[]$, the state-conditioned output distributions $B_j[]$ and the state priors $\pi_j[]$. These models determine the classifier, $\boldsymbol{C}$, via the classification rule that maps the training pattern, $\boldsymbol{x_i}$ to the class $C(\boldsymbol{x_i})$ as follows:

- Given the pattern, $\boldsymbol{x_i}$, a "path score" is computed for each path $\boldsymbol{p}$ (defined as one possible sequence of states, s_0, s_1..) in each model H_j:
$$\ell(x_i, H_j, p) = \log \pi_j[s_0] + \sum_{t=0}^{T_i-2} \log A_j[s_t, s_{t+1}] + \sum_{t=0}^{T_i-1} \log B_j[s_t, x_i(t)].$$

- The emission probability $B_j[]$ is a mixture of K gaussian distributions:
$$B_j[s, x_i(t)] = \sum_{k=0}^{K-1} \frac{c_{j,s,k}}{\sqrt{(2\pi)^{N_f} |\Sigma_{j,s,k}|}} e^{-\frac{[x_i(t)-\mu_{j,s,k}]^T \Sigma_{j,s,k}^{-1} [x_i(t)-\mu_{j,s,k}]}{2}}.$$

- The path with the highest score (the winning path) is determined and $\boldsymbol{x_i}$ is mapped to the class of the HMM that the winner path belongs to:
$$C(x_i) = \arg\max_j \left[\arg\max_p [\ell(x_i, H_j, p)] \right].$$

The classifier operation can be viewed in terms of competition between paths. The observation, $\boldsymbol{x_i}$, is ultimately labeled by the class index of the HMM to which the winning path belongs.

The design objective for is to jointly optimize the model parameters Λ_j composed of:

Priors, $\pi_j[s_0]$,

State transition matrices, $A_j[s_t, s_{t+1}]$,

Distribution means $\mu_{j,s,k}[m], m \in [1, N_f]$,

Distribution variances $\Sigma_{j,s,k}[m][n], m, n \in [1, N_f]$,

Distribution weights, $c_{j,s,k}$,

So as to minimize the empirical classification error rate:
$$\underset{\Lambda_j}{Min} \left\{ P_e = 1 - \frac{1}{N} \sum_{i=1}^{N} \delta(C(x_i), c_i) \right\}.$$

Here $\delta()$ is the Kronecker delta function: $\delta(u,v)=1$ if $u=v$, and $\delta(u,v)=0$ elsewhere.

As noted above, a significant design problem is due to the piecewise constant nature of P_e, which precludes the use of direct descent-based optimization. The popular ML approach circumvents this difficulty by replacing the true cost function with a sub-optimal design objective. GPD and other MCE approaches smooth the true cost function to allow descent-based optimization, but are still susceptible to poor local optima.

3. DETERMINISTIC ANNEALING

The fundamental principle underlying the DA approach to HMM design is the randomization of the "best-path" rule during the design. The original (non-random) rule which associates pattern $\boldsymbol{x_i}$ with a unique winning state sequence $\boldsymbol{p}$ is replaced by a randomized rule that chooses state sequence $\boldsymbol{p}$ in model H_j *in probability*. Specifically, the winning probability of path $\boldsymbol{p}$ is given by the Gibbs distribution,
$$P(p, H_j \mid x_i) = \frac{e^{\gamma \ell(x_i, H_j, p)}}{\sum_{j=1}^{N_H} \sum_{p \in H_j} e^{\gamma \ell(x_i, H_j, p)}}.$$

We note that paths with higher scores are more probable winners in the competition. 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 indicates that we assign higher probabilities of winning to paths of higher scores. In the limiting case of $\gamma \to \infty$, the random classification rule reverts to the non-random "best path" classifier that assigns all the winning probability to the path with the highest score.

The "Gibbs" parametric form of this distribution is not arbitrary, but is derivable from information-theoretic principles [6, 8,14]. We should re-emphasize that the random classifier paradigm is adopted only during the training phase. The DA algorithm ultimately produces a regular, non-random HMM-based classifier.

The expected misclassification rate of the random classifier is given by:
$$< P_e > = 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{N_H} \left[\delta(C(H_j), c_i) \sum_{p \in H_j} P(p, H_j \mid x_i) \right]}{N},$$
where $C(H_j)$ is the class associated with the HMM H_j.

Straightforward minimization of the expected misclassification rate with respect to all the HMM parameters and the scale parameter γ is possible although such a method would be highly susceptible to shallow local minimum traps. We propose instead to introduce the notion of annealing which involves an entropy-constrained formulation.

Instead of simply optimizing the misclassification cost $<P_e>$ during the design process, we do so while enforcing a constraint on randomness, which is measured by the (conditional) Shannon entropy:

$$\mathcal{H} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_H}\sum_{p\in H_j} P\left(p, H_j \mid x_i\right)\log P\left(p, H_j \mid x_i\right).$$

Thus we minimize the expected misclassification $<P_e>$ while constraining the entropy to a prescribed level, $\mathcal{H} = \hat{\mathcal{H}}$. We then gradually lower the entropy level while tracking the minimum. The constrained optimization problem of minimizing $<P_e>$ at a given entropy level is equivalent to the unconstrained Lagrangian minimization:

$$\min_{\Lambda_j,\gamma}\left\{L = < P_e > - T\mathcal{H}\right\}, \qquad (1)$$

where T is the corresponding Lagrange multiplier. The parameter T is gradually reduced from a high value to zero while tracking the minimum of L. This is directly analogous to the process of annealing in physics. The parameter T is naturally referred to as the "temperature". As $T \to 0$, the optimization reduces to the unconstrained minimization of $<P_e>$ which forces $\gamma \to \infty$ 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.

Thus, the optimal probability distribution over the classes evolves during the design. When the entropy (and temperature) is high, the probability distribution over the paths is uniform, and all paths in the trellis are equally probable. However, as the entropy is reduced, the distribution becomes more discriminating and assigns higher probability to more likely paths. Eventually, only the most likely path is assigned a non-zero probability and the classifier becomes a hard (non-random) classifier. The DA procedure is summerized in Figure 1.

The re-estimation process for a given value of T uses a gradient descent algorithm to optimize the model parameter set $\{\Lambda_j\}$ and the smoothing factor γ.

The derivative of criterion L with respect to parameter θ is given by:

$$\frac{\partial L}{\partial \theta} = \frac{\gamma}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_H}\sum_{p\in H_j}\left\{\begin{array}{l} P(p, H_j \mid x_i)\dfrac{\partial P(p, H_j \mid x_i)}{\partial \theta} \\ \times[T\gamma\ell(x_i, H_j, p) - \delta(C(H_j), c_i)]\end{array}\right\}$$

There are two different cases for consideration:

for $\theta = \Lambda_j \in \{c_{j,s,k}, \mu_{j,s,k}[m], \mu_{j,s,k}[m][n]\}$

$$\frac{\partial P(p, H_j \mid x_i)}{\partial \Lambda_j} = \frac{\partial \ell(x_i, H_j, p)}{\partial \Lambda_j} - \sum_{j=1}^{N_H}\sum_{p\in H_j} P(p, H_j \mid x_i)\frac{\partial \ell(x_i, H_j, p)}{\partial \Lambda_j},$$

and $\theta = \gamma$:

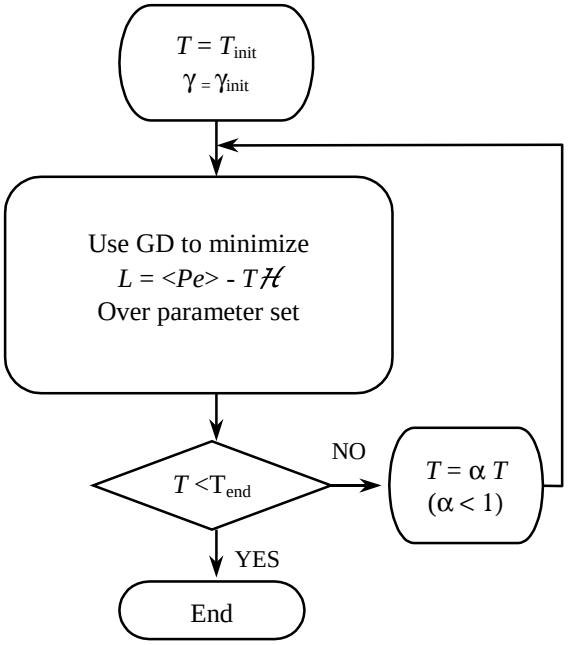

Figure 1:DA procedure Flowchart.

$$\frac{\partial P(p, H_j \mid x_i)}{\partial \gamma} = \ell(x_i, H_j, p) - \sum_{j=1}^{N_H}\sum_{p\in H_j} P(p, H_j \mid x_i)\ell(x_i, H_j, p)$$

The re-estimation formula for HMM training can be written in an efficient forward-backward form similar to [11].

It is worthwhile to note some connections between the GPD approach and the DA optimization. If in (1), we impose $T = 0$ and fix γ to a constant value, the resulting criterion is a smoothed version of the actual MCE criterion, and the smoothness of this criterion is determined by the parameter γ. This criterion is very similar to the one used in the GPD method. It can further be shown [11] that for particular values of the constants used in GPD criterion [2], the two criteria are equivalent. In this sense, GPD may be viewed as a special, degenerate case of the DA procedure as it minimizes its criterion at zero temperature, and for a particular value of the smoothness factor γ. The DA method, however, involves the important effect of annealing where the temperature is gradually reduced while the classifier parameters and the smoothing factor are optimized at each temperature.

4. EXPERIMENTAL RESULTS

We have compared the performance DA to that of GPD and standard ML on the challenging task of recognizing spoken utterances of letters belonging to the E-set: {*b,c,d,e,g,p,t,v,z*}. The E-set classification problem is notoriously difficult due to its high confusability. Misclassification within the E-set has been identified as

the most significant cause of errors in the more general problem of spelled word recognition which has several applications such as automated telephone forwarding systems and automated directory assistance [4].

The experiments were carried out on speech data from the ISOLET database, which is a part of the CLSU corpora. The E-set portion of this database was divided into a training set (utterances by 60 male and 60 female speakers) and a test set (utterances by 15 male and 15 female speakers). Two utterances of each letter by each speaker were used. The speech signal was divided into 32 ms length frames with a 16 ms inter-frame overlap; 10 MFCC coefficients and their first-order time derivatives (ΔMFCC coefficients) were extracted from each frame. The HMM classifier consists of nine left-to-right HMMs. Two different classifier configurations were tested. The first was a minimal two states per HMM configuration, and the second consisted of six states per HMM. In both cases, the state-conditional output distribution was Gaussian.

Table 1 compares the error rates obtained by the three different optimization schemes: ML, GPD and DA. Clearly, DA yields the best classification error rates in both configurations. The DA error rates are consistently lower than ML and GPD on both training and test sets.

Number of States	Data Set	ML	GPD	DA
2	Train	25.41	21.25	2.96
	Test	28.70	26.67	17.04
6	Train	24.86	19.54	1.76
	Test	28.89	27.04	15.74

Table 1: *Error rates obtained by competing design methods (ML, GPD and DA) on the E-set in two configurations (two states / six states).*

5. CONCLUSIONS AND FUTURE WORK

This paper was motivated by the difficulties encountered in direct optimization of the classification error rate for continous HMM classifier design. A powerful optimization method based on deterministic annealing was developed to attack this shortcoming. Experimental results on E-set letter recognition show the promise of the approach and demonstrate performance gains over two standard classifier design methods, namely maximum likelihood and generalized probabilistic descent. Work is currently in progress to extend the deterministic annealing approach to the design of semi-continuous HMM classifiers.

6. REFERENCES

[1] Chang P.-C. and Juang B.-H. (1993), Discriminative training of dynamic programming based speech recognizers. *IEEE Trans. On Speech and Audio Processing*, Vol. 1, No 2, April 1993, pp 135-143.

[2] Juang B.-H., Chou W. and Lee C.-H. (1997), Minimum classification error rate methods for speech recognition. *IEEE Trans. On Speech and Audio Processing* Vol. 5, No. 3, May 1997, pp. 257-265.

[3] Juang B.-H. and Katagiri S. (1992), Discriminative learning for minimum error classification. *IEEE Trans. On Signal Processing*, Vol. 40, No 12, Dec 1992, pp 3043-3054.

[4] Junqua J.-C.(1997), SmarTspell: a multipass recognition system for name retrieval over the telephone», *IEEE Trans. On Speech and Audio Processing*, Vol. 5, No. 2, 1997, pp. 173-182.

[5] Katagiri S., Lee C.-H. and Juang B.-H. (1991), New discriminative training algorithms based on the generalized probabilisticdescend method, *Proc. Workshop on Neural Networks for Signal Processing, Princeton NJ, Sept 1991*, pp. 299-308.

[6] Miller D., Rao A.V., Rose K. and Gersho A. (1996), A global optimization technique for statistical classifier design, *IEEE Trans. On Signal Processing*, Vol. 44, No 12, Dec 1996.

[7] Rao A.V., Miller D., Rose K. and Gersho A.(1996), A generalized VQ method for combined compression and estimation, *Proc. ICASSP 1996*, pp. 2032-2035.

[8] Rao A.V., Miller D., Rose K. and Gersho A.(1997), Mixture of experts regression modeling by deterministic annealing, *IEEE Trans. On Signal Processing*, Vol. 45, No 11, Nov 1997, pp 2811-2820.

[9] Rao A.V. and Rose K. and Gersho A. (1997), Design of robust HMM speech recognizers using deterministic *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, Dec. 1997, pp. 466-473.

[10] Rao A.V., Miller D., Rose K. and Gersho A.. (1999), A deterministic annealing approach for parsimonious design of piecewise regression models, *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 21, No. 2, Feb. 1999, pp. 159-173.

[11] Rao A.V. and Rose K. (1999), Deterministically annealed design of hidden Marov model Speech Recognizers. Submitted *to IEEE Trans. On Speech and Audio Processing.*

[12] Rose K., Gurewitz E. and Fox G.C. (1992), Vector quantization by deterministic annealing, *IEEE Trans. On Information theory* Vol. 38, 1992, pp. 1249-1258.

[13] Rose K., Gurewitz E. and Fox G.C. (1993), Constrained clustering as an optimization method, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. 15, 1993, pp. 785-794.

[14] Rose K. (1998), Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, *IEEE Trans. On Information theory* Vol. 86, No. 11, Nov. 1998, pp. 2210-2239.