Design of robust HMM speech recognizers using deterministic annealing *

Ajit Rao, Kenneth Rose, and Allen Gersho
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106.

Abstract

We attack the difficult problem of optimizing a hidden Markov model (HMM) based speech recognizer to minimize its misclassification rate. In conventional HMM recognizer design, the training data is divided into subsets of identically labeled tokens and the HMM for each label is designed from the corresponding subset using a maximum likelihood (ML) objective. However, ML is a mismatched objective and ML design does not minimize the recognizer's misclassification rate. The misclassification rate is difficult to optimize directly because the cost surface is riddled with shallow local minima that tend to trap naive descent methods. Here, we propose an approach which is based on the powerful technique of deterministic annealing (DA) to minimize the misclassification cost while avoiding shallow local minima. In the DA approach, the classifier's decision is randomized during design and its expected misclassification rate is minimized while enforcing a level of "randomness" measured by the Shannon entropy. The entropy constraint is gradually withdrawn (annealing) and in the limit, the cost function converges to the misclassification rate of a regular "non-random" recognizer. This algorithm is implementable by a low complexity forward-backward procedure similar to the Baum-Welch re-estimation used in ML design. Our experiments on speaker-independent isolated word speech recognition of clean and noise-corrupted utterances of letters of the difficult E-set = { b,c,d,e,g,p,t,v,z} demonstrate

*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.,

0-7803-3698-4/97/$10.00 © 1997 IEEE.

that DA-designed recognizers offer consistent and substantial improvements in accuracy over ML-designed recognizers.

1 Introduction

Hidden Markov modeling (HMM) of speech utterances is a key paradigm in many conventional speech recognition systems. While the objective in speech recognition applications is accurate classification of utterances, the design of HMM-based speech recognition systems has traditionally been equated to and based on maximum likelihood modeling of speech. Recently, several speech recognition researchers [1, 2, 3, 4, 5] have recognized the inherent mismatch between the goals of the recognition problem (minimization of the misclassification rate) and the objective of the modeling problem (maximization of the likelihood). The problem is exacerbated because: (i) the training data is often inadequate to estimate the distribution optimally and (ii) the HMM is not necessarily the optimal stochastic model for speech. It should perhaps be noted that a similar mis-match between the true objective and the commonly used design objective has been pointed out for the related problem of pattern classifier and regression function design. This realization has led to new approaches [3, 6, 7] that directly optimize the true objective and demonstrate significant improvements in performance.

Robustness and accuracy of speech recognizers can be substantially improved by joint optimization of all design parameters to directly minimize the misclassification rate. However, there are several important design difficulties that must be overcome. Firstly, the misclassification cost is not a smooth function of the design variables. This eliminates the possibility of using a simple local descent method for optimization. In fact, the cost surface is highly complex and riddled with shallow local minima which tend to trap even sophisticated descent algorithms. Secondly, joint optimization of the recognizer parameters is computationally complex, even for an off-line design.

The novel approach we propose in this paper is based on the powerful technique of deterministic annealing (DA) for clustering and related problems which was first presented in [8, 9]. Of particular relevance here is the DA extension to include structural constraints which was applied to the design of pattern classifiers [7], regression functions [6] and a new class of source coding systems [10]. In this paper, the DA approach is extended, non-trivially, to the design of HMM-based speech recognizers. The method is used in conjunction with minimum classification error (MCE) training [3] and can be applied to both discrete and continuous output HMM systems. Moreover, DA is implementable by a low-complexity forward-backward algorithm similar to the Baum-Welch re-estimation steps of the popular but suboptimal maximum likelihood design method.

2 HMM-based speech recognition

Consider the following typical isolated-word recognition problem : We are given a training set $\mathcal{T} \equiv \{(\mathbf{y}_1, c_1), (\mathbf{y}_2, c_2), ..(\mathbf{y}_N, c_N)\}$ of labelled *training patterns.* The pattern, $\mathbf{y}_i$ corresponds to an utterance of the word, c_i which belongs to a finite-sized dictionary, $\mathcal{C} \equiv \{1, 2, ..M\}$. Moreover $\mathbf{y}_i$ is a vector of l_i observations which may be discrete or continuous. Continuous observations are usually a sequence of *feature vectors* (typically consisting of cepstral coefficients or linear prediction coefficients and their derivatives) extracted from consecutive time frames of the speech utterance. If discrete-observation HMMs must be used, the feature vectors are vector-quantized to entries in a pre-designed finite length codebook of *prototype vectors* and the sequence of quantization indexes obtained by this process is defined as the discrete observation vector.

The HMM recognition system consists of a set of HMMs $\{H_j,\ j = 1, 2..M\}$, one per word in the dictionary. The model H_j is fully specified by the parameter set Λ_j which includes the state transition probabilities, state-dependent output distributions, and the initial probabilities of the states. The HMM system that we design here uses the common "best path" discriminant approach [1] which can be implemented by the computationally efficient Viterbi search algorithm.

Given a training pattern $\mathbf{y}_i$ we determine for each HMM H_j a discriminant $d_j(\mathbf{y}_i)$ which is the log likelihood (based on the HMM model) of the most likely state sequence:

$$d_j(\mathbf{y}_i) = \max_{\mathbf{s} \in \mathcal{S}_{l_i}(H_j)} l(\mathbf{y}_i, \mathbf{s}, H_j). \tag{1}$$

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

$$C(\mathbf{y}_i) = arg \max_j d_j(\mathbf{y}_i). \tag{2}$$

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.

To design this HMM-based recognizer, we must jointly optimize the HMM parameters $\{\Lambda_j\}$ to minimize the empirical misclassification rate measured over the training set,

$$\min_{\{\Lambda_j\}} P_e = 1 - \frac{1}{N} \sum_{i=1}^{N} \delta(C(\mathbf{y}_i), c_i). \tag{3}$$

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

[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.

A significant difficulty in this optimization arises from the nature of the cost P_e which is a piecewise constant function of the HMM parameter set. Consequently the gradients with respect to the HMM parameters are zero almost everywhere, thus preventing us from using gradient descent based optimization techniques. Recently, Juang et al [3] have proposed the Generalized Probabilistic descent (GPD) approach to circumvent this difficulty - in GPD, the piecewise cost function is replaced by a smooth approximation to it. While the modified GPD cost function is continuously differentiable, thus allowing descent-based optimization, in practice, there are numerous shallow local minima on the complex cost surface that can easily trap optimization methods based on simple descent. Here, we propose an alternative approach based on the powerful optimization method of deterministic annealing which simultaneously tackles the piecewise nature of the cost function and the problem of shallow local minima traps.

3 Deterministic Annealing

The fundamental idea behind the DA approach is to randomize the "best-path" classification rule during design. In effect, we replace the original (non-random) rule which associates a pattern $\mathbf{y}_i$ to a unique winning state sequence $\mathbf{s}$, by a randomized rule that chooses a state sequence $\mathbf{s}$ in model H_j with a probability,

$$P(\mathbf{s}, H_j|\mathbf{y}_i) = \frac{e^{\gamma l(\mathbf{y}_i, \mathbf{s}, H_j)}}{\sum_{j'} \sum_{\mathbf{s}' \in \mathcal{S}_{l_i}(H_{j'})} e^{\gamma l(\mathbf{y}_i, \mathbf{s}', H_{j'})}}. \quad (4)$$

This parametric form for $P(\mathbf{s}, H_j|\mathbf{y}_i)$ is referred to as the "Gibbs" distribution. Note that paths with higher "scores", $l(\mathbf{y}_i, \mathbf{s}, H_j)$ are more probable to be chosen as output. 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 state sequences with higher log likelihoods. In the limiting case of $\gamma \rightarrow \infty$, the random classification rule reverts to the non-random "best path" classifier, which assigns a non-zero probability of winning only to the path with the highest log likelihood as in (1). The parametric form of this distribution is not arbitrary, but can be derived in a systematic manner from information-theoretic principles (see [11] for derivation in the context of HMMs, as well as earlier derivations of DA [6, 7, 10]). At this point, we re-emphasize that the random classifier paradigm is adopted only during design - ultimately the DA algorithm will produce a regular, non-random HMM-based classifier.

The average misclassification rate of the random classifier is given by:

$$< P_e >= 1 - \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathbf{s} \in \mathcal{S}_{l_i}(H_{c_i})} P(\mathbf{s}, H_j|\mathbf{y}_i) \quad (5)$$

Straightforward minimization of the expected misclassification (5) with respect to all the HMM parameters and the scale parameter γ is possible although such a method is highly susceptible to shallow local minima traps. We prefer 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 the randomness which is measured by the Shannon entropy

$$H = -\frac{1}{N}\sum_i \sum_j \sum_{\mathbf{s}\in\mathcal{S}_{l_i}(H_j)} P(\mathbf{s}, H_j|\mathbf{y}_i)\log P(\mathbf{s}, H_j|\mathbf{y}_i). \tag{6}$$

Thus we minimize the expected misclassification $< P_e >$ while constraining the entropy to a prescribed 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:

$$\min_{\{\Lambda_j\},\gamma} L =< P_e > -TH \tag{7}$$

where T is the corresponding Lagrange parameter. The parameter, T, is gradually reduced from a high value to zero while tracking the local minima of L. This is directly analogous to the process of annealing in physics. The parameter T is naturally referred to as the "temperature". As $T \rightarrow 0$, 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.

An important aspect of the proposed method is the discovery of an efficient forward-backward algorithm to determine the gradient parameters for the optimization. This algorithm substantially cuts down on computational complexity and memory requirements. The complexity of the DA method scales similarly to the maximum likelihood method with respect to the number of states and training vectors.

4 Experimental Results

We have compared the performance of the DA algorithm and the standard maximum likelihood (ML) algorithm for the difficult E-set recognition task (recognition of spoken utterances of the letters $b, c, d, e, g, p, t, v, z$). The E-set classification problem is well-known to be difficult because of the confusability of the alphabet. Misclassification within the E-set has been recognized as the most significant cause of errors in the more general problem of letter recognition which has several applications such as automated telephone forwarding systems and automated directory assistance [12]. In many practical situations, the difficulty of the problem is further aggravated by noisy background conditions.

The experiments were carried out on both clean and noise-corrupted (white noise and car noise) speech data from the ISOLET database which is a part of the CLSU corpora [2]. The speech data consists of E-set letters spoken by 30 speakers (15 male and 15 female). Every speaker uttered each letter twice. For each utterance, two noise-corrupted versions were obtained by adding synthetic white noise and recorded car noise to the clean speech.

The speech was sampled at 16 KHz and divided into frames of 512 samples (32ms). Consecutive frames overlap by 256 samples (16ms). An FFT was performed on each speech frame and Mel-scaled FFT cepstral coefficients (MFCC) [13] were extracted. The MFCC coefficients are known to have the advantage of robustness to noise and ease of computation over other features. The feature set consists of 14 MFCC coefficients and their first-order time derivatives (ΔMFCC coefficients). We recognize that the front-end feature extraction process can be optimized further to improve the misclassification rate, but feature extraction is not our focus here.

In our experiments, we designed recognizers based on discrete HMMs. The 28-dimensional feature obtained from the MFCC analysis was quantized using a codebook of 64 "prototypes" For each dataset (clean speech, white noise background, car noise background), the codebook was designed independently using a Generalized Lloyd algorithm [14]. The recognizer consisted of nine HMMs, each configured in a four-state left-to-right architecture.

Table 1 compares the error rates obtained by the maximum likelihood (ML) method and the DA design method in each of the background conditions. Clearly, DA yields classification error rates that are consistently lower - the improvement is by a factor of 2-3.

Background Condition	Clean	Car Noise	White Noise
P_e (ML)	22.22%	38.52%	28.52%
P_e (DA)	7.04%	18.52%	9.63%

Table 1: Comparison of misclassification rates (P_e) obtained by the maximum likelihood (ML) method and the deterministic annealing (DA) method for E-set recognition under different background conditions.

5 Conclusions

We have presented a novel algorithm to optimize HMM-based speech recognizers, based on direct minimization of the misclassification rate. At the heart of this algorithm is the powerful deterministic annealing design method. The

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

new approach consistently outperforms the popular Maximum Likelihood algorithm at the cost of a modest increase in design complexity. Simulation results show that a dramatic decrease in misclassification rate is achieved by DA design of the speech recognizer.

References

[1] 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.

[2] H. Boulard, Y. Konig, N. Morgan, "REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities - Application to Transition-Based Connectionist Speech Recognition", ICSI technical Report TR94-064, Internat, Computer Science Inst. CA.

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

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

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

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

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

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

[9] K. Rose, E. Gurewitz, G.C. Fox, "Constrained clustering as an optimization method", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.15, p.785-794, 1993.

[10] A. Rao, D. Miller, K. Rose, A. Gersho, " A Generalized VQ Method for Combined Compression and Estimation", in *Proc. ICASSP-96* pp. 2032 - 2035.

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

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

[13] J. W. Picone, "Signal modeling techniques in speech recognition", *Proc. IEEE*, vol. 81, pp.1215-1247, 1993.

[14] Y. Linde, A. Buzo, R. M. Gray "An algorithm for vector quantizer design", *IEEE Trans. on Comm.*, COM-28:84–95, 1980.