REMAP - EXPERIMENTS WITH SPEECH RECOGNITION

Yochai Konig[†‡] *Hervé Bourlard*[†*] *Nelson Morgan*[‡†]
{konig,bourlard,morgan}@icsi.berkeley.edu
[†] International Computer Science Institute, Berkeley, CA
[‡] EECS Department, University of California at Berkeley, Berkeley, CA
[*] Faculté Polytechnique de Mons, Mons, Belgium

ABSTRACT

In this report we present experimental and theoretical results using a framework for training and modeling continuous speech recognition systems based on the theoretically optimal Maximum a Posteriori (MAP) criterion. This is in constrast to most state-of-the-art systems which are trained according to a Maximum Likelihood (ML) criterion. Although the algorithm is quite general, we applied it to a particular form of hybrid system combining Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs) in which the ANN targets and weights are iteratively re-estimated to guarantee the increase of the posterior probability of the correct model, hence actually minimizing the error rate. More specifically, this training approach is applied to a transition-based model that uses local conditional transition probabilities (i.e., the posterior probability of the current state given the current acoustic vector and the previous state) to estimate the posterior probabilities of sentences. Experimental results on isolated and continuous speech recognition tasks show an increase in the estimates of posterior probabilities of the correct sentences after training, and significant decreases in error rates in comparison to a baseline system.

1. INTRODUCTION

1.1. Maximum a Posteriori (MAP) Framework

In statistical pattern classification, it is known that the system leading to the minimum probability of error is the one that is trained to maximize the a posteriori probability of the correct class conditioned on the evidence [1] and uses that same criterion during recognition. Thus, in the case of speech recognition, if $X = \{x_1, \ldots, x_n, \ldots, x_N\}$ represents the input sequence to be classified, M_i $(i = 1, \ldots, I)$ the possible models, L the parameter set of the language model (i.e., both a lexicon and a probabilistic grammar), and Θ the acoustic model parameters, X will be optimally assigned to the sentence associated with model M_j if

$$M_j = \underset{M_i}{\operatorname{argmax}} \; P(M_i|X, L, \Theta), \quad i = 1, \ldots, I \tag{1}$$

The ideal training algorithm should determine the set of parameters $(\hat{\Theta}, \hat{L})$ that will maximize $P(M_{w_j}|X_j, L, \Theta)$ for all training utterances X_j $(j = 1, \ldots, J)$, associated with M_{w_j}[1], i.e.,

$$(\hat{\Theta}, \hat{L}) \;=\; \underset{(\Theta, L)}{\operatorname{argmax}} \; \prod_{j=1}^{J} P(M_{w_j}|X_j, L, \Theta) \tag{2}$$

with the following constraint:

$$\sum_{i=1}^{I} P(M_i|X, L, \Theta) \;=\; 1 \tag{3}$$

for every X, and where the sum over i represents the sum over all possible models. Note that this constraint makes the Maximum a Posteriori (MAP) criterion (2) discriminant. Indeed, when increasing the posterior probability of the correct model, the total probability mass assigned to the other models will automatically be reduced.

1.2. Maximum Likelihood (ML) Framework

Despite the optimality of the MAP criterion, most speech recognition systems are trained according to a maximum likelihood criterion that maximizes, in the parameter space, the likelihood of the data given some model. In hidden Markov models (HMMs), this likelihood can be represented as $P(X|M, \Theta)$. Classically, the likelihood formulation of (1) is obtained by applying Bayes' rule:

$$P(M|X, L, \Theta) = \frac{P(X|M, L, \Theta) P(M|L, \Theta)}{P(X|L, \Theta)} \tag{4}$$

For practical reasons, it is assumed that:

1. The parameters Θ of the acoustic model are independent of the parameters L of the language model, yielding

$$P(X|M, L, \Theta) \;\approx\; P(X|M, \Theta) \tag{5}$$

$$P(M|L, \Theta) \;\approx\; P(M|L) \tag{6}$$

2. Despite the fact that Θ and L vary during training, $P(X|L, \Theta)$ is assumed to be constant (leading to the poor discrimination properties of ML based systems). To understand this, imagine that a change in Θ happened to increase the likelihood of alternate models as well; in this case both the numerator and the denominator of (4) might increase.

1.3. Hybrid HMM/ANN Systems

In recent years there has been a significant body of work, both theoretical and experimental, which has aimed to overcome some of the limitations of the current HMM-based systems by combining HMMs and ANNs. In particular, we have shown that fairly simple layered structures can be used to estimate local emission probabilities for HMMs [2]. This approach is now usually referred to as a hybrid HMM/ANN system. Although the initial architecture (which we have called the Discriminant HMM) was developed to estimate global posterior probabilities $P(M|X)$, theoretical as well as implementation problems led us to a simplified version of this approach that was still based

[1] M_{w_j} represents the model associated with the specific input sequence X_j that is known at training time.

on a likelihood criterion discriminantly trained at the local (HMM state) level. A number of speech recognition systems based on this latter approach have been proved, in controlled tests, to be to be both effective in terms of accuracy (comparable or better than equivalent state-of-the-art systems) and efficient in terms of CPU and memory run-time requirements [3].

1.4. Posterior-based Hybrid HMM/ANN

In [4], we presented a new hybrid HMM/ANN approach that directly optimizes the acoustic parameter set Θ according to the MAP criterion, i.e., maximizing $P(M|X,\Theta)$ where M is the correct HMM associated with X. In principle this approach should minimize the error rate; cross-validation will be used to guarantee that minimization of the error rate happens not only on the training data, but also on an independent test set. This algorithm, which we call REMAP (RECURSIVE ESTIMATION AND MAXIMIZATION OF A POSTERIORI PROBABILITIES), iteratively re-estimates ANN targets and weights to guarantee an increase of the posterior probability of the correct sequence. We show in [4] that estimation of the new ANN targets can be done using "forward" and "backward" recurrences that are reminiscent of the EM-based Forward-Backward training algorithm of standard HMMs.

In [5] we reported initial experimental results on an isolated word recognition task. Here, we report extended experimental results on the earlier task and on a small continuous speech recognition task (all the experiments were done using only acoustic information). In addition, we describe theoretical ideas on how to incorporate language information into our framework. We begin by briefly reviewing our transition-based model, DISCRIMINANT HMM (DHMM), and our training algorithm, REMAP.

2. DISCRIMINANT HMM (DHMM)

REMAP is developed in the context of a transition-based model (though it is also applicable to non-transition-based models). Furthermore, the model uses local conditional transition probabilities (i.e., the posterior probability of the current state given the current acoustic vector and the previous state) to estimate the global posterior of sentence models. Thus, it is a true recognition model, i.e., it directly maps from acoustic sequences to sentences, unlike Hidden Markov Models (HMMs) that model the inverse modeling (the likelihood of producing an acoustic sequence).

In [2] it was shown that it is possible to compute the global posterior probability $P(M|X,L,\Theta)$ of (1) and (2) as:

$$P(M|X,L,\Theta) = \sum_{\forall\Gamma_j} P(M,\Gamma_j|X,L,\Theta) \quad (7)$$

$$= \sum_{\forall\Gamma_j} P(M|\Gamma_j,X,L,\Theta)P(\Gamma_j|X,L,\Theta) \quad (8)$$

in which "$\forall\Gamma_j$" represents all possible (legal) state sequences in M. Let $q_{j,n}$ denotes the specific state visited at time n for path Γ_j, with $q_{j,n} \in \mathcal{Q} = \{q^1,\ldots,q^k,q^\ell,\ldots,q^K\}$, the set of all possible HMM states making up all possible models M. Considering the second factor of (8) as the acoustic model and assuming that it is independent of the language model parameters (coupled with other standard assumptions [4]), we can then rewrite it as:

$$P(\Gamma_j|X,\Theta) = \prod_{n=1}^{N} P(q_{j,n}|q_{j,n-1},x_n,\Theta) \quad (9)$$

The first factor in (8) can be considered independent of the acoustic sequence X (since the state sequence is assumed) and will be further discussed later in this paper. An example of the model is given in Figure 1.

These new acoustic models, referred to as DISCRIMINANT HMMs (DHMM)[2], are thus now described in terms of CONDITIONAL TRANSITION PROBABILITIES $P(q_n^\ell|q_{n-1}^k,x_n)$, in which q_n^ℓ stands for the specific state q^ℓ of $\mathcal{Q}$ hypothesized at time n. As with traditional hybrid HMM/ANN systems, conditional transition probabilities can be estimated by an ANN (in our case a multilayer perceptron) with K output units and in which the acoustic input x_n[3] is complemented by a set of additional input units representing the state q^ℓ hypothesized at the previous time step $n-1$. The conditional transition probabilities are thus functions of Θ, the ANN parameter set, and can be written as $P(q_n^\ell|q_{n-1}^k,x_n,\Theta)$.

3. REMAP FOR DISCRIMINANT HMMS

3.1. Motivations

Discriminant HMMs as described above use conditional transition probabilities as the key building block for acoustic recognition. It is, however, well known that estimating transitions accurately is a difficult problem [6]. In our previous hybrid systems, the targets used for ANN training are typically given by the best segmentation resulting from a Viterbi alignment. This procedure thus yields rigid transition targets, which may not be optimal in the case of training (and testing!) of conditional transition probabilities.

One possible solution to this problem is to use a "full" MAP algorithm taking all possible paths into account to estimate conditional transition probabilities. This would lead to smooth estimates of ANN targets and (implicitly) to more training examples (including "negative" examples) since all the vectors of each training sentence will be assigned, with different probabilities, to all possible transitions permitted by the associated HMM.

3.2. Problem Formulation

Global MAP training of Discriminant HMMs should find the optimal parameter set Θ maximizing (2). In the following derivation we omit the dependency on the language model L; this will be further discussed in Section 5. Although, in principle, we could use a generalized back-propagation-like gradient procedure in Θ to maximize (2) (see, e.g., [7]), an EM-like algorithm should have better convergence properties, and would preserve the statistical interpretation of the ANN outputs. In this case, "full" MAP training of transition-based HMM/ANN hybrids requires a solution to the following problem: given a trained ANN at iteration t providing a parameter set Θ^t and, consequently, estimates of $P(q_n^\ell|x_n,q_{n-1}^k,\Theta^t)$, how can we determine new ANN targets that:

1. will be smooth estimates of conditional transition probabilities, $\forall$ possible (k,ℓ) state transition pairs in M and $\forall n \in [1,n]$.
2. when used in training the ANN for iteration $t+1$, will lead to new estimates of Θ^{t+1} and $P(q_n^\ell|x_n,q_{n-1}^k,\Theta^{t+1})$ that are guaranteed to incrementally increase (2)?

[2] It could be argued that these models are no longer HMMs but more like "stochastic finite state acceptors".

[3] As done with previous hybrid HMM/ANN systems, x_n will usually be replaced by $X_{n-c}^{n+d} = \{x_{n-c},\ldots,x_n,\ldots,x_{n+d}\}$ to take some acoustic context into account.

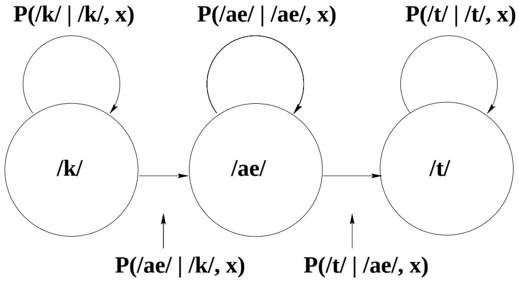

Figure 1. An example of a Discriminant HMM for the word "cat". The variable x refers to a specific acoustic observation x_n at time n.

In [4], we prove that a re-estimation of ANN targets that guarantee convergence to a local maximum of (2) is given by[4]:

$$P^*(q_n^\ell|x_n, q_{n-1}^k) = P(q_n^\ell|X, q_{n-1}^k, \Theta^t, M) \quad (10)$$

which means that the new ANN target associated with x_n and a specific transition $q^k \rightarrow q^\ell$ has to be calculated as the probability of that specific transition CONDITIONED ON THE WHOLE TRAINING SENTENCE X and the associated model M. Roughly speaking, our global optimization goal (2) is realized through the estimation of the targets (that correspond to a "local" acoustic window, e.g., 100 ms) by using the whole utterance. In [4], we further prove that alternating ANN target estimation (the "estimation" step) and ANN training (the "maximization" step) is guaranteed to incrementally increase (2) over t.

The remaining problem is to find an efficient algorithm to express $P(q_n^\ell|X, q_{n-1}^k, M)$ in terms of $P(q_n^\ell|x_n, q_{n-1}^k, M)$. This can be obtained by observing that:

$$P(q_n^\ell|X, q_{n-1}^k, M) = \frac{p(q_{n-1}^k, q_n^\ell, M|X)}{\sum_\ell p(q_{n-1}^k, q_n^\ell, M|X)} \quad (11)$$

in which the terms on the right hand side can be computed with α (forward) and β (backward) EM-like recurrences using only local conditional transition probabilities.

3.3. REMAP Training Algorithm

The general scheme of the REMAP training of hybrid HMM/ANN systems can finally be summarized as follow. Starting from some initial net providing $P(q_n^\ell|x_n, q_{n-1}^k, \Theta^t)$, $t = 0$, $\forall$ possible (k, ℓ)-pairs[5]:

1. Compute ANN targets $P(q_n^\ell|X, q_{n-1}^k, \Theta^t, M)$ according to (11), $\forall$ possible (k, ℓ) state transition pairs in M and $\forall n \in [1, n]$.
2. For all x_n's in X, train the ANN to minimize the relative entropy between the outputs and targets. This provides us with a new set of parameters Θ^t, for $t = t + 1$.
3. Iterate from 1 until convergence.

This procedure is thus composed of two steps: an Estimation (E) step, corresponding to step 1 above, and a Maximization (M) step, corresponding to step 2. Convergence of this training scheme can be proved [4]. In this regard, it is reminiscent of the EM algorithm [8]. However, in the standard EM algorithm, the M step involves the actual maximization of the likelihood function. In a related approach, usually referred to as GENERALIZED EM (GEM) algorithm, the M step does not actually maximize the likelihood but simply increases it (by using, e.g., a gradient procedure). Similarly, REMAP increases the global posterior function during the M step (in the direction of targets that actually maximize that global function), rather than actually maximizing it. Recently, a similar approach was suggested for mapping input sequences to output sequences [9].

[4]In the following, we consider only one particular training sequence X associated with one particular model M. It is, however, easy to see that all of our conclusions remain valid for the case of several training sequences X_i, $i = 1, \ldots, I$.

[5]For instance, by training up such a net from a labeled database like TIMIT.

System	Error Rate	Posterior
Baseline Hybrid	3.4%	-
DHMM, pre-REMAP	2.7%	0.1269
1 REMAP iteration	2.5%	0.1731
2 REMAP iterations	2.5%	0.1773

Table 1. Training and testing on noisy isolated digits.

4. EXPERIMENTS AND RESULTS

We report on experiments with isolated and continuous speech, where recognition was based on acoustic information. The isolated speech recognition task we started with is the Digits+ corpus in use at ICSI, which is a subset of a larger database recorded over a clean telephone line at Bellcore. It is composed of 200 speakers saying the words "zero" through "nine", "oh", "no", and "yes". For the additive noise in these experiments, we used automotive sound that was recorded over a cellular telephone. Noise was randomly selected from this source and then added to the clean speech waveforms (10db S/N ratio). In order to better utilize the data we use a jack-knife procedure. For each of four experiments, three fourths of the data was used for training and cross-validation, and one fourth was used for testing. Specifically, in each experiment we use 1720 utterances for training, 230 for cross-validation and 650 (from 50 speakers) for testing. All our nets have 214 inputs: 153 inputs for the acoustic features, and 61 to represent the previous state (one unit for every possible previous state). The acoustic features are combined from 9 frames with 17 features each (RASTA-PLP8 + delta features + delta log gain) computed with an analysis window of 25ms computed every 12.5 ms (overlapping windows) and the sampling rate was 8Khz. The nets have 200 hidden units and 61 outputs. The combined results for all the four cuts are summarized in Table 1. Note that the row entitled "Baseline Hybrid" refers to an ANN trained on targets of 1's and 0's that have been obtained from a forced Viterbi procedure by our standard HMM/ANN system as described in [2]; the row entitled "DHMM, pre-REMAP" means a Discriminant HMM using the same training approach, with hard targets determined by the first system, and additional inputs to represent the previous state. The rightmost column gives the average probability of the correct model over all test words as determined during recognition. Our recognition rate after the

System	Error Rate
DHMM, pre-REMAP	14.9%
1 REMAP iteration	13.6%
2 REMAP iterations	13.2%

Table 2. Training and testing on continuous numbers, no syntax, no durational models.

first and second iterations of REMAP is significantly better (at $p < 0.05$ level) than the baseline hybrid system. Although for this task the contribution of the REMAP step is small, combining it with the transition-based, posterior framework as done in the Discriminant HMM, gives a significant improvement.

Our next step was to test whether this improved performance can also be obtained with continuous speech. For this purpose we chose the Numbers'93 corpus. It is a continuous-speech database collected by CSLU at the Oregon Graduate Institute. It consists of numbers spoken naturally over telephone lines on the public-switched network [10]. The Numbers'93 database consists of 2167 speech files of spoken numbers produced by 1132 callers. We used 877 of these utterances for training and 657 for cross-validation and testing (200 for cross-validation) saving the remaining utterances for final testing purposes. There are 36 words in the vocabulary, namely *zero, oh, 1, 2, 3,...,20, 30, 40, 50,...,100, 1000, a, and, dash, hyphen,* and *double*. Our results are summarized in Table 2.

The improvement in the recognition rate as a result of REMAP iterations is significant at $p < 0.05$. However all the experiments were done using acoustic information alone. Using our (baseline) hybrid system under equal conditions, i.e., no duration information and no language information, we get 31.6% word error; adding the duration information back we get 12.4% word error. We are currently experimenting with enforcing minimum duration constraints to our framework.

5. LANGUAGE MODEL

Starting from (8), and assuming that given the state sequence and the language model, we can omit the dependence on the acoustic sequence, we get:

$$P(M_i|X,\Theta,L) \approx \sum_{\Gamma} P(\Gamma|X,L,\Theta)P(M_i|\Gamma,L,\Theta) \qquad (12)$$

Using Bayes' rule, we also have:

$$P(M_i|X,\Theta,L) \approx \sum_{\Gamma} P(\Gamma|X,\Theta,L)\frac{P(\Gamma|M,\Theta L)P(M|L,\Theta)}{P(\Gamma|L,\Theta)}$$

If we further assume that the effect of the language model can be ignored in the acoustic term, i.e., $P(\Gamma|X,\Theta,L) \approx P(\Gamma|X,\Theta)$, we finally have:

$$P(M_i|X,\Theta,L) \approx \sum_{\Gamma} P(\Gamma|X,\Theta)\frac{P(\Gamma|M,L,\Theta)P(M|L,\Theta)}{P(\Gamma|L,\Theta)}$$

in which $P(\Gamma|X,\Theta)$ is computed as described in Section 2. $P(M|L,\Theta)$ can be assumed independent of the acoustic model parameters and can be estimated using standard language modeling techniques. In principle $P(\Gamma|M,L,\Theta)$ and $P(\Gamma|L,\Theta)$ can be estimated during training by dynamic programming techniques similar to our α and β recurrences [4], and the ratio of these two terms represents the additional state transition information that is gained by knowing the specific word sequence.

6. CONCLUSIONS AND FUTURE WORK

We presented a discriminant training algorithm for hybrid HMM/ANN systems based on a global MAP criterion. Our results on small isolated and continuous speech recognition tasks show an increase in the estimates of the posterior probabilities of the correct sentences after training, and significant decreases in error rates in comparison to a baseline system. However, all of our experiments were done using acoustic information only. We have also described a way to incorporate the language model in our framework, which will be tested in the near future.

Acknowledgments

We would like to thank Kristine Ma and Su-Lin Wu for their help with the DIGITS+ database. We gratefully acknowledge the support of the Office of Naval Research, URI No. N00014-92-J-1617 (via UCB), the European Commission via ESPRIT project 20077 (SPRACH), and ICSI and FPMs in general for supporting this work.

REFERENCES

[1] R. O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, Inc, 1973.

[2] H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach.* Kluwer Academic Publishers, 1994.

[3] J. M. Steeneken and D. H. Van Leeuwen. Multilingual assessment of speaker independent large vocabulary speech-recognition systems: the SQALE project (speech recognition quality assessment for language engineering). In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, Madrid, Spain, September 1995.

[4] H. Bourlard, Y. Konig, and N. Morgan. REMAP: Recursive estimation and maximization of a posteriori probabilities, application to transition-based connectionist speech recognition. Technical Report TR-94-064, International Computer Science Institute, Berkeley, CA, 1994.

[5] H. Bourlard, Y. Konig, and N. Morgan. REMAP: Recursive estimation and maximization of a posteriori probabilities in connectionist speech recognition. In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, Madrid, Spain, September 1995.

[6] J. R. Glass. *Finding Acoustic Regularities in Speech Applications to Phonetic Recognition.* PhD thesis, M.I.T, May 1988.

[7] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe. Global optimization of a neural network-hidden Markov model hybrid. *IEEE trans. on Neural Networks*, 3(2):252–258, March 1992.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B*, 34:1–38, 1977.

[9] Y. Bengio and P. Frasconi. An input output HMM architecture. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT press, Cambridge, 1995.

[10] R.A. Cole, M. Fanty, and T. Lander. Telephone speech corpus development at CSLU. In *Proceedings Int'l Conference on Spoken Language Processing*, Yokohama, Japan, September 1994.