REMAP: RECURSIVE ESTIMATION AND MAXIMIZATION OF A POSTERIORI PROBABILITIES IN CONNECTIONIST SPEECH RECOGNITION

Hervé Bourlard [‡,†], *Yochai Konig* [‡,*], *and Nelson Morgan* [‡,*]

[‡] International Computer Science Institute, Berkeley, CA
[†] Faculté Polytechnique de Mons, Mons, Belgium
[*] University of California at Berkeley, Berkeley, CA
Emails: bourlard, konig, morgan@icsi.berkeley.edu

ABSTRACT

In this paper, we briefly describe REMAP, an approach for the training and estimation of posterior probabilities, and report its application to speech recognition. REMAP is a recursive algorithm that is reminiscent of the EXPECTATION MAXIMIZATION (EM) [5] algorithm for the estimation of data likelihoods. Although very general, the method is developed in the context of a statistical model for transition-based speech recognition using ARTIFICIAL NEURAL NETWORKS (ANN) to generate probabilities for HIDDEN MARKOV MODELS (HMMs). In the new approach, we use local conditional posterior probabilities of transitions to estimate global posterior probabilities of word sequences. As with earlier hybrid HMM/ANN systems we have developed, ANNs are used to estimate posterior probabilities. In the new approach, however, the network is trained with targets that are themselves estimates of local posterior probabilities. Initial experimental results support the theory by showing an increase in the estimates of posterior probabilities of the correct sentences after REMAP iterations, and a decrease in error rate for an independent test set.

1. INTRODUCTION

Today, 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 $P(X|M,\Theta)$, in which $X = \{x_1, \dots, x_n, \dots, x_N\}$ is a sequence of acoustic vectors, M is a HMM, and Θ is the parameter set on which optimization is performed. The goal of the recognition process is to determine the most probable sequence of words M given what has been uttered X. The optimal solution to this problem is given by the MAXIMUM A POSTERIORI (MAP) criterion based on the MAP probability $P(M|X,\Theta)$. Classically, the likelihood formulation of the problem is obtained by applying Bayes' rule:

$$P(M|X,\Theta) = \frac{P(X|M,\Theta)P(M|\Theta)}{P(X|\Theta)} \qquad (1)$$

For practical reasons, it is assumed that:

1. The parameters of $P(X|M)$ (acoustic model) are independent of the parameters of $P(M)$ (language model). Consequently, these probabilities are respectively denoted $P(M|X,\Theta)$ and $P(M|\Theta^*)$ and are estimately independently of each other. The full consequences of this assumption fall outside the scope of this paper. However, it is clear that trained acoustic probabilities will be affected by the prior probability of sequences of linguistic units as represented in the acoustic training data.
2. The denominator $P(X)$ does not depend on any of the parameters Θ or Θ^*. This assumption is reasonable during recognition. However, since $P(X|\Theta)$ can be expressed as a sum of joint acoustic and model probabilities, the value of the denominator of (1) will be modified during training, and ignoring this will reduce discrimination between correct and incorrect models.

In recent years there has been a significant body of work, both theoretical and experimental, that has established the viability of ARTIFICIAL NEURAL NETWORKS (ANNs) as a useful technology for speech recognition. In particular, we have shown that fairly simple layered structures, which we lately have termed BIG DUMB NEURAL NETWORKS (BDNNs), can be used to estimate local probabilities for HMMs [3]. This approach is now usually referred to as a HYBRID HMM/ANN SYSTEM. Although the theoretical architecture (which we have called the DISCRIMINANT HMM) was initially developed for 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 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, on 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. Recently, such a system has been evaluated under both the North American ARPA program and the European LRE SQALE project (20,000 word vocabulary, speaker independent continuous speech recognition). In the preliminary results of the SQALE evaluation (reported in [7]) the system was found to perform slightly better than any other leading European system and required an order of magnitude less CPU resources to complete the test.

The initial Discriminant HMM theory has recently been extended to accommodate full MAP training of hybrid HMM/ANN systems. In this paper, we present this new hybrid HMM/ANN approach that directly optimizes the 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), generates successive estimates of new (local) posterior probabilities as targets for an ANN training step to guarantee an iterative increase of the global posteriors. We show in [2] that estimation of the new ANN targets can be done using "forward" and "backward" recurrences that are reminiscent of the EXPECTATION MAXIMIZATION (EM) algorithm. In the experiments reported here, we use a modified approach that only uses a "forward" recurrence for both training and recognition.

Unlike most previous hybrid HMM/ANN systems that we and others have developed, the new formulation determines the most probable word sequence, rather than the utterance corresponding to the most probable state sequence. Also, in addition to using all possible state sequences, the proposed training algorithm uses posterior probabilities at both local and global levels and is discriminant in nature.

2. DISCRIMINANT HMM

In [3], summarizing earlier work (such as [4]), we showed that it was possible to compute the global a posteriori probability $P(M|X,\Theta)$ of a Discriminant HMM M given an acoustic vector sequence $X = \{x_1, \ldots, x_n, \ldots, x_N\}$ as:

$$P(M|X,\Theta) = \sum_{\forall\Gamma_j} P(M, q_{j,1}, q_{j,2}, \ldots, q_{j,N}|X,\Theta) \qquad (2)$$

in which "$\forall\Gamma_j$" represents all possible (legal) state sequences in M, $q_{j,n}$ 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. Each term in (2) can further be decomposed into:

$$P(q_{j,1}, q_{j,2}, \ldots, q_{j,N}|X,\Theta) P(M_i|q_{j,1}, q_{j,2}, \ldots, q_{j,N}, X, \Theta) \qquad (3)$$

and, under the assumptions stated in [3], we have

$$P(q_{j,1}, q_{j,2}, \ldots, q_{j,N}|X,\Theta) = \prod_{n=1}^{N} P(q_{j,n}|q_{j,n-1}, x_n, \Theta) \qquad (4)$$

The second factor in (3) can be considered independent of the acoustic sequence X (since the state sequence is given). As discussed in [2], depending on what we encode into the acoustic models, this latter factor will represent phonological, lexical and/or syntactical information.

Discriminant HMMs are thus 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[1] 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 will 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.

Additionally, for these conditional transition probabilities, there is a disparity between the training input space of the ANN and the input space that may be hypothesized during recognition. For this case, the ANN input space includes both the local acoustic vectors and the previous state category. During training, the network only processes input consisting of "correct" pairs of acoustic vectors and correct previous state, while in recognition the net should generalize to all possible combinations of acoustic vectors and previous states (since, during dynamic programming, all transitions permitted by the HMM topologies will be hypothesized for each acoustic vector in X). However, some hypothesized inputs may correspond to an impossible condition that will never have been observed, such as the acoustics of the temporal center of a vowel in combination with a previous state that corresponds to a plosive. It is unfortunately possible that the interpolative capabilities of the network may not be sufficient to give these "impossible" pairs a sufficiently low probability during recognition [2]. This problem is can be viewed as a lack of negative examples (i.e., impossible transitions for some given acoustic data).

One possible solution to these problems 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

$$\prod_{i=1}^{I} P(M_i|X_i,\Theta) \qquad (5)$$

in which M_i represents the Markov model associated with each training utterance X_i, with $i = 1, \ldots, I$.

Although, in principle, we could use a generalized back-propagation-like gradient procedure in Θ to maximize (5) (see, e.g., [1]), 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 $\forall n \in [1,n]$.

[1] 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.

2. when 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 (5)?

In [2], we prove that a re-estimate of ANN targets that guarantee convergence to a local maximum of (5) is given by[2]:

$$P^*(q_n^\ell|x_n, q_{n-1}^k) = P(q_n^\ell|X, q_{n-1}^k, \Theta^t, M) \qquad (6)$$

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.

In [2], we further prove that alternating ANN target estimation (the "estimation" step) and ANN training (the "maximization" step) is guaranteed to incrementally increase (5) over t.[3]

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:

$$P(q_n^\ell|X, q_{n-1}^k, M) = \frac{p(q_{n-1}^k, q_n^\ell, M|X)}{\sum_\ell p(q_{n-1}^k, q_n^\ell, M|X)} \qquad (7)$$

The terms on the right hand side can be computed from α (forward) or β (backward) EM-like recurrences using only local conditional transition probabilities. For the experiments reported in this paper, we used a forward recursion only, in which $\alpha_n(k) = \sum_{\Gamma_n \in M_i} P(\Gamma_n, q_n^k|X_1^n)$ can be expressed in terms of $\alpha_{n-1}(k)$'s (where q^k's are possible predecessor states of q^ℓ in M) and local conditional transition probabilities; Γ_n refers to the set of subsequences associated with the first n frames. For training, this recursion is modified to only permit contributions from paths with a particular transition (k, l) at time n in order to compute the terms required for (7).

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[4]:

1. Compute ANN targets $P(q_n^\ell|X_j, q_{n-1}^k, \Theta^t, M)$ according to (7), $\forall$ 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. In this regards, it is reminiscent of the EM algorithm. However, EM is an iterative approach to maximum likelihood estimation, while REMAP is an iterative approach to maximum a posteriori probability estimation. Also, 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.

Convergence of this training scheme can however be proved [2]. As for the EM, the convergence proof relies on the definition of an auxiliary function with the following properties:

1. When increased, the global MAP is also guaranteed to increase.
2. For a given (fixed) set of parameters Θ^t, cancelling the partial derivative of that function with respect to the conditional transition probabilities (i.e., actually maximizing the auxiliary function) yields new targets (6).
3. Training the net with these new targets (which, of course, won't be precisely reached) guarantees an increase of the auxiliary function and, consequently, of the global posteriors.

4. EXPERIMENTS AND RESULTS

We have begun to test this theoretical formulation on practical tasks. The speech recognition task we started with is the Digits+ corpus in use at ICSI. It is composed of 200 speakers saying the words "zero" through "nine", "oh", "no", and "yes". Each word was recorded in isolation over a clean telephone line at Bellcore. 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 our pilot 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.

Results for the pilot test set are summarized in Table 1. Note that the row entitled "Classic Hybrid" refers to an ANN trained on targets that are 1's and 0's that have been obtained from a forced Viterbi procedure by our standard HMM/ANN system as described in [3]; the row entitled "Disc. HMM, 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.

As predicted by the theory, Table 1 shows an increase of the posterior probability for each iteration, accompanied by a decrease in error rate. Since the sum of all possible model posteriors is equal to one (which is not true for data likelihoods), an increase in posteriors for

[2] 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, $i = 1, \ldots, I$.

[3] Note here that one "iteration" does not stand for one iteration of the ANN training but for one estimation-maximization iteration for which a complete ANN training will be required.

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

System	Error Rate	Posterior
Classical Hybrid	3.1%	-
Disc. HMM, pre-REMAP	2.9%	0.108
1 REMAP iteration	2.3%	0.161
2 REMAP iterations	2.3%	0.175
3 REMAP iterations	2.2%	0.180

Table 1: Training and testing on noisy speech.

the correct class means that the posteriors for alternative models decrease. Inspection of the posteriors for the incorrect digit models show that REMAP does indeed decrease their probability while increasing the probability of the correct models. For this case at least, a single REMAP iteration appears to be sufficient to accomplish this improvement.

5. DISCUSSION AND FUTURE WORK

A wide range of discriminant approaches (e.g., MMI, GPD) to speech recognition have been studied by researchers. A significant difficulty that has remained in applying these approaches to continuous speech recognition has been the requirement to run computationally intensive algorithms on all of the rival sentences. Since this is not generally feasible, compromises must always be made in practice. For instance, estimates for all rival sentences can be derived from a list of the "N-best" utterance hypotheses, or by using an ergodic model containing all possible phonemes. While thus far we have only applied REMAP to isolated word recognition, for which all rival sentences can be considered, the techniques should also apply to continuous speech without the need for such approximations.

To summarize our current results:

- We have a method for MAP training and estimation of sequences.
- This can be used in a new form of hybrid system using HMMs and an estimator of posterior probabilities. A convenient estimator, which we have used, is a neural network trained with back propagation to minimize the relative entropy between target distributions and network outputs. As with the standard HMM/ANN hybrid approach, the estimated probabilities are local (conditioned on the local acoustic signal, though also conditioned on the previous state in the discriminant HMM case). However, in the case of REMAP the network estimators are trained with probabilistic targets that are themselves estimates of posterior probabilities.
- Initial experiments actually show an increase in global posteriors, accompanied (as expected) by a reduction in error rate for this process.

Much work is still required to optimize the practical heuristics for this method. While our results look promising, their improvement over our older approach is not statistically significant (at the $p < .05$ level, assuming a normal approximation to the binomial distribution for the error); this was a pilot experiment in which we only had 650 test utterances. However, every measure we examined (error rate, average posterior probability of correct models, average posterior probabilities of incorrect models, number of words with increased posteriors for the correct model) improved with each iteration and was consistent with our expectations from the theory.

This result was obtained with a system in which we have not yet optimized the new method for a number of factors (e.g., the input window size) that have long been optimized for the older hybrid system. We also have occasionally made some simplifying assumptions in order to implement this algorithm; we need to study the effect of these choices. Once we have done these optimizations and extended the evaluation to the full Digits+ task of 2600 utterances (using a jackknife approach so that we can test on all utterances independently from the training set), we will extend the current work to continuous speech recognition and explore the use of language models with REMAP. The current status, as shown in this article, is that new theoretical groundwork has been established, and that an implementation does appear to improve recognition in at least one small but nontrivial speech recognition task.

Acknowledgments

We would like to thank Kristine Ma and Su-Lin Wu for their help with the DIGITS+ database. The earlier (mid-80's) work in discriminant HMMs on which this is based was done in collaboration with Chris Wellekens. We gratefully acknowledge the support of the Office of Naval Research, URI No. N00014-92-J-1617 (via UCB), ESPRIT project 6487 (WERNICKE) (through ICSI), and ICSI in general for supporting this work.

6. REFERENCES

[1] Bengio, Y. R., De Mori, R., Flammia, G., & Kompe, R. (1992). "Global optimization of a neural-hidden Markov model hybrid," *IEEE Trans. on Neural Networks*, vol. 3, pp. 252-258.

[2] Bourlard, H., Konig, Y., & Morgan, N. (1994). "REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities – Application to Transition-Based Connectionist Speech Recognition," ICSI Technical Report TR-94-064, Intl. Computer Science Institute, Berkeley, CA.

[3] Bourlard, H. and Morgan, N. (1994). Connectionist Speech Recognition – A Hybrid Approach, Kluwer Academic Publishers.

[4] Bourlard, H. and Wellekens, C.J. (1990). "Links between Markov models and multilayer perceptrons," IEEE Trans. on PAMI, vol. 12, no. 12, pp. 1167-1178.

[5] Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, pp. 1-38.

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

[7] Steeneken, J.M. and Van Leeuwen, D.A. (1995). "Multi-Lingual Assessment of Speaker Independent large vocabulary speech-recognition systems: the SQALE project (speech recognition quality assessment for language engineering)," Proceedings of EUROSPEECH'95 (Madrid), September 1995.