A NEW TRAINING ALGORITHM FOR HYBRID HMM/ANN SPEECH RECOGNITION SYSTEMS

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

[†]Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez, B-7000 Mons, Belgium
Email: bourlard@tcts.fpms.ac.be
and
[‡]International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA.

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

Our previous hybrid HMM/ANN approach [2] used ANNs to generate local posteriors that were divided by priors to get scaled likelohoods for use in standard HMMs[1]. After many years of development, that approach has been shown quite successful on large speech recognition problems as test on large reference databases[2].

The new training algorithm discussed in this paper, referred to as REMAP (Recursive Estimation and Maximization of A Posteriori Probabilities), uses local conditional posterior probabilities of transitions (estimated by a particular form of ANN) to maximize during training (or estimate during recognition) global posterior probabilities of word sequences. Hence minimizing global posterior probabilities of rival word sequences and, in theory, the error rate. Thus, although we still use ANNs to estimate posterior probabilities, these are no longer divided by priors and the network is now trained with targets that are themselves local posterior probabilities that are iteratively re-estimated to guarantee a monotonic increase of the global posterior probabilities of the correct sentences. This kind of algorithm is thus expected to have better global discriminant properties than the previously developed hybrid HMM/ANN systems. On top of better discriminant properties, we believe that the model could also have additional potential benefits in terms of its modeling properties.

2 APPROACH

If M is a model and X a speech utterance, the optimal training and recognition criterion of a statistical classifier should be based on $P(M|X,\Theta)$, where Θ is the set of parameters. However, in standard HMM approaches, this is often replaced by a training based on $P(X|M,\Theta)$ (with independant training of $P(M|\Theta^*)$), resulting in likelihood training and maximum a posteriori decoding.

In [2] we show that it is however possible to express $P(M|X,\Theta)$ in terms of local "conditional transition probabilities" $P(q_n^\ell|q_{n-1}^k,x_n,\Theta)$, in which q_n^ℓ stands for the specific class q^ℓ (associated with a particular

[1]We mainly used multilayer perceptrons as ANNs, but recurrent neural networks have also been used successfully by others [8].

[2]The most recent results, those of the EU funded SQALE evaluations, show the hybrid approach slightly ahead of more traditional HMM systems. The hybrid system was evaluated on both British and American English tasks, using a 20,000 word vocabulary and a trigram language model, along with the other leading European systems produced by LIMSI (France), Philips (Germany) and Cambridge University/HTK (UK) [9]. Additionally, the hybrid system was efficient in its runtime CPU and memory requirements.

ANN output) hypothesized at time n, and x_n for the acoustic vector at time n as described in [2]. Initially referred to as "Discriminant HMM", such a system could also be referred[3] to as Stochastic Finite State Acceptor (SFSA).

An example of SFSA using conditional transition probabilities is given in Figure 1.

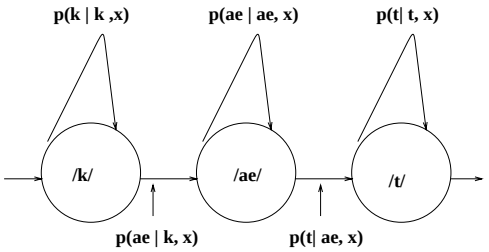

Figure 1: **An example of SFSA for the word "cat". The variable x refers to a specific acoustic observation x_n at time n.**

The conditional transition probabilities defined above can be estimated (and parametrized) by a particular form of ANN (Figure 2) with the acoustic data and the previous state (as hypothesized by the topology of M) at the input. Θ represents then the set of ANN parameters.

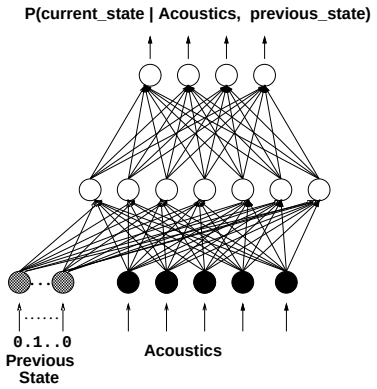

Figure 2: **An ANN (multilayer perceptron) that estimates local conditional transition probabilities**

REMAP has been developed to iteratively estimate the (ANN) parameters of the SFSA according to a global Maximum a Posteriori (MAP) criteria, i.e., to maximize the posterior probability of the correct sentence while, by definition of posterior probabilities, reducing the posterior probabilities of all the rival models. In [3] we describe the mathematical theory behind REMAP, and proved its convergence properties. In the full paper, we will briefly review the algorithm before reporting on our recent experimental results.

[3]Thanks to a suggestion of John Bridle.

The main motivations behind this work are manyfold. First, we wanted to extend our hybrid HMM/ANN system by returning to earlier Discriminant HMM theory [2]. Secondly, we wanted to have an algorithm to train these models in a globally discriminant way, without having to divide the ANN outputs by priors to estimate scaled likelihoods for "standard" HMMs. Finally, we are interested in working with transition-based recognizers, in which case we want to get smooth estimates of conditional transition probabilities to train an ANN. Finally, as discussed in Section 5, we believe that such a model could have additional modeling properties.

3 REMAP TRAINING of SFSA

3.1 Motivations

The SFSA theory as described above uses transition-based probabilities as the key building block for acoustic recognition. However, it is well known that estimating transitions accurately is a difficult problem [6]. Due to the inertia of the articulators, the boundaries between phones are blurred and overlapped in continuous speech. In our baseline hybrid HMM/ANN system, targets were typically obtained by using a standard forced Viterbi alignment (segmentation). For a transition-based system as defined above, this procedure would thus yield rigid transition targets, which is not realistic.

Another problem related to the Viterbi-based training of the ANN probability estimator is the lack of coverage of the input space during training. Indeed, during training (based on hard transitions), the ANN only processes inputs 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 all possible models and transitions will be hypothesized for each acoustic input. For example, some hypothesized inputs may correspond to an impossible condition that has thus never 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.

One possible solution to these problems is to use a full MAP algorithm to find transition probabilities at each frame for all possible transitions by a forward-backward-like algorithm [7], taking all possible paths into account.

3.2 Problem Formulation

As described above, global maximum a posteriori training of SFSAs should find the optimal parameter set Θ maximizing

$$\prod_{j=1}^{J} P(M_j|X_j, \Theta) \qquad (1)$$

in which M_j represents the Markov model associated with each training utterance X_j, with $j = 1, \ldots, J$.

Suppose that we are given a trained ANN at iteration t providing a parameter set Θ^t and, consequently, estimates of $P(q_n^\ell | x_n, q_{n-1}^k, \Theta^t)$. In this case, training of the SFSA by a global MAP criterion requires new ANN targets that satisfy the following two conditions:

1. They must be smooth estimates of conditional transition probabilities $q_{n-1}^k \rightarrow q_n^\ell$, $\forall k, \ell \in [1, K]$ and $\forall n \in [1, N]$, and

2. when training the ANN for iteration $t+1$, they must 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 the global posterior probability $P(M_i | X, \Theta)$.

In [3], we prove that a re-estimate of ANN targets that guarantee convergence to a local maximum of (1) is given by:

$$P^*(q_n^\ell | x_n, q_{n-1}^k, M) = P(q_n^\ell | X, q_{n-1}^k, \Theta^t, M) \qquad (2)$$

where we have estimated the left-hand side using a mapping from the previous state and the local acoustic data to the current state, thus making the estimator realizable by an ANN with a local acoustic window. Thus, we will want to estimate the transition probability conditioned on the *local* data (as ANN targets) by using the transition probability conditioned on *all* of the data.

In [3], we further prove that alternating ANN target estimation (the "estimation" step) and ANN training (the "maximization" step) is guaranteed to incrementally increase (1) over t.[4]

The remaining problem is to find an efficient algorithm to express $P(q_n^\ell | X, q_{n-1}^k, M)$ in terms of $P(q_n^\ell | x_n, q_{n-1}^k)$ so that the next iteration targets can be found. In [3] we shown that there are efficient (forward-backward like) recursions to do this in terms of all possible paths through the SFSA.

3.2.1 REMAP Training

The general scheme of the REMAP training of SFSA-based systems can be summarized as follows:

1. Start 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].

2. Compute ANN targets $P(q_n^\ell | X_j, q_{n-1}^k, \Theta^t, M_j)$ $\forall$ training sentences X_j associated with SFSA M_j, $\forall$ possible (k, ℓ) state transition pairs in M_j and $\forall$ x_n, $n = 1, \ldots, N$ in X_j (see next point).

3. For every x_n in the training database, train the ANN to minimize the relative entropy between the outputs and targets. See [3] for more details. This provides us with a new set of parameters Θ^t, for $t = t + 1$.

4. Iterate from 2 until convergence.

This procedure is thus composed of two steps: an Estimation (E) step, corresponding to step 2 above, and a Maximization (M) step, corresponding to step 3 above. In this regards, it is reminiscent of the Estimation-Maximization (EM) algorithm as discussed in [5]. 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 by Bengio and Frasconii for mapping input sequences to output sequences [1].

4 EXPERIMENTS

Recognition is based on the REMAP forward recurrences to compute $P(M|X)$ from $P(q_n^\ell | q_{n-1}^k, x_n, \Theta)$. We report on experiments with isolated and continuous speech based on acoustic information. The isolated speech recognition task we started with is the Digits+ corpus, 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 by dividing the data into four equal cuts. For each experiment, three cuts were used for training and cross-validation, and one cut was used for testing. The combined results for all the four cuts are summarized in Table 1. Note that the row entitled "Classic 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 "Dis. HMM, pre-REMAP" means a SFSA 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 first and second iterations of REMAP is significantly better (at $p < 0.05$

[4] 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.

[5] This can be done, for instance, by training up such a net from a hand-labeled database like TIMIT or from some initial forward-backward estimator of equivalent local probabilities (usually referred to as "gamma" probabilities in the Baum-Welch procedure).

System	Error Rate	Posterior
Classical Hybrid	3.4%	-
Dis. HMM, 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.

level) than the classical 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 SFSA, gives overall significant improvement.

5 DISCUSSIONS

5.1 Better Discrimination

As described above, REMAP corresponds to an iterative approach that is guaranteed to increase the posterior probability of the correct utterance. Since the total probability of all possible utterances is 1, the resulting system may be considered discriminant, even though the system has not observed most possible incorrect sequences. The approach uses a probability estimator that is trained to distinguish between categories associated with states, and thus is a local estimator. In other words, a formalism has been developed that permits estimators to be trained for discrimination between local windows of an observed feature sequence in such a way that they will also be optimized for discrimination of the correct complete sequence (e.g., sentence) from rival candidates. The training is based on an iterative procedure that is reminiscent of the Baum-Welch procedure used for estimating sequence likelihoods [3].

5.2 Better Modeling Properties

In addition to the better discriminant properties discussed above, SFSA/REMAP systems have many other potential advantages that should be investigated here, including:

- Better modeling capabilities (as opposed to pure speech description) — It is indeed clear that current HMM improvements are often achieved simply by increasing the size of the training databases, the number of context-dependent phonemes, and the number of underlying parameters. Since in SFSAs the observations, as well as all possible factors that could influence those observations or correlate with them, appear as conditionals of local probabilities, it may be expected that many dependencies could be captured without having to increase the number of basic speech units. For example, this could be the case with (phonetic) context dependencies and speech characteristics like speech rate and vocal effort.

- The formalism used for SFSAs and their (discriminant) training algorithm yield a model where transition probabilities as well as some parts of the language model information are intimately linked, it can be expected that we do not need some of the scaling factors usually used (and empirically optimized) in standard HMM approaches.

References

[1] Bengio, Y., and Frasconi, P., "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.

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

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

[4] Cole, R.A., Fanty, M., and Lander, T. (1994). "Telephone Speech Corpus Development at CSLU," *Proc. Intl. Conf. on Spoken Language Processing* (Yokohama, Japan).

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

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

[7] Liporace, L.A., "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. on Information Theory*, vol. IT-28, no. 5, pp. 729-734, 1982.

[8] Robinson, T., Hochberg, M. and Renals, S., "The use of recurrent networks in continuous speech recognition," *Automatic Speech and Speaker Recognition – Advanced Topics*, (C.H. Lee, K.K. Paliwal and F. K. Soong, eds.), Kluwer Academic Publishers, 1996.

[9] 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 (European Conference on Speech Technology* (Madrid), pp. 1271- 1274, September 1995.