ESTIMATION OF GLOBAL POSTERIORS AND FORWARD-BACKWARD TRAINING OF HYBRID HMM/ANN SYSTEMS

J. Hennebert(5,2), C. Ris(1), H. Bourlard(3,2), S. Renals(4) and N. Morgan(2)

(1) TCTS, FPMs, B-7000 Mons, Belgium
(2) ICSI, Berkeley CA 94704, USA
(3) IDIAP, 1920 Martigny, Switzerland
(4) Computer Science, University of Sheffield, Sheffield S1 4DP, UK
(5) CIRC, EPFL, 1015 Lausanne, Switzerland

ABSTRACT

The results of our research presented in this paper are two-fold. First, an estimation of global posteriors is formalized in the framework of hybrid HMM/ANN systems. It is shown that hybrid HMM/ANN systems, in which the ANN part estimates local posteriors, can be used to modelize global model posteriors. This formalization provides us with a clear theory in which both REMAP and "classical" Viterbi trained hybrid systems are unified. Second, a new forward-backward training of hybrid HMM/ANN systems is derived from the previous formulation. Comparisons of performance between Viterbi and forward- backward hybrid systems are presented and discussed.

1. INTRODUCTION

In [1, 2] it was shown that it is possible to express the global posterior probability $P(M|X,\Theta)$ of a model (stochastic finite state acceptor) M given the acoustic data X and the parameters Θ in terms of the local posteriors (conditional transition probabilities) $P(q_l^n|q_k^{n-1}, x_n, \Theta)$ (where q_k^n denotes the specific state q_k of M at time n) and the language model priors. An application of the generalized EM algorithm applied to stochastic finite acceptors, known as REMAP, was introduced to iteratively estimate the parameter set Θ. The global posterior probability of the correct model can be optimized by optimizing the local posterior probabilities through re-estimating targets for the ANN probability estimator.

In this paper: (1) we demonstrate that the original HMM/ANN system [3, 4] trained using local criteria indeed optimizes the global posterior probability, given certain well-defined assumptions; (2) we use the REMAP algorithm to derive a forward-backward training algorithm for the original HMM/ANN system; (3) we demonstrate the performance of these algorithms on the task-independent Phonebook database.

2. ESTIMATION OF GLOBAL POSTERIORS

2.1. REMAP formulation

The objective of the REMAP formulation is to produce an estimate of the global posterior probability of a model M given the acoustic data $X = X_1^N = \{x_1, x_2, \ldots, x_N\}$ (and the parameter set Θ):

$$P(M|X) = \sum_{\ell_1=1}^{L} \cdots \sum_{\ell_N=1}^{L} P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N, M|X) \quad (1)$$

where $q_{\ell_n}^n$ is HMM state ℓ_n visited at time n, and the summation is over all possible state sequences (the Viterbi approximation maximizes over the best state sequences).

If we consider a particular state sequence, the posterior probability of the state sequence and the model may be decomposed into the product of an acoustic model and a prior over models ("language model" and state sequences):

$$\begin{aligned} & P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N, M|X) \\ & \quad = P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N|X) P(M|X, q_{\ell_1}^1, \ldots, q_{\ell_N}^N) \\ & \quad \simeq \underbrace{P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N|X)}_{\text{ac.model}} \underbrace{P(M|q_{\ell_1}^1, \ldots, q_{\ell_N}^N)}_{\text{prior}} \quad (2) \end{aligned}$$

The X dependence in the second factor in (2) is dropped since the hidden part (the state sequence) is hypothesized. With the usual assumptions of a first-order Markov process and conditionals on X limited to local context X_{n-c}^{n+c} we can simplify the two factors in (2):

$$\begin{aligned} P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N|X) &= P(q_{\ell_1}^1|X) P(q_{\ell_2}^2|X, q_{\ell_1}^1) \ldots \\ & \quad \ldots P(q_{\ell_N}^N|X, q_{\ell_1}^1, \ldots, q_{\ell_{N-1}}^{N-1}) \\ &= \prod_{n=1}^{N} P(q_{\ell_n}^n|X, Q_1^{n-1}) \end{aligned}$$

$$\simeq \prod_{n=1}^{N} P(q_{\ell_n}^n | X_{n-c}^{n+d}, q_{\ell_{n-1}}^{n-1}) \quad (3)$$

$$\begin{aligned} P(M|q_{\ell_1}^1, \ldots, q_{\ell_N}^N) \\ &= \frac{P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N | M_i) P(M_i)}{P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N)} \\ &\simeq P(M) \left[\prod_{n=1}^{N} \frac{P(q_{\ell_n}^n | q_{\ell_{n-1}}^{n-1}, M)}{P(q_{\ell_n}^n | q_{\ell_{n-1}}^{n-1})} \right] \quad (4) \end{aligned}$$

Using these simplifications we can approximate (1):

$$P(M|X) \simeq P(M) \sum_{\ell_1, \ldots, \ell_N} \left[\prod_{n=1}^{N} P(q_{\ell_n}^n | X_{n-c}^{n+d}, q_{\ell_{n-1}}^{n-1}) \frac{P(q_{\ell_n}^n | q_{\ell_{n-1}}^{n-1}, M)}{P(q_{\ell_n}^n | q_{\ell_{n-1}}^{n-1})} \right] (5)$$

and the Viterbi approximation may be obtained by replacing the sum over state sequences $(\ell_1, \ldots, \ell_N)$ with a maximization. This formulation has two sets of prior probabilities where $P(q_\ell^n | q_k^{n-1})$ represent the training data priors and $P(q_\ell^n | q_k^{n-1}, M)$ the Markov model priors. The training data priors are independent of the HMM topology. The Markov model priors are actually the so-called transition probabilities between states. Assuming time-invariant HMMs (as usually done in standard HMMs), these priors can be written as $P(q_\ell | q_k)$ and $P(q_\ell | q_k, M)$.

The REMAP [1] training algorithm uses local conditional transition probabilities $P(q_k^n | X, q_l^n - 1)$ (estimated by a particular form of MLP) to maximize during training (or estimate during recognition) the global posterior probability of the word sequences.

2.2. Original HMM/ANN system

The above formulation was derived in the context of stochastic finite state acceptor models (also known as discriminative HMMs). However, by removing the dependency on the previous state in (5) we arrive at a hybrid system similar to those previously developed, (e.g., in [3] and [4]). In this case, (5) becomes:

$$P(M|X) \simeq \sum_{\ell_1, \ldots, \ell_N} \left[\prod_{n=1}^{N} P(q_{\ell_n}^n | X_{n-c}^{n+d}) \frac{P(q_{\ell_n}^n | M)}{P(q_{\ell_n}^n)} \right] P(M) \quad (6)$$

which gives a clear justification for dividing the local posterior estimate by the training data priors to arrive at the scaled likelihoods that are used in the decoding. This demonstrates that given the previously stated assumptions the initial hybrid HMM/ANN systems do produce an estimate of the global posterior $P(M|X)$. This is not entirely straightforward, since although we use scaled likelihoods of the form $P(q_{\ell_n}^n | X_{n-c}^{n+d}) / P(q_{\ell_n}^n)$ as in (6), the first-order Markov model prior $P(q_{\ell_n}^n | q_{\ell_{n-1}}^{n-1}, M)$ in (5) is used in favour of the zeroth-order Markov model prior $P(q_{\ell_n}^n | M)$ in (6). Equations (5) and (6) also provide us with a clear way of properly including language model information $[P(M)]$ into the formalism (as part of other local prior information).

2.3. Discussion

The development above is based on the local posterior probabilities $P(q_{\ell_n}^n | X_{n-c}^{n+d})$. If the local likelihoods are used (as in usual in HMMs) the following expression can be written (with exactly the same assumptions) :

$$P(M|X) \simeq P(M) \sum_{\ell_1, \ldots, \ell_N} \left[\prod_{n=1}^{N} P(X_{n-c}^{n+d} | q_{\ell_n}^n) \frac{P(q_{\ell_n}^n | q_{\ell_{n-1}}^{n-1}, M)}{P(X_{n-c}^{n+d})} \right] \quad (7)$$

Using Bayes rule, we can show that expression (6) and (7) are then similar :

$$\frac{P(X_{n-c}^{n+d} | q_{\ell_n}^n)}{P(X_{n-c}^{n+d})} = \frac{P(q_{\ell_n}^n | X_{n-c}^{n+d})}{P(q_{\ell_n}^n)} \quad (8)$$

The difference between the hybrid and the likelihood approaches lies at the local level. The hybrid system estimates local posteriors and is then discriminant at the frame level. The likelihood system estimates local probability density functions. Both systems can give us an estimate of the global posterior. Classically, the denominator $P(X)$ in (7) is dropped from the equations because it is constant at recognition time.

3. FORWARD-BACKWARD ESTIMATION

In the hybrid systems previously developed (e.g. [3] and [4]), we used Viterbi training in which the summation over state sequences in (5) or (6) is replaced by a maximization over state sequences. However, we can now derive a forward-backward algorithm for hybrid HMM/ANN training without using the Viterbi approximation. This is an application of the Generalized EM algorithm, where the missing data is the state sequence (as usual in HMM estimation), the E-step is the estimation of ANN targets using a forward-backward recurrence and the M-step is the MLP training. This is a *generalized* EM algorithm since the M-step is not exact.

3.1. Recurrences

We can write down forward (α) and backward (β) recurrences:

$$\begin{aligned}\alpha_n(\ell) &= \frac{p(X_1^n, q_\ell^n|M)}{p(X_1^n)} \quad (9)\\ &= \left[\sum_k \alpha_{n-1}(k) p(q_\ell|q_k)\right] \frac{p(x_n|q_\ell)}{p(x_n)}\\ \beta_n(\ell) &= \frac{p(X_{n+1}^N|q_\ell^n, X_1^n, M)}{p(X_{n+1}^N)} \quad (10)\\ &= \sum_k \beta_{n+1}(k) P(q_k|q_\ell) \frac{p(x_{n+1}|q_k)}{p(x_{n+1})}\end{aligned}$$

which are similar to the standard recurrences used in HMMs, apart from the scaling factor $p(x_n)$. This scaling factor is necessary:

1. To have the α (or β) recursion estimating $\frac{p(X|M)}{p(X)} = \frac{P(M|X)}{P(M)}$.
2. To have the α and β recurrences expressed in terms of $\frac{p(x_n|q_k)}{p(x_n)} = \frac{P(q_k|x_n)}{P(q_k)}$, which is the only value that can be estimated by the ANN (provided that we can get an estimate of $P(q_k)$ – see below).

Assuming that we can estimate state priors in the full forward-backward framework, the ANN targets may then be re-estimated using the following:

$$\begin{aligned}P(q_k^n|X, M) = \gamma_n(k) &= \frac{p(q_k^n, X|M)}{p(X|M)} \quad (11)\\ &= \frac{\alpha_n(k)\beta_n(k)}{\sum_\ell \alpha_n(\ell)\beta_n(\ell)}\end{aligned}$$

As for REMAP, convergence can be proved. This approach has been previously employed for speech recognition in [6] and for handwriting recognition in [5]. However in the last case an explicit Viterbi segmentation was assumed when estimating the priors $P(q_k)$ required by (6).

3.2. Priors and durations

As a generalization of what has been done with Viterbi-based hybrid HMM/ANN systems, priors $P(q_k)$ can be estimated as:

$$P(q_k) = \frac{\sum_{n=1}^N P(q_k^n|X, M)}{N} = \frac{\sum_{n=1}^N \gamma_n(k)}{N} \quad (12)$$

which allows us to compute forward and backward recurrences and to iterate the training process.

At recognition time, a duration modelling is usually used in order to enhance the performance of the system. Such a duration model needs the estimation of duration histograms which is straightforward in case of Viterbi. In the forward-backward context, we can define the state duration in a particular utterance ω_i as :

$$d_{\omega_i}(q_k) = \sum_{n=1}^N \gamma_n(k) \quad (13)$$

At the contrary of Viterbi, forward-backward durations can take non-integer values.

3.3. Discussion

The Viterbi procedure considers the best state sequence sequence through the HMM, which means that we take a hard decision about which state q_k is visited at time n. In other words, we can express $\gamma_n(k)$ in (11) while working with Viterbi, simply setting $\gamma_n(k) = 1$ if state q_k is visited at time n and setting $\gamma_n(k) = 0$ if the state is not visited. The forward-backward procedure can then be seen as a smoother version of the Viterbi procedure, since we have "soft" decision regarding which state q_k is visited at time n. We usually talk about *hard segmentation* when working with Viterbi and *soft segmentation* when working with forward-backward.

Taking smooth decision at the frame level makes more sense, especially at the boundaries between stationary parts of signal, and when the speech signal is degraded. For this reason, we expect advantages of using a forward-backward criterion when training in difficult conditions : few training data, noisy data, strong coarticulation effects, bad or flat initialization of the parameter set ...

4. EXPERIMENTS

We demonstrate the performance of these algorithms on the task-independent Phonebook database [7]. Phonebook is a phonetically rich isolated word telephone-speech database. It consists of more than 92,000 utterances and almost 8,000 different words, with an average of 11 talkers for each word. Each speaker of a demographically-representative set of over 1,300 native speakers of American English made a single telephone call and read 75 words. The database contains 106 lists of 75 words. Each list is referred by a 2 letter label (for example `aa`, `ab`, ...). The speakers and words are different for each list.

We defined two training sets for our experiments:

1. a small training set of 9,000 utterances and a cross-validation set (used to adapt the MLP training) of 2,000 utterances
2. a training set of 19,000 utterances (21 lists: `*a *h *m *q *t`) and a cross-validation set of 7,000 utterances (8 lists: `*o *y`)

Recognition experiments were performed on a medium size lexicon (600 words) test set of 6,500 utterrances (8 lists: `*d *r`). We used for acoustic features 12 log-rasta PLP (+ delta-features + delta-energy) [9].

The hybrid HMM/ANN system was based on 56 context independent phone HMMs. The CMU dictionary has been used for the phonetic transcription. We used a multilayer perceptron with 234 inputs (9 frames of input context) and 56 outputs (see [8] for more details). For training set 1, a MLP of 600 hidden units has been used. For training set 2, a MLP of 1000 hidden units has been used. A minimum duration model equal to half of the average duration for each phone has been derived and used in both Viterbi and forward-backward cases.

In the table below, these preliminary results show a clear advantage of forward-backward (F-B) training over Viterbi training for the small training set. No significant difference is observed in the case of the larger training set. This result confirms our expectation regarding the behaviour of the forward-backward procedure when used with a small training set.

Word error rate	Training set 1	Training set 2
Viterbi	13.7%	9.8%
F-B	12.2%	10.1%

Table 1: Error rates on isolated word recognition (600 lexicon words) with hybrid HMM/ANN system and log-rasta plp features. Comparison between Viterbi and full foward-backward training.

5. CONCLUSIONS

The theoretical perspective developed in this paper provides us with a better, unified, view of hybrid stochastic HMM/ANN systems and their relationships to standard HMMs and stochastic finite state acceptor. It is shown that such systems, in which the ANN part estimates local posteriors, can be used to modelize global model posteriors.

This better formalisation inspired us to derive a new forward-backward training dedicated to hybrid systems. The training includes ANN target, priors and duration estimation. Finally, comparisons of performance between Viterbi and forward-backward hybrid systems are presented and discussed.

6. ACKNOWLEDGMENTS

We thank the European Community for their support in this work (SPRACH Long Term Research Project 20077).

7. REFERENCES

[1] Bourlard, H., Konig, Y. and Morgan, N., "REMAP : Recursive Estimation and Maximization of A Posteriori Probabilities in connectionist speech recognition", *Proc. EUROSPEECH'95*, Madrid, September 1995.

[2] Bourlard, H., Konig, Y. and Morgan, N., "A Training Algorithm for Statistical Sequence Recognition with Applications to Transition-Based Speech Recognition," *IEEE Signal Processing Letters*, vol. 3, no. 7, pp. 203-205, 1996.

[3] Renals, S., Morgan, N., Bourlard, H., Cohen, M. and Franco, H., "Connectionist Probability Estimators in HMM Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 161-174, 1994.

[4] Robinson, T., Hochberg, M. and Renals, S., "The use of recurrent networks in continuous speech recognition," in *Automatic Speech and Speaker Recognition – Advanced Topics*, pp. 233-258, Kluwer Academic Publishers, 1996.

[5] Senior, A. and Robinson, T., "Forward-Backward Retraining of Recurrent Neural Networks," in *Advances in Neural Information Processing Systems 8*, MIT Press, pp. 743-749, 1996.

[6] Yan,Y., Fanty M., and Cole R., "Speech Recognition Using Neural Networks with Forward Backward Probability Generated Targets," *Proc. ICASSP'97*, IV-3241-3244, Munich.

[7] Pirelli, J. et al., "Phonebook: A Phonetically-rich isolated word telephone speech database," *Proc. ICASSP'95*, Detroit.

[8] Dupont, S., Bourlard, H., Deroo, O., Fontaine, V. and Boite, J.-M., "Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on 'Phonebook' and Related Improvements," to be published in *Proc. of ICASSP'97.*

[9] Hermansky, H., Morgan, N., "Rasta Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, special issue on Robust Speech Recognition, vol.2, no.4, pp. 578-589, Oct.94