REMAP MODELING FOR CONNECTIONIST SPEECH RECOGNITION

Yochai Konig, Hervé Bourlard[a], and Nelson Morgan
International Computer Science Institute

Berkeley, CA 94704, USA.

[a] Also affiliated with Faculté Polytechnique de Mons, Mons, Belgium

Summary

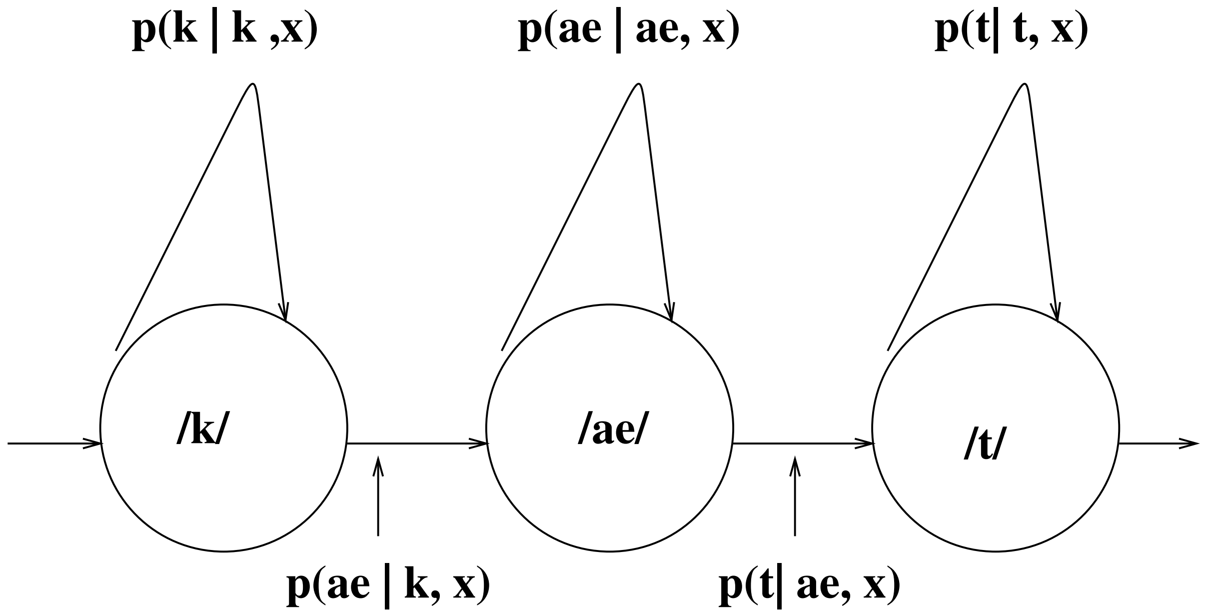
- We can train hybrid HMM/ANN system in a globally discriminant way by estimating ANN parameters that maximize the global posterior probabilities, i.e. minimize the utterance error rate.
- In training we use posterior probabilities as targets ("soft targets") versus labels ("hard targets") in our standard HMM/ANN system.
- In recognition we use only posterior probabilities versus scaled likelihoods in our standard system.
- Preliminary experiments show an improvement in recognition results.

Algorithm

- **Goal-** To increase $P(M|X)$ of the correct model. X - sequence of acoustic vectors, M- sentence model.
- **Question-** How to incorporate this global goal in the local training of the ANN?
- **Idea-** REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities. ANN targets are re-estimated iteratively to guarantee a continuous increase of the global posterior. The global posteriors of all possible models sum up to one, so we get discriminant training.

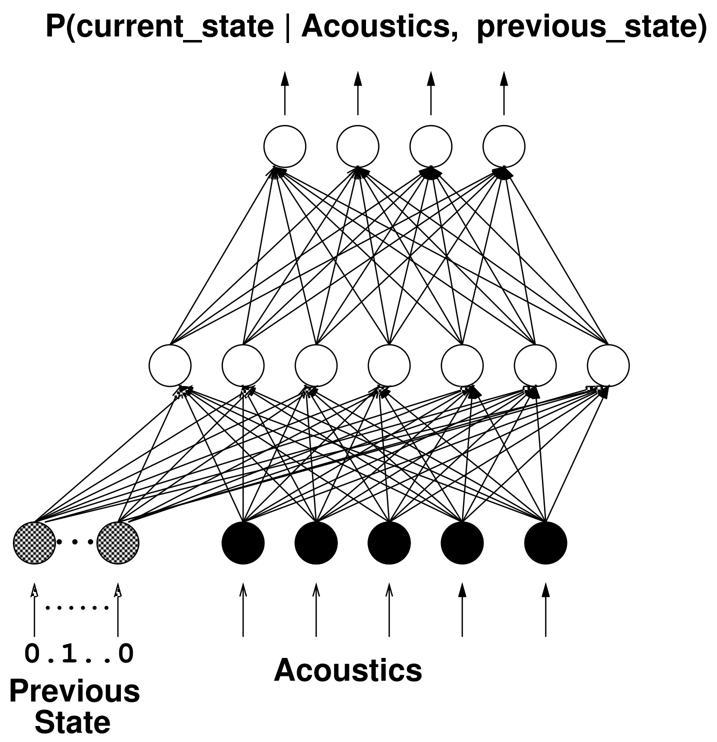
Discriminant HMM - An example of "cat"

- It can be shown that $P(M|X)$ can be expressed in terms of $p(q_n^\ell | q_{n-1}^k, X_{n-c}^{n+d})$, where X_{n-c}^{n+d} is a window of acoustic vectors, and q_{n-1}^k represents being at state k at time $n-1$.

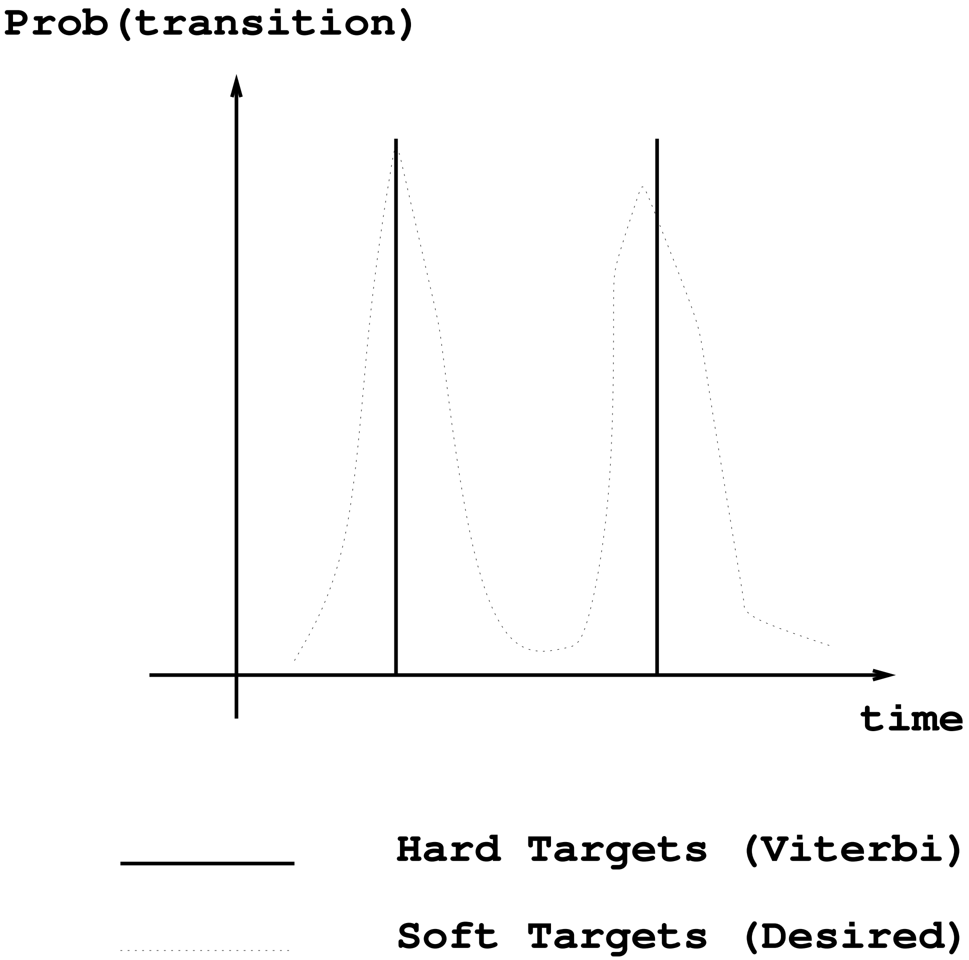


Local Transition Probabilities

- An MLP that estimates these local conditional transition probabilities.

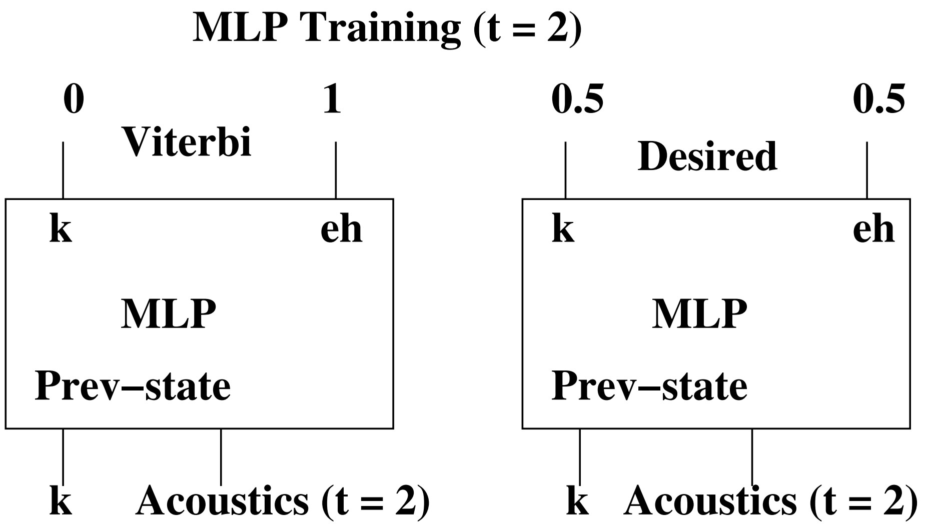


Motivation - Soft Targets



Soft Targets - Details

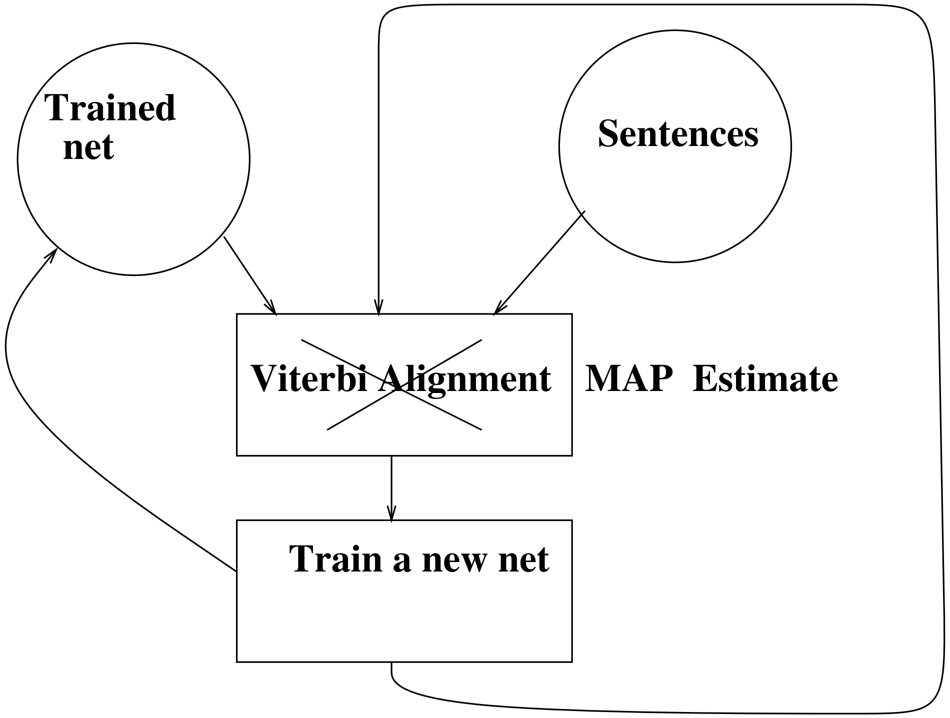
Time	1	2	3
Viterbi	k−>k	k−>eh	eh−>eh
Desired	k−>k 0.7 k−>eh 0.3	k−>k 0.5 k−>eh 0.5	k−>k 0.2 k−>eh 0.8



REMAP Algorithm - Idea

- **E-step** Estimate new transition targets given the current MLP.
- **M-step** Train new MLP to maximize performance according to new targets.
- Iterate until the increase of the a posteriori probability of the correct model is too small.

Before and After REMAP



Targets (Viterbi) : /k/ –> /k/ /k/ –> /ae/

Targets (MAP) : /k/ –> /k/ 0.8 /k/ –> /k/ 0.3
/k/ –> /ae/ 0.2 /k/ – > /ae/ 0.7
/ae/ –> /ae/ 0.8
/ae/ –> /t/ 0.2

REMAP Algorithm - Details

- Start from some initial net providing $P(q_\ell^n | X_{n-c}^{n+d}, q_k^{n-1})$, $\forall$ possible (k, ℓ)-pairs.
- **E-step** Run recurrences to compute MLP targets $P(q_\ell^n | X, q_k^{n-1})$, $\forall$ possible (k, ℓ)-pairs.
- **M-step** For every x_n in the training database, train MLP with output targets equal to $P(q_\ell^n | X, q_k^{n-1})$, $\forall$ possible q_k at the input or for a limited subset as imposed by the HMM topology.
- Iterate from E-step until convergence, or according to cross-validation results.

Proof - Outline

- Defining an auxiliary function such that maximizing that function is equivalent to maximizing the global posterior probability of the correct model.
- Finding new targets for training the MLP that maximize the auxiliary function.
- Showing that training the MLP with these new targets leads to an increase in the value of the auxiliary function.

Experimental Methods

- **Task-** Digits+ database: "one" through "nine", "zero", "oh", "no", and "yes". Isolated words over a clean phone line. Added Noise: 10DB S/N. 200 Speakers, 1720 training utterances, 230 cross-validation, 650 testing.
- **Nets-** 214 inputs, 153 inputs- acoustic features, 61 - previous state. 200 hidden, 61 outputs.
- **Acoustic Features-** RASTA-PLP8 + delta features + delta log gain. Analysis window - 25 ms estimated every 12.5 ms. 8 Khz sampling, telephone bandwidth.

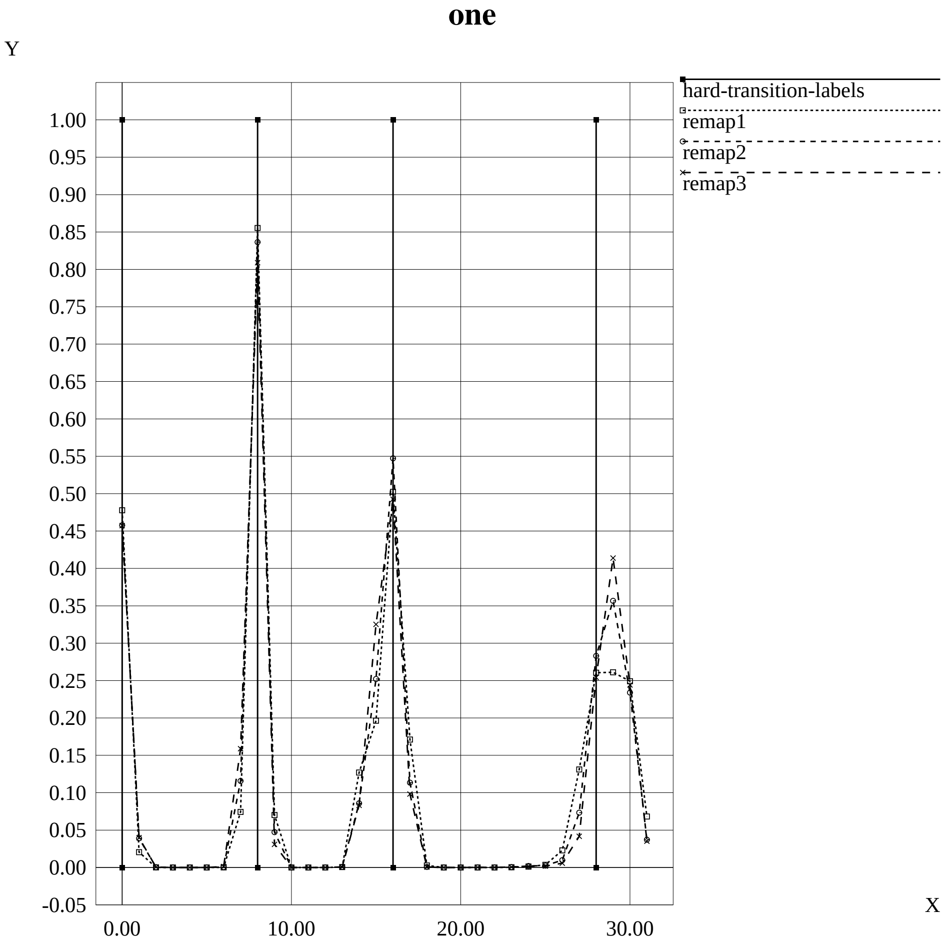
Experiments - Results

System	Error Rate	Average Posterior
Classical Hybrid	3.1%	-
Discriminant HMM, pre-REMAP	2.9%	0.110
1 REMAP iteration	2.3%	0.161
2 REMAP iterations	2.3%	0.174
3 REMAP iterations	2.2%	0.180

Table 1: Results in word error (wrong words)

The Effect of REMAP

- Y-axis shows the probability of a transition (changing state) for every frame in the utterance “one”.



Conclusions

- The EM-like REMAP algorithm is a general solution to the problem of parameter estimation with incomplete data according to the Maximum A Posteriori criterion in hybrid HMM/MLP systems.
- We have applied REMAP to transition-based connectionist speech recognition system, specifically to the Discriminant HMM.
- We have shown recognition improvement on a small but non-trivial task. We plan to test our theory on more difficult tasks.