REMAP MODELING FOR
CONNECTIONIST SPEECH
RECOGNITION

Yochai Konig, Hervé Bourlard\textsuperscript{a}, and Nelson Morgan
International Computer Science Institute
Berkeley, CA 94704, USA.

\textsuperscript{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 $p(q^k_n | q^{k-1}_n, X^{n+d}_{n-c})$, where $X^{n+d}_{n-c}$ is a window of acoustic vectors, and $q^{k-1}_n$ represents being at state $k$ at time $n - 1$. 

---

![Diagram showing transitions between states](attachment:diagram.png)
Local Transition Probabilities

- An MLP that estimates these local conditional transition probabilities.
Motivation - Soft Targets

Prob(transition)

--- Hard Targets (Viterbi)

............... Soft Targets (Desired)
Soft Targets - Details

<table>
<thead>
<tr>
<th>Time</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viterbi</td>
<td>k→k</td>
<td>k→eh</td>
<td>eh→eh</td>
</tr>
<tr>
<td>Desired</td>
<td>k→k 0.7</td>
<td>k→k 0.5</td>
<td>k→k 0.2</td>
</tr>
<tr>
<td></td>
<td>k→eh 0.3</td>
<td>k→eh 0.5</td>
<td>k→eh 0.8</td>
</tr>
</tbody>
</table>

MLP Training (t = 2)

0
Viterbi

k

MLP

Prev-state

0.5
Desired

k

MLP

Prev-state

0.5

k

Acoustics (t = 2)

k

Acoustics (t = 2)
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/ 0.8
/k/ → /ae/ 0.2
/k/ → /ae/ 0.8

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

Trained net → Sentences

Viterbi Alignment → MAP Estimate

Train a new net
**REMAP Algorithm - Details**

- Start from some initial net providing $P(q^k_{\ell} | X_{n-c}^{n+d}, q^{n-1}_k)$, ∀ possible $(k, \ell)$-pairs.

- **E-step** Run recurrences to compute MLP targets $P(q^n_{\ell} | X, q^{n-1}_k)$, ∀ possible $(k, \ell)$-pairs.

- **M-step** For every $x_n$ in the training database, train MLP with output targets equal to $P(q^n_{\ell} | X, q^{n-1}_k)$, ∀ 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

<table>
<thead>
<tr>
<th>System</th>
<th>Error Rate</th>
<th>Average Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Hybrid</td>
<td>3.1%</td>
<td>-</td>
</tr>
<tr>
<td>Discriminant HMM, pre-REMAP</td>
<td>2.9%</td>
<td>0.110</td>
</tr>
<tr>
<td>1 REMAP iteration</td>
<td>2.3%</td>
<td>0.161</td>
</tr>
<tr>
<td>2 REMAP iterations</td>
<td>2.3%</td>
<td>0.174</td>
</tr>
<tr>
<td>3 REMAP iterations</td>
<td>2.2%</td>
<td>0.180</td>
</tr>
</tbody>
</table>

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”.

![Graph showing the effect of REMAP](image-url)
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.