**Segment-level training based on Confidence Measures for Hybrid HMM/ANN Speech Recognition**

S. Pavankumar Dubagunta$^{1,2}$ and Mathew Magimai Doss$^1$

$^1$Idiap Research Institute, Martigny, Switzerland  
$^2$École polytechnique fédérale de Lausanne (EPFL), Switzerland

---

**Background: Conventional Hybrid HMM Systems**

In hybrid hidden Markov model (HMM) based speech recognition, the scaled likelihood of an acoustic observation $x_t$ given a HMM state $q_i$ at time $t$, labelled $l_i$, is estimated as:

$$p(x_t | q_i = l_i) = \sum_{d=1}^{D} \frac{p(x_t | d)^{l_i}}{p(d)} \cdot \sum_{c=1}^{C} p(d | c)^{l_i} \cdot a_{q_i | c}$$

**Confidence Estimation using Local Posteriors**

Conventionally, given the segmentation, the artificial neural network (ANN) is trained using one hot encodings of the targets and minimizing frame level cross-entropy. This can be expressed as:

$$E_f(t) = \text{KL}(y_{t} \parallel \hat{y}_{t}) = \log p(x^d | y_{t}) - \log p(x^d | \hat{y}_{t})$$

where $D$ are estimated from the state segment counts instead of frame label counts. This can be minimised:

$$E_f(s) = -\text{CM}(s) = \sum_{t \in s} \frac{e_{x,t}^{(s)}}{v_{x,t}^{(s)}} \cdot \text{KL}(y_{t} \parallel \hat{y}_{t})$$

**Results**

<table>
<thead>
<tr>
<th>System</th>
<th>AMI (Word level)</th>
<th>Mediaparl (Word level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLLR+spk/vec ANN</td>
<td>35.6</td>
<td>21.3</td>
</tr>
<tr>
<td>MLLR+spk/vec MFC</td>
<td>23.9</td>
<td>19.8</td>
</tr>
<tr>
<td>MLLR+spk/vec</td>
<td>23.9</td>
<td>19.8</td>
</tr>
</tbody>
</table>

**Conclusion**

The proposed linguistic-segment-level training of ANNs based on confidence measures improves robustness to duration variations in the training data set.