IDIAP RESEARCH REPORT

Joint Training of Multi-Stream HMMs

Samy Bengio [a]

IDIAP–RR 05-22

May 2005

[a] IDIAP Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-22

Joint Training of Multi-Stream HMMs

Samy Bengio

May 2005

Abstract. This report describes a novel technique to jointly train efficiently several streams of data describing the same sequence of events using a unified EM algorithm.

1 Introduction

Multi-stream HMMs is a very popular and efficient technique to decompose a complex problem such as automatic speech recognition into several streams of information, where each stream of data is trained independently of the others, while all streams are jointly used during decoding.

In this report, we propose a method to trained all streams simultaneously in order to maximize the joint probability of all data streams. This can be done by adding an additional, but reasonable assumption, which states that given the value of the state variable at time t, all streams up to time t are independent from each other (hence, the state variable contains all the correlation between all streams up to time t).

2 Notation

Let

- N be the number of streams.
- $x^n_{a:b}$ be the observations of stream n between time a and b

3 Usual Derivation for Each Stream

Let us start by computing the normal forward and backward equations for each stream. First the forward equation:

$$\begin{aligned} \alpha_n(i,t) &= p(x^n_{1:t}, q_t = i) && (1)\\ &= p(x^n_t | q_t = i) \sum_j p(q_t = i | q_{t-1} = j) \alpha_n(j, t-1) && (2) \end{aligned}$$

which can be used to compute the likelihood of one stream:

$$\begin{aligned} L_n &= p(x^n_{1:T}) && (3)\\ &= \sum_i p(x^n_{1:T}, q_T = i) = \sum_i \alpha_n(i,T) && (4) \end{aligned}$$

and then the backward equation:

$$\begin{aligned} \beta_n(i,t) &= p(x^n_{t+1:T} | q_t = i) && (5)\\ &= \sum_j p(x^n_{t+1} | q_{t+1} = j) p(q_{t+1} = j | q_t = i) \beta_n(j, t+1) && (6) \end{aligned}$$

which, together with the forward equation, can be used to compute the posterior probability of a state given the stream observation:

$$\begin{aligned} \gamma_n(i,t) &= p(q_t = i | x^n_{1:T}) && (7)\\ &= \frac{p(x^n_{1:t}, x^n_{t+1:T}, q_t = i)}{p(x^n_{1:T})} \\ &= \frac{\alpha_n(i,t)\beta_n(i,t)}{L_n} && (8) \end{aligned}$$

4 Derivation for Joint Streams

In this section, we derive the new EM training algorithm for multi-streams.

On top of the usual HMM assumptions accepted for the single stream case (emissions at time t only depend on the state, and the current state only depends on the previous state), we need to add one more assumption, which basically says that, given the value of the current state, the random variables of all streams up to time t are independent:

$$p(x^1_{1:t}, x^2_{1:t}, \cdots, x^N_{1:t}|q_t = i) = \prod_{n=1}^{N} p(x^n_{1:t}|q_t = i) \tag{9}$$

4.1 Joint Likelihood and Auxiliary Function

As usual when deriving an EM algorithm, we start by defining the following indicator variable:

$$z_{i,t} = \begin{cases} 1 & \text{if } q_t = i \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The joint complete log likelihood of the streams and the state l_c is:

$$\begin{aligned} l_c &= \log p(X^1, X^2, \cdots, X^N, Q) && (11)\\ &= \log p(X^1, X^2, \cdots, X^N|Q) + \log P(Q) \\ &= \log p(X^1|Q) + \log p(X^2|Q) + \cdots + \log p(X^N|Q) + \log P(Q) \\ &= \sum_n \log p(X^n|Q) + \log P(Q) \\ &= \sum_t \sum_n \sum_i z_{i,t} \log p(x^n_t|q_t = i) + \\ &\quad \sum_t \sum_i \sum_j z_{i,t} z_{j,t-1} \log p(q_t = i|q_{t-1} = j) && (12) \end{aligned}$$

and the corresponding auxiliary variable becomes:

$$\begin{aligned} A &= E_Q\left[\log p(x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T}, Q)|X^{1:N}\right] && (13)\\ &= \sum_t \sum_n \sum_i E_Q\left[z_{i,t}|X^{1:N}\right] \log p(x^n_t|q_t = i) + \\ &\quad \sum_t \sum_i \sum_j E_Q\left[z_{i,t} z_{j,t-1}|X^{1:N}\right] \log p(q_t = i|q_{t-1} = j) && (14) \end{aligned}$$

The forward recursion becomes:

$$\begin{aligned} \alpha(i,t) &= p(x^1_{1:t}, x^2_{1:t}, \cdots, x^N_{1:t}, q_t = i) && (15)\\ &= p(x^1_{1:t}, x^2_{1:t}, \cdots, x^N_{1:t}|q_t = i) p(q_t = i) \\ &= p(x^1_{1:t}|q_t = i) p(x^2_{1:t}|q_t = i) \cdots p(x^N_{1:t}|q_t = i) p(q_t = i) \\ &= \frac{p(x^1_{1:t}|q_t = i)p(q_t = i)}{p(q_t = i)} \frac{p(x^2_{1:t}|q_t = i)p(q_t = i)}{p(q_t = i)} \cdots \frac{p(x^N_{1:t}|q_t = i)p(q_t = i)}{p(q_t = i)} p(q_t = i) \\ &= \frac{p(x^1_{1:t}, q_t = i)}{p(q_t = i)} \frac{p(x^2_{1:t}, q_t = i)}{p(q_t = i)} \cdots \frac{p(x^N_{1:t}, q_t = i)}{p(q_t = i)} p(q_t = i) \\ &= \frac{\alpha_1(i,t)}{p(q_t = i)} \frac{\alpha_2(i,t)}{p(q_t = i)} \cdots \frac{\alpha_N(i,t)}{p(q_t = i)} p(q_t = i) \\ &= \frac{\prod_{n=1}^{N} \alpha_n(i,t)}{p(q_t = i)^{N-1}} && (16) \end{aligned}$$

We can recover the joint likelihood of the streams as:

$$
\begin{aligned}
L &= p(x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T}) && (17)\\
&= \sum_i p(x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T}, q_T = i) \\
&= \sum_i \alpha(i, T) && (18)
\end{aligned}
$$

Let us also compute the joint posterior probability of being in a state given all data of all streams, the so-called gamma:

$$
\begin{aligned}
E_Q\left[z_{i,t}|X^{1:N}\right] &= \gamma(i,t) && (19)\\
\gamma(i,t) &= p(q_t = i|x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T}) && (20)\\
&= \frac{p(x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T}, q_t = i)}{p(x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T})} \\
&= \frac{p(x^1_{t+1:T}|q_t = i) \cdots p(x^N_{t+1:T}|q_t = i) p(x^1_{1:t}, \cdots, x^N_{1:t}, q_t = i)}{L} \\
&= \frac{\beta_1(i,t) \cdots \beta_N(i,t) \alpha(i,t)}{L} \\
&= \frac{\alpha(i,t) \prod_n \beta_n(i,t)}{L} && (21)
\end{aligned}
$$

Note that we need to compute $p(q_t = i)$, which can be done recursively as follows:

$$
\begin{aligned}
p(q_t = i) &= \sum_j p(q_t = i, q_{t-1} = j) && (22)\\
&= \sum_j p(q_t = i|q_{t-1} = j) p(q_{t-1} = j) && (23)
\end{aligned}
$$

We can also similarly compute the posterior of a transition as follows:

$$
\begin{aligned}
E_Q\left[z_{i,t} z_{j,t-1}|X^{1:N}\right] &= p(q_t = i, q_{t-1} = j|x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T}) && (24)\\
&= \frac{p(x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T}, q_t = i, q_{t-1} = j)}{p(x^1_{1:T}, x^2_{1:T}, \cdots, x^N_{1:T})} \\
&= \frac{\alpha(j, t-1) p(q_t = i|q_{t-1} = j) \prod_n p(x^n_t|q_t = i) \beta_n(i,t)}{L}
\end{aligned}
$$

5 Conclusion

While most multi-stream HMM techniques assume that each stream is first trained independently, we have shown in this report a principled way to jointly trained all streams in order to optimize the joint probability of all streams.