HLDA-MMI Update

Contents

0.1 Objective function

The objective function optimised in HLDA is the likelihood of the data given the correct-model HMM and transformed by the accepted rows $\mathbf{A}_{[p]}$ of the transform $\mathbf{A}$, multiplied by the likelihood given a single Gaussian for the rejected dimensions, and multiplied by the determinant $|A|$ of the transform. The objective function is:

$$\log p(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathbf{A}_{[p]}) + \log p(\mathcal{O}|\mathcal{M}_{\text{rej}}, \mathbf{A}_{[n-p]}) + T \log |A| \tag{1}$$

where $\mathcal{M}_{\text{num}}$ is the model of the correct transcription, $\mathcal{M}_{\text{rej}}$ is a single Gaussian for the rejected dimensions and T is the length of the training data. The auxiliary function used to optimise normal HLDA is

$$\mathcal{F}_{HLDA}(\mathbf{A};\mathcal{M}) = \frac{1}{2}\sum_m \gamma_m \log \frac{|A|^2}{|\text{diag}(\mathbf{A}_{[p]}\mathbf{W}_{(m)}\mathbf{A}_{[p]}^T)||\text{diag}(\mathbf{A}_{[n-p]}\Sigma\mathbf{A}_{[n-p]})|}, \tag{2}$$

where $\mathbf{A}_{[p]}$ is the accepted p rows of transform matrix $\mathbf{A}$ and $\mathbf{A}_{[n-p]}$ is the rejected $n-p$ rows, $\mathbf{W}_{(m)}$ are the within-class covariances for Gaussian mixtures m and γ_m are the occupation counts; Σ is the global covariance of the data.

In MMI-based HLDA we are trying to the likelihood of the data given the correct transcription's HMM, minus a constant $\alpha \leq 1$ times the likelihood for the HMM aligned to a lattice $\mathcal{M}_{\text{den}}$ derived from a recognition model. The overall objective function is:

$$\begin{array}{c}\log p(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathbf{A}_{[p]}) - \alpha \log p(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathbf{A}_{[p]}) \\ +(1-\alpha)\log p(\mathcal{O}|\mathcal{M}_{\text{rej}}, \mathbf{A}_{[n-p]}) + T(1-\alpha)\log |A|\end{array} \tag{3}$$

where $p(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathbf{A}_{[p]})$ is the likelihood of the accepted dimensions of the transformed data given the numerator model (i.e. the HMM of the correct transcription), $p(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathbf{A}_{[p]})$ the likelihood given the recognition model (all sentences),

$\log p(\mathcal{O}|\mathcal{M}_{\text{rej}}, \mathbf{A}_{[n-p]})$ the likelihood of the rejected dimensions of the data given $\mathcal{M}_{\text{rej}}$ which is a single diagonal Gaussian, and T is the length of the training data.

This objective function is not as easy to optimise as the HLDA objective function. Suppose we define $\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]})$ as a notation for the standard auxiliary function for $\log p(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathbf{A}_{[p]})$ given a HMM updated using the ML statistics derived from model $\mathcal{M}_{\text{mle}}$ and (previous) transform $\hat{\mathbf{a}}$, and an alignment to numerator model $\mathcal{M}_{\text{num}}$ also derived from previous transform $\hat{\mathbf{a}}$. ($\mathcal{M}_{\text{num}}$ and $\mathcal{M}_{\text{mle}}$ are the same for MMI, but are kept separate because they differ in MPE training).

$\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]})$ has the correct properties to maximise $\log p(\emptyset|\mathcal{M}_{\text{num}}, \mathbf{A}_{[p]})$ since the two are equal at the old value of $\mathbf{A}$ which is $\hat{\mathbf{A}}$ but $\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]}) \leq \log p(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]})$ everywhere else (refer to the Baum-Welch algorithm). This ensures that increasing the auxiliary function will increase the objective function. Unfortunately the same property does not hold for the denominator case, where $-\alpha\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]}) \geq -\alpha \log p(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathbf{A}_{[p]})$ due to the negation. The problem is that if the transform changes significantly from the one used to gather statistics from the data ($\hat{\mathbf{A}}$), $\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]})$ starts to become an underestimate of the real likelihood $\log p(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathbf{A}_{[p]})$.

However, the auxiliary function given by:

$$\begin{aligned}\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]}) - \alpha\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]}) \\ +(1-\alpha)\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{rej}}, \mathbf{A}_{[n-p]}) + T(1-\alpha)\log|A|\end{aligned} \tag{4}$$

is useful because around the old value of the transform $\hat{\mathbf{A}}$ it has the same differential w.r.t. elements of the transform $\mathbf{A}$ as the real objective function does. Therefore if it converges it will converge to a locally optimum value of $\mathbf{A}$ (although it will not converge). In order to obtain an auxiliary which will converge, we can add to it some function $\mathcal{I}(\mathbf{A}, \bar{\mathbf{A}})$ which has its maximum value at $\mathbf{A} = \bar{\mathbf{A}}$. The auxiliary function will still converge to the right point, if it converges, since the only bias in $\mathcal{I}$ is towards the previous value $\bar{\mathbf{A}}$ which is allowed to change; after a number of iterations of EM the new auxiliary function should converge to the correct point.

A suitable choice of $\mathcal{I}(\mathbf{A}, \bar{\mathbf{A}})$ will lead to good convergence; in particular, a good choice will be one which penalises changes in alignment by forcing the new transform to be similar to some old transform $\bar{\mathbf{A}}$. The old transform $\bar{\mathbf{A}}$ might be the transform $\hat{\mathbf{A}}$ used to gather statistics from the data, but it is distinguished from $\hat{\mathbf{A}}$ since it might be advantageous to use a transform derived from HLDA for $\bar{\mathbf{A}}$ on the first iteration, even if the data was aligned using non-HLDA features. This might provide a better starting point. In the case of multiple iterations of MMI-HLDA, $\bar{\mathbf{A}}$ should be the the transform from the previous iteration.

A suitable form for $\mathcal{I}(\mathbf{A}, \bar{\mathbf{A}})$ is the following:

$$E \quad (\mathcal{H}(\mathcal{O}|\mathcal{M}_{\text{mle}}, \mathcal{M}_{\text{mle}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]}, \bar{\mathbf{A}}_{[p]}) + \mathcal{H}(\mathcal{O}|\mathcal{M}_{\text{rej}}, \mathbf{A}_{[n-p]}, \bar{\mathbf{A}}_{[n-p]}) + T \log |A|), \tag{5}$$

where E is a smoothing constant use to ensure good convergence (e.g. E might be 0.1, 0.5 or 1, or perhaps zero if $\alpha < 1$). $\mathcal{H}(\mathcal{O}|\mathcal{M}_{\text{mle}}, \mathbf{A}, \hat{\mathbf{A}}_{[p]}, \bar{\mathbf{A}}_{[p]})$, which is a function of the variable $\mathbf{A}_{[p]}$ with the other arguments being considered as constants, is defined as the likelihood of the accepted dimensions of the data, given the alignment of Gaussians to data data given by $\mathcal{M}_{\text{mle}}$ and $\hat{\mathbf{A}}$, and the data and means transformed using the old transform $\bar{\mathbf{A}}$, but the variance of the Gaussians obtained from the data transformed using the new transform $\mathbf{A}$. $\mathcal{I}(\mathbf{A}, \hat{\mathbf{A}})$ has its highest value where $\mathbf{A} = \hat{\mathbf{A}}$; any difference leads to a mismatch between the variances and data and reduces the value of the smoothing function[1]

The two transforms $\hat{\mathbf{A}}$ and $\bar{\mathbf{A}}$ are distinguished because we might use a diagonal transform as $\hat{\mathbf{A}}$ for alignment of the data but might want to use a transform derived from HLDA for $\bar{\mathbf{A}}$ because it is a better starting point for MMI-HLDA optimisation, and the alignment of the data is unlikely to change much between the two.

The auxiliary function is now:

$$\begin{aligned}
\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathcal{M}_{\text{mle}}, \mathbf{A}, \hat{\mathbf{A}}) \ &- \ \alpha \mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{den}}, \mathcal{M}_{\text{mle}}, \mathbf{A}, \hat{\mathbf{A}}) \\
&+ (1-\alpha)\mathcal{G}(\mathcal{O}_{[n-p]}|\mathcal{M}_{\text{rej}}, \mathbf{A}) \\
+ \ E\mathcal{H}(\mathcal{O}|\mathcal{M}_{\text{mle}}, \mathbf{A}, \hat{\mathbf{A}}, \bar{\mathbf{A}}) \ &+ \ E\mathcal{H}(\mathcal{O}|\mathcal{M}_{\text{rej}}, \mathbf{A}, \bar{\mathbf{A}}) \\
&+ (1 - \alpha + E)T \log |A|
\end{aligned} \tag{6}$$

0.1.1 Auxiliary functions for Gaussian likelihoods

The statistics gathered from the data are the occupation counts γ_m, the Gaussian covariances $\mathbf{W}_{(m)}$ and the Gaussian means μ_m. These will be distinguished by three superscripts: num (numerator), den (denominator) and mle (maximum likelihood estimation statistics). The num and den statistics are gathered from the correct model $\mathcal{M}_{\text{num}}$ and a lattice of recognised data $\mathcal{M}_{\text{den}}$ respectively; the mle statistics are the same as the numerator in this case but are kept separate here to enable extension to MPE training.

As mentioned above, the function $\mathcal{G}(\mathcal{O}|\mathcal{M}_{\text{num}}, \mathbf{A}_{[p]}, \hat{\mathbf{A}}_{[p]})$ is the normal auxiliary function used for Baum-Welch estimation, except a distinction is made between the model set $\mathcal{M}_{\text{num}}$ or $\mathcal{M}_{\text{den}}$ used to evaluate the likelihood and the model $\mathcal{M}_{\text{mle}}$

[1] Actually this may not be quite true where variance flooring is used, but even with variance flooring the smoothing function will not want $\mathbf{A}$ to differ substantially from the old transform $\bar{\mathbf{A}}$.

used to estimate the Gaussian means and variances. Writing it out as a function of the statistics from the training data,

$$\begin{aligned}&\mathcal{G}(\mathcal{O}_{[p]}|\mathcal{M}_{\text{num}}, \mathcal{M}_{\text{mle}}, \mathbf{A}, \hat{\mathbf{A}}) = \\ &-\tfrac{1}{2}\textstyle\sum_m \gamma_m^{\text{num}} \sum_{k=1}^{K} \left[\log|\text{diag}(\Sigma^{(m)})| + \right. \\ &\left.\text{diag}(\Sigma^{(m)})^{-1} \cdot (\mathbf{A}_{[p]}\mathbf{X}^{\text{num}}_{(m)}\mathbf{A}^T_{[p]})\right],\end{aligned} \tag{7}$$

where the within-class variances $\Sigma^{(m)}$ are equal to $\mathbf{A}_{[p]}\mathbf{W}^{\text{mle}}_{(m)}\mathbf{A}^T_{[p]}$, and the notation $\mathbf{A}\cdot\mathbf{B}$ for matrices $\mathbf{A}$ and $\mathbf{B}$ of the same size is a matrix dot product, defined as the result of multiplying corresponding elements and adding them all up as in a vector dot product. $\mathbf{X}^{\text{num}}_{(m)}$ is defined as the matrix $\mathbf{W}^{\text{num}}_{(m)} + (\mu_m^{\text{num}} - \mu_m^{\text{mle}})(\mu_m^{\text{num}} - \mu_m^{\text{mle}})^T$, which is the variance of the numerator data around the ML mean.

This function represents the likelihood of the accepted dimensions of the Gaussians, with both Gaussian parameters and data transformed using matrix $\mathbf{A}_{[p]}$. Replacing num with den gives the corresponding function for the denominator. For the rejected dimensions,

$$\mathcal{G}(\mathcal{O}_{[p]}|\mathcal{M}_{\text{rej}}, \mathbf{A}) = -\frac{1}{2}T\left[\log|\text{diag}(\Sigma)| + \text{diag}(\Sigma)^{-1} \cdot (\mathbf{A}_{[n-p]}\Sigma^{\text{data}}\mathbf{A}^T_{[n-p]})\right],$$

where T is the length of the training data, Σ^{data} the global variance of the data and Σ is the transformed variance $\mathbf{A}_{[n-p]}\Sigma\mathbf{A}^T_{[n-p]}$ of the rejected dimensions of the data.

The smoothing function is as follows:

$$\begin{aligned}&\mathcal{H}(\mathcal{O}_{[p]}|\mathcal{M}_{\text{mle}}, \mathbf{A}, \hat{\mathbf{A}}, \bar{\mathbf{A}}) = \\ &-\tfrac{1}{2}\textstyle\sum_m \gamma_m^{\text{num}} \sum_{k=1}^{K} \left[\log|\text{diag}(\bar{\Sigma}^{(m)})| + \text{diag}(\bar{\Sigma}^{(m)})^{-1} \cdot (\mathbf{A}_{[p]}\mathbf{W}^{\text{num}}_{(m)}\mathbf{A}^T_{[p]})\right],\end{aligned} \tag{8}$$

where the within-class variances $\bar{\Sigma}^{(m)}$ are equal to $\bar{\mathbf{A}}_{[p]}\mathbf{W}^{\text{mle}}_{(m)}\bar{\mathbf{A}}^T_{[p]}$. This function is the Gaussian likelihood of the ML training data and means transformed using the new transform $\mathbf{A}_{[p]}$ but with the Gaussian variances obtained using the old transform $\bar{\mathbf{A}}_{[p]}$. For the rejected dimensions, it is:

$$\mathcal{H}(\mathcal{O}_{[p]}|\mathcal{M}_{\text{rej}}, \mathbf{A}, \bar{\mathbf{A}}) = -\frac{1}{2}T\left[\log|\text{diag}(\bar{\Sigma})| + \text{diag}(\bar{\Sigma})^{-1} \cdot (\mathbf{A}_{[n-p]}\Sigma^{\text{data}}\mathbf{A}^T_{[n-p]})\right],$$

where the variance $\bar{\Sigma}$ is equal to $\bar{\mathbf{A}}_{[n-p]}\Sigma^{\text{data}}\bar{\mathbf{A}}^T_{[n-p]}$, i.e. obtained using the old transform $\bar{\mathbf{A}}_{[p]}$, but the data is transformed using the new transform$\mathbf{A}$.

0.2 Row-by-row updates.

The technique to be used for optimisation of the transform is to update each row $\mathbf{a}_r$ of $\mathbf{A}$. The technique previously used in [2] was based on leaving the Gaussian variances constant while updating $\mathbf{a}_r$ based on its effect on the data. This does not work for the denominator case (or for the numerator case, if $\mathcal{M}_{\text{num}} \neq \mathcal{M}_{\text{mle}}$, i.e. for MPE) because the variances are not necessarily quite at their optimum point w.r.t the likelihood when set using the current transform. The problem is solved by taking advantage of the fact that diagonal variances are used, and using a linear approximation for the relation between each variance element Σ_{rr} and the likelihood function.

Suppose the current value of the matrix is $\mathbf{A}$ and we are looking for a new value of $\mathbf{a}_r$. Let us define $\mathbf{W}^{\text{num}}_{(m)} + (\mu^{\text{num}}_m - \mu^{\text{mle}}_m)^T(\mu^{\text{num}}_m - \mu^{\text{mle}}_m)$, which appears in Equation 7, as $\mathbf{X}^{\text{num}}_{(m)}$; this is independent of $\mathbf{A}$. The variance given the current value of $\mathbf{A}$ is $\mathbf{A}_{[p]}\mathbf{W}^{\text{mle}}_{(m)}\mathbf{A}^T_{[p]}$. Define this as $\Sigma^{(m)}$, and calculate the inverse of its diagonal $\text{diag}(\Sigma^{(m)})^{-1}$. The dependence of $\mathcal{G}$ on $\mathbf{a}_r$ due to changes in the variance of the data can be expressed as

$$-0.5\gamma^{\text{num}}_m . \frac{1}{{\Sigma^{(m)}}_{rr}} \mathbf{X}^{\text{num}}_{(m)} \cdot \mathbf{a}_r \mathbf{a}^t_r$$

where $\mathbf{A} \cdot \mathbf{B}$ is the matrix dot product as described above. The dependence of $\mathcal{G}$ on $\mathbf{a}_r$ due to changes in the Gaussian likelihood (using a linear approximation) can be expressed as

$$-0.5\gamma^{\text{num}}_m . k_m \mathbf{W}^{\text{mle}}_{(m)} \cdot \mathbf{a}_r \mathbf{a}^t_r$$

, where $k_m = \frac{1}{{\Sigma^{(m)}}_{rr}} - \frac{\mathbf{a}_r \mathbf{X}^{\text{num}}_{(m)} \mathbf{a}^t_r}{{\Sigma^{(m)}}^2_{rr}}$; note that k_m can be positive or negative but would be zero if $\mathcal{M}_{\text{num}} = \mathcal{M}_{\text{mle}}$.

Also taking into account the use of variance flooring, the dependence of the functions $\mathcal{G}$ on $\mathbf{a}_r$ can be condensed into a matrix

$$\mathbf{G} = \sum_m \gamma^{\text{num}}_m \left[\frac{1}{\text{floor}({\Sigma^{(m)}}_{rr})} \mathbf{X}^{\text{num}}_{(m)} + k_m \mathbf{W}^{\text{mle}}_{(m)} \right]$$

where

$$k_m = \left\{ \begin{array}{l} \text{floor}({\Sigma^{(m)}}_{rr}) = {\Sigma^{(m)}}_{rr} \rightarrow \frac{1}{{\Sigma^{(m)}}_{rr}} - \frac{\mathbf{a}_r \mathbf{X}^{\text{num}}_{(m)} \mathbf{a}^t_r}{{\Sigma^{(m)}}^2_{rr}} \\ \text{floor}({\Sigma^{(m)}}_{rr}) \neq {\Sigma^{(m)}}_{rr} \rightarrow 0 \end{array} \right\},$$

$\mathbf{X}^{\text{num}}_{(m)}$ is the augmented variance $\mathbf{W}^{\text{num}}_{(m)} + (\mu^{\text{num}}_m - \mu^{\text{mle}}_m)^T(\mu^{\text{num}}_m - \mu^{\text{mle}}_m)$ incorporating the difference in means and ${\Sigma^{(m)}}_{rr} = \mathbf{a}_r \mathbf{W}^{\text{mle}}_{(m)} \mathbf{a}^T_r$ is dimension r of the variance of Gaussian m using the current value of $\mathbf{A}$.

We can define $\mathbf{G}^{\text{num}}$ and $\mathbf{G}^{\text{den}}$ to correspond to the numerator and denominator parts of the objective function, with $\mathbf{G}^{\text{num}}$ calculated as above and $\mathbf{G}^{\text{den}}$ using the denominator statistics instead.

Differentiating the smoothing function $\mathcal{H}$ to give a similar matrix $\mathbf{G}^{\text{mle}}$ gives

$$\mathbf{G}^{\text{mle}} = \sum_m \gamma^{\text{mle}}_m \left(\frac{1}{\text{floor}(\bar{\Sigma}^{(m)}_{rr})} \mathbf{W}^{\text{mle}}_{(m)} \right)$$

where $\bar{\Sigma}_{rr}^{(m)} = \bar{\mathbf{a}}_r \mathbf{W}_{(m)}^{\text{mle}} \bar{\mathbf{a}}_r^T$ is dimension r of the variance calculated using the "old" transform $\bar{\mathbf{A}}$. A term corresponding to k_m does not appear since the variance of the Gaussian does not depend on the new transform $\mathbf{A}$ and so its differential with respect to it is zero.

The rest of the dependence on $\mathbf{a}$ arises from the term $(1-\alpha+E)T\log|A|$ in Equation 6. Define $\mathbf{c}_r$ as $(1-\alpha+E)T$ times row r of the matrix of cofactors of $|A|$. The term $\log(\mathbf{c}_r\mathbf{a}_r^T)$ captures the dependence of the objective function on $\mathbf{a}_r$ via $\log|A|$.

Now, if we define $\mathbf{G} = \mathbf{G}^{\text{num}} - \alpha\mathbf{G}^{\text{den}} + E\mathbf{G}^{\text{mle}}$, the new row $\mathbf{a}_r$ can be calculated by maximising the function:

$$\log(\mathbf{c}_r\mathbf{a}_r^T) - 0.5\mathbf{a}_r\mathbf{G}\mathbf{a}_r^T.$$

Differentiating w.r.t. $\mathbf{a}_r$ gives:

$$\frac{\mathbf{c}_r^T}{\mathbf{c}_r\mathbf{a}_r^T} = \mathbf{G}\mathbf{a}_r^T \tag{9}$$

Multipling on the left by the inverse of $\mathbf{G}$ gives

$$\frac{\mathbf{G}^{-1}\mathbf{c}_r^T}{\mathbf{c}_r\mathbf{a}_r^T} = \mathbf{a}_r^T.$$

$\mathbf{a}_r$ is given by $\frac{1}{\beta}\mathbf{c}_r\mathbf{G}^{-1}$ where $\beta = \mathbf{c}_r\mathbf{a}_r^T$, and β can be found by multiplying Equation 9 on the left by $\mathbf{c}_r$, giving

$$\frac{\mathbf{c}_r\mathbf{G}^{-1}\mathbf{c}_r^T}{\mathbf{c}_r\mathbf{a}_r^T} = \mathbf{c}_r\mathbf{a}_r^T$$

so $\beta = \sqrt{\mathbf{c}_r\mathbf{G}^{-1}\mathbf{c}_r^T}$.

The above is valid for updating the accepted rows of $\mathbf{A}$. For the rejected rows (which also need to be updated), the matrix $\mathbf{G}$ should instead be calculated as follows:

$$\mathbf{G} = T((1-\alpha)\frac{1}{\Sigma_{rr}} + E\frac{1}{\bar{\Sigma}_{rr}})\Sigma^{\text{data}}$$

where Σ^{data} is the global variance of the data, $\Sigma_{rr} = \mathbf{a}_r\Sigma^{\text{data}}\mathbf{a}_r^T$ is dimension r of the data transformed by the current value of $\mathbf{A}$, $\bar{\Sigma}_{rr} = \bar{\mathbf{a}}_r\Sigma^{\text{data}}\bar{\mathbf{a}}_r^T$ is dimension r of the data transformed using the "old" transform $\bar{A}$.

Bibliography

[1] Mark J.F Gales, “Maximum Likelihood Multiple Projection Schemes for Hidden Markov Models, ” *Technical report no. 365, Cambridge University Engineering Dept.*, July 2001.

[2] Mark J.F Gales, “Semi-tied covariance matrices for hidden Markov models” *IEEE Transactions on Speech and Audio Processing 7:272-281*, 1999