exchangeable prior, labels are arbitrary entities that do not necessarily remain consistent over Gibbs iterations), we cannot simply integrate over multiple Gibbs-sampled state sequences. We propose two solutions to this problem. The first, which we refer to as the *likelihood metric*, is to simply choose from a fixed set of Gibbs samples the one that produces the largest likelihood given the estimated parameters (marginalizing over state sequences), and then produce the corresponding Viterbi state sequence. This heuristic, however, is sensitive to overfitting and will, in general, be biased toward solutions with more states.

An alternative, and more robust, metric is what we refer to as the *minimum expected Hamming distance*. We first choose a large reference set $\mathcal{R}$ of state sequences produced by the Gibbs sampler and a possibly smaller set of test sequences $\mathcal{T}$. Then, for each sequence $z^{(i)}$ in the test set $\mathcal{T}$, we compute the empirical mean Hamming distance between the test sequence and the sequences in the reference set $\mathcal{R}$; we denote this empirical mean by $\hat{H}_i$. We then choose the test sequence $z^{(j^*)}$ that minimizes this expected Hamming distance. That is,

$$z^{(j^*)} = \arg \min_{z^{(i)} \in \mathcal{T}} \hat{H}_i.$$

The empirical mean Hamming distance $\hat{H}_i$ is a *label-invariant loss function* since it does not rely on labels remaining consistent across samples—we simply compute

$$\hat{H}_i = \frac{1}{|\mathcal{R}|} \sum_{z^{(j)} \in \mathcal{R}} \text{Hamm}\big(z^{(i)}, z^{(j)}\big),$$

where $\text{Hamm}(z^{(i)}, z^{(j)})$ is the Hamming distance between sequences $z^{(i)}$ and $z^{(j)}$ after finding the optimal permutation of the labels in test sequence $z^{(i)}$ to those in reference sequence $z^{(j)}$. At a high level, this method for choosing state sequence samples aims to produce segmentations of the data that are *typical* samples from the posterior. Jasra, Holmes and Stephens (2005) provide an overview of some related techniques to address the label-switching issue. Although we could have chosen any label-invariant loss function to minimize, we chose the Hamming distance metric because it is closely related to the official NIST *diarization error rate* (DER) that is calculated during the evaluations. The final metric by which the speaker diarization algorithms are judged is the *overall* DER, a weighted average over the set of meetings based on the length of each meeting.

In Figure 12(a) we report the DER of the chain with the largest likelihood given the parameters estimated at the 10,000th Gibbs iteration for each of the 21 meetings, comparing the sticky and original HDP-HMM with DP emissions. We see that the sticky model's temporal smoothing provides substantial performance gains. Although not depicted in this paper, the likelihoods based on the parameter estimates under the original HDP-HMM are almost always higher than those under the sticky model. This phenomenon is due to the fact that without the sticky parameter, the HDP-HMM over-segments the data and thus produces parameter estimates more finely tuned to the data, resulting in higher likelihoods.