Gibbs samples, there are still state sequence sample paths with very rapid dynamics. The result of this fragmentation into redundant states is a slight reduction in predictive performance on test sequences, as in the multinomial emission case. See Figure 11(b).

8. Speaker diarization results. Recall the *speaker diarization* task from Section 2, which involves segmenting audio recordings from the NIST Rich Transcription 2004–2007 database into speaker-homogeneous regions while simultaneously identifying the number of speakers. In this section we present our results on applying the sticky HDP-HMM with DP emissions to the speaker diarization task.

A minimum speaker duration of 500 ms was set by associating two preprocessed MFCCs with each hidden state. We also tied the covariances of within-state mixture components (i.e., each speaker-specific mixture component was forced to have identical covariance structure), and used a nonconjugate prior on the mean and covariance parameters. We placed a normal prior on the mean parameter with mean equal to the empirical mean and covariance equal to 0.75 times the empirical covariance, and an inverse-Wishart prior on the covariance parameter with 1000 degrees of freedom and expected covariance equal to the empirical covariance. Our choice of a large degrees of freedom is akin to an empirical Bayes approach in that it concentrates the mass of the prior in reasonable regions based on the data. Such an approach is often helpful in high-dimensional applied problems since our sampler relies on forming new states (i.e., speakers) based on parameters drawn from the prior. Issues of exploration in this high-dimensional space increase the importance of the setting of the base measure. For the concentration parameters, we placed a Gamma$(12, 2)$ prior on γ, a Gamma$(6, 1)$ prior on $\alpha + \kappa$, and a Gamma$(1, 0.5)$ prior on σ. The self-transition parameter ρ was given a Beta$(500, 5)$ prior. For each of the 21 meetings, we ran 10 chains of the blocked Gibbs sampler for 10,000 iterations for both the original and sticky HDP-HMM with DP emissions. We used a sticky HDP-HMM truncation level of $L = 15$, where the DP-mixture-of-Gaussians emission distribution associated with each of these L HMM states was truncated to $L' = 30$ components. Our choice of L significantly exceeds the typical number of speakers, which in the NIST database tends to be between 4 and 6. In practice, our sampler never approached using the full set of possible states and emission components.

In order to explore the importance of capturing the temporal dynamics, we also compare our sticky HDP-HMM performance to that of a Dirichlet process mixture of Gaussians that simply pools together the data from each meeting, ignoring the time indices associated with the observations. We considered a truncated Dirichlet process mixture model with $L = 50$ components and a Gamma$(6, 1)$ prior on the concentration parameter γ. The base measure was set as in the sticky HDP-HMM.

For the NIST speaker diarization evaluations, the goal is to produce a single segmentation for each meeting. Due to the label-switching issue (i.e., under our