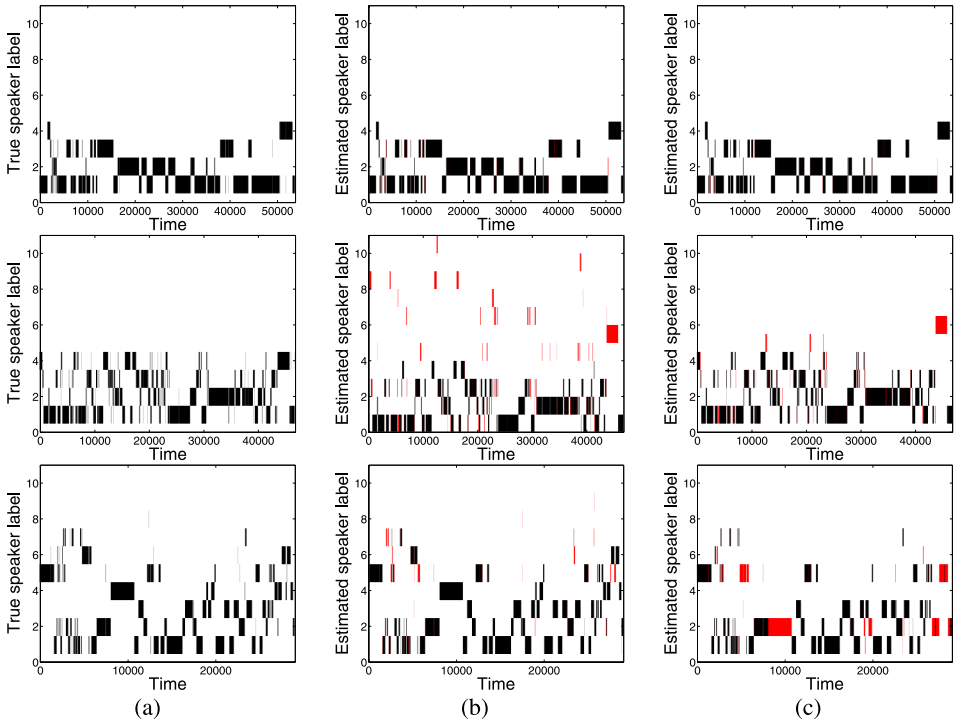

FIG. 13. *Qualitative results for meetings AMI_20041210-1052 (meeting* 1, *top), CMU_20050228-1615 (meeting* 3, *middle) and NIST_20051102-1323 meeting (meeting* 16, *bottom).* (a) *True state sequence with the post-processed regions of overlapping- and nonspeech time steps removed.* (b) *and* (c) *Plotted only over the time-steps as in* (a), *the state sequences inferred by the sticky HDP-HMM with DP emissions at Gibbs iteration* 10,000 *chosen using the most likely and minimum expected Hamming distance metrics, respectively. Incorrect labels are shown in red. For meeting* 1, *the maximum likelihood and minimum expected Hamming distance diarizations are similar, whereas in meeting* 3 *we clearly see the sensitivity of the maximum likelihood metric to overfitting. The minimum expected Hamming distance diarization for meeting* 16 *has more errors than that of the maximum likelihood due to poor mixing rates and many samples failing to identify a speaker.*

to this trend is the NIST_20051102-1323 meeting (meeting 16). For the sticky model, the state sequence using the maximum likelihood metric had very low DER [see Figure 13(b)]; however, there were many chains that merged speakers and produced segmentations similar to the one in Figure 13(c), resulting in such a sequence minimizing the expected Hamming distance. See Section 9 for a discussion on the issue of merged speakers. Running meeting 16 for 50,000 Gibbs iterations improved the performance, as depicted by the revised results in Figure 12(c). We summarize our overall performance in Table 1, and note that (when using the 50,000 Gibbs iterations for meeting 16 and 10,000 Gibbs iterations for all other