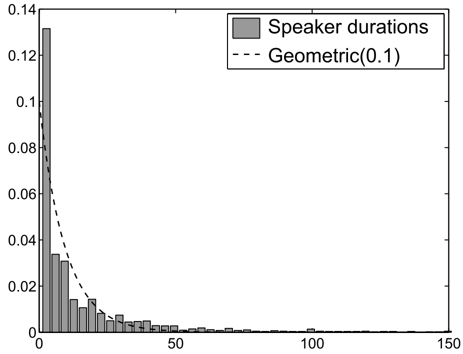

FIG. 2. *Normalized histogram of speaker durations of the preprocessed audio features from the* 21 *meetings in the NIST database. A* Geom(0.1) *density is also shown for comparison.*

(MFCCs),[1] computed over a 30 ms window every 10 ms, as a feature vector. After these features are computed, a speech/nonspeech detector is run to identify and remove observations corresponding to nonspeech. (Nonspeech refers to time intervals in which nobody is speaking.) The preprocessing step of removing nonspeech observations is important in ensuring that the fitted acoustic models are not corrupted by nonspeech information.

When working with this data set, we discovered that the high frequency content of these features contained little discriminative information. Since minimum speaker durations are rarely less than 500 ms, we chose to define the observations as averages over 250 ms, nonoverlapping blocks. This preprocessing stage also aids in achieving speaker dynamics at the correct granularity (as opposed to finer temporal scale features leading to inferring within-speaker dynamics in addition to global speaker changes). In Figure 2 we plot a histogram of the speaker durations of our preprocessed features based on the ground truth labels provided for each of the 21 meetings. From this plot, we see that a geometric duration distribution fits this data reasonably well. This motivates our approach of simply increasing the prior probability of self-transitions within a Markov framework rather than moving to the more complicated semi-Markov formulation of speaker transitions.

Another key feature of the speaker diarization data is the fact that the speaker specific emissions are not well approximated by a single Gaussian; see Figure 3. This observation has led many researchers to consider a mixture-of-Gaussians speaker model, as previously described. As demonstrated in Section 8, we show

[1] Mel-frequency cepstral coefficients (MFCCs) comprise a representation of the short-term power spectrum of a sound on the mel scale (a nonlinear scale of frequency based on the human auditory system response). Specifically, the computation of an MFCC typically involves (i) taking the Fourier transform of a windowed excerpt of a signal, (ii) mapping the log powers of the obtained spectrum onto the mel scale and (iii) performing a discrete cosine transform of the mel log powers. The MFCCs are the amplitudes of the resulting spectrum.