The Annals of Applied Statistics
2011, Vol. 5, No. 2A, 1020–1056
DOI: 10.1214/10-AOAS395

A STICKY HDP-HMM WITH APPLICATION TO SPEAKER DIARIZATION[1]

BY EMILY B. FOX, ERIK B. SUDDERTH, MICHAEL I. JORDAN AND ALAN S. WILLSKY

Duke University, Brown University, University of California, Berkeley and Massachusetts Institute of Technology

We consider the problem of *speaker diarization*, the problem of segmenting an audio recording of a meeting into temporal segments corresponding to individual speakers. The problem is rendered particularly difficult by the fact that we are not allowed to assume knowledge of the number of people participating in the meeting. To address this problem, we take a Bayesian nonparametric approach to speaker diarization that builds on the hierarchical Dirichlet process hidden Markov model (HDP-HMM) of Teh et al. [*J. Amer. Statist. Assoc.* **101** (2006) 1566–1581]. Although the basic HDP-HMM tends to over-segment the audio data—creating redundant states and rapidly switching among them—we describe an augmented HDP-HMM that provides effective control over the switching rate. We also show that this augmentation makes it possible to treat emission distributions nonparametrically. To scale the resulting architecture to realistic diarization problems, we develop a sampling algorithm that employs a truncated approximation of the Dirichlet process to jointly resample the full state sequence, greatly improving mixing rates. Working with a benchmark NIST data set, we show that our Bayesian nonparametric architecture yields state-of-the-art speaker diarization results.

1. Introduction. A recurring problem in many areas of information technology is that of segmenting a waveform into a set of time intervals that have a useful interpretation in some underlying domain. In this article we focus on a particular instance of this problem, namely, the problem of *speaker diarization*. In speaker diarization, an audio recording is made of a meeting involving multiple human participants and the problem is to segment the recording into time intervals associated with individual speakers [Wooters and Huijbregts (2007)]. This segmentation is to be carried out without a priori knowledge of the number of speakers involved in the meeting; moreover, we do not assume that we have a priori knowledge of the speech patterns of particular individuals.

Received April 2010; revised August 2010.

[1]Supported in part by MURIs funded through AFOSR Grant FA9550-06-1-0324 and ARO Grant W911NF-06-1-0076, by AFOSR under Grant FA9559-08-1-0180 and by DARPA IPTO Contract FA8750-05-2-0249.

Key words and phrases. Bayesian nonparametrics, hierarchical Dirichlet processes, hidden Markov models, speaker diarization.