SPEECH SEGMENTATION USING HIERARCHICAL DIRICHLET PROCESSES[1]

Amir Hossein Harati Nejad Torbati and Joseph Picone

Dept. of Electrical and Computer Engineering
College of Engineering
Temple University, Philadelphia, USA
amir.harati@gmail.com, picone@temple.edu

Marc Sobel

Department of Statistics
Fox School of Business and Management
Temple University, Philadelphia, USA
marc.sobel@temple.edu

ABSTRACT

Speech recognition systems have historically used context-dependent phones as acoustic units because they perform well and allow leveraging of linguistic information such as pronunciation lexicons. However, it is desirable in some cases to automatically discover acoustic units, particularly when dealing with a new language for which minimal linguistic resources exist. The process of discovering acoustic units usually consists of two stages: segmentation and clustering. In this paper, we introduce a nonparametric Bayesian approach for segmentation in which Hidden Markov models (HMMs) with an unbounded number of states are used to segment the utterance. An 11% improvement in finding boundaries is demonstrated. A self-similarity measure over segments shows an 88% improvement compared to manual segmentation.

Index Terms— nonparametric Bayesian models, hierarchical Dirichlet processes, speech segmentation

1. INTRODUCTION

Acoustic unit selection is a critical issue in many speech recognition applications where there are limited linguistic resources or training data available for the target language. For example, recently IARPA's Babel program [1] sponsored a competition to create a speech to text system in a mystery language in one week of time using very limited resources. Though traditional context-dependent phone models perform well when there is ample data, automatic discovery of acoustic units offers the potential to provide good performance for resource deficient languages with complex linguistic structures (e.g., African click languages).

Most approaches to automatic discovery of acoustic units [2]-[4] do this in two steps: segmentation and clustering. Segmentation is accomplished using a heuristic method that detects changes in energy and/or spectrum. Similar segments are then clustered using an agglomerative method such as a decision tree. Advantages of this approach include the potential for higher performance than that obtained using traditional linguistic units, and the ability to automatically discover pronunciation lexicons.

In this paper, we propose the use of nonparametric Bayesian methods for segmentation. In this problem, the number of units (or segments) is unknown. One approach is to exhaustively search through a model space consisting of many possible parameterizations. An alternative approach is based on a nonparametric Bayesian statistical model [5] in which the model complexity can be inferred directly from the data. Segmenting an utterance into acoustic units can be approached in a manner similar to that used in speaker diarization [6], where the goal is to segment an audio into regions that correspond to a specific speaker. Fox et al. [6] used one state per speaker and demonstrated segmentation without knowing the number of speakers a priori. Here, we demonstrate that a similar approach can be used to segment the utterance into acoustic units.

Our approach is demonstrated in Figure 1 for an example utterance from the TIMIT Corpus [7]. The segmentation is performed using an extension of Hidden Markov models with an unbounded number of states and mixtures. This model is known as infinite HMM or more recently a Hierarchical Dirichlet Process HMM (HDP-HMM) [6]. It uses a hierarchical Bayesian model to define a nonparametric Bayesian HMM [8].

Relation to Prior Work: We propose a new algorithm for the segmentation of the speech based on a nonparametric Bayesian approach [5] known as an HDP-HMM [6]. Previously a dynamic programming method was applied that incorporated a heuristic stopping criterion [2]-[4]. Recently, Lee & Glass [9] proposed a nonparametric Bayesian approach for unsupervised segmentation of

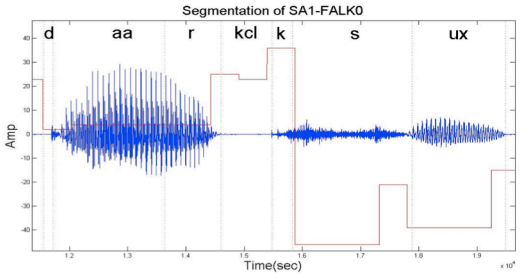

Figure 1. Segmentation of a speech utterance produced through a process of automatic unit discovery is shown by overlaying the duration and index of each unit on the waveform. The height of each rectangle overlay simply indicates the index of that unit.

1. This research was supported in part by the National Science Foundation through Major Research Instrumentation Grant No. CNS-09-58854.

speech. A Dirichlet Process Mixture (DPM) model was used. In order to obtain phoneme-like segments, they modeled each segment using a 3-state HMM. A Gibbs sampler was employed to estimate the segment's boundaries along with their parameters. In our model, we model each segment using one state of an HMM and let HDP-HMM discover the optimal number of segments.

2. HIERARCHICAL DIRICHLET PROCESSES

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain. In the following discussion we will denote the state of the Markov chain at time t with z_t. An observation at time t is conditionally independent of the state of the HMM, and is denoted by $x_t \sim F\left(\theta_{z_t,s_t}\right)$ where s_t is the mixture index. In an HMM, we do not know the exact identity of the previous state. Instead, we could have reached z_t from any state with some probability. In an infinite HMM, the set of predecessor states is infinite, so instead of a transition matrix, we have distribution for the predecessor states which is modeled as a Dirichlet process (DP). We denote this distribution as π_j. The Markovian property of an HMM is denoted by $z_t \sim \pi_{z_{t-1}}$, which implies the current state is only a function of the previous state.

An HDP-HMM is an HMM with unbounded number of states. Since we want the set of predecessor states to be reused at each point in time, so that we can return to various states via a process similar to a self-transition in an HMM, the DPs should be somehow linked together. In order to make sharing of states possible, the base distribution for each DP should be discrete and at the same time have broad support, which simply means all DPs share a common distribution that is a drawn from a DP. This structure is referred as Hierarchical Dirichlet Process (HDP) [8].

Unlike an HDP in which an association of data to a group is assumed to be known a priori, we must infer this association in an HDP-HMM. A major problem with the original formulation of HDP-HMM is state persistence. HDP-HMM has a tendency to create many redundant states and switch rapidly among them [6]. This is mitigated by introducing a sticky parameter, κ, to the definition of HDP-HMM, as shown in Eq. (1):

$$
\begin{aligned}
&\beta \mid \gamma \sim GEM(\gamma) \\
&\pi_j \mid \alpha, \beta \sim DP(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa}) \\
&\psi_j \mid \sigma \sim GEM(\sigma) \\
&\theta_{kj}^{**} \mid H, \lambda \sim H(\lambda) \\
&z_t \mid z_{t-1}, \{\pi_j\}_{j=1}^{\infty} \sim \pi_{z_{t-1}} \\
&s_t \mid \{\psi_j\}_{j=1}^{\infty}, z_t \sim \psi_{z_t} \\
&x_t \mid \{\theta_{kj}^{**}\}_{k,j=1}^{\infty}, z_t \sim F\left(\theta_{z_t s_t}\right)
\end{aligned}
\tag{1}
$$

This parameter encourages consecutive data to belong to the same group (in HMM terms, it increases the probability of a self-transition). The original HDP-HMM formulation can be derived by setting κ = 0. In Figure 2, we depict a graphical representation of this model [6]. Observations are generated from a parametric distribution denoted by θ_{kj}^{**}. Indices j and k are determined by the state and mixture numbers.

In Eq. (1) we show a particular construction of a DP, known as a Griffiths, Engen and McCloskey (GEM) model, or stick-breaking construction, which generates a DP by successively sampling a beta distribution over the remaining part of a stick with an initial length equal to one. The distribution, β, is the base distribution that links all DPs together, and can be interpreted as the expected value of transition distribution. z_t, s_t and x_t are state , mixture index and the observation respectively. This model has been successfully used in several segmentation tasks [6].

The final ingredient in this model is an inference algorithm. Eq. (1) describes a generative model. Inference algorithms are used to infer the values of the latent variables, in this case z_t and s_t. There are several popular approaches for inference including the block sampler [6] used in this work. This sampler employs a Markovian structure of the model to improve its performance. A variation of the forward-backward procedure is used that enables us to sample the state sequence $z_{1:T}$ at once. However, a block sampler needs a fixed truncation level K_z to be specified in advance. This truncation level represents the maximum number of states that the inference algorithm can find. It should be noted that the resulting algorithm is different from a parametric Bayesian HMM because it induces a sparse subset of the K_z possible states [6]. Similarly, a fixed truncation level K_s is used to represent the maximum number of mixtures per state. In practice if both K_z and K_s are sufficiently large the results will be the same as if we use an infinite truncation level.

In our HDP-HMM model, each state of the HMM represents a segment. Since HDP-HMM has an unbounded number of states, the model can infer the number of segments automatically from the data. Modeling each segment with a state of an HMM means that the algorithm segments speech into stationary parts. The resulting segments are usually shorter than phoneme-like segments.

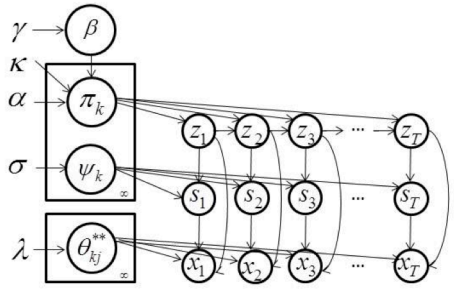

Figure 2. A graphical representation of an HDP-HMM is shown that integrates a mixture distribution model with an infinite HMM.

3. EXPERIMENTS

To evaluate the proposed algorithm, we used data extracted from the TIMIT database [7] that consists of 3696 utterances. This data was chosen because of the existence of highly accurate manual segmentations and also existence of published results. Each utterance was converted into standard MFCC features, and then L frames of data are averaged to produce one output frame. This averaging process is done to ensure that segments have a minimum duration of L frames. Typically, L varies from 1 to 3, corresponding to minimum durations of 10 to 30 ms.

The resulting feature vector was then used as the input to an HDP-HMM for segmentation. A conjugate prior is used to ensure that the posterior distribution remains in the same family of distributions as the prior. Since the posterior distribution in our model is a multivariate normal, we use the normal inverse Wishart distribution for the prior.

In the HDP-HMM model, there are several parameters that must be adjusted among them the truncation level for the number of states (K_z), and the truncation level for the number of mixtures (K_s) per state are more important. K_z and K_s should be set to be larger than the expected number of states and number of mixtures per state. Computational complexity increases linearly with the size of the training data, but quadratically with K_z and K_s.

To measure the performance of the segmentation we followed the approach used in [9] with a tolerance window of 20 ms. In this approach, the discovered boundaries of the segments are compared to the manually segmented reference boundaries. The number of co-occurrences of segments boundaries and phoneme boundaries is called recall. The percent of declared boundaries that coincides with phoneme boundaries is called precision. A single numeric score that represents the combination of these two is referred to as the F-score. It is defined as:

$$\text{F-score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}, \quad (2)$$

and can be interpreted as a weighted average of precision and recall. A comparison of HDP-HMM to other state of the art systems is shown in Table 1. The first row represents a system that performs unsupervised segmentation with no prior information about number of segments for each utterance. The second row represents a system that implements a semi-supervised approach. The third row provides results for a recently proposed nonparametric Bayesian approach [9]. HDP-HMM performs particularly well on recall, which implies that it is finding boundaries that better match the reference phoneme boundaries. The improvement in recall is over 11%. The precision is lower, however, which means there are slightly more false alarms. This is not unexpected since its determination of acoustic units is driven by the complexity of the data. Here we did not count segments which existed inside phoneme boundaries as false alarms since they can be merged with neighboring segments without violating phoneme boundaries, and these segments do not change the overall results significantly.

Table 1. The segmentation performance of the HDP-HMM model is compared to several other nonparametric approaches. HDP-HMM excels in recall while maintaining an acceptable precision.

Algorithm	Recall	Precision	F-score
Dusan & Rabiner (2006) [10]	75.2	66.8	70.8
Qiao et al. (2008) [11]	77.5	76.3	76.9
Lee & Glass (2012) [9]	76.2	76.4	76.3
HDP-HMM	**86.5**	68.5	76.6

Another approach to measure the quality of the segments is to measure the similarity of segments with the same identity. A similarity score, S, can be defined as:

$$S(s_1, s_2) = \begin{cases} s_1 = \frac{1}{MN} \sum_{\substack{I=\{i\} \\ M=|I|}} \sum_{\substack{J=\{j\}=\{j: class(j)=class(i), i \neq j\} \\ N=|J|}} \left| corr(x_i, x_j) \right| \\ s_2 = 1 - \frac{1}{M'N'} \sum_{\substack{I=\{i\} \\ M'=|I|}} \sum_{\substack{J=\{j\}=\{j: class(j) \neq class(i)\} \\ N'=|J|}} \left| corr(x_i, x_j) \right| \end{cases} \quad (3)$$

where s_1 is the in-class similarity score and is defined as the average over the correlation between different instances of segments with identical labels. Similarly, s_2 is the out-of-class dissimilarity score. The quality of segmentation is higher when both numbers are closer to one. It should be noted that the similarity score functions much like a likelihood score — it increases monotonically with an increase in the number of classes. Therefore, for a meaningful comparison, the number of classes being compared for two algorithms must be the same.

In Table 2, we demonstrate the impact these parameters have on segmentation performance. N_s and N_c are the number of discovered states and the number classes respectively. Similarity scores for the manual segmentations and the HDP-HMM algorithm are shown in the last two columns of Table 2. The number of classes for the manual segmentations is fixed to 61, the number of phones used to

Table 2. A demonstration of the HDP-HMM approach to automatic discovery of acoustic units. The in-class similarity scores for the proposal algorithm are significantly higher than those for the manual segmentations.

Experiment	Params. (N_s / N_c)	Manual Segmentations	HDP-HMM
K_z=100, K_s=1, L=1	70/70	(0.44,0.72)	(0.82,0.73)
K_z=100, K_s=1, L=2	33/33	(0.44,0.72)	(0.77,0.73)
K_z=100, K_s=1, L=3	23/23	(0.44,0.72)	(0.75,0.72)
K_z=100, K_s=5, L=1	55/139	(0.44,0.72)	(0.90,0.72)
K_z=100, K_s=5, L=2	53/73	(0.44,0.72)	(0.87,0.72)
K_z=100, K_s=5, L=3	43/51	(0.44,0.72)	(0.83,0.72)

Table 3. Samples of the lexicons are shown for several parameter configurations. The labels in the second and third columns are arbitrarily assigned to acoustic units. There is a reasonable amount of consistency between words with similar phonetic transcriptions.

Exp.	Word	FALK0	FCJF0
K_Z=100 K_s=1 L=1	She	81-2-7-41	27-67-40-41-68
	Wash	45-25-29-54-59-30-94-81	41-45-25-29-54-73-8-4-27-81-17
	Water	29-54-59-28-71-72-98	29-54-28-98
	All	60-54-80-41	29-54-80-41
K_Z=100 K_s=1 L=2	She	60-18-79-70	27-67-40-41-68
	Wash	75-10-51-91-52-60-61	75-10-51-91-19-54-60-61
	Water	10-51-3-99	10-51-3
	All	10-51-70	10-51-70
K_Z=100 K_s=5 L=1	She	35-75-43-89	35-76-43-89
	Wash	70-29-48-47-88-7-100-35-41	70-48-47-88-7-15-6-35-41
	Water	48-47-88-73-50-57-45	47-88-39-47
	All	25-87-7-43	47-30-43
K_Z=100 K_s=5 L=3	She	24-6-86	17-38-6-30-58
	Wash	43-26-30-73-24	5-43-26-30-76-10-17-59-78
	Water	43-26-30-50-69	26-50-80
	All	26-30-69-55	26-69

mark the corpus. For HDP-HMM, N_s is typically change between 20 and 75 while N_c varies between 23 and 139 depending on the configuration settings.

Note that increasing the number of classes results in an increase in the in-class similarity scores, but the out-of-class dissimilarity scores remain relatively constant. If we consider the last row of the table, we observe that the number of classes (51) is roughly comparable to the number of phones (61), yet the similarity score for HDP-HMM is 88% larger (0.83 vs. 0.44). This suggests that HDP-HMM segmentation is a promising approach. We also believe this demonstrates that the segments discovered by our algorithm can be modeled with simpler models when incorporated into more complex systems such as a speech recognition system.

In Table 3, excerpts from automatically discovered lexicons are shown for four different parameter configurations. This data was the result of processing utterance SA1 for speakers FALK0 and FCJF0. The labels shown are arbitrarily assigned during the discovery process. Though we don't expect the value of the label to be repeated for a different set of data, we can see that there is a general similarity in the sequence of labels for similar words spoken by different speakers. For example, word "all" in the first row of the table is represented by segments "60-54-80-41" for FALK0 and "29-54-80-41" for FCJF0.

Further analysis revealed that the segments 60 and 29 are also acoustically close. The normalized distance between the mean of the Gaussian distributions that represent each segment is 11.6 while the average distance between two arbitrary segments is 41.1. This indicates that segments 29 and 60 are accounting for slightly different pronunciations of the initial phone.

Segments derived using the proposed algorithm follow an *N*-gram statistical structure. For example, in the second row of Table 3, segment 79 always follows segment 18, and segment 12 always follows segments 70, 79 and 68 (which are very close in terms of acoustic distance).

The first two experiments use a single Gaussian emission for each state (K_s=1). The last two experiments use Gaussian mixtures (K_s=5) where the maximum number of mixtures per state is K_s. The flexibility added by the mixture model improves the consistency of the segmentation. For example, by comparing the word "she" for the first and third experiments in Table 3, we observe that the segmentations for both speakers are much more similar in the third experiment (K_s=5) than the first experiment (K_s=1). Recall that in this model the number of mixtures per state can vary, and the number of derived classes grows only as needed based on the complexity of the data. Hence, the model essentially adapts to the data.

Figure 1 demonstrates that the boundaries found by HDP-HMM approximately coincide with boundaries found from manual segmentation, supporting the results reported in Table 1. However, in some cases the discovered segments combine several phonemes (e.g., /aa r/) while in other instances a single phoneme is divided into more than one segment (e.g., /s/). This splitting does not violate the phoneme boundaries and can be interpreted as a finer representation of the phoneme. This is supported by the fact that for a comparable number of classes the similarity score is higher for the automatically discovered segments. This suggests that the splitting/merging phenomenon inherent to the HDP-HMM improves the segmentation process and the resulting segments can generate a set of acoustic units that represent the data more consistently.

4. CONCLUSIONS

We have investigated the application of an HDP-HMM model to segmentation of speech. It was shown that this segmentation model produces meaningful and consistent results. Discovered boundaries generally coincide with the boundaries for manually segmented phonemes. It was shown that for a comparable number of classes, the HDP-HMM model improves segmentation self-similarity score by more than 88%. Moreover, we have shown that our algorithm improves recall rate over other state of the art algorithms by more than 11%.

Future research will be focused on clustering segments produced by HDP-HMM and automatic generation of a corresponding lexicon. This step can also be implemented using a nonparametric Bayesian approach. This is the last crucial step in achieving our goal of a system based entirely on nonparametric Bayesian approaches.

5. REFERENCES

[1] M. Harper, "IARPA Solicitation IARPA-BAA-11-02," *http://www.iarpa.gov/solicitations_babel.html*, 2011.

[2] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Communication*, vol. 29, no. 2–4, pp. 99–114, 1999.

[3] B. Ma, et al., "An Acoustic Segment Modeling Approach to Automatic Language Identification," in *Proc. of INTERSPEECH*, 2005, pp. 2829–2832.

[4] K. Paliwal, "Lexicon-building methods for an acoustic sub-word based speech recognizer," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1990, pp. 729– 732.

[5] A. Harati, J. Picone, and M. Sobel, "Applications of Dirichlet Process Mixtures to Speaker Adaptation," in *Proceedings of ICASSP*, 2012, pp. 4321–4324.

[6] E. Fox, et al., "A Sticky HDP-HMM with Application to Speaker Diarization.," *The Annalas of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[7] V. Zue, et al., "Acoustic Segmentation and Phonetic Classification in the SUMMIT System," in *Proceedings of ICASSP*, 1989, pp. 389–392.

[8] Y. Teh and M. Jordan, "Hierarchical Bayesian Nonparametric Models with Applications," in *Bayesian Nonparametrics: Principles and Practice*, Cambridge-UK: Cambridge University Press, 2010, pp. 158–207.

[9] C. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the ACL*, 2012, pp. 40–49.

[10] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *Proceedings of INTERSPEECH*, 2006, pp. 1317–1320.

[11] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithms and comparisons," in *Proceedings of ICASSP*, 2008, pp. 3989–3992.