**SPEECH Segmentation** **USING HIERARCHICAL Dirichlet Processes1**

*Amir Hossein Harati Nejad Torbati and Joseph Picone Marc Sobel*

Dept. of Electrical and Computer Engineering Department of Statistics

College of Engineering Fox School of Business and Management

Temple University, Philadelphia, USA Temple University, Philadelphia, USA

amir.harati@gmail.com, picone@temple.edu 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 compared to other state of the algorithms is observed; moreover a self-similarity measure over segments shows 88% improvement in compare to the manual segmentation on the TIMIT Corpus.

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

# 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.



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.

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. 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 acoustical 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:** In this paper, we propose a new algorithm for the segmentation of the speech. We apply 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 speech. Unlike us they used a Dirichlet Process Mixture (DPM). In order to obtain phoneme-like segments, they modeled each segment using a 3 state HMM. They have used a Gibbs sampler to estimate the segments along with their parameters.

# 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 *zt*. An observation at time *t* is conditionally independent of the state of the HMM, and is denoted by  where st is the mixture index. In an HMM, we do not know the exact identity of the previous state. Instead, we could have reached *zt* 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 , 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) [7].

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.  :



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 as a special case by setting κ = 0. In Figure 2, a graphical representation of this model is depicted [6]. Observations are generated from a parametric distribution denoted by . Indices *j* and *k* are determined by the state and mixture numbers.

In Eq.  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. *zt*, *st* and *xt* are state , mixture index and the observation respectively. This model has been successfully used in several speech segmentation tasks [6].

The final ingredient in this model is an inference algorithm. Eq.  describes a generative model. Inference algorithms are used to infer the values of the latent variables, in this case *zt* and *st*. 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 *z1:T* at once. However, a block sampler needs a fixed truncation level *Kz* 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 *Kz* possible states [6]. Similarly, a fixed truncation level *Ks* is used to represent the maximum number of mixtures per state. In practice if both *Kz* and *Ks* are sufficiently large the results will be the same as if we use an infinite truncation level.



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

The proposed approach in this paper is based on 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 and therefore resulted segments are usually shorter than phoneme-like segments.

Table 1- The segmentation performance of our model in compare to state of the art unsupervised and semi-supervised models.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

# Experiments

To evaluate the proposed algorithm, we used data extracted from the TIMIT database [7] which 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 msec.

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 (Kz), and the truncation level for the number of mixtures (Ks) per state are more important. Kz and Ks 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 Kz and Ks.

To measure the performance of the segmentation we followed the approach used in [9] with a tolerance window of 20msc. The idea is to measure the discovered boundaries of the segments to a ground truth (i.e. manually phonetic transcription). The number of co-occurrences of segments boundaries and phoneme boundaries is called recall. Percent of declared boundaries that coincides with phoneme boundaries called precision. A single number score, named F-score, can also be defined as :



Table 1 shows a comparison of HDP-HMM algorithm with other state of the art systems, including a nonparametric Bayesian method proposed in [9]. [10] is the baseline unsupervised system in which the number of boundaries for each utterance is not known in advance and [11] is a semi-supervised algorithm where the number of boundaries is assumed to be known. As it can be seen HDP-HMM has better performance in compare to other methods especially the recall rate is significantly higher which mean better co-occurrence with phoneme boundaries. F-score is comparable to other approaches; however, precision is lower which means there are more false alarms. Also this can be explained by modeling segments with a Gaussian distribution while phonemes (and phoneme-like segments like in [9]) use 3 state HMMs to model each segment.

Another approach to measure the quality of the segments is by measuring the similarity of segments with the same identity. This approach has not been used in other publications and therefore we used it to compare the quality of our segments against phonemes (i.e. considering phonemes as a manual segmentation.) A similarity score has been defined which consists of a pair of numbers: (1) a measure of the similarity of segments with identical labels, a measure of the dissimilarity of segments with non-identical labels. The quality of segmentation is higher when both numbers are closer to one. 

In this equation, *s1* is the in-class similarity score and is defined as the average over the absolute correlation between different instances of segments with identical labels. Similarly, *s2* is the out-of-class dissimilarity score.

It should be noted that the similarity score functions much like a likelihood score –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. Ns and Nc are respectively the number of discovered states and the number classes. From experiments we can see Ns is typically change between 20 and 75. 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 mark the corpus. For HDP‑HMM, the number of classes 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 shows that the HDP‑HMM segmentation is promising.

In Table 3, excerpts from automatically discovered lexicons are shown for four different parameter configurations. This data resulted from processing utterance SA1 for speakers FALK0 and FCJF0. The labels shown are arbitrarily assigned during the automatic 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” for the first experiment is represented with segments “60‑54‑80‑41” for FALK0 and “29‑54‑80‑41” for FCJF0.

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.**  **(Ns / Nc)** | **Manual Segmentations** | **HDP-HMM** |
| Kz=100, Ks=1, L=1 | 70/70 | (0.44,0.72) | (0.82,0.73) |
| Kz=100, Ks=1, L=2 | 33/33 | (0.44,0.72) | (0.77,0.73) |
| Kz=100, Ks=1, L=3 | 23/23 | (0.44,0.72) | (0.75,0.72) |
| Kz=100, Ks=5, L=1 | 55/139 | (0.44,0.72) | (0.90,0.72) |
| Kz=100, Ks=5, L=2 | 53/73 | (0.44,0.72) | (0.87,0.72) |
| Kz=100, Ks=5, L=3 | 43/51 | (0.44,0.72) | (0.83,0.72) |

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).

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** |
| Kz=100 Ks=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 |
| Kz=100 Ks=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 |
| Kz=100 Ks=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 |
| Kz=100 Ks=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 |

The first two experiments use a single Gaussian emission for each state (Ks=1). The last two experiments use Gaussian mixtures (Ks=5) where the maximum number of mixtures per state is Ks. The flexibility added by the mixture model improves the consistency of the segmentation. For example, by comparing the word “she” for experiments one and third we can see for the third experiment, segmentations for both speakers are much more similar than segmentations for the first experiment. 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. The model essentially adapts to the data.

Figure 1 demonstrates that the boundaries founded by the proposed method approximately coincide with boundaries found from manually segmentation of the speech utterance into phonemes which is in agreement with our results reported in Table 1. However, in some cases the automatically 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/). We demonstrated that when the number of classes is comparable to the number of phonemes the similarity score is higher for the automatically discovered segments. This suggests that the splitting/merging phenomena inherent to the HDP‑HMM improve the segmentation process and the resulting segments can generate a set of acoustic units that represent the data more consistently.

# ConclusionS

We have investigated application of an HDP-HMM model to segmentation of speech data. It was shown that this segmentation model produces meaningful and consistent results. Discovered boundaries found by our algorithm generally coincide with the boundaries for manually segmented phonemes. However, sometimes, an automatically discovered segment can cover more (or less) than an entire phoneme. 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 the recall rate by more than 11% in compare to other state of the art systems which includes a recently proposed nonparametric Bayesian algorithm.

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, thereby achieving our goal of a system entirely based on nonparametric Bayesian approaches.

# 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 50th Annual Meeting of the Association for Computational Linguistics, 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 phoeme segmentation: Objectives, algorithms and comparisons,” in proceedings of ICASSP, 2008, pp. 3989 – 3992.