**Segmentation of Speech Utterances using HDP-HMM**

 *Amir Hossein Harati Nejad Torbati and Joe 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

One of the important assumptions in designing a speech recognizer is the availability of some sort of acoustic units. This assumption is not always correct and generally we are interested to discover acoustic units automatically. The process of discovering acoustic units usually consists of two stages. The first stage is segmentation and the second one is clustering. In this paper, we propose a nonparametric Bayesian approach for the segmentation problem which unlike most other works is not based on heuristic approaches. A small dataset extracted from SA part of TIMIT is used to demonstrate the capability of our approaches.

**Index Terms—** nonparametric Bayesian models, Automatic acoustic discovery, Segmentation, Hidden Markov models

#  Introduction

Automatic discovery of acoustical units is of some interest in speech recognition applications. Most algorithms approach this problem in two steps of segmentation and clustering ‎[1],‎[2],‎[3] and ‎[4].

Usually segmentation is accomplished using a heuristic method (e.g. energy changes of the spectrum). After the initial segmentation, similar segments are clustered in several groups using a heuristic clustering algorithm (e.g. tree). Each group represents an acoustical unit. It is expected that acoustical units and dictionary generated directly from speech works better than traditional linguistic based units (e.g. phonemes); moreover, for many languages there is not enough linguistic or even written resources and therefore automatically generated units are the only available option.

Sponsor section: Owls nest staff

 In this paper, the goal is to investigate the first step using nonparametric Bayesian approach. Unlike most other works, our approach is not heuristic. In this approach we use HMMs to segment an utterance into several homogenous regions. HMMs provide a very powerful toolkit for segmentation, however, the number of states should be known in advance. Since the number of segments is not known a priori, we cannot use HMMs directly. Fortunately, a new variant of HMMs, HDP-HMM, is recently proposed based on nonparametric Bayesian statistics ‎[5] and ‎[6]. HDP-HMM provides a variant of HMM with unbounded number of states. The basic idea for segmenting an utterance into acoustical units is similar to speaker diarization problem ‎[6]. In that problem, the goal is to segment an audio file into speaker homogenous areas. One of the recent and successful approaches toward that problem was based on HDP-HMM ‎[6]. In that application, each state represents a speaker, experiments show that the model successfully can capture speakers without knowing the number of speakers a priori.

Figure 1 shows the result of segmentation using the proposed algorithm for SA1 and speaker FALK0 from TIMIT dataset and with minimum segment length equal to 30 msec. As it can be deduced from this figure the segments founded by the algorithm is consistent with the intuition. It should be noted that these segments are not the final acoustical units and might be regrouped after the clustering stage. One important assumption is that acoustical units consist of relatively homogenous regions. This assumption automatically encourages the using of more units since each unit is relatively simple and cannot be accounted for complex acoustical events individually.

In the subsequent sections, we show that HDP-HMM can handle the segmentation problem and despite of different accents can find similar results (to some extend) for words spoken by different speakers.



Figure 1-Segmentation of Utterance SA1- speaker FALK0 using the proposed algorithm

#  HDP-HMM

In this section, we review the definition of HDP-HMM; however, the backgrounds and details are not discussed in this paper. Interested reader can refer to ‎[7] and ‎[6]; for a more compact review refer to ‎[8].

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain ‎[9]. In the following discussion we will denote the state of the Markov chain at time  with  and the state-specific transition distribution for state by.The Markovian structure means. Observations are conditionally independent given the state of the HMM and are denoted by.

HDP-HMM is an extension of HMM in which the number of states can be unbounded. The idea is relatively simple; at each statewe should be able to go to an infinite number of states so the transition distribution should be a draw from a Dirichlet Process (DP). On the other hand, we want reachable states from one state to be shared among all states so these DPs should somehow be linked together. The result is a Hierarchical Dirichlet Process (HDP). In an HDP-HMM each state corresponds to a group and therefore, unlike HDP in which an association of data to groups is assumed to be known a priori, we are interested to infer this association. The major problem with original HDP-HMM is the state persistence. HDP-HMM has a tendency to make many redundant states and switch rapidly among them ‎[6]. This problem is solved by introducing a sticky parameter to the definition of HDP-HMM ‎[6]. Equation shows the definition of a sticky HDP-HMM with unimodal emissions.is a sticky hyper-parameter and generally can be learned from data. Original HDP-HMM is a special case with. From this equation we can see for each state (group) we have a simple unimodal emission distribution. This limitation can be addressed using a more general model defined in . In this model, a DP is associated with each state and a model with augmented stateis obtained. Figure 2 shows a graphical representation.



Figure 2- Graphical model of HDP-HMM ‎[6]

 

 

 

## Inference and learniNG

Inference can be implemented in several ways. In this paper we have used a block sampler which proposed by ‎[6]. This sampler uses Markovian structure of the model to improve its performance. In this algorithm a fixed truncation level Kz is used for the number of states. However, it should be noted that the result is different from a classical parametric Bayesian HMM since the truncated HDP priors induce a shared sparse subset of the Kz possible states ‎[6]. In short, we obtain an approximation to the nonparametric Bayesian HDP-HMM with maximum number of possible states set to Kz. However, for almost all applications this should not cause any problem if we set Kz reasonably high. Similarly, a fixed truncation level Ks is used for number of mixtures.

# Experiments

## Data

The data is extracted from TIMIT database ‎[10] and consists of SA1 sentence for 2 speakers: FALK0 and FCJF0. The reason to keep the dataset small was to investigate the HDP-HMM segmentation approach quickly before using this approach for a more challenging dataset. The existence of the manual transcription for TIMIT was another important factor. This helps us to align the segments with words rather easily. In other words, dictionary can be generated with just aligning. For other datasets we have to use a forced-alignment procedure to do the same thing.

## Experiment setup

First each utterance is converted into MFCC features (window length=25 msec, frame rate=10 msc), then frames converted into block of size L (without overlap) by averaging. The result is fed into HDP-HMM and segmented into several states. Conjugate priors used through all experiments.

There are several parameters that we change through these experiments.

* block\_size (L) : number of frames that make a block.
* Kz: truncation level for number of states.
* Ks: truncation level for number of mixtures.

It should also be noted that segments labels are arbitrary and specifically there is no relation for segment labels among different experiments (Tables).

## Results

Table 1 shows a part of a lexicon generated from our algorithm using (L=1,Kz=100,Ks=1) for FALK0 and FCJF0. From this table, we can see there is a general similarity in sense of segment sequence for similar words spoken by different speakers. It can be seen that some of the segments are speaker specific. For example, word “all” is represented with segments (60,54,80,41) for FALK0 and (29,54,80,41) for FCJF0, however, a closer look shows that segments 60 and 29 are acoustically very close. The normalized distance between the mean of Gaussian distributions that represents each segment is 11.64 while the average distance between two arbitrary segments is 41.13. The correlation between the two means is 80.15% while the average absolute correlation is around 28.77%. This means segments 29 and 60 are representing different flavor of the same unit and might merge in a later clustering stage. Another interesting thing is the fact that some segments are relatively short (shorter than 30 msec which is the minimum length of many state of the art system for different models.) This suggests that using an appropriate variable length models might be helpful in modeling acoustic units.

Table 1- Lexicon generated using (L=1,Kz=100,Ks=1)

|  |  |
| --- | --- |
| Speaker |  |
| FALK0 | sil: 81{14}she: 81{10}-2-7{2}-41{7}your: 33-79{6}-41{2}-94wash: 45{6}-25{6}-29{3}-54{2}-59{3}-30{2}-94-81{8}water: 25-29{4}-54{3}-59{4}-28-71{2}-72{8}-98all: 60{11}-54{3}-80{2}-41{4}year: 41{3}-74{16}-79{2}-89{2}-71-72{3}-76{8}-83 |
| FCJF0 | sil: 81{10)-17-81{5}-27she: 27{7}-67{2}-40-41{6}-68your: 4-27{4}-67{2}-40-41{7}wash: 41-45{3}-25{5}-29{7}-54{3}-73{2}-8{2}-4-27{6}-81-17{3}water: 29{9}-54{2}-28{7}-98{5}all: 29{12}-54-80{2}-41{2}year: 41{2}-74{10}-89{7}-71{8} |

Table 2 shows the result for (L=2,Kz=100,Ks=1). It seems segments follow an n-gram statistical structure. For example segment 79 always follows segment 18, or segment 12 always follows segments 70, 79 and 68 (which are very close in an acoustical distance sense.). Because of the larger block size, the accuracy of the mapping is lower, this can be specifically seen by comparing words “wash” and “water” from Table 1 and Table 2 we can see in this experiment segment 75 which is closely related to “w” is not presented for word “wash” for both speakers. However for pervious experiment segments 45 and 25 (distance 11) are presented for “wash” and “water” for speaker FALK0 while for speaker FCJF0 segment 45 only exists for word “wash”. By listening to the audio file it seems that FCJF0 skip “w” for “water”. The fact that, pervious experiment can detect “w” for speaker FALK0 and for word “water”, and this experiment cannot do the same, shows that increasing the minimum length of segments can reduce the accuracy of the mapping. If we proceed and use L=3, we see some words (e.g. water) will be represented with just one segment which is also used to represent other words. This practically can render the segmentation useless.

Table 3 shows the results for (L=1,Kz=100,Ks=4). In this case we allow the algorithm to model each segment with a mixture of Gaussians (up to 4 components). Letting each segment to be molded by more than one Gaussian means each segment can potentially models more than one acoustical event. The flexibility of modeling segments using multiple mods can violate the assumption that each segment is relatively simple and homogenous. However, the results show a more reliable and consistent segmentation. One possible way of using this more flexible model is to define each segment with both state and mixture numbers. In this way, each segment will be modeled using a single Gaussian and therefore the simplicity is assured. A later clustering stage can merge similar segments that labeled as different segments in the segmentation stage. Table 4 shows the result for (L=3,Kz=100,Ks=10). Interestingly, we see in this case no word is represented by a single segment unlike the case when Ks were 1 which means the results are more consistent and accurate.

Table 2- Lexicon generated using (L=2,Kz=100,Ks=1)

|  |  |
| --- | --- |
| Speaker |  |
| FALK0 | sil: 37{7}she: 60{5}-18-79-70{3}your: 79-25{3}-70wash:75{6}-10{2}-51{2}-91-52-60{3}-61water: 10{3}—51{3}-3{2}-99{4}all:10{7}-51-70{2}year: 70{11}-48{2}-99{5}-87 |
| FCJF0 | sil: 37{9)-60she: 27{7}-67{2}-40-41{6}-68your: 54-60-18-70{3}-12wash: 75{4}-10{4}-51-91-19-54-60{3}-61water: 10{5}-51-3{6}all: 10{7}-51-70year: 70{8}-48{5} |

Table 3- Lexicon generated using (L=1,Kz=100,Ks=4)

|  |  |
| --- | --- |
| Speaker |  |
| FALK0 | sil: 50{8}-22{4}-100-35she: 35{10{-75{2}-43{5}-89your: 72-91{6}-2{2}-45wash: 70{2}-29{6}-48{4}-47{3}-88{4}-7{2}-100{2}-35{7}-41{2}water: 48-47{4}-88{7}-73{2}-50{3}-57{6}-45all: 25{11}-87{3}-7{2}-43{4}year: 43{18}-31{3}-23{3}-18{9}-13 |
| FCJF0 | sil: 100{5}-50{2}-22-100{10}-35 she: 35{8}-76{3}-43{3}-89{3}your: 31-6{2}-35{2}-76{2}-42{5}-84{2}wash:70{4}-48{5}-47{7}-88{3}-7{2}-15{3}-6{2}-35{5}-41{3}water: 47{9}-88{2}-39{12}-47all: 47{12}-30{3}-43{2}year: 43{12}-76{6}-13{9} |

# Conclusion

In this paper, we investigated the usage of HDP-HMM model for segmenting speech utterances into homogenous sections which can be used as a first step of a nonparametric Bayesian approach for automatic acoustical unit discovery algorithm. It has been shown that HDP-HMM can produce meaning full and consistent results. From experiments, we can claim allowing each state of HDP-HMM to have multiple mixtures (instead of one Gaussian) can improve the consistency of the results. However this can complicate the clustering step. One solution to this problem is to define (and label) segments using both state number and mixture number. In this way, segments remain simple (presented using a Gaussian distribution) and at the same time the segmentation remains consistent and more reliable.

The next step of discovering acoustic units is to cluster segments produced by HDP-HMM and generate a lexicon. This step can also be implemented using nonparametric Bayesian approach which results in a complete nonparametric Bayesian solution for acoustic unit discovery problem. This problem is now under investigation and the result will be published accordingly.

# Reference

1. 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.
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. M. Bacchiani, M. Ostendorf, Y. Sagisaka, and K. Paliwal, “Design of a speech recognition system based on acoustically derived segmental units,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996, pp. 443– 446.
4. G. Goussard and T. Niesler, “Automatic discovery of subword units and pronunciations for automatic speech recognition using TIMIT,” in *Proceedings of the twenty-first annual symposium of the Pattern Recognition Association of South Africa*, 2010, pp. 93–99.
5. Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical Dirichlet Processes,” *Journal of the American Statistical Association*, vol. 101, no. 47, pp. 1566–1581, 2006.

Table 4- Lexicon generated using (L=3,Kz=100,Ks=10)

|  |  |
| --- | --- |
| Speaker |  |
| FALK0 | sil: 45{3}-68{2}she: 24{3}-6-86{2}your: 57{2}-80wash: 43{4}-26-30-73-24{3}water: 43-26-30{2}-50{2}-69{2}all: 26{4}-30-69-55year: 55{6}-57-50-69{3}-7 |
| FCJF0 | sil: 68{6}she: 17{2}-38-6-30-58your: 9-38-70-69{2}wash: 5-43{3}-26{2}-30-76-10-17-59-78water: 26{3}-50{3}-80{2}all: 26{4}-69year: 30-55{3}-57-56{4} |

1. E. Fox, E. Sudderth, M. Jordan, and A. Willsky, “A Sticky HDP-HMM with Application to Speaker Diarization.,” *The Annalas of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.
2. Y. Teh and M. Jordan, “Hierarchical Bayesian Nonparametric Models with Applications,” in *Bayesian Nonparametrics: Principles and Practice*, S. W. Hjort, C. Holmes, P. Mueller, Ed. Cambridge-UK: Cambridge University Press, 2010, pp. 158–207.
3. A. Harati, “Hierarchical Dirichlet Processes and Infinite HMMs,” in *PhD Preliminary Exam, Department of Electrical and Computer Engineering, Temple University*, 2012.
4. L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 879–893, 1989.
5. V. Zue, J. Glass, M. Phillips, and S. Seneff, “Acoustic Segmentation and Phonetic Classification in the SUMMIT System,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1989, pp. 389– 392.