**SPEECH Segmentation 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

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 there is minimal linguistic resources. The process of discovering acoustic units usually consists of two stages: segmentation and clustering. In this paper, we introduce a nonparametric Bayesian approach for the segmentation. We have used a recently proposed infinite Hidden Markov Models (HMMs) which enables us to use the power HMMs for segmenting the signal without specifying the number of segments a priori. Results show the proposed algorithm performs pretty well and can segment the speech data into rather homogenous and consistent parts.

**Index Terms—** nonparametric Bayesian models, speech segmentation, automatic discovery of acoustic units

#  Introduction

Acoustic unit selection is a critical issue in a range of speech recognition applications where there is limited linguistic information or training data available for the target language. For example, recently IARPA’s Babel program  is sponsoring a competition to create a speech recognition system in a mystery language in one week of time using a very small amount of training data. Though traditional context-dependent phone models perform well when there is ample data, automatic discovery of acoustic units offers the potential to provide comparable performance using much fewer parameters, or even better performance for languages with complex linguistic structures (e.g., African click languages).

Most conventional approaches to automatic discovery of acoustical units do this in two steps: segmentation and clustering . Segmentation is typically accomplished using a heuristic method that detects changes in the temporal evolution of the 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 traditional linguistic units, and the ability to automatically discover pronunciation lexicons.

In this paper, we propose the use of nonparametric Bayesian methods for the segmentation portion of this problem. In problems like automatic discovery of acoustic units, the number of units and therefore segments are unknown. One solution is to compare many models by assuming different number of units for each model. However, this approach is computationally expensive and the criteria for the comparison is controversial [???]. An alternative solution is to use nonparametric Bayesian approach, where the number of parameters is unbounded and can be inferred from the data directly.

In this approach we use hidden Markov models (HMMs) to segment an utterance into several homogenous regions. HMMs represent a very powerful model of the temporal behavior of a speech signal. However, the number of states for each model is typically defined a priori and does not vary based on the complexity of the data. HDP-HMMs is an extension of HMMs in which the number of states is unbounded [?4?]. The idea is relatively simple; at each state  we should be able to go to an infinite number of states, therefore the transition distribution should be drawn 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 be linked together. The result is a hierarchical DP (HDP) in which a DP acts as common prior for all other DPs.

Segmenting an utterance into acoustic units can be approached in a manner similar to the speaker diarization problem . In speaker diarization, the goal is to segment an audio file into regions which correspond to a specific speaker. In the HDP-HMM approach, each state represents a speaker. Experiments show that the model can successfully capture speakers without knowing the number of speakers a priori. In parallel to speaker diarization problem, we hypothesize that a similar approach can capture the underlying acoustical units without advance knowledge about their numbers.

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.

----

#  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 ,, and .

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 and . 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 . 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 . 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.



Figure 1. A typical segmentation produced through a process of automatic discovery. It can be deduced that discovered segments are consistent with intuitive expectation.

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.

#  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 and ; for a more compact review refer to .

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  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 . This problem is solved by introducing a sticky parameter to the definition of HDP-HMM . 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 . 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 . 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 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. IARPA Babel
2. 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.
3. Chin-Lee’s work published in ICASSP.
4. 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.
5. 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.
6. A. Harati, “Hierarchical Dirichlet Processes and Infinite HMMs,” in *PhD Preliminary Exam, Department of Electrical and Computer Engineering, Temple University*, 2012.
7. 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.
8. 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.

----

The Dirichlet Process Mixture (DPM) model is a popular application of nonparametric Bayesian methodologies ‎[2]. DPM provides a framework to infer the number of clusters and their parameters jointly. Furthermore, DPM is a data‑driven method. The structure evolves as more data become available. demonstrates this phenomenon using a simulation in which the amount of data changes from 20 to 2000 data points and the corresponding discovered clusters increase from one to three. Because of this adaptability to the amount of available data the probability of over-fitting or under-fitting is significantly reduced.

Dirichlet processes have been previously successfully applied to language modeling problems in speech recognition ‎[3]. However, they have not been extensively applied to the acoustic modeling problem due, in part, to the computational issues involved in parameter estimation. Recent advances in fast computational methods, such as variational inference methods ‎[4], have enabled application to computationally intensive tasks such as acoustic training.

As a proof of concept, we investigated the use of the DPM model for speaker adaptation. Specifically, we explored replacing the process of building regression trees in a maximum likelihood speaker adaption (MLLR) system ‎[5] with a DPM-based model. This is a task well-suited to the DPM model because acoustic model adaptation involves gradually adapting Gaussian means trained on large amounts of data to speaker-specific means. All experiments have been designed using the DARPA 1000-word Resource Management (RM) task ‎[6] and the HTK speech recognizer ‎[7]. We have used a publicly available MATLAB library [8] for variational inference.

#  DIRICHLET Process MIXTURE MODELS

The traditional solution to determining the underlying distribution of some observed data is to assume a finite exponential mixture model and infer the parameters. However, the number of components (clusters) must be determined using computationally expensive model selection methods ‎[1]‎[2]. Results have been marginal at best and the clusters often do not capture the underlying characteristics of the data, but rather simply minimize some global distortion measure. An alternative solution that attempts to preserve underlying modalities is the Dirichlet Process Mixture (DPM) model.

The general form of a K-component mixture model is:

  . (1)

In this formulation,are the mixing proportions and must be positive and sum to one,is usually a parametric distribution (i.e. Gaussian) with parameters . The finite mixture model can be expressed as a hierarchical model ‎[1]:

 (2)

In this formulation α is pseudocount hyper-parameter of Dirichlet priors. $H$ is the prior distribution over the parameters$ θ\_{k}$and δ is the Kronecker delta function. For a Gaussian mixture model, $θ\_{k}=\{μ\_{k},Λ\_{k}\}$ where $μ\_{k }$is the mean vector and $Λ\_{k}$is the covariance matrix. is chosen to be the conjugate prior of $f$. For a Gaussian distribution it would be a normal-inverse-Wishart distribution.

It can be shown ‎[1]‎[2] that when , $G\~DP(α,H)$ would be a Dirichlet process (DP) with a base distribution of $H$and a concentration parameter $ α$. One of the most important properties of $G$ is its discrete nature that results in the clustering property of a DP.

The predictive distribution for a DP is given by [1]:

 (3)

In this formulationis the total number of observations and is the number of previous observations for cluster. It states that the probability of assigning a new observation to cluster  is proportional to its size and the probability of initiating a new cluster is proportional to .

Direct computation of the posterior probability in ‎[3] is intractable; therefore some kind of approximation should be used. The most popular approaches are based on Monte Carlo Markov chain (MCMC) methodologies and particularly Gibbs Sampling methods ‎[1]‎[2]. However, MCMC based methods can be slow to converge and cannot be used in large-scale problems ‎[1]‎[2]‎[4]. A different class of alternative approaches is based on variational inference, in which we recast the inference problem in terms of optimization ‎[8] and then relax the optimization problem to obtain a tractable solution. Mean-field algorithms, which restrict the variational distribution to a factorization model, have been used in inference from DPM ‎[4]‎[8].

In variational inference, the posterior probability  is approximated $ P(Z|X)$with an arbitrary function$ q(Z)$. In other words, because the exact form of  is not known, an approximation is assumed. In the case of mean-field algorithm this approximation is also factorized. The log marginal probability is given by ‎[8]:

  (4)

The goal of optimization problem is to maximize the lower bound  or equivalently minimize the Kullback–Leibler (KL) divergence with respect to. It has been shown that the general solution to this optimization problem follows the form of [9]:

 (5)

The above expression does not give an explicit closed‑form solution; instead it provides the means to obtain the solution iteratively. Because of the convexity of the bound, the convergence is guaranteed ‎[8]. However, it might converge to a local solution. In ‎[4], the authors used a truncated stick-breaking representation for the variational distribution. One of the downsides of this approach is that variational families are not nested over truncation level T ‎[10] and as a result this must be optimized.

This issue is addressed using an accelerated variational Dirichlet process mixture (AVDP) algorithm ‎[10] that can handle extremely large data sets. In ‎[11] two other extensions of the variational inference named collapsed variational stick-breaking (CSB) and collapsed Dirichlet priors (CDP) have been introduced. CSB is based on stick‑breaking representation, but the difference here is to integrate out mixture weights. For CSB, the truncation level T should be specified. CDP, on the other hand, is based on a finite symmetric Dirichlet distribution approximation of a Dirichlet process. For this algorithm, the size of Dirichlet distribution K needs to be specified. Both of these algorithms give comparable results and are considerably faster than Gibbs sampling. In this research we have used AVDP, CSB and CDP inference algorithms.

#  Application To SpEAker AdaptATion

Maximum Likelihood Linear Regression (MLLR) is a well‑known speaker adaptation technique in which mean transformation matrices are estimated using a tree‑based clustering process ‎[5]. Clustering is usually accomplished using a regression class tree which is constructed using a centroid splitting algorithm. This algorithm begins with a single node and recursively grows a tree using an ML-based distance measure. However, this is an ad hoc algorithm and its performance is sensitive to the specific training recipe and the amount of data. Further, it is difficult to determine when the algorithm should be stopped.

In this paper, we explore the use of DPM as an alternate clustering algorithm to investigate the potential advantages of this approach. The procedure we employ is as follows:

****

Figure 2. A comparison of regression tree and ADVP approaches for monophone models.

1. Train speaker independent (SI) models, collecting all mixture components and their frequency of occurrence.
2. Generate samples for each component and cluster them using one of the DPM inference algorithms.
3. Construct a tree structure of the final result.
4. Assign clusters to each component.

Training SI model is done using HTK ‎[7]. MFCC features plus energy and their first and second derivatives have been used. After training, Gaussian components are extracted. Data points are generated for each component proportional to the number of occurrences in the training data and then clustered using one of the DPM inference methods.

Clusters were reorganized in form of a tree for two reasons. First, we need a method to merge clusters to deal with insufficient data. Second, we need this mechanism to be consistent with HTK. The difference with a regular regression tree is in the construction process. While centroid splitting algorithm is a top‑down approach, the proposed algorithm starts from the terminal nodes that are obtained using DPM and merges them using a bottom‑up Euclidean distance‑based approach. Finally components are labeled using a majority-voting scheme. The result of this section is used to compute transformation matrices using a maximum likelihood approach.

**4. RESULTS AND DISCUSSION**

In this section we summarize the results of several experiments conducted on a speaker adaptation task based on the Resource Management Corpus ‎[6].

**4.1. Monophone Models**

In this experiment, monophone models with a single Gaussian mixture have been trained. Before training on the speaker‑dependent data, speaker‑independent models were trained. MLLR models were trained for 12 different speakers. There are 600 training utterance for each speaker in the dataset. shows the word error rate (WER) as number of utterances in training dataset for each speaker changes for both the regression tree and ADVP approaches.

It is evident that ADVP works significantly better than regression tree. The reason is related to the number of clusters discovered by each method. To some extent, result of ADVP resembles broad phonetic classes. For example, distributions related to phoneme “w” and “r” (which are both liquids) are in the same cluster. This is also true for phonemes belonging to other classes however there are some exceptions too since the clustering is done automatically in the feature space and based on the similarities of the distributions.

**4.2. Cross-word Models**

In , we compare results for the speaker adaptation task, but this time cross-word models are used. Several DPM approaches are evaluated: ADVP, CSB and CDP. ADVP performs slightly better than the regression tree approach for medium amounts of training data but works slightly worse when the amount of training data increased. For example, WER for 100 utterances of training data drops from 5.17% to 4.77% and for 600 utterances increases from 3.79% to 4.3%.

It can be seen that CDP and CSB work slightly better than the regression tree approach when the amount of training data is increased. For example, WER decreases from 3.79% for regression tree to 3.53% for both of CDP and CSB algorithms.

The clusters generated using DPM have acoustically and phonetically meaningful interpretations. For instance, CSB generates 20 clusters in which distributions corresponding to states two and four of the “silence” model consist a single cluster. The distribution related to the center state of “silence” model (which is also tied to the distribution of short pause model) is not a member of this particular cluster. This is not unexpected since many words in the corpus are articulated with no pause between them and therefore, the model corresponding to that would be more similar to speech models.

 shows the number of discovered clusters for each method. DPM‑based clustering generally works better when we have a large amount of training data. Therefore, we can combine these systems in order to get better results. The result of cascading ADVP, CDP and CSB with regression tree is presented in . It is evident that cascading a DPM clustering with a regression tree‑based clustering slightly improves the results, demonstrating that the two approaches can complement one another.

**5. CONCLUSION**

****

Figure 3. A comparison of WERs between regression tree-based MLLR and several DPM inference algorithms for cross-word acoustic models. ADVP works better for medium data sizes, while other DPM methods work better for large amounts of data.

In this paper we examined an application of Dirichlet Process Mixtures to an acoustic modeling problem in speaker adaptation. Experimental results were promising, showing that the DPM approach can outperform the classic MLLR approach and decrease the error rate up to 10%. DPM appears to do better when there are larger amounts of data, suggesting application to LVCSR tasks could produce promising results.

In order to fully utilize the potential of nonparametric Bayesian methods we need to redesign the architecture of the speech recognizers and investigate a better reestimation process. Also, in this work we assigned each “distribution” to just one cluster. An obvious extension is to use some form of soft tying. Series and parallel combinations of these systems might also lead to improved results.

**6. References**

1. E. Sudderth, “Graphical models for visual object recognition and tracking,” Ph.D. dissertation, Massachusetts Institute of Technology, May 2006.
2. J. Paisley, “Machine learning with Dirichlet and beta process priors: Theory and Applications”, Ph.D. Dissertation, Duke University, May 2010.
3. P. Liang, M. I. Jordan, and D. Klein. In T. O'Hagan and M. West “Probabilistic grammars and hierarchical Dirichlet processes”, *The Handbook of Applied Bayesian Analysis*, Oxford University Press, 2010.
4. D. M. Blei and M. I. Jordan, “Variational inference for Dirichlet process mixtures,” *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.
5. C. J. Leggetter, “Improved acoustic modeling for HMMs using linear transformations,” Ph.D. Dissertation, University of Cambridge, February 1995.

1. P. Price, W. Fisher, J. Bernstein, and D. Pallett, “The DARPA 1000-Word Resource Management Database for continuous speech recognition,” *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 651-654, New York, New York, USA, April 1988.

Figure 4. The number of discovered clusters is shown. ADVP generates the fewest clusters.

1. *http://htk.eng.cam.ac.uk/docs/docs.shtml*
2. *https://sites.google.com/site/kenichikurihara/academic-software/variational-dirichlet-process-gaussian-mixture-model*
3. C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, New York, USA, 2007.
4. K. Kurihara, M. Welling, and N. Vlassis, “Accelerated variational Dirichlet process mixtures,” *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, Massachusetts, USA, 2007 (editors: B. Schölkopf and J.C. Hofmann).
5. K. Kurihara, M. Welling, and Y. W. Teh, “Collapsed variational Dirichlet process mixture models,” *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, Jan. 2007.