**Department of Electrical and Computer Engineering**

Proposal

for

**Non-Parametric Bayesian approaches for Acoustic Modeling**

submitted to:

Dr. Joseph Picone, Examining Committee Chair

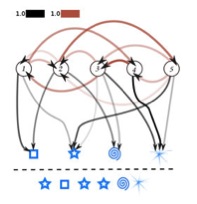
Dr. Iyad Obeid, Committee Member, Department of Electrical and Computer Engineering

Dr. Marc Sobel, Committee Member, Department of Statistics

Dr. Chang-Hee Won, Committee Member, Department of Electrical and Computer Engineering

Dr. Slobodan Vucetic, Committee Member, Department of Computer and Information Sciences

Jan 6, 2013



prepared by:

Amir Harati, PhD Candidate

PhD Advisor: Dr. Joseph Picone, Professor and Chair

Department of Electrical and Computer Engineering

Temple University

College of Engineering

1947 North 12th Street

Philadelphia, Pennsylvania 19122

Tel: 215-204-7597

Email: amir.harati@gmail.com

For further information, please contact Dr. Joseph Picone (email: picone@temple.edu).



**Executive Summary**

The Bayesian framework is an intuitive approach to model the nature. The goal of Bayesian analysis is to reduce the uncertainty about unobserved variables by combining prior knowledge with observations. A fundamental limitation of Bayesian or non-Bayesian approaches is the inflexibility of the model. The accuracy of parameters improves using more data but the model cannot use the data efficiently to learn new structure. One way to tackle this problem is to define many different models and then to select the most likely model based on the observed data. However, the model selection process is computationally expensive and often requires large amounts of data and moreover, the selection criterion is not well defined.

Recently, nonparametric Bayesian methods have become a popular alternative. In such approaches, we do not fix the complexity (e.g. the number of mixture components in a mixture model) *a priori*, and instead place a prior over the complexity (or structure). This prior usually biases the system towards sparse or low complexity, solutions. This helps to control the number of parameters in the model yet allows the structure to be learned during a data-driven training process. Therefore models are not fixed and can utilize new data that become available over the time.

In speech recognition technology, we deal with the complexity problem at many levels. Examples in acoustic modeling include the number of states and the number of mixture components in a hidden Markov model. Also, the number of models (and parameter-sharing between these models) is often determined as a compromise between complexity and computational issues. In language modeling, we must estimate the probabilities of unseen events in very large but sparse N‑gram models. Nonparametric Bayesian modeling has been used to smooth such *N*-gram language models.

In this proposal, our goal is to investigate the applications of nonparametric Bayesian modeling in acoustic modeling problems. Four different (but related) problems have been studied and four approaches within the nonparametric Bayesian framework proposed.

One of these applications is speaker adaption. For this application we propose a nonparametric Bayesian clustering approach, in section four, to replace the regression tree in maximum likelihood linear regression (MLLR) algorithm. Another application, which discussed in section five, is speech segmentation and automatic sub-word discovery and lexicon building. For this application we propose algorithms based on nonparametric Bayesian modeling to automatically segment and cluster speech signals and consecutively discover sub-word units and corresponding dictionary that maps words into these units.

In section six, we propose a new type of HDP-HMM along with its inference algorithm to model topologically restricted left-right models. This model will be used to model sub-word units and therefore to replace HMMs in a conventional speech recognizer. In section seven, we turn our attention to define a nonparametric Bayesian framework for acoustic modeling and training. Particularly, we introduce a recipe to consistently use left-right HDP-HMM to model sub-word units for a continues speech recognizer. Moreover, a nonparametric Bayesian approach will be proposed to replace the phonetic tree used in state of the art speech recognizers to tie triphone states.

Overall, our contribution will be fourth fold: 1-Studying different applications of nonparametric Bayesian approaches in acoustic modeling. 2- A new approach to sub-word modeling based on nonparametric Bayesian modeling 3-Introduning a new type of HDP-HMM (left-right HDP-HMM) model and its inference algorithm. 4- A nonparametric Bayesian recipe and algorithm to train acoustic models for a state of the art speech recognizer.

Contents

[1. Introduction 1](#_Toc347164364)

[2. Nonparametric Bayesian 3](#_Toc347164365)

[2.1. Dirichlet Process 3](#_Toc347164366)

[2.2. Hierarchical Dirichlet Process 4](#_Toc347164367)

[2.2.1. Stick-Breaking Construction 4](#_Toc347164368)

[2.2.2. Chinese Restaurant Franchise 5](#_Toc347164369)

[2.3. HDP-HMM 7](#_Toc347164370)

[2.3.1. CRF with Loyal Customers 8](#_Toc347164371)

[2.4. Inference Algorithm 8](#_Toc347164372)

[2.4.1. Direct Sampler 8](#_Toc347164373)

[2.4.2. Block Sampler 11](#_Toc347164375)

[2.4.3. Learning Hyper-parameters 14](#_Toc347164376)

[3. Acoustic Modeling 17](#_Toc347164377)

[3.1. Acoustic Modeling in sate of the Art Automatic Speech Recognizers 17](#_Toc347164378)

[4. Speaker Adaption 18](#_Toc347164379)

[4.1. Problem statement 18](#_Toc347164380)

[4.2. Related Works 19](#_Toc347164381)

[4.3. Proposed Approach 19](#_Toc347164382)

[5. Speech Segmentation and Acoustical Unit Learning 20](#_Toc347164383)

[5.1. Problem statement 20](#_Toc347164384)

[5.2. Related Works 21](#_Toc347164385)

[5.3. Proposed Approach 21](#_Toc347164386)

[6. Left-to-right HDP-HMM models 23](#_Toc347164387)

[6.1. Problem Statement 23](#_Toc347164388)

[6.2. Related works 23](#_Toc347164389)

[6.3. Proposed Approach 24](#_Toc347164390)

[7. Nonparametric Bayesian training 24](#_Toc347164391)

[7.1. Problem statement 24](#_Toc347164392)

[7.2. Proposed approach 26](#_Toc347164393)

[7.2.1. Training left-right HDP-HMM 26](#_Toc347164394)

[7.2.2. Tying states 26](#_Toc347164395)

[8. Research Plan 27](#_Toc347164396)

[9. Conclusion 27](#_Toc347164397)

[10. Reference 27](#_Toc347164398)

# Introduction

Balancing unique behaviors such as a speaker’s accent with generalized behavior such as an expected formant location that is tied to a phoneme’s identity, is one of the most challenging aspects of speech processing. In applications such as speech recognition, the number of modalities is large and the space of potential solutions vast. For example, varying the number of states in a hidden Markov model often tends to smear information across states rather than allow states to retain an identity modeling a specific phonetic event. Similarly, clustering of formants using Gaussian mixture models often results in clusters that are averaged across unrelated individual events. Such problems can be mitigated using technologies such as phonetic decision trees, but this often results in intricate and elaborate training processes (Harati et al., 2012).

Generally, determining model complexity is among the most difficult problems in pattern recognition. An oversimplified model cannot describe the data and a very complex model generally is prone to over-fitting. Model selection techniques usually need a huge amount of data and are computationally expensive (Bishop, 2007). Any selection methodology needs a criterion for selecting a preferred model. There is not a widely-accepted consensus on this criterion (Ghahramani, 2010). Hence, this process is application specific and involves searching through a discrete space (e.g., a combinational search over models). The final result is sensitive to the criterion used to guide the search.

Nonparametric Bayesian methods provide a mathematically elegant framework that allows inference of model structure and complexity without diluting the purity of modes or clusters (Sudderth 2006). In a fully Bayesian framework, hyper-parameters along with model parameters can be learned automatically from the data. In other words, the data can speak for itself. Unlike in a model selection problem, the optimization of the model parameters is a continuous optimization problem and hence is more tractable.

Hierarchical modeling can be used to increase the power of nonparametric Bayesian models (Teh, et al., 2006). First, hierarchical modeling provides better control over the large number of degree of freedom that exists in nonparametric models (Teh & Jordan, 2010). Second, it makes it possible to use simple building blocks (e.g., a Dirichlet process) to construct models that have rich probabilistic structures (Teh & Jordan, 2010).

In speech recognition, like other pattern recognition applications, selection of the appropriate model complexity and the optimal hyper-parameters are among the most difficult and time-consuming parts of the process, and has a direct effect on performance of the system. Model complexity is not just confined to the complexity of an individual HMM or mixture component but it also includes the overall complexity of the system. A typical state of the art speech recognition system has a large number of degrees of freedom, often utilizing over 10M parameters that must be estimated during training. These parameters must be estimated using a complicated bootstrapping process. A major goal of this paper is to propose a formalization of this process in which a nonparametric extension is constructed within a hierarchical framework.

Among many possible hierarchical Bayesian nonparametric models, in this paper we only consider the hierarchical Dirichlet process (HDP) (Teh, et al., 2006). The motivation for defining an HDP can be understood better by considering the problem of modeling related grouped data. In this problem we are interested in modeling several groups of related data using mixture models. In a traditional nonparametric Bayesian solution we can use a Dirichlet process (DP) prior for each group. This solution can indeed solve the problem by modeling each group using a mixture model, but the resulting mixtures are not linked.

In many applications, for a variety of reasons to be explained later, we want to share components among groups. For example, in topic modeling application, each document can be thought as a group (Teh, et al., 2004). Moreover, under an exchangeability assumption (e.g. bag of words) (Teh & Jordan, 2010), we can model each document as a probability distribution across topics (Teh, et al., 2004). In this case, each topic is a probability distribution across words. It should be noted that a document can have several topics with different strength. Because the number of topics is unbounded the problem fits within the nonparametric framework. Specifically, it is an example of a Dirichlet process mixture (DPM) model. However, if we want different documents to share topics then we have to define another layer that links these individual DPMs together. In other words, there should be a common pool that contains all possible topics (unbounded); each document will be generated by first selecting topics from this common pool randomly and then generating words according to the topic specific distributions. The details of this model will be discussed in following sections.

Hidden Markov models are a time series generalization of a mixture model. As stated above, a DPM can also be considered as a nonparametric extension of a mixture model. Therefore, we expect to have a similar structure for nonparametric HMMs. An analogous structure exists, but it is based on hierarchical Dirichlet process (Teh, et al., 2006) and therefore is called HDP-HMM. Details of this definition will be elaborated in subsequent sections of this paper. However, to understand the motivation behind this definition we can imagine a segmentation problem where the number of segments is not known a priori and each segment can be represented by one state of an HMM.

In this paper, we propose several applications of nonparametric Bayesian approach for acoustic modeling problem. In the second part of this paper, nonparametric Bayesian methods used in the subsequent sections will briefly be introduced. In section three we discuss and introduce the acoustic modeling problem. After these introductory sections, we will focus on four primary applications of nonparametric Bayesian methods that are the subject of this proposal. It should be noted that problem are ordered from easiest to the most difficult.

In section four, an application of nonparametric Bayesian clustering to speaker adaption problem will be proposed. This problem is perhaps the easiest application in this paper and was intended to study the feasibility of the framework in speech recognition problems. The proposed method replaces the regression tree in maximum likelihood linear regression (MLLR) algorithm.

In section five, we study the segmentation problem. Segmentation is among the most fundamental problems in speech and signal processing. In this section, an approach for automatically segmenting speech utterances will be proposed. Despite of its importance, segmentation by itself has little practical importance. Hence, in this section we also propose an approach to apply nonparametric Bayesian approach to segment and cluster speech utterances in order to automatically discover acoustic sub-word units that could replace more traditionally used units like phonemes and finally we propose a method to generate a lexicon to map words into these sub-word units.

In section six, we turn our attention into the very important problem of nonparametric Bayesian modeling of individual sub-word units. This problem traditionally tackled using left-right HMMs with fixed number of states and with predetermined number of Gaussians per state in state of the art speech recognizers. In this section we propose a new topologically constraint HDP-HMM, which we call left-right HDP-HMM, and its corresponding inference algorithm to solve the mentioned problem within the nonparametric Bayesian framework. The proposed model will learn both the number of states and number of mixtures automatically from the data.

Finally in section seven, we present an approach for training a complete speech recognizer within the nonparametric Bayesian framework. This approach, as will be discussed later, will use the left-right HDP-HMMs to model each individual sub-word unit. Moreover, it can be used to train continues speech recognizers using available speech corpus and using only utterance level transcriptions. We also introduce a data driven nonparametric Bayesian approach to replace phonetic trees for state tying.

# Nonparametric Bayesian

## Dirichlet Process

A Dirichlet process (DP) is a distribution over distributions, or more precisely over discrete distributions. Formally, a Dirichlet processis “defined to be the distribution of a random probability measureoversuch that for any finite measurable partitionofthe random distribution is distributed as finite dimensional Dirichlet distribution” (The et al., 2006):



A constructive definition for Dirichlet process is given by Sethuraman (Sethuraman, 1994) which is known as stick-breaking construction. This construction explicitly shows that draws from a DP are discrete with probability one.



can be interpreted as a random probability measure over positive integers and is denoted by. In both of these definitions, or base distribution, is the mean of the DP, andis the concentration parameter which can be understood as the inverse of variance.

Another way to look at the DP is through the Polya urn scheme. In this approach, we have to consider i.i.d. draws from a DP and consider the predictive distribution over these draws (Teh et al., 2006):



In the urn interpretation of equation , we have an urn with several balls of different colors in it. We draw a ball and put it back in the urn and add another ball of the same color to the urn. With probability proportional towe draw a ball with a new color. To make the clustering property more clear, we should introduce a new set of variables that represent distinct values of the atoms. Letto be the distinct values andbe the number of associated with. We would now have:



Another useful interpretation of is the Chinese restaurant process (CRF). In CRF we have a Chinese restaurant with infinite number of tables. A new customer comes into the restaurant and can either sit around one of the occupied tables with probability proportional to the number of people already sitting there or initiate a new table with probability proportional to. In this metaphor, each customer is a data point and each table is a cluster.

## Hierarchical Dirichlet Process

A Hierarchical Dirichlet Process (HDP) is the natural extension of a Dirichlet process for problems with multiple groups of data. Usually, data is split into groups a priori. For example, consider a collection of documents. If words are considered as data points, each document would be a group. We want to model data inside a group using a mixture model. However, we are also interested to tie groups to each other, i.e. to share clusters across all groups. Let’s assume that we have an indexed collection of DPs with a common base distribution. Unfortunately this simple model cannot solve the problem since for continues  different  necessary have no atoms in common. The solution is to use a discrete  with broad support. In other words,  is itself a draw from a Dirichlet process. HDP is defined by (Teh & Jordan, 2010) equation .



In this definition provides prior distribution for factor.  governs the variability of  around andcontrols the variability of around . , and are hyper-parameters of HDP. Definition shows the first representation of HDP. Another representation can be obtained by introducing an indicator variable as shown in equation .

Figure 1 shows the graphical models of both of these representations.



### Stick-Breaking Construction

Because is a Dirichlet distribution it has a stick-breaking representation:



Where  and. Since support of is contained in within the support of  we can write a similar equation to for:



Then we have:





### Chinese Restaurant Franchise

The Chinese restaurant franchise (CRF) is the natural extension of Chinese restaurant process for HDPs. In CRF, we have a franchise with several restaurants and a franchise wide menu. The first customer in restaurant *j* sits at one of the tables and orders an item from the menu. Other customers either sit at one of the occupied tables and eat the food served at that table or sit at a new table and order their own food from the menu. Moreover, the probability of sitting at a table is proportional to the number of customers already seated at that table. In this metaphor, restaurants correspond to groups and customerin restaurant** corresponds to (customers are distributed according to). Tables are i.i.d. variables distributed according toand finally foods are i.i.d. variables distributed according to. If customerat restaurantsits at tableand that table serves dish, we will have. In another way, each restaurant represents a simple DP and therefore a cluster over data points. At the franchise level we have another DP but this time clustering is over tables.



Figure 1-(a) HDP representation of (b) Alternative indicator variable representation (The et al., 2004)

Now let introduce several variables that will be used throughout this paper. is the number of customers in restaurant , seated around table,and who eat dish.is the number of tables in restaurant serving dish  and is the number of unique dishes served in the entire franchise. Marginal counts are denoted with dots. For example,is the number of customers in restauranteating dish.

CRF can be characterized by its state which consists of the dish labels, the tables  and dishes . As a function of the state of the CRF, we also have the number of customers, the number of tables, customer labels and table labels (Teh & Jordan, 2010). The posterior distribution ofis given by:



Where is the total number of tables in the franchise andis the total number of tables serving dish. Equation shows the posterior for.is the total number of customers in restaurant andis the total number of customers in restauranteating dish.



Conditional distributions can be obtained by integrating outandrespectively. By integrating outfrom we obtain:



And by integrating outfrom we obtain:



A draw from can be obtained using and a draw from can be obtained using .





From and we see that the posterior of is a mixture of atoms corresponding to dishes and an independent draw from andis a mixture of atoms at and an independent draw from (Teh & Jordan, 2010).

## HDP-HMM

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain (Rabiner, 1989). In the following discussion we will denote the state of the Markov chain at time  with  and the state-specific transition distribution for stateby.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 infinite. 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 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 an HDP. In an HDP-HMM each state corresponds to a group (restaurant) 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 (Teh et al., 2006). This problem is solved by introducing a sticky parameter to the definition of HDP-HMM (Fox et al., 2011). 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 (Fox et al., 2011)

### CRF with Loyal Customers

The metaphor for the Chinese restaurant franchise for sticky HDP-HMM is a franchise with loyal customers. In this case each restaurant has a special dish which is also served in other restaurants. If a customer is going to restaurant then it is more likely that he eats the specialty dish there. His children also go to the same restaurant and eat the same dish. However, if eats another dish () then his children go to the restaurant indexed byand more likely eat their specialty dish. Thus customers are actually loyal to dishes and tend to go to restaurants where their favorite dish is the specialty.

## Inference Algorithm

### Direct Sampler

This sampler is adapted from (Fox et al, 2011) and (Fox et al, 2010). In this section we present the sampler for HDP-HMM with DP emission.. The algorithm is divided into two steps: the first step is to sample the augmented stateand the second is to sample.In order to sample  we need to have the posterior. By inspecting Figure 2 and using the chain rule we can write the following relationship for this posterior:



The reason that we have summed over in the last line is because we are interested to calculate the likelihood for each state. This equation also tells us that we should first sample the state and then conditioned on the current state, sample the mixture component for that state. For Gaussian emissions we can write (ref):







The algorithm is as follows:

1. Given a previous set of and
2. For all.
3. For each of thecurrently instantiated states compute:

* The predictive conditional distributions for each of the  currently instantiated mixture components for this state, and also for a new component and for a new state.





* The predictive conditional distribution of the HDP-HMM state without knowledge of the current mixture component.



1. Sample:



1. Sample conditioned on:



1. If increase theand transform as



1. Ifincrement.
2. Update the cache. If there is a state withor removeand decrease. If remove the componentand decrease.
3. Sample auxiliary variables by simulating a CRF:
4. For eachsetand. For each customer in restauranteating dish(), sample:



1. Incrementand if increment.
2. For each,sample the override variables in restaurant:



1. Set the number of informative tables in restaurant:



1. Sample:



1. Optionally sample hyper-parametersand.

### Block Sampler

The problem with the direct assignment sampler mentioned in the previous section is the slow convergence rate since we sample states sequentially. The sampler can also group two temporal sets of observations related to one underlying state into two separate states. However, in the last sampling scheme we have not used the Markovian structure to improve the performance. In this section a variant of forward-backward procedure is incorporated in the sampling algorithm that enables us to sample the state sequenceat once. To achieve this goal, a fixed truncation level should be accepted which in a sense reduces the model into a parametric model (Fox et al, 2011). 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 possible states (Fox el al, 2011). In short, we obtain an approximation to the nonparametric Bayesian HDP-HMM with maximum number of possible states set to . For almost all applications this should not cause any problem if we set  reasonably high. The approximation used in this algorithm is the degree  weak limit approximation to the DP (Ishwaran & Zarepour, 2002) which is defined as:



Using is approximated as (Fox et al, 2010):



We can write:



And posteriors are :



In is the number of transitions from state to stateand is the same as .

Finally an orderweak limit approximation is used for the DP prior on the emission parameters:



The forward-backward algorithm for the joint sample  andgiven can be obtained by:



The right side of equation has two parts: forward and backward probabilities (Rabiner, 1989).The forward probability includes  and backward probability includes. The forward probabilities approximated with, therefore for backward probabilities we have:



As a result we would have (Fox et al, 2010) :



where for Gaussian emission for components are given by 

The algorithm is as follows (Fox et al, 2010):

1. Given the previous and.
2. For, initialize messages to 
3. Forand compute



1. Sample the augmented state sequentially and start from:

Set andforand

For all compute:



1. Sample augmented state:



1. Increase andand add to the cached statistics.



1. Sample  similar to the previous algorithm
2. Update :



1. For :

* Sample and:



* For  sample:



1. Set and
2. Optionally sample hyper-parametersand.

### Learning Hyper-parameters

Hyper-parameters includingandcan also be inferred like other parameters of the model (Fox et al. , 2010).

#### Posterior for

Consider the probability of data to sit behind table:



This equation can be written by considering equation and . From this equation we can say customer table assignment follows a DP with concentration parameter. Antoniak (Antoniak, 1974) has shown that if  then the distribution of the number of unique values of  resulting from draws from has the following form:



Where is the Stirling number of the first kind. Using these two equations the distribution of the number of tables in the restaurantis as follows:



The posterior overis as follows:



The reason for the last line is that is not a function of and therefore can be ignored.

By substitution of  and also by considering that  we obtain:



Finally by considering the fact that we have placed a prior on we can write:

   
Wherecan be either one or zero. For marginal probabilities we obtain:







#### Posterior of

Similar to the discussion for if we want to find the distribution of the unique number of dishes served in the whole franchise we would have. Therefore for the posterior distribution of we can write:



By considering the fact that that prior overiswe can finally write:



And finally for the marginal distributions we have:







#### Posterior of

The posterior foris obtained in a similar way to. We use two auxiliary variablesand and the final marginalized distributions are:







It should be noted that in cases where we use auxiliary variables we prefer to iterate several times before moving to the next iteration of the main algorithm.

#### Posterior of

By definition  and by considering the fact that the prior on is and we can write:



# Acoustic Modeling

Generally speaking, the goal of a speech recognizer is to map the acoustic data into word sequence. This problem can be formulated, simplistically, with :



In this formulation, is the probability of a particular word sequence given acoustical observations, and the goal is to find a sequence W that maximizes this probability.  is the language model and indicates what is the prior probability of words.  is the probability of the observed acoustic data and usually can be ignored and finally  is the acoustic model. Therefore generally we can divide the problem into two separate sub-problems and solve each one independently. Our focus in this research will be the acoustical modeling problem.

## Acoustic Modeling in sate of the Art Automatic Speech Recognizers

In this section, we review the approach that is used to tackle the acoustic modeling problem in most state of the art Automatic Speech Recognizers (ASR).

The basic idea for acoustic modeling is to find a mapping between word sequences and acoustic observations. In early systems (Furui, 1986), each word modeled separately. This approach is relatively simple and works satisfactory for small vocabulary and isolated speech recognition tasks, however, it is not scalable to continues large vocabulary tasks. The problem is related to the selected acoustic units (i.e. words). Since the number of words in a typical language is very large and increases over the time, modeling all words independently is not feasible. An alternative approach is to break down words into some finite set of units common to all possible words and then just model these units. People used different units such as phonemes (Lee, 1990), syllables (Ganapathiraju et al., 2001) and acoustically inspired units (Paliwal, 1990). Phonemes are the most popular and easy to use units and most successful commercial systems are based on them.

After selecting type of the units (e.g. phonemes) we have to select a statistical (or generally a machine learning) model for these units. Given a set of trained models and some new observations we test all models against the observations and select the model with the highest score (e.g. likelihood). The most successful models used in state of the art ASRs are left-to-right or Bakis hidden Markov models (HMMs) with Gaussian or mixture of Gaussians emissions (Rabiner, 1989). An HMM is a generalization of mixture model where latent variables are not independent of each other and are related with a Markov chain. This makes them particularly attractive to model sequential observations. Most systems use a simple HMM with some predetermined number of states (e.g. 3) for all units and also with some predetermined number of mixture components per state.

State of the art speech recognizers usually use some form of context dependent unit instead of simple context independent units. For example, phoneme based systems usually has 42 context independent phonemes but in order to improve the quality of models we can incorporate the left and right context and define context dependent units (i.e. triphones). However, the number of units grows exponentially with increasing the depth of the context. For example, number of triphones is 42\*42\*42=74088. This means training context dependent models face a serious data sparsity problem. In any practical situation, many models will never have any observation and many more will have just a few examples and therefore estimated parameters will have large variances. In fact the resulted system will perform worse than a context independent system for a given amount of training data. This problem has been tackled by tying models and states together so similar models share data which is a trade of between model accuracy and amount of data. The most successful approach to tie states is based on a phonetic decision tree which is a binary tree with phonetic questions attached to its nodes (Young et al., 2006). The tying is happening between corresponding states of all triphones with the same central state. For each state of a phoneme a tree grown from a single node that contains all the corresponding states of all triphones for that phoneme. The tree grown by asking phonetic questions and stop when the number of data points in a node reaches to a minimum amount or dividing a node does not increase the likelihood significantly. After this step, we will have enough data for all states of all triphones.

Therefore a general recipe to train acoustic models in a contemporary ASR is as follow:

* The first step is to prepare the data. We need to obtain some transcribed speech utterances and convert them into appropriate features representation (e.g. Mel-frequency cepstral coefficients –MFCC). We also need a dictionary that contains all possible words and their corresponding sub words (e.g. phonemes) decomposition.
* The next step is train all context independent phonetic models using the transcribed data and using EM algorithm. This step is usually performed using the self-organizing property of HMMs. In other words, we let HMMs to segment data into different models and states.
* After training good monophone models, the next step is to clone monophones into triphones by simply copy the emission distribution and transition matrix for all triphones with same central state and then train them using the available data.
* The fourth step is to tie states (as mentioned above) and train the resulted models for several more iteration using EM algorithm.

In this research, our goal is to investigate the applications of nonparametric Bayesian methods which discussed in previous section in acoustic modeling. In a typical speech recognizer which described above, there are several tasks (specially clustering, segmentation and model topology) that can be viewed as potential candidates for nonparametric Bayesian modeling. In following sections we describe our proposed approaches for each of these tasks.

# Speaker Adaption

## Problem statement

In typical state of the art speech recognition systems a single model is developed using data for all speakers to cover the variance across dialects, speaking styles and speaker specific features. At the other hand, it is a well-known fact that a speaker dependent system trained on sufficiently large data set always outperforms the speaker independent system (Ishii et al.,1996). However, since it is not practical to train a separate speech recognizer for each speaker a middle ground solution looks attractive. Speaker adaption algorithms are designed to adapt speaker independent models to a specific speaker. There are many different approaches proposed for speaker adaption. However, in this section we just review Maximum Likelihood Linear Regression (MLLR) algorithm and introduce a nonparametric Bayesian extension of this algorithm.

## Related Works

Basic idea for MLLR is to transfer the speaker independent model’s parameter (mean and variance of Gaussian mixtures in HMM models) to a new set of speaker dependent set (Gales, 1996). However since the amount of adaption data is limited several tradeoff should be adapted. One such compromise is to just map “mean” parameter and another is to cluster several models together and transform them using the same data (Gales, 1996).It is also desirable to make these clustering dynamic therefore when more data became available we use a more fine-grain transformation. Toward this goal the clustering is achieved using a regression tree where the root contains all models and at each level tree divides into two sets.

## Proposed Approach

As mentioned, MLLR consists of two different stages. The first stage is clustering and the second stage is to compute transformation matrices for each cluster. 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 research, we propose the use of DPM as an alternate clustering algorithm to investigate the potential advantages of this approach. The procedure we employ is as follows:

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.

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 regular speech recognizers. 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. Figure 3compares the result of proposed approach with regression tree for monophone models. From this figure we can say the proposed methods works better than an ordinary tree.

****

Figure 3- A comparison of regression tree and DPM based clustering (Harati et al., 2012). Inference implemented using ADVP algorithm.

# Speech Segmentation and Acoustical Unit Learning

## Problem statement

Speech segmentation is perhaps the most fundamental process in speech recognition. In fact the popularity of HMMs is a result of their segmentation properties. By segmentation, we mean to find boundaries between different events in speech. If the event is a word then speech segmentation is become equal to speech recognition. If the event is a phoneme then the segmentation is equal to phoneme recognition and if the event is a speaker then the segmentation is equal to speaker recognition and diarization (Fox et al., 2011). Of course direct segmentation is not usually used for any of these tasks because defining the event is not a trivial task. For example, there are no criteria that show a portion of speech is corresponding to a word and so implementing a speech recognizer using just a direct segmentation is not practical. As a result most application of direct speech segmentation (which implicitly used in all speech recognizers via HMM or similar technology) is for unsupervised tasks. For example, finding similar segments in a given speech utterance which can be hypothesized to be a particular word of interest (word spotting and spoken term detection) (Lee & Glass, 2012), or to separate speech from non-speech.

Another interesting application of speech segmentation is for acoustic unit discovery. 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 (Harper, 2011) 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 (Bacchiani & Ostendorf, 1999) 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.

Both of the clustering and segmentation sub-problems are good candidates for nonparametric Bayesian modeling. In the following we discussed the related works and our proposed approach.

## Related Works

As mentioned in the last section, classical methods for acoustic unit discovery involve segmentation and clustering. The segmentation is implemented using dynamic programming method that incorporates a heuristic stopping criterion (Bacchiani & Ostendorf, 1999) and clustering implemented using a heuristic agglomerative method. Recently, Lee & Glass (2012) proposed a nonparametric Bayesian approach for unsupervised segmentation of 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.

Another related problem is speaker diarization. In this problem, the goal is to segment the speech utterance into speaker homogeneous area. Fox et al. (2011) have used HDP-HMM model to solve this problem by modeling each speaker as a single state. It has been shown that the results are comparable to the state of the art speaker diarization systems (Fox et al., 2011).

## Proposed Approach

Our approach for speech segmentation is based on HDP-HMM model. We propose to segment the speech using an ergodic HMM. In this model, each state models an acoustic unit. Figure4 demonstrates an example on some primary experiments based on this model (Harati et al., 2013). From this figure we can see the discovered boundaries approximately coincide with phoneme boundaries. Table1 compares the performance of the proposed algorithm with some other state of the art algorithms. 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:





Figure 4-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

From this table we can see 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%.

Although theoretically HDP-HMM should label segments to their corresponding clusters automatically, our initial results show this labeling is not reliable and so we need to perform another clustering stage. We propose to investigate different clustering methods including a nonparametric Bayesian approach (e.g. DPM) for this step.

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) | 75.2 | 66.8 | 70.8 |
| Qiao et al. (2008) | 77.5 | 76.3 | 76.9 |
| Lee & Glass (2012) | 76.2 | 76.4 | 76.3 |
| Proposed Approach | **86.5** | 68.5 | 76.6 |

Automatically discovered units are not very useful unless we can define a dictionary that maps words into the units. Therefore the next step is to align the transcription with the discovered segments (and hence units) and generate a lexicon. We are planning to use forced alignment or using the manually transcribed data to map words into acoustic units.

The performance of the system will be measured into two different ways:

1. To measure the unit classification error. This will show how units modeled using our approach perform without considering errors that can introduced in lexicon generation step.
2. To measure word error rate (WER) for a system trained completely using our proposed units. This method of measuring performance is more interesting from a practical point of view; however, the performance will be a function of lexicon building stage too.

By looking at different steps of acoustic model training from the last section we see, in this section, a nonparametric Bayesian approach is suggested for the first step (generating sub-word units and lexicon) of the general training recipe.

# Left-to-right HDP-HMM models

## **Problem Statement**

Perhaps the most important element of acoustic modeling is the statistical approach used to model sub-word units. Most state of the art ASR systems use left-to-right HMMs with Gaussian mixtures emissions to model phonetic units (Rabiner, 1989). Usually, the number of sates is fixed for all models (for example to 3) and mixtures trained progressively by starting from one mixture per state and increase the number of mixtures until further increment does not improve the likelihood of the training data. Number of mixtures per state is also a fixed parameter for all states and models.

Because of the simplicity and existence of efficient algorithms, this parametric HMM models have been used extensively in many different applications, however, it is evident that setting the parameters (number of states and number of mixtures per state) and even topology of the model a priori is heuristic and experimental. Moreover, all models usually have the similar structure which is not the optimum choice.

During the past few decades, there were several attempts to address some of these problems; however, because of the lack of better models and also limited computation power these efforts were not very successful. In the next section, some of these works will be reviewed.

## Related works

As discussed above, HMMs parameterized both in their topology (e.g. number of states) and emission distributions. Most attempts to relax these parameterizations were focused on the second aspect. Bourlard (1993) and others proposed to replace Multilayer Perceptron (MLP) with Gaussian Mixture Models (GMMs). It has proven (Bourlard & Morgan, 1993) when used for classification; MLPs generate reasonable estimates of posterior distribution of the output classes conditioned on the input patterns. These hybrid HMM-MLP systems works slightly better than traditional HMM-GMMs , but the gain is not practically significant to justify moving from well-established technology to a new one with unknown problems. Moreover, most of the gain fades by using more complicated systems (e.g. speaker adaption, post processing). Another example of this approach is reported in (Lefèvre, 2003) and (Shang, 2009) where nonparametric density estimators have been used to replace the GMMs. Again the improvements are marginal at best. All of these approaches can be classified as nonparametric non-Bayesian methods. Being non-Bayesian makes them especially prone to overfitting or over-smoothing.

Henter et al (2012) introduce a new model named Gaussian Process dynamical model (GPDM) to completely replace HMMs in acoustic modeling. The new model is nonparametric Bayesian and is based on Gaussian process and supposedly solves some of the problems traditionally associated with hidden Markov models such as duration modeling and stepwise constant evolution (Henter et al., 2012). However, this model is used only in speech synthesis and there is no result reported for speech recognition problems using this model. One of the possible, issues with this model for speech recognition is the lack of any kind recognition algorithm that could use new model at this moment. (Since the recognition algorithms are developed for HMM structure).

## Proposed Approach

We have introduced the nonparametric Bayesian counterpart of HMMs, HDP-HMMs, previously. Therefore one natural way to extend nonparametric methods in acoustic modeling is to replace HMMs with HDP-HMMs. However, HDP-HMM is fully ergodic model (all states are connected to each other) while in speech application we usually need a more constraint topology to model a time sequence. Especially, left-right topology is proved to be useful in speech recognition and similar applications. We are proposing a new type of HDP-HMM that is restricted in this sense. Therefore, the model is still learning its structure (number of states and possible skip transitions) while it remains within the left-right family of HMMs. There are two approaches to do this, the first approach is to use a regular HDP-HMM and then convert it into a left-right structure and the second one is to directly define a left-right HDP-HMM. Here we propose to develop the second approach while comparing the result with the first approach. Therefore developing a left-right HDP-HMM and its inference algorithm is one of the proposed contributions of this research.

One of the intrinsic differences between ergodic HMMs and left-right HMMs is that the former model just one sequence of events. These events can happen in different orders but if we have two separate sequence we have to model them separately. Left-right HMMs, on the other hand, model ordered sequence of events with a start and an end. Therefore a single HMM can model several sequences. This also opens the door into two interesting issues: First, the state’s labels will not be arbitrarily and therefore there is no label switching problem. Secondly, as a consequence it looks like it is possible (though perhaps with some heuristics) to use a straight forward parallel inference (training) strategy. Investigating, this possibility is another contribution of this research.

Overall, the proposed model will non-parametrically estimate the number of states and also number of mixtures per state. Since each state will have a different number of components for Gaussian mixture which determined directly from the data it is expected to estimated distribution be very close to the true distribution for that state. It should also be noted that unlike HMM-MLP most of the new complexity of our model is added to the training part and recognition part essentially remains the same.

For this section we will test our proposed model against regular HMM models for a phoneme recognition or isolated word recognition application. The reason to choose these applications is to focus our investigation on modeling capabilities of the proposed model. A general speech recognizer consists of many different parts (e.g. Viterbi decoder) that can alter the results. We will discuss about testing the performance of our model in continues speech recognizer in the next section.

# Nonparametric Bayesian training

## Problem statement

In pervious sections, we have discussed about the general recipe for training acoustic units in a state of the art speech recognizer. In Section 6 we introduced a left-right HDP-HMM model to replace ordinary HMMs in speech recognizer. In this section we will introduce a recipe to train these new models in a more general nonparametric Bayesian framework.

One of the interesting features of standard acoustic model training is the flat start (Young et al., 2006). Flat start means we can initialize HMM models using global calculations (e.g. means and covariance) over the training data and for each speech utterance connect its corresponding phoneme HMMs together and train them as a big HMM. This method makes it possible to avoid using phoneme level transcription (which are very difficult to produce) for training acoustic model training. Therefore we want the training procedure for nonparametric Bayesian model also has this convenient property.

Acoustic units usually trained in progressive steps, starting from very simple models and gradually training more and more complex ones. Broadly speaking the training procedure is as follow:

1. Boot strap and flat start: This step defines the basic models and initializes them.
2. Training monophones: This step trains monophone models.
3. Defining triphones and tie states: This step makes a much more complex model starting from simpler models (tying will be discussed in the following paragraph.)
4. Train tied state triphones.
5. Optionally use adaption techniques to adapt speaker independent models into speaker dependent models.

One of the important challenges in training more complex systems is the data sacristy problem. Context dependent models like triphones can model acoustic events more accurately (relative to systems using context independent models.). However, each model has less data and so estimating the parameters correctly become a serious problem. Moreover, some of the triphones will never be observed in a given training dataset. To deal with these problems, people suggest tying either model or components of the models together. Tying similar models seems a good idea but it turns out that tying states is much more effective (Beulen at al., 1997). There were two main stream approaches to tie states:

The first approach is the data driven approach:

1. First a list of all triphones existed in the dataset is produced.
2. Using monophone models trained in pervious steps, these triphones models initialized by cloning from monophone models.
3. After training these triphone models, corresponding states of all triphones with similar center phoneme grouped.
4. For each group, a clustering algorithm is applied. The clustering algorithm has two steps. First cluster similar states (based on Euclidian distance) and then merge clusters with few data points to closest cluster.
5. Train tied models.
6. For triphones not existed in the data use back-off modeling (back off to diphones or monophones.)

Alternatively, we can use phonetic trees to cluster the data. In this case, we first group all corresponding states of all triphones with similar center phoneme. We also provide a pool of phonetic questions (e.g. is the left phoneme a stop? ). The clustering is as follow:

1. Put all states in the root node of the tree.
2. Find the best question that divide the node into two nodes and maximize the local likelihood scores.
3. Keep doing step two for all nodes until increments in the likelihood fall below a threshold. The resulted nodes are called terminal nodes and all states with in a terminal node will tied together.
4. If data points in a node is less than a threshold combine it with its parent node.
5. Unseen models can be clustered by starting from the root and answering the question until we get to a terminal node.

Both of these approaches have been used successfully in state of the art speech recognition systems. Particularly phonetic tree based approach due to its simplicity and effectiveness has become a very successful and popular technology.

## Proposed approach

In this section we describe the procedure to train left-right HDP-HMM models introduced in the previous section. Moreover, we describe a procedure within the nonparametric Bayesian paradigm to tie states. But it should be noted that training HDP-HMM models and tying are two separate problems and so can be used independently. The training algorithm is independent of the sub-word unit used for speech recognition; therefore, in the following we will restrict our discussion to a phonetic based system. However, using other units (including acoustic units) is the same.

### Training left-right HDP-HMM

As discussed before, it is very important to have a training procedure that let us train our models without having phonetic level transcriptions. To this end, we introduce a variable Zi that contains the model id for each data point Xi . For a given speech utterance, the algorithm is as follows:

1. Initialize Zi either randomly or boot strap using a conventional system.
2. The result is several sub-sequences. Each sub-sequence will have a unique Zi. Therefore a sequence of Xi will converted into a sequence of sub-sequences Wj.
3. For a given sequence of data use the transcription to generate a list of models.
4. Regroup sub-sequences Wj based on their corresponding Zj and distribute each group to the corresponding HDP-HMM model (MZi).
5. Train each HDP-HMM using the inference algorithm. Training each left-right HDP-HMM involves several sequences of data . Fortunately, since each left-right HDP-HMM has a start state (the left most one) using multiple sequence in inference algorithm does not change the algorithm too much. For each new sequence we just need to start from the left most state (by forcing the first data point to belong to that state).
6. After all Models trained, we should re-estimate the Zi  for all Xi . This can be done using Viterbi algorithm or in a Bayesian framework.
7. After several iterations and after convergence we can fix the topology of each model.

### Tying states

After training context independent models, we can use phonetic trees to cluster states of the trained models and tie them together. Alternatively, we can use a nonparametric Bayesian approach which is closely related to data driven approach described previously. Here we describe the proposed algorithm:

1. Given the monophone models, train all existed triphones in the data set and also segment the data into different states.
2. Group all corresponding states of all triphones with the same central phoneme.
3. Each of these groups will contain all the data associated with states inside the group.
4. In each group use Dirichlet Process Mixture (DPM) to cluster the data. It is also possible to use a Hierarchical Dirichlet Process (HDP) across different groups.
5. Merge small clusters into closest cluster.
6. Use back-off modeling for unseen triphones.

# Research Plan

***Feb 1- March 30:***

1. Implementing left-right HDP-HMMs and the corresponding inference algorithm.
2. Use HDP-HMMs to segment speech data from TIMIT.

***April 1-April 30:***

1. Experiments using left-right HDP-HMMs and compare to the baseline system.
2. Clustering and automatic unit discovery using segmentations produced from TIMIT dataset.

***May 1-May 31:***

1. Diagnosing possible problems related to left-right HDP-HMM implementation.
2. Generating the lexicon for automatic discovered units and use them in state of the art speech recognizer and compare with baseline system.

***June 1-July 31:***

1. Wrap up the left-right HDP-HMM and its inference algorithm.
2. Diagnose and debugs problems related to the automatic unit discovery and lexicon building.

***August 1- September 30:***

1. Wrap up the speech segmentation and automatic unit discovery.
2. Review our previous works on the speech adaption applications (already done) and wrap it up.
3. Implementing the nonparametric training framework for continues speech recognition (first section.)

***October 1- November 30:***

1. Diagnosing the training framework and run preliminary experiments.
2. Wrap up all other parts of the dissertation.

***December 1-December 30:***

1. Wrap up the first part of the training frame works and implement the second part (state tying).
2. Run experiments related to this section.

***January 1- January 31 :***

1. Wrap up the training frame work.
2. Finalize the draft of the dissertation.

# Conclusion

In this paper, we investigated several applications of nonparametric Bayesian approach in acoustic modeling problem. The applications were sorted from easy to difficult. The first application that we propose to investigate was the speaker adaption problem. For this application, we proposed to use a bottom-up approach based on Dirichlet Process Mixture (DPM) to replace the top-down regression tree of MLLR algorithm. The second application was speech segmentation and automatic sub-word discovery. For this application, we proposed to use nonparametric Bayesian methods for segmentation and clustering and also to generate a lexicon that maps words into discovered units. The third application is to use a nonparametric Bayesian model to model each sub-word unit. In this section we propose a new type of HDP-HMM named left-right HDP-HMM and its corresponding inference algorithm. Finally, we proposed a nonparametric Bayesian framework and training recipe to use left-right HDP-HMMs in a continues speech recognizer application.

Nonparametric Bayesian statistics is one of the new promising approaches in machine learning and data modeling. It brings a good mix of flexibility and being biased toward simpler models (Occam's razor). By considering the exponential trends in data generation and computational power we can see approaches like nonparametric Bayesian are necessary tools to harness this enormous power. In this proposal, we proposed to investigate several applications in acoustic modeling problem, however, there are many directions that can be persuade for the future. One important and practical problem is to use the massive parallel processing powers (both clusters and GPUs) to accelerate the speed of inference algorithms. As of now, the main problem associated with nonparametric Bayesian approaches is their expensive computational cost. Because of this some groups already start to adapt parallel training techniques for the inference algorithm (Williamson et al., 2012) & (Suchard et al., 2010).

Another direction, especially in speech processing and similar applications, is to look into more complicated hierarchical models. Defining new models, under a Bayesian framework, is relatively straightforward. However designing efficient inference algorithm is a challenge. Also using models efficiently and intelligently in various problems might be a more difficult problem than just defining new models. For example, a new component to the proposed approach in this paper is to add another level of hierarchical clustering to cluster the data within a particular model based on acoustic similarities and differences. In such a way, we can train several instance for each model with better accuracy (for example it has been shown that having gender specific models significantly improves the recognition rate; this approach can be considered as a generalization of gender specific modeling) and since we use hierarchical framework we can tie the models and share the data in various ways. Particularly, by considering the vast amount of speech data (e.g. youtube) that became available during the past few years and by considering the huge acoustic diversity existed in this data (e.g. different speakers, environments) the importance of the this suggested direction becomes more clear.

# Reference

Antoniak, C. (1974). Mixtures of Dirichlet Process with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(7), 1152–1174

Bacchiani, M., & Ostendorf, M. (1999). Joint lexicon, acoustic unit inventory and model design. *Speech Communication*, 29(2-4), 99–114.

Beulen, K., Bransch, E., & Ney, H. (1997). state tying for context dependent phoneme models. *proceeding of Fifth European Conference on Speech Communication and Technolog* (pp. 1179–1182). Rhodes, Greece.

Bishop, C. (2011). *Pattern Recognition and Machine Learning* (2nd ed., p. 738). New York, New York, USA: Springer.

Bourlard, H., & Morgan, N. (1993). *Connectionist Speech Recognition A Hybrid Approach*. Springer.

Dusan, S., & Rabiner, L. (2006). On the relation between maximum spectral transition positions and phone boundaries. *Proceedings of INTERSPEECH* (pp. 1317–1320). Pittsburgh, Pennsylvania, USA

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2010). Supplement to “ A Sticky HDP-HMM with Application to Speaker Diarization”. *The Annals of Applied Statistics*, *S*(2A), S1–S32.

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2011). A Sticky HDP-HMM with Application to Speaker Diarization. *The Annalas of Applied Statistics*, 5(2A), 1020–1056.

Furui, S. (1986.). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(1), 52 – 59.

Gales, M. J. F. (1996). *model-based techniques for noise robust speech recognition*. Cambridge University.

Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G., & Picone, J. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(4), 358–366.

Ghahramani, Z. (2010). Bayesian Hidden Markov Models and Extensions. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (pp. 56–56). Uppsala, Sweden.

Harati, A., Picone, J., & Sobel, M. (2012). Applications of Dirichlet Process Mixtures to Speaker Adaptation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4321–4324). Kyoto, Japan.

Harati, A., Picone, J., & Sobel, M. (2013). Speech Segmentation Using Hierarchical Dirichlet Processes. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (p. TBD). Vancouver, Canada.

Harper, M. (2011). IARPA Solicitation IARPA-BAA-11-02. *IARPA BAA*.

Henter, G. E., Frean, M. R., & Kleijn, W. B. (2012). Gaussian process dynamical models for nonparametric speech representation and synthesis. *IEEE International Conference on Acoustics Speech and Signal Processing* (pp. 4505– 4508). Kyoto, Japan.

Ishii, J., Tonomura, M., & Matsunaga, S. (1996). Speaker Adaptation Using Tree Structured Shared-State HMMs. *Fourth International Conference on Spoken Language* (pp. 1149–1152). Philadelphia, Pennsylvania, USA

Ishwaran, H., & Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2), 269–283

Lee, C., & Glass, J. (2012). A Nonparametric Bayesian Approach to Acoustic Model Discovery. *Proceedings of the Association for Computational Linguistics* (pp. 40–49). Jeju, Republic of Korea.

Lee, K.-F. (n.d.). Context-independent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(4), 599– 609

Lefèvre, F. (n.d.). Non-parametric probability estimation for HMM-based automatic speech recognition. *Computer Speech & Language*, 17(2-3), 113–136.

Paliwal, K. (1990). Lexicon-building methods for an acoustic sub-word based speech recognizer. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 729–732). Albuquerque, New Mexico, USA

Qiao, Y., Shimomura, N., & Minematsu, N. (2008). Unsupervised optimal phoneme segmentation: Objectives, algorithms and comparisons. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 3989–3992). Las Vegas, Nevada, USA

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2), 879–893

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 639–650.

Shang, L. (n.d.). Nonparametric Discriminant HMM and Application to Facial Expression Recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2090– 2096). Miami, FL, USA

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., West, M., & Cron, A. (2010). Understanding GPU Programming for Statistical Computation: Studies in Massively Parallel Massive Mixtures. *Journal of Computational and Graphical Statistics*, 19(2), 419–438.

Sudderth, E. (2006). *Graphical Models for Visual Object Recognition and Tracking*. Massachusetts Institute of Technology

Teh, Y., & Jordan, M. (2010). Hierarchical Bayesian Nonparametric Models with Applications. In S. W. Hjort, C. Holmes, P. Mueller (Ed.), *Bayesian Nonparametrics: Principles and*

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). *Hierarchical Dirichlet Processes*.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(47), 1566–1581

Williamson, S. A., Dubey, A., & Xing, E. P. (2012). Exact and efficient parallel inference for nonparametric mixture models. *arXiv preprint arXiv:1211.7120*

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., et al. (2006). *The HTK Book* (p. 384). Cambridge, UK.