# A Nonparametric Bayesian Approach to Acoustic Model Discovery

**Chia-ying Lee and James Glass**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{chiaying,jrg}@csail.mit.edu

## Abstract

We investigate the problem of acoustic modeling in which prior language-specific knowledge and transcribed data are unavailable. We present an unsupervised model that simultaneously segments the speech, discovers a proper set of sub-word units (e.g., phones) and learns a Hidden Markov Model (HMM) for each induced acoustic unit. Our approach is formulated as a Dirichlet process mixture model in which each mixture is an HMM that represents a sub-word unit. We apply our model to the TIMIT corpus, and the results demonstrate that our model discovers sub-word units that are highly correlated with English phones and also produces better segmentation than the state-of-the-art unsupervised baseline. We test the quality of the learned acoustic models on a spoken term detection task. Compared to the baselines, our model improves the relative precision of top hits by at least 22.1% and outperforms a language-mismatched acoustic model.

## 1 Introduction

Acoustic models are an indispensable component of speech recognizers. However, the standard process of training acoustic models is expensive, and requires not only language-specific knowledge, e.g., the phone set of the language, a pronunciation dictionary, but also a large amount of transcribed data. Unfortunately, these necessary data are only available for a very small number of languages in the world. Therefore, a procedure for training acoustic models without annotated data would not only be a breakthrough from the traditional approach, but would also allow us to build speech recognizers for any language efficiently.

In this paper, we investigate the problem of unsupervised acoustic modeling with only spoken utterances as training data. As suggested in Garcia and Gish (2006), unsupervised acoustic modeling can be broken down to three sub-tasks: segmentation, clustering segments, and modeling the sound pattern of each cluster. In previous work, the three sub-problems were often approached sequentially and independently in which initial steps are not related to later ones (Lee et al., 1988; Garcia and Gish, 2006; Chan and Lee, 2011). For example, the speech data was usually segmented regardless of the clustering results and the learned acoustic models.

In contrast to the previous methods, we approach the problem by modeling the three sub-problems as well as the unknown set of sub-word units as latent variables in one nonparametric Bayesian model. More specifically, we formulate a Dirichlet process mixture model where each mixture is a Hidden Markov Model (HMM) used to model a sub-word unit and to generate observed segments of that unit. Our model seeks the set of sub-word units, segmentation, clustering and HMMs that best represent the observed data through an iterative inference process. We implement the inference process using Gibbs sampling.

We test the effectiveness of our model on the TIMIT database (Garofolo et al., 1993). Our model shows its ability to discover sub-word units that are highly correlated with standard English phones and to capture acoustic context information. For the segmentation task, our model outperforms the state-of-



*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 40–49,
Jeju, Republic of Korea, 8-14 July 2012. 

the-art unsupervised method and improves the relative F-score by 18.8 points (Dusan and Rabiner, 2006). Finally, we test the quality of the learned acoustic models through a keyword spotting task. Compared to the state-of-the-art unsupervised methods (Zhang and Glass, 2009; Zhang et al., 2012), our model yields a relative improvement in precision of top hits by at least 22.1% with only some degradation in equal error rate (EER), and outperforms a language-mismatched acoustic model trained with supervised data.

## 2 Related Work

**Unsupervised Sub-word Modeling** We follow the general guideline used in (Lee et al., 1988; Garcia and Gish, 2006; Chan and Lee, 2011) and approach the problem of unsupervised acoustic modeling by solving three sub-problems of the task: segmentation, clustering and modeling each cluster. The key difference, however, is that our model does not assume independence among the three aspects of the problem, which allows our model to refine its solution to one sub-problem by exploiting what it has learned about other parts of the problem. Second, unlike (Lee et al., 1988; Garcia and Gish, 2006) in which the number of sub-word units to be learned is assumed to be known, our model learns the proper size from the training data directly.

Instead of segmenting utterances, the authors of (Varadarajan et al., 2008) trained a single state HMM using all data at first, and then iteratively split the HMM states based on objective functions. This method achieved high performance in a phone recognition task using a label-to-phone transducer trained from some transcriptions. However, the performance seemed to rely on the quality of the transducer. For our work, we assume no transcriptions are available and measure the quality of the learned acoustic units via a spoken query detection task as in Jansen and Church (2011).

Jansen and Church (2011) approached the task of unsupervised acoustic modeling by first discovering repetitive patterns in the data, and then learned a whole-word HMM for each found pattern, where the state number of each HMM depends on the average length of the pattern. The states of the whole-word HMMs were then collapsed and used to represent acoustic units. Instead of discovering repetitive patterns first, our model is able to learn from any given data.

**Unsupervised Speech Segmentation** One goal of our model is to segment speech data into small sub-word (e.g., phone) segments. Most unsupervised speech segmentation methods rely on acoustic change for hypothesizing phone boundaries (Scharenborg et al., 2010; Qiao et al., 2008; Dusan and Rabiner, 2006; Estevan et al., 2007). Even though the overall approaches differ, these algorithms are all one-stage and bottom-up segmentation methods (Scharenborg et al., 2010). Our model does not make a single one-stage decision; instead, it infers the segmentation through an iterative process and exploits the learned sub-word models to guide its hypotheses on phone boundaries.

**Bayesian Model for Segmentation** Our model is inspired by previous applications of nonparametric Bayesian models to segmentation problems in NLP and speaker diarization (Goldwater, 2009; Fox et al., 2011); particularly, we adapt the inference method used in (Goldwater, 2009) to our segmentation task. Our problem is, in principle, similar to the word segmentation problem discussed in (Goldwater, 2009). The main difference, however, is that our model is under the continuous real value domain, and the problem of (Goldwater, 2009) is under the discrete symbolic domain. For the domain our problem is applied to, our model has to include more latent variables and is more complex.

## 3 Problem Formulation

The goal of our model, given a set of spoken utterances, is to jointly learn the following:

- Segmentation: To find the phonetic boundaries within each utterance.
- Nonparametric clustering: To find a proper set of clusters and group acoustically similar segments into the same cluster.
- Sub-word modeling: To learn a HMM to model each sub-word acoustic unit.

We model the three sub-tasks as latent variables in our approach. In this section, we describe the observed data, latent variables, and auxiliary variables

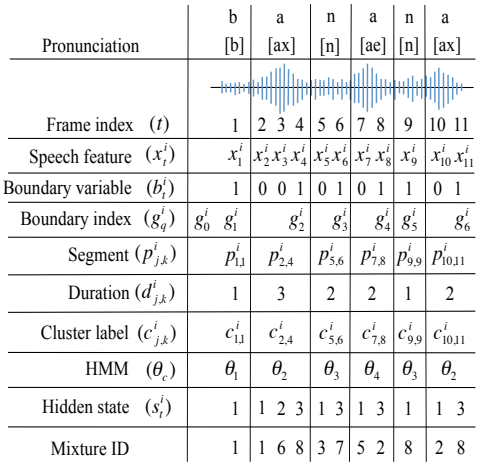


Figure 1: An example of the observed data and hidden variables of the problem for the word *banana*. See Section 3 for a detailed explanation.

of the problem and show an example in Fig. 1. In the next section, we show the generative process our model uses to generate the observed data.

**Speech Feature ($x_t^i$)** The only observed data for our problem are a set of spoken utterances, which are converted to a series of 25 ms 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) and their first- and second-order time derivatives at a 10 ms analysis rate. We use $x_t^i \in \mathbb{R}^{39}$ to denote the $t^{th}$ feature frame of the $i^{th}$ utterance. Fig. 1 illustrates how the speech signal of a single word utterance *banana* is converted to a sequence of feature vectors $x_1^i$ to $x_{11}^i$.

**Boundary ($b_t^i$)** We use a binary variable $b_t^i$ to indicate whether a phone boundary exists between $x_t^i$ and $x_{t+1}^i$. If our model hypothesizes $x_t^i$ to be the last frame of a sub-word unit, which is called a *boundary frame* in this paper, $b_t^i$ is assigned with value 1; or 0 otherwise. Fig. 1 shows an example of the boundary variables where the values correspond to the true answers. We use an auxiliary variable $g_q^i$ to denote the index of the $q^{th}$ boundary frame in utterance $i$. To make the derivation of posterior distributions easier in Section 5, we define $g_0^i$ to be the beginning of an utterance, and $L_i$ to be the number of boundary frames in an utterance. For the example shown in Fig. 1, $L_i$ is equal to 6.

**Segment ($p_{j,k}^i$)** We define a segment to be composed of feature vectors between two boundary frames. We use $p_{j,k}^i$ to denote a segment that consists of $x_j^i, x_{j+1}^i \cdots x_k^i$ and $d_{j,k}^i$ to denote the length of $p_{j,k}^i$. See Fig. 1 for more examples.

**Cluster Label ($c_{j,k}^i$)** We use $c_{j,k}^i$ to specify the cluster label of $p_{j,k}^i$. We assume segment $p_{j,k}^i$ is generated by the sub-word HMM with label $c_{j,k}^i$.

**HMM ($\theta_c$)** In our model, each HMM has three emission states, which correspond to the beginning, middle and end of a sub-word unit (Jelinek, 1976). A traversal of each HMM must start from the first state, and only left-to-right transitions are allowed even though we allow skipping of the middle and the last state for segments shorter than three frames. The emission probability of each state is modeled by a diagonal Gaussian Mixture Model (GMM) with 8 mixtures. We use $\theta_c$ to represent the set of parameters that define the $c^{th}$ HMM, which includes state transition probability $a_c^{j,k}$, and the GMM parameters of each state emission probability. We use $w_{c,s}^m \in \mathbb{R}$, $\mu_{c,s}^m \in \mathbb{R}^{39}$ and $\lambda_{c,s}^m \in \mathbb{R}^{39}$ to denote the weight, mean vector and the diagonal of the inverse covariance matrix of the $m^{th}$ mixture in the GMM for the $s^{th}$ state in the $c^{th}$ HMM.

**Hidden State ($s_t^i$)** Since we assume the observed data are generated by HMMs, each feature vector, $x_t^i$, has an associated hidden state index. We denote the hidden state of $x_t^i$ as $s_t^i$.

**Mixture ID ($m_t^i$)** Similarly, each feature vector is assumed to be emitted by the state GMM it belongs to. We use $m_t^i$ to identify the Gaussian mixture that generates $x_t^i$.

## 4 Model

We aim to discover and model a set of sub-word units that represent the spoken data. If we think of utterances as sequences of repeated sub-word units, then in order to find the sub-words, we need a model that concentrates probability on highly frequent patterns while still preserving probability for previously unseen ones. Dirichlet processes are particulary suitable for our goal. Therefore, we construct our model as a Dirichlet Process (DP) mixture model, of which the components are HMMs that are used

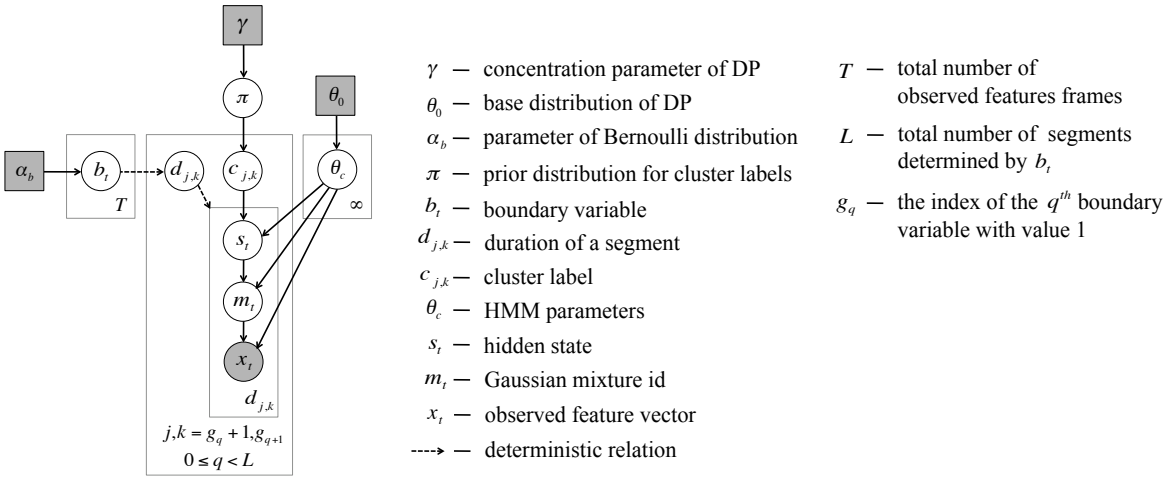


Figure 2: The graphical model for our approach. The shaded circle denotes the observed feature vectors, and the squares denote the hyperparameters of the priors used in our model. The dotted arrows indicate deterministic relations. Note that the Markov chain structure over the $s_t$ variables is not shown here due to limited space.

to model sub-word units. We assume each spoken segment is generated by one of the clusters in this DP mixture model. Here, we describe the generative process our model uses to generate the observed utterances and present the corresponding graphical model. For clarity, we assume that the values of the boundary variables $b_t^i$ are given in the generative process. In the next section, we explain how to infer their values.

Let $p^i_{g_q^i+1,g_{q+1}^i}$ for $0 \leq q \leq L_i - 1$ be the segments of the $i^{th}$ utterance. Our model assumes each segment is generated as follows:

1. Choose a cluster label $c^i_{g_q^i+1,g_{q+1}^i}$ for $p^i_{g_q^i+1,g_{q+1}^i}$. This cluster label can be either an existing label or a new one. Note that the cluster label determines which HMM is used to generate the segment.

2. Given the cluster label, choose a hidden state for each feature vector $x_t^i$ in the segment.

3. For each $x_t^i$, based on its hidden state, choose a mixture from the GMM of the chosen state.

4. Use the chosen Gaussian mixture to generate the observed feature vector $x_t^i$.

The generative process indicates that our model ignores utterance boundaries and views the entire data as concatenated spoken segments. Given this viewpoint, we discard the utterance index, $i$, of all variables in the rest of the paper.

The graphical model representing this generative process is shown in Fig. 2, where the shaded circle denotes the observed feature vectors, and the squares denote the hyperparameters of the priors used in our model. Specifically, we use a Bernoulli distribution as the prior of the boundary variables and impose a Dirichlet process prior on the cluster labels and the HMM parameters. The dotted arrows represent deterministic relations. For example, the boundary variables deterministically construct the duration of each segment, $d$, which in turn sets the number of feature vectors that should be generated for a segment. In the next section, we show how to infer the value of each of the latent variables in Fig. 2[1].

## 5 Inference

We employ Gibbs sampling (Gelman et al., 2004) to approximate the posterior distribution of the hidden variables in our model. To apply Gibbs sampling to our problem, we need to derive the conditional posterior distributions of each hidden variable of the model. In the following sections, we first derive the sampling equations for each hidden variable and then describe how we incorporate acoustic cues to reduce the sampling load at the end.

[1]Note that the value of $\pi$ is irrelevant to our problem; therefore, it is integrated out in the inference process

### 5.1 Sampling Equations

Here we present the sampling equations for each hidden variable defined in Section 3. We use $P(\cdot|\cdots)$ to denote a conditional posterior probability given observed data, all the other variables, and hyperparameters for the model.

**Cluster Label ($c_{j,k}$)** Let $C$ be the set of distinctive label values in $c_{-j,k}$, which represents all the cluster labels except $c_{j,k}$. The conditional posterior probability of $c_{j,k}$ for $c \in C$ is:

$$P(c_{j,k} = c|\cdots) \propto P(c_{j,k} = c|c_{-j,k};\gamma)P(p_{j,k}|\theta_c) = \frac{n^{(c)}}{N-1+\gamma}P(p_{j,k}|\theta_c) \quad (1)$$

where $\gamma$ is a parameter of the DP prior. The first line of Eq. 1 follows Bayes' rule. The first term is the conditional prior, which is a result of the DP prior imposed on the cluster labels [2]. The second term is the conditional likelihood, which reflects how likely the segment $p_{j,k}$ is generated by HMM$_c$. We use $n^{(c)}$ to represent the number of cluster labels in $c_{-j,k}$ taking the value $c$ and $N$ to represent the total number of segments in current segmentation.

In addition to existing cluster labels, $c_{j,k}$ can also take a new cluster label, which corresponds to a new sub-word unit. The corresponding conditional posterior probability is:

$$P(c_{j,k} \neq c, c \in C|\cdots) \propto \frac{\gamma}{N-1+\gamma}\int_\theta P(p_{j,k}|\theta)\,d\theta \quad (2)$$

To deal with the integral in Eq. 2, we follow the suggestions in (Rasmussen, 2000; Neal, 2000). We sample an HMM from the prior and compute the likelihood of the segment given the new HMM to approximate the integral.

Finally, by normalizing Eq. 1 and Eq. 2, the Gibbs sampler can draw a new value for $c_{j,k}$ by sampling from the normalized distribution.

**Hidden State ($s_t$)** To enforce the assumption that a traversal of an HMM must start from the first state and end at the last state[3], we do not sample hidden state indices for the first and the last frame of a segment. For each of the remaining feature vectors in a segment $p_{j,k}$, we sample a hidden state index according to the conditional posterior probability:

$$P(s_t = s|\cdots) \propto P(s_t = s|s_{t-1})P(x_t|\theta_{c_{j,k}}, s_t = s)P(s_{t+1}|s_t = s) = a^{s_{t-1},s}_{c_{j,k}}P(x_t|\theta_{c_{j,k}}, s_t = s)a^{s,s_{t+1}}_{c_{j,k}} \quad (3)$$

where the first term and the third term are the conditional prior – the transition probability of the HMM that $p_{j,k}$ belongs to. The second term is the likelihood of $x_t$ being emitted by state $s$ of HMM$_{c_{j,k}}$. Note for initialization, $s_t$ is sampled from the first prior term in Eq. 3.

**Mixture ID ($m_t$)** For each feature vector in a segment, given the cluster label $c_{j,k}$ and the hidden state index $s_t$, the derivation of the conditional posterior probability of its mixture ID is straightforward:

$$P(m_t = m|\cdots) \propto P(m_t = m|\theta_{c_{j,k}}, s_t)P(x_t|\theta_{c_{j,k}}, s_t, m_t = m) = w^m_{c_{j,k},s_t}P(x_t|\mu^m_{c_{j,k},s_t}, \lambda^m_{c_{j,k},s_t}) \quad (4)$$

where $1 \leq m \leq 8$. The conditional posterior consists of two terms: 1) the mixing weight of the $m^{th}$ Gaussian in the state GMM indexed by $c_{j,k}$ and $s_t$ and 2) the likelihood of $x_t$ given the Gaussian mixture. The sampler draws a value for $m_t$ from the normalized distribution of Eq. 4.

**HMM Parameters ($\theta_c$)** Each $\theta_c$ consists of two sets of variables that define an HMM: the state emission probabilities $w^m_{c,s}, \mu^m_{c,s}, \lambda^m_{c,s}$ and the state transition probabilities $a^{j,k}_c$. In the following, we derive the conditional posteriors of these variables.

**Mixture Weight $w^m_{c,s}$:** We use $\underline{w}_{c,s} = \{w^m_{c,s}|1 \leq m \leq 8\}$ to denote the mixing weights of the Gaussian mixtures of state $s$ of HMM $c$. We choose a symmetric Dirichlet distribution with a positive hyperparameter $\beta$ as its prior. The conditional posterior probability of $\underline{w}_{c,s}$ is:

$$P(\underline{w}_{c,s}|\cdots) \propto P(\underline{w}_{c,s};\beta)P(\mathbf{m}_{c,s}|\underline{w}_{c,s}) \propto Dir(\underline{w}_{c,s};\beta)Mul(\mathbf{m}_{c,s};\underline{w}_{c,s}) \propto Dir(\underline{w}_{c,s};\beta') \quad (5)$$

where $\mathbf{m}_{c,s}$ is the set of mixture IDs of feature vectors that belong to state $s$ of HMM $c$. The $m^{th}$ entry of $\beta'$ is $\beta + \sum_{m_t \in \mathbf{m}_{c,s}} \delta(m_t, m)$, where we use $\delta(\cdot)$

[2] See (Neal, 2000) for an overview on Dirichlet process mixture models and the inference methods.

[3] If a segment has only 1 frame, we assign the first state to it.

$$
\begin{aligned}
P(p_{l,t}, p_{t+1,r}|\mathbf{c}^-, \boldsymbol{\theta}) &= P(p_{l,t}|\mathbf{c}^-, \boldsymbol{\theta}) P(p_{t+1,r}|\mathbf{c}^-, c_{l,t}, \boldsymbol{\theta}) \\
&= \left[ \sum_{c \in C} \frac{n^{(c)}}{N^- + \gamma} P(p_{l,t}|\theta_c) + \frac{\gamma}{N^- + \gamma} \int_\theta P(p_{l,t}|\theta)\, d\theta \right] \\
&\times \left[ \sum_{c \in C} \frac{n^{(c)} + \delta(c_{l,t}, c)}{N^- + 1 + \gamma} P(p_{t+1,r}|\theta_c) + \frac{\gamma}{N^- + 1 + \gamma} \int_\theta P(p_{t+1,r}|\theta)\, d\theta \right] \\
P(p_{l,r}|\mathbf{c}^-, \boldsymbol{\theta}) &= \sum_{c \in C} \frac{n^{(c)}}{N^- + \gamma} P(p_{l,r}|\theta_c) + \frac{\gamma}{N^- + \gamma} \int_\theta P(p_{l,r}|\theta)\, d\theta
\end{aligned}
$$

Figure 3: The full derivation of the relative conditional posterior probabilities of a boundary variable.

to denote the discrete Kronecker delta. The last line of Eq. 5 comes from the fact that Dirichlet distributions are a conjugate prior for multinomial distributions. This property allows us to derive the update rule analytically.

**Gaussian Mixture $\mu_{c,s}^m, \lambda_{c,s}^m$:** We assume the dimensions in the feature space are independent. This assumption allows us to derive the conditional posterior probability for a single-dimensional Gaussian and generalize the results to other dimensions.

Let the $d^{th}$ entry of $\mu_{c,s}^m$ and $\lambda_{c,s}^m$ be $\mu_{c,s}^{m,d}$ and $\lambda_{c,s}^{m,d}$. The conjugate prior we use for the two variables is a normal-Gamma distribution with hyperparameters $\mu_0$, $\kappa_0$, $\alpha_0$ and $\beta_0$ (Murphy, 2007).

$$
\begin{aligned}
&P(\mu_{c,s}^{m,d}, \lambda_{c,s}^{m,d}|\mu_0, \kappa_0, \alpha_0, \beta_0) \\
&= N(\mu_{c,s}^{m,d}|\mu_0, (\kappa_0 \lambda_{c,s}^{m,d})^{-1}) Ga(\lambda_{c,s}^{m,d}|\alpha_0, \beta_0)
\end{aligned}
$$

By tracking the $d^{th}$ dimension of feature vectors $x \in \{x_t | m_t = m, s_t = s, c_{j,k} = c, x_t \in p_{j,k}\}$, we can derive the conditional posterior distribution of $\mu_{c,s}^{m,d}$ and $\lambda_{c,s}^{m,d}$ analytically following the procedures shown in (Murphy, 2007). Due to limited space, we encourage interested readers to find more details in (Murphy, 2007).

**Transition Probabilities $a_c^{j,k}$:** We represent the transition probabilities at state $j$ in HMM $c$ using $\underline{a}_c^j$. If we view $\underline{a}_c^j$ as mixing weights for states reachable from state $j$, we can simply apply the update rule derived for the mixing weights of Gaussian mixtures shown in Eq. 5 to $\underline{a}_c^j$. Assume we use a symmetric Dirichlet distribution with a positive hyperparameter $\eta$ as the prior, the conditional posterior for $\underline{a}_c^j$ is:

$$
P(\underline{a}_c^j | \cdots) \propto Dir(\underline{a}_c^j; \eta')
$$

where the $k^{th}$ entry of $\eta'$ is $\eta + n_c^{j,k}$, the number of occurrences of the state transition pair $(j, k)$ in segments that belong to HMM $c$.

**Boundary Variable ($b_t$)** To derive the conditional posterior probability for $b_t$, we introduce two variables:

$$
\begin{aligned}
l &= (\arg\max_{g_q} g_q < t) + 1 \\
r &= \arg\min_{g_q} t < g_q
\end{aligned}
$$

where $l$ is the index of the closest turned-on boundary variable that precedes $b_t$ plus 1, while $r$ is the index of the closest turned-on boundary variable that follows $b_t$. Note that because $g_0$ and $g_L$ are defined, $l$ and $r$ always exist for any $b_t$.

Note that the value of $b_t$ only affects segmentation between $x_l$ and $x_r$. If $b_t$ is turned on, the sampler hypothesizes two segments $p_{l,t}$ and $p_{t+1,r}$ between $x_l$ and $x_r$. Otherwise, only one segment $p_{l,r}$ is hypothesized. Since the segmentation on the rest of the data remains the same no matter what value $b_t$ takes, the conditional posterior probability of $b_t$ is:

$$
P(b_t = 1 | \cdots) \propto P(p_{l,t}, p_{t+1,r}|\mathbf{c}^-, \boldsymbol{\theta}) \quad (6)
$$

$$
P(b_t = 0 | \cdots) \propto P(p_{l,r}|\mathbf{c}^-, \boldsymbol{\theta}) \quad (7)
$$

where we assume that the prior probabilities for $b_t = 1$ and $b_t = 0$ are equal; $\mathbf{c}^-$ is the set of cluster labels of all segments except those between $x_l$ and $x_r$ ; and $\boldsymbol{\theta}$ indicates the set of HMMs that have associated segments. Our Gibbs sampler hypothesizes $b_t$'s value by sampling from the normalized distribution of Eq. 6 and Eq. 7. The full derivations of Eq. 6 and Eq. 7 are shown in Fig. 3.

Note that in Fig. 3, $N^-$ is the total number of segments in the data except those between $x_l$ and $x_r$.

For $b_t = 1$, to account the fact that when the model generates $p_{t+1,r}$, $p_{l,t}$ is already generated and owns a cluster label, we sample a cluster label for $p_{l,t}$ that is reflected in the Kronecker delta function. To handle the integral in Fig. 3, we sample one HMM from the prior and compute the likelihood using the new HMM to approximate the integral as suggested in (Rasmussen, 2000; Neal, 2000).

### 5.2 Heuristic Boundary Elimination

To reduce the inference load on the boundary variables $b_t$, we exploit acoustic cues in the feature space to eliminate $b_t$'s that are unlikely to be phonetic boundaries. We follow the pre-segmentation method described in Glass (2003) to achieve the goal. For the rest of the boundary variables that are proposed by the heuristic algorithm, we randomly initialize their values and proceed with the sampling process described above.

## 6 Experimental Setup

To the best of our knowledge, there are no standard corpora for evaluating unsupervised methods for acoustic modeling. However, numerous related studies have reported performance on the TIMIT corpus (Dusan and Rabiner, 2006; Estevan et al., 2007; Qiao et al., 2008; Zhang and Glass, 2009; Zhang et al., 2012), which creates a set of strong baselines for us to compare against. Therefore, the TIMIT corpus is chosen as the evaluation set for our model. In this section, we describe the methods used to measure the performance of our model on the following three tasks: sub-word acoustic modeling, segmentation and nonparametric clustering.

**Unsupervised Segmentation** We compare the phonetic boundaries proposed by our model to the manual labels provided in the TIMIT dataset. We follow the suggestion of (Scharenborg et al., 2010) and use a 20-ms tolerance window to compute recall, precision rates and F-score of the segmentation our model proposed for TIMIT's training set. We compare our model against the state-of-the-art unsupervised and semi-supervised segmentation methods that were also evaluated on the TIMIT training set (Dusan and Rabiner, 2006; Qiao et al., 2008).

**Nonparametric Clustering** Our model automatically groups speech segments into different clusters. One question we are interested in answering is whether these learned clusters correlate to English phones. To answer the question, we develop a method to map cluster labels to the phone set in a dataset. We align each cluster label in an utterance to the phone(s) it overlaps with in time by using the boundaries proposed by our model and the manually-labeled ones. When a cluster label overlaps with more than one phone, we align it to the phone with the largest overlap.[4] We compile the alignment results for 3696 training utterances[5] and present a confusion matrix between the learned cluster labels and the 48 phonetic units used in TIMIT (Lee and Hon, 1989).

**Sub-word Acoustic Modeling** Finally, and most importantly, we need to gauge the quality of the learned sub-word acoustic models. In previous work, Varadarajan et al. (2008) and Garcia and Gish (2006) tested their models on a phone recognition task and a term detection task respectively. These two tasks are fair measuring methods, but performance on these tasks depends not only on the learned acoustic models, but also other components such as the label-to-phone transducer in (Varadarajan et al., 2008) and the graphone model in (Garcia and Gish, 2006). To reduce performance dependencies on components other than the acoustic model, we turn to the task of spoken term detection, which is also the measuring method used in (Jansen and Church, 2011).

We compare our unsupervised acoustic model with three supervised ones: 1) an English triphone model, 2) an English monophone model and 3) a Thai monophone model. The first two were trained on TIMIT, while the Thai monophone model was trained with 32 hour clean read Thai speech from the LOTUS corpus (Kasuriya et al., 2003). All of the three models, as well as ours, used three-state HMMs to model phonetic units. To conduct spoken term detection experiments on the TIMIT dataset, we computed a posteriorgram representation for both training and test feature frames over the

[4] Except when a cluster label is mapped to /vcl/ /b/, /vcl/ /g/ and /vcl/ /d/, where the duration of the release /b/, /g/, /d/ is almost always shorter than the closure /vcl/. In this case, we align the cluster label to both the closure and the release.

[5] The TIMIT training set excluding the sa-type subset.

| $\gamma$ | $\alpha_b$ | $\beta$ | $\eta$ | $\mu_0$ | $\kappa_0$ | $\alpha_0$ | $\beta_0$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.5 | 3 | 3 | $\boldsymbol{\mu}^d$ | 5 | 3 | $3/\boldsymbol{\lambda}^d$ |

Table 1: The values of the hyperparameters of our model, where $\boldsymbol{\mu}^d$ and $\boldsymbol{\lambda}^d$ are the $d^{th}$ entry of the mean and the diagonal of the inverse covariance matrix of training data.

HMM states for each of the four models. Ten keywords were randomly selected for the task. For every keyword, spoken examples were extracted from the training set and were searched for in the test set using segmental dynamic time warping (Zhang and Glass, 2009).

In addition to the supervised acoustic models, we also compare our model against the state-of-the-art unsupervised methods for this task (Zhang and Glass, 2009; Zhang et al., 2012). Zhang and Glass (2009) trained a GMM with 50 components to decode posteriorgrams for the feature frames, and Zhang et al. (2012) used a deep Boltzmann machine (DBM) trained with pseudo phone labels generated from an unsupervised GMM to produce a posteriorgram representation. The evaluation metrics they used were: 1) P@N, the average precision of the top N hits, where N is the number of occurrences of each keyword in the test set; 2) EER: the average equal error rate at which the false acceptance rate is equal to the false rejection rate. We also report experimental results using the P@N and EER metrics.

**Hyperparameters and Training Iterations** The values of the hyperparameters of our model are shown in Table 1, where $\boldsymbol{\mu}^d$ and $\boldsymbol{\lambda}^d$ are the $d^{th}$ entry of the mean and the diagonal of the inverse covariance matrix computed from training data. We pick these values to impose weak priors on our model.[6] We run our sampler for 20,000 iterations, after which the evaluation metrics for our model all converged. In Section 7, we report the performance of our model using the sample from the last iteration.

## 7 Results

Fig. 4 shows a confusion matrix of the 48 phones used in TIMIT and the sub-word units learned from 3696 TIMIT utterances. Each circle represents a mapping pair for a cluster label and an English phone. The confusion matrix demonstrates a strong

[6]In the future, we plan to extend the model and infer the values of these hyperparameters from data directly.

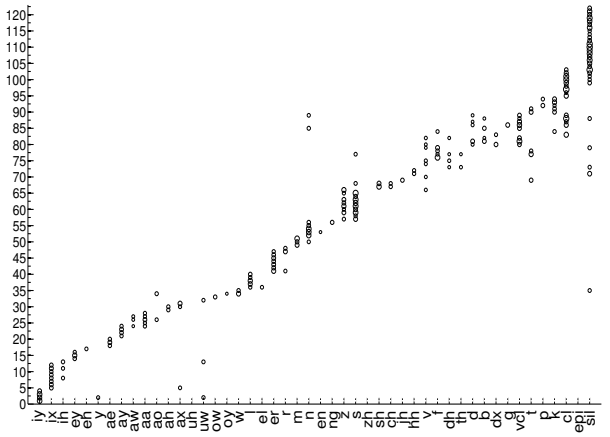


Figure 4: A confusion matrix of the learned cluster labels from the TIMIT training set excluding the sa type utterances and the 48 phones used in TIMIT. Note that for clarity, we show only pairs that occurred more than 200 times in the alignment results. The average co-occurrence frequency of the mapping pairs in this figure is 431.

correlation between the cluster labels and individual English phones. For example, clusters 19, 20 and 21 are mapped exclusively to the vowel /ae/. A more careful examination on the alignment results shows that the three clusters are mapped to the same vowel in a different acoustic context. For example, cluster 19 is mapped to /ae/ followed by stop consonants, while cluster 20 corresponds to /ae/ followed by nasal consonants. This context-dependent relationship is also observed in other English phones and their corresponding sets of clusters. Fig. 4 also shows that a cluster may be mapped to multiple English phones. For instance, clusters 85 and 89 are mapped to more than one phone; nevertheless, a closer look reveals that these clusters are mapped to /n/, /d/ and /b/, which are sounds with a similar place of articulation (i.e. labial and dental). These correlations indicate that our model is able to discover the phonetic composition of a set of speech data without any language-specific knowledge.

The performance of the four acoustic models on the spoken term detection task is presented in Table 2. The English triphone model achieves the best P@N and EER results and performs slightly better than the English monophone model, which indicates a correlation between the quality of an acoustic model and its performance on the spoken term detection task. Although our unsupervised model does not perform as well as the supervised English

| unit(%) | P@N | EER |
|---|---|---|
| English triphone | 75.9 | 11.7 |
| English monophone | 74.0 | 11.8 |
| Thai monophone | 56.6 | 14.9 |
| Our model | 63.0 | 16.9 |

Table 2: The performance of our model and three supervised acoustic models on the spoken term detection task.

acoustic models, it generates a comparable EER and a more accurate detection performance for top hits than the Thai monophone model. This indicates that even without supervision, our model captures and learns the acoustic characteristics of a language automatically and is able to produce an acoustic model that outperforms a language-mismatched acoustic model trained with high supervision.

Table 3 shows that our model improves P@N by a large margin and generates only a slightly worse EER than the GMM baseline on the spoken term detection task. At the end of the training process, our model induced 169 HMMs, which were used to compute posteriorgrams. This seems unfair at first glance because Zhang and Glass (2009) only used 50 Gaussians for decoding, and the better result of our model could be a natural outcome of the higher complexity of our model. However, Zhang and Glass (2009) pointed out that using more Gaussian mixtures for their model did not improve their model performance. This indicates that the key reason for the improvement is our joint modeling method instead of simply the higher complexity of our model.

Compared to the DBM baseline, our model produces a higher EER; however, it improves the relative detection precision of top hits by 24.3%. As indicated in (Zhang et al., 2012), the hierarchical structure of DBM allows the model to provide a descent posterior representation of phonetic units. Even though our model only contains simple HMMs and Gaussians, it still achieves a comparable, if not better, performance as the DBM baseline. This demonstrates that even with just a simple model structure, the proposed learning algorithm is able to acquire rich phonetic knowledge from data and generate a fine posterior representation for phonetic units.

Table 4 summarizes the segmentation performance of the baselines, our model and the heuristic

| unit(%) | P@N | EER |
|---|---|---|
| GMM (Zhang and Glass, 2009) | 52.5 | 16.4 |
| DBM (Zhang et al., 2012) | 51.1 | 14.7 |
| Our model | 63.0 | 16.9 |

Table 3: The performance of our model and the GMM and DBM baselines on the spoken term detection task.

| unit(%) | Recall | Precision | F-score |
|---|---|---|---|
| Dusan (2006) | 75.2 | 66.8 | 70.8 |
| Qiao et al. (2008)* | 77.5 | 76.3 | 76.9 |
| Our model | 76.2 | 76.4 | 76.3 |
| Pre-seg | 87.0 | 50.6 | 64.0 |

Table 4: The segmentation performance of the baselines, our model and the heuristic pre-segmentation on TIMIT training set. *The number of phone boundaries in each utterance was assumed to be known in this model.

pre-segmentation (pre-seg) method. The language-independent pre-seg method is suitable for seeding our model. It eliminates most unlikely boundaries while retaining about 87% true boundaries. Even though this indicates that at best our model only recalls 87% of the true boundaries, the pre-seg reduces the search space significantly. In addition, it also allows the model to capture proper phone durations, which compensates the fact that we do not include any explicit duration modeling mechanisms in our approach. In the best semi-supervised baseline model (Qiao et al., 2008), the number of phone boundaries in an utterance was assumed to be known. Although our model does not incorporate this information, it still achieves a very close F-score. When compared to the baseline in which the number of phone boundaries in each utterance was also unknown (Dusan and Rabiner, 2006), our model outperforms in both recall and precision, improving the relative F-score by 18.8%. The key difference between the two baselines and our method is that our model does not treat segmentation as a stand-alone problem; instead, it jointly learns segmentation, clustering and acoustic units from data. The improvement on the segmentation task shown by our model further supports the strength of the joint learning scheme proposed in this paper.

## 8 Conclusion

We present a Bayesian unsupervised approach to the problem of acoustic modeling. Without any prior

knowledge, this method is able to discover phonetic units that are closely related to English phones, improve upon state-of-the-art unsupervised segmentation method and generate more precise spoken term detection performance on the TIMIT dataset. In the future, we plan to explore phonological context and use more flexible topological structures to model acoustic units within our framework.

## Acknowledgements

The authors would like to thank Hung-an Chang and Ekapol Chuangsuwanich for training the English and Thai acoustic models. Thanks to Matthew Johnson, Ramesh Sridharan, Finale Doshi, S.R.K. Branavan, the MIT Spoken Language Systems group and the anonymous reviewers for helpful comments.

## References

Chun-An Chan and Lin-Shan Lee. 2011. Unsupervised hidden Markov modeling of spoken queries for spoken term detection without speech recognition. In *Proceedings of INTERSPEECH*, pages 2141 – 2144.

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Sorin Dusan and Lawrence Rabiner. 2006. On the relation between maximum spectral transition positions and phone boundaries. In *Proceedings of INTERSPEECH*, pages 1317 – 1320.

Yago Pereiro Estevan, Vincent Wan, and Odette Scharenborg. 2007. Finding maximum margin segments in speech. In *Proceedings of ICASSP*, pages 937 – 940.

Emily Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. 2011. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*.

Alvin Garcia and Herbert Gish. 2006. Keyword spotting of arbitrary words using minimal speech resources. In *Proceedings of ICASSP*, pages 949–952.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallet, Nancy L. Dahlgren, and Victor Zue. 1993. Timit acoustic-phonetic continuous speech corpus.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, second edition.

James Glass. 2003. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137 – 152.

Sharon Goldwater. 2009. A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, 112:21–54.

Aren Jansen and Kenneth Church. 2011. Towards unsupervised training of speaker independent acoustic models. In *Proceedings of INTERSPEECH*, pages 1693 – 1696.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64:532 – 556.

Sawit Kasuriya, Virach Sornlertlamvanich, Patcharika Cotsomrong, Supphanat Kanokphara, and Nattanun Thatphithakkul. 2003. Thai speech corpus for Thai speech recognition. In *Proceedings of Oriental COCOSDA*, pages 54–61.

Kai-Fu Lee and Hsiao-Wuen Hon. 1989. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37:1641 – 1648.

Chin-Hui Lee, Frank Soong, and Biing-Hwang Juang. 1988. A segment model based approach to speech recognition. In *Proceedings of ICASSP*, pages 501–504.

Kevin P. Murphy. 2007. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, University of British Columbia.

Radford M. Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Yu Qiao, Naoya Shimomura, and Nobuaki Minematsu. 2008. Unsupervised optimal phoeme segmentation: Objectives, algorithms and comparisons. In *Proceedings of ICASSP*, pages 3989 – 3992.

Carl Edward Rasmussen. 2000. The infinite Gaussian mixture model. *In Advances in Neural Information Processing Systems*, 12:554–560.

Odette Scharenborg, Vincent Wan, and Mirjam Ernestus. 2010. Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. *Journal of the Acoustical Society of America*, 127:1084–1095.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of ACL-08: HLT, Short Papers*, pages 165–168.

Yaodong Zhang and James Glass. 2009. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams. In *Proceedings of ASRU*, pages 398 – 403.

Yaodong Zhang, Ruslan Salakhutdinov, Hung-An Chang, and James Glass. 2012. Resource configurable spoken query detection using deep Boltzmann machines. In *Proceedings of ICASSP*, pages 5161–5164.