A Non-parametric Approach for Acoustic Model Discovery

Chia-ying Lee and James Glass

MIT Computer Science and Artificial Intelligence Lab
Spoken Language Systems Group
Acoustic Model
Acoustic Model
Training an Acoustic Model

- Manually transcribed data are required
Training an Acoustic Model

- Manually transcribed data are required
  - Phone transcriptions

```
/b/  /ax/  /n/  /ae/  /n/  /ax/
```

Acoustic Model
Training an Acoustic Model

- Manually transcribed data are required
  - Phone transcriptions
  - Word transcriptions

Acoustic Model

/b/ /æx/ /n/ /æ/ /n/ /æx/
banana
Towards Unsupervised Training

- Can we train an acoustic model with just speech input?
Towards Unsupervised Training

- Can we train an acoustic model with just speech input?
Towards Unsupervised Training

- Can we train an acoustic model with just speech input?
Related Work

- **Inspiration**
  - A Bayesian framework for word segmentation: Exploring the effects of context [Goldwater et al., Cognition 2009]
  - Towards unsupervised training of speaker independent acoustic models [Jansen and Church, INTERSPEECH 2011]
  - Unsupervised learning of acoustic sub-word units [Varadarajan et al., ACL 2008]
  - A segment model based approach to speech recognition [Lee et al., ICASSP1988]
Related Work

• **Inspiration**
  - A Bayesian framework for word segmentation: Exploring the effects of context [Goldwater et al., Cognition 2009]

• **Unsupervised acoustic modeling**
  - Towards unsupervised training of speaker independent acoustic models [Jansen and Church, INTERSPEECH 2011]
  - Unsupervised learning of acoustic sub-word units [Varadarajan et al., ACL 2008]
  - Keyword spotting of arbitrary words using minimal speech resources [Garcia and Gish, ICASSP 2006]
  - A segment model based approach to speech recognition [Lee et al., ICASSP 1988]
Challenges

/b/  /æ/  /n/  /æ/  /n/  /æ/
Challenges

- Unknown phone boundaries

/b/ /\ax/ /n/ /ae/ /n/ /\ax/
Challenges

- Unknown phone boundaries
- Unknown phone identities
Challenges

- Unknown phone boundaries
- Unknown phone identities
- Unknown phone set
Generative Story

- A simple explanation of how a spoken utterance is generated

- Assumptions
  - HMM-based mixture model
  - Speech segments are i.i.d
Generative Story

- A simple explanation of how a spoken utterance is generated

HMM1  HMM2  ...  HMMi  HMMi+1  ...
Generative Story

- A simple explanation of how a spoken utterance is generated.
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

• A simple explanation of how a spoken utterance is generated
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

- A simple explanation of how a spoken utterance is generated
Generative Story

• A simple explanation of how a spoken utterance is generated

• Main latent variables
  - Phone boundaries ($b$)

HMM1  HMM2  ...  HMMi  HMMi+1  ...

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_1$</td>
<td>$b_9$</td>
<td>$b_{16}$</td>
<td>$b_{28}$</td>
<td>$b_{37}$</td>
</tr>
<tr>
<td>i+1</td>
<td>1</td>
<td>i</td>
<td>2</td>
<td>i</td>
</tr>
<tr>
<td>/b/</td>
<td>/ax/</td>
<td>/n/</td>
<td>/ae/</td>
<td>/n/</td>
</tr>
</tbody>
</table>
Generative Story

- A simple explanation of how a spoken utterance is generated

- Main latent variables
  - Phone boundaries ($b$)
  - Cluster labels ($c$)
Generative Story

- A simple explanation of how a spoken utterance is generated

- Main latent variables
  - Phone boundaries ($b$)
  - Cluster labels ($c$)
  - HMM parameters ($\theta$)
Generative Story

- A simple explanation of how a spoken utterance is generated

- Main latent variables
  - Phone boundaries ($b$)
  - Cluster labels ($c$)
  - HMM parameters ($\theta$)
  - # of HMMs

\[
\begin{array}{c}
\theta_1 \\
\text{HMM1} \\
\hline
\theta_2 \\
\text{HMM2} \\
\hdots
\end{array}
\]
Unknown Number of HMMs

- An unknown set of phone units
Unknown Number of HMMs

- An unknown set of phone units
  - Impose a Dirichlet Process prior to guide inference on the number of HMMs
Unknown Number of HMMs

- An unknown set of phone units
  - Impose a Dirichlet Process prior to guide inference on the number of HMMs

- Is Dirichlet process (DP) a proper prior for this task?
  - Does phone frequency inherit power law?
Phone Frequency -- Monophone

Monophone frequency of TIMIT si data ~ 1890 utterances
Phone Frequency -- Triphone

freq = constant / rank
Triphone frequency of TIMIT si data ~ 1890 utterances
Unknown Number of HMMs

• An unknown set of phone units
  - Impose a Dirichlet Process prior to guide inference on the number of HMMs

• Is Dirichlet process (DP) a proper prior for this task?
  - Does phone frequency inherit power law?
Unknown Number of HMMs

- An unknown set of phone units
  - Impose a Dirichlet Process prior to guide inference on the number of HMMs

- Is Dirichlet process (DP) a proper prior for this task?
  - Does phone frequency inherit power law?
  - DP should be a reasonable prior to start with
Generative Story

- A simple explanation of how a spoken utterance is generated

Main latent variables
- Phone boundaries \((b)\)
- Cluster labels \((c)\)
- HMM parameters \((\theta)\)
- # of HMMs
Generative Model

\[ \alpha: \text{concentration parameter of DP} \]
\[ G_0: \text{base distribution of DP} \]
\[ \beta \sim \text{GEM}(\alpha) \]
\[ \alpha_b: \text{prior for } b_t \]
\[ x: \text{observations} \]
\[ L_i: \text{length of the } i\text{-th segment} \]
\[ N: \text{total number of segments} \]
\[ T: \text{total number of frames} \]

\[ \begin{align*}
  \alpha & \rightarrow \beta \\
  \beta & \sim \text{GEM}(\alpha) \\
  \alpha_b & \rightarrow b_t \\
  b_t & \rightarrow L_i \\
  L_i & \rightarrow C_i \\
  C_i & \rightarrow \theta_k \\
  \theta_k & \rightarrow \infty \\
  \theta_k & \rightarrow X_j \\
  X_j & \rightarrow x \\
  x & \rightarrow \text{data} \\
  j = 1 \ldots L_i \\
  i = 1 \ldots N 
\end{align*} \]

\( \rightarrow \) deterministic relation
Generative Model

\[ \alpha: \text{concentration parameter of DP} \]
\[ G_0: \text{base distribution of DP} \]
\[ \beta \sim \text{GEM}(\alpha) \]
\[ \alpha_b: \text{prior for } b_t \]
\[ x: \text{observations} \]
\[ L_i: \text{length of the } i\text{-th segment} \]
\[ N: \text{total number of segments} \]
\[ T: \text{total number of frames} \]
\[ \rightarrow \text{deterministic relation} \]
\[ \bigcirc \text{latent variables that will be inferred} \]
Inference Procedure

- Iterate \( n \) times
  - \( n = 20,000 \) in our experiments

1. Initialize boundary variables \((b_t)\) randomly
2. Sample \( c_i \) for each segment
3. Sample HMM parameters \((\theta_i)\)
4. Sample for each \( b_t \)
Inference Procedure

- Iterate $n$ times
  - $n = 20,000$ in our experiments
DP as a Prior for Cluster Labels ($c$)

- A Chinese restaurant process representation
  - Each table is a phonetic unit
  - Each speech segment is a customer $s_i = [X_t, X_{t+1}, \ldots, X_{t+L_i}]$
DP as a Prior for Cluster Labels ($c$)

- A Chinese restaurant process representation
  - Each table is a phonetic unit
  - Each speech segment is a customer $s_i = [x_t, x_{t+1}, \ldots x_{t+L_i}]$

$$c_1 = 1$$
A Chinese restaurant process representation

- Each table is a phonetic unit
- Each speech segment is a customer  $s_i = [x_t, x_{t+1}, \ldots, x_{t+L_i}]$

$$c_1 = 1$$  $$c_2 = 2$$
DP as a Prior for Cluster Labels ($c$)

- A Chinese restaurant process representation
  - Each table is a phonetic unit
  - Each speech segment is a customer $s_i = [x_t, x_{t+1}, \ldots x_{t+L_i}]$

$$c_1 = 1 \quad c_2 = 2 \quad c_3 = 1$$
DP as a Prior for Cluster Labels (c)

- A Chinese restaurant process representation
  - Each table is a phonetic unit
  - Each speech segment is a customer  $s_i = [x_t, x_{t+1}, \ldots, x_{t+Li}]$

\[ c_1 = 1 \quad c_2 = 2 \quad c_4 = 3 \quad c_t = K \]
\[ c_3 = 1 \quad c_8 = 2 \quad c_5 = 3 \]
\[ c_9 = 1 \]
Posterior Distribution for $c_i$

- For a new segment $(s_i)$, the posterior probability distribution of $c_i$:
For a new segment ($s_i$), the posterior probability distribution of $c_i$:
- $s_i$ sits at an occupied table $\implies$ $s_i$ is not a new phone
For a new segment \( s_i \), the posterior probability distribution of \( c_i \):

- \( s_i \) sits at an occupied table \( \rightarrow \) \( s_i \) is not a new phone
For a new segment \((s_i)\), the posterior probability distribution of \(c_i\):
- \(s_i\) sits at an occupied table \(\rightarrow\) \(s_i\) is not a new phone
For a new segment \((s_i)\), the posterior probability distribution of \(c_i\):

- \(s_i\) sits at an occupied table \(\rightarrow\) \(s_i\) is not a new phone

\[
p(c_i = k, 1 \leq k \leq K | \ldots) \propto \frac{n_k}{N - 1 + \alpha} p(s_i | \theta_k)
\]

- \(n_k\): number of customers at table \(k\)
- \(N\): number of customers seen so far
- \(\alpha\): concentration parameter of DP

Posterior probability  \(\propto\)  DP prior  \(\propto\)  likelihood
For a new segment \((s_i)\), the posterior probability distribution of \(c_i\):

- \(s_i\) sits at an occupied table \(\rightarrow\) \(s_i\) is not a new phone

\[ p(c_i = k, 1 \leq k \leq K | \cdots) \propto \frac{n_k}{N - 1 + \alpha} p(s_i | \theta_k) \]

- \(s_i\) opens a new table \(\rightarrow\) \(s_i\) is a new phone
• For a new segment \((s_i)\), the posterior probability distribution of \(c_i\):

- \(s_i\) sits at an occupied table \(\rightarrow\) \(s_i\) is not a new phone

\[
p(c_i = k, 1 \leq k \leq K | \cdots) \propto \frac{n_k}{N - 1 + \alpha} p(s_i | \theta_k)
\]

- \(s_i\) opens a new table \(\rightarrow\) \(s_i\) is a new phone

\[
p(c_i = K + 1 | \cdots) \propto \frac{\alpha}{N - 1 + \alpha} \int_{\theta} p(s_i | \theta) d\theta
\]
For a new segment \((s_i)\), the posterior probability distribution of \(c_i\):

- \(s_i\) sits at an occupied table \(\rightarrow\) \(s_i\) is not a new phone
  \[
p(c_i = k, 1 \leq k \leq K \mid \cdots) \propto \frac{n_k}{N - 1 + \alpha} p(s_i \mid \theta_k)
  \]
- \(s_i\) opens a new table \(\rightarrow\) \(s_i\) is a new phone
  \[
p(c_i = K + 1 \mid \cdots) \propto \frac{\alpha}{N - 1 + \alpha} \int_{\theta} p(s_i \mid \theta) d\theta
  \]
Inference Procedure

- Iterate \( n \) times
  - \( n = 20,000 \) in our experiments

1. Initialize boundary variables \((b_t)\) randomly
2. Sample \( c_i \) for each segment
3. Sample HMM parameters \((\theta_i)\)
4. Sample for each \( b_t \)

Gibbs sampling
Inference for HMM Parameters ($\theta$)

- HMM is used to model each phone
  - Three states with only left-to-right and self transitions
  - Always start from the first state
  - A 8-mixture diagonal GMM is used for the emission distributions

\[
\begin{align*}
\sum_{i=1}^{8} w_{1,i} N(u_{1,i}, \sigma_{1,i}^2) & \quad \sum_{i=1}^{8} w_{2,i} N(u_{2,i}, \sigma_{2,i}^2) & \quad \sum_{i=1}^{8} w_{3,i} N(u_{3,i}, \sigma_{3,i}^2)
\end{align*}
\]
Inference for HMM Parameters ($\theta$)

- HMM is used to model each phone
  - Three states with only left-to-right and self transitions
  - Always start from the first state
  - A 8-mixture diagonal GMM is used for the emission distributions

- Latent variables
  - Transition probabilities ($a$)
  - Mixture weights ($w$)
  - Mean ($\mu$)
  - Variance ($\sigma^2$)

\[
\begin{align*}
\sum_{i=1}^{8} w_{1,i} N(u_{1,i}, \sigma_{1,i}^2) & \quad \sum_{i=1}^{8} w_{2,i} N(u_{2,i}, \sigma_{2,i}^2) & \quad \sum_{i=1}^{8} w_{3,i} N(u_{3,i}, \sigma_{3,i}^2)
\end{align*}
\]
Priors and Posteriors for HMM

- **Priors**
  - Dirichlet distributions for transition probabilities ($a$) and mixture weights ($w$)
  - Normal-gamma distributions for Gaussian parameters ($\mu, \sigma^2$)
Priors and Posteriors for HMM

- **Priors**
  - Dirichlet distributions for transition probabilities \((a)\) and mixture weights \((w)\)
  - Normal-gamma distributions for Gaussian parameters \((\mu, \sigma^2)\)

- **Posteriors**
  - Gather relevant counts from customer segments
Priors and Posteriors for HMM

- **Priors**
  - Dirichlet distributions for transition probabilities ($\alpha$) and mixture weights ($\omega$)
  - Normal-gamma distributions for Gaussian parameters ($\mu, \sigma^2$)

- **Posteriors**
  - Gather relevant counts from customer segments
  - Update prior distributions
  - Sample new values for the latent variables
Inference Procedure

- Iterate \( n \) times
  - \( n = 20,000 \) in our experiments

1. Initialize boundary variables \((b_t)\) randomly
2. Sample \( c_i \) for each segment
3. Sample HMM parameters \((\theta_i)\)
4. Sample for each \( b_t \)
Inference on Phone Boundaries ($b$)

- **Boundary variables**
  - Naively, every frame can be a phone boundary
Inference on Phone Boundaries ($b$)

- **Boundary variables**
  - Naively, every frame can be a phone boundary
  - Boundary variables take binary values

![Diagram showing binary values for boundary variables]

<table>
<thead>
<tr>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>...</th>
<th>$b_9$</th>
<th>$b_{10}$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Prior and Posterior for Phone Boundaries

- Prior
  - Fixed prior probabilities $p(b_t = 1) = \alpha_b$ and $p(b_t = 0) = 1 - \alpha_b$
Prior and Posterior for Phone Boundaries

- **Prior**
  - Fixed prior probabilities $p(b_t = 1) = \alpha_b$ and $p(b_t = 0) = 1 - \alpha_b$

- **Posterior: examine one boundary variable ($b_t$) at a time**
  - Fix the current values of other boundary variables
  - Consider both 0 and 1 for $b_t$ and the respective segmentation outcomes
Prior and Posterior for Phone Boundaries

• **Prior**
  - Fixed prior probabilities \( p(b_t = 1) = \alpha_b \) and \( p(b_t = 0) = 1 - \alpha_b \)

• **Posterior:** examine one boundary variable \((b_t)\) at a time
  - Fix the current values of other boundary variables
  - Consider both 0 and 1 for \( b_t \) and the respective segmentation outcomes

\[
\begin{align*}
b_t &= 1 \\
St-1 & \quad & St+1
\end{align*}
\]
Prior and Posterior for Phone Boundaries

- **Prior**
  - Fixed prior probabilities \( p(b_t = 1) = \alpha_b \) and \( p(b_t = 0) = 1 - \alpha_b \)

- **Posterior**: examine one boundary variable \((b_t)\) at a time
  - Fix the current values of other boundary variables
  - Consider both 0 and 1 for \( b_t \) and the respective segmentation outcomes

\[
p(b_t = 1 | \cdots) \propto p(b_t = 1)p(s_{t-1} | c^-, \theta)p(s_{t+1} | c^-, \theta)
\]

- \( c^- \): cluster labels of all other segments
- \( \theta \): the set of HMMs
Prior and Posterior for Phone Boundaries

- **Prior**
  - Fixed prior probabilities $p(b_t = 1) = \alpha_b$ and $p(b_t = 0) = 1 - \alpha_b$

- **Posterior**: examine one boundary variable ($b_t$) at a time
  - Fix the current values of other boundary variables
  - Consider both 0 and 1 for $b_t$ and the respective segmentation outcomes
Prior and Posterior for Phone Boundaries

- **Prior**
  - Fixed prior probabilities $p(b_t = 1) = \alpha_b$ and $p(b_t = 0) = 1 - \alpha_b$

- **Posterior**: examine one boundary variable $(b_t)$ at a time
  - Fix the current values of other boundary variables
  - Consider both 0 and 1 for $b_t$ and the respective segmentation outcomes

$$p(b_t = 0 | \ldots) \propto \alpha_b p(b_t = 0)p(s_t | c^-, \Theta)$$
Prior and Posterior for Phone Boundaries

- **Prior**
  - Fixed prior probabilities $p(b_t = 1) = \alpha_b$ and $p(b_t = 0) = 1 - \alpha_b$

- **Posterior**: examine one boundary variable ($b_t$) at a time
  - Fix the current values of other boundary variables
  - Consider both 0 and 1 for $b_t$ and the respective segmentation outcomes

Generate a sample for $b_t$

$$
\begin{align*}
p(b_t = 1 | \cdots) & \propto \frac{p(b_t = 1)p(s_{t-1} | c^-, \theta)p(s_{t+1} | c^-, \theta)}{p(b_t = 0 | \cdots) \propto \frac{p(b_t = 0)p(s_t | c^-, \theta)}{}}
\end{align*}
$$
Acoustic Landmarks

- Naively, every frame can be a phone boundary
  - In fact, some frames are more likely to be boundaries and some are less likely
  - Compute landmarks [Glass et al. 2003] and only do inference on landmarks
  - A language-independent method
Acoustic Landmarks

- Naively, every frame can be a phone boundary
  - In fact, some frames are more likely to be boundaries and some are less likely
  - Compute landmarks [Glass et al. 2003] and only do inference on landmarks
  - A language-independent method
Acoustic Landmarks

• Naively, every frame can be a phone boundary
  - In fact, some frames are more likely to be boundaries and some are less likely
  - Compute landmarks [Glass et al. 2003] and only do inference on landmarks
  - A language-independent method

• Advantage
  - Reduce inference load
Experiments

- Data set
  - TIMIT training and test sets
  - Multi-speaker, clean read speech, 16kHz sampling rate
Experiments

- **Data set**
  - TIMIT training and test sets
  - Multi-speaker, clean read speech, 16kHz sampling rate

- **Qualitative assessment**
  - Correlation between induced phone units and English phones
  - Compare results of 300 and 3696 utterances
Experiments

• Data set
  - TIMIT training and test sets
  - Multi-speaker, clean read speech, 16kHz sampling rate

• Qualitative assessment
  - Correlation between induced phone units and English phones
  - Compare results of 300 and 3696 utterances

• Quantitative assessment
  - Spoken term detection
  - Phone segmentation
Discovered Phone Units -- 300 utterances

- 43 phone units discovered from 300 TIMIT utterances
  - Phone units are correlated with English broad phone classes
Discovered Phone Units -- 300 utterances

- 43 phone units discovered from 300 TIMIT utterances
  - Phone units are correlated with English broad phone classes

```
front    back    semi    nasal    fricative    stop
vowel    vowel   vowel
```

![Discovered phone units diagram](image)
Discovered Phone Units -- 300 utterances

- 43 phone units discovered from 300 TIMIT utterances
  - Phone units are correlated with English broad phone classes
Discovered Phone Units -- 300 utterances

- 43 phone units discovered from 300 TIMIT utterances
  - Phone units are correlated with English broad phone classes
Discovered Phone Units -- 300 utterances

- 43 phone units discovered from 300 TIMIT utterances
  - Phone units are correlated with English broad phone classes
Discovered Phone Units -- 3696 utterances

- 123 phone units discovered from 3696 TIMIT utterances
  - A finer correlation between discovered phones and English phones
Discovered Phone Units -- 3696 utterances

- 123 phone units discovered from 3696 TIMIT utterances
  - A finer correlation between discovered phones and English phones
Discovered Phone Units -- 3696 utterances

- 123 phone units discovered from 3696 TIMIT utterances
  - A finer correlation between discovered phones and English phones
Discovered Phone Units -- 3696 utterances

- 123 phone units discovered from 3696 TIMIT utterances

  - A finer correlation between discovered phones and English phones
Discovered Phone Units -- 3696 utterances

- 123 phone units discovered from 3696 TIMIT utterances
  - A finer correlation between discovered phones and English phones

Context-dependent:
   /ae/ + /m/, /n/
   /ae/ + stops
Spoken Term Detection

- Given a spoken query \( (w) \), find all spoken documents that contain \( w \)
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units
Spoken Term Detection

- Given a spoken query \( w \), find all spoken documents that contain \( w \)
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units

\[ x \]: a single frame of feature vector

\[ State_{i,j} \]: the \( j \)-th state of the \( i \)-th HMM
Spoken Term Detection

- Given a spoken query (w), find all spoken documents that contain w
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units

\[ \text{posterior-gram}(x) = \sum_{i=1}^{K} \sum_{j=1}^{3} \frac{p(\text{State}_{i,j} \mid x)}{\sum_{i=1}^{K} \sum_{j=1}^{3} p(\text{State}_{i,j} \mid x)} \text{ for } 1 \leq i \leq K \text{ and } 1 \leq j \leq 3 \]

\[ K \text{: the total number of HMMs} \]
Spoken Term Detection

- Given a spoken query (w), find all spoken documents that contain w
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units
  - Apply dynamic time warping to keyword detection [Zhang et al, 2009]
Spoken Term Detection

- Given a spoken query (w), find all spoken documents that contain w
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units
  - Apply dynamic time warping to keyword detection [Zhang et al, 2009]
  - 10 selected keywords
Spoken Term Detection

- Given a spoken query (w), find all spoken documents that contain w
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units
  - Apply dynamic time warping to keyword detection [Zhang et al, 2009]
  - 10 selected keywords

P@N: the average precision of top N hits

<table>
<thead>
<tr>
<th></th>
<th>P@N</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thai Monophone Model (Supervised)</td>
<td>56.6</td>
<td>14.9</td>
</tr>
<tr>
<td>Our model</td>
<td>63.0</td>
<td>16.9</td>
</tr>
<tr>
<td>English Monophone (Supervised)</td>
<td>74.0</td>
<td>11.8</td>
</tr>
<tr>
<td>Zhang 2009 (GMM) (Unsupervised)</td>
<td>52.5</td>
<td>16.4</td>
</tr>
</tbody>
</table>
Spoken Term Detection

- Given a spoken query (w), find all spoken documents that contain w
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units
  - Apply dynamic time warping to keyword detection [Zhang et al, 2009]
  - 10 selected keywords

P@N: the average precision of top N hits

<table>
<thead>
<tr>
<th>Model</th>
<th>P@N</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Monophone (Supervised)</td>
<td>74.0</td>
<td>11.8</td>
</tr>
<tr>
<td>Thai Monophone Model (Supervised)</td>
<td>56.6</td>
<td>14.9</td>
</tr>
<tr>
<td>Our model</td>
<td>63.0</td>
<td>16.9</td>
</tr>
</tbody>
</table>
Spoken Term Detection

- Given a spoken query (w), find all spoken documents that contain w
  - 3696 utterances for discovering phone units
  - Compute posterior-grams on the HMM states of the discovered phone units
  - Apply dynamic time warping to keyword detection [Zhang et al, 2009]
  - 10 selected keywords

<table>
<thead>
<tr>
<th></th>
<th>P@N</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Monophone (Supervised)</td>
<td>74.0</td>
<td>11.8</td>
</tr>
<tr>
<td>Thai Monophone Model (Supervised)</td>
<td>56.6</td>
<td>14.9</td>
</tr>
<tr>
<td>Our model</td>
<td>63.0</td>
<td>16.9</td>
</tr>
<tr>
<td>Zhang 2009 (GMM) (Unsupervised)</td>
<td>52.5</td>
<td>16.4</td>
</tr>
<tr>
<td>Zhang 2012 (DBM) (Unsupervised)</td>
<td>51.1</td>
<td>14.7</td>
</tr>
</tbody>
</table>
Phone Segmentation

- TIMIT training set
Phone Segmentation

- TIMIT training set

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dusan et al. (2006)</td>
<td>75.2</td>
<td>66.8</td>
<td>70.8</td>
</tr>
<tr>
<td>Qiao et al. (2008)</td>
<td>77.5</td>
<td>76.3</td>
<td>76.9</td>
</tr>
<tr>
<td>Our model</td>
<td>76.2</td>
<td>76.4</td>
<td>76.3</td>
</tr>
<tr>
<td>Landmarks</td>
<td>87.0</td>
<td>50.6</td>
<td>64.0</td>
</tr>
</tbody>
</table>
Conclusions

- An unsupervised framework for discovering acoustic model
  - Assume phone frequency adheres to power law
  - Use Dirichlet Process to guide inference on the unknown set of phones
Conclusions

- An unsupervised framework for discovering acoustic model
  - Assume phone frequency adheres to power law
  - Use Dirichlet Process to guide inference on the unknown set of phones

- Experimental results
  - Discovered units are highly correlated with standard phones
  - More accurate spoken term detection performance among top hits (P@N)
  - Segmentation results beat the state-of-the-art unsupervised method
Conclusions

• **An unsupervised framework for discovering acoustic model**
  - Assume phone frequency adheres to power law
  - Use Dirichlet Process to guide inference on the unknown set of phones

• **Experimental results**
  - Discovered units are highly correlated with standard phones
  - More accurate spoken term detection performance among top hits (P@N)
  - Segmentation results beat the state-of-the-art unsupervised method

• **Towards unsupervised training methods**
Conclusions

- **An unsupervised framework for discovering acoustic model**
  - Assume phone frequency adheres to power law
  - Use Dirichlet Process to guide inference on the unknown set of phones

- **Experimental results**
  - Discovered units are highly correlated with standard phones
  - More accurate spoken term detection performance among top hits (P@N)
  - Segmentation results beat the state-of-the-art unsupervised method

- **Towards unsupervised training methods**
Conclusions

- **An unsupervised framework for discovering acoustic model**
  - Assume phone frequency adheres to power law
  - Use Dirichlet Process to guide inference on the unknown set of phones

- **Experimental results**
  - Discovered units are highly correlated with standard phones
  - More accurate spoken term detection performance among top hits (P@N)
  - Segmentation results beat the state-of-the-art unsupervised method

- **Towards unsupervised training methods**
Thank you.
Future Work

- **Explore context information**
  - Revisit the assumption that phones are generated independently

- **Learn proper HMM structures from data**
  - Replace the fixed 3-state and 8 GMM structure

- **Apply to more languages**
  - Looking into the OGI corpus
  - Babel data