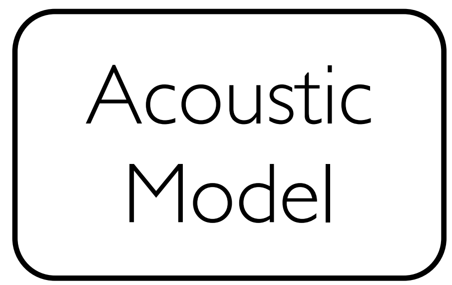
A Non-parametric Approach for Acoustic Model Discovery

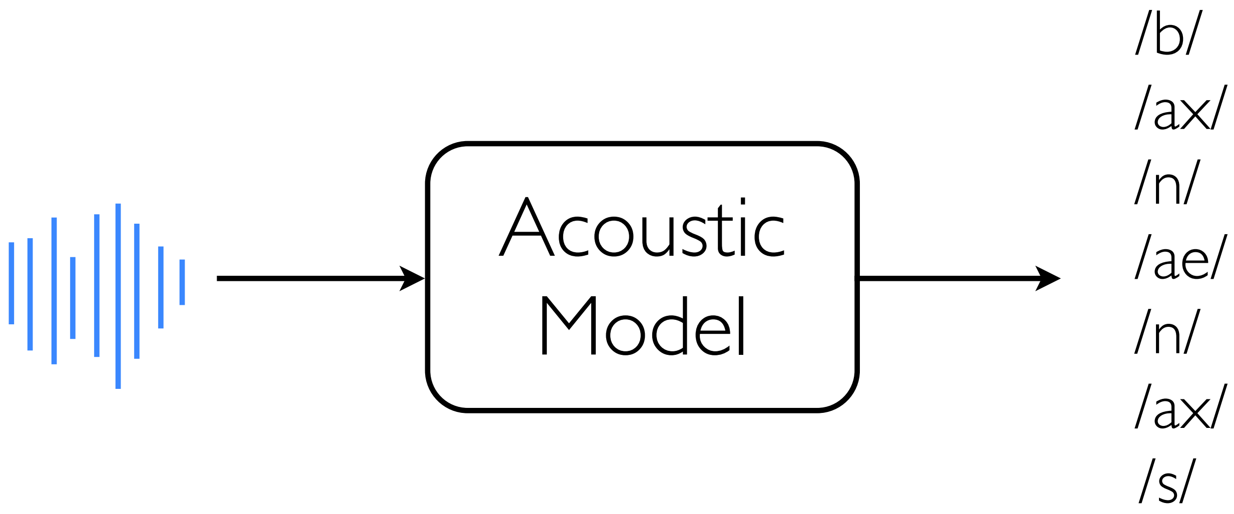
Chia-ying Lee and James Glass

MIT Computer Science and Artificial Intelligence Lab
Spoken Language Systems Group

Acoustic Model



Acoustic Model



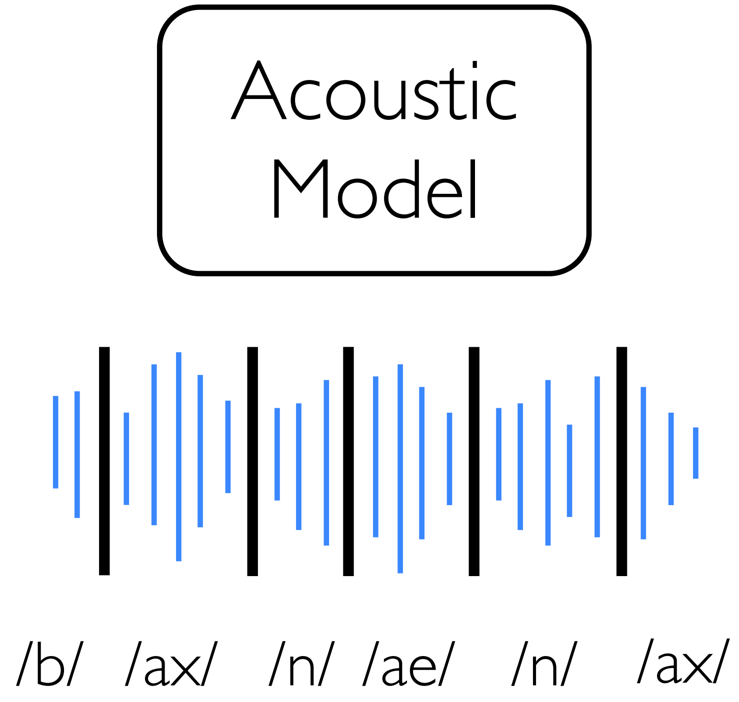
Training an Acoustic Model

- Manually transcribed data are required

Acoustic
Model

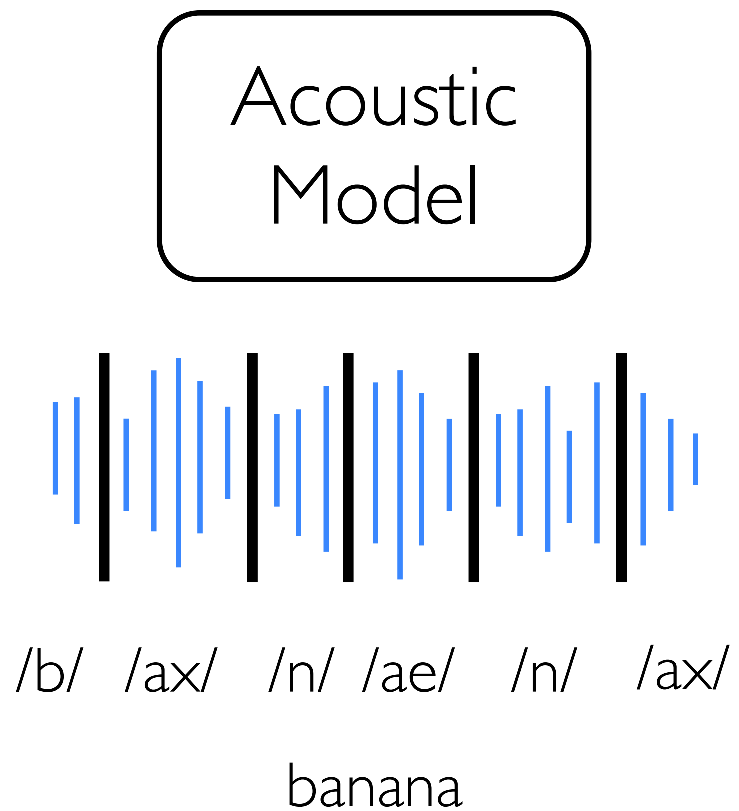
Training an Acoustic Model

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



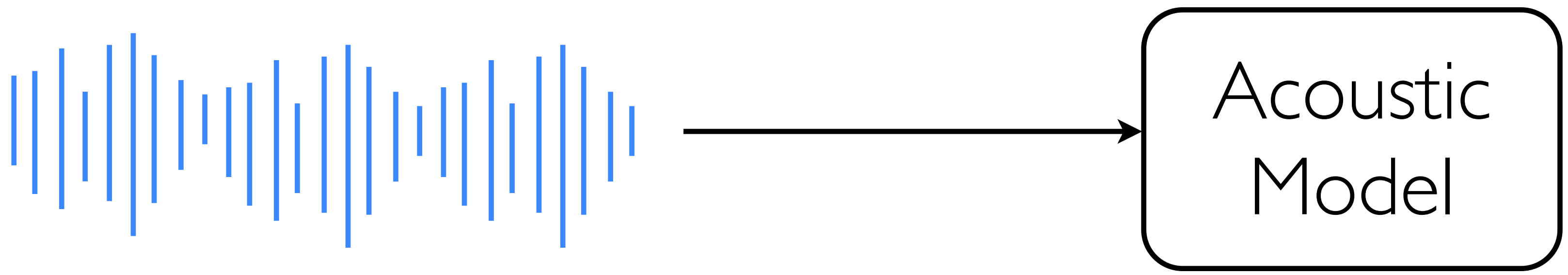
Training an Acoustic Model

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



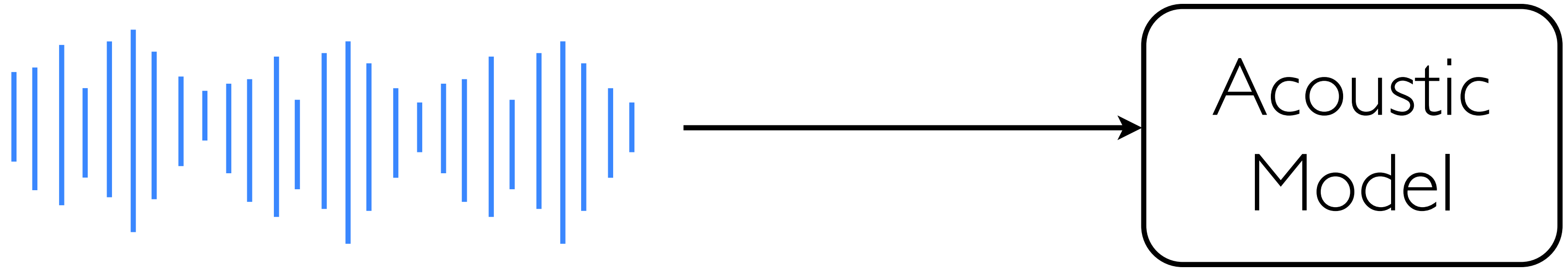
Towards Unsupervised Training

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



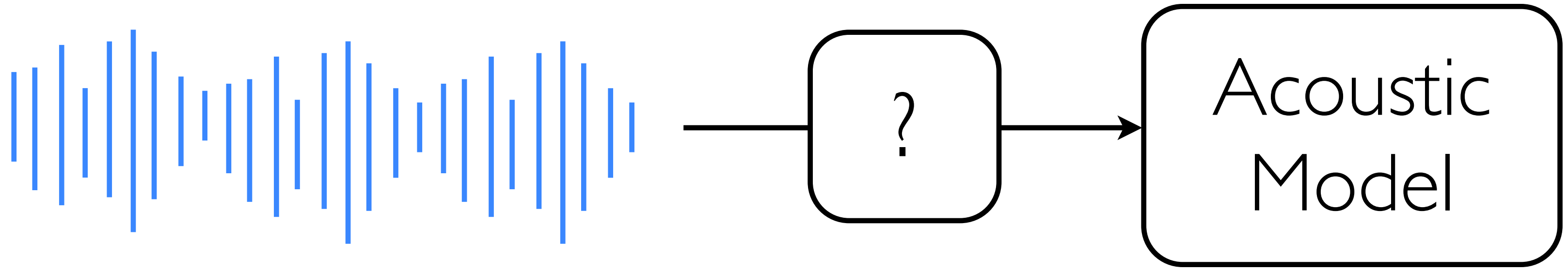
Towards Unsupervised Training

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



Towards Unsupervised Training

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



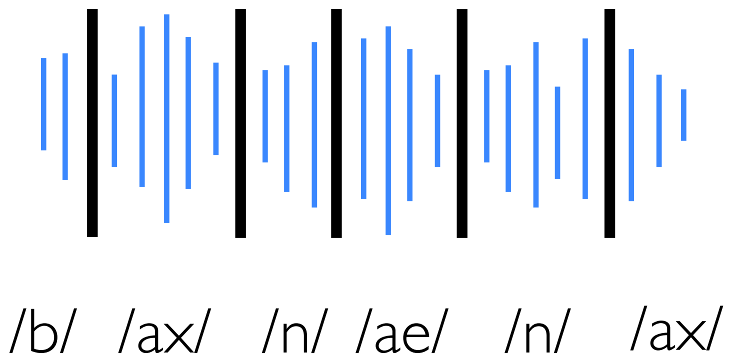
Related Work

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

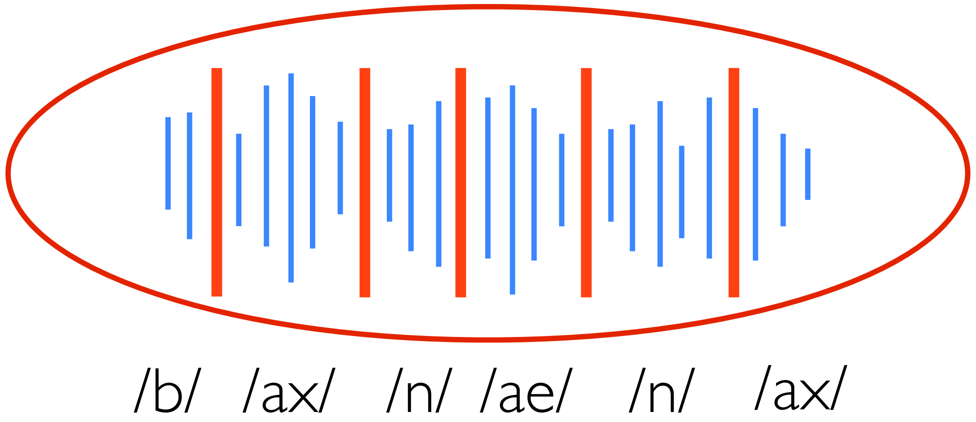
Related Work

- 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., *ICASSP1988*]

Challenges

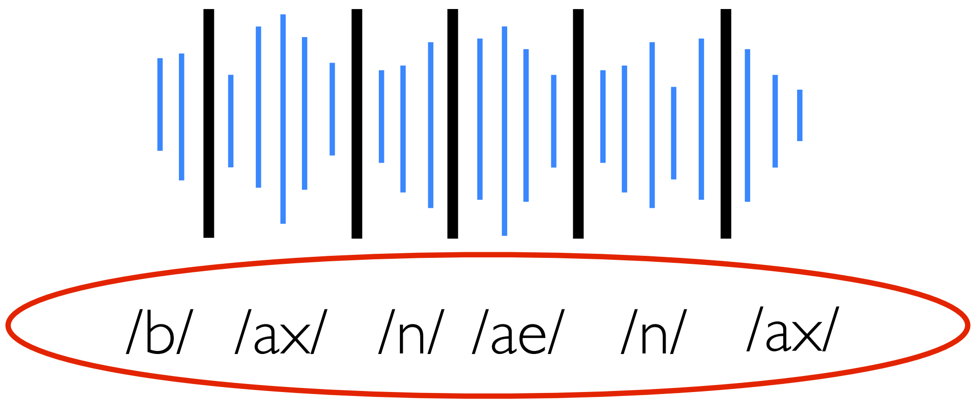


Challenges



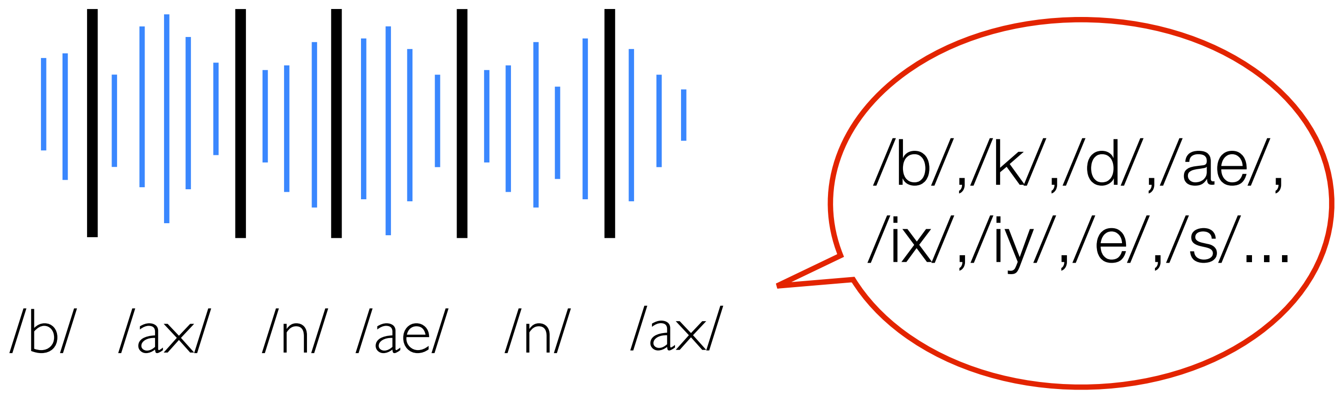
- Unknown phone boundaries

Challenges



- Unknown phone boundaries
- Unknown phone identities

Challenges



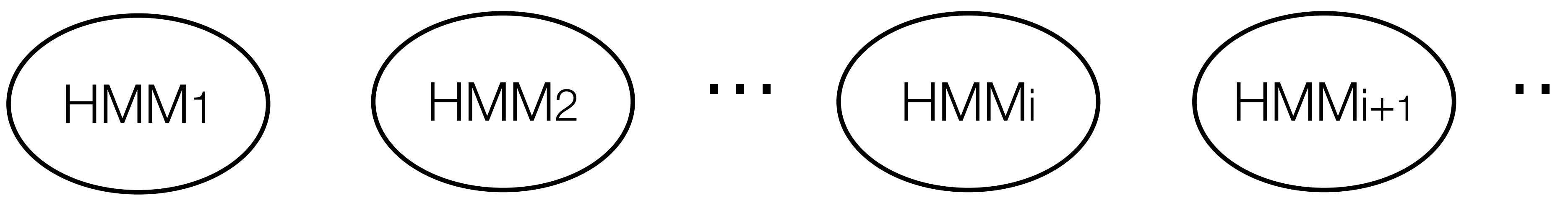
- Unknown phone boundaries
- Unknown phone identities
- Unknown phone set

Generative Story

- A simple explanation of how a spoken utterance is generated
- Assumptions
  - HMM-based mixture model
  - Speech segments are i.i.d

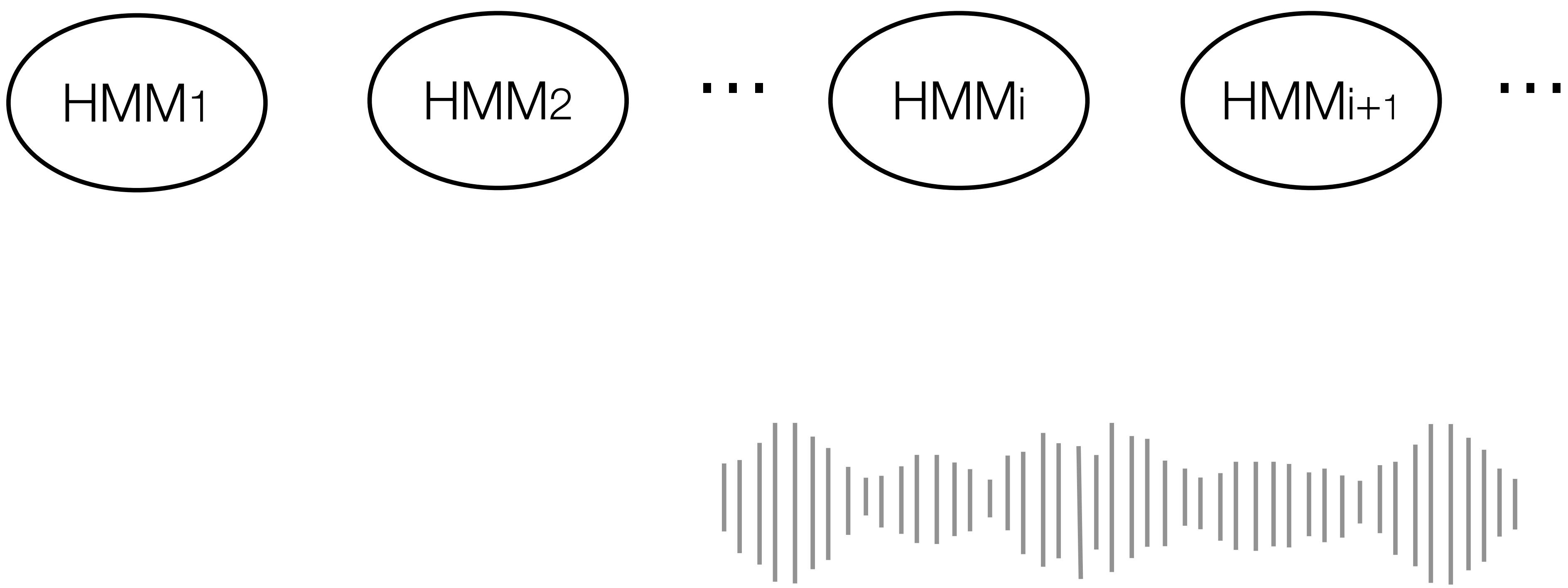
Generative Story

- A simple explanation of how a spoken utterance is generated



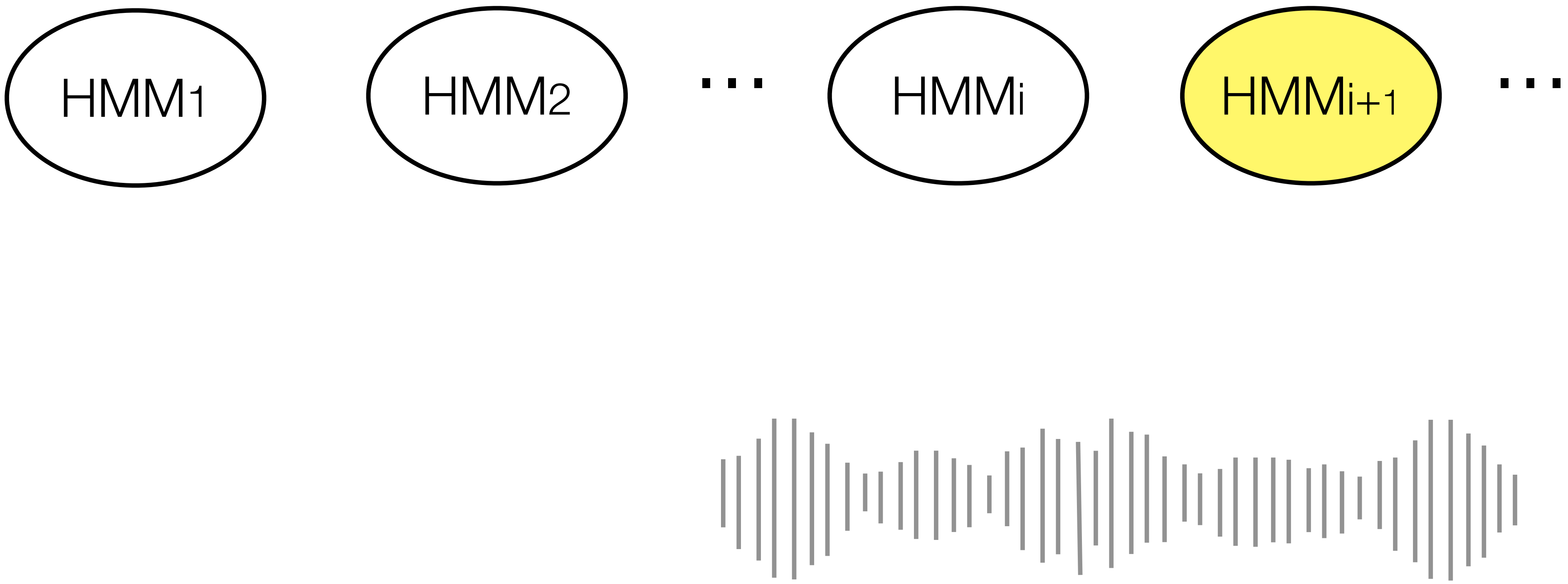
Generative Story

- A simple explanation of how a spoken utterance is generated



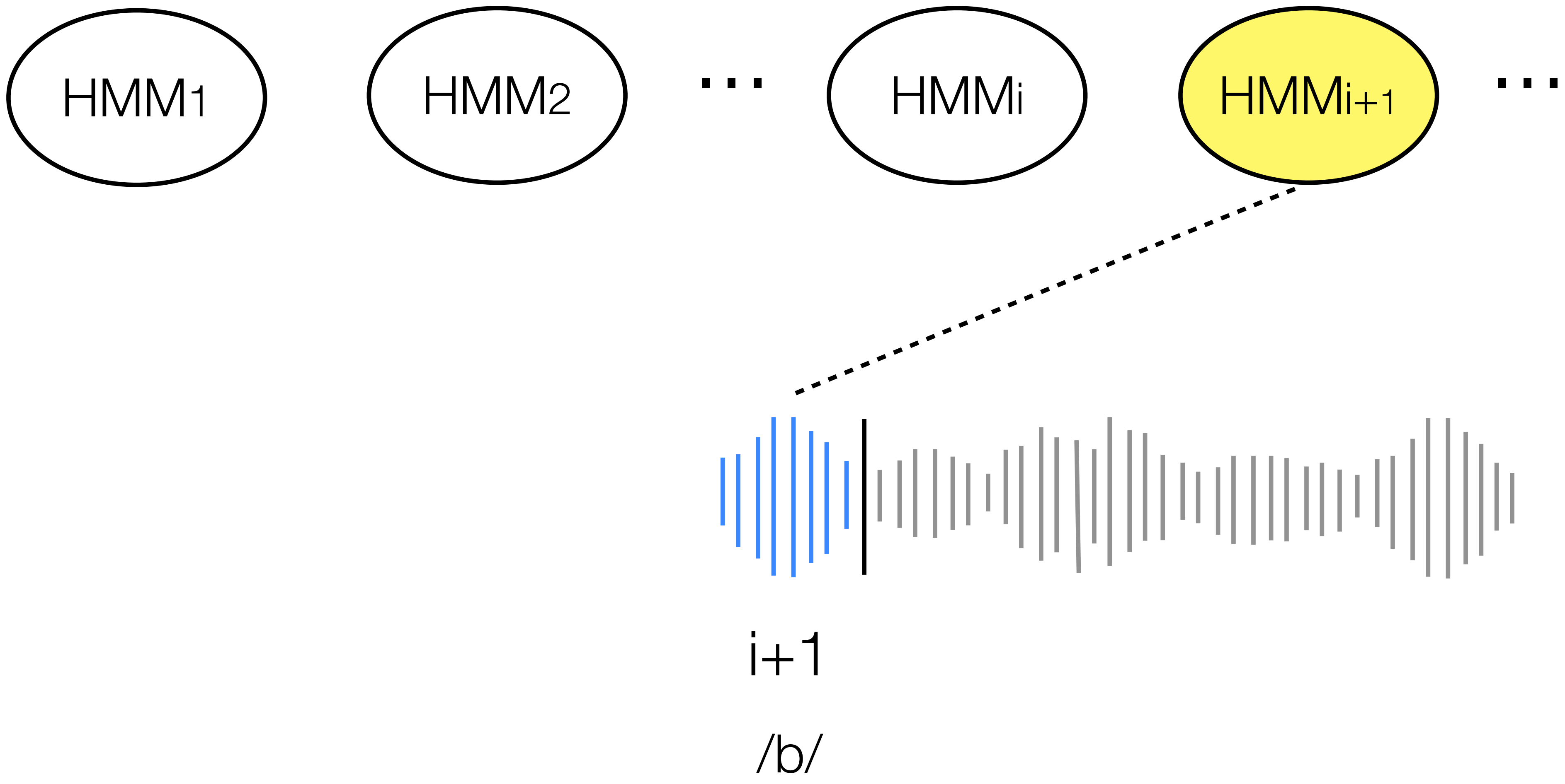
Generative Story

- A simple explanation of how a spoken utterance is generated



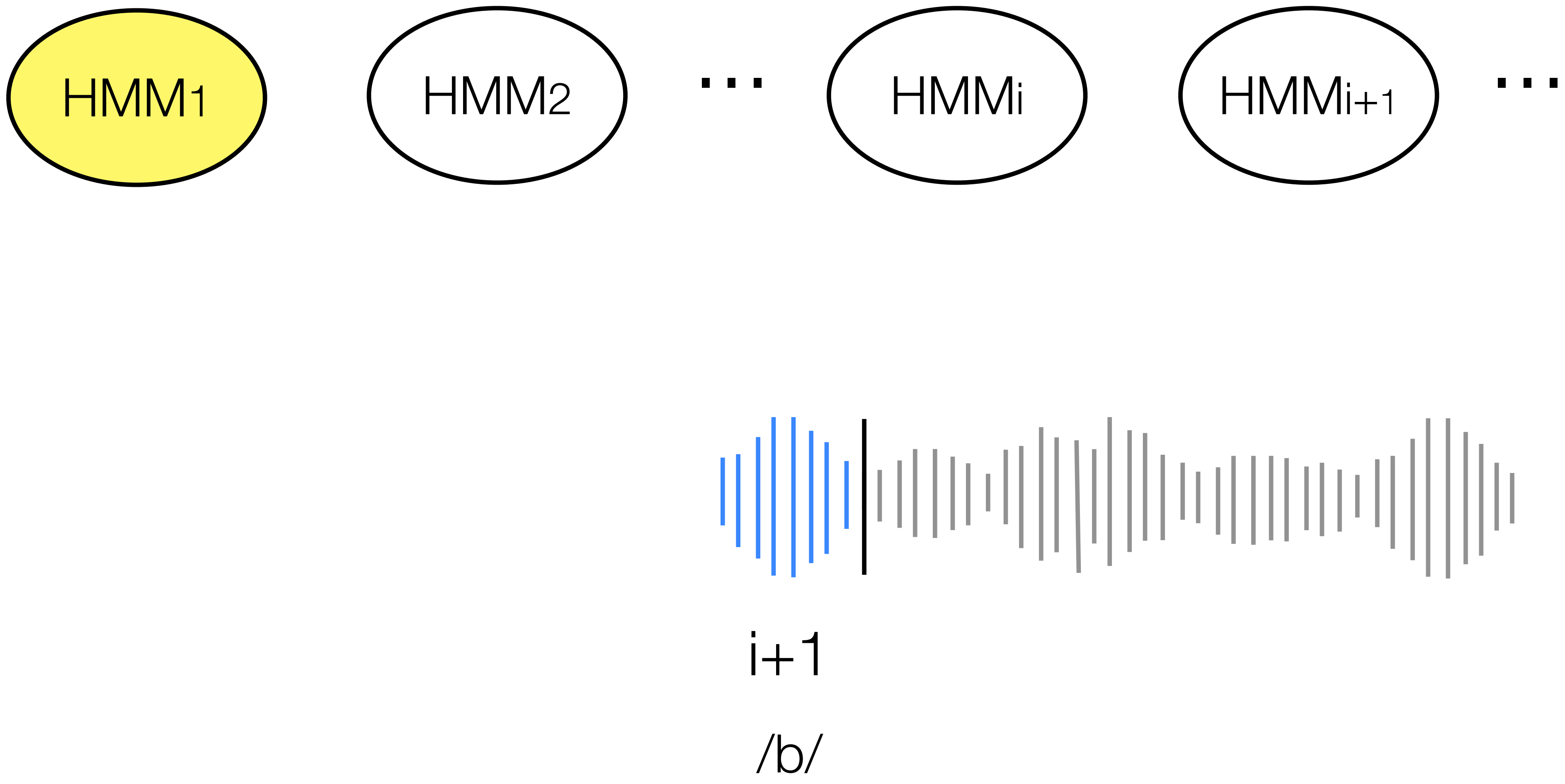
Generative Story

- A simple explanation of how a spoken utterance is generated



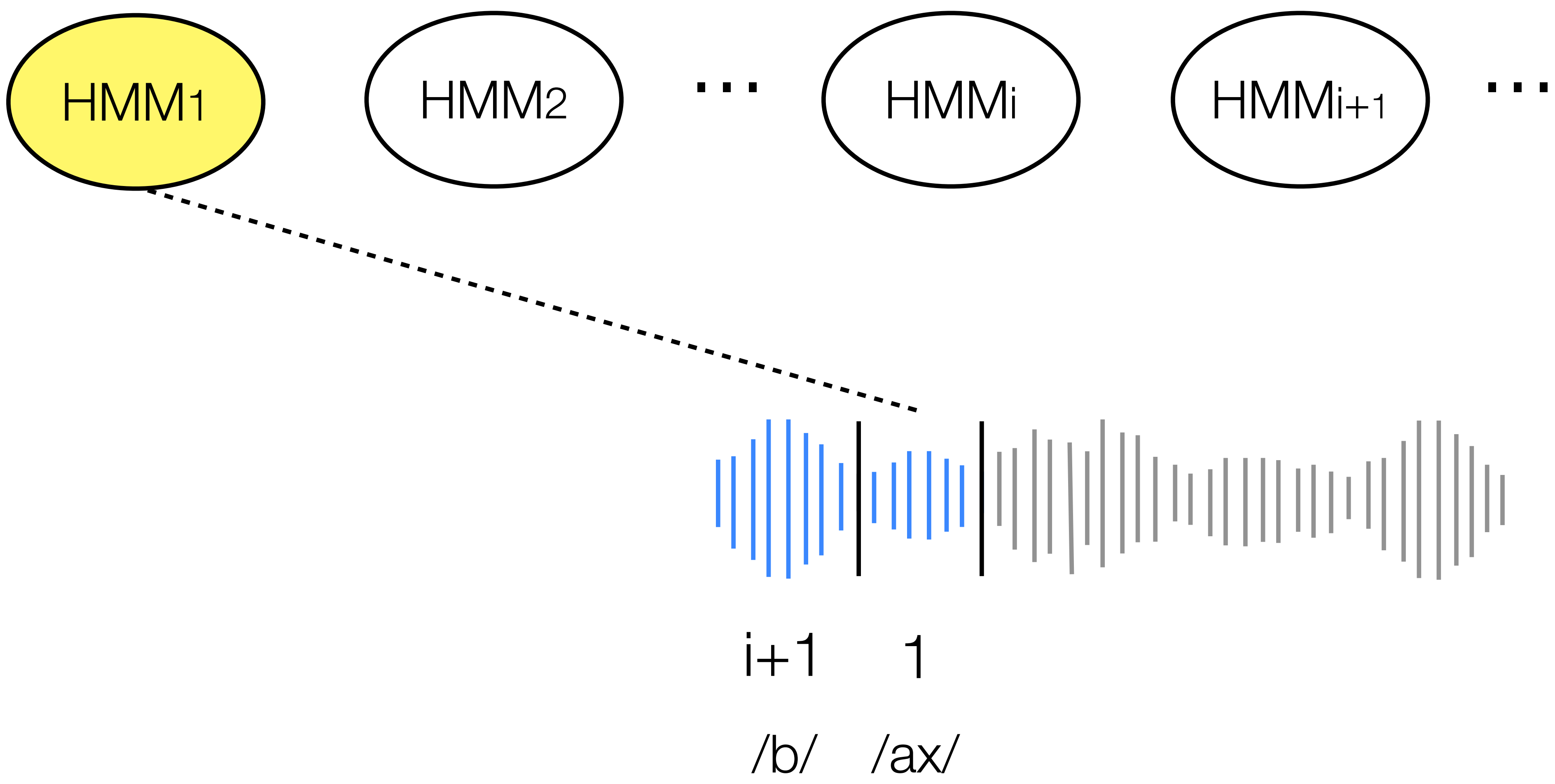
Generative Story

- A simple explanation of how a spoken utterance is generated



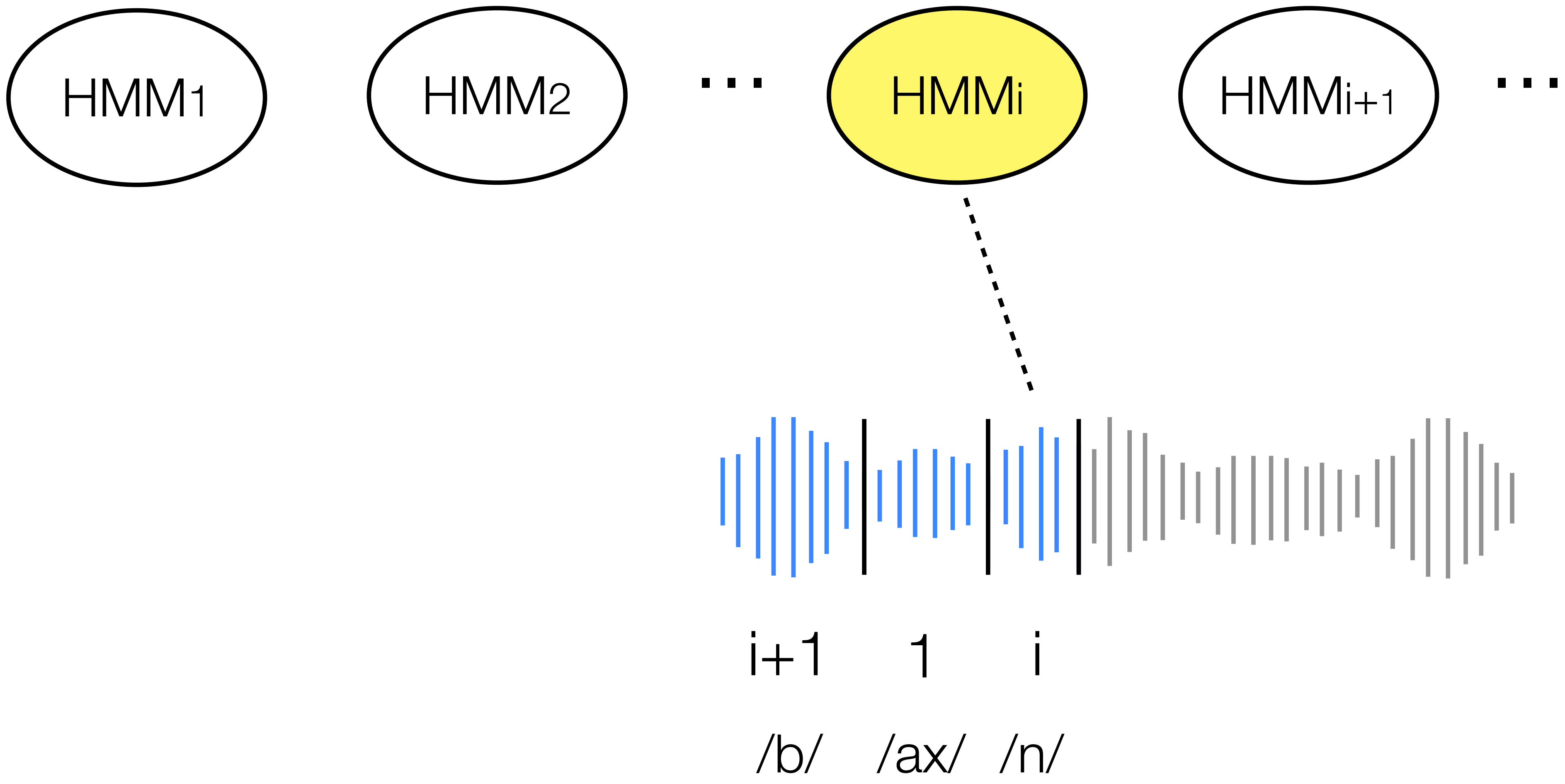
Generative Story

- A simple explanation of how a spoken utterance is generated



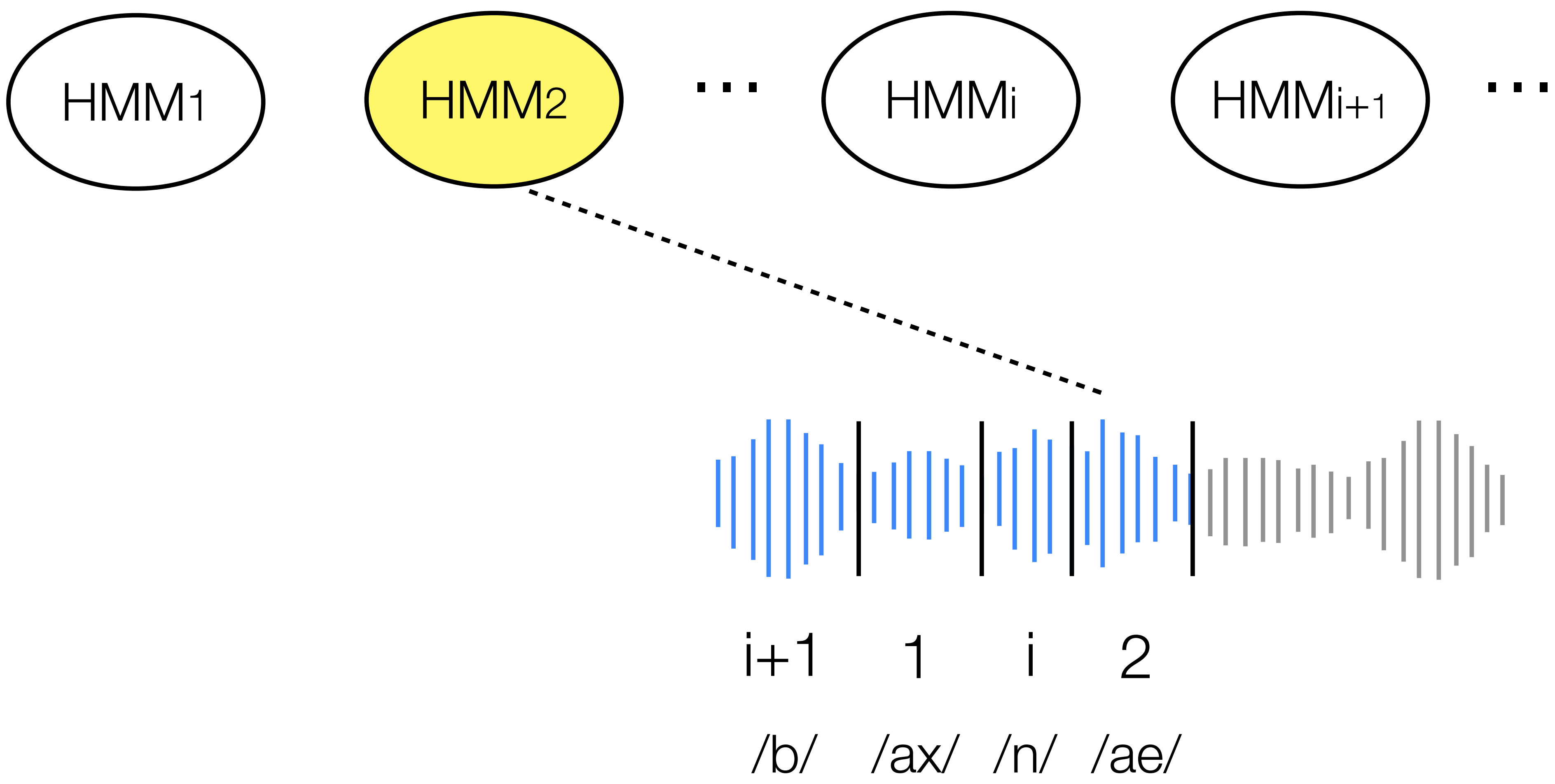
Generative Story

- A simple explanation of how a spoken utterance is generated



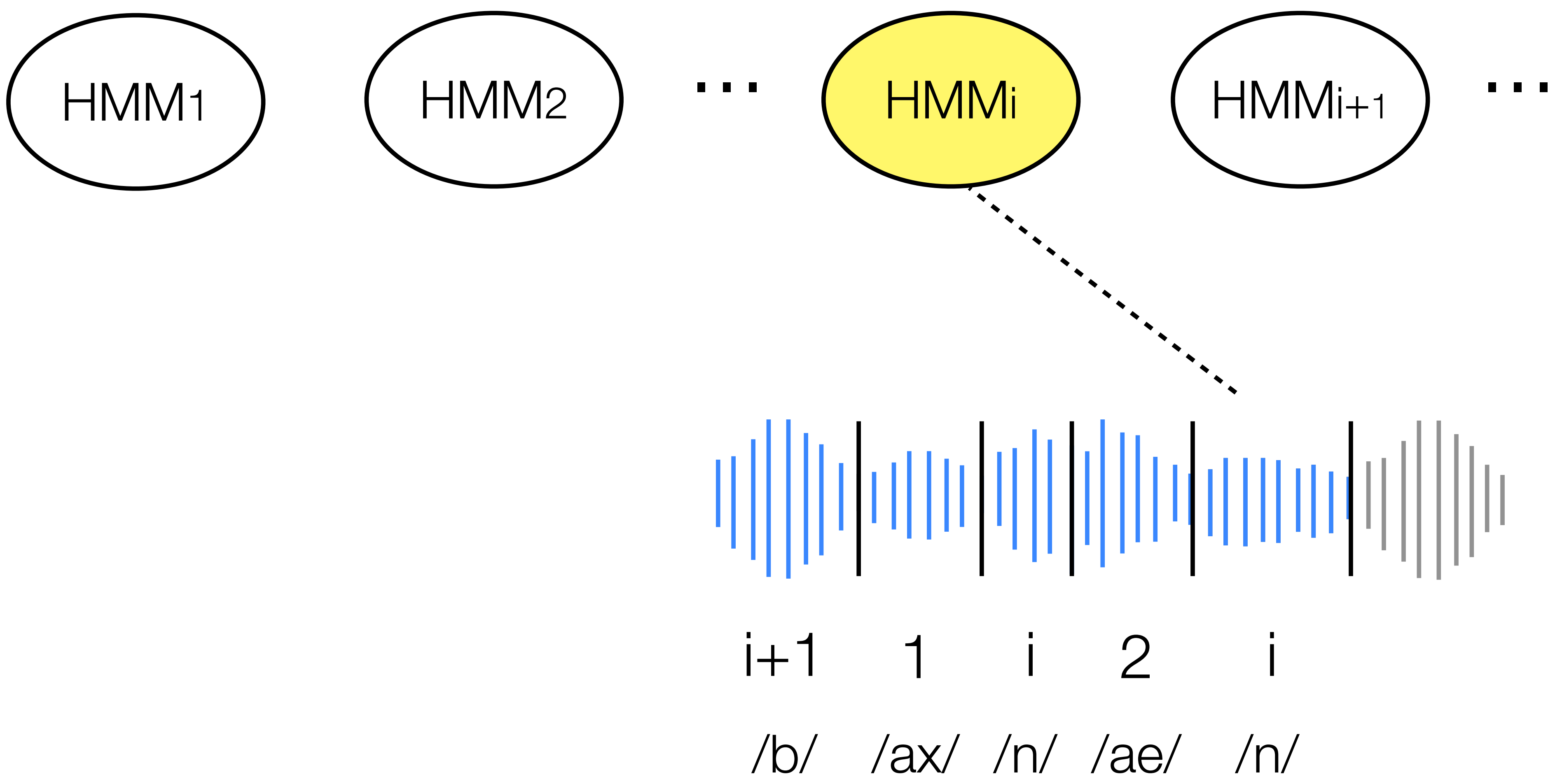
Generative Story

- A simple explanation of how a spoken utterance is generated



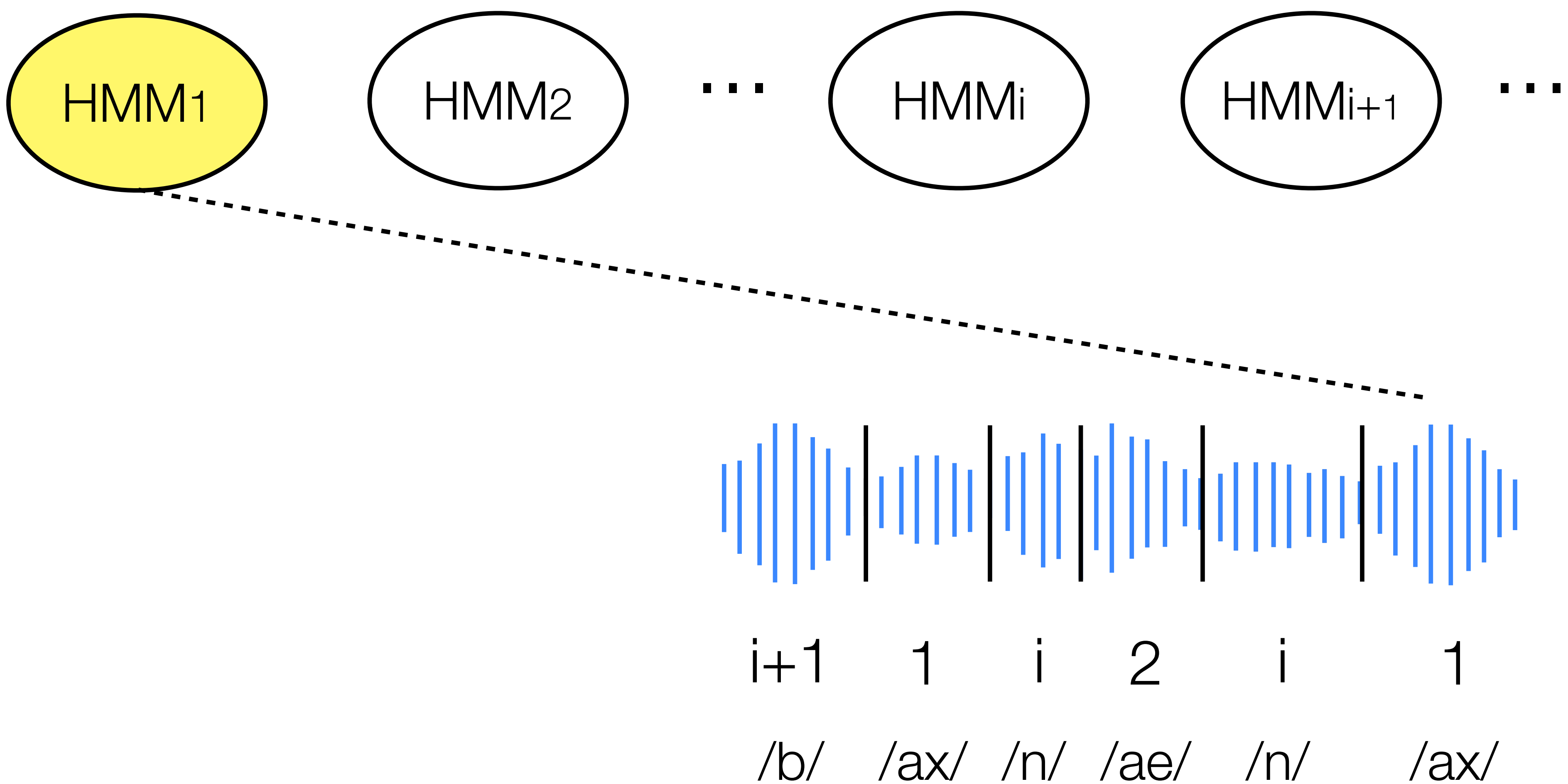
Generative Story

- A simple explanation of how a spoken utterance is generated



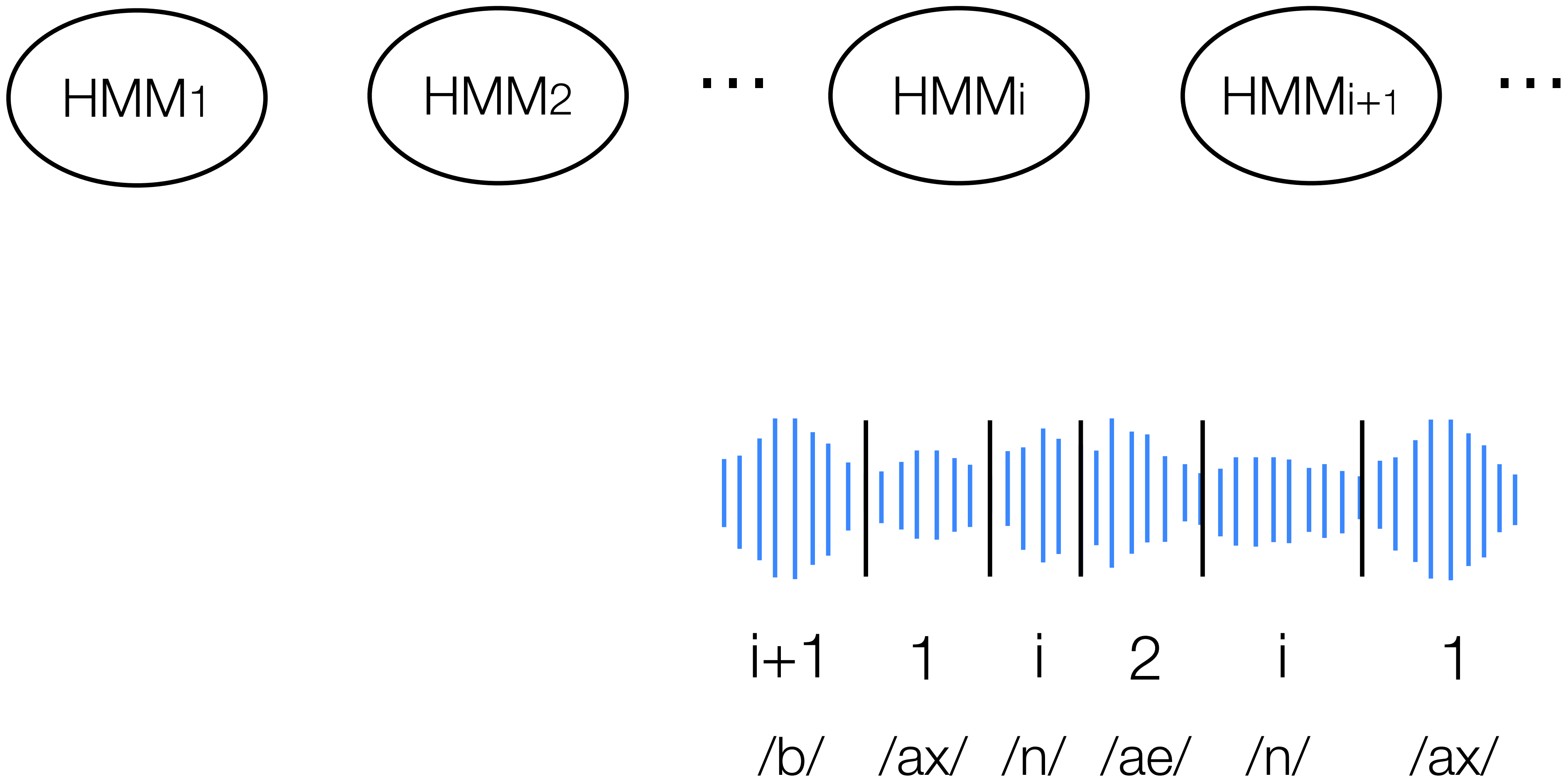
Generative Story

- A simple explanation of how a spoken utterance is generated



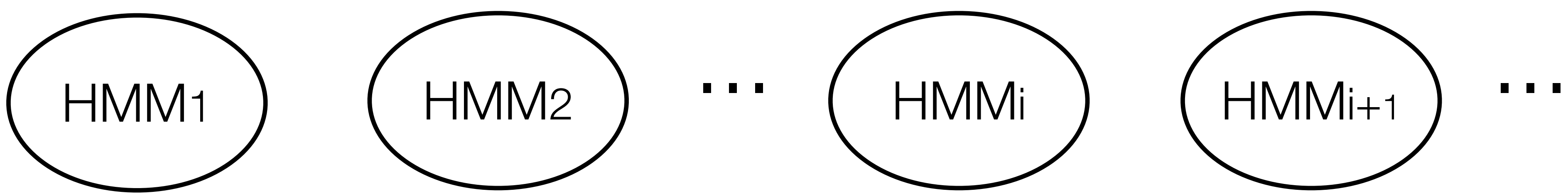
Generative Story

- A simple explanation of how a spoken utterance is generated

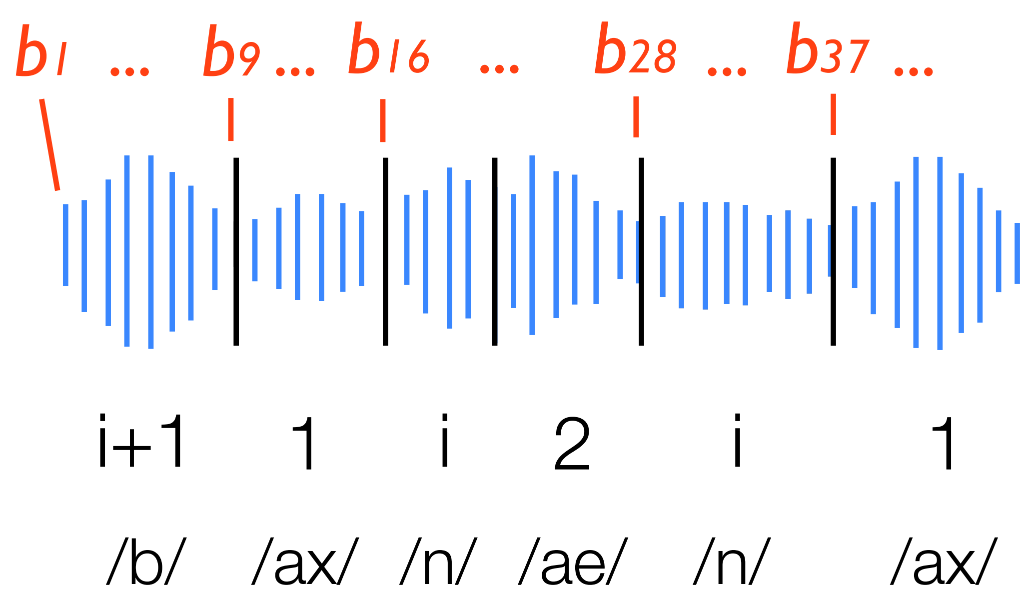


Generative Story

- A simple explanation of how a spoken utterance is generated

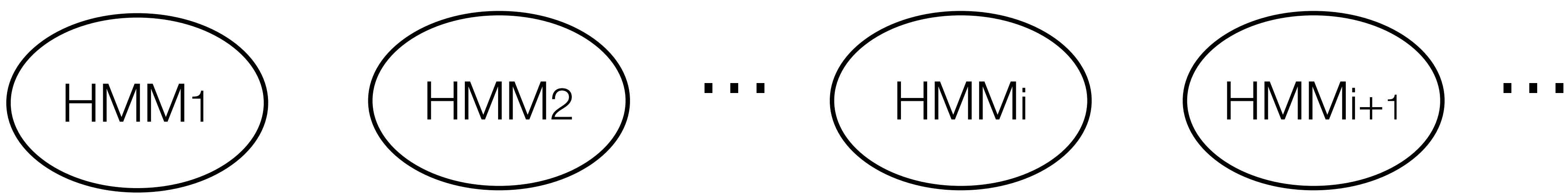


- Main latent variables
  - Phone boundaries (***b***)

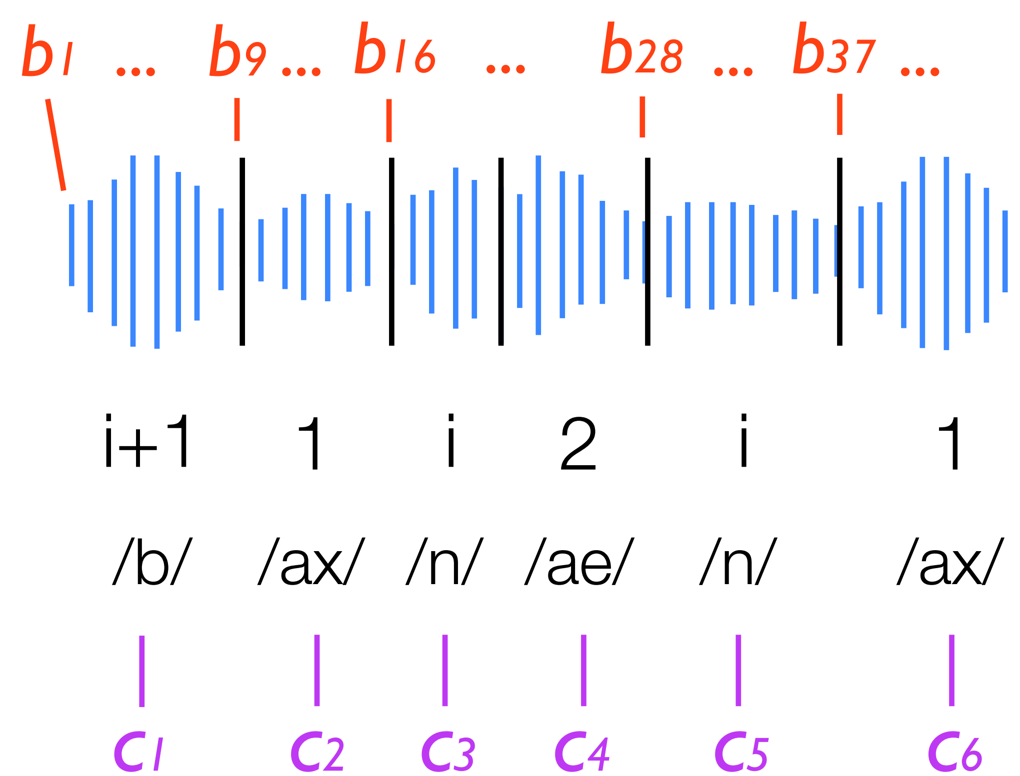


Generative Story

- A simple explanation of how a spoken utterance is generated

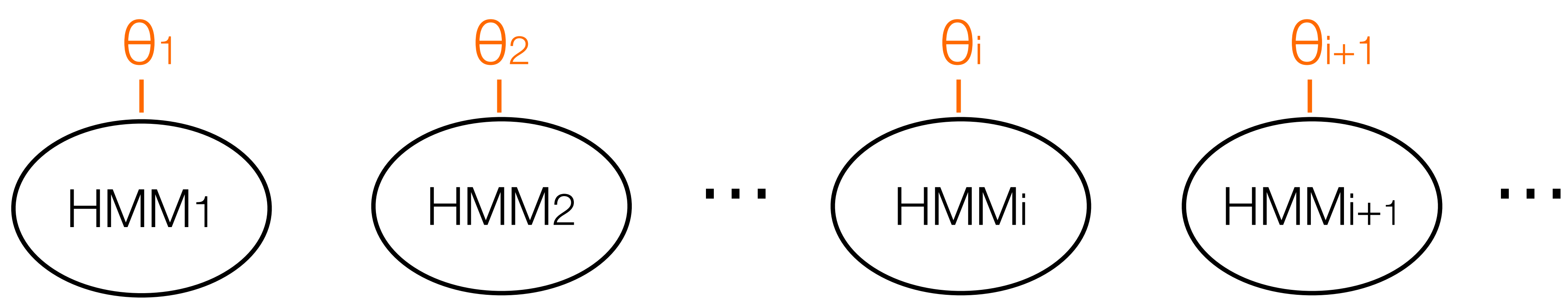


- Main latent variables
  - Phone boundaries (***b***)
  - Cluster labels (***c***)

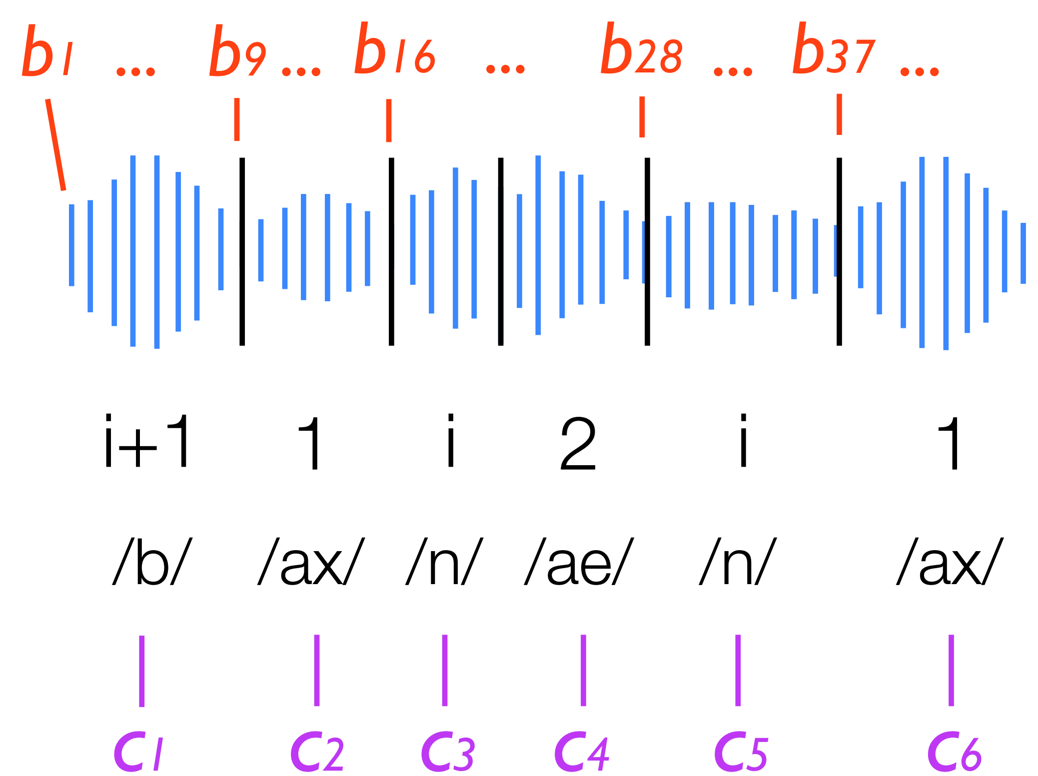


Generative Story

- A simple explanation of how a spoken utterance is generated

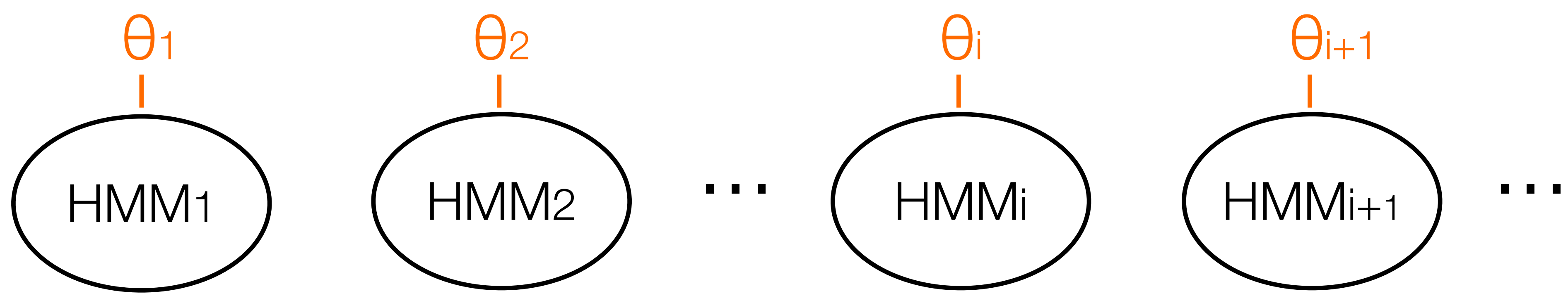


- Main latent variables
  - Phone boundaries (***b***)
  - Cluster labels (***c***)
  - HMM parameters (**θ**)

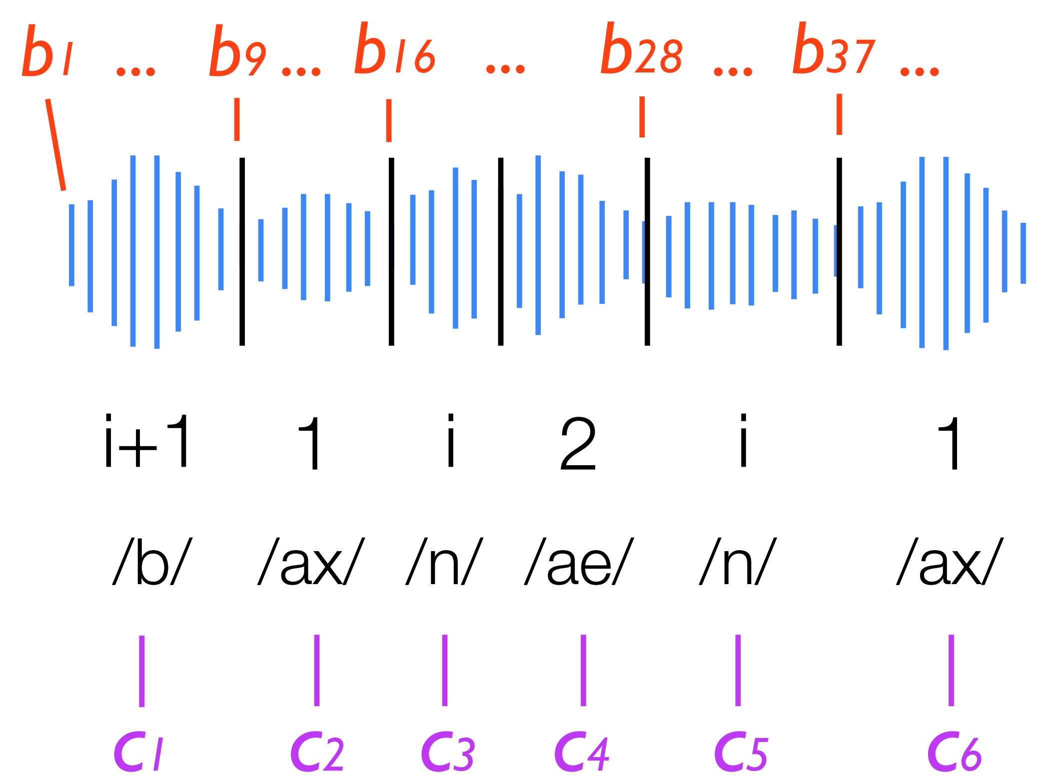


Generative Story

- A simple explanation of how a spoken utterance is generated



- Main latent variables
  - Phone boundaries (***b***)
  - Cluster labels (***c***)
  - HMM parameters (**θ**)
  - # of HMMs



Unknown Number of HMMs

- An unknown set of phone units

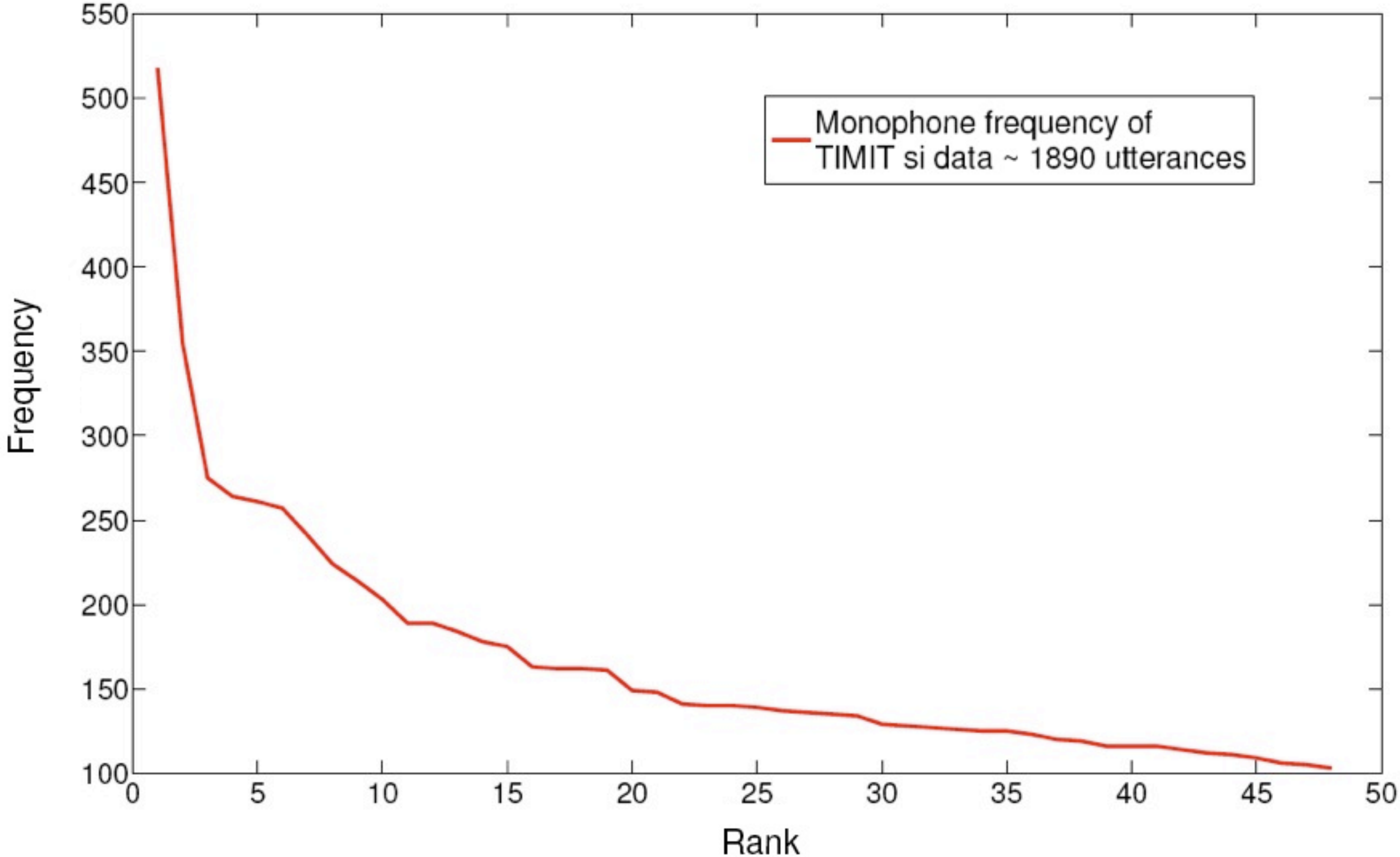
Unknown Number of HMMs

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

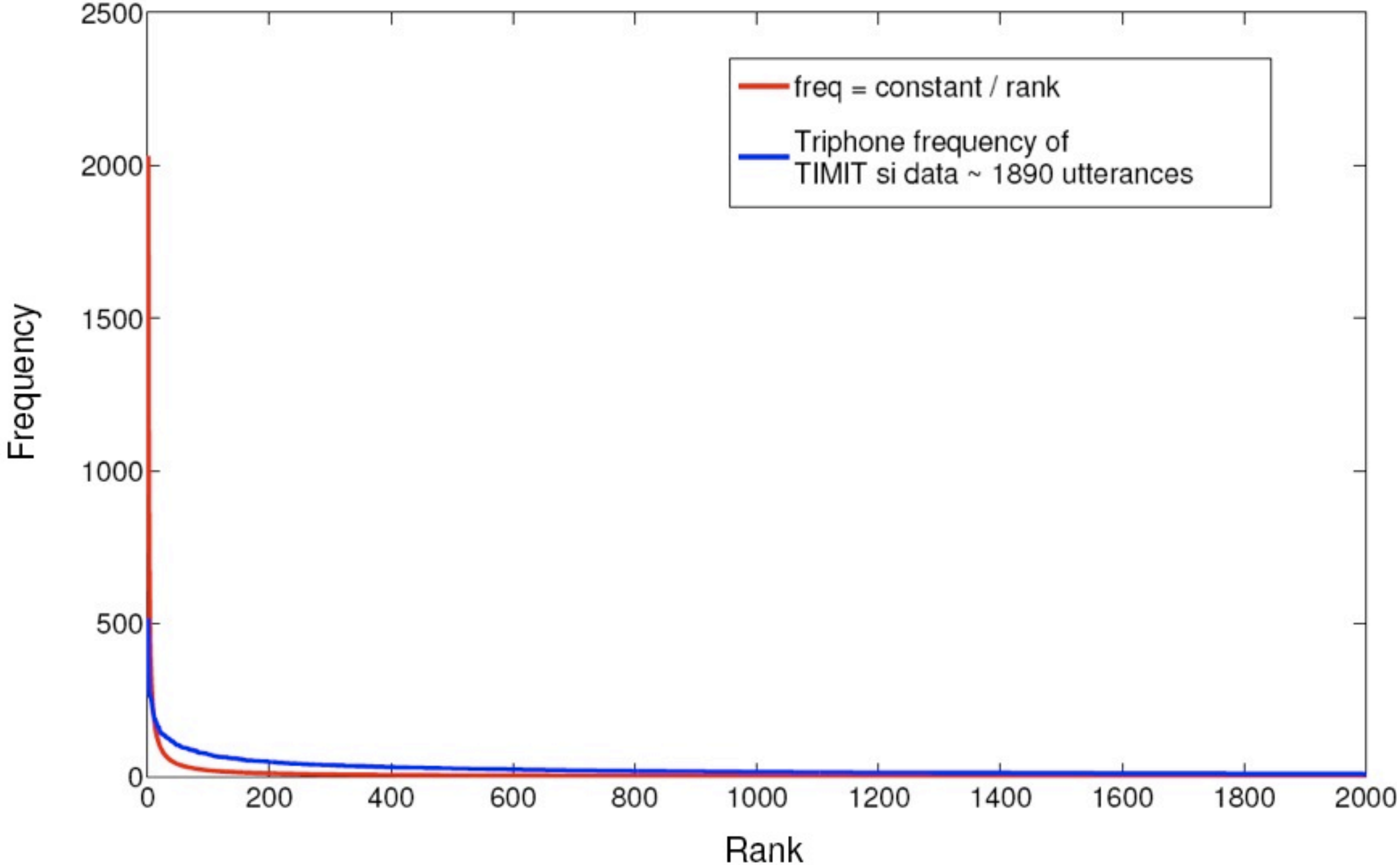
Unknown Number of HMMs

- 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



Phone Frequency -- Triphone



Unknown Number of HMMs

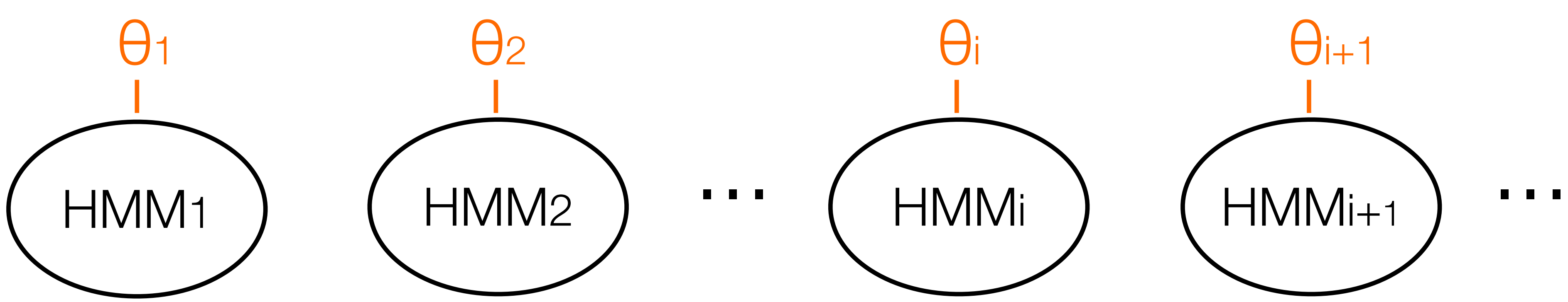
- An unknown set of phone units
  - Impose a Dirichlet Process prior to guide inference on the number of HMMs
- Is Dirichlet process (DP) a proper prior for this task?
  - Does phone frequency inherit power law?

Unknown Number of HMMs

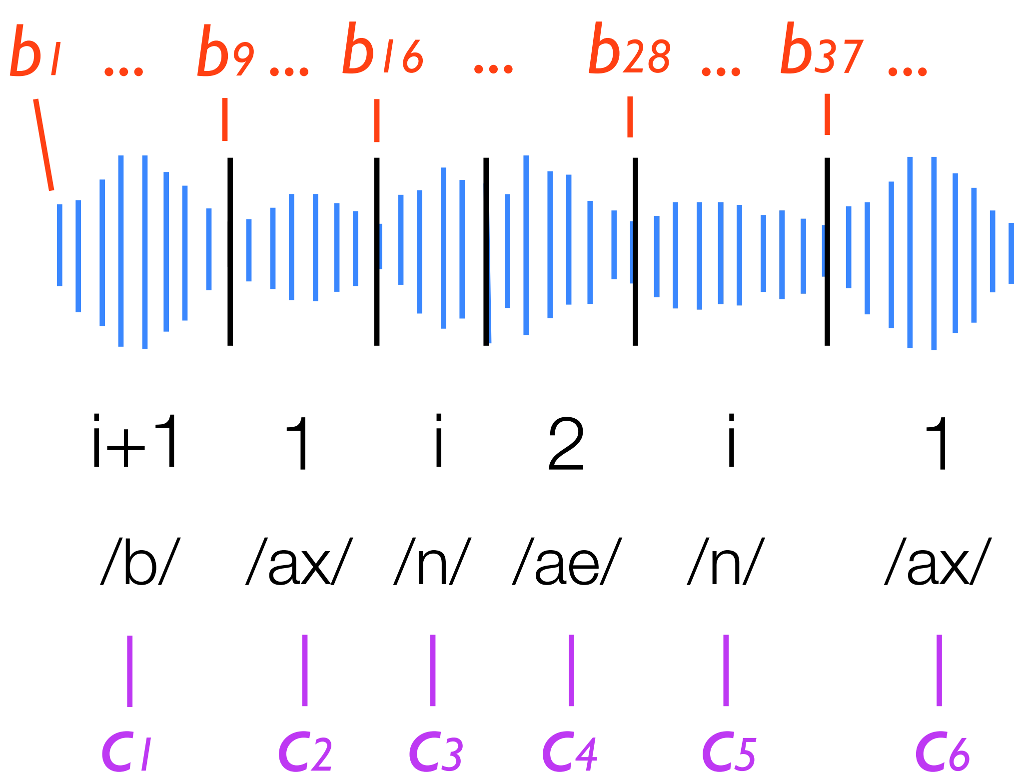
- An unknown set of phone units
  - Impose a Dirichlet Process prior to guide inference on the number of HMMs
- Is Dirichlet process (DP) a proper prior for this task?
  - Does phone frequency inherit power law?
  - DP should be a reasonable prior to start with

Generative Story

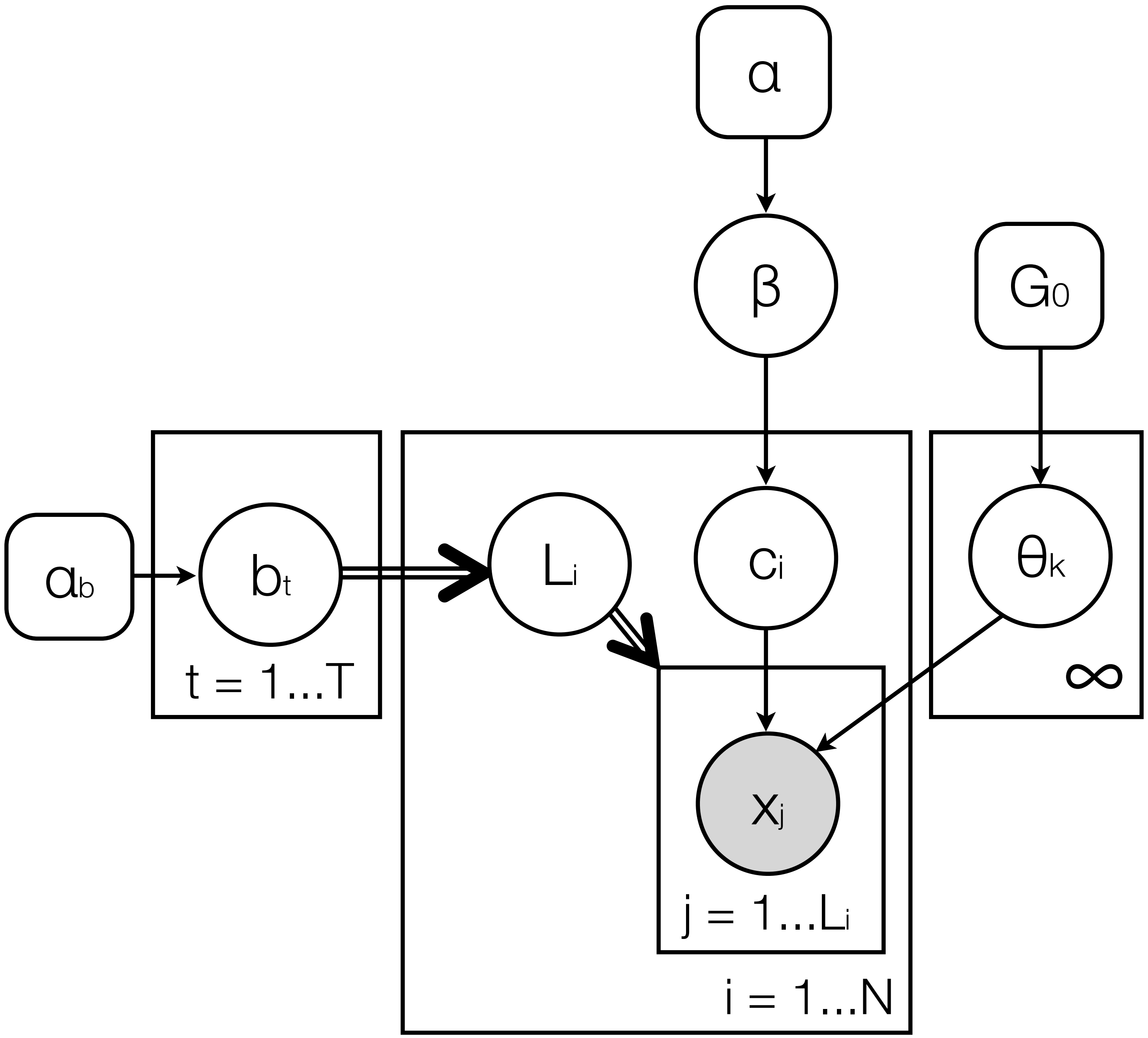
- A simple explanation of how a spoken utterance is generated



- Main latent variables
  - Phone boundaries (***b***)
  - Cluster labels (***c***)
  - HMM parameters (**θ**)
  - # of HMMs



Generative Model



α: concentration parameter of DP

G_0: base distribution of DP

β~GEM(α)

$α_b$: prior for b_t

x: observations

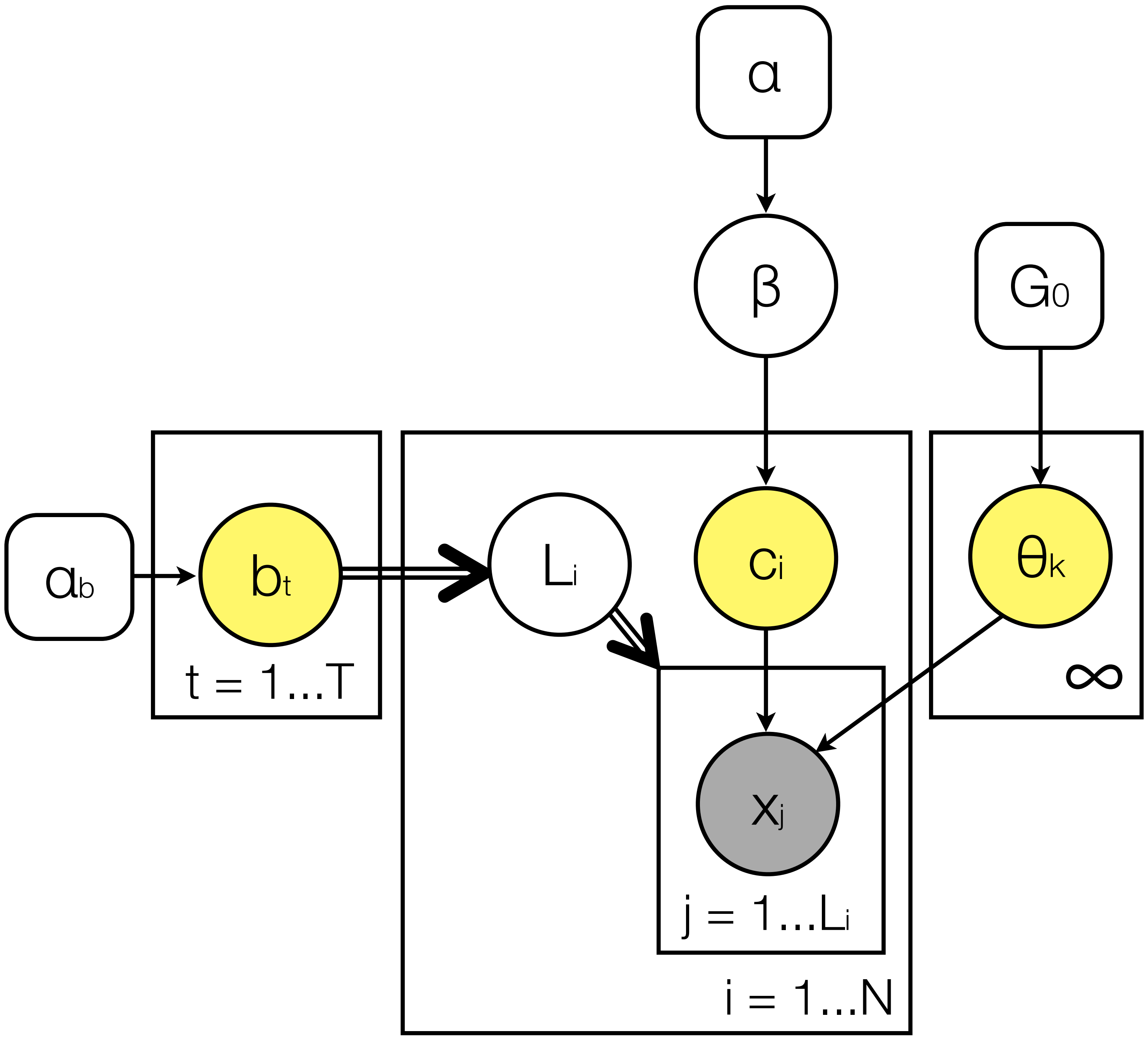
L_i: length of the i-th segment

N: total number of segments

T: total number of frames

$\Longrightarrow$ deterministic relation

Generative Model



α: concentration parameter of DP

G_0: base distribution of DP

$\beta \sim GEM(\alpha)$

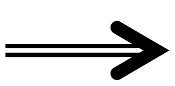
α_b: prior for b_t

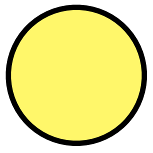
x: observations

L_i: length of the i-th segment

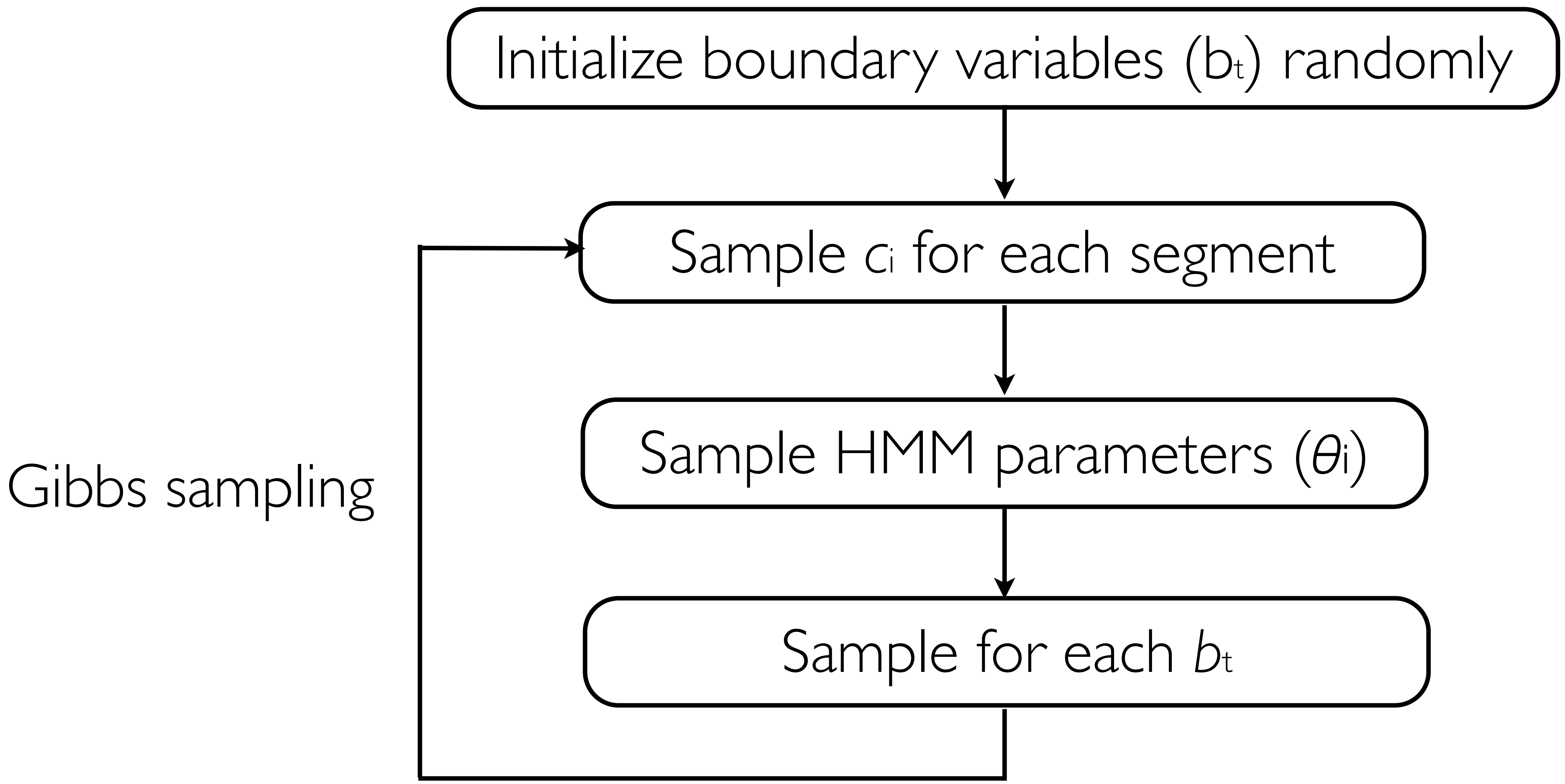
N: total number of segments

T: total number of frames


deterministic relation

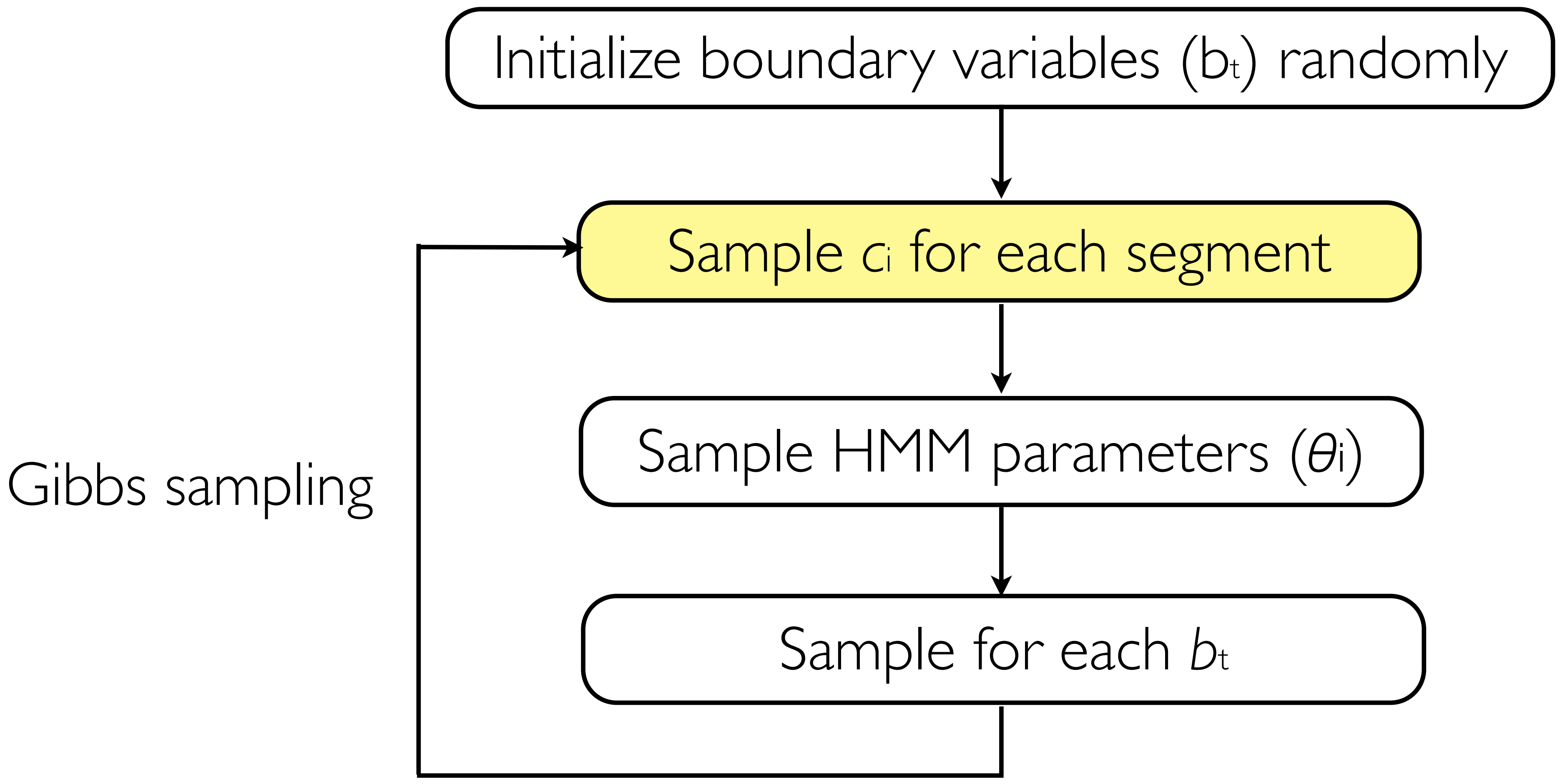

latent variables that will be inferred

Inference Procedure



- **Iterate n times**
  - n = 20,000 in our experiments

Inference Procedure



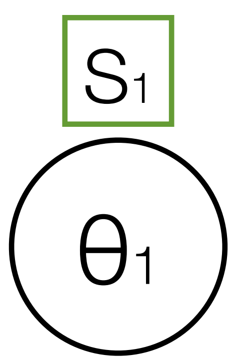
- **Iterate n times**
  - n = 20,000 in our experiments

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*)

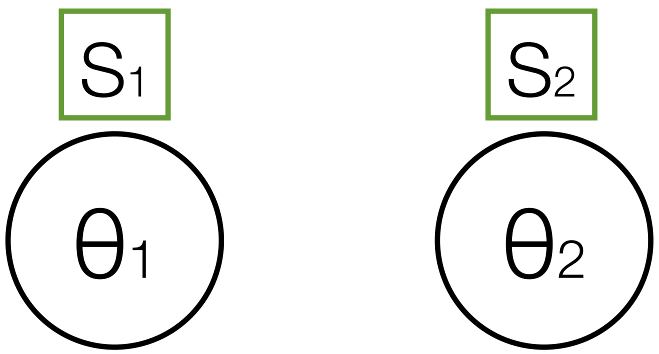
- A Chinese restaurant process representation
  - Each table is a phonetic unit
  - Each speech segment is a customer $s_i = [x_t, x_{t+1}, \ldots x_{t+L_i}]$



$c_1 = 1$

DP as a Prior for Cluster Labels (c)

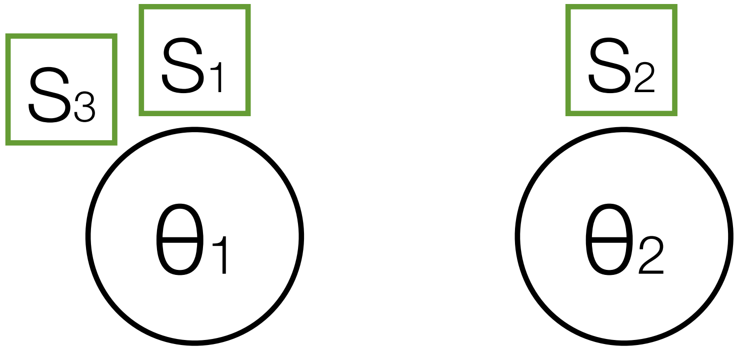
- A Chinese restaurant process representation
  - Each table is a phonetic unit
  - Each speech segment is a customer $s_i = [x_t, x_{t+1}, \ldots x_{t+L_i}]$



$c_1 = 1$ $c_2 = 2$

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 $s_i = [x_t, x_{t+1}, \ldots x_{t+L_i}]$

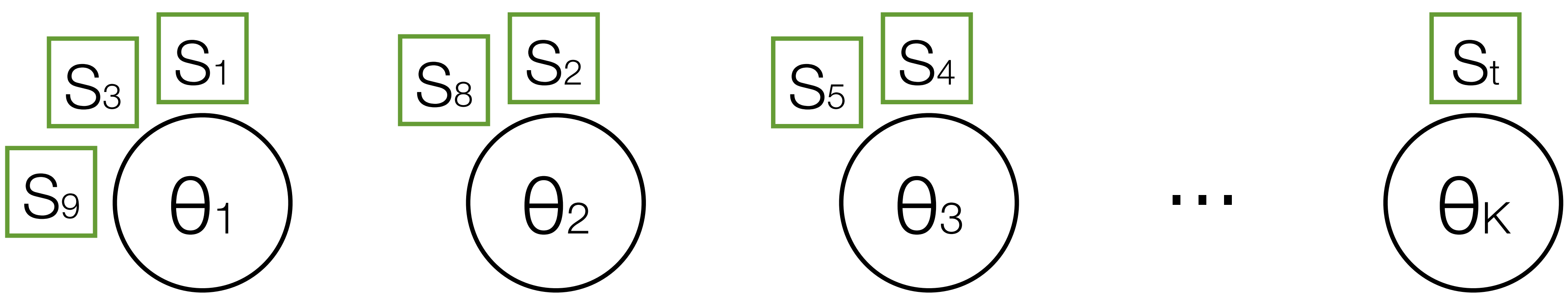


$c_1 = 1$ $c_2 = 2$

$c_3 = 1$

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 $s_i = [x_t, x_{t+1}, \ldots x_{t+L_i}]$

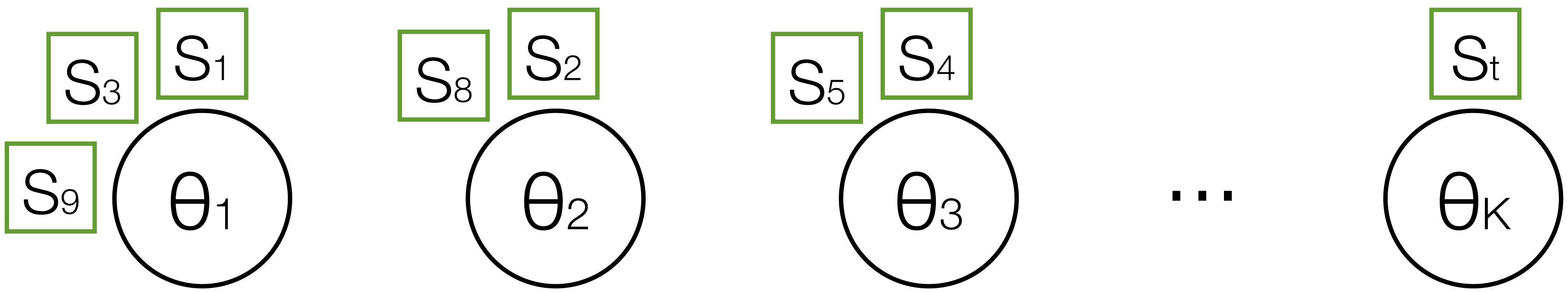


$c_1 = 1$ $c_2 = 2$ $c_4 = 3$ $c_t = K$

$c_3 = 1$ $c_8 = 2$ $c_5 = 3$

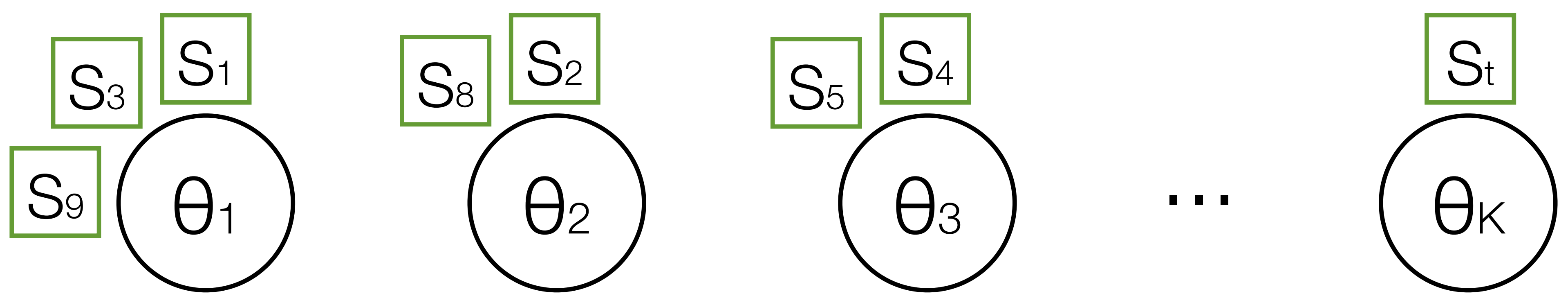
$c_9 = 1$

Posterior Distribution for c_i



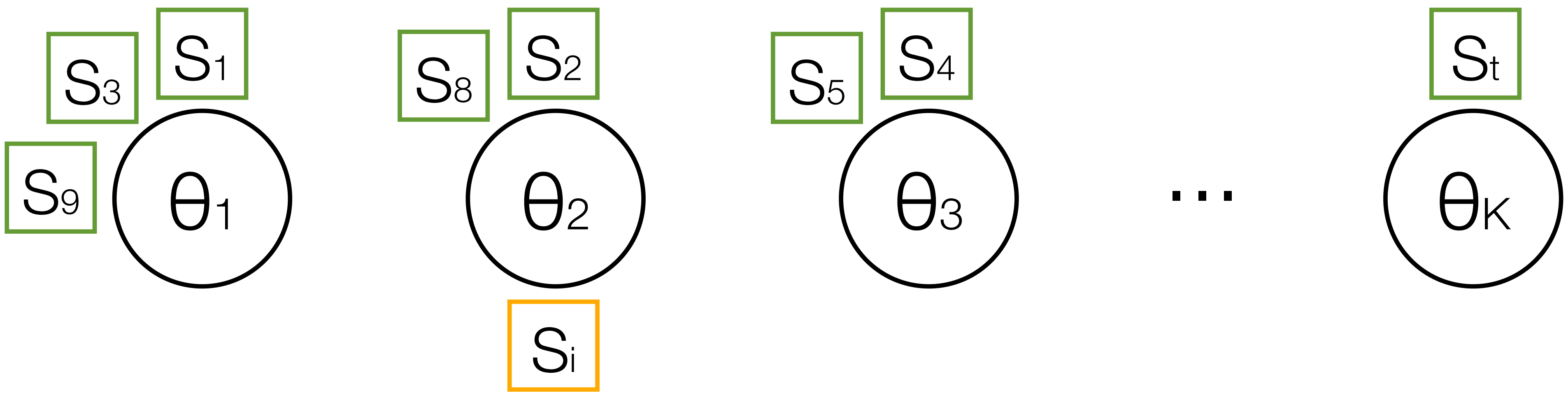
- For a new segment (s_i), the posterior probability distribution of c_i :

Posterior Distribution for c_i



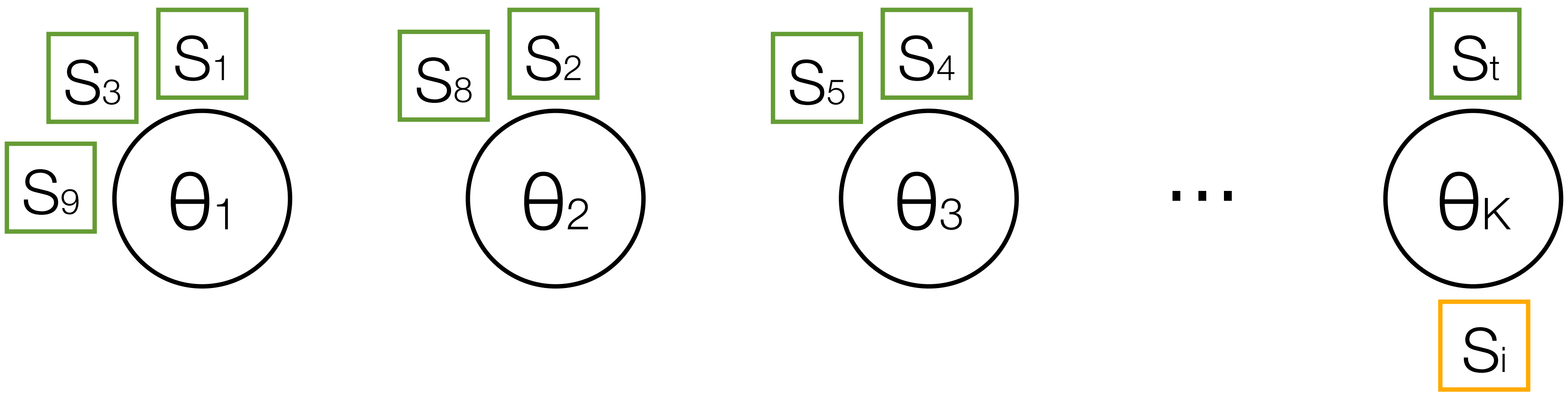
- For a new segment (s_i), the posterior probability distribution of c_i :
  - s_i sits at an occupied table ⟶ s_i is not a new phone

Posterior Distribution for c_i



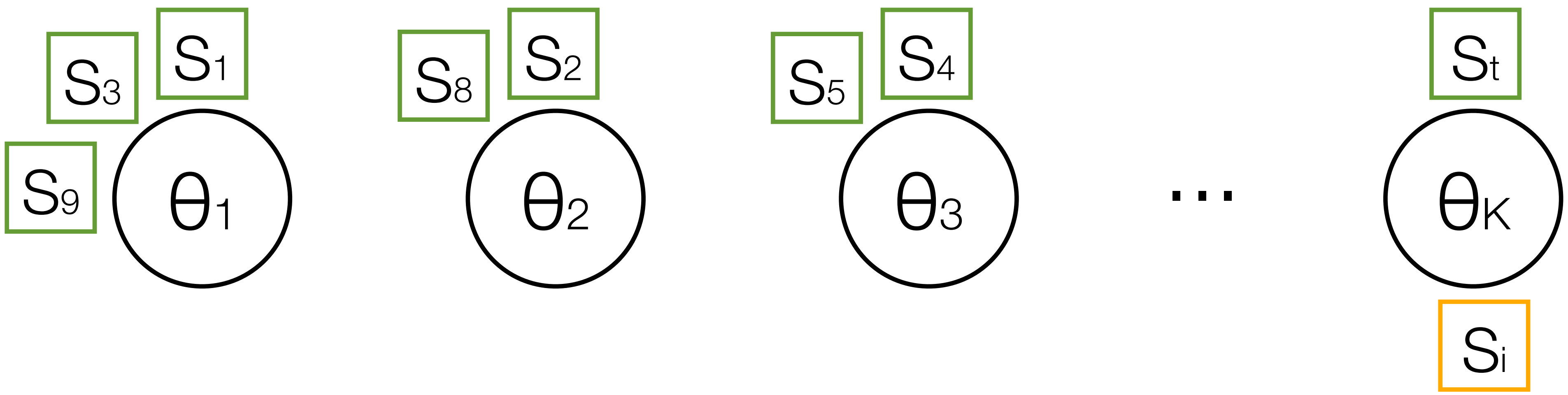
- **For a new segment (s_i), the posterior probability distribution of c_i :**
  - s_i sits at an occupied table ⟶ s_i is not a new phone

Posterior Distribution for c_i



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

Posterior Distribution for c_i



- For a new segment (s_i), the posterior probability distribution of c_i :

  - s_i sits at an occupied table $\longrightarrow$ s_i is not a new phone

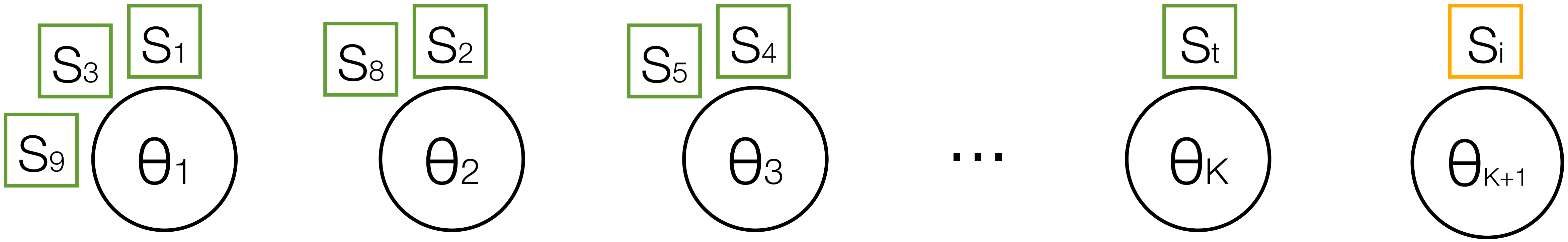
$$\underbrace{p(c_i = k, 1 \le k \le K \mid \cdots)}_{\text{posterior probability}} \propto \underbrace{\frac{n_k}{N - 1 + \alpha}}_{\text{DP prior}} \underbrace{p(s_i \mid \theta_k)}_{\text{likelihood}}$$

n_k : number of customers at table k

N : number of costumers seen so far

α : concentration parameter of DP

Posterior Distribution for c_i

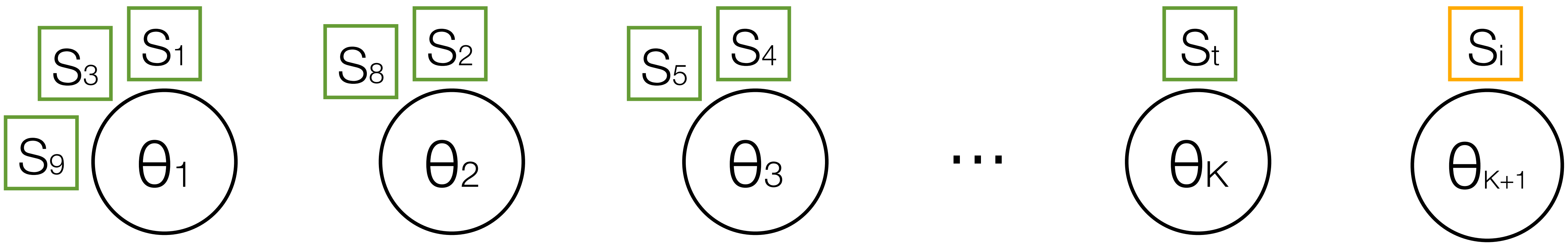


- For a new segment (s_i), the posterior probability distribution of c_i :
  - s_i sits at an occupied table $\longrightarrow$ s_i is not a new phone

$$p(c_i = k, 1 \le k \le K \mid \cdots) \propto \frac{n_k}{N-1+\alpha} p(s_i \mid \theta_k)$$

  - s_i opens a new table $\longrightarrow$ s_i is a new phone

Posterior Distribution for c_i



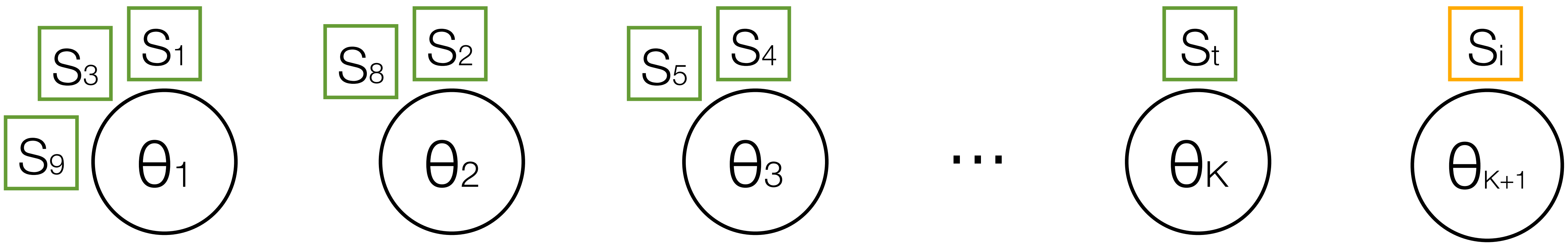
- For a new segment (s_i), the posterior probability distribution of c_i :
  - s_i sits at an occupied table ⟶ s_i is not a new phone

$$p(c_i = k, 1 \le k \le K \mid \cdots) \propto \frac{n_k}{N-1+\alpha} p(s_i \mid \theta_k)$$

  - s_i opens a new table ⟶ s_i is a new phone

$$p(c_i = K+1 \mid \cdots) \propto \frac{\alpha}{N-1+\alpha} \int_\theta p(s_i \mid \theta) d\theta$$

Posterior Distribution for c_i



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

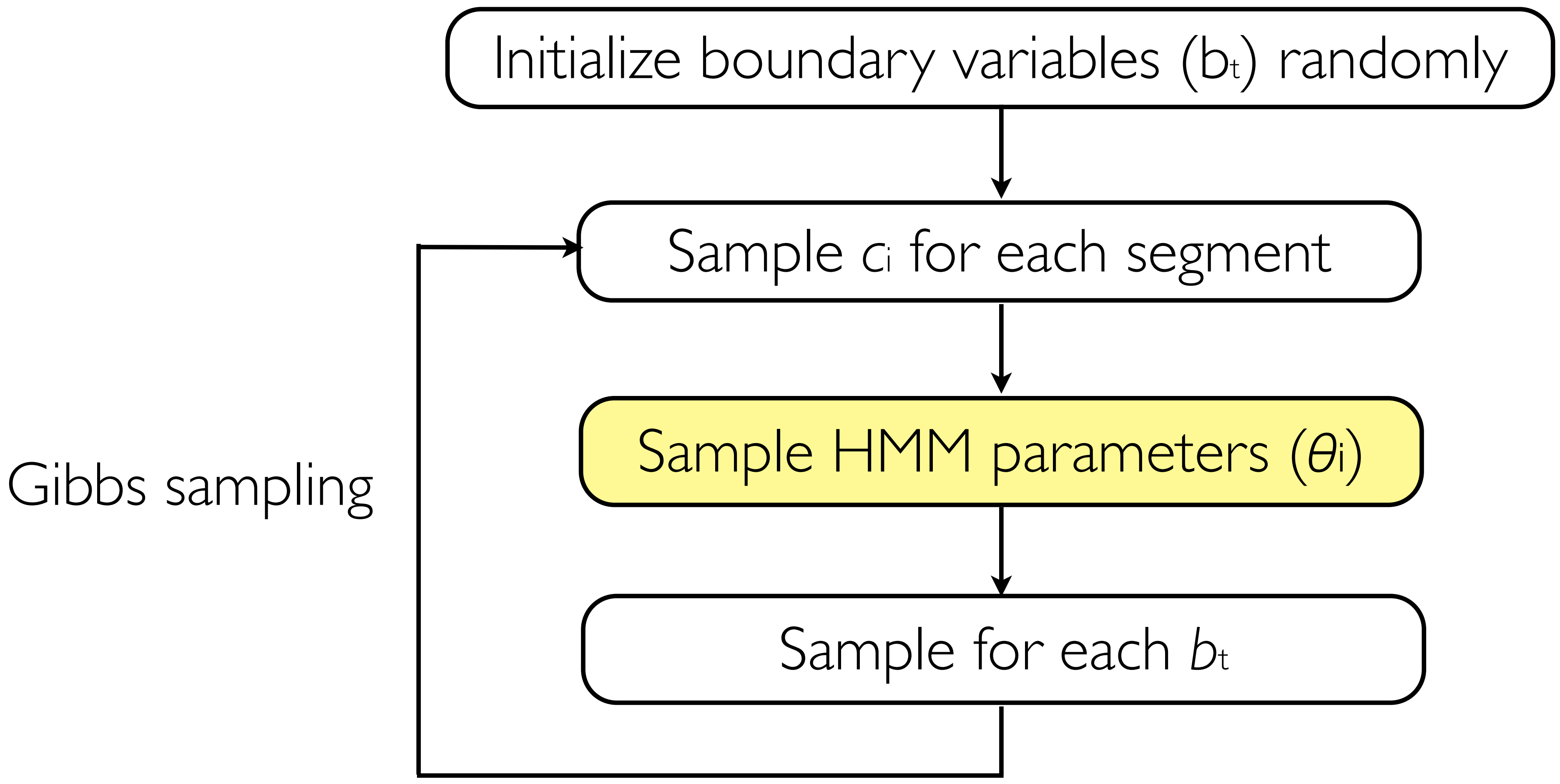
$$p(c_i = k, 1 \leq k \leq K \mid \cdots) \propto \frac{n_k}{N-1+\alpha} p(s_i \mid \theta_k)$$

  - s_i opens a new table $\longrightarrow$ s_i is a new phone

$$p(c_i = K+1 \mid \cdots) \propto \frac{\alpha}{N-1+\alpha} \int_\theta p(s_i \mid \theta) d\theta$$

Generate a sample for c_i

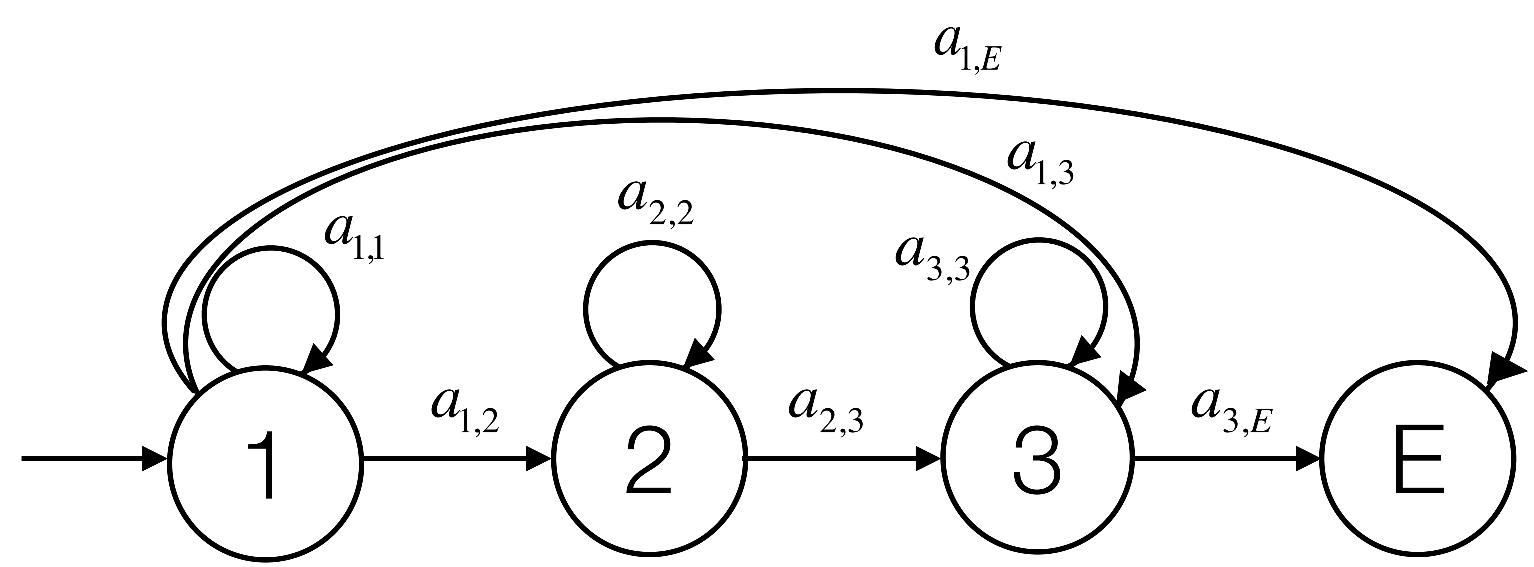
Inference Procedure



- **Iterate n times**
  - n = 20,000 in our experiments

Inference for HMM Parameters (θ)

- **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

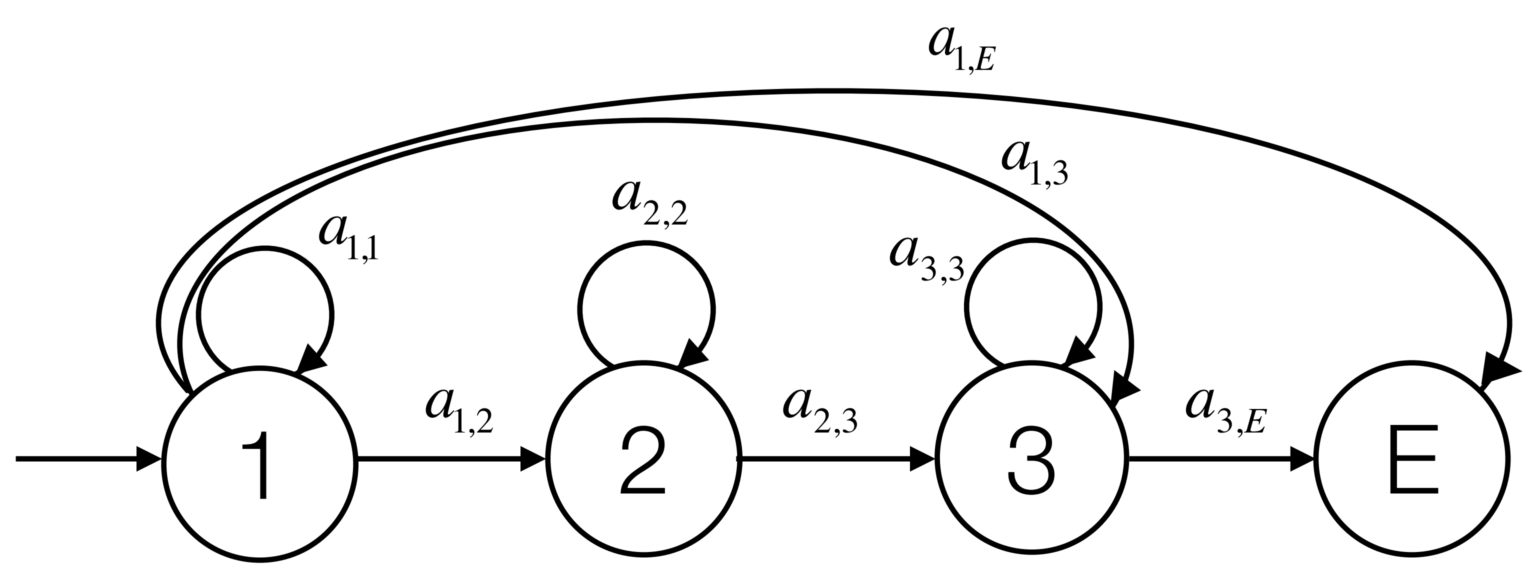


$$\sum_{i=1}^{8} w_{1,i} N\left(u_{1,i}, \sigma_{1,i}^2\right) \quad \sum_{i=1}^{8} w_{2,i} N\left(u_{2,i}, \sigma_{2,i}^2\right) \quad \sum_{i=1}^{8} w_{3,i} N\left(u_{3,i}, \sigma_{3,i}^2\right)$$

Inference for HMM Parameters (θ)

- **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 ($\boldsymbol{w}$)
  - Mean ($\boldsymbol{\mu}$)
  - Variance ($\boldsymbol{\sigma}^2$)



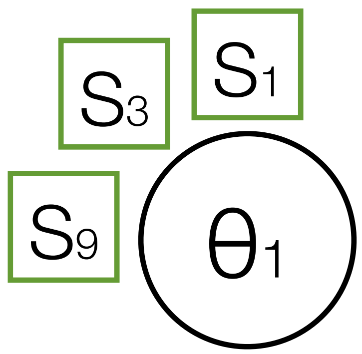
$$\sum_{i=1}^{8} w_{1,i} N\left(u_{1,i}, \sigma_{1,i}^2\right) \quad \sum_{i=1}^{8} w_{2,i} N\left(u_{2,i}, \sigma_{2,i}^2\right) \quad \sum_{i=1}^{8} w_{3,i} N\left(u_{3,i}, \sigma_{3,i}^2\right)$$

Priors and Posteriors for HMM

- Priors
  - Dirichlet distributions for transition probabilities (a) and mixture weights ($\boldsymbol{w}$)
  - Normal-gamma distributions for Gaussian parameters (μ, σ^2)

Priors and Posteriors for HMM

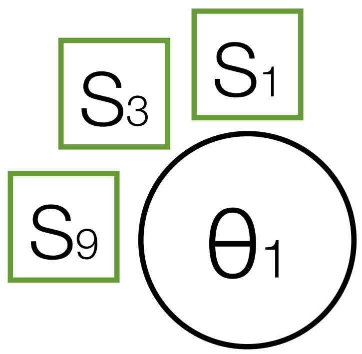
- Priors
  - Dirichlet distributions for transition probabilities (a) and mixture weights (w)
  - Normal-gamma distributions for Gaussian parameters (μ, σ^2)
- Posteriors
  - Gather relevant counts from customer segments



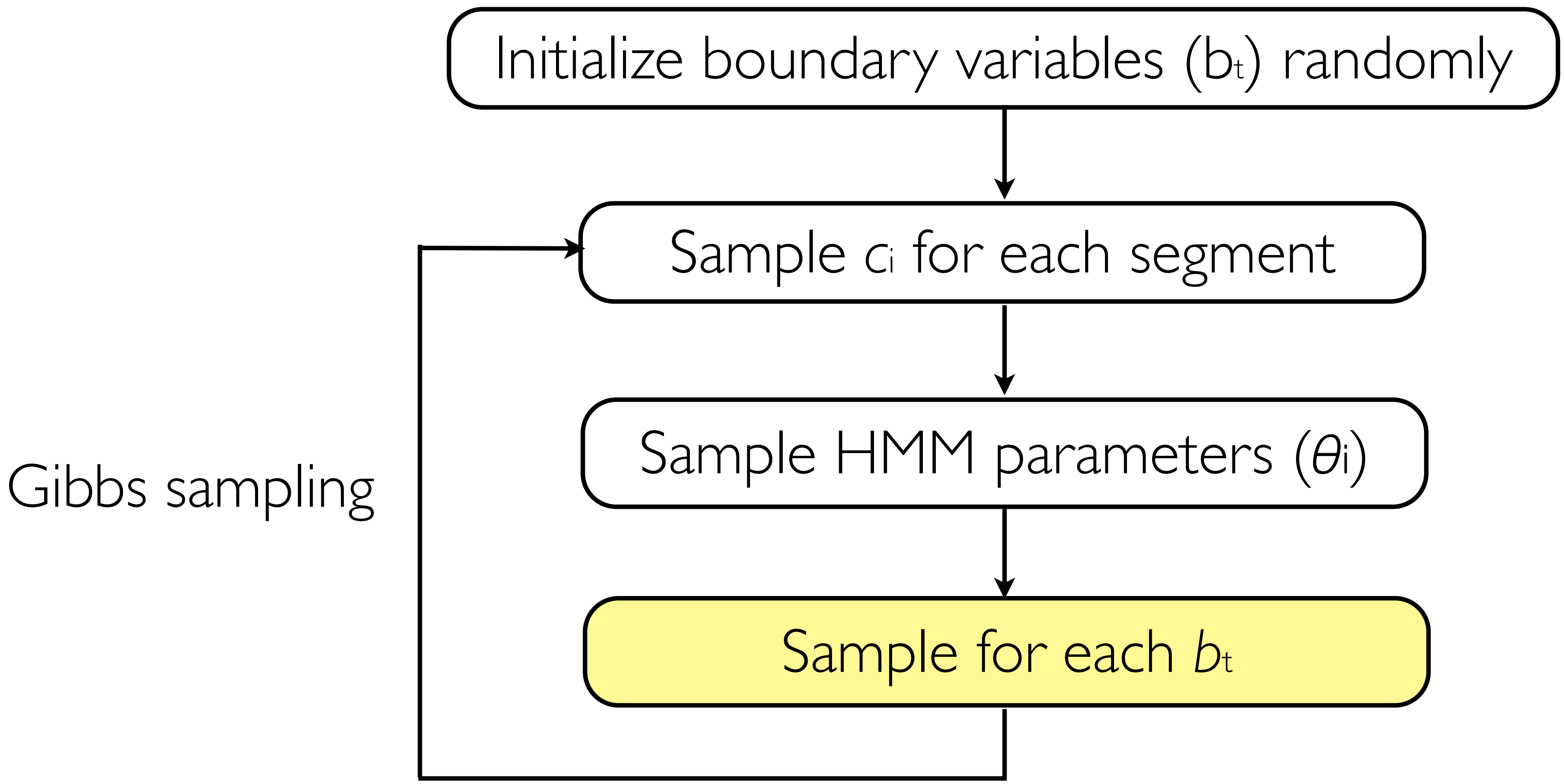
Priors and Posteriors for HMM

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

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



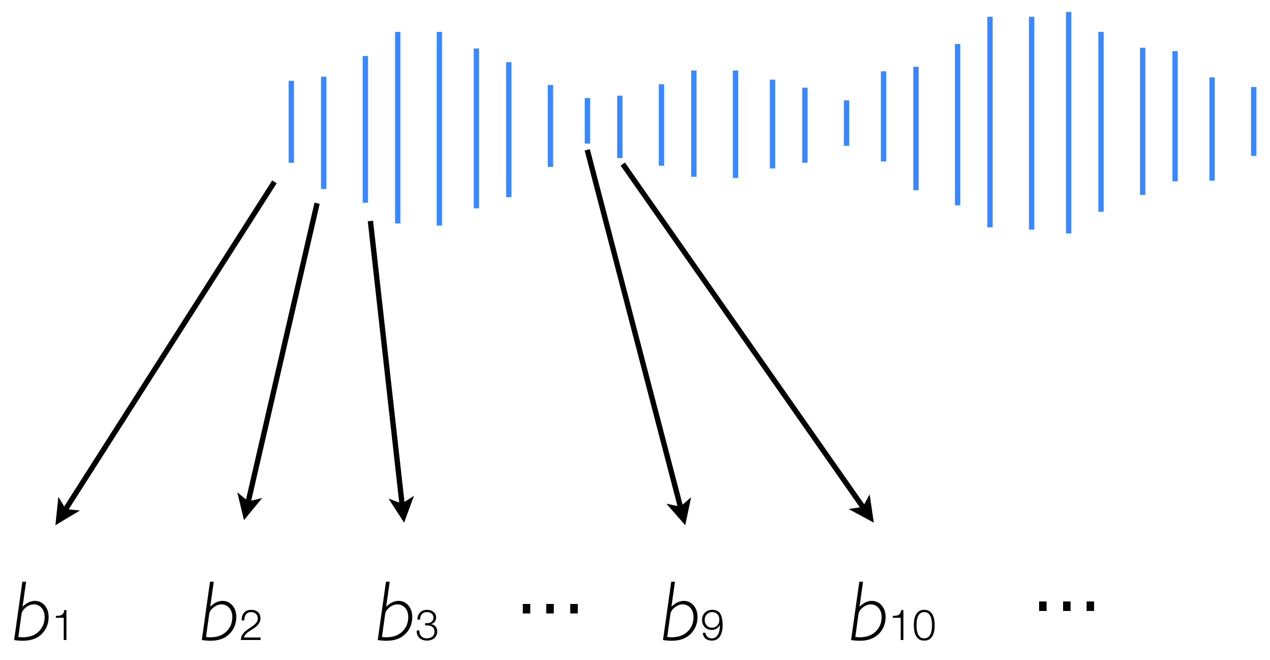
Inference Procedure



- **Iterate n times**
  - n = 20,000 in our experiments

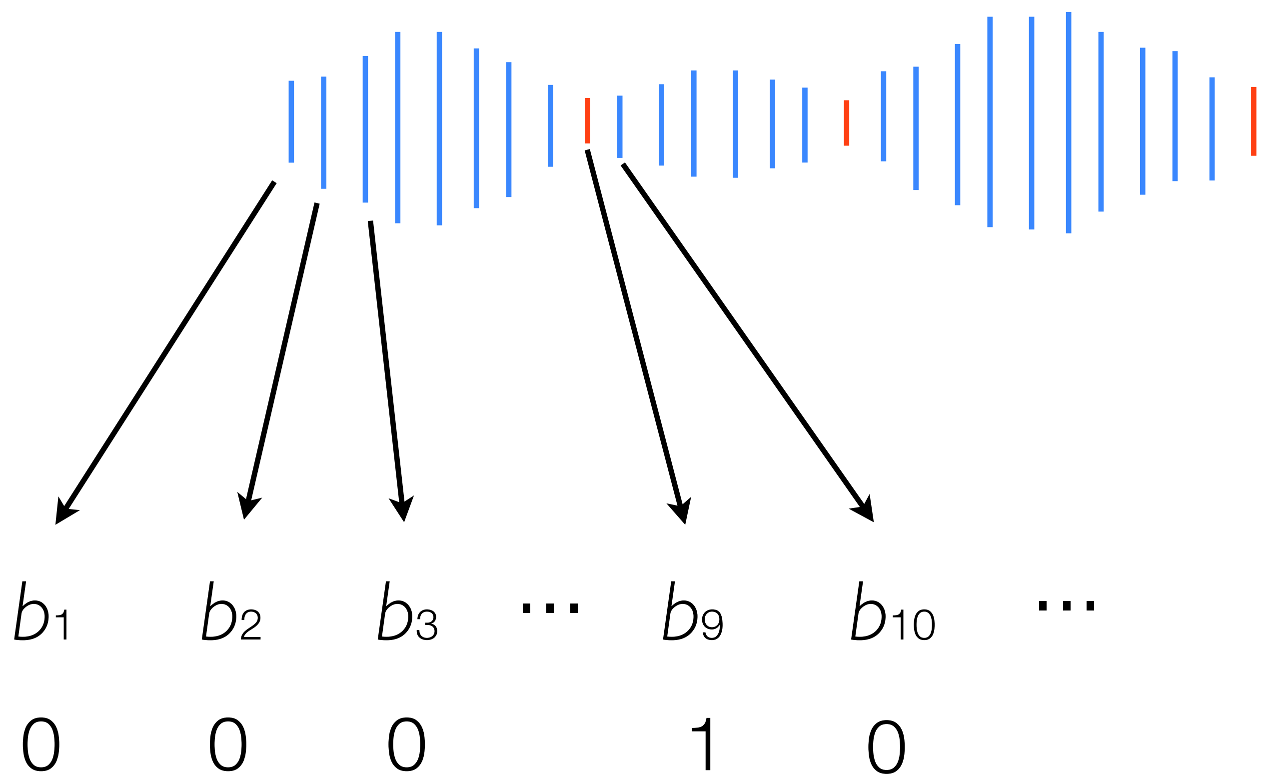
Inference on Phone Boundaries (b)

- **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

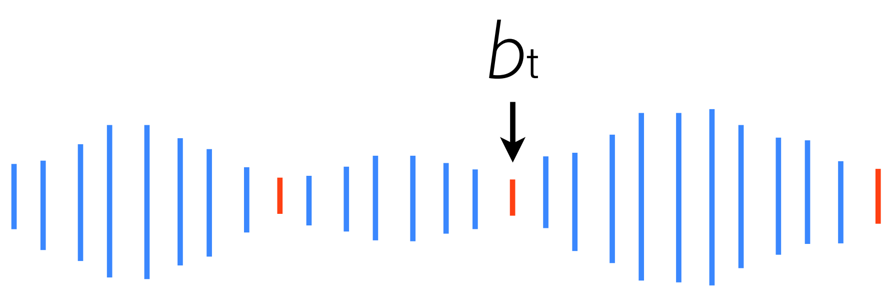


Prior and Posterior for Phone Boundaries

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

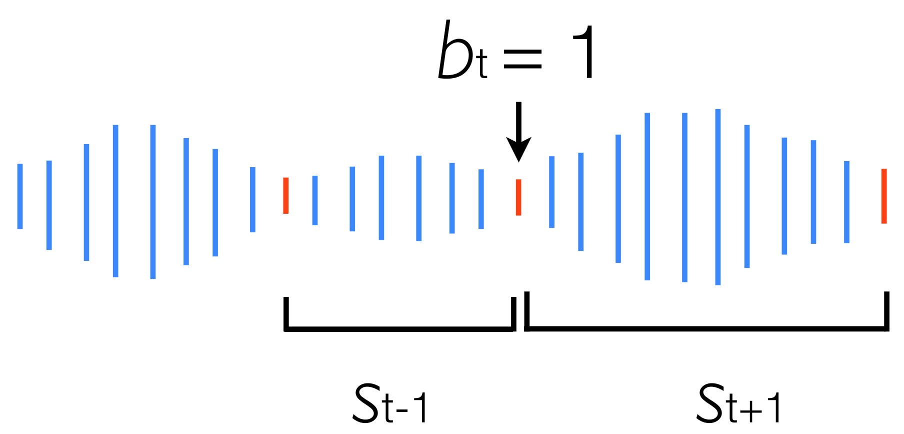
Prior and Posterior for Phone Boundaries

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



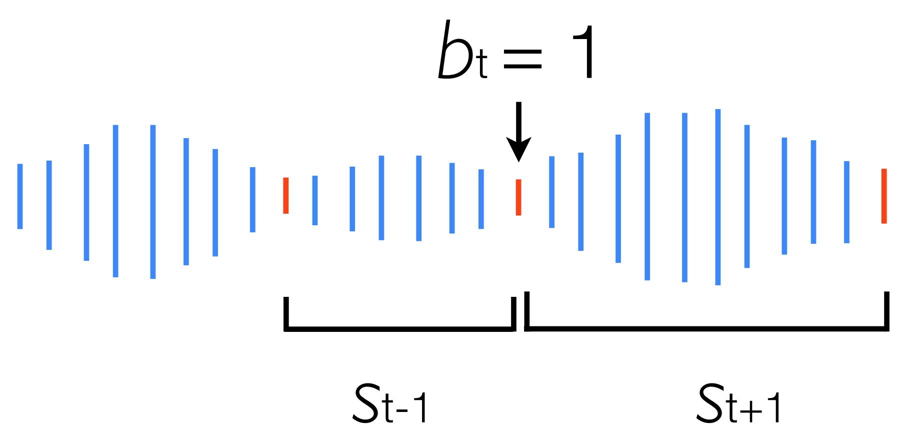
Prior and Posterior for Phone Boundaries

- 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



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



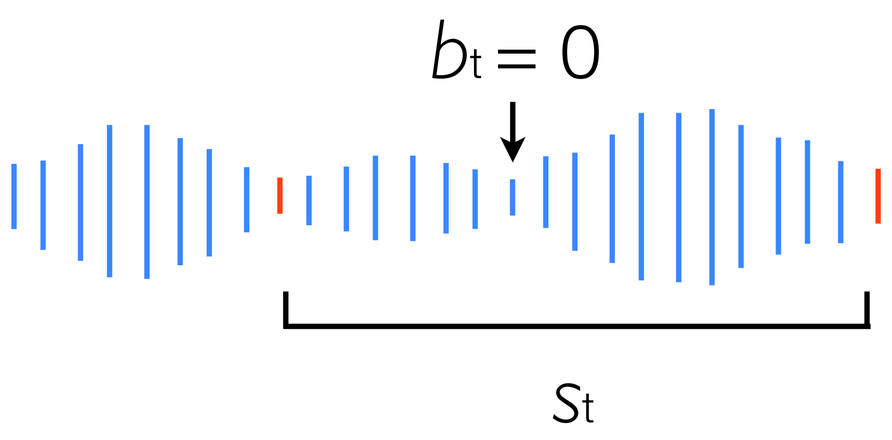
$$p(b_t = 1 \mid \cdots) \propto p(b_t = 1)\, p(s_{t-1} \mid c^-, \underline{\theta})\, p(s_{t+1} \mid c^-, \underline{\theta})$$

c^- : cluster labels of all other segments

$\underline{\theta}$: the set of HMMs

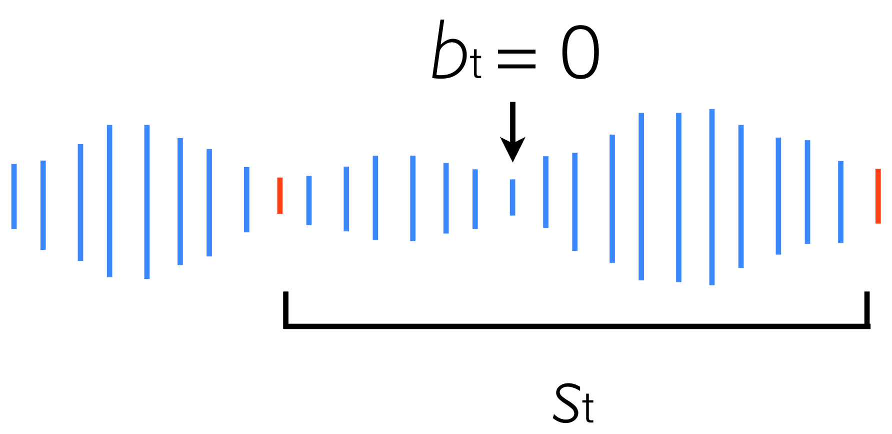
Prior and Posterior for Phone Boundaries

- **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



$$p(b_t = 0 \mid \cdots) \propto p(b_t = 0)\, p(s_t \mid c^-, \underline{\theta})$$

Prior and Posterior for Phone Boundaries

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

$$p\left(b_t=1 \mid \cdots\right) \propto p\left(b_t=1\right) p\left(s_{t-1} \mid c^-, \underline{\theta}\right) p\left(s_{t+1} \mid c^-, \underline{\theta}\right)$$

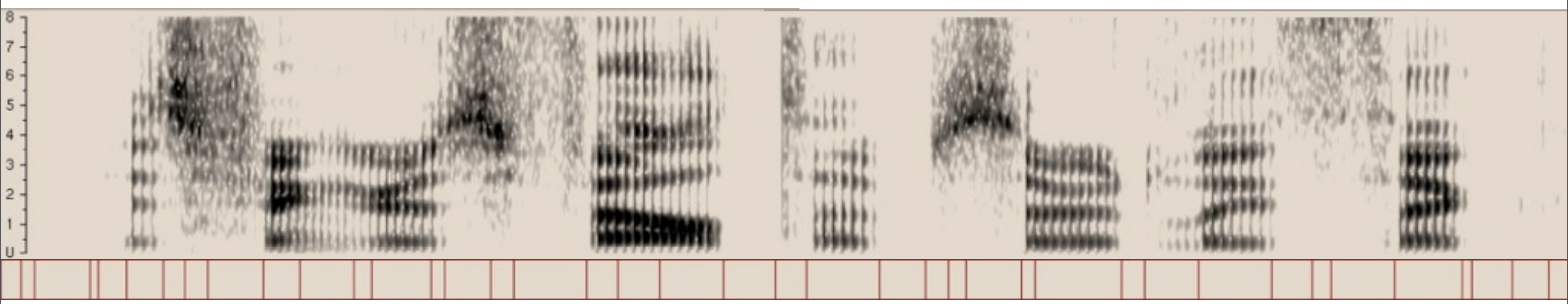
$$p\left(b_t=0 \mid \cdots\right) \propto p\left(b_t=0\right) p\left(s_t \mid c^-, \underline{\theta}\right)$$

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

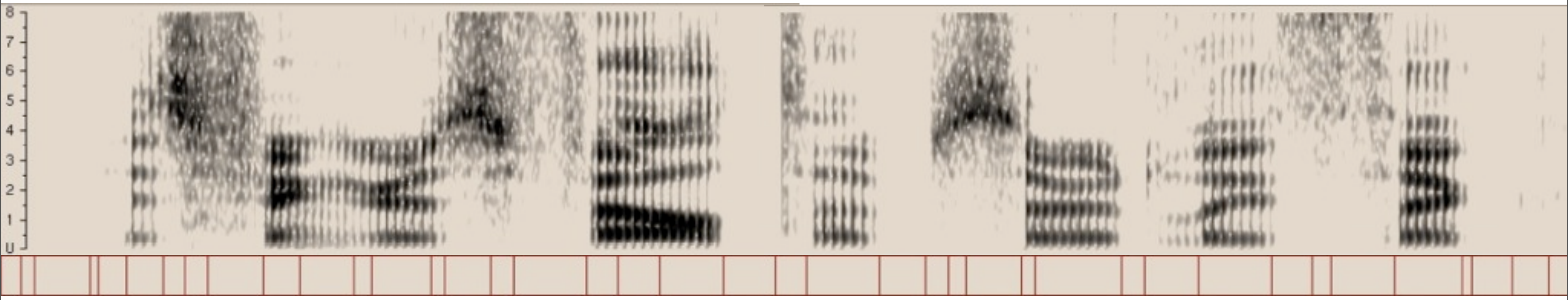
Acoustic Landmarks

- 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

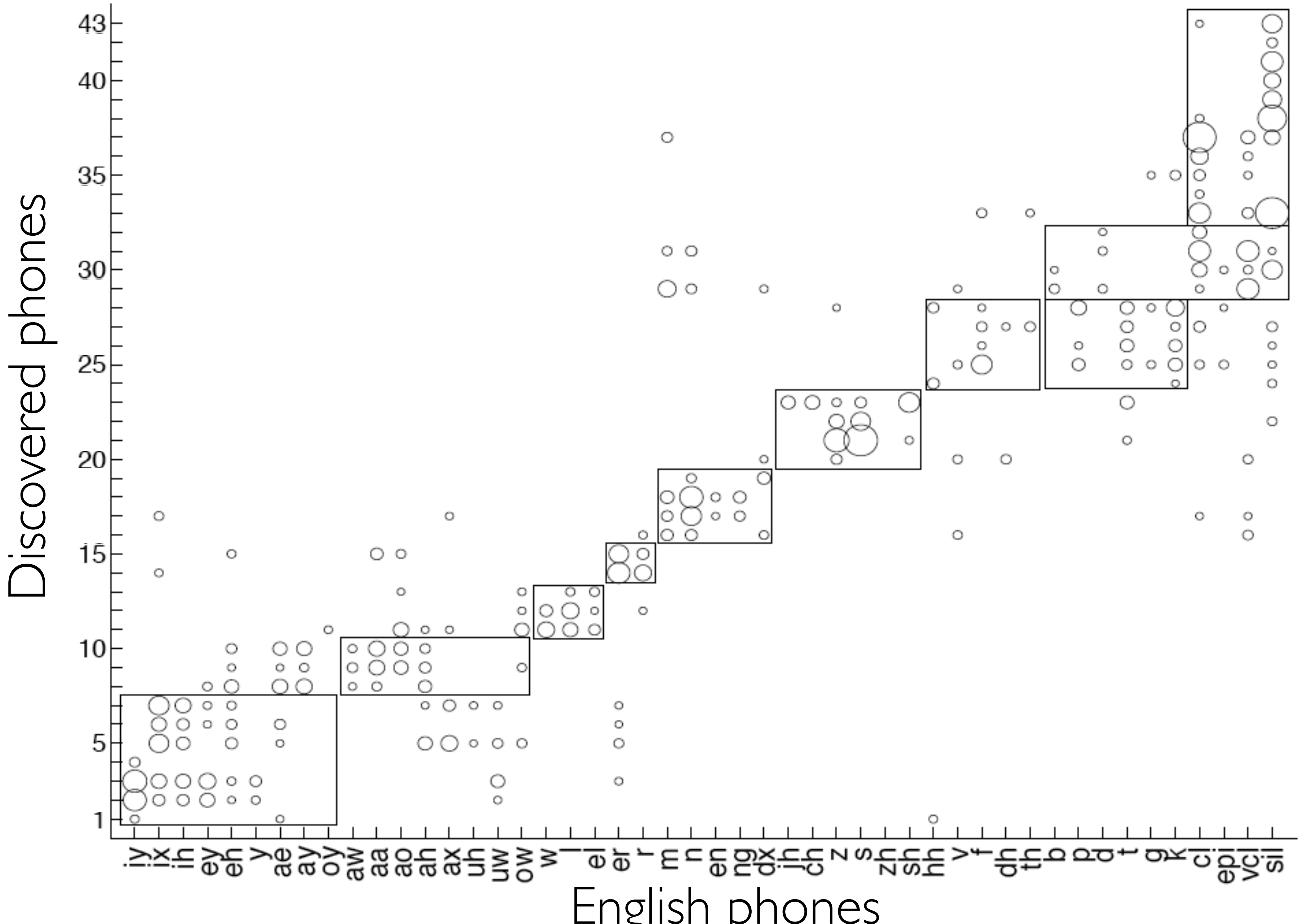
- **Data set**
  - TIMIT training and test sets
  - Multi-speaker, clean read speech, 16kHz sampling rate
- **Qualitative assessment**
  - Correlation between induced phone units and English phones
  - Compare results of 300 and 3696 utterances

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

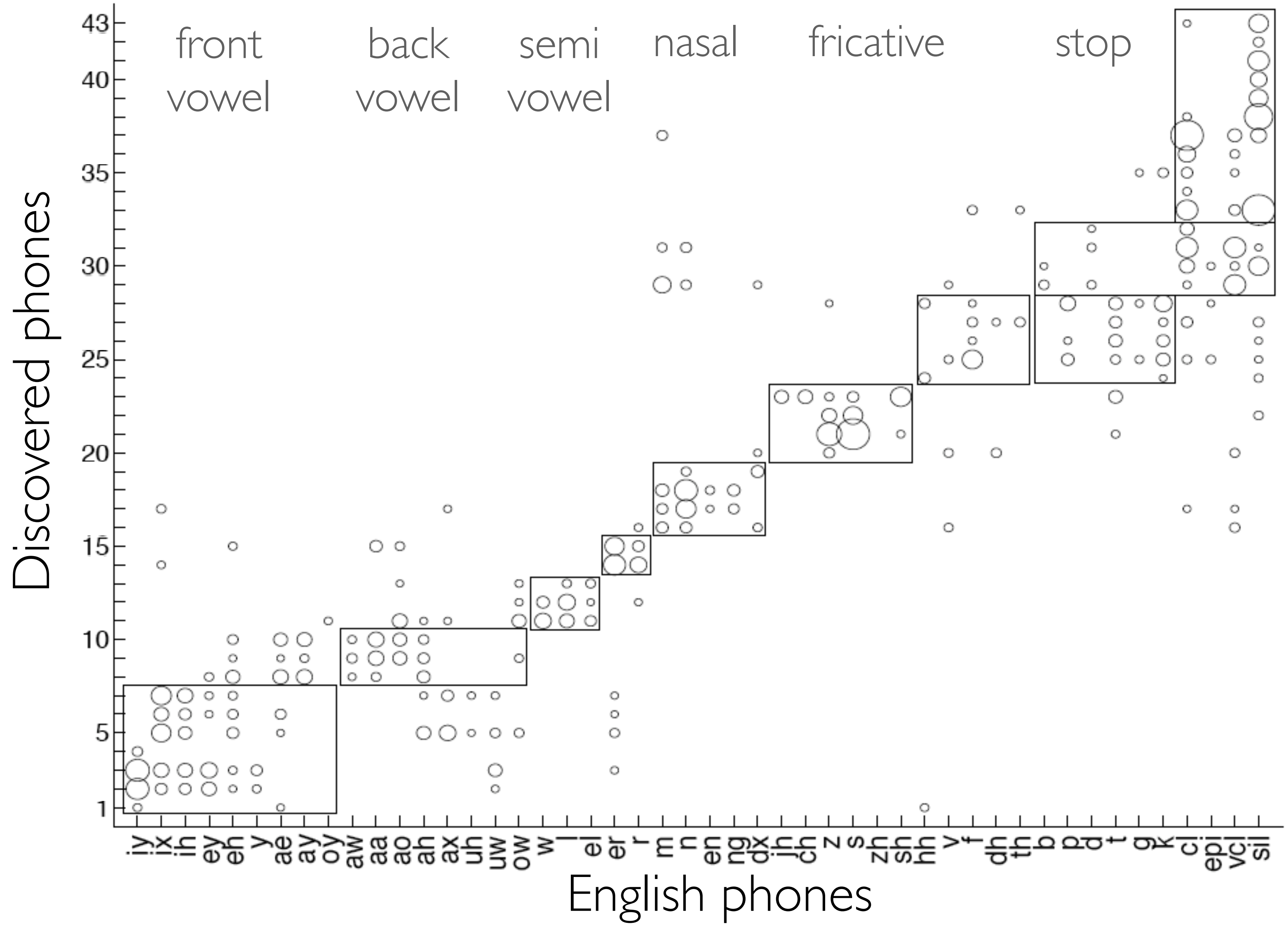
Discovered Phone Units -- 300 utterances

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



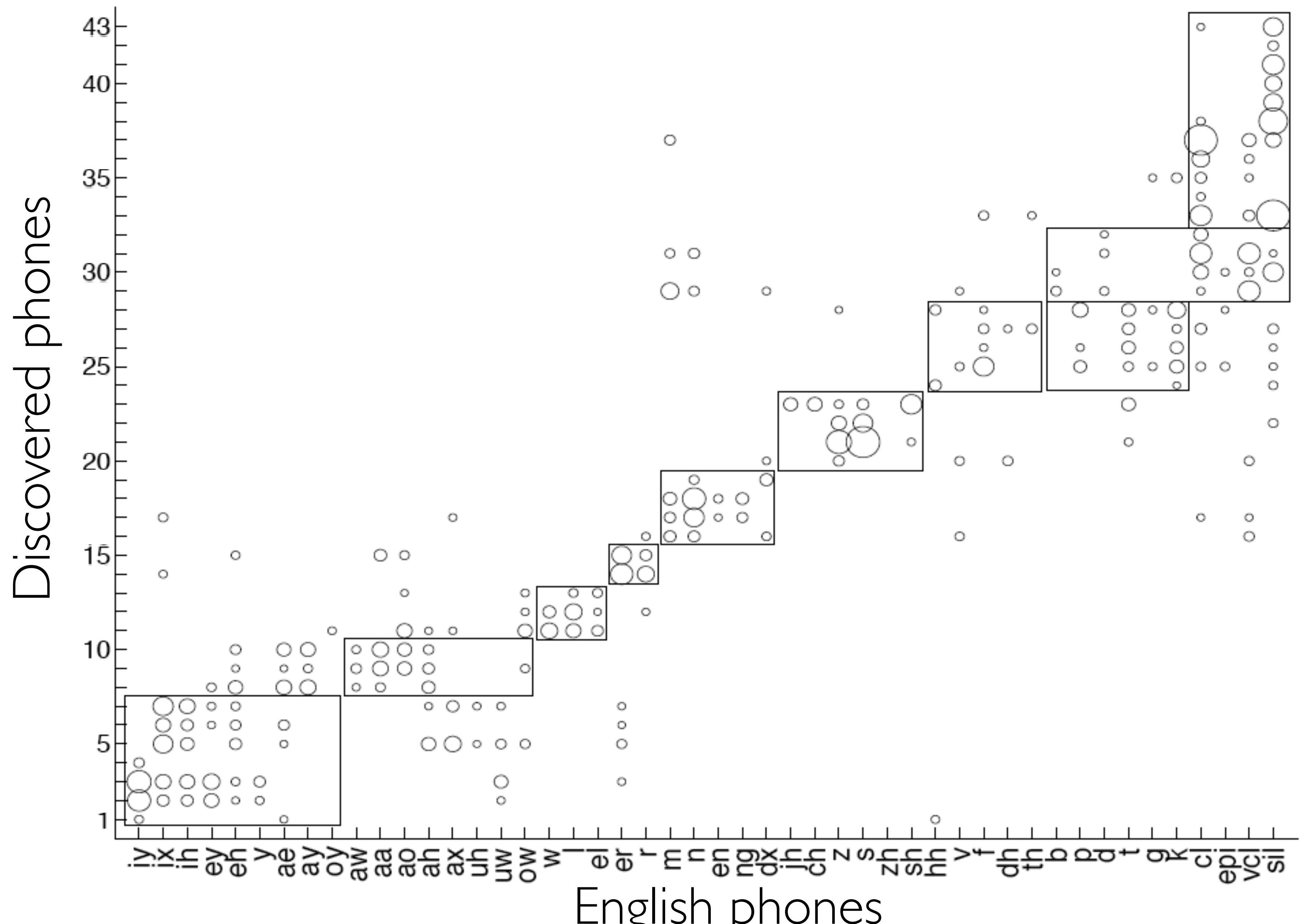
Discovered Phone Units -- 300 utterances

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



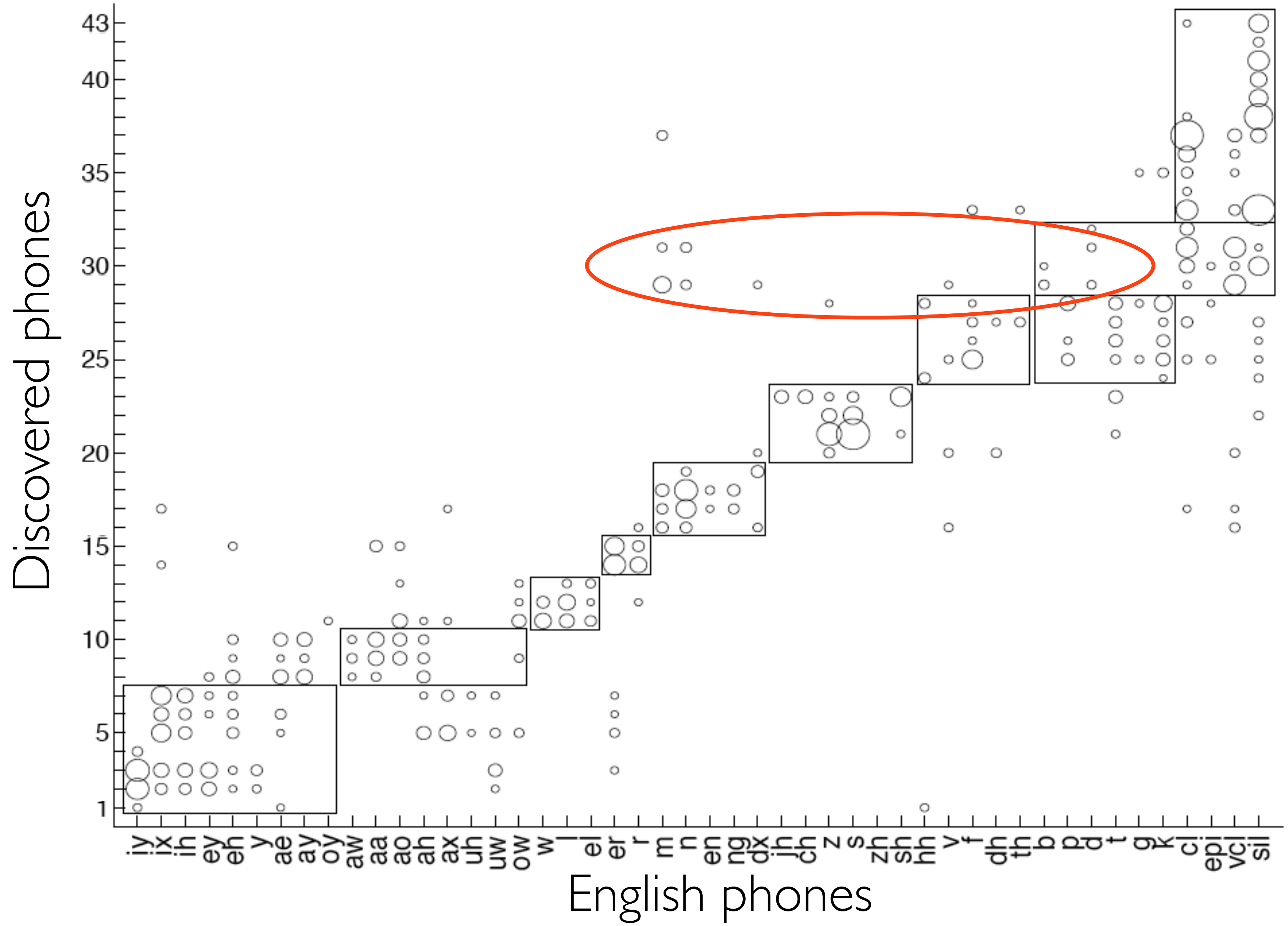
Discovered Phone Units -- 300 utterances

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



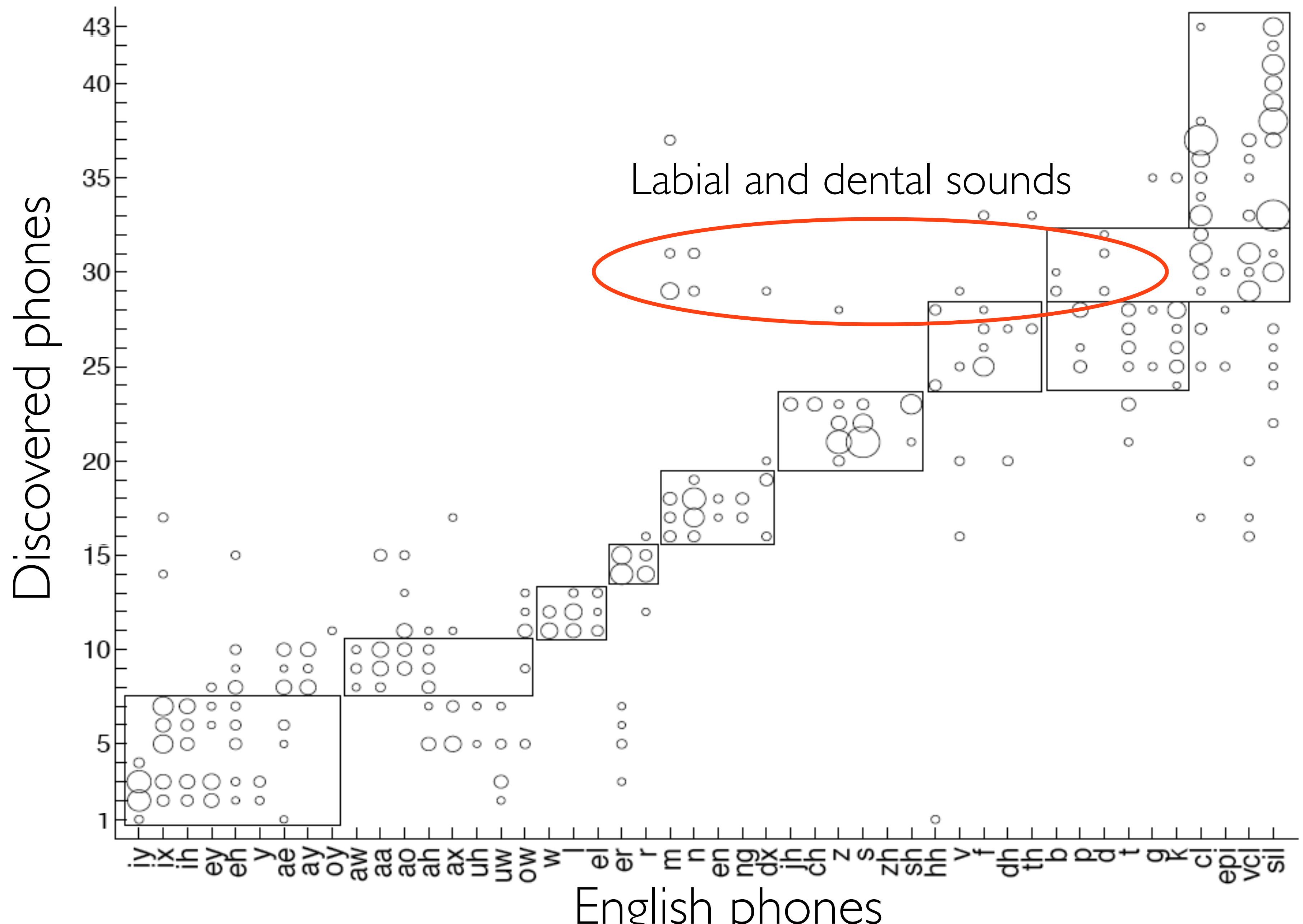
Discovered Phone Units -- 300 utterances

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



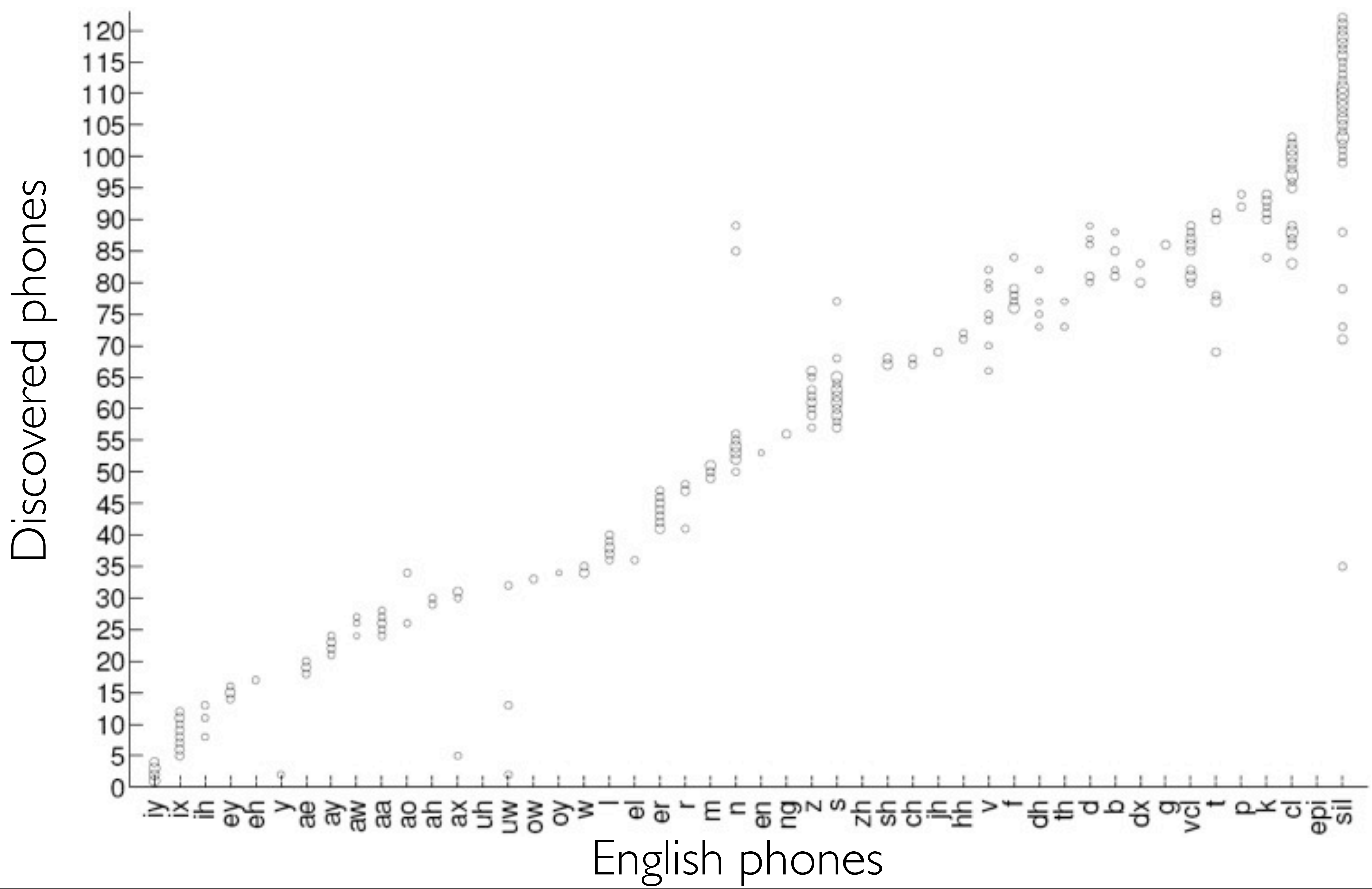
Discovered Phone Units -- 300 utterances

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



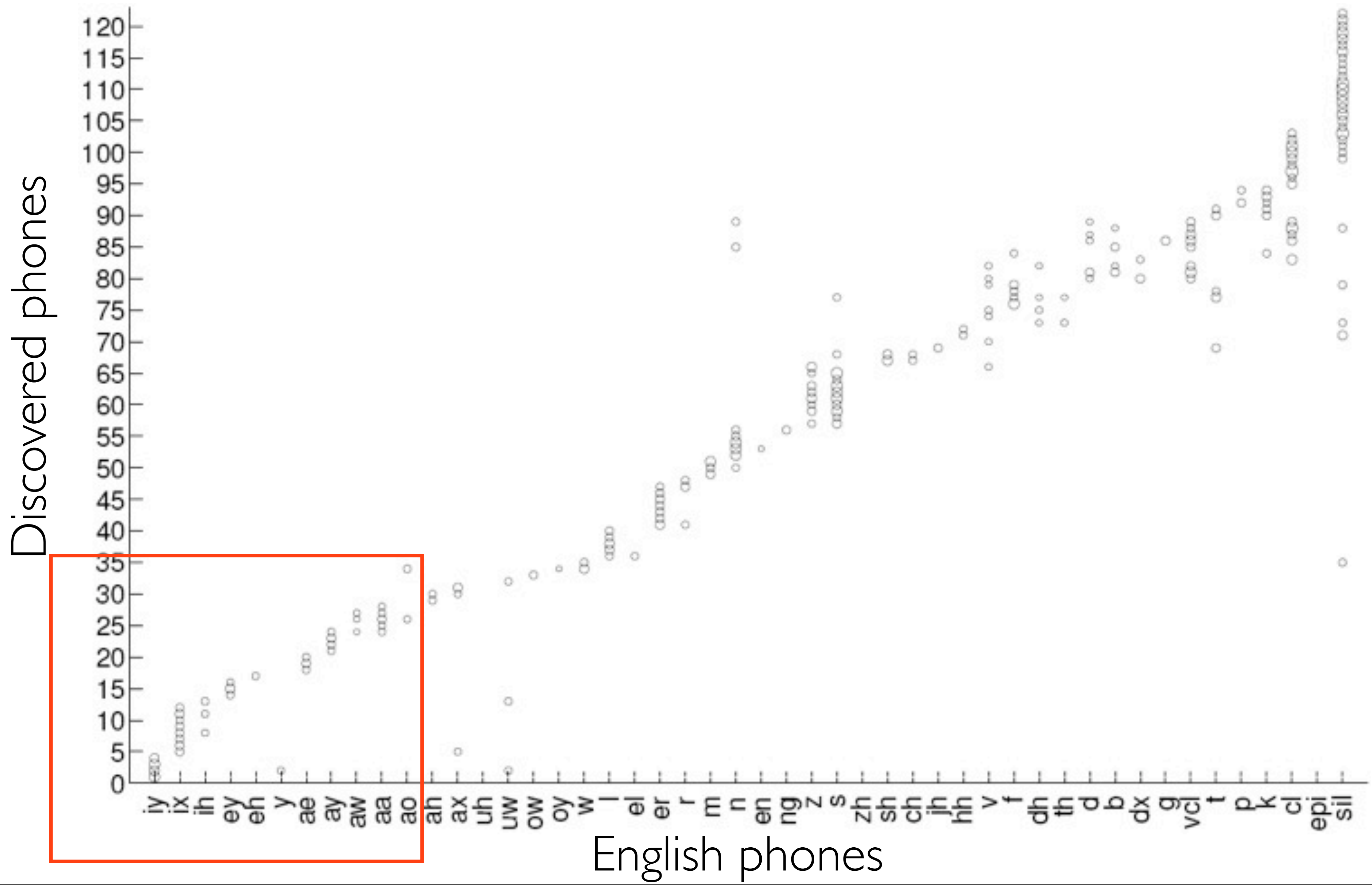
Discovered Phone Units -- 3696 utterances

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



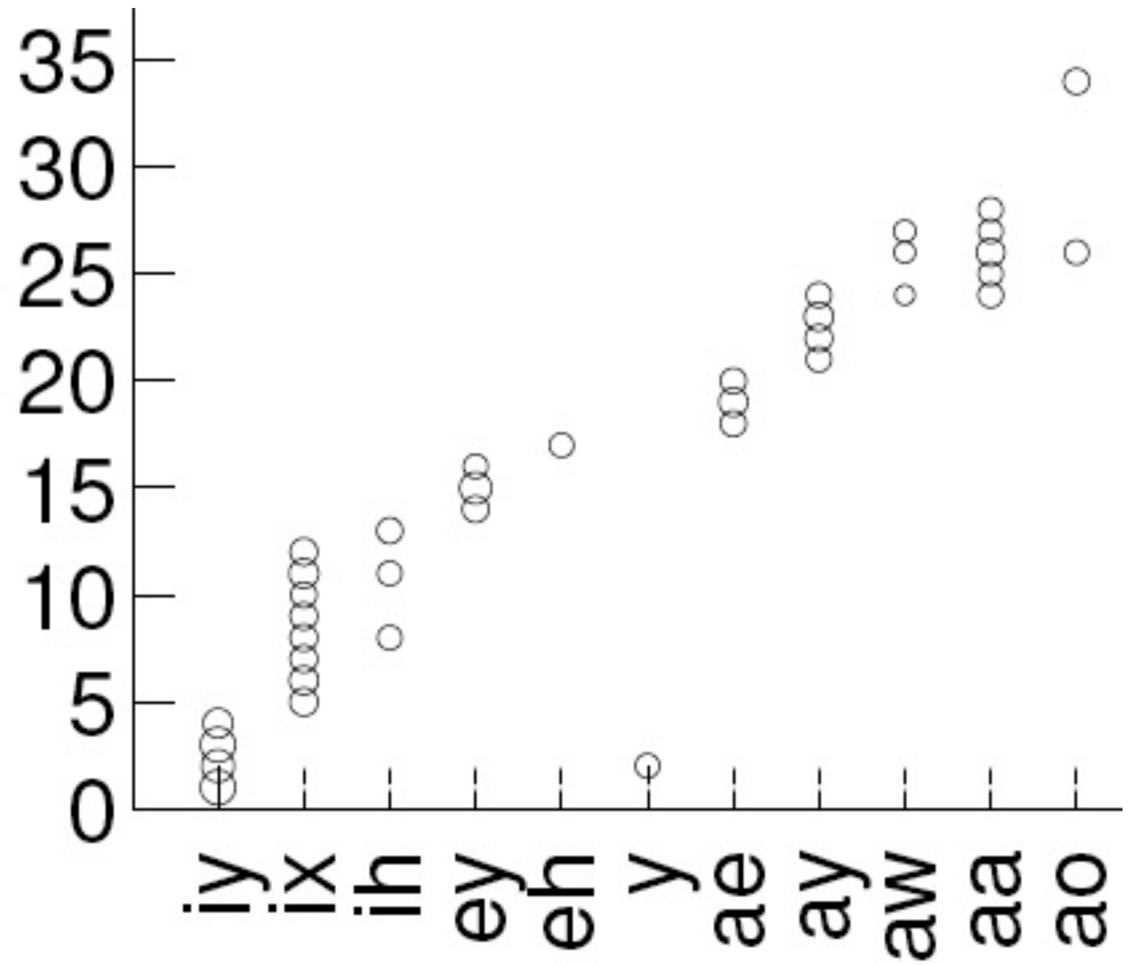
Discovered Phone Units -- 3696 utterances

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



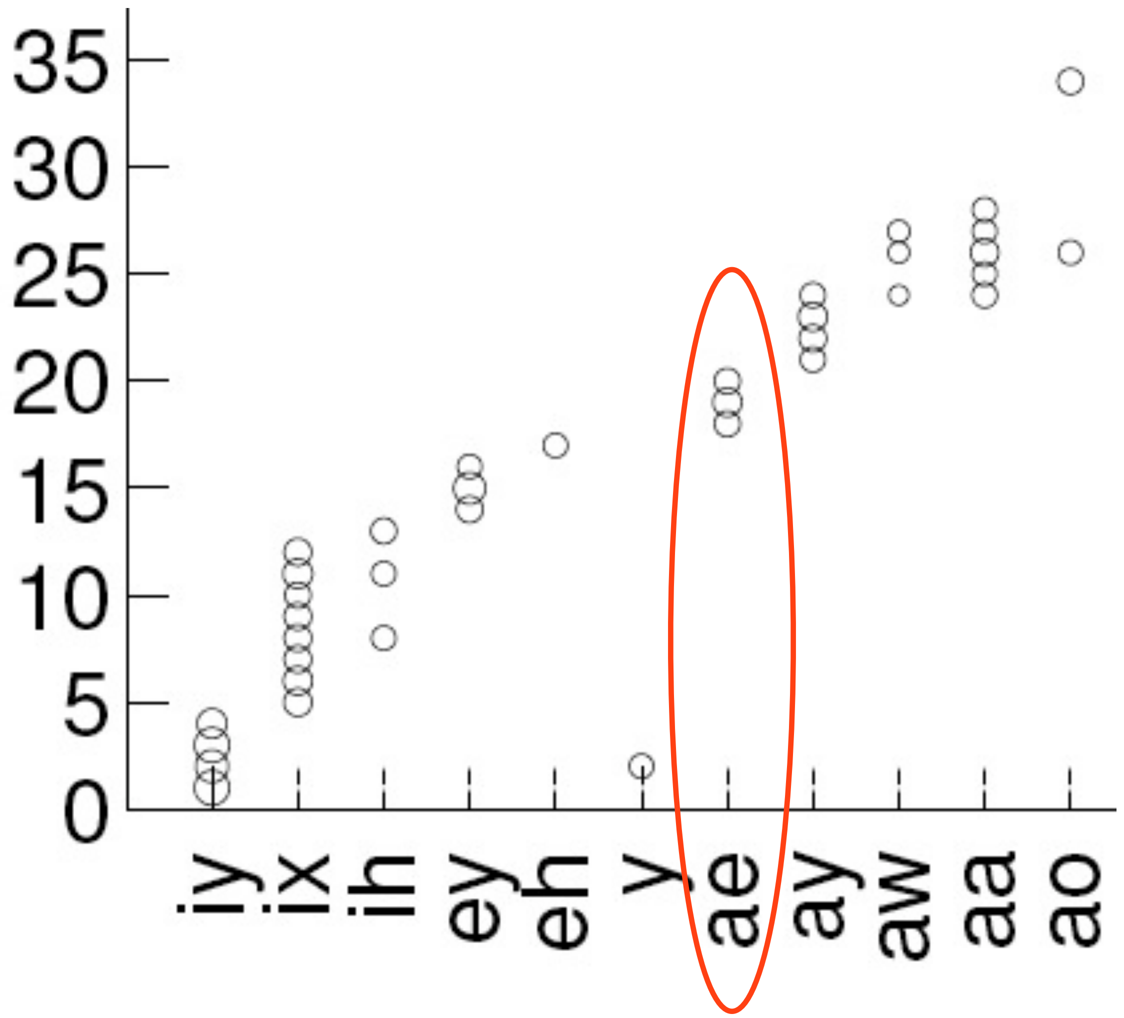
Discovered Phone Units -- 3696 utterances

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



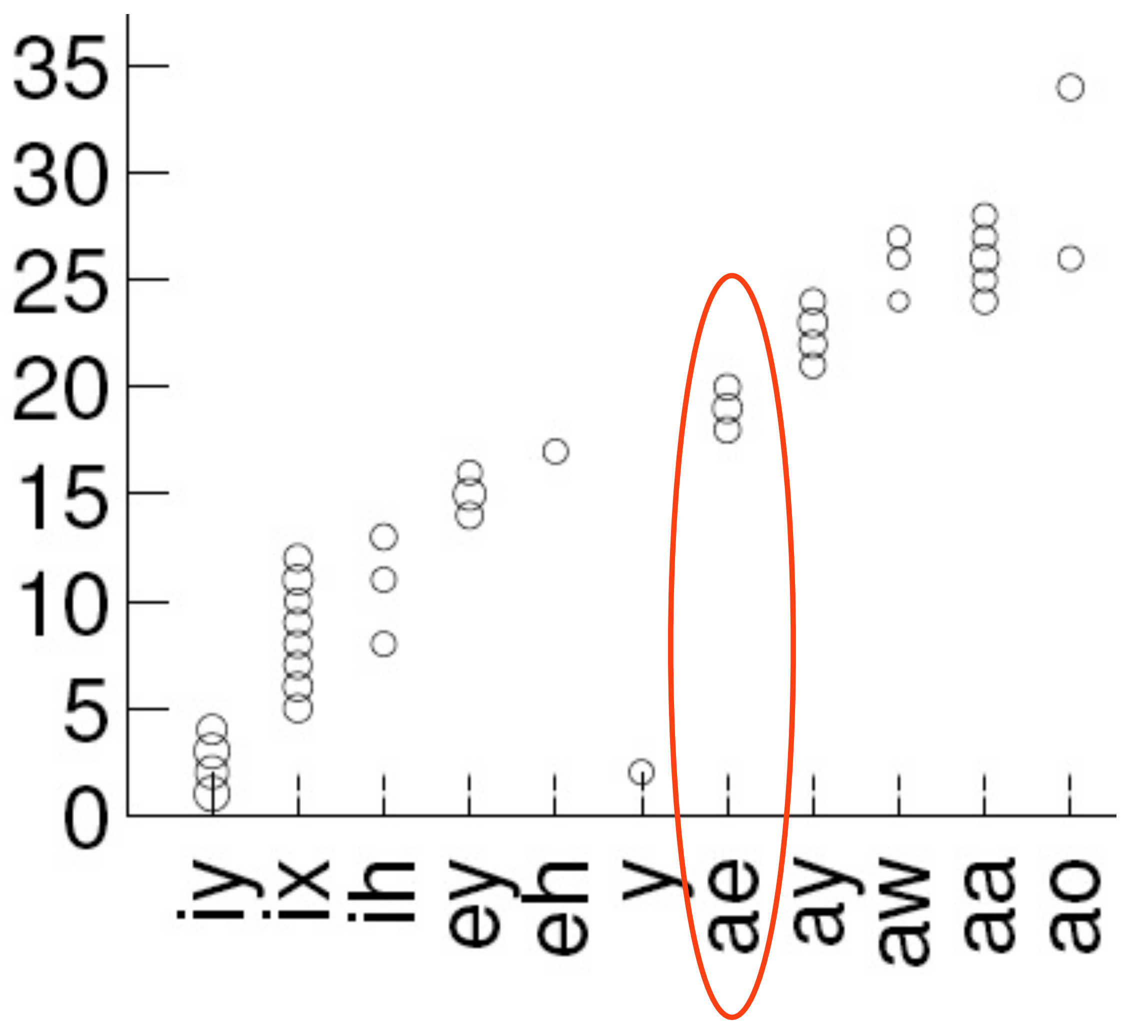
Discovered Phone Units -- 3696 utterances

- 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

x : a single frame of feature vector

$State_{i,j}$: the j-th state of the i-th HMM

$$\text{posterior-gram}(x) = \left[\frac{p(State_{i,j} \mid x)}{\sum_{i=1}^{K} \sum_{j=1}^{3} p(State_{i,j} \mid x)} \right] \text{ for } 1 \le i \le K \text{ and } 1 \le j \le 3$$

K : 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

P@N	EER

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

	P@N	EER
English Monophone (Supervised)	74.0	11.8
Thai Monophone Model (Supervised)	56.6	14.9
Our model	63.0	16.9

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

	P@N	EER
English Monophone (Supervised)	74.0	11.8
Thai Monophone Model (Supervised)	56.6	14.9
Our model	63.0	16.9
Zhang 2009 (GMM) (Unsupervised)	52.5	16.4
Zhang 2012 (DBM) (Unsupervised)	51.1	14.7

Phone Segmentation

- TIMIT training set

Phone Segmentation

- TIMIT training set

	Recall	Precision	F-score
Dusan et al. (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
Landmarks	87.0	50.6	64.0

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**

Acoustic Model

Lexicon

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

Acoustic Model

Lexicon

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**

Acoustic Model

Lexicon

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