



Context Modelling for HMM-Based Speech Synthesis

Kai Yu

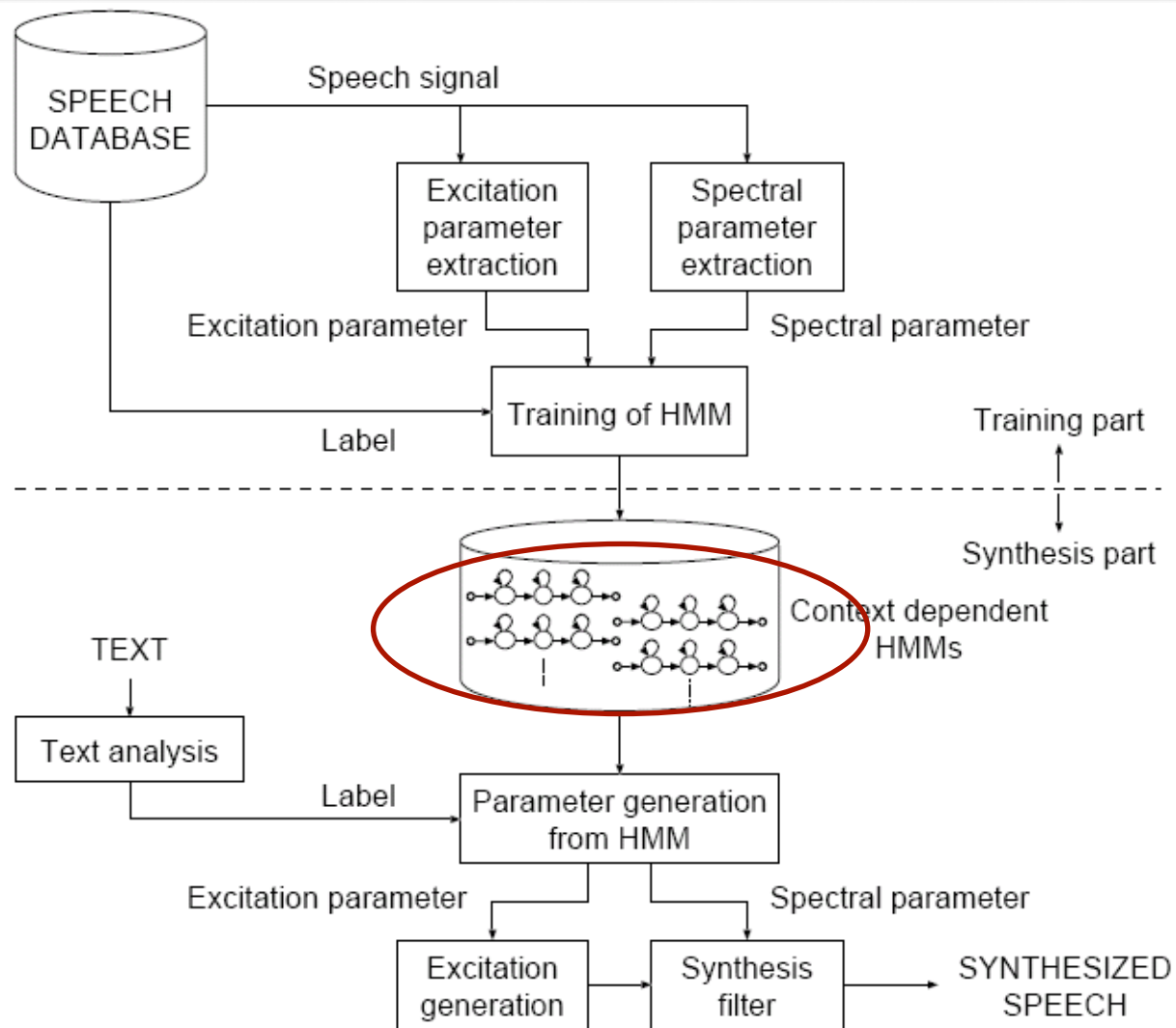
Machine Intelligence Lab
Cambridge University Engineering Department

Mar. 16, 2011

Overview

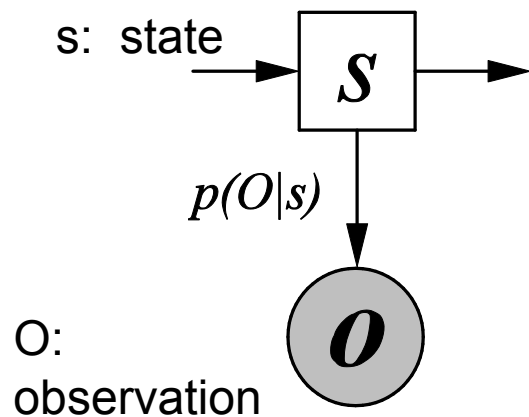
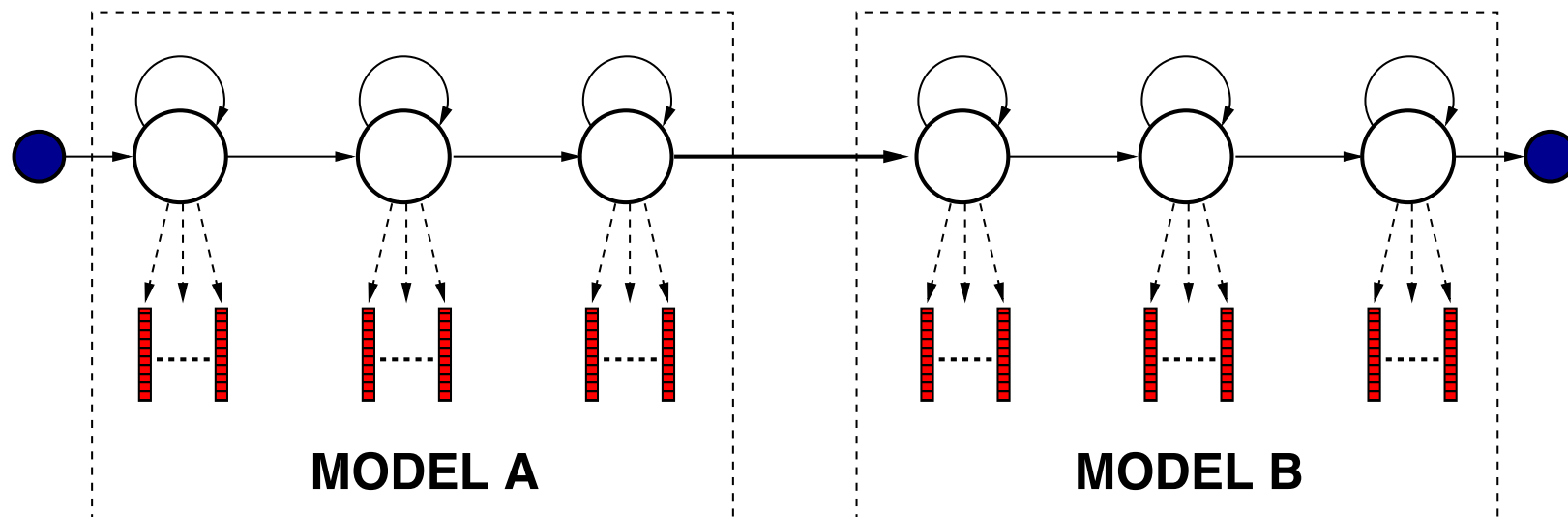
- Rich context features in speech synthesis
- Context dependent HMM modelling
 - Decision tree based state clustering
- Context groups and factorization
- **Structured context modelling framework**
 - **Adaptive training with factorized decision tree**
 - **Product of expert**
- **Discussion welcome!**

HMM based Statistical Speech Synthesis (HTS)



HMM-based Speech Synthesis

Composite HMMs



$$p(\mathbf{O}|\mathcal{M}) = \sum_{\theta} a_{\theta_0, \theta_1} \prod_{t=1}^T a_{\theta_{t-1}, \theta_t} \mathcal{N}(\mathbf{o}_t; \mu_{\theta_t}, \Sigma_{\theta_t})$$

$$\mathcal{M} = \{\mathcal{M}_a, \dots, \mathcal{M}_b\}$$

Rich contexts of phone in speech synthesis



■ Context features of **eh**

Neighbouring phones	Left: s
	Right: n
Position	2 nd phone from word start
	4 th phone from word end
Stress/Accent	Current phone stressed
Linguistic role	Noun, object
Emphasis	Current word emphasized
	Previous word not emphasized

Issues with Rich Contexts - Complexity

- Significantly increased model complexity
 - One HMM per context

```
aa^aa-v+dh=ax@2_1/A:1_0_1/B:1-0-2@1-1&11-3#9-2$2-1!1-2;8-2|aa/C:0+0+2/D:content_1/E:in+1@10+3&7+1#1+2/F:det_1/G:0_0/H:13=12@1=1|L-L%/I:0=0/J:13+12-1
```

- Typical context dimension: 55
- Typical full context HMMs:
 - 34110 – ARCTIC 1 hr, 1K sent
 - $> 1e+23$ – All possible combinations
- Robustness of parameter estimation
- Unseen new contexts during synthesis (Generalization ability)

Issues with Rich Contexts – Acoustic Effect

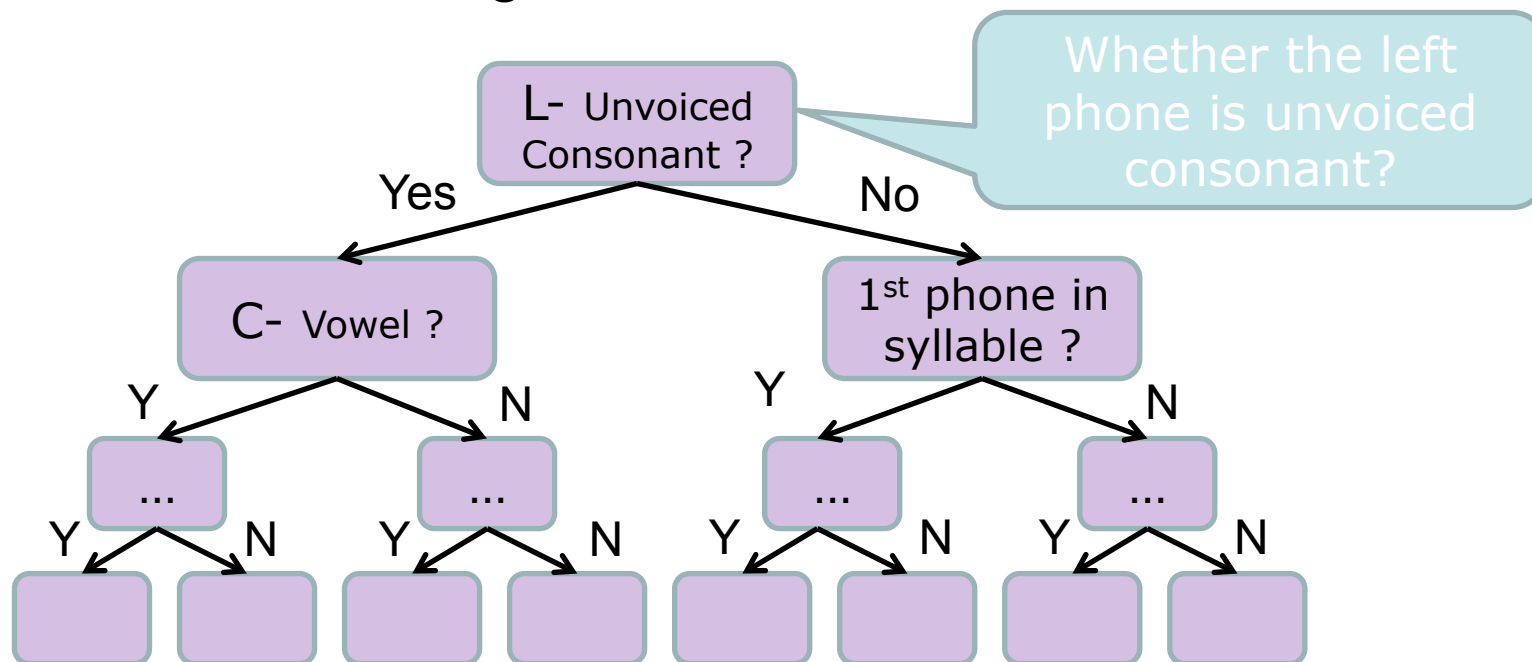
- Contexts are of different acoustic effect natures
- Affected acoustic property
 - Source (F0) – acoustic unit counts/phonetic/position
 - Spectrum – phonetic/syllable
 - Duration – position/phonetic
- Strength
 - Emphasis v.s. word stress
- Homogeneity (description) range
 - Phone/Syllable/Word/Phrase
 - Sentence/Corpus - Speaker/Emotion
 - Sometimes range boundary is not clear

Issues with Rich Contexts – Label Inaccuracy

- From text to rich contexts
 - Dictionary mapping – phone, accent, stress
 - Linguistic analysis – syllable, phrase, content words, PoS
 - Counting – number, position
 - Rich out-of-text information – emotion, accent, etc.
- Analysis (conversion) is not accurate
 - Multiple pronunciations or uncommon realization
 - Automatic label generation errors
 - Labelling inconsistency (emotion, emphasis)
- How to improve accuracy of context label generation
- Exact effect of inaccurate context labels?

Decision tree based state clustering

- Why decision tree based clustering
 - Effective treatment of unseen contexts
 - Reduced model complexity via parameter sharing
- Yes/No context-specific questions
- State-based clustering instead of model based



Procedure of Decision Tree Based State Clustering

- Build mono-phone HMMs with single Gaussian per state
- Initialize full context-dependent HMMs
- For each state (stream) index, build one tree to construct parameter sharing structure
 - Pull all data together to form a single Gaussian dist. as the root node
 - For each leaf node, select a context question to split the node into two
 - Likelihood of the whole data set will increase
 - The selected question is the one maximizing the likelihood increase
 - Repeat the process until stopping criterion is met
- Gaussian parameters within each leaf node are tied

Decision question selection

- Decision question example

Phonetic	Left phone is vowel?
	Current phone is "aa"?
	Current syllable is stressed?
Position	Current phone is the 3 rd one of the syllable from backward?
Number	The number of phones in the current syllable is 3?

- Efficient likelihood calculation from statistics

$$\begin{aligned}\mathcal{L}(\mathcal{S}) &= -\frac{1}{2} \sum_{t,s \in \mathcal{S}} \gamma_s(t) \log \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}(\mathcal{S}), \boldsymbol{\Sigma}(\mathcal{S})) \\ &= -\frac{\gamma(\mathcal{S})}{2} (\log |\boldsymbol{\Sigma}(\mathcal{S})| + K)\end{aligned}$$

Stopping criteria

- Likelihood increase less than threshold
- Occupancy of leaf nodes less than threshold
- Trade-off between model complexity and likelihood increase
 - Minimum description length

$$l(\mathcal{M}_i) = -\log p(\mathcal{D}|\mathcal{M}_i^{\text{ML}}) + \lambda \frac{\alpha_i}{2} \log N_{\mathcal{D}} + K$$

Likelihood of data given ML estimate

Number of free parameters

Data points or total occupancy

Manual scaling factor

Problems with straightforward context modelling

- Weak contexts such as natural emphasis may be completely ignored
- Re-clustering of all contexts is required in case of any context change
- Effects of contexts are modelled sequentially rather than simultaneously
- Training data is fragmented with the tree growing
- Phone (state/stream) level likelihood may not be consistent with context range
- Incorporating new context will lead to exponentially increased parameters

Context groups and factorization

- Context questions statistics

	Phone Identity	Position	Counts	Accent/ Stress	Part-of-Speech	Emphasis
mgc	192	8	16	3	3	0
lf0	368	98	10	320	25	0
dur	221	49	137	4	14	0

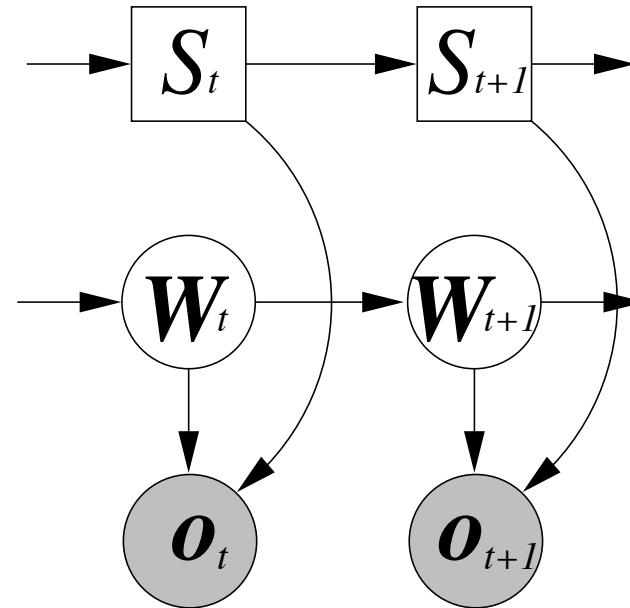
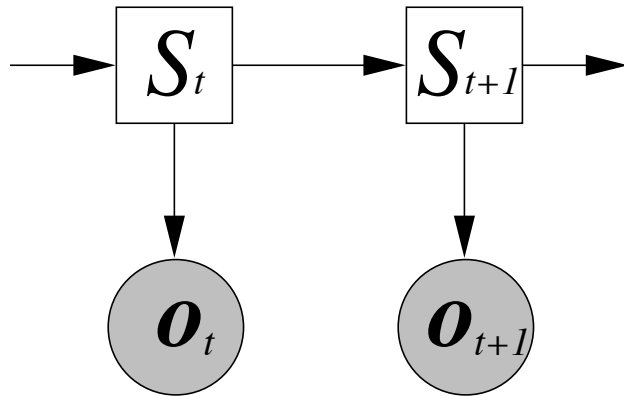
- Effect of contexts are different
- Structured modelling
 - Model the relationship and interaction of contexts
 - Wider coverage due to combination effect

Structured context modelling

- **Issues to address**
 - **Assumptions of relationship between contexts**
 - **Model structure and parameters estimation**
 - **Model usage during synthesis**
 - **State clustering**
- **Adaptive HMM framework**
- **Product of Expert (PoE) framework**

Adaptive HMM for context modelling

- Adaptive HMMs



- Multiple sets of model parameters
 - Each model set is associated with one context group
 - Transforms are used to modify HMM parameters
 - Model context relationship

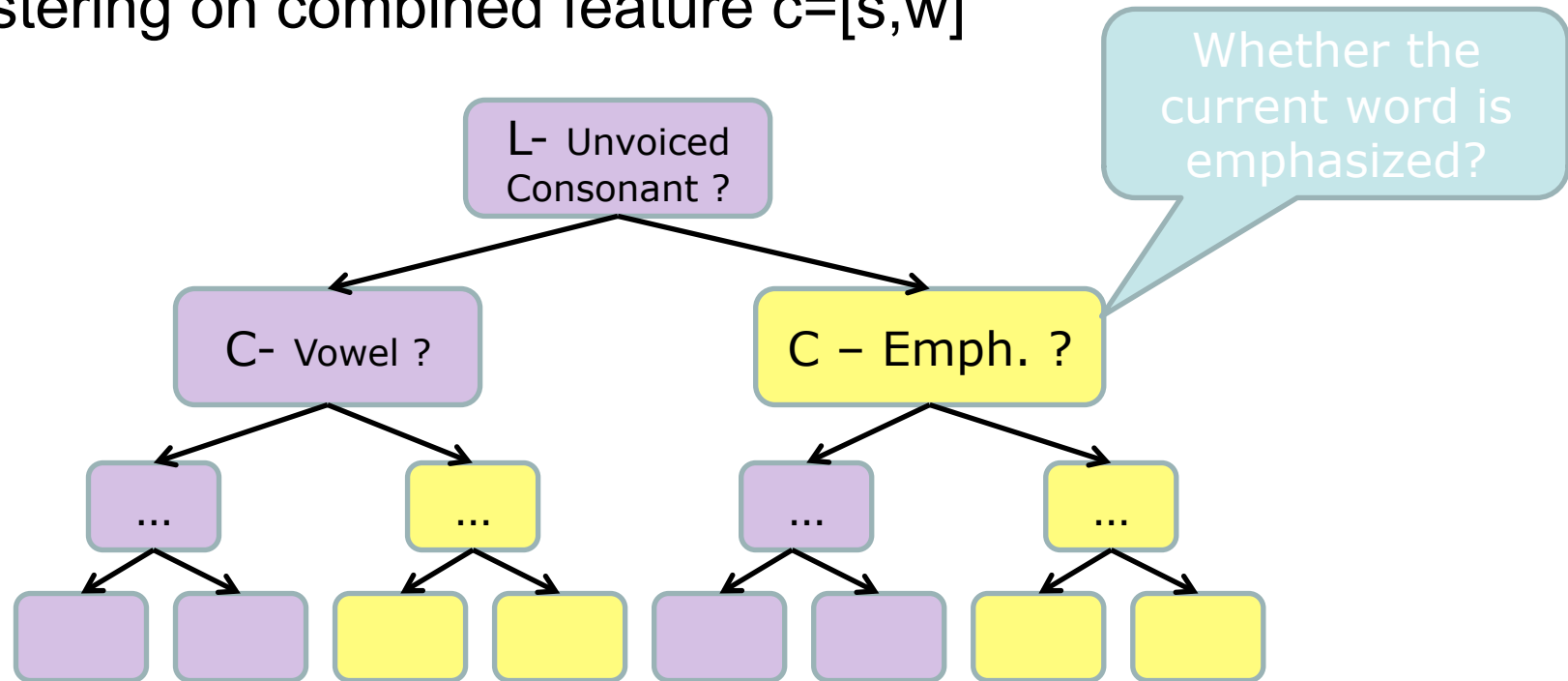
Context-Specific Homogeneity

$$p(\mathbf{o}|s, w) = p(\mathbf{o}|\mathcal{M}(s, w)) \quad \mathcal{M}(s, w) = \mathcal{F}_w(\mathcal{M}_s)$$

- s and w are context FACTORS, each can contain several context features
- Homogeneity assumption:
 - \mathcal{M}_s and \mathcal{F}_w are unchanged within homogeneous block
 - Homogeneous range can be phone/sentence/corpus
- Full context $c=[s,w]$
 - Number of full contexts: $N_c = N_s \times N_w$
 - Number of factorized contexts: $N_s + N_w$

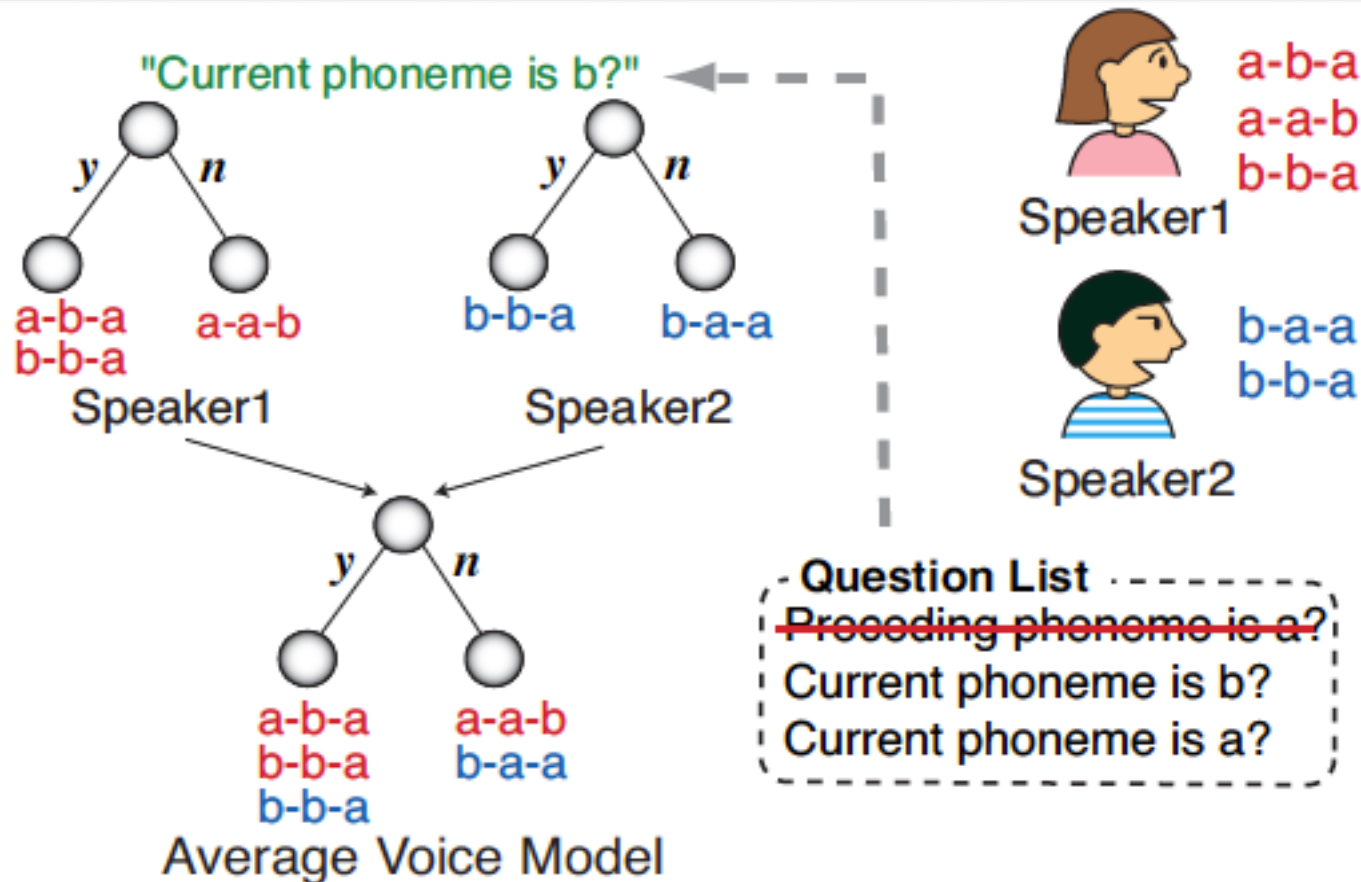
Parameter Tying with Structured Context Modelling

- Decision tree based state clustering
- Clustering on combined feature $c=[s,w]$



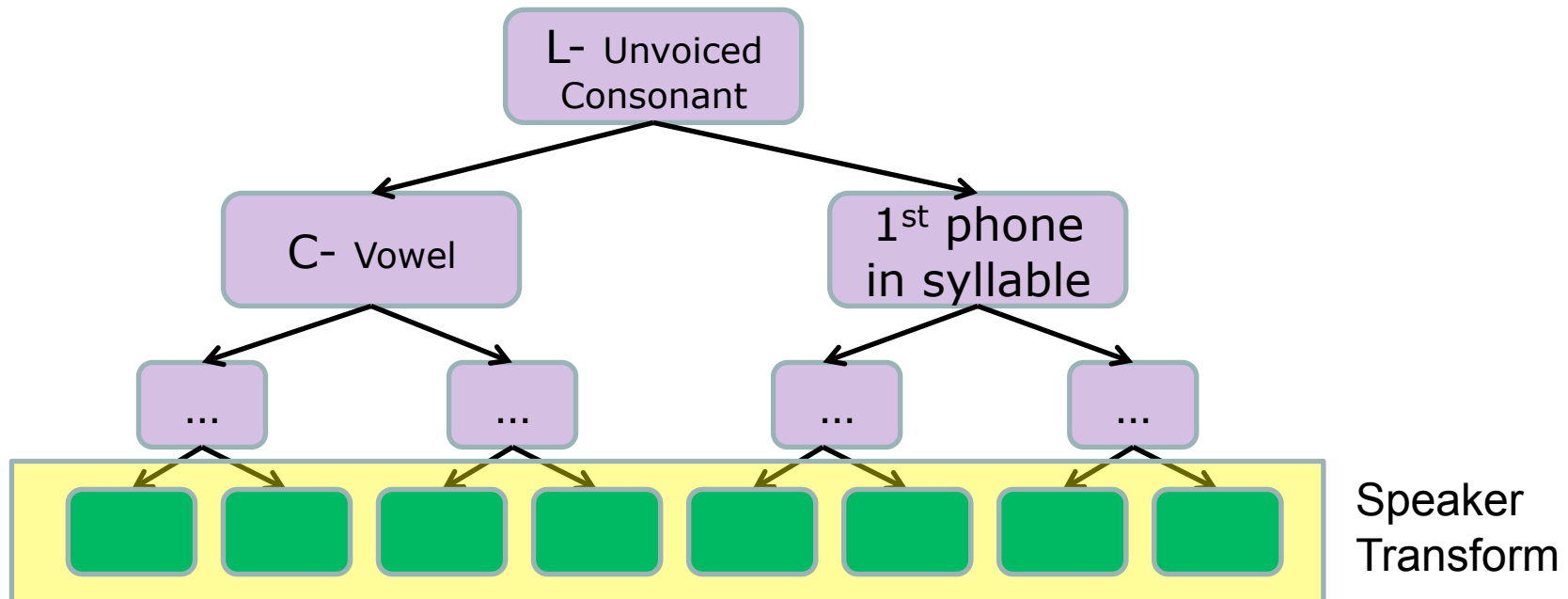
- Structured clustering
 - Shared decision tree
 - Factorized decision tree

Shared Decision Tree



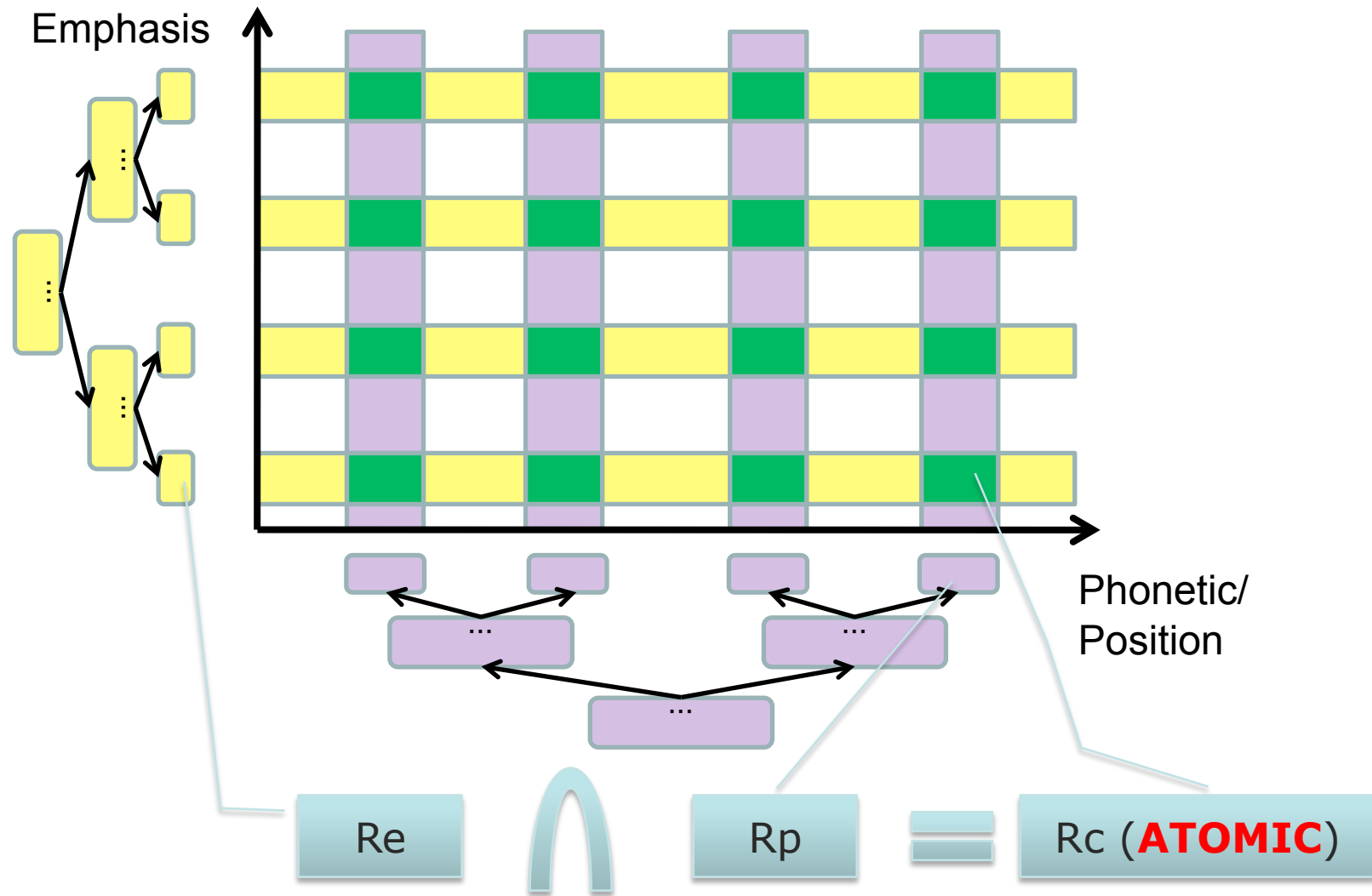
- Common context questions for all speakers to avoid unbalanced split during clustering
- Common decision trees for all speakers

Speaker Adaptation in ASR

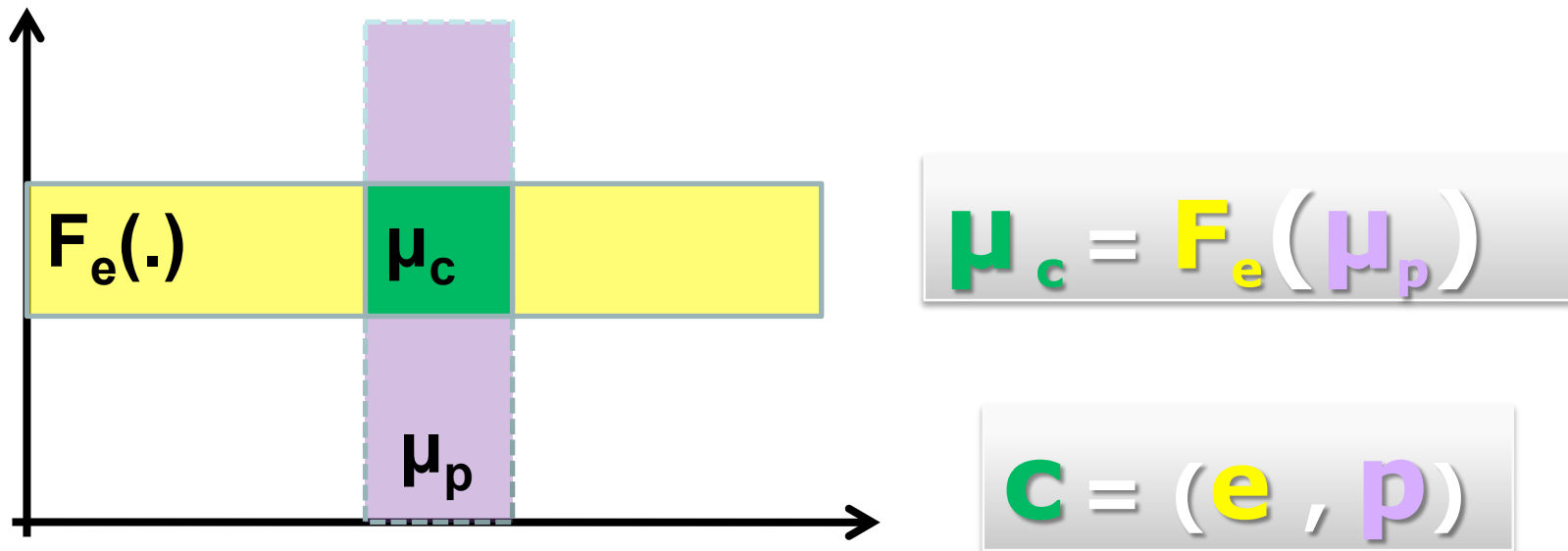


- No question selection
- Common phonetic context trees for all speakers
- Speaker adaptation is a special case of context adaptive HMM framework

Factorized Decision Tree – Emphasis as e.g.

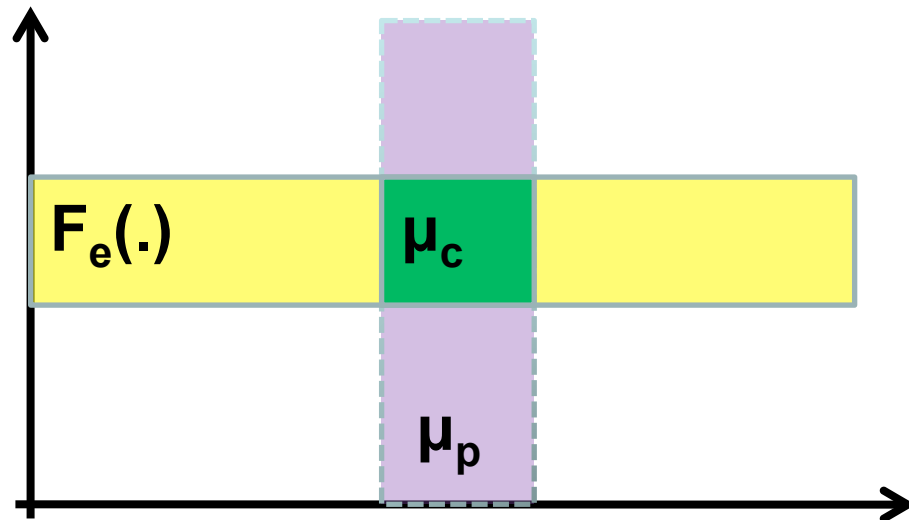


Model Structure of Context Adaptive Training



- Base Gaussians associated with base context tree
- Transform associated with emphasis context tree
- Param. of atomic nodes are combination of the two
- Two sets of para. estimation interleaves

Linear Transform Based Approach



$$\mu_c = F_e(\mu_p)$$

MLLR

$$\hat{\mu}_{r_c} = A_{r_e} \mu_{r_p} + b_{r_e}$$
$$\hat{\Sigma}_{r_c} = \Sigma_{r_p}$$

CMLLR

$$\hat{\mu}_{r_c} = A_{r_e} \mu_{r_p} + b_{r_e}$$
$$\hat{\Sigma}_{r_c} = A_{r_e} \Sigma_{r_p} A_{r_e}^T$$

- Context relationship is assumed to be (piece-wise) linear
- Powerful in terms of context transformation
- Hard to model more than two context factors

Parameter Estimation – Linear Transform Based Context Adaptive Training

$$\hat{\Lambda}_{r_c} = \mathcal{F}_{r_e}(\Lambda_{r_p}) \quad r_c = r_p \cap r_e \quad \hat{\mu}_m = A_{r_e(m)} \mu_{r_p(m)} + b_{r_e(m)} = W_{r_e(m)} \xi_{r_p(m)}$$

$$\hat{\Sigma}_m = \Sigma_{r_p(m)}$$

The flowchart illustrates the derivation of parameters from observed data. It starts with the equation $W_{r_e,d} = G_{r_e,d}^{-1} k_{r_e,d}$, which leads to $\mu_{r_p} = G_{r_p}^{-1} k_{r_p}$. The final step is $\Sigma_{r_p} = \text{diag} \left(\frac{\sum_{t,m \in r_p} \gamma_m(t) (o_t - \hat{\mu}_m)(o_t - \hat{\mu}_m)^T}{\sum_{t,m \in r_p} \gamma_m(t)} \right)$. The right side of the slide provides the definitions for $G_{r_e,d}$, $k_{r_e,d}$, G_{r_p} , and k_{r_p} .

$$G_{r_e,d} = \sum_t \sum_{m \in r_e} \frac{\gamma_m(t)}{\sigma_{dd}^{r_p(m)}} \xi_{r_p(m)} \xi_{r_p(m)}^T$$

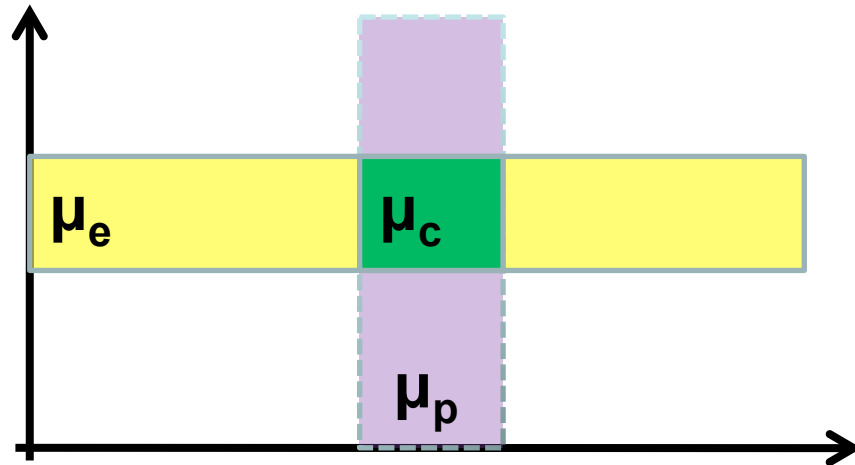
$$k_{r_e,d} = \sum_t \sum_{m \in r_e} \frac{\gamma_m(t) o_{t,d}}{\sigma_{dd}^{r_p(m)}} \xi_{r_p(m)}$$

$$G_{r_p} = \sum_t \sum_{m \in r_p} \gamma_m(t) A_{r_e(m)}^T \Sigma_m^{-1} A_{r_e(m)}$$

$$k_{r_p} = \sum_t \sum_{m \in r_p} \gamma_m(t) A_{r_e(m)}^T \Sigma_m^{-1} (o_t - b_{r_e(m)})$$

$$\Sigma_{r_p} = \text{diag} \left(\frac{\sum_{t,m \in r_p} \gamma_m(t) (o_t - \hat{\mu}_m)(o_t - \hat{\mu}_m)^T}{\sum_{t,m \in r_p} \gamma_m(t)} \right)$$

Cluster based approach



$$\mu_c = F_c(\mu_e, \mu_p)$$

$$\hat{\mu}_{r_c} = \lambda_1 \mu_{r_p} + \lambda_2 \mu_{r_e}$$

$$\hat{\Sigma}_{r_c} = \Sigma_{r_p}$$

- Less powerful due to simple interpolation weights
- Regression base-class can be used for interpolation weights
- Easy to be used for more than two factors

Parameter Estimation – Cluster Based Context Adaptive Training

$$\hat{\Lambda}_{r_c} = \mathcal{F}_{r_t}(\Lambda_{r_p}, \Lambda_{r_e}) \quad r_c = r_p \cap r_e \quad \hat{\mu}_m = \lambda_{r_t(m)}^{(p)} \mu_{r_p(m)} + \lambda_{r_t(m)}^{(e)} \mu_{r_e(m)} = M_m \lambda_{r_t(m)}$$

$$\hat{\Sigma}_m = \Sigma_{r_p(m)}$$

$$\lambda_{r_t} = G_{r_t}^{-1} k_{r_t}$$

$$G_{r_t} = \sum_t \sum_{m \in r_t} \gamma_m(t) M_m^T \Sigma_{r_p(m)}^{-1} M_m$$

$$k_{r_t} = \sum_{m \in r_t} M_m^T \Sigma_{r_p(m)}^{-1} \sum_t \gamma_m(t) o_t$$

$$\hat{\mu} = G^{-1} k \quad \hat{\mu} = [\hat{\mu}_{r_p=1}^T \cdots \hat{\mu}_{r_p=N(p)}^T \quad \hat{\mu}_{r_e=1}^T \cdots \hat{\mu}_{r_e=N(e)}^T]^T$$

$$\hat{\Sigma}_{r_p} = \text{diag} \left(\frac{\sum_{t, m \in r_p} \gamma_m(t) (o_t - \hat{\mu}_m)(o_t - \hat{\mu}_m)^T}{\sum_{t, m \in r_p} \gamma_m(t)} \right)$$

State Clustering for Factorized Decision Tree

- Independent construction
 - Context factors are completely independent
 - Easy implementation
- Dependent construction
 - Construct decision tree for one factor given the decision tree structure of the other factor
 - Interleave between multiple sets of model parameters
- Simultaneous construction
 - At each split, all trees are optimized inter-dependently

Dependent Decision Tree Construction

- MLLR based context adaptive training as example

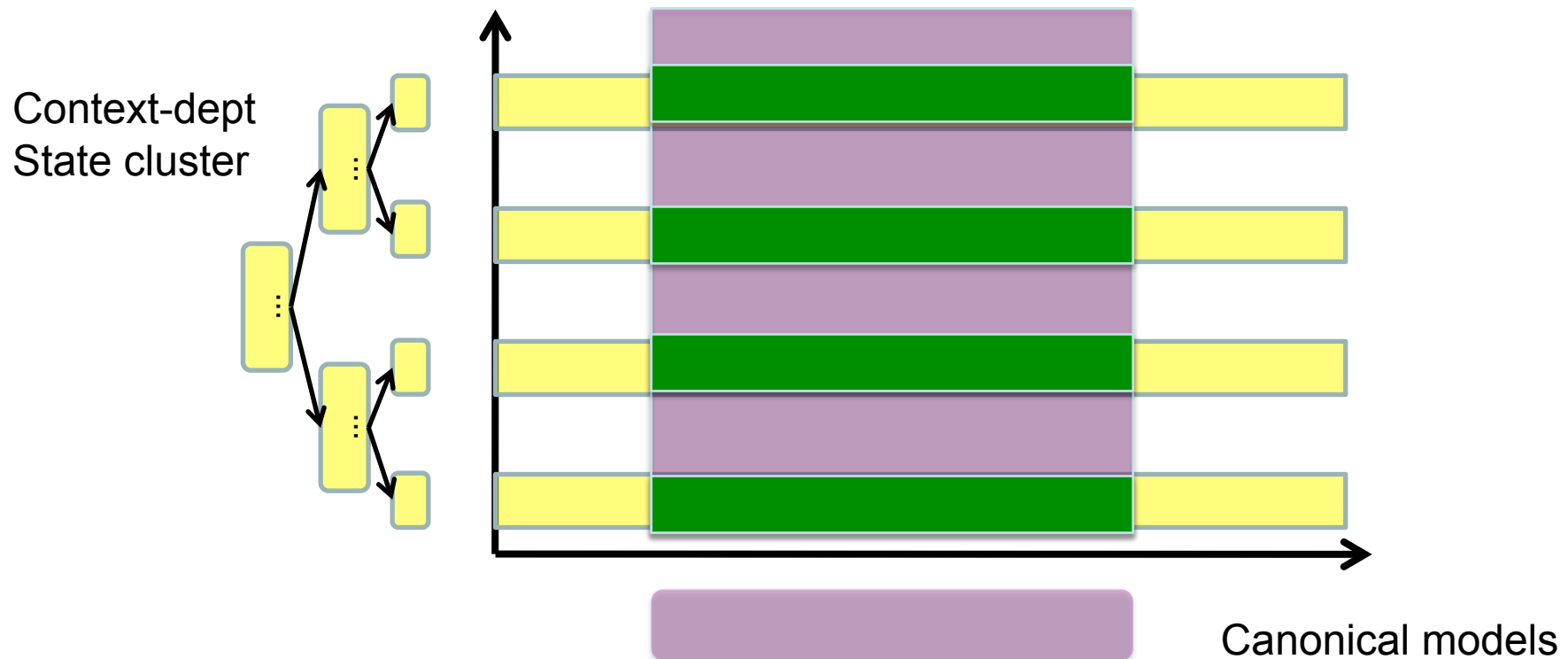
$$\mathcal{L}(S) = -\frac{1}{2} \sum_{t,s \in S} \gamma_s(t) \left(\log |\Sigma(S)| + (\mathbf{o}_t - \mathbf{A}_{r_s} \mu(S) - \mathbf{b}_{r_s})^\top \Sigma^{-1}(S) (\mathbf{o}_t - \mathbf{A}_{r_s} \mu(S) - \mathbf{b}_{r_s}) + K \right)$$

- Decision tree structure of MLLR is fixed
- 1. Estimate mean/cov for full context-dept HMMs given MLLR
- 2. Split each leaf node using applicable context questions
- 3. Calculate likelihood increase of each split
 - Re-estimate new mean/cov of each leaf node
 - Evaluate likelihood of the whole data set
- 4. Choose the question yielding the largest likelihood increase and split the corresponding leaf node
- 5. Go to 3 until the stopping criterion is met

Simultaneous Decision Tree Construction

- 1. Estimate parameters for full context-dependent HMMs
- 2. Create root nodes for each context factor
- 3. Split all leaf nodes of all trees using applicable context questions
- 4. Calculate likelihood increase of each split
 - Identify the parameter set associated with the tree
 - Re-estimate the parameters given the split and the other sets of parameters
 - Evaluate likelihood of the whole data set
- 5. Choose the question and the leaf node which yields the largest likelihood increase and split it
- 6. Go to 3 until the stopping criterion is met

Canonical State Model



- Construct “canonical” state models
- Each context-dependent model is transformed from the canonical state models

General Form of Canonical State Model

- Canonical state model is a large global GMM

$$p(\mathbf{o}|s_g) = \sum_{m=1}^M c_g^{(m)} \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_g^{(m)}, \boldsymbol{\Sigma}_g^{(m)})$$

- Context specific state models are transformed from it

$$p(\mathbf{o}|s) = \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$$

$$\boldsymbol{\mu}_s = \mathcal{F}_\mu(\boldsymbol{\Lambda}_{s_g}; \theta_s)$$

$$\boldsymbol{\Sigma}_s = \mathcal{F}_\Sigma(\boldsymbol{\Lambda}_{s_g}; \theta_s)$$

$$\boldsymbol{\Lambda}_{s_g} = \{c^{(m)}, \boldsymbol{\mu}_g^{(m)}, \boldsymbol{\Sigma}_g^{(m)}\}, m = 1, \dots, M$$

Forms of state specific transform

- Gaussian Selection

$$\mathcal{F}_\mu(\Lambda_{s_g}; \theta_s) = \sum_m \delta(m - m_s) \mu_m \quad \mathcal{F}_\Sigma(\Lambda_{s_g}; \theta_s) = \sum_m \delta(m - m_s) \Sigma_m$$

- Parameter Interpolation

$$\mathcal{F}_\mu(\Lambda_{s_g}; \theta_s) = \sum_m \lambda_{sm} \mu_m \quad \mathcal{F}_\Sigma(\Lambda_{s_g}; \theta_s) = \sum_m \delta(m - m_s) \Sigma_m$$

- Linear transform

$$\mathcal{F}_\mu(\Lambda_{s_g}; \theta_s) = \sum_m \delta(m - m_s) \mathbf{A}_s \mu_m + \mathbf{b}_s \quad \mathcal{F}_\Sigma(\Lambda_{s_g}; \theta_s) = \sum_m \delta(m - m_s) \Sigma_m$$

- Combined Transformations

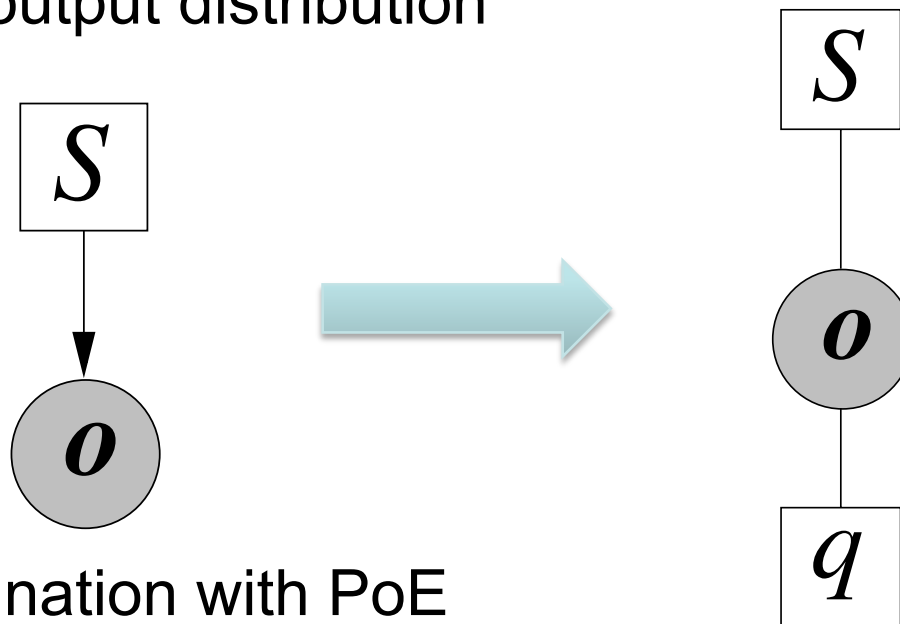
$$\mathcal{F}_\mu(\Lambda_{s_g}; \theta_s) = \mathbf{A}_s \sum_m \lambda_{sm} \mu_m + \mathbf{b}_s \quad \mathcal{F}_\Sigma(\Lambda_{s_g}; \theta_s) = \sum_m \delta(m - m_s) \Sigma_m$$

Comments on Canonical State Model

- Canonical state model is a more general form of context adaptive training
- Initialization of canonical state model
 - Data driven
 - Prior knowledge
- State clustering given canonical state model
 - Standard decision tree clustering with adapted full context model
 - Dependent decision tree clustering of context transformations

Product of Expert for Context Modelling

- PoE for state output distribution



- Context combination with PoE
 - Multiple context groups are modelled separately
 - Contexts are always time synchronous
 - Directly model acoustic property

$$p(\mathbf{o}|c_1, \dots, c_S) = \frac{1}{Z} \prod_{s=1}^S p(\mathbf{o}|\mathcal{M}_s) \quad Z = \int_{\mathbf{o}} \prod_{s=1}^S p(\mathbf{o}|\mathcal{M}_s) d\mathbf{o}$$

Use Gaussian as expert

$$p(\mathbf{o}|c_1, \dots, c_S) = \frac{1}{Z} \prod_{s=1}^S \mathcal{N}(\mathbf{o}; \mu_s, \Sigma_s) = \mathcal{N}(\mathbf{o}; \mu, \Sigma)$$

- c_1, \dots, c_S are context FACTORS
- Full context $\mathbf{c}=[c_1, \dots, c_S]$
 - Number of full contexts: $\prod_{s=1}^S N_s$
 - Number of factorized contexts: $\sum_{s=1}^S N_s$
- Product of Gaussian results in Gaussian

$$\mu = \Sigma \left(\sum_{s=1}^S \Sigma_s^{-1} \mu_s \right) \quad \Sigma = \left(\sum_{s=1}^S \Sigma_s^{-1} \right)$$

- Easy to calculate likelihood

Parameter Estimation In PoG

$$Q = -\frac{1}{2} \sum_{t,c} \gamma_c(t) \left(\log |\Sigma_c| + (\mathbf{o}_t - \mu_c)^T \Sigma_c^{-1} (\mathbf{o}_t - \mu_c) \right) \quad c = [c_1, \dots, c_S]$$

- c is full context label
- Given covariance matrices and other comp. mean vectors, mean update has closed-form solution

$$\mu_s = \left(\sum_{t, c_i=s} \gamma_c(t) \Sigma_c \right)^{-1} \left(\sum_{t, c_i=s} \gamma_c(t) (\mathbf{o}_t - \mathbf{b}_{c_i}) \right)$$
$$\Sigma_c = \left(\sum_{i=1}^S \Sigma_{c_i}^{-1} \right)^{-1} \quad \mathbf{b}_{c_i} = \Sigma_c \left(\sum_{j=1, j \neq i}^S \Sigma_{c_j}^{-1} \mu_{c_j} \right)$$

- Simultaneous mean update and covariance update does not have closed-form solution
- Gradient descent approach to be used

State clustering with PoG

- Independent construction
 - Build decision tree using separate question sets
 - Perform intersection to get atomic leaf node
- Dependent construction
 - Interleave between multiple sets of model parameters
- Simultaneous construction
 - At each split, all trees are optimized inter-dependently
- No closed form parameter re-estimation -> more computational cost -> approximation required

Wrap Up

- Rich context modelling is crucial for HMM-based speech synthesis
- Straightforward full context modelling has limitations
- Structured context modelling is interesting
 - Context adaptive training and canonical state model concerns context relationship
 - Acoustic condition adaptation is a special form of context adaptive training
 - Product of expert directly models context acoustic property
- State clustering with structured context representation can have various forms