Context Modelling for HMM-Based Speech Synthesis

Kai Yu

Machine Intelligence Lab
Cambridge University Engineering Department

Mar. 16, 2011
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)

- Speech signal
- Excitation parameter extraction
- Spectral parameter extraction
- Training of HMM
- Excitation parameter
- Spectral parameter
- Label
- Context dependent HMMs
- Parameter generation from HMM
- Synthesis filter
- SYNTHESIZED SPEECH

TEXT
Text analysis

- Excitation generation
- Spectral parameter
HMM-based Speech Synthesis

Composite HMMs

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

\[
\mathcal{M} = \{\mathcal{M}_a, \cdots, \mathcal{M}_b\}
\]
## Rich contexts of phone in speech synthesis

### Context features of **eh**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbouring phones</td>
<td>Left: s</td>
</tr>
<tr>
<td></td>
<td>Right: n</td>
</tr>
<tr>
<td>Position</td>
<td>2\textsuperscript{nd} phone from word start</td>
</tr>
<tr>
<td></td>
<td>4\textsuperscript{th} phone from word end</td>
</tr>
<tr>
<td>Stress/Accent</td>
<td>Current phone stressed</td>
</tr>
<tr>
<td>Linguistic role</td>
<td>Noun, object</td>
</tr>
<tr>
<td>Emphasis</td>
<td>Current word emphasized</td>
</tr>
<tr>
<td></td>
<td>Previous word not emphasized</td>
</tr>
</tbody>
</table>
Significantly increased model complexity

- One HMM per context

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

![Decision tree diagram](image)

Whether the left phone is unvoiced consonant?
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

<table>
<thead>
<tr>
<th>Phonectic</th>
<th>Left phone is vowel?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Current phone is “aa”?</td>
</tr>
<tr>
<td></td>
<td>Current syllable is stressed?</td>
</tr>
<tr>
<td>Position</td>
<td>Current phone is the 3\textsuperscript{rd} one of the syllable from backward?</td>
</tr>
<tr>
<td>Number</td>
<td>The number of phones in the current syllable is 3?</td>
</tr>
</tbody>
</table>

- Efficient likelihood calculation from statistics

\[
\mathcal{L}(S) = -\frac{1}{2} \sum_{t,s \in S} \gamma_s(t) \log \mathcal{N} \left( o_t; \mu(S), \Sigma(S) \right) \\
= -\frac{\gamma(S)}{2} \left( \log |\Sigma(S)| + K \right)
\]
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

![Equation]

\[ l(\mathcal{M}_i) = - \log p(\mathcal{D}|\mathcal{M}_i^{ML}) + \lambda \frac{\alpha_i}{2} \log N_D + K \]

- Likelihood of data given ML estimate
- Number of free parameters
- Data points or total occupancy

Manual scaling factor
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

<table>
<thead>
<tr>
<th>Phone Identity</th>
<th>Position</th>
<th>Counts</th>
<th>Accent/Stress</th>
<th>Part-of-Speech</th>
<th>Emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>mgc</td>
<td>192</td>
<td>8</td>
<td>16</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>lf0</td>
<td>368</td>
<td>98</td>
<td>10</td>
<td>320</td>
<td>25</td>
</tr>
<tr>
<td>dur</td>
<td>221</td>
<td>49</td>
<td>137</td>
<td>4</td>
<td>14</td>
</tr>
</tbody>
</table>

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

\[ p(o|s, w) = p(o|M(s, w)) \quad M(s, w) = F_w(M_s) \]

- \( s \) and \( w \) are context FACTORS, each can contain several context features
- Homogeneity assumption:
  - \( M_s \) and \( F_w \) 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 $c=[s,w]$
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.

Emphasis

Phonetic/Position

Re  Rp  Rc (ATOMIC)
- 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
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

\[
\hat{\Lambda}_{rc} = \mathcal{F}_{re} (\Lambda_{rp}) \quad r_c = r_p \cap r_e
\]
\[
\hat{\mu}_m = A_{re(m)} \mu_{rp(m)} + b_{re(m)} = W_{re(m)} \xi_{rp(m)}
\]
\[
\hat{\Sigma}_m = \Sigma_{rp(m)}
\]

\[
W_{re,d} = G_{re,d} \kappa_{re,d}
\]
\[
\mu_{rp} = G_{rp}^{-1} \kappa_{rp}
\]
\[
G_{re,d} = \sum_t \sum_{m \in r_e} \frac{\gamma_m(t)}{\sigma_{rp(m)}^2} \xi_{rp(m)} \xi_{rp(m)}^T
\]
\[
k_{re,d} = \sum_t \sum_{m \in r_e} \frac{\gamma_m(t) o_{t,d}}{\sigma_{rp(m)}^2} \xi_{rp(m)}
\]
\[
G_{rp} = \sum_t \sum_{m \in r_p} \gamma_m(t) A_{re(m)}^T \Sigma_m^{-1} A_{re(m)}
\]
\[
k_{rp} = \sum_t \sum_{m \in r_p} \gamma_m(t) A_{re(m)}^T \Sigma_m^{-1} (o_t - b_{re(m)})
\]
\[
\Sigma_{rp} = \text{diag} \left( \frac{\sum_{t,m \in r_p} \gamma_m(t) (o_t - \hat{\mu}_m)(o_t - \hat{\mu}_m)^T}{\sum_{t,m \in r_p} \gamma_m(t)} \right)
\]
Cluster based approach

- 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

\[ \hat{\Lambda}_{rc} = \mathcal{F}_{rt}(\Lambda_{rp}, \Lambda_{re}) \quad r_c = r_p \cap r_e \]

\[ \hat{\mu}_m = \lambda_{rt(m)}^{(p)} \mu_{rp(m)} + \lambda_{rt(m)}^{(e)} \mu_{re(m)} = M_m \lambda_{rt(m)} \]

\[ \hat{\Sigma}_m = \Sigma_{rp(m)} \]

\[ G_{rt} = \sum_t \sum_{m \in r_t} \gamma_m(t) M_m^T \Sigma_{rp(m)}^{-1} M_m \]

\[ k_{rt} = \sum_{m \in r_t} M_m^T \Sigma_{rp(m)}^{-1} \sum_t \gamma_m(t) o_t \]

\[ \hat{\lambda}_{rt} = G_{rt}^{-1} k_{rt} \]

\[ \hat{\mu} = G^{-1} k \]

\[ \hat{\mu} = \left[ \hat{\mu}_{r=1}^T \ldots \hat{\mu}_{r=N(p)}^T \hat{\mu}_{e=1} \ldots \hat{\mu}_{e=N(e)} \right]^T \]

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

SSS (2011/03/16) Kai Yu - Context Modelling For HMM Based Speech Synthesis
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
  \[ \mathcal{L}(S) = -\frac{1}{2} \sum_{t, s \in S} \gamma_s(t) \left( \log |\Sigma(S)| + (o_t - A_{r_s}\mu(S) - b_{r_s})^T \Sigma^{-1}(S)(o_t - A_{r_s}\mu(S) - b_{r_s}) + K \right) \]

- 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
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

\[ p(o|s_g) = \sum_{m=1}^{M} c_g^{(m)} \mathcal{N}(o; \mu_g^{(m)}, \Sigma_g^{(m)}) \]

- Context specific state models are transformed from it

\[ p(o|s) = \mathcal{N}(o; \mu_s, \Sigma_s) \]

\[ \mu_s = \mathcal{F}_\mu(\Lambda_{sg}; \theta_s) \]

\[ \Sigma_s = \mathcal{F}_\Sigma(\Lambda_{sg}; \theta_s) \]

\[ \Lambda_{sg} = \{c^{(m)}, \mu_g^{(m)}, \Sigma_g^{(m)}\}, m = 1, \ldots, M \]
Forms of state specific transform

- Gaussian Selection

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

- Parameter Interpolation

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

- Linear transform

\[ \mathcal{F}_\mu(\Lambda_{sg}; \theta_s) = \sum_m \delta(m - m_s) A_s \mu_m + b_s \quad \mathcal{F}_\Sigma(\Lambda_{sg}; \theta_s) = \sum_m \delta(m - m_s) \Sigma_m \]

- Combined Transformations

\[ \mathcal{F}_\mu(\Lambda_{sg}; \theta_s) = A_s \sum_m \lambda_{sm} \mu_m + b_s \quad \mathcal{F}_\Sigma(\Lambda_{sg}; \theta_s) = \sum_m \delta(m - m_s) \Sigma_m \]
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

\[
p(o|c_1, \cdots, c_S) = \frac{1}{Z} \prod_{s=1}^{S} p(o|M_s) \quad Z = \int \prod_{s=1}^{S} p(o|M_s) \, do
\]
Use Gaussian as expert

\[ p(o|c_1, \ldots, c_S) = \frac{1}{Z} \prod_{s=1}^{S} N(o; \mu_s, \Sigma_s) = N(o; \mu, \Sigma) \]

- \(c_1, \ldots c_S\) are context FACTORS
- Full context \(c=[c_1, \ldots c_S]\)
  - Number of full contexts:
  - Number of factorized contexts:
- 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

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

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

\[
\mu_s = \left( \sum_{t,c_i=s} \gamma_c(t) \Sigma_c \right)^{-1} \left( \sum_{t,c_i=s} \gamma_c(t) (o_t - b_{ci}) \right)
\]

\[
\Sigma_c = \left( \sum_{i=1}^{S} \Sigma_{c_i}^{-1} \right)^{-1} \quad b_{ci} = \Sigma_c \left( \sum_{j=1, j \neq i} \Sigma_{c_j}^{-1} \mu_{cj} \right)
\]

- 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
- 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