HIDDEN SEMI-MARKOV MODEL BASED SPEECH SYNTHESIS

Authors: Heiga ZEN †
Keiichi TOKUDA †
Takashi MASUKO ††
Takao KOBAYASHI ††
Tadashi KITAMURA †

Presenter: Heiga ZEN †

† Nagoya Inst. of Tech. †† Tokyo Inst. of Tech.
Outline

Introduce HSMM instead of HMM into HTS

- Summarize unit selection and HMM-based approach
- System overview of the HTS
  - Training part
  - Synthesis part
  - Problems of current system
- Solution ⇒ Introduce HSMM into HTS
- Experiments
- Conclusion & Future works
- Demonstrations
Corpus-based speech synthesis system

Unit selection and concatenation (e.g., CHATR)
- High quality (sometimes discontinuous)
- Require large memory and disk space
- Difficult to convert their voice characteristics

Speech synthesis from HMMs themselves (e.g., HTS)
- Vocoded speech (smooth and stable)
- Small footprint (less than 1MB)
- Easy to convert its voice characteristics
System overview of HTS

Training part

- Speech signal
  - F0 Extraction
    - F0
    - Mel-cepstral Analysis
      - Mel-cepstral Coefficients

- Label
  - Training of HMM
    - Parameter Generation from HMM
      - Label
      - F0
      - Mel-cepstral Coefficients

Synthesis part

- TEXT
  - Text Analysis
    - Label

- Excitation Generation
  - MLSA Filter
  - Synthesized Speech
Training part of HTS

- Training data
  - Initialization and Reestimation
  - Copy CI-HMMs to CD-HMMs
    - Embedded Reestimation
      - Tree-based clustering (F0)
      - Tree-based clustering (Spectra)
        - Embedded Reestimation
          - CD-label sequence
            - Tree-based clustering (Duration)

Context-Dependent HMMs and Duration Models

- Spectra
- F0
- Duration
Synthesis part of HTS

Text analysis

State Duration Distributions

Sentence HMM

State Durations

Mel-cepstrum

F0

Excitation Generation

MLSA Filter

SYNTHESIZED SPEECH

Parameter Generation from HMM

Context-Dependent HMMs and Duration Models
Inconsistencies (1)

Last iteration of embedded reestimation

HMMs

Estimate

Consistent

Inconsistent !

statistical variables

$\alpha_t(i)$ : Fw. prob.

$\beta_t(k)$ : Bw. prob.

Duration Models

Clustering
Inconsistencies (2)

Training ⇒ *Without* explicit duration model
Synthesis ⇒ *With* explicit duration model

Training

```
q_1 -> q_2 -> q_3
```

```
o_1 -> q_1
q_2 -> q_3
o_2 -> q_3
o_3 -> q_3
```

Synthesis

```
q_1 -> q_2 -> q_3
```

```
q_1^D
q_2^D
q_3^D
```

```
q_1^D -> q_1
q_2^D -> q_2
q_3^D -> q_3
```

Graphical model representations

Different kind of generative model!

Inconsistency between training and synthesis part!
Hidden semi-Markov model (HSMM)

- HMM + explicit duration model $\Rightarrow$ HSMM
- HSMM as acoustic models for ASR
  [Russel & Moore;'85, Levinson;'86, Rabinar;'90]
- Computational cost is high
  $\Rightarrow$ Not used in state-of-the-art ASR systems
  For HTS, computational cost is not so high!
- Introduce HSMM into HTS (training part)
  $\Rightarrow$ Estimate model params. with explicit dur. models
  Inconsistencies can be solved!
Training procedures

HMM-based system

- Training data
- Initialization and Reestimation
- Copy CI-HMMs to CD-HMMs
- Embedded Reestimation
- Tree-based clustering (Spectra)
- Tree-based clustering (F0)
- Tree-based clustering (Duration)
- Embedded Reestimation
- Context-Dependent HMMs and Duration Models

HSMM-based system

- Training data
- Initialization and Reestimation
- Copy CI-HMMs to CD-HMMs
- Embedded Reestimation
- Tree-based clustering (Spectra)
- Tree-based clustering (F0)
- Tree-based clustering (Duration)
- Embedded Reestimation
- Context-Dependent HSMMs
## Experimental conditions

<table>
<thead>
<tr>
<th>Training data</th>
<th>ATR Japanese speech database B-set Speaker MHT, MSH, FYI, and FTK first 450 utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test data</td>
<td>Remaining 53 utterances</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>16 kHz</td>
</tr>
<tr>
<td>Window</td>
<td>25-ms Blackman window</td>
</tr>
<tr>
<td>Frame rate</td>
<td>5-ms</td>
</tr>
<tr>
<td>Spectral analysis</td>
<td>24-order Mel-cepstral analysis</td>
</tr>
<tr>
<td>Dynamic feature</td>
<td>calculated from ±1 frames</td>
</tr>
<tr>
<td>Feature vector</td>
<td>$c(0) \sim c(24)$, $\log F0$, and its $\Delta$, $\Delta\Delta$</td>
</tr>
<tr>
<td>Topology</td>
<td>5-state left-to-right HMM / HSMM Spectrum : single Gaussian distribution</td>
</tr>
<tr>
<td></td>
<td>F0 : multi-space probability distribution</td>
</tr>
</tbody>
</table>
# Constructed Models

## #leaf-nodes after tree-based context clustering

<table>
<thead>
<tr>
<th>speaker</th>
<th>model</th>
<th>Spec.</th>
<th>$F_0$</th>
<th>Dur.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTK</td>
<td>HMM</td>
<td>956</td>
<td>1392</td>
<td>404</td>
</tr>
<tr>
<td></td>
<td>HSMM</td>
<td>963</td>
<td>1427</td>
<td>343</td>
</tr>
<tr>
<td>FYM</td>
<td>HMM</td>
<td>870</td>
<td>1365</td>
<td>368</td>
</tr>
<tr>
<td></td>
<td>HSMM</td>
<td>874</td>
<td>1360</td>
<td>343</td>
</tr>
<tr>
<td>MHT</td>
<td>HMM</td>
<td>969</td>
<td>1133</td>
<td>338</td>
</tr>
<tr>
<td></td>
<td>HSMM</td>
<td>969</td>
<td>1150</td>
<td>313</td>
</tr>
<tr>
<td>MYI</td>
<td>HMM</td>
<td>728</td>
<td>1234</td>
<td>377</td>
</tr>
<tr>
<td></td>
<td>HSMM</td>
<td>737</td>
<td>1217</td>
<td>361</td>
</tr>
</tbody>
</table>

#model-parameters were almost the same
Subjective listening test results

<table>
<thead>
<tr>
<th>Test type</th>
<th>Paired comparison test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>8 graduate students</td>
</tr>
<tr>
<td>Test sentences</td>
<td>20 test sentences were chosen at random</td>
</tr>
</tbody>
</table>

![Diagram showing subjective listening test results for different models: HMM and HSMM. The confidence intervals are represented by bars and the 95% confidence interval is indicated by a solid line.](image-url)
Conclusion

Introduce HSMM instead of HMM into HTS

- HMM + explicit duration model $\Rightarrow$ HSMM
- Spec., F0 & Dur. were reestimated simultaneously
- Solve some inconsistencies in the current system
- Synthetic speech quality $\Rightarrow$ Equal or better

Future works

- Try other distributions for duration modeling
  (Gamma, logarithmic Gaussian, nonparametric, etc.)
- Make system more consistent…
Synthesized speech samples

Japanese

"Chiisana unagiyani nekkino yona monoga minagiru"

HMM 🎧 HSMM 🎧

"Dorobo demo haittakato isshun bokuwa omotta"

HMM 🎧 HSMM 🎧

"Tokaiedewa deau hitono hotondoga misiranu hitode aru"

HMM 🎧 HSMM 🎧

English (trained from CMU ARCTIC database)

"I remembered the red wine of the Italian rancho, and shuddered inwardly."

HMM 🎧 HSMM 🎧

"I said, and dismissed the matter as not worth thinking about."

HMM 🎧 HSMM 🎧

"Then came my boy code."

HMM 🎧 HSMM 🎧
Thanks !