An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005

Heiga Zen†
Tomoki Toda‡‡

† Nagoya Institute of Technology
‡‡ Nara Institute of Science & Technology
Main paradigms of corpus-based speech synthesis

Sample-based (e.g., unit selection)
- High quality (sometimes discontinuous)
- Large footprint
- Difficult to convert its voice characteristics

Statistics-based (e.g., HMM-based synthesis)
- Vocoded (smooth & stable)
- Small footprint (less than 2MB)
- Easy to convert its voice characteristics

We developed the HMM-based synthesis (Nitech-HTS) & participated in the Blizzard Challenge 2005
Overview of the basic HMM-based speech synthesis

Training part

- SPEECH DATABASE
- Speech signal
- F0 Extraction
- Mel-cepstral Analysis
- Mel-cepstrum
- log F0
- Label
- Training HMM
- HMMs & duration models

Synthesis part

- TEXT
- Text Analysis
- Label
- log F0
- Pulse/Noise Excitation
- Excitation
- MLSA Filter
- SYNTHESIZED SPEECH
Overview of the Nitech-HTS voice

**Training part**
- Speech signal
  - F0 Extraction
  - log F0
  - Label
  - TRAINING HSMM
  - STRAIGHT Analysis
  - STRAIGHT Mel-cepstrum & Aperiodicity

**Synthesis part**
- TEXT
  - Festival
  - Label
  - Parameter Generation
    - considering GV
  - Mixed Excitation
  - Mixed Excitation
  - MLSA Filter
  - SYNTHESESIZED SPEECH

1. Front-end
2. Acoustic modeling
3. Param. generation
Overview of the Nitech-HTS voice

Training part

Speech signal

F0 Extraction

F0

log F0

STRAIGHT Analysis

STRAIGHT Mel-cepstrum & Aperiodicity

Training HSMM

Hidden semi-Markov model

Synthesis part

TEXT

Festival

Label

log F0 & Aperiodicity

Parameter Generation considering GV

Mixed Excitation

Excitation

MLSA Filter

SYNTHESIZED SPEECH

1. Front-end

2. Acoustic modeling

3. Param. generation
Overview of the Nitech-HTS voice

Training part

1. Front-end
2. Acoustic modeling
3. Param. generation

Synthesis part

1. Front-end
2. Acoustic modeling
3. Param. generation

Speech signal

F0 Extraction

F0 Analysis

STRAIGHT Mel-cepstrum & Aperiodicity

F0

log F0

Label

HSMMs

Festival

TEXT

1. Front-end
2. Acoustic modeling
3. Param. generation

log F0 & Aperiodicity

Mixed Excitation

Excitation

MLSA Filter

SYNTHESIZED SPEECH

Param. gen. considering GV

Parameter Generation considering GV

HSMMS

1. Front-end
2. Acoustic modeling
3. Param. generation

SPEECH DATABASE

SYNTHESIZED SPEECH

F0

log F0 & Aperiodicity

Label

STRAIGHT Mel-cepstrum & Aperiodicity
Overview of the STRAIGHT

STRAIGHT [Kawahara et al;’99] - High quality vocoder

Waveform

F0 extraction

Fixed-point analysis

Analysis

F0 adaptive spectral smoothing in the time-frequency region

Synthetic waveform

Synthesis

Mixed excitation with phase manipulation

F0

Smoothed spectrum

Aperiodic factors
Hidden semi-Markov model (HSMM)

Basic HMM-based speech synthesis system

<table>
<thead>
<tr>
<th>Training</th>
<th>Without explicit duration model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthesis</td>
<td>With explicit duration model</td>
</tr>
</tbody>
</table>

Inconsistency between training & synthesis
⇒ Degrading quality of synthetic speech

Introducing the HSMM into the HTS [Zen et al.;'04]
- HMM + explicit duration model ⇒ HSMM
- Estimating state output & dur. prob. simultaneously
⇒ Above inconsistency can be solved
Parameter generation considering global variance

Parameter generation from HMM [Tokuda et al.;'00]
- Generating param. trajectory maximizing its output prob.
- Generated params. are often over-smoothed
  ⇒ Muffled speech

Parameter generation considering GV [Toda et al.;'05]
- Generateing param. trajectory maximizing both its output & GV probs.
- GV prob. works as a penalty to prevent over-smoothing
  ⇒ Improves quality of the HTS
## Evaluation of the STRAIGHT, HSMM & GV

<table>
<thead>
<tr>
<th>Training data</th>
<th>ATR Japanese speech database B-set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speakers MYI &amp; FYM, 450 utterances</td>
</tr>
<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>Topology</td>
<td>5-state left-to-right HMM / HSMM</td>
</tr>
<tr>
<td>Test type</td>
<td>Mean Opinion Score (MOS)</td>
</tr>
<tr>
<td>Subjects</td>
<td>12 students</td>
</tr>
<tr>
<td>Test sentences</td>
<td>Randomly selected 10 test sentences</td>
</tr>
</tbody>
</table>
## Settings to evaluate the STRAIGHT, HSMM & GV

<table>
<thead>
<tr>
<th></th>
<th>Spectral analysis</th>
<th>Excitation</th>
<th>Model</th>
<th>GV</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>24 Mel-Cep</td>
<td>Pulse/Noise</td>
<td>HMM</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>24 STRAIGHT Mel-Cep</td>
<td>Mixed</td>
<td>HMM</td>
<td>No</td>
</tr>
<tr>
<td>C</td>
<td>24 Mel-Cep</td>
<td>Pulse/Noise</td>
<td>HSMM</td>
<td>No</td>
</tr>
<tr>
<td>D</td>
<td>24 STRAIGHT Mel-Cep</td>
<td>Mixed</td>
<td>HSMM</td>
<td>No</td>
</tr>
<tr>
<td>E</td>
<td>39 STRAIGHT Mel-Cep</td>
<td>Mixed</td>
<td>HSMM</td>
<td>No</td>
</tr>
<tr>
<td>F</td>
<td>24 Mel-Cep</td>
<td>Pulse/Noise</td>
<td>HSMM</td>
<td>Yes</td>
</tr>
<tr>
<td>G</td>
<td>24 STRAIGHT Mel-Cep</td>
<td>Mixed</td>
<td>HSMM</td>
<td>Yes</td>
</tr>
<tr>
<td>H</td>
<td>39 STRAIGHT Mel-Cep</td>
<td>Mixed</td>
<td>HSMM</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Experimental results (Speaker MYI)

<table>
<thead>
<tr>
<th>Spectra</th>
<th>Model</th>
<th>Par Gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 mcep.</td>
<td>HMM</td>
<td>non-GV</td>
</tr>
<tr>
<td>24 ST.</td>
<td>HMM</td>
<td>non-GV</td>
</tr>
<tr>
<td>24 mcep.</td>
<td>HSMM</td>
<td>non-GV</td>
</tr>
<tr>
<td>24 ST.</td>
<td>HSMM</td>
<td>non-GV</td>
</tr>
<tr>
<td>39 ST.</td>
<td>HSMM</td>
<td>GV</td>
</tr>
<tr>
<td>24 mcep.</td>
<td>HSMM</td>
<td>GV</td>
</tr>
<tr>
<td>24 ST.</td>
<td>HSMM</td>
<td>GV</td>
</tr>
<tr>
<td>39 ST.</td>
<td>HSMM</td>
<td>GV</td>
</tr>
</tbody>
</table>

Average MOS

- A: 2.0
- B: 3.1
- C: 2.0
- D: 3.2
- E: 3.2
- F: 3.1
- G: 4.7
- H: 4.7

Spectra: STRAIGHT & GV
Experimental results (Speaker MYI)

<table>
<thead>
<tr>
<th>Spectra Model</th>
<th>Par Gen</th>
<th>Average MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 mcep. HMM</td>
<td>non-GV</td>
<td>2.0</td>
</tr>
<tr>
<td>24 ST. HMM</td>
<td>non-GV</td>
<td>3.1</td>
</tr>
<tr>
<td>24 mcep. HSMM</td>
<td>non-GV</td>
<td>3.2</td>
</tr>
<tr>
<td>24 ST. HSMM</td>
<td>non-GV</td>
<td>3.2</td>
</tr>
<tr>
<td>39 ST. HSMM</td>
<td>non-GV</td>
<td>3.1</td>
</tr>
<tr>
<td>24 mcep. HSMM</td>
<td>GV</td>
<td>4.7</td>
</tr>
<tr>
<td>24 ST. HSMM</td>
<td>GV</td>
<td>4.7</td>
</tr>
<tr>
<td>39 ST. HSMM</td>
<td>GV</td>
<td>4.7</td>
</tr>
</tbody>
</table>
Experimental results (Speaker FYM)

Average MOS

<table>
<thead>
<tr>
<th>Spectra Model</th>
<th>Par Gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 mcep.</td>
<td>non-GV</td>
</tr>
<tr>
<td>24 ST.</td>
<td>non-GV</td>
</tr>
<tr>
<td>24 mcep.</td>
<td>non-GV</td>
</tr>
<tr>
<td>24 ST.</td>
<td>non-GV</td>
</tr>
<tr>
<td>39 ST.</td>
<td>non-GV</td>
</tr>
<tr>
<td>24 mcep.</td>
<td>GV</td>
</tr>
<tr>
<td>24 ST.</td>
<td>GV</td>
</tr>
<tr>
<td>39 ST.</td>
<td>GV</td>
</tr>
</tbody>
</table>

HSMM & Order
## Setting of the Nitech-HTS voices

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training data</strong></td>
<td>CMU ARCTIC database&lt;br&gt;Speakers BDL, SLT, CLB &amp; RMS&lt;br&gt;1132 utterances</td>
</tr>
<tr>
<td><strong>Sampling rate</strong></td>
<td>16 kHz</td>
</tr>
<tr>
<td><strong>Window</strong></td>
<td>25-ms Blackman window</td>
</tr>
<tr>
<td><strong>Frame rate</strong></td>
<td>5-ms</td>
</tr>
<tr>
<td><strong>Spectral analysis</strong></td>
<td>39-order STRAIGHT Mel-cepstral analysis</td>
</tr>
<tr>
<td><strong>Feature vector</strong></td>
<td>Spectra: $c(0) \sim c(39)$, their $\Delta$ &amp; $\Delta\Delta$&lt;br&gt;F0: $\log F0$, its $\Delta$ &amp; $\Delta\Delta$&lt;br&gt;Aperiodicity: Average on 5 frequency bands ($0-1 \cdot 1-2 \cdot 2-4 \cdot 4-6 \cdot 6-8$kHz), their $\Delta$ &amp; $\Delta\Delta$</td>
</tr>
<tr>
<td><strong>Topology</strong></td>
<td>5-state left-to-right HSMM</td>
</tr>
</tbody>
</table>
Real time ratio

Machine: Pentium-4 1.6 GHz

About 3 times faster than the real time
Listening test results (MOS)

- **Speech synthesis experts:**
  - Natural: 3.19
  - Nitech-HTS: 3.11
  - The best of the others: 4.71

- **Real users from web:**
  - Natural: 4.53
  - Nitech-HTS: 3.06
  - The best of the others: 2.86

- **US-English native undergraduates:**
  - Natural: 4.76
  - Nitech-HTS: 3.38
  - The best of the others: 2.97
<table>
<thead>
<tr>
<th></th>
<th>conv01</th>
<th>guten01</th>
<th>news01</th>
<th>mrt01</th>
<th>sus01</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDL</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
</tr>
<tr>
<td>CLB</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
</tr>
<tr>
<td>RMS</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
</tr>
<tr>
<td>SLT</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
</tr>
</tbody>
</table>
Conclusion

Nitech-HTS voice for the Blizzard Challenge 2005

- Integrate state-of-the-art technologies
  ▪ STRAIGHT
  ▪ Hidden semi-Markov model
  ▪ Parameter generation considering global variance

- Constructed voices
  ▪ Small footprint (less than 2MB)
  ▪ 3 times faster than the real time
  ▪ Listening test results were better than we expected