The Nitech-NAIST HMM-based Speech Synthesis System for the Blizzard Challenge 2006

Heiga Zen† Tomoki Toda†† Keiichi Tokuda†

† Nagoya Institute of Technology
†† Nara Institute of Science & Technology
Background

Evaluating speech synthesis systems

- Quality highly depends on datasets
- Difficult to compare technologies themselves

Blizzard Challenge

- Compares TTS systems on common speech datasets
  - 2005: CMU ARCTIC databases (1 hour)
  - 2006: ATR US-English speech database (5 hours)
- Large scale subjective listening tests

Nitech & NAIST jointly participated in this challenge
Overview of Nitech-HTS 2005

Training part

- Speech signal
- F0 Extraction
- STRAIGHT Analysis
- STRAIGHT Mel-cepstrum & Aperiodicity
- Parameter Generation considering GV
- Mixed Excitation
- MLSA Filter
- Synthesized SPEECH

Synthesis part

- TEXT
- Festival
- Label
- log F0 & Aperiodicity
- STRAIGHT Mel-cepstrum
Overview of Nitech-NAIST-HTS 2006

**Training part**

1. **Speech signal**
   - Speech DATABASE
   - F0 Extraction
     - log F0
     - Label
   - STRAIGHT Analysis
     - F0
   - STRAIGHT MGC-LSP & Aperiodicity
   - Training HSMM

**Synthesis part**

1. **TEXT**
   - Festival
   - Parameter Generation considering GV
     - Log F0 & Aperiodicity
   - Mixed Excitation
   - Excitation
   - MGLSA Filter
   - SYNTHESIZED SPEECH
Overview of Nitech-NAIST-HTS 2006

Training HSMM

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

Synthesis part

Parameter Generation considering GV

MGLSA Filter

SYNTHESIZED SPEECH

Festival

log F0 & Aperiodicity

Mixed Excitation

Excitation

STRAIGHT MGC-LSP & Aperiodicity

STRAIGHT Mel-Generalized Cep. based LSP (MGC-LSP)

TEXT

SPEECH DATABASE

Speech signal

F0 Extraction

F0

log F0

Label

Training HSMM

STRAIGHT MGC-LSP

STRAIGHT Mel-Gen. Cep. based LSP

(STRAIGHT MGC-LSP)
Overview of Nitech-NAIST-HTS 2006

Training part

SPEECH DATABASE

Speech signal

F0 Extraction

F0

log F0

STRAIGHT MGC-LSP & Aperiodicity

Maximum Likelihood Linear Transform

HSMMs

Parameter Generation considering GV

log F0 & Aperiodicity

Mixed Excitation

Excitation

MGLSA Filter

SYNTHESIZED SPEECH

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

Synthesis part

TEXT

Festival

Label

SYNTHETIZED SPEECH DATABASE

F0

Extraction

STRAIGHT MGC-LSP

Analysis

STRAIGHT MGC-LSP
Overview of Nitech-NAIST-HTS 2006

Training part

- Speech signal
- F0 Extraction
- STRAIGHT F0 Analysis
- STRAIGHT MGC-LSP & Aperiodicity
- Training HSMM

Synthesis part

- TEXT
- Festival
- Parameter Generation considering GV
- Full covariance GV pdf
- Mixed Excitation
- MGLSA Filter
- SYNTHESIZED SPEECH
Mel-Generalized-Cepstral (MGC) Analysis

Spectral model in MGC analysis [Tokuda et al.;'94]

\[
H(z) = \begin{cases} 
\left( 1 + \gamma \sum_{m=0}^{M} c(m) \tilde{z}^{-m} \right)^{1/\gamma}, & -1 \leq \gamma < 0 \\
\exp \sum_{m=0}^{M} c(m) \tilde{z}^{-m}, & \gamma = 0 
\end{cases}
\]

\[
\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1
\]

- Includes all-pole and cepstral representations
- Can control frequency resolution by $\alpha$

\[
\text{e.g.) } (\gamma, \alpha) = (-1, 0) \Rightarrow \text{LPC}, \quad (\gamma, \alpha) = (0, 0) \Rightarrow \text{Cepstrum}
\]

\[
\text{e.g.) } \alpha = 0.42 \Rightarrow \text{Mel scale (16 kHz sampling)}
\]
Examples of Speech Spectra

Speech spectra represented by MGC ($M = 24$)

<table>
<thead>
<tr>
<th>$\alpha = 0.0$ (Linear)</th>
<th>$\alpha = 0.42$ (Mel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>Mel-LPC</td>
</tr>
<tr>
<td>Generalized -Cepstrum</td>
<td>Mel-Generalized -Cepstrum</td>
</tr>
<tr>
<td>Cepstrum</td>
<td>Mel-Cepstrum</td>
</tr>
</tbody>
</table>

$\gamma = -1$ (all-pole)  $\gamma = -1/3$  $\gamma = 0$ (cepstral)
Mel-Generalized-Cepstrum based LSP (MGC-LSP)

If $\gamma \neq 0$, MGC can be reformulated as ...

$$H(z) = \frac{\tilde{K}}{\{C(\tilde{z})\}^n} \quad n = -1/\gamma$$

Consider $C(\tilde{z})$ as an LPC polynomial $\Rightarrow$ MGC-LSP

- MGC-LSP has **better quantization/interpolation property** than Mel-Cepstrum, LSP, & MGC [Koishida et al.;'96]
- Applying MGC-LSP to HTS [Marume et al.;'06]
  - **Significant improvements** over Mel-Cepstrum
  - Evaluated in the basic system rather than the latest one
Speech Parameter Generation Considering GV

Objective function

\[ \mathcal{L} = w \log P( \mathbf{Wc} \mid q, \lambda) + \log P(\nu(\mathbf{c}) \mid \lambda_{\nu}) \]

- \( \mathbf{c} \): Speech param. vector sequence (static feature only)
- \( \nu(\mathbf{c}) \): GV of \( \mathbf{c} \) (intra-utterance variance)
- \( P(\mathbf{Wc} \mid q, \lambda) \): Output prob. of \( \mathbf{c} \) from an HSMM \( \lambda \) for \( q \)
- \( P(\nu(\mathbf{c}) \mid \lambda_{\nu}) \): Output prob. of GV of \( \mathbf{c} \)

Generates \( \mathbf{c} \) which maximizes this objective function

- Reduces buzziness in synthesized speech (^_^)
- Sometimes artificial sounds are generated (>_<)
Artificial Sounds Caused by GV

Why artificial sounds are synthesized by using GV?

Diagonal cov. matrices are used for state output pdfs

Correlations among cepstral features are ignored

Each cepstral feature is optimized independently

Relations among cepstral features may not be kept

Using full covariance for state output pdfs

Can capture correlations among cepstral features

Requires huge amount of training data
Artificial Sounds Caused by GV

Why artificial sounds are synthesized by using GV?

Diagonal cov. matrices are used for state output pdfs

⇒ Correlations among cepstral features are ignored
Artificial Sounds Caused by GV

Why artificial sounds are synthesized by using GV?
Diagonal cov. matrices are used for state output pdfs

⇒ Correlations among cepstral features are ignored
⇒ Each cepstral feature is optimized independently
Artificial Sounds Caused by GV

Why artificial sounds are synthesized by using GV?

Diagonal cov. matrices are used for state output pdfs

⇒ Correlations among cepstral features are ignored
⇒ Each cepstral feature is optimized independently
⇒ Relations among cepstral features may not be kept
Artificial Sounds Caused by GV

Why artificial sounds are synthesized by using GV?

Diagonal cov. matrices are used for state output pdfs
  ⇒ Correlations among cepstral features are ignored
  ⇒ Each cepstral feature is optimized independently
  ⇒ Relations among cepstral features may not be kept

Using full covariance for state output pdfs
  • Can capture correlations among cepstral features (^_^)
  • Requires huge amount of training data (>_<)
Maximum Likelihood Linear Transform (MLLT)

Structured precision (inverse cov.) matrix modeling

- Well approximate full covariance models
- Smaller #params. than full covariance models
- Models: MLLT, STC, EMLLT, SPAM, etc.

Maximum Likelihood Linear Transform [Gopinath;'98]

\[ \Sigma_j^{-1} = A^\top \Lambda_j A \]

- \(A\) is shared over all Gaussians in the model set
- Iterative closed-form update formula [Gales;'99]
Full Covariance GV PDF

Objective function

\[ \mathcal{L} = w \log P(Wc | q, \lambda) + \log P(v(c) | \lambda_v) \]

\[ P(v(c) | \lambda_v) = \mathcal{N}(v(c) | \mu_v, \Sigma_v) \]

- Diagonal covariance is used for \( \Sigma_v \)
  \( \Rightarrow \) Unable to capture correlations

- Try to use full covariance models & its approximations
  \( \Rightarrow \) Apply full covariance & FA
# Evaluations

<table>
<thead>
<tr>
<th>Training data</th>
<th>ATR US-English speech database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 male speaker</td>
</tr>
<tr>
<td></td>
<td>4,273 utterances</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>16 kHz</td>
</tr>
<tr>
<td>Frame rate</td>
<td>5 ms</td>
</tr>
<tr>
<td>Spectral analysis</td>
<td>512-order STRAIGHT analysis</td>
</tr>
<tr>
<td>Feature vector</td>
<td>0-39 Mel-Cepstrum ($\gamma = 0$), $\Delta$ &amp; $\Delta\Delta$</td>
</tr>
<tr>
<td></td>
<td>Log Gain, 1-39 MGC-LSP ($\gamma = -1/3$), $\Delta$ &amp; $\Delta\Delta$</td>
</tr>
<tr>
<td></td>
<td>Log Gain, 1-39 Mel-LSP ($\gamma = -1$), $\Delta$ &amp; $\Delta\Delta$</td>
</tr>
<tr>
<td></td>
<td>log $F0$, $\Delta$ &amp; $\Delta\Delta$</td>
</tr>
<tr>
<td>Average aperiodicities on 5 frequency bands</td>
<td>(0-1·1-2·2-4·4-6·6-8kHz), $\Delta$ &amp; $\Delta\Delta$</td>
</tr>
<tr>
<td>Topology</td>
<td>5-state left-to-right HSMM</td>
</tr>
</tbody>
</table>
## Conditions

<table>
<thead>
<tr>
<th>Test</th>
<th>Objective</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mel-Cep. vs MGC-LSP vs Mel-LSP (w/o GV)</td>
<td>MOS</td>
</tr>
<tr>
<td>2</td>
<td>Mel-Cep. vs MGC-LSP vs Mel-LSP (with GV)</td>
<td>MOS</td>
</tr>
<tr>
<td>3</td>
<td>w/o GV vs diagonal vs FA vs full (GV pdf)</td>
<td>MOS</td>
</tr>
<tr>
<td>4</td>
<td>2005 vs 2006</td>
<td>Pref.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subjects</th>
<th>17 Japanese students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test sentences</td>
<td>Blizzard Challenge '05 (guten+conv, 100)</td>
</tr>
<tr>
<td></td>
<td>Randomly selected 10 sentences</td>
</tr>
</tbody>
</table>
Experimental Results (1)

Without GV

Mean Opinion Scores

Diag MLLT Diag MLLT Diag MLLT

I 95% confidence interval

all-pole Mel-LSP \((\gamma = -1)\) MGC-LSP \((\gamma = -1/3)\) Mel-Cepstrum \((\gamma = 0)\) cepstral
Experimental Results (1)

Without GV

Mean Opinion Scores

<table>
<thead>
<tr>
<th></th>
<th>Diag</th>
<th>MLLT</th>
<th>Diag</th>
<th>MLLT</th>
<th>Diag</th>
<th>MLLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-pole</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mel-LSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma = -1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGC-LSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma = -1/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mel-Cepstrum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma = 0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MGC-LSP achieved the best score

95% confidence interval
Experimental Results (1)

Without GV

<table>
<thead>
<tr>
<th></th>
<th>Diag</th>
<th>MLLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mel-LSP (° = -1)</td>
<td>Diag</td>
<td>MLLT</td>
</tr>
<tr>
<td>MGC-LSP (° = -1/3)</td>
<td>Diag</td>
<td>MLLT</td>
</tr>
<tr>
<td>Mel-Cepstrum (° = 0)</td>
<td>Diag</td>
<td>MLLT</td>
</tr>
</tbody>
</table>

MLLT degraded the performance
Experimental Results (2)

With GV

Mean Opinion Scores

Diag  MLLT  Diag  MLLT  Diag  MLLT

Mel-LSP ($\gamma = -1$)

MGC-LSP ($\gamma = -1/3$)

Mel-Cepstrum ($\gamma = 0$)

95% confidence interval
Experimental Results (2)

With GV

Mean Opinion Scores

- Mel-Cepstrum ($\gamma = 0$)
- Mel-LSP ($\gamma = -1$)
- MGC-LSP ($\gamma = -1/3$)

MGC-LSP achieved the best score (Diag)
Experimental Results (2)

With GV

Mean Opinion Scores

Diag  MLLT  Diag  MLLT  Diag  MLLT

Mel-Cepstrum with MLLT was best

I 95% confidence interval

all-pole  Mel-LSP (γ = −1)  MGC-LSP (γ = −1/3)  Mel-Cepstrum (γ = 0)  cepstral
Experimental Results (3)

Full covariance GV pdf

Mean Opinion Scores

I 95% confidence interval

<table>
<thead>
<tr>
<th>None</th>
<th>Diag</th>
<th>FA</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>3.0</td>
<td>3.5</td>
<td>3.75</td>
</tr>
</tbody>
</table>
Experimental Results (3)

Full covariance GV pdf

Mean Opinion Scores

<table>
<thead>
<tr>
<th>None</th>
<th>Diag</th>
<th>FA</th>
<th>Full</th>
</tr>
</thead>
</table>

95% confidence interval

Use of GV was effective
Experimental Results (3)

Full covariance GV pdf

Mean Opinion Scores

I 95% confidence interval

Full was best, but no significant difference
Experimental Results (4)

2005 vs 2006 (Training data: 5 hours)

2005: Mel-Cepstrum + Diag Cov. + diag GV pdf
2006: Mel-Cepstrum + MLLT Cov. + full GV pdf

2006 system was significantly better than 2005 system
Our HMM-based system was still competitive even a large corpus (5 hours) was used.
Conclusion

Nitech-NAIST-HTS 2006

- Mel-Generalized-Cepstrum based LSP (MGC-LSP)
- Maximum Likelihood Linear Transformation (MLLT)
- Full covariance GV pdf
  ⇒ Mel-Cepstrum, MLLT, & Full cov. GV pdf were used

Blizzard Challenge 2006

- Relatively large corpus (5 hours)
- 14 groups participated
- Nitech-NAIST-HTS 2006 was still competitive

Sample1  Sample2  Sample3

http://hts.ics.nitech.ac.jp/nitech-naist-hts_blizzard2006/
Thanks for listening!! (´ ▽ `)

Any questions? (・ω・ )
System Description

- Footprint: 5 MBytes
- Voice building time: 1 day (9 × PentiumD 3.2GHz)
- Synthesis time: 1 x RT
- Front-end: Festival speech synthesis system
- No manual correction
  (segmentations, text analyzer output, F0, etc)
- No tuning