HMM-based Speaker Interpolation
Advanced Signal Processing Seminar SS 2008

Susanne Rexeis  Matthias Straka

29.04.2008
HMM-based Speaker Interpolation

Introduction

Modeling

Interpolation

Summary
Model different emotional expressions and speaking styles
- Interpolation between styles
- *Goal*: make synthesized speech sound more natural
- Other ideas are based on variation of pitch, loudness and speed
- This approach: context based decision trees
HMM phoneme modeling

Prosodic approaches (Variation of F0 level, loudness, speech tempo)

Tree-based context clustering
  - Style dependend modeling
  - Style mixed modeling
Tree-based context clustering

Why?

- Reduction of distributions
- Splitting conditions
  - phonetic context of phoneme
  - linguistic context of phoneme
- Automatic generation of tree with MDL
- Can handle phonemes in unseen context
Style dependent and style mixed modeling
What’s the difference?

- Style dependent modeling
  - One model for each speaking style
  - Models are connected to one root node

- Style mixed modeling
  - One model for all speaking styles
Style dependend and style mixed modeling
The resulting trees
Style dependent and style mixed modeling
Which one is better?

- Advantages of style dependent modeling
  - High reduction of distributions
  - Easy to add new styles

- Advantages of style mixed modeling
  - Higher reduction of distributions
  - Evaluation results similar to slightly better

- Main disadvantage of style mixed modeling
  - Adding of new styles requires new tree
Style dependent and style mixed modeling
Evaluation results

- 503 test sentences of each style for training
- male and female speaker
- synthesis of 53 unseen sentences in different styles
- 9 test subjects
- 8 randomly drawn test sentences/tester
- style classification

<table>
<thead>
<tr>
<th>Synthetic Speech</th>
<th>Classification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>98.3 0.6 0.0 0.0 1.1</td>
</tr>
<tr>
<td>Rough</td>
<td>6.9 82.3 0.0 0.0 10.8</td>
</tr>
<tr>
<td>Joyful</td>
<td>1.1 0.0 94.9 0.0 4.0</td>
</tr>
<tr>
<td>Sad</td>
<td>0.6 1.1 0.0 94.9 3.4</td>
</tr>
</tbody>
</table>

Figure: Classification results of style dependent modeling, male speaker

<table>
<thead>
<tr>
<th>Synthetic Speech</th>
<th>Classification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>98.9 0.0 0.0 0.0 1.1</td>
</tr>
<tr>
<td>Rough</td>
<td>2.8 89.8 0.0 1.1 6.3</td>
</tr>
<tr>
<td>Joyful</td>
<td>0.6 0.0 96.0 0.0 3.4</td>
</tr>
<tr>
<td>Sad</td>
<td>0.0 0.6 0.0 96.0 3.4</td>
</tr>
</tbody>
</table>

Figure: Classification results of style mixed modeling, male speaker
Interpolate between two or more styles of speech
- Speakers – From male to female and everything in between
- Dialects – American English to Indian English
- Emotions – Happy to angry

Interpolation of Gaussian PDFs
Interpolation
Basics

- $N$ Speaking styles $S_1, S_2, \ldots, S_N$
- Mean vectors $\mu_k$ and covariance matrices $U_k$
- $N$ HMMs $\lambda_1, \lambda_2, \ldots, \lambda_N$
- Weights $a_1, a_2, \ldots, a_N$ with $\sum_{k=1}^{N} a_k = 1$
- $\tilde{\mu}$ and $\tilde{U}$ as the interpolation result
Interpolation
Three different methods

(a) Interpolation among observations

\[ \tilde{\mu} = \sum_{k=1}^{N} a_k \mu_k \]
\[ \tilde{U} = \sum_{k=1}^{N} a_k^2 U_k \]
(b) Interpolation among output distributions

\[ \tilde{\mu} = \sum_{k=1}^{N} a_k \mu_k \]

\[ \tilde{U} = \sum_{k=1}^{N} a_k \left( U_k + \mu_k \mu_k^T \right) - \tilde{\mu} \tilde{\mu}^T \]
(c) Interpolation based on Kullback information measure

\[ \tilde{\mu} = \left( \sum_{k=1}^{N} a_k U_k^{-1} \right)^{-1} \left( \sum_{k=1}^{N} a_k U_k^{-1} \mu_k \right)^{-1} \]

\[ \tilde{U} = \left( \sum_{k=1}^{N} a_k U_k^{-1} \right)^{-1} \]
Interpolation
How to do it exactly?

- Change coefficients \((a_1, a_2)\) gradually from \((1, 0)\) to \((0, 1)\)
- For equally structured models \(\lambda_k\): interpolate directly from \(\lambda_k\)
- Most of the time this is not possible. Alternative:
  - Transform text into context-dependent phoneme labels (synthesis stage)
  - Create HMMs with identical topologies for each style
  - Determine parameters for spectrum, \(F_0\) and state duration
  - Interpolate with these parameters
Interpolation
How to do it exactly?
Used four styles
- neutral
- sad
- joyful
- rough

42 phonemes and various phonetic and linguistic contexts

Parameter extraction
- 25ms windows
- 25 mel-cepstral coefficients

Style modeling with semi-hidden Markov Models (5 left-to-right states)
Summary
What we have learned today

- Two ways to model speaking styles using decision trees
  - Style-dependent modeling
  - Style-mixing modeling
- Building decision trees
- Interpolating between styles
  - Three interpolation equations
- What can we do with it?
  - Style interpolation
  - Gender interpolation
That’s All Folks
... and now it’s your turn to ask questions

Thank You
For Your Attention