HMM-based Speaker Interpolation

Advanced Signal Processing Seminar SS 2008

Susanne Rexeis Matthias Straka

29.04.2008

Introduction
Overview

Introduction
Motivation

- 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

Style Modeling
Different approaches

- HMM phoneme modeling
- Prosodic appoaches (Variation of F0 level, loudness, speech tempo)
- Tree-based context clustering
  - Style dependend modeling
  - Style mixed modeling

Tree-based context clustering
Why?

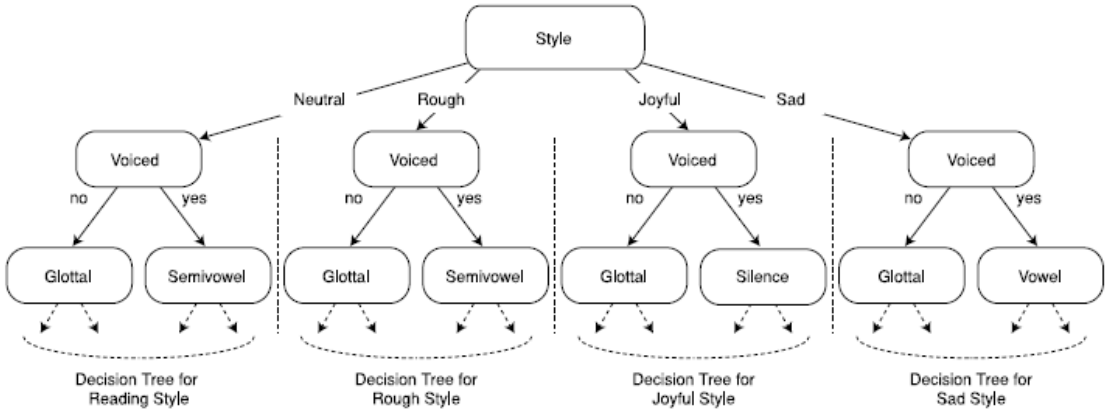
- Reduction of distributions
- Splitting conditions
  - phonetic context of phoneme
  - linguistic context of phoneme
- Automatic generation of tree with MDL
- Can handle phonemes in unseen context

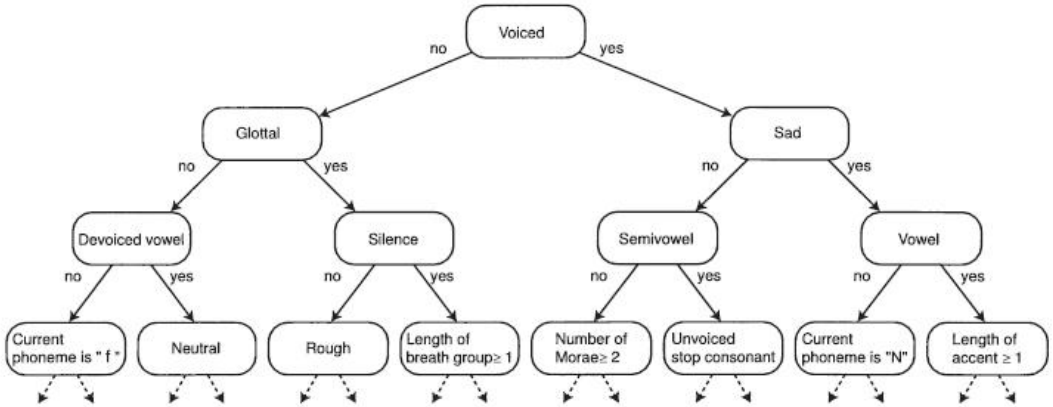
Style dependend and style mixed modeling
What's the difference?

- 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 dependend 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 dependend 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

Synthetic Speech	Classification (%)				
	Neutral	Rough	Joyful	Sad	Other
Neutral	98.3	0.6	0.0	0.0	1.1
Rough	6.9	82.3	0.0	0.0	10.8
Joyful	1.1	0.0	94.9	0.0	4.0
Sad	0.6	1.1	0.0	94.9	3.4

Figure: Classification results of style dependent modeling, male speaker

Synthetic Speech	Classification (%)				
	Neutral	Rough	Joyful	Sad	Other
Neutral	98.9	0.0	0.0	0.0	1.1
Rough	2.8	89.8	0.0	1.1	6.3
Joyful	0.6	0.0	96.0	0.0	3.4
Sad	0.0	0.6	0.0	96.0	3.4

Figure: Classification results of style mixed modeling, male speaker

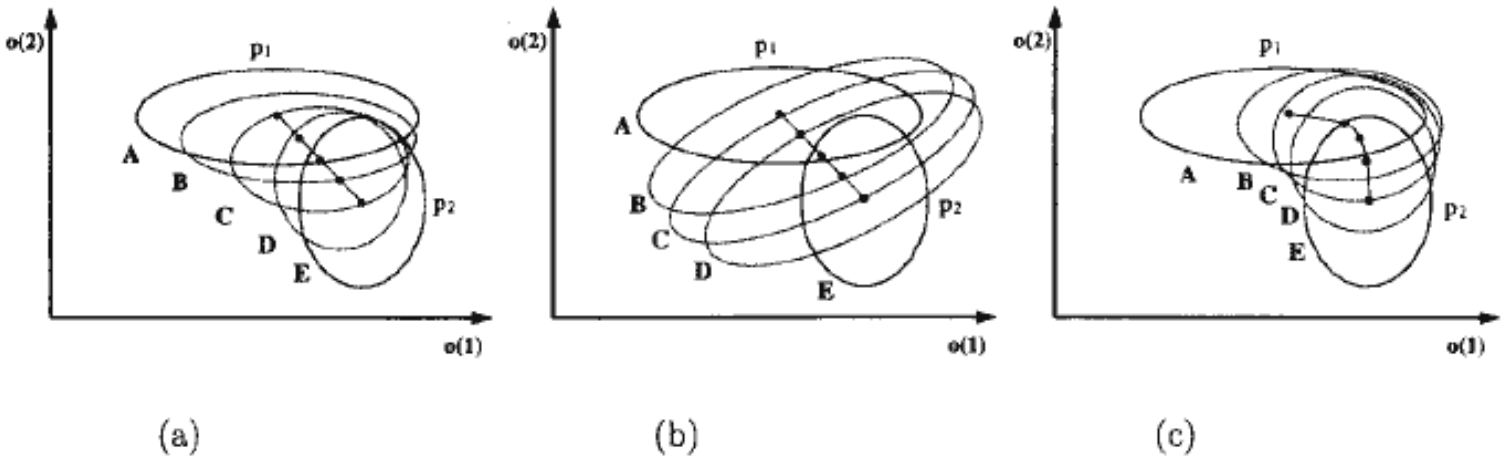
Interpolation
We have a model, now what?

- 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 μ_k and covariance matrices $\mathbf{U}_k$
- 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 $\widetilde{\mathbf{U}}$ as the interpolation result

Interpolation
Three different methods

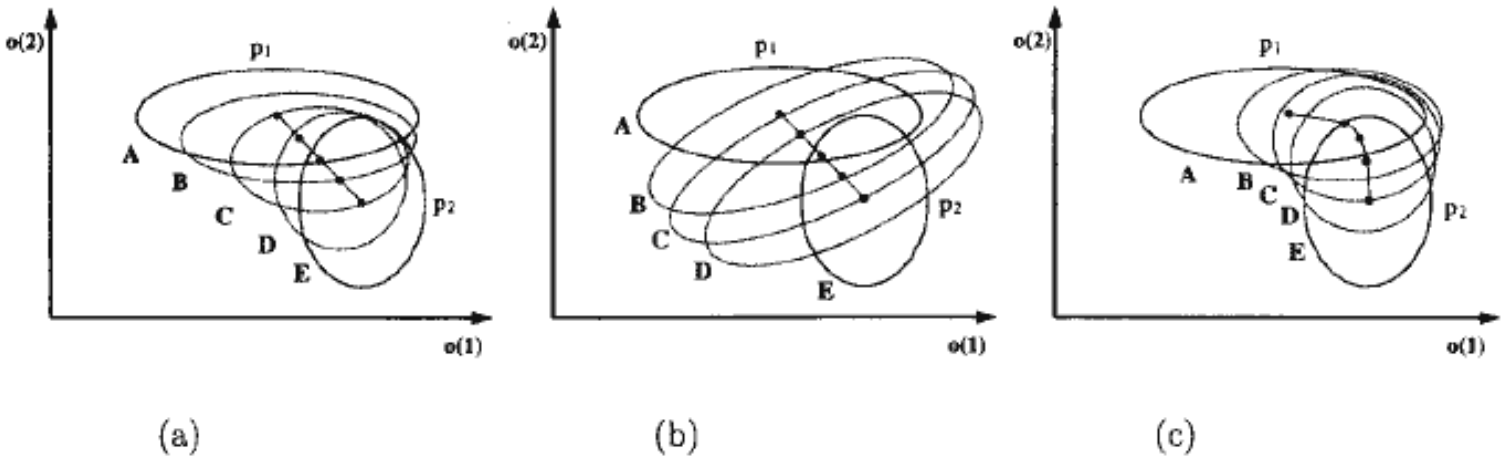


(a) Interpolation among observations

$$\tilde{\mu} = \sum_{k=1}^{N} a_k \mu_k \qquad \tilde{\mathbf{U}} = \sum_{k=1}^{N} a_k{}^2 \mathbf{U}_k$$

Interpolation

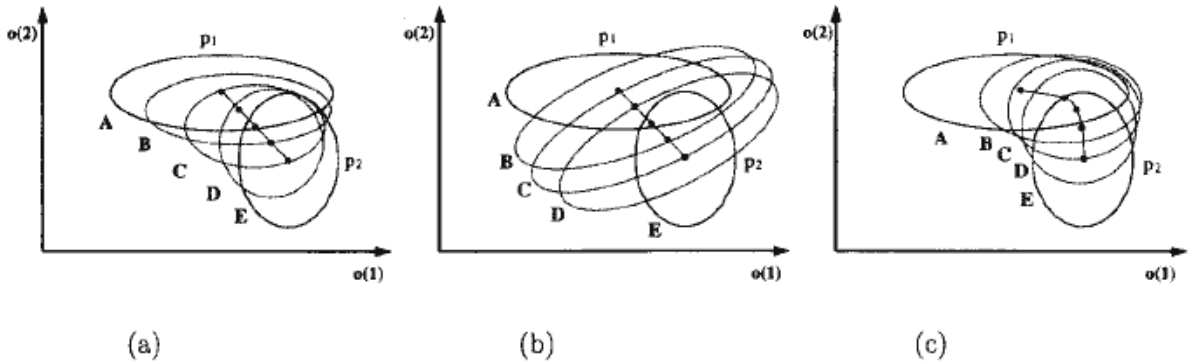
Three different methods



(b) Interpolation among output distributions

$$\tilde{\mu} = \sum_{k=1}^{N} a_k \mu_k \qquad \tilde{\mathbf{U}} = \sum_{k=1}^{N} a_k \left(\mathbf{U}_k + \mu_k \mu_k{}^T \right) - \tilde{\mu}\tilde{\mu}^T$$

Interpolation
Three different methods



(c) Interpolation based on Kullback information measure

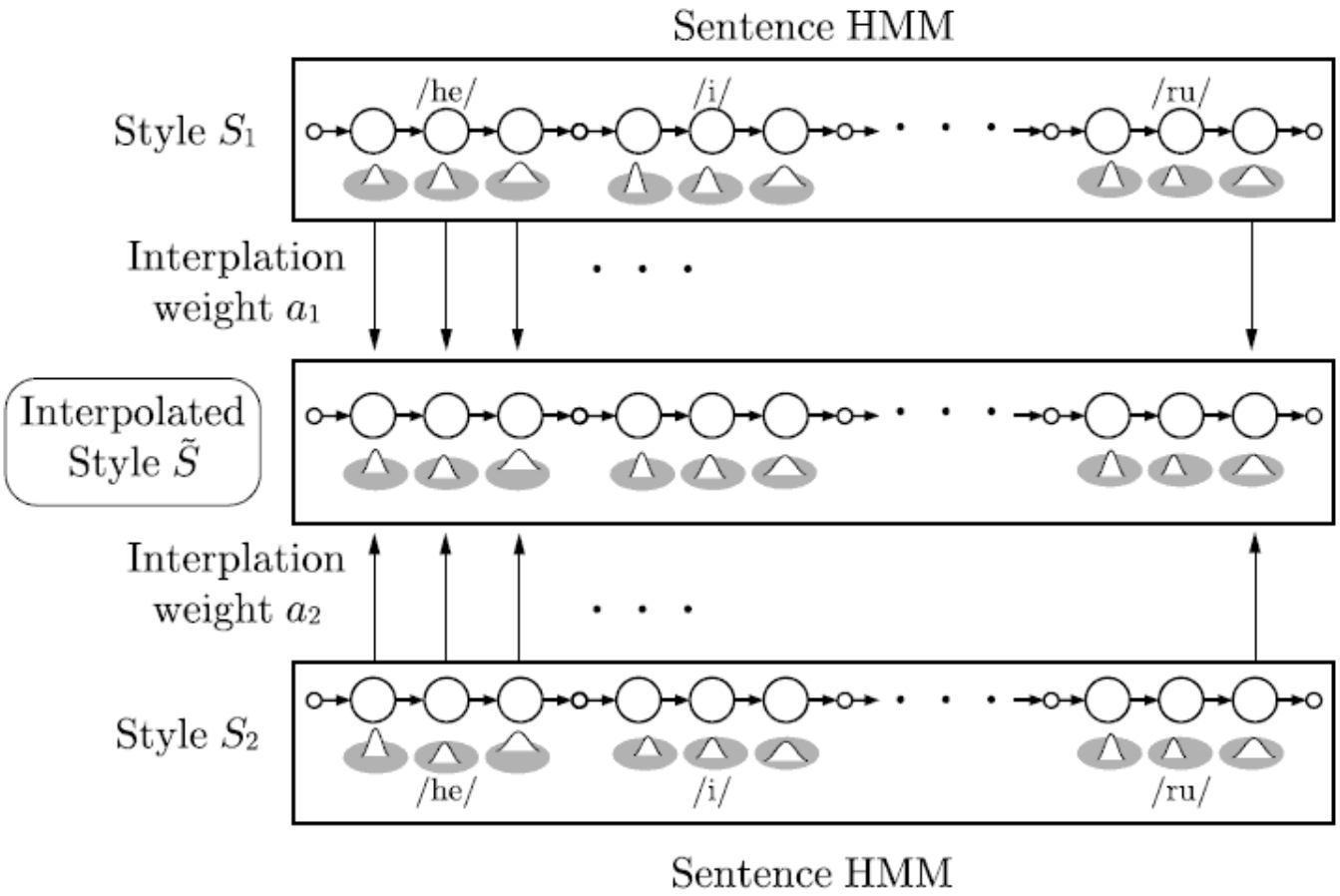
$$\tilde{\mu} = \left(\sum_{k=1}^{N} a_k \mathbf{U}_k^{-1}\right)^{-1} \left(\sum_{k=1}^{N} a_k \mathbf{U}_k^{-1} \mu_k\right)^{-1}$$

$$\tilde{\mathbf{U}} = \left(\sum_{k=1}^{N} a_k \mathbf{U}_k^{-1}\right)^{-1}$$

Interpolation
How to do it exactly?

- Change coefficients (a_1, a_2) gradually from $(1, 0)$ to $(0, 1)$
- For equally structured models λ_k: interpolate directly from λ_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

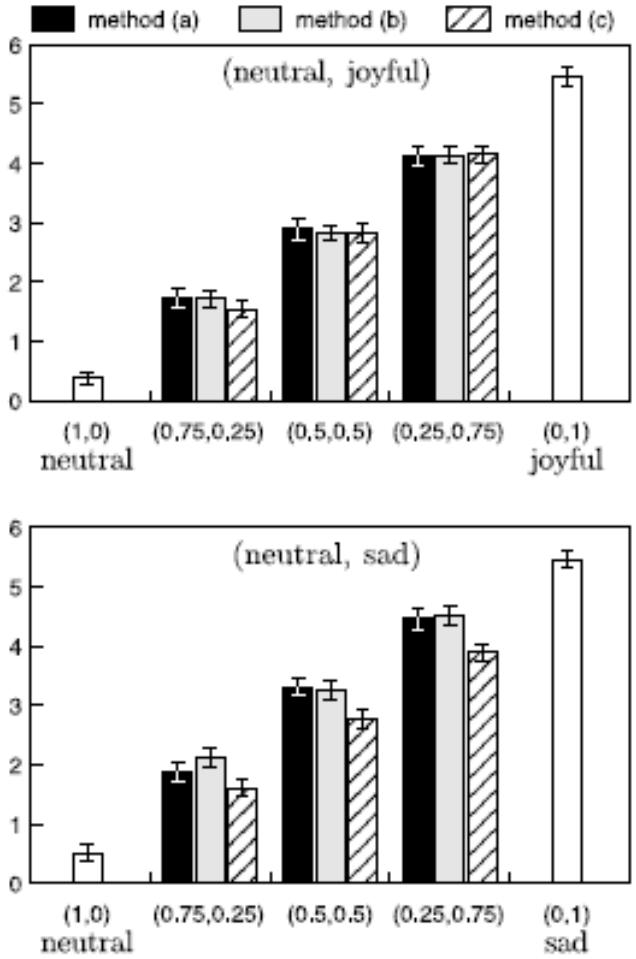
Interpolation
How to do it exactly?



Interpolation
Evaluation – Setup

- 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)

Interpolation
Evaluation – Results



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