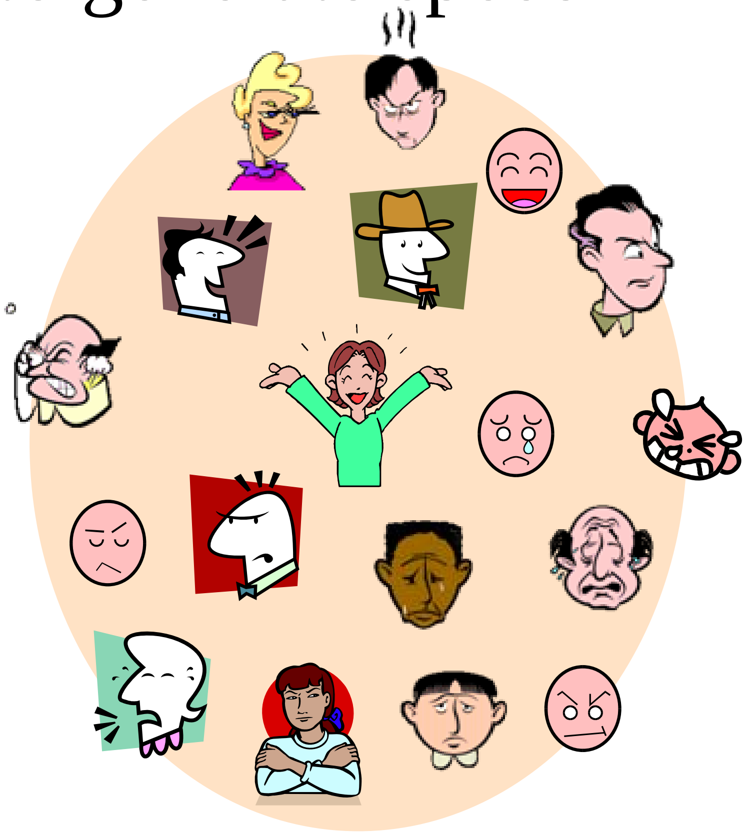
An HMM-Based Approach to Flexible Speech Synthesis

Keiichi Tokuda

Nagoya Institute of Technology

Towards Human-like Talking Machines

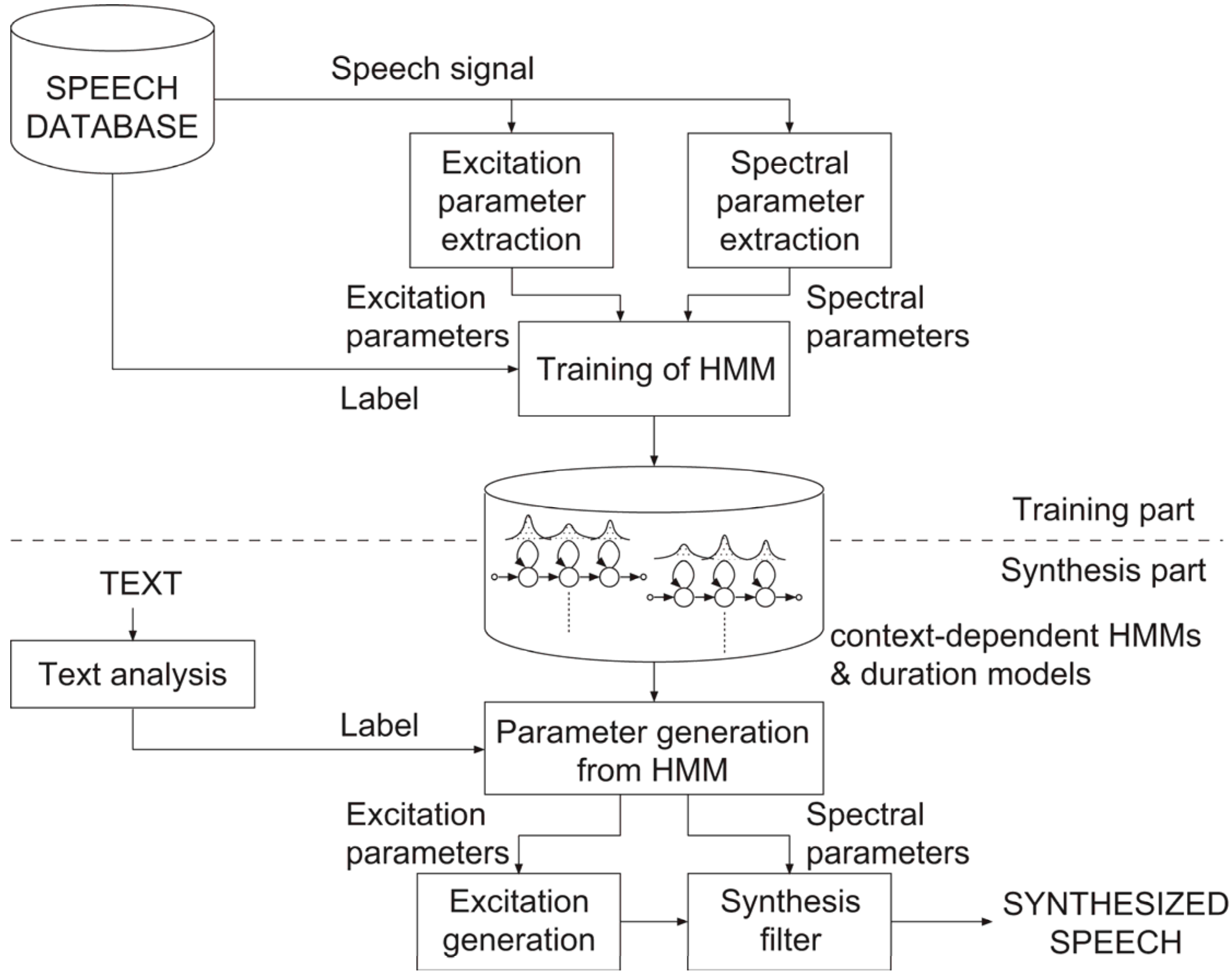
- For realizing natural human-computer interaction, speech synthesis systems are required to have an ability to generate speech with:
  - arbitrary speaker's voice
  - various speaking styles
  - emphasis
  - emotional expressions
  - and so on



Corpus-Based Speech Synthesis

- **Unit selection approach**
  - High quality speech can be synthesized using waveform concatenation algorithms.
  - To obtain various voices, a large amount of speech data is necessary.
- **HMM-based approach**
  - Generate speech parameters from statistics.
  - Voice quality can easily be changed by transforming HMM parameters.

System Overview



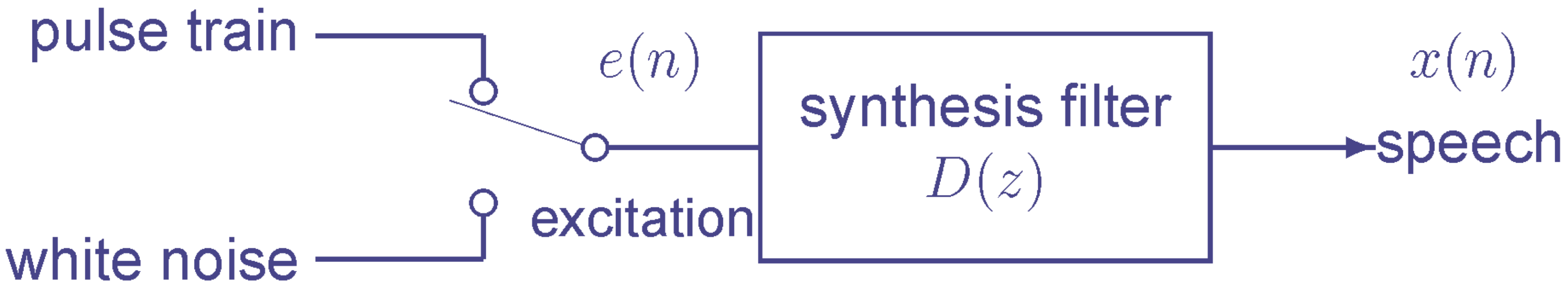
Overview of This Talk

- Basic Techniques
  - Vocoding technique
  - Speech Parameter generation algorithm
  - F0 pattern modeling
- Recent improvements and evaluation
- Relation to the unit selection approach
- Flexibility of the approach
  - Speaker adaptation (mimicking voices)
  - Speaker Interpolation (mixing voices)
  - Eigenvoices (producing voices), etc.

Overview of This Talk

- Basic Techniques
  - Vocoding technique
  - Speech Parameter generation algorithm
  - F0 pattern modeling
- Recent improvements and evaluation
- Relation to the unit selection approach
- Flexibility of the approach
  - Speaker adaptation (mimicking voices)
  - Speaker Interpolation (mixing voices)
  - Eigenvoices (producing voices), etc.

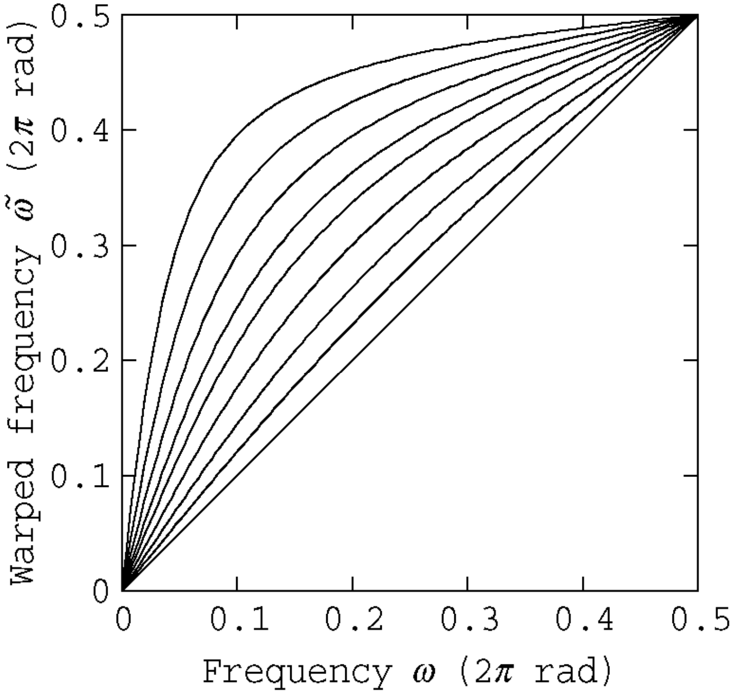
Source-Filter Model



D(*z*) should be defined by the state output vector of HMM, e.g., mel-cepstrum, lsp's

Synthesis Filter Model

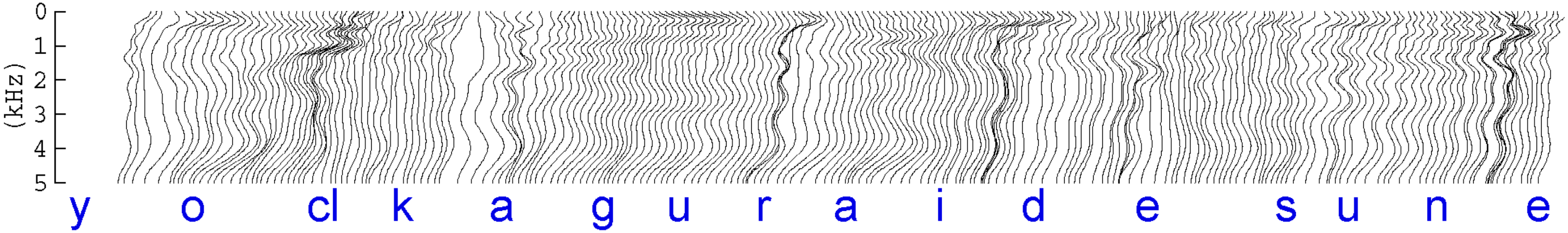
$$D(z) = \exp \sum_{m=0}^{M} c(m)\, \tilde{z}^{-m}, \qquad \tilde{z}^{-1} = \left. \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \right|_{z=e^{-j\omega}} = e^{-j\tilde{\omega}}$$



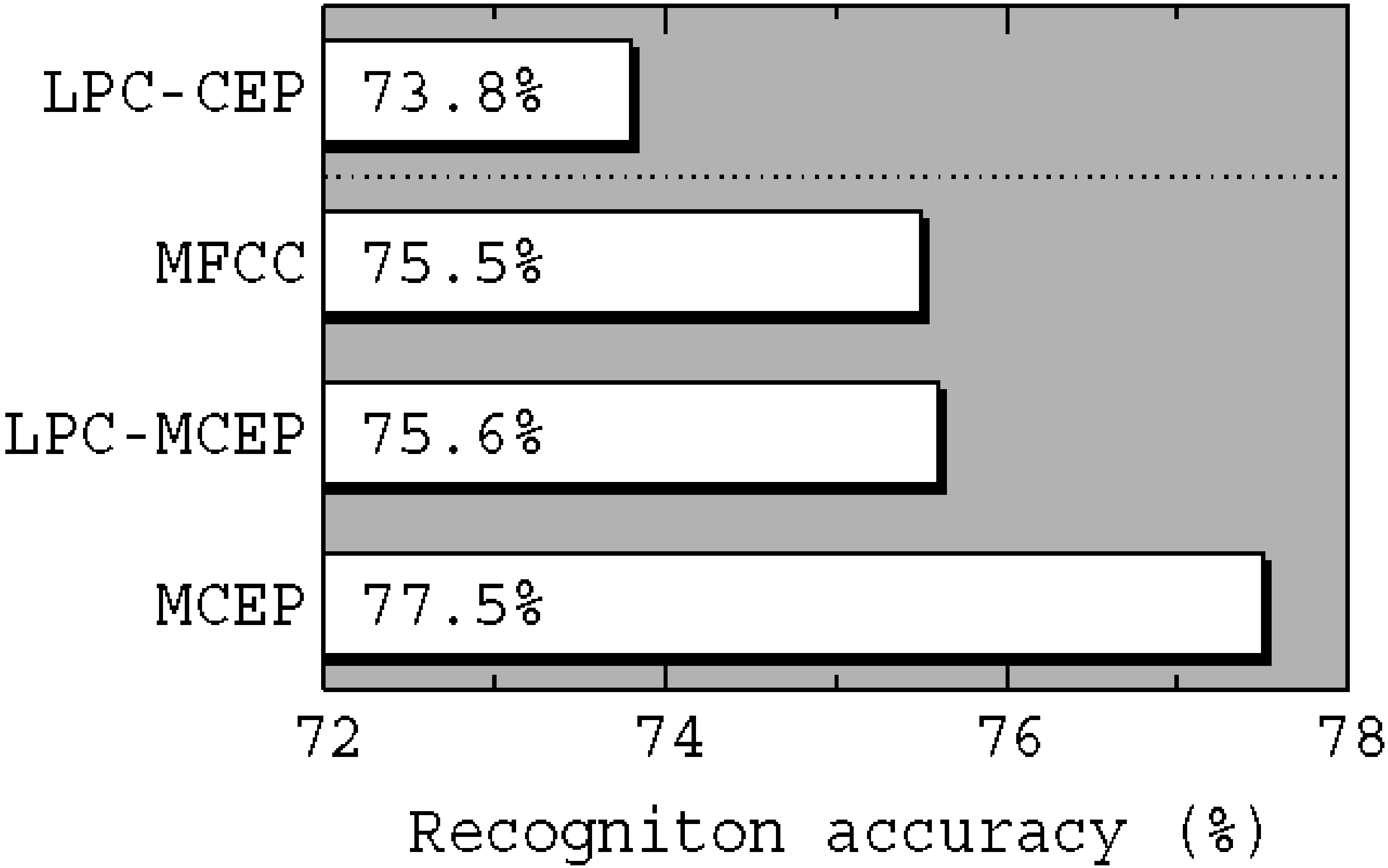
Objective Function

$$\boldsymbol{c} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{x} \mid \boldsymbol{c})$$

$$\begin{aligned} \boldsymbol{x} &= [x(0),\, x(1),\, \ldots,\, x(N-1)]' \\ \boldsymbol{c} &= [c(0),\, c(1),\, \ldots,\, c(M)]' \end{aligned}$$

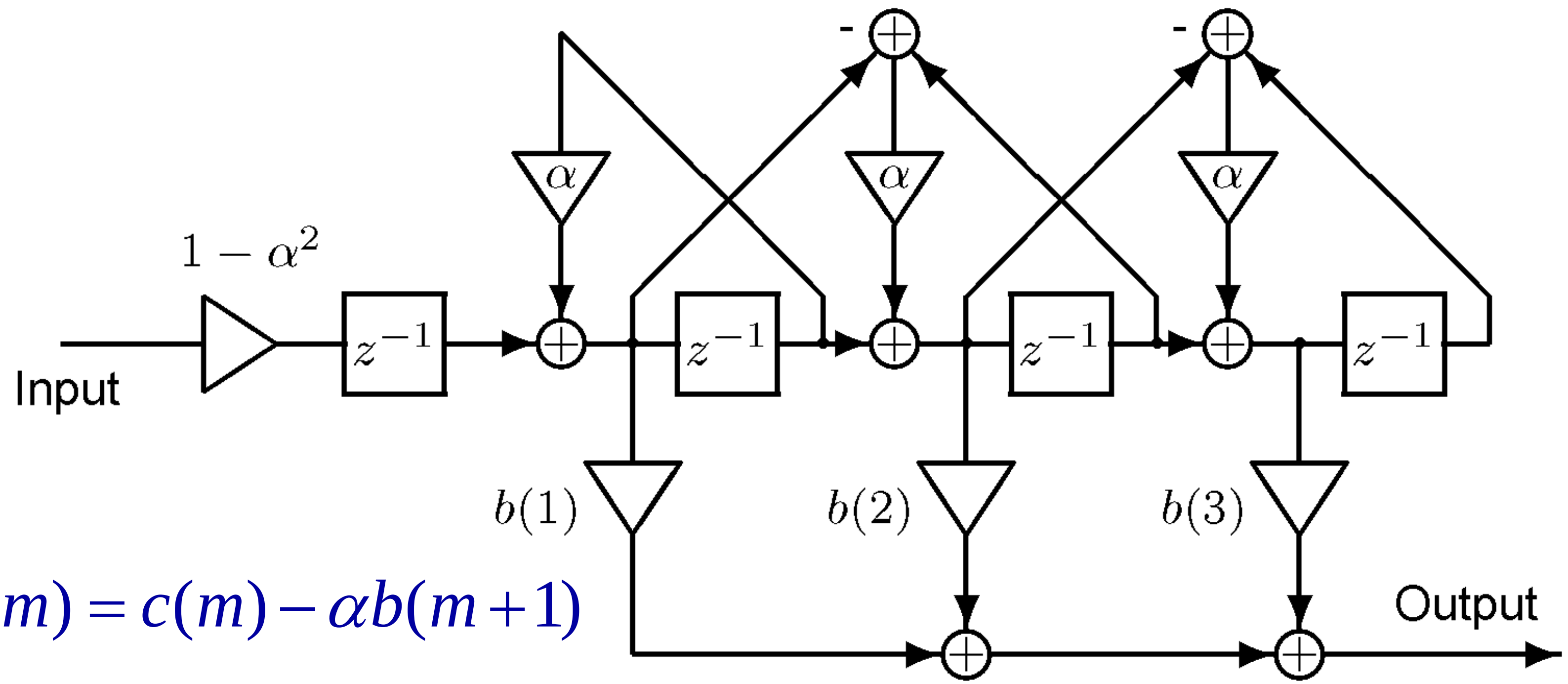


Evaluation in Speech Recognition



Synthesis Filter

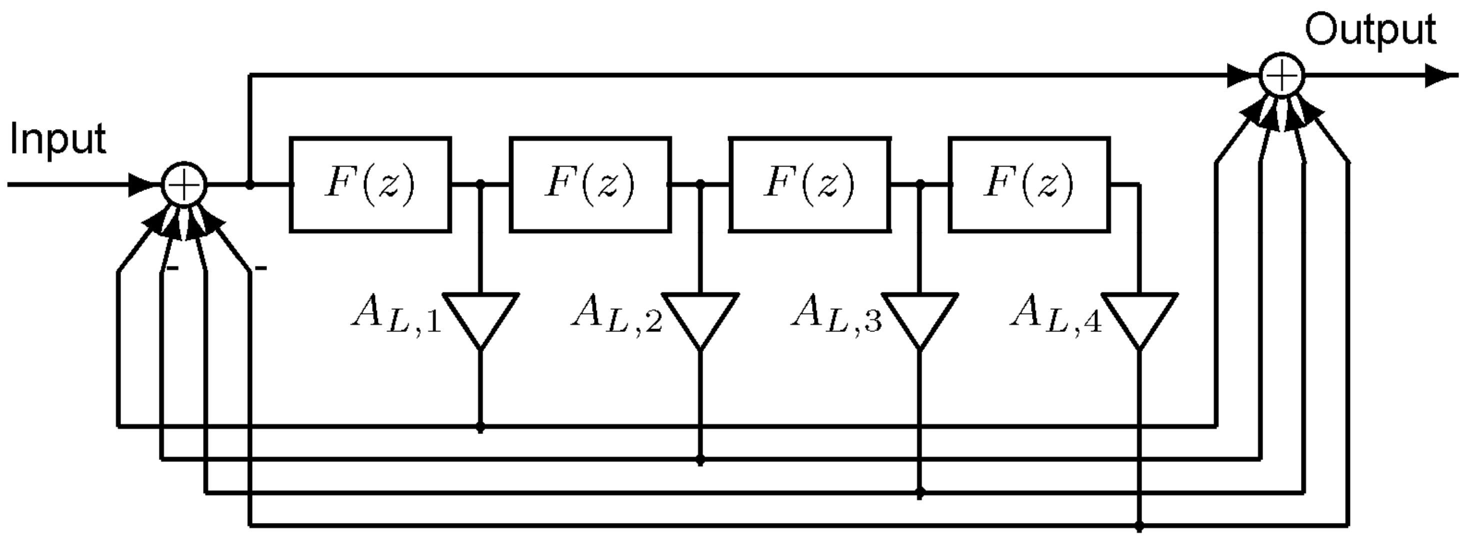
$$D(z) = \exp F(z), \qquad F(z) = \sum_{m=0}^{M} c(m)\, \tilde{z}^{-m}$$



$$b(m) = c(m) - \alpha b(m+1)$$

MLSA Filter

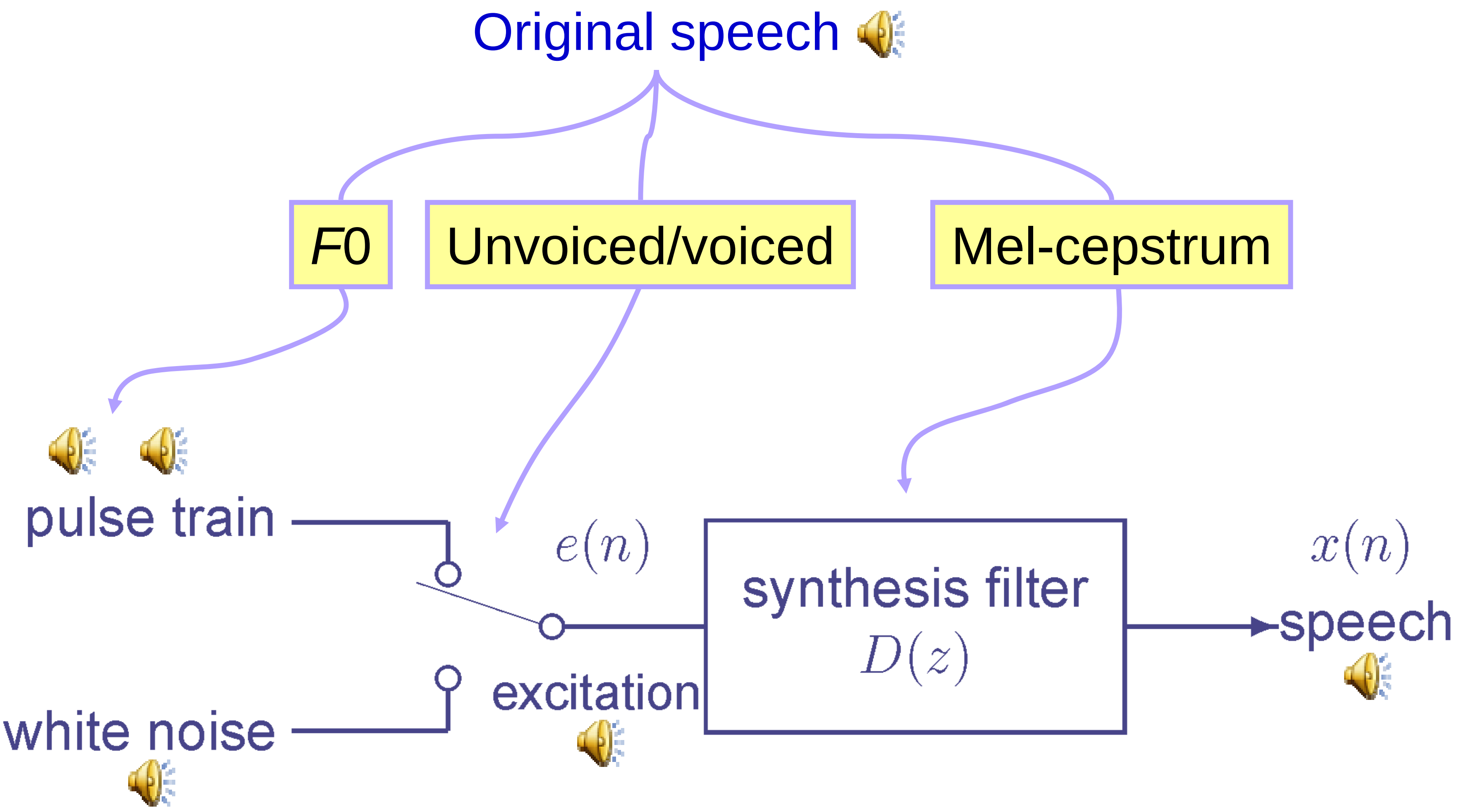
$$D(z) = \exp F(z) \simeq \frac{1 + \sum_{l=1}^{L} A_{L,l} \{F(z)\}^l}{1 + \sum_{l=1}^{L} A_{L,l} \{-F(z)\}^l}$$



Features of MLSA Filter

- Filter coefficients given by mel-cepstrum
- Sufficient approximation accuracy
  - ⇨ maximum spectral error 0.24dB
- Guaranteed stability
- Computationally efficient
  - ⇨ $O(M)$ multiply-add operations a sample

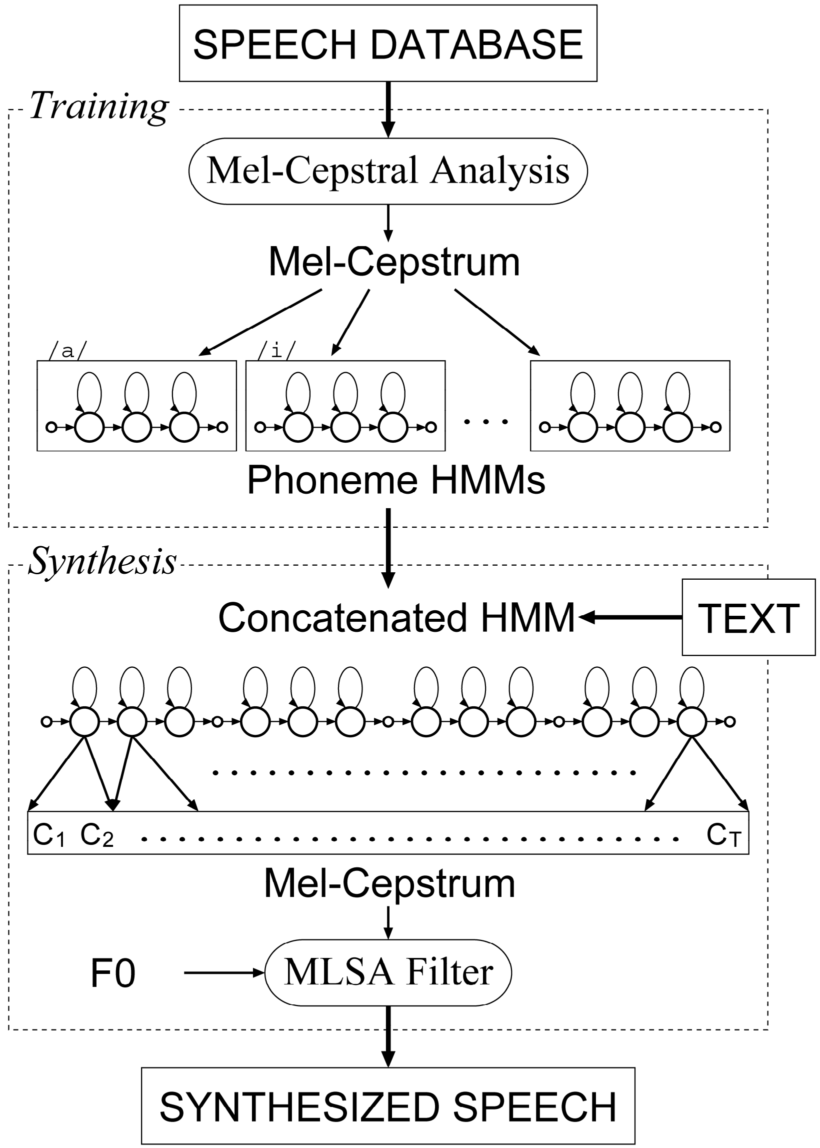
Vocoded Speech Samples



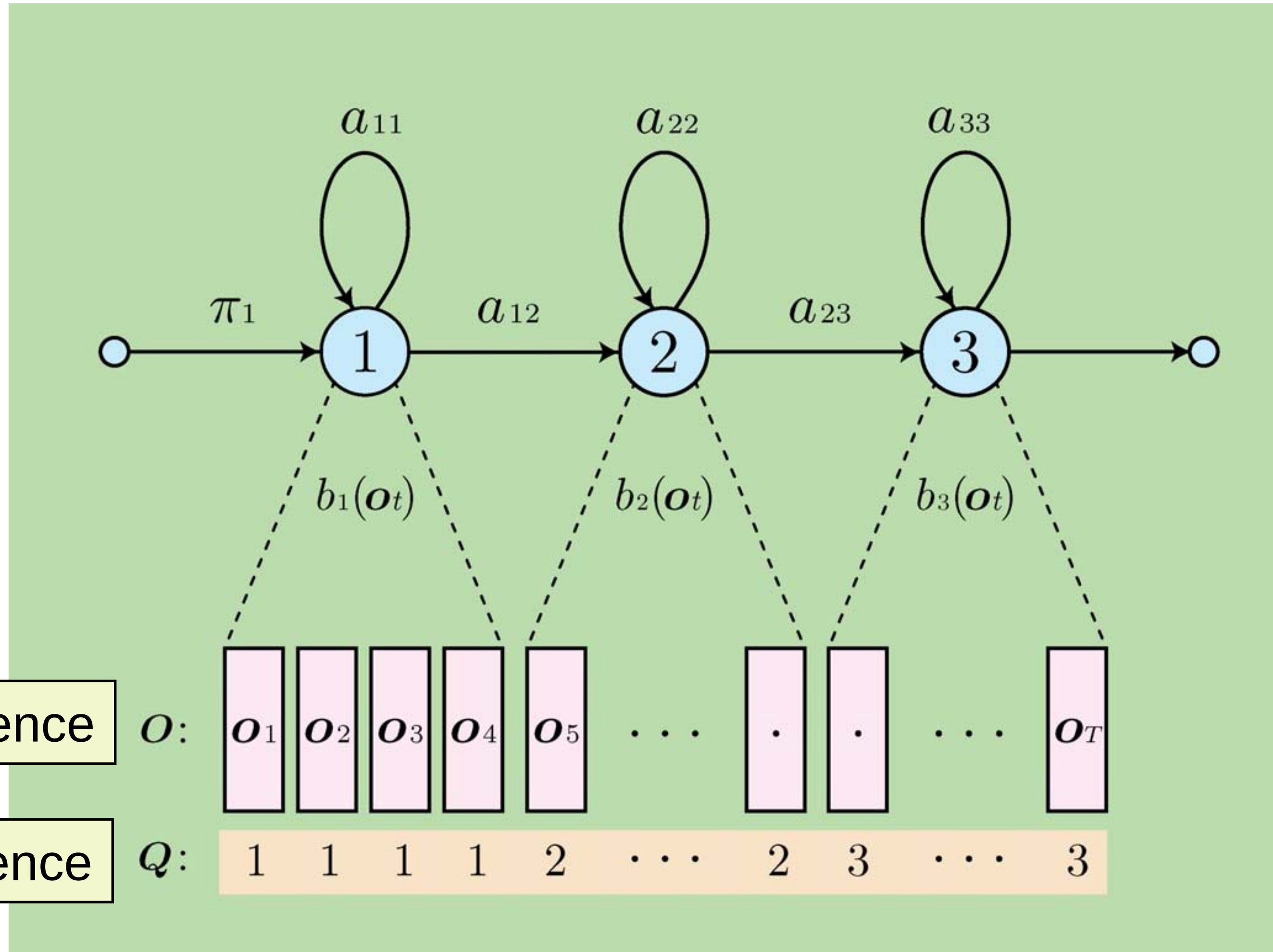
Overview of This Talk

- Basic Techniques
  - Vocoding technique
  - Speech Parameter generation algorithm
  - F0 pattern modeling
- Recent improvements and evaluation
- Relation to the unit selection approach
- Flexibility of the approach
  - Speaker adaptation (mimicking voices)
  - Speaker Interpolation (mixing voices)
  - Eigenvoices (producing voices), etc.

System Overview (only spectrum part)



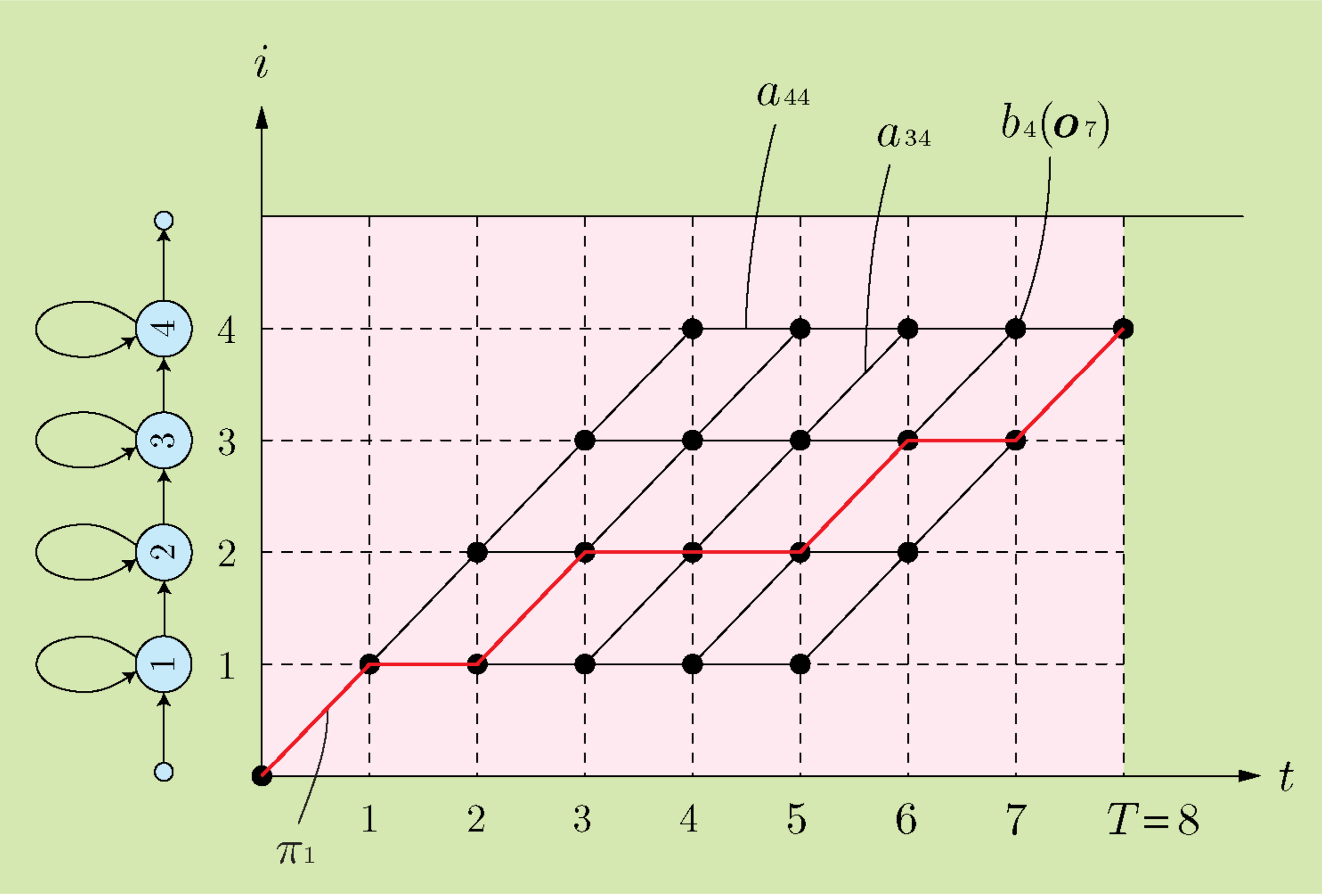
Hidden Markov Model: HMM



State output sequence

State sequence

Output Probability of HMM



$$P(\boldsymbol{O} \mid \lambda) = \sum_{\boldsymbol{Q}} P(\boldsymbol{O}, \boldsymbol{Q} \mid \lambda) = \sum_{\boldsymbol{Q}} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{o}_t)$$

Speech Parameter Generation

For given HMM λ, determine a speech parameter vector sequence $\boldsymbol{O} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top\right]^\top$ which maximizes

$$
\begin{aligned}
P(\boldsymbol{O} \mid \lambda) &= \sum_{\boldsymbol{Q}} P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda) P(\boldsymbol{Q} \mid \lambda) \\
&\simeq \max_{\boldsymbol{Q}} P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda) P(\boldsymbol{Q} \mid \lambda)
\end{aligned}
$$

$$\Downarrow$$

$$
\begin{aligned}
\boxed{\boldsymbol{Q}_{\max} = \arg\max_{\boldsymbol{Q}} P(\boldsymbol{Q} \mid \lambda)} \\
\boldsymbol{O}_{\max} = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O} \mid \boldsymbol{Q}_{\max}, \lambda)
\end{aligned}
$$

Determination of State Durations

$$P(\boldsymbol{Q} \mid \lambda) = \prod_{i=1}^{K} p_i(d_i)$$

Standard HMM ⇨ $p_i(d_i)$: geometric distribution

⇩

Gaussian with mean m_i and variance σ_i^2

$$d_i = m_i, \quad i = 1, 2, \ldots, K$$

Speech Parameter Generation

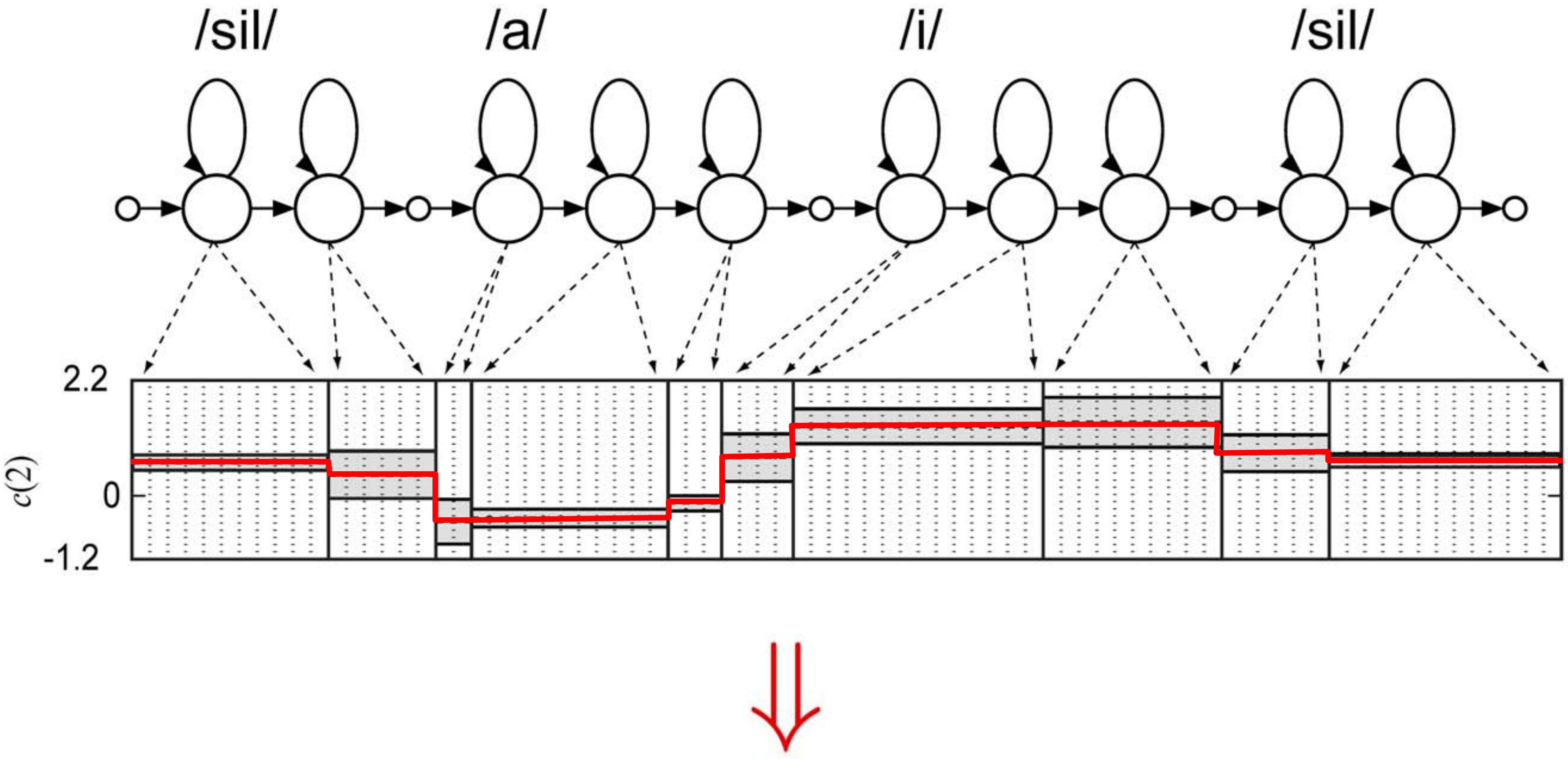
For given HMM λ, determine a speech parameter vector sequence $\boldsymbol{O} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top\right]^\top$ which maximizes

$$
\begin{aligned}
P(\boldsymbol{O} \mid \lambda) &= \sum_{\boldsymbol{Q}} P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda) P(\boldsymbol{Q} \mid \lambda) \\
&\simeq \max_{\boldsymbol{Q}} P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda) P(\boldsymbol{Q} \mid \lambda)
\end{aligned}
$$

$$\Downarrow$$

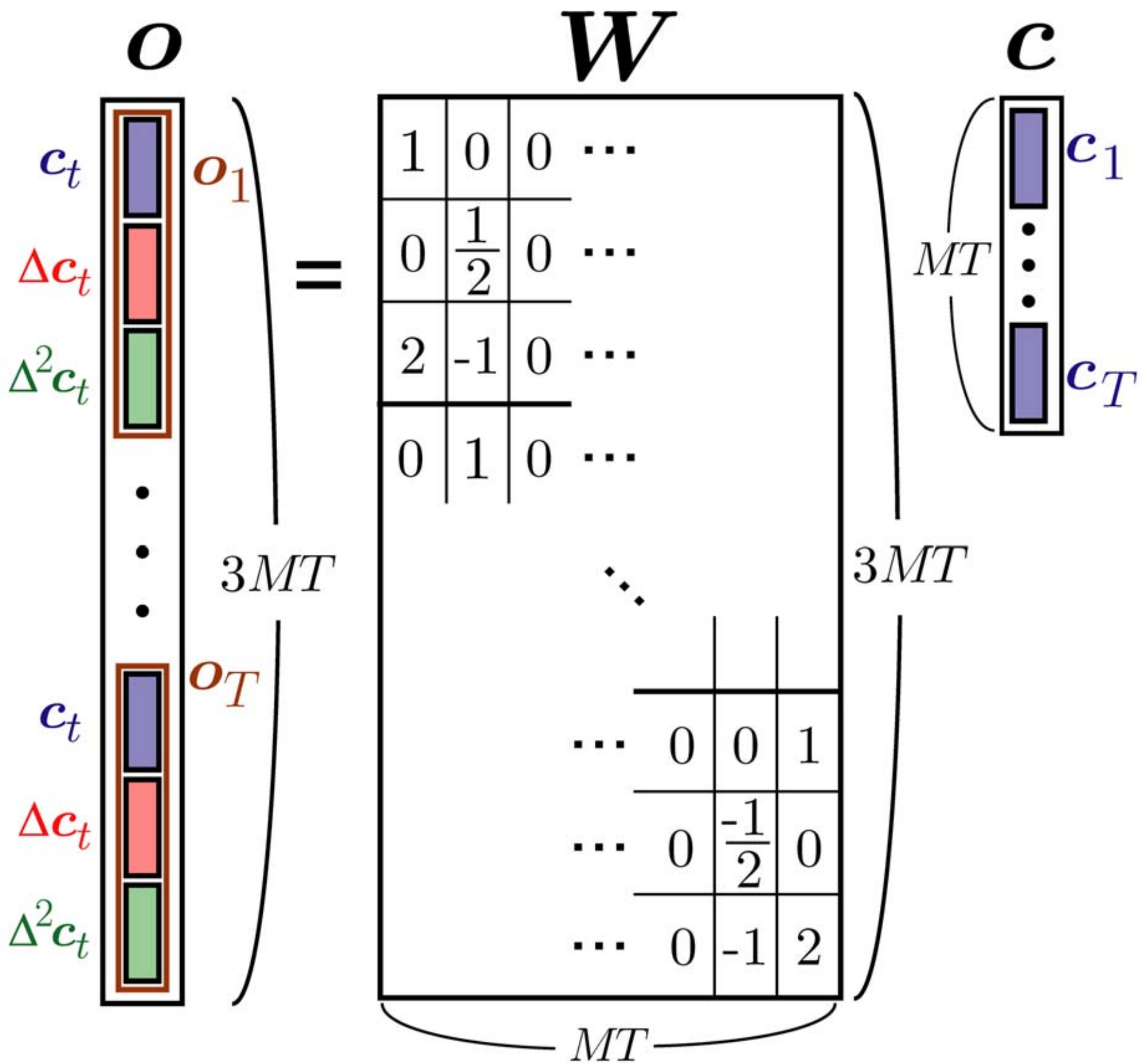
$$
\begin{aligned}
\boldsymbol{Q}_{\max} &= \arg\max_{\boldsymbol{Q}} P(\boldsymbol{Q} \mid \lambda) \\
\boldsymbol{O}_{\max} &= \arg\max_{\boldsymbol{O}} P(\boldsymbol{O} \mid \boldsymbol{Q}_{\max}, \lambda)
\end{aligned}
$$

Without Dynamic Feature



O becomes a sequence of mean vectors.

Integration of Dynamic Feature



Solution for The Problem

By setting

$$\frac{\partial \log P(\boldsymbol{W}\boldsymbol{C} \,|\, \boldsymbol{Q}_{max}, \lambda)}{\partial \boldsymbol{C}} = \boldsymbol{0}$$

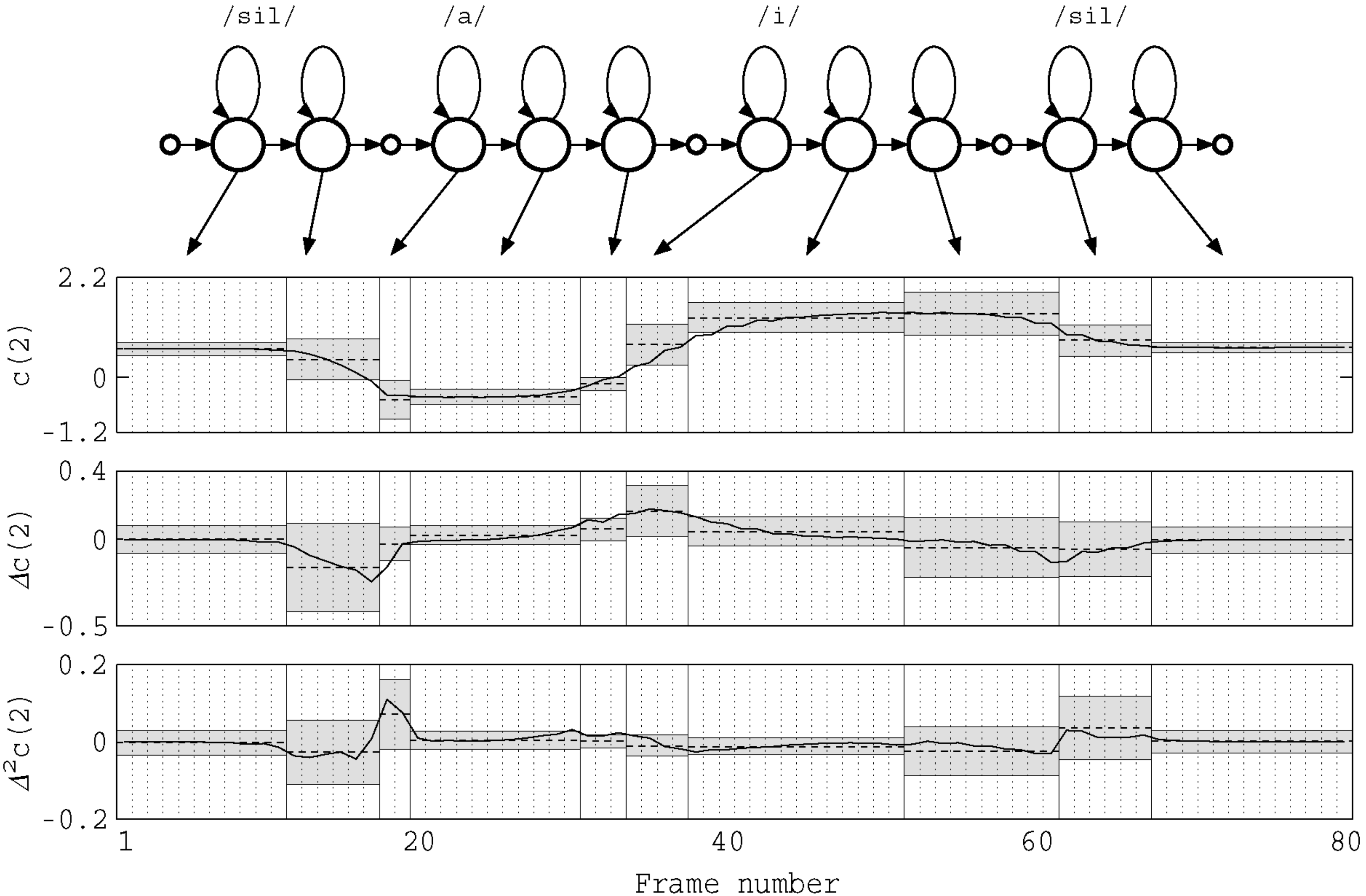
we obtaine

$$\boldsymbol{W}^{\top}\boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{C} = \boldsymbol{W}^{\top}\boldsymbol{U}^{-1}\boldsymbol{M}$$

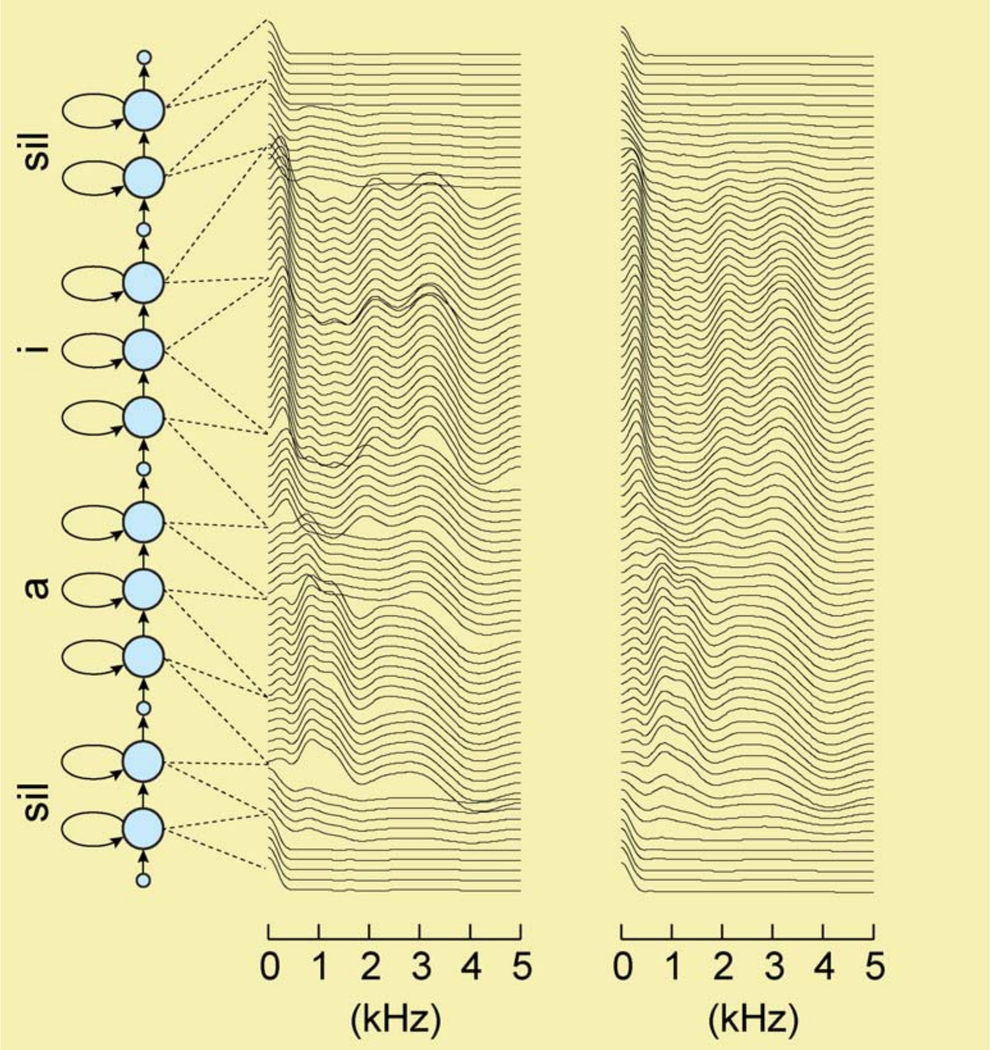
where

$$\begin{aligned} \boldsymbol{C} &= \left[\boldsymbol{c}_1^{\top}, \boldsymbol{c}_2^{\top}, \ldots, \boldsymbol{c}_T^{\top}\right]^{\top} \\ \boldsymbol{M} &= \left[\boldsymbol{\mu}_{q_1}^{\top}, \boldsymbol{\mu}_{q_2}^{\top}, \ldots, \boldsymbol{\mu}_{q_T}^{\top}\right]^{\top} \\ \boldsymbol{U}^{-1} &= \text{diag}\left[\boldsymbol{U}_{q_1}^{-1}, \boldsymbol{U}_{q_2}^{-1}, \ldots, \boldsymbol{U}_{q_T}^{-1}\right] \end{aligned}$$

Generated Speech Parameter

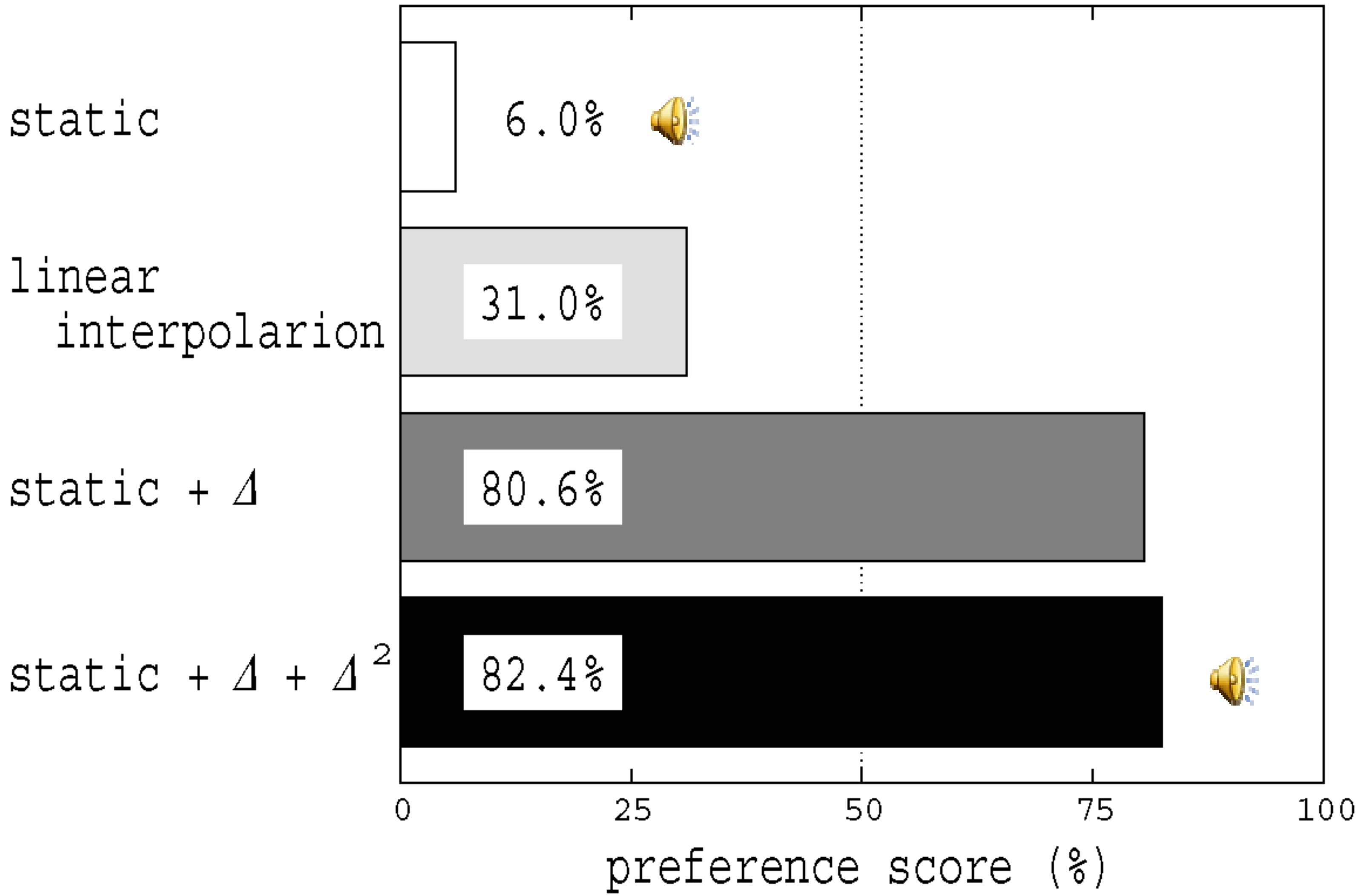


Generated Spectra



(a) (b)

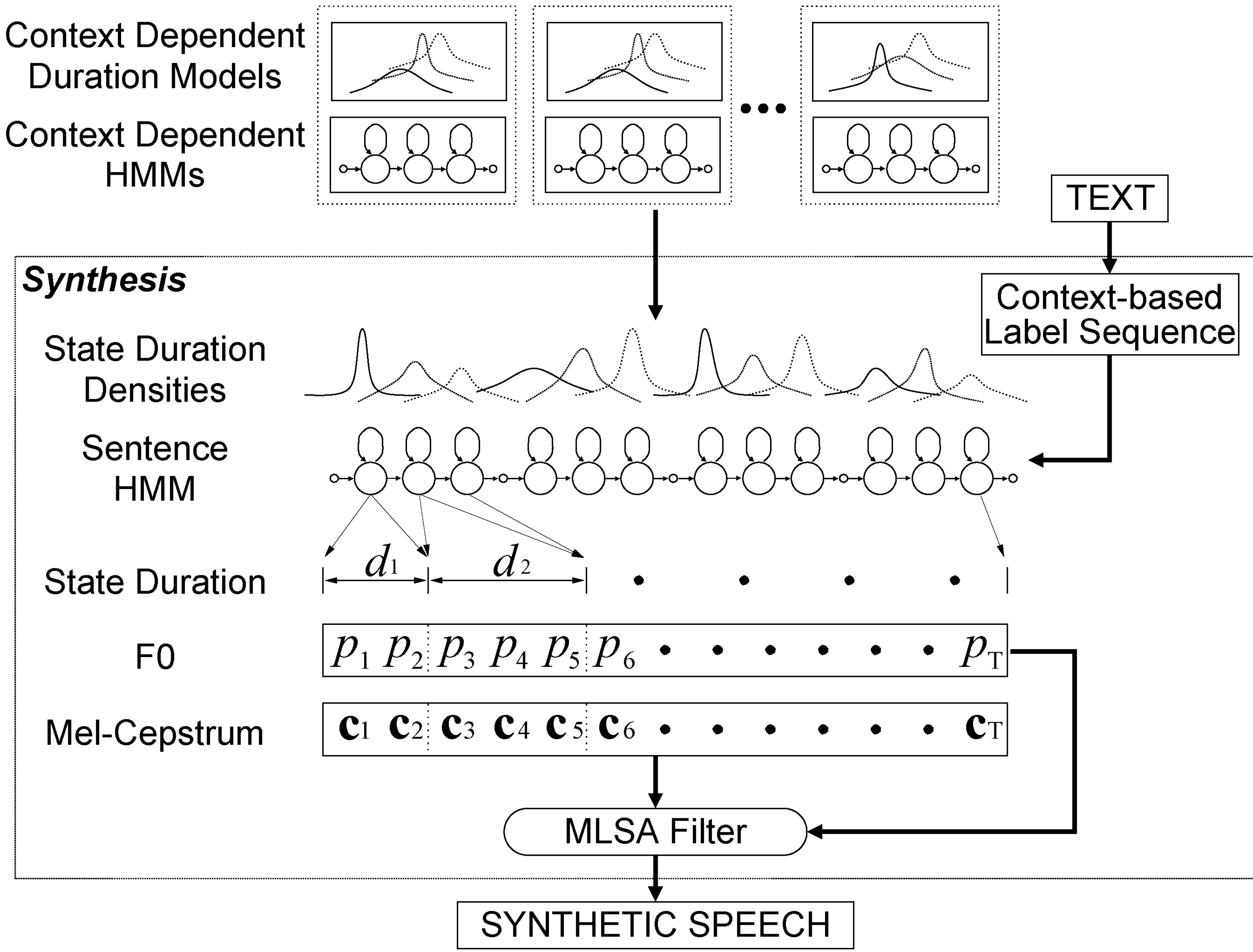
Effect of Dynamic Features



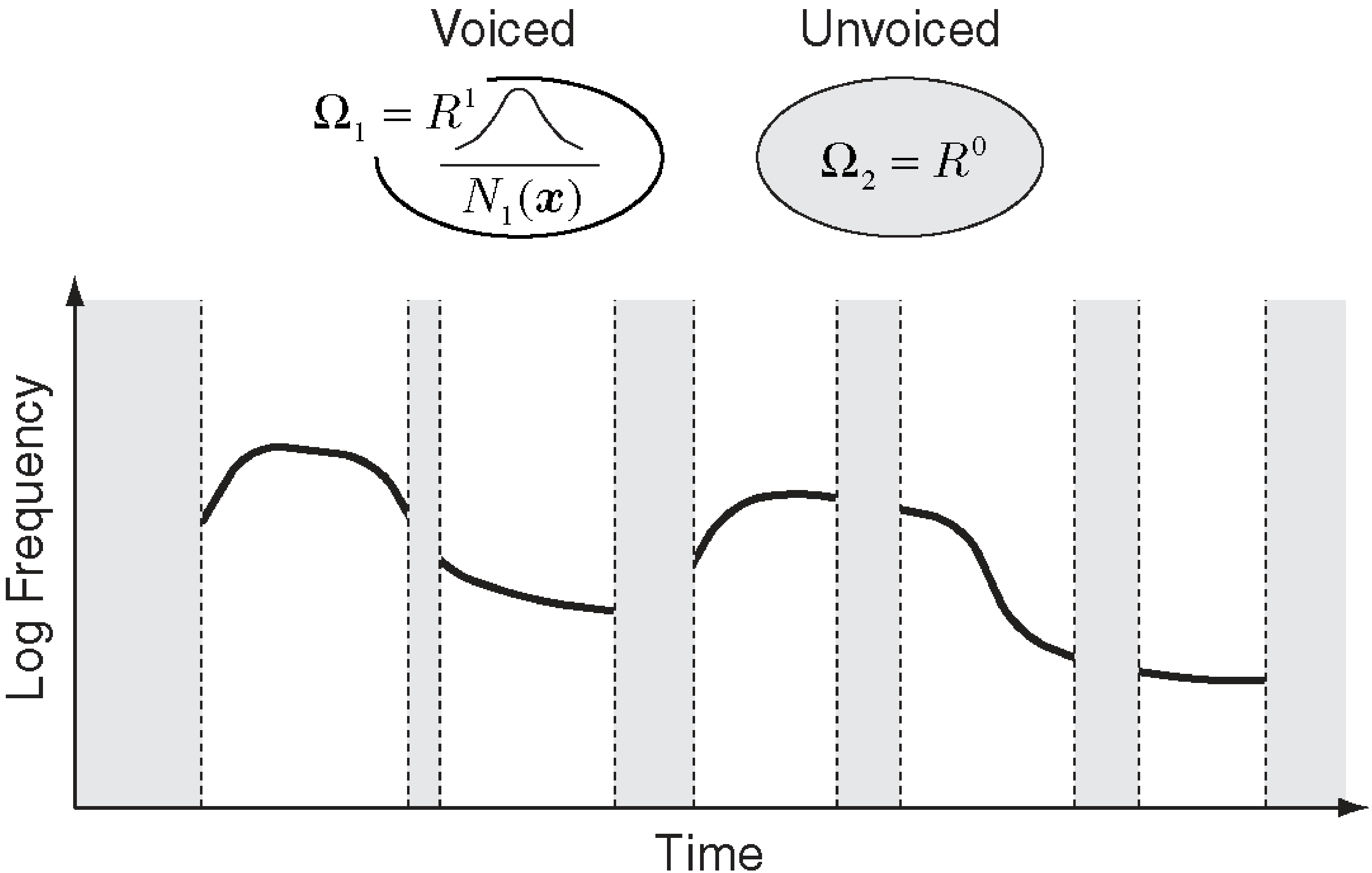
Overview of This Talk

- Basic Techniques
  - Vocoding technique
  - Speech Parameter generation algorithm
  - F0 pattern modeling
- Recent improvements and evaluation
- Relation to the unit selection approach
- Flexibility of the approach
  - Speaker adaptation (mimicking voices)
  - Speaker Interpolation (mixing voices)
  - Eigenvoices (producing voices), etc.

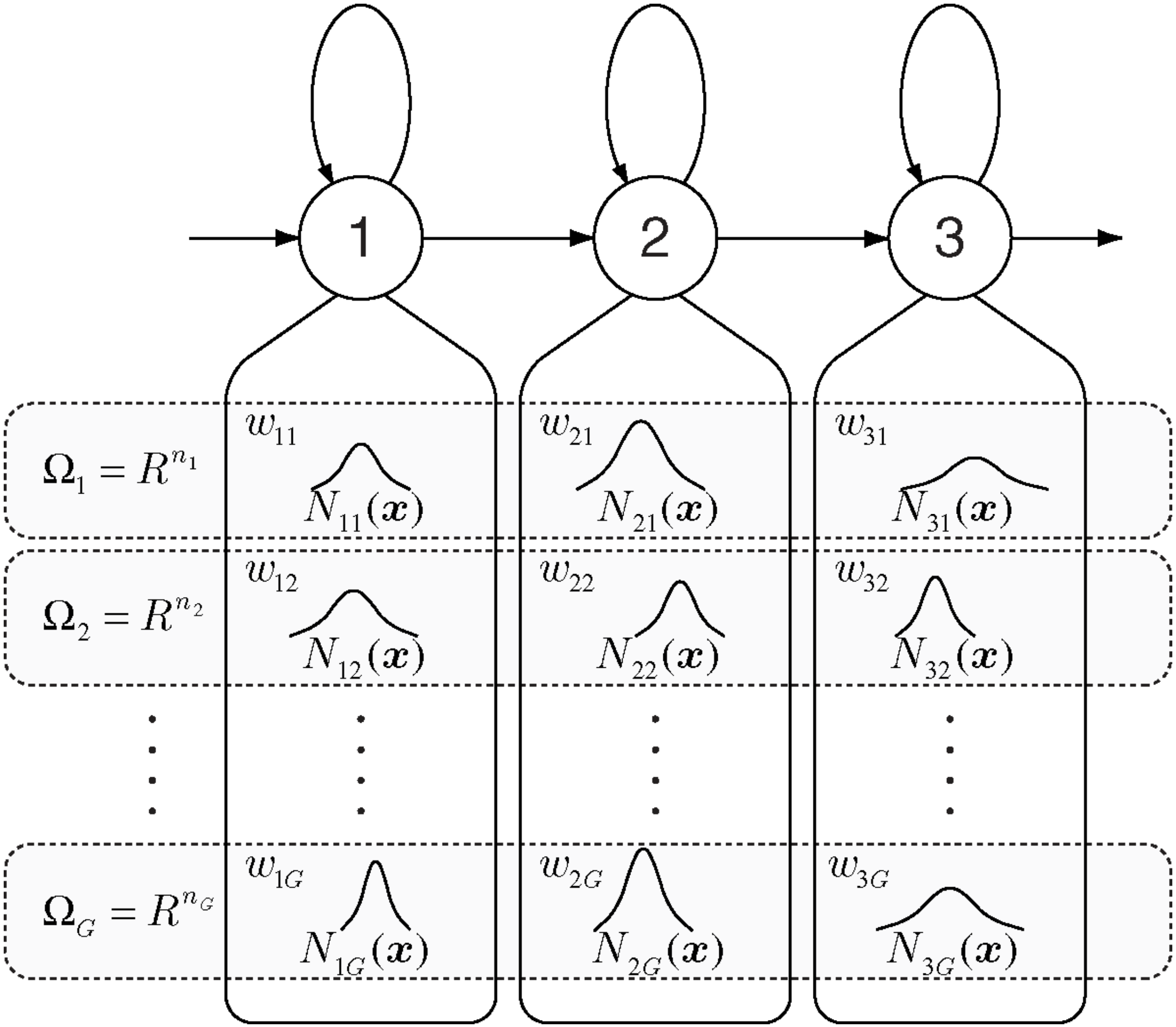
HMM-Based Speech Synthesis System



Observation of *F0*

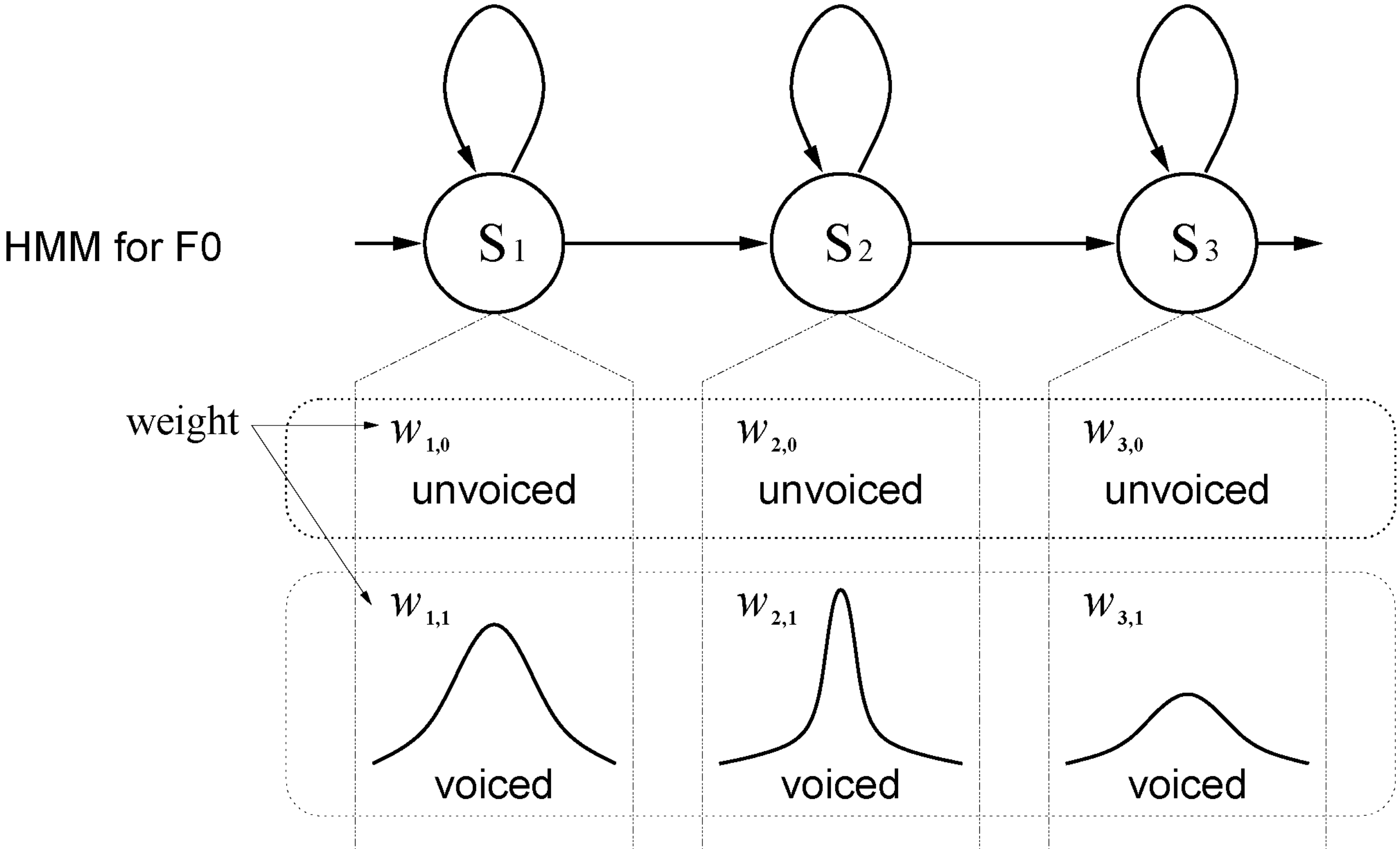


MSD-HMM

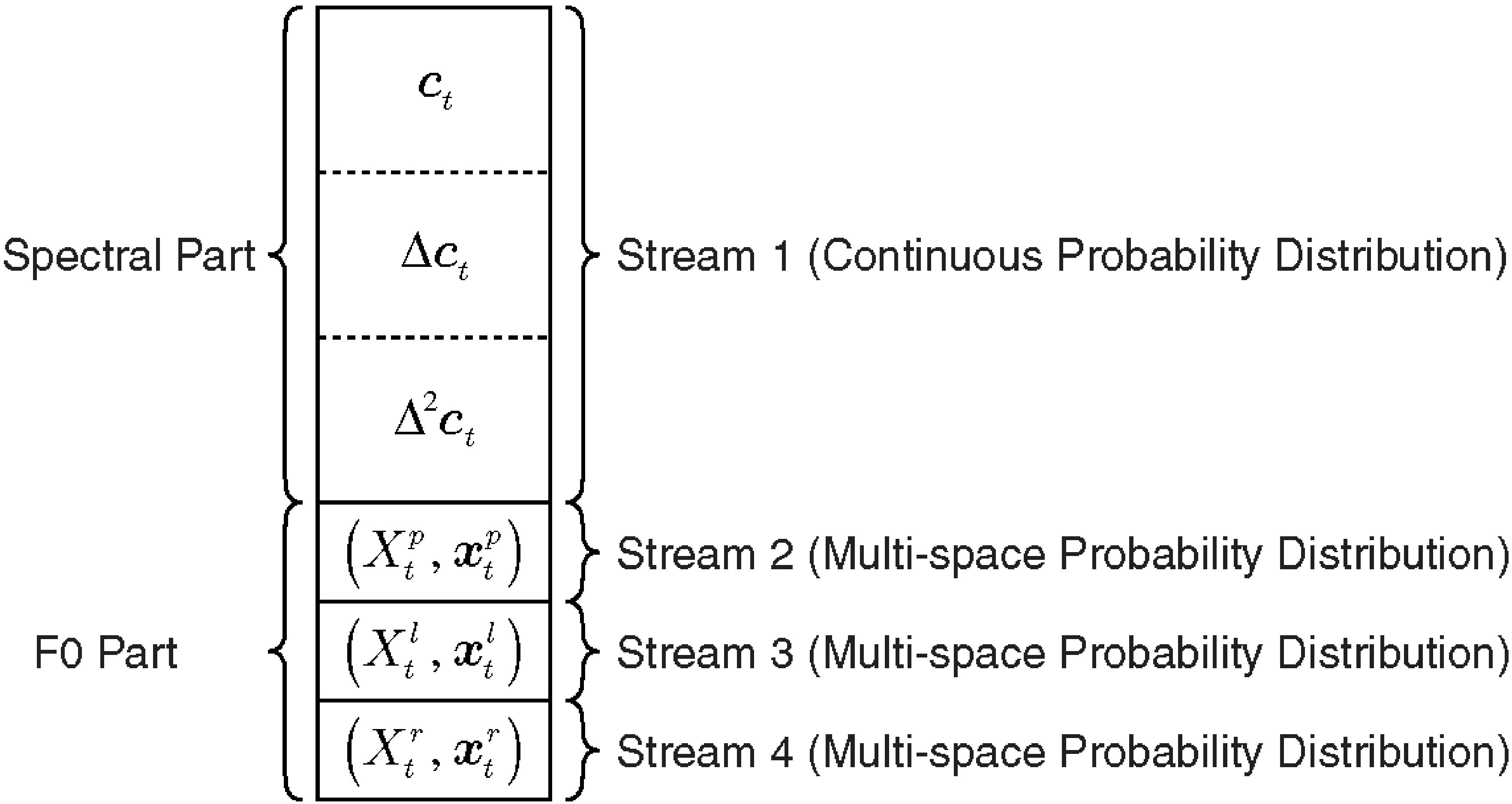


1
2
3
w_{11}
w_{21}
w_{31}
$\Omega_1 = R^{n_1}$
$N_{11}(\boldsymbol{x})$
$N_{21}(\boldsymbol{x})$
$N_{31}(\boldsymbol{x})$
w_{12}
w_{22}
w_{32}
$\Omega_2 = R^{n_2}$
$N_{12}(\boldsymbol{x})$
$N_{22}(\boldsymbol{x})$
$N_{32}(\boldsymbol{x})$
w_{1G}
w_{2G}
w_{3G}
$\Omega_G = R^{n_G}$
$N_{1G}(\boldsymbol{x})$
$N_{2G}(\boldsymbol{x})$
$N_{3G}(\boldsymbol{x})$

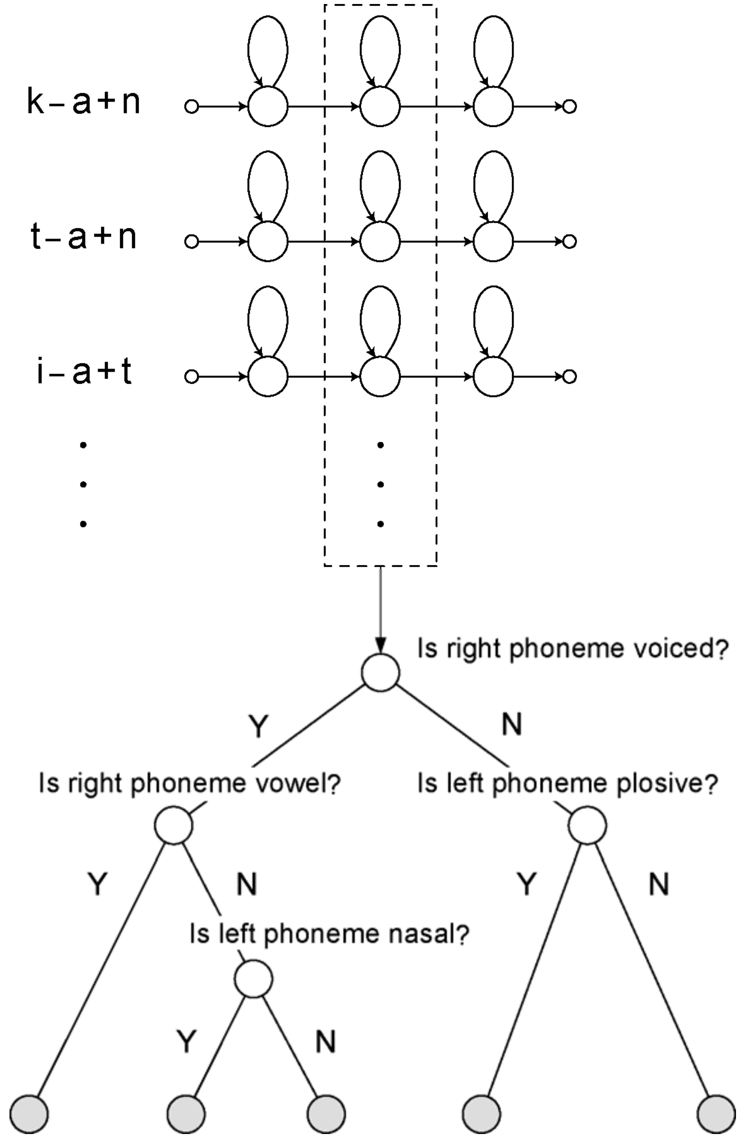
MSD-HMM for *F0* Modeling



State Output Vector



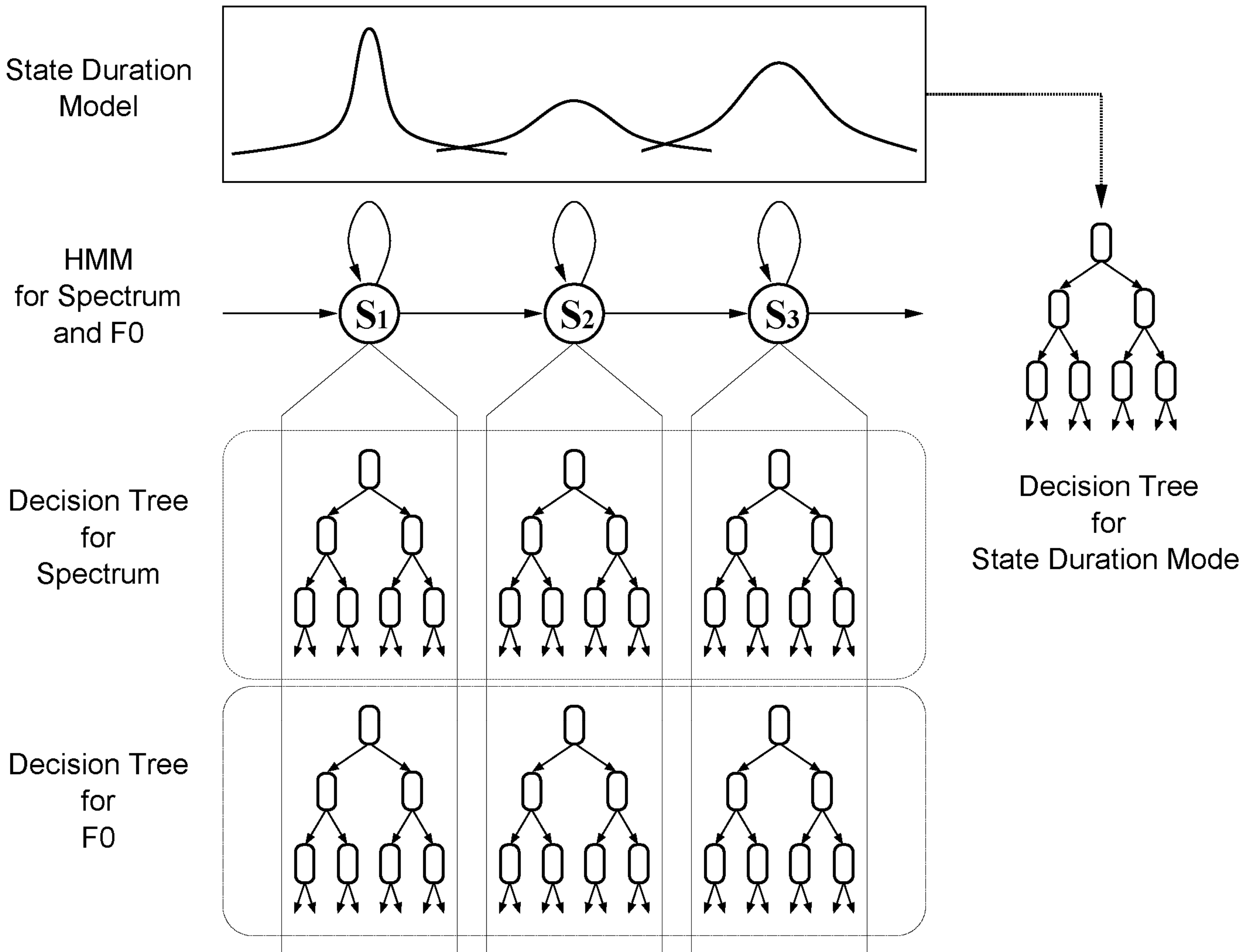
Context Clustering



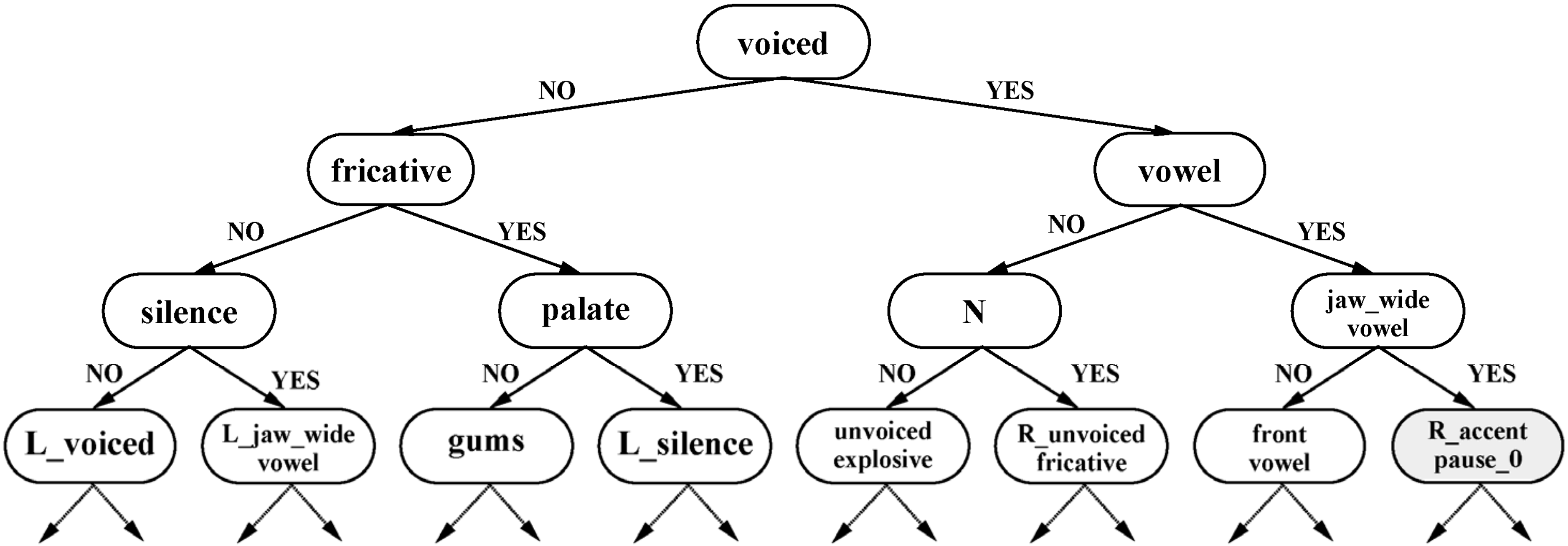
Context Clustering: Factors

- {preceding, current, succeeding} phoneme
- Position of current phoneme in current syllable
- Number of phonemes at {preceding, current, succeeding} syllable
- Accent of {preceding, current, succeeding} syllable
- Position of current syllable in current word
- Number of {preceding, succeeding} stressed syllables in current phrase
- Number of {preceding, succeeding} accented syllables in current phrase
- Number of syllables {from previous, to next} stressed syllable
- Number of syllables {from previous, to next} accented syllable
- Vowel within current syllable
- Guess at part of speech of {preceding, current, succeeding} word
- Number of syllables in {preceding, current, succeeding} word
- Position of current word in current phrase
- Number of {preceding, succeeding} content words in current phrase
- Number of words {from previous, to next} content word
- Number of syllables in {preceding, current, succeeding} phrase
- Position in major phrase
- ToBI endtone of current phrase

Context Clustering: HMM Structure

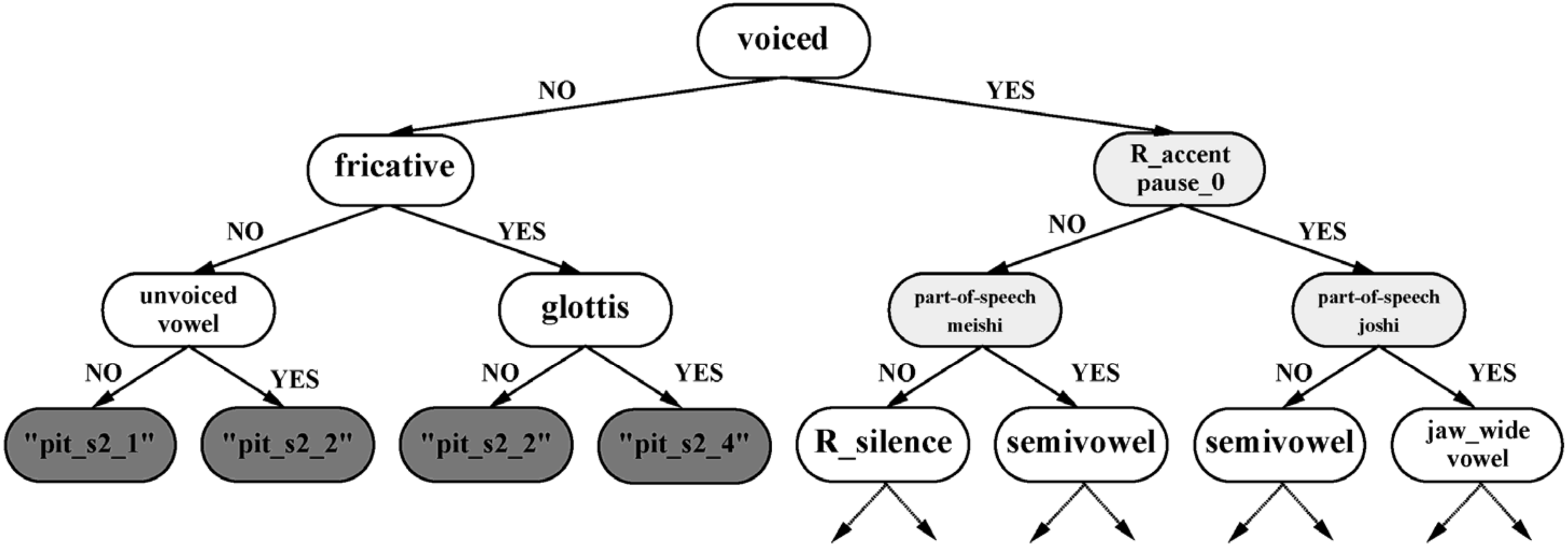


Tree for Spectrum (1st state)



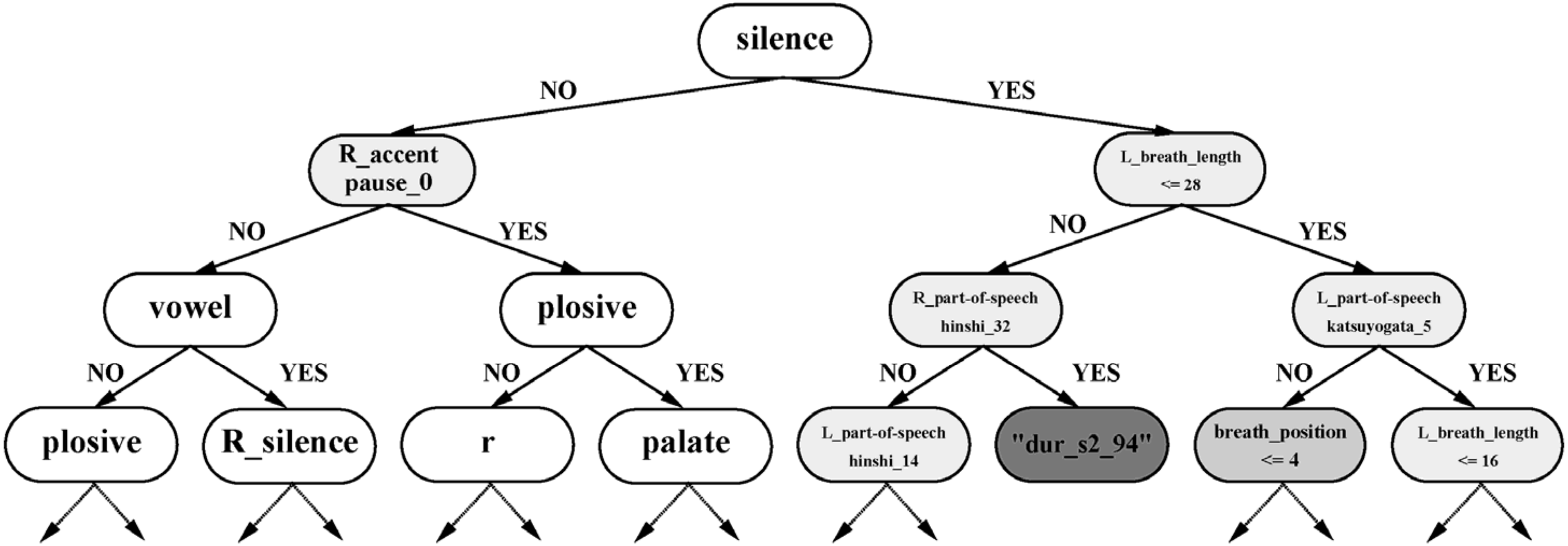
- Questions about phonetic attributes

Tree for F0 (1st state)



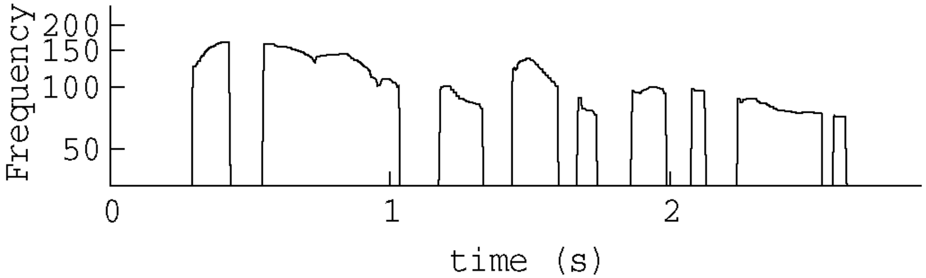
- Questions about linguistic attributes

Tree for State Duration

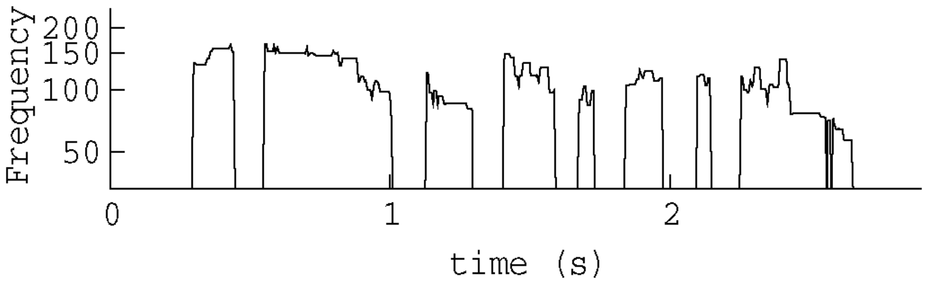


- Linguistic questions for pause
- Phonetic questions for speech

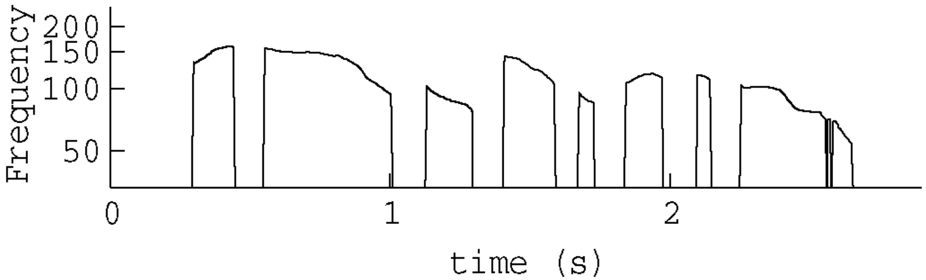
Generated *F0*



natural speech



without dynamic features



with dynamic features (Δ+Δ^2)

Effect of Dynamic Feature

Subjective Evaluation Result (Preference Score)

		Dynamic feature of spectrum: with	Dynamic feature of spectrum: without
Dynamic feature of $F0$	with	**91.3%**	**37.5%**
	without	**35.8%**	**11.8%**

「小さな鰻屋に，熱気のようなものがみなぎる」
“Chiisana unagiyani, nekkinoyouna monoga minagiru”

Overview of This Talk

- Basic Techniques
  - Vocoding technique
  - Speech Parameter generation algorithm
  - F0 pattern modeling
- Recent improvements and evaluation
- Relation to the unit selection approach
- Flexibility of the approach
  - Speaker adaptation (mimicking voices)
  - Speaker Interpolation (mixing voices)
  - Eigenvoices (producing voices), etc.

Recent Improvements

- Introduction of "hidden semi-Markov models"
- STRAIGHT vocoding
- Parameter generation considering global variance (GV)

⇨ Now, it is competitive to state-of-the-art unit selection systems

- Basic system
- 2005
- 2006

One-hour training data

Five-hour training data

Evaluation: Blizzard Challenge

- Speech Recognition

  Comparison on common datasets has been improving the core technology, e.g., DARPA, NIST

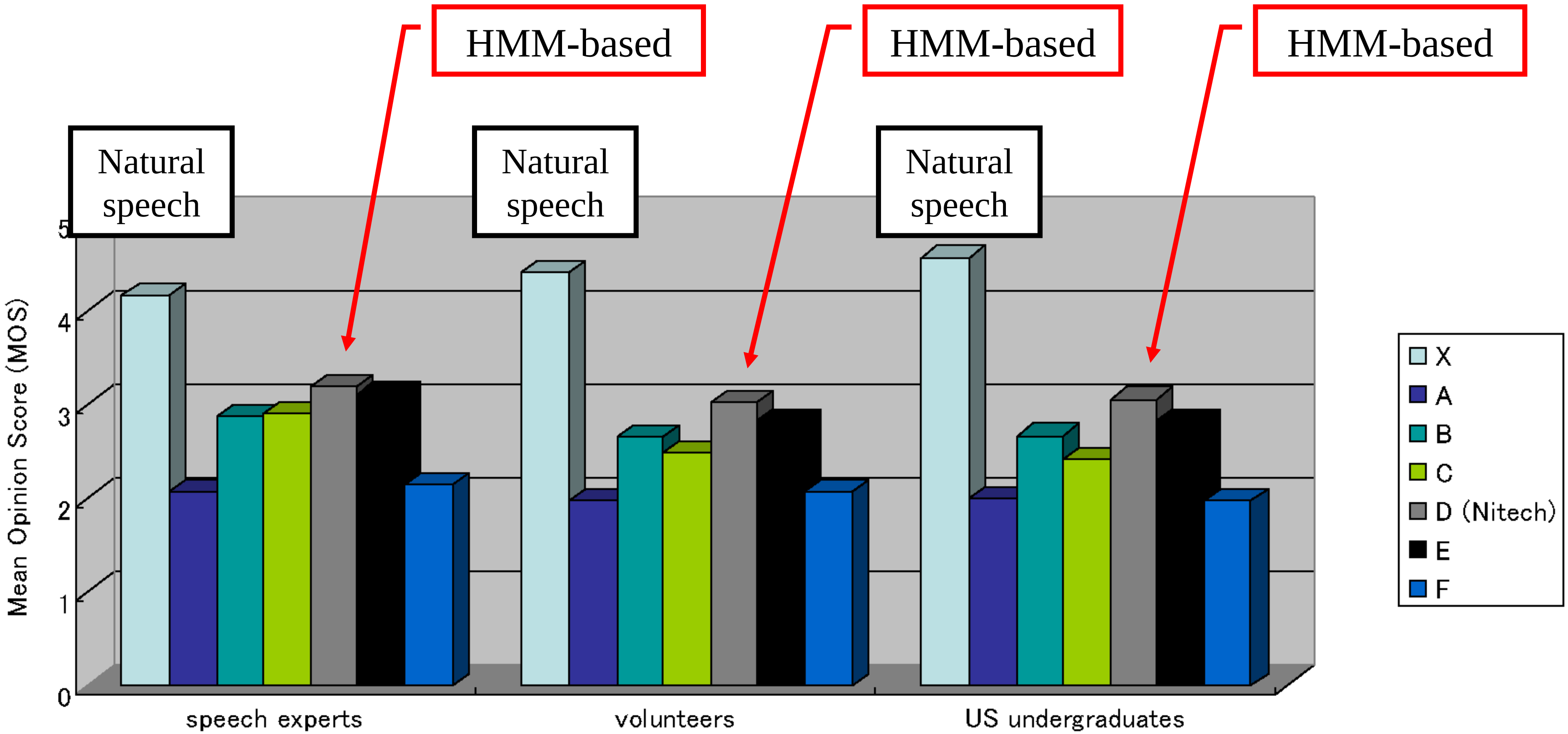
- Speech Synthesis

  It is necessary to compare speech synthesis techniques on common datasets

→ Blizzard Challenge 2005 and 2006

Results of Blizzard Challenge 2005

ARCTIC set (one hour training data)



The Software

Modified HTK (HTS) + SPTK

Modifications to HTK:

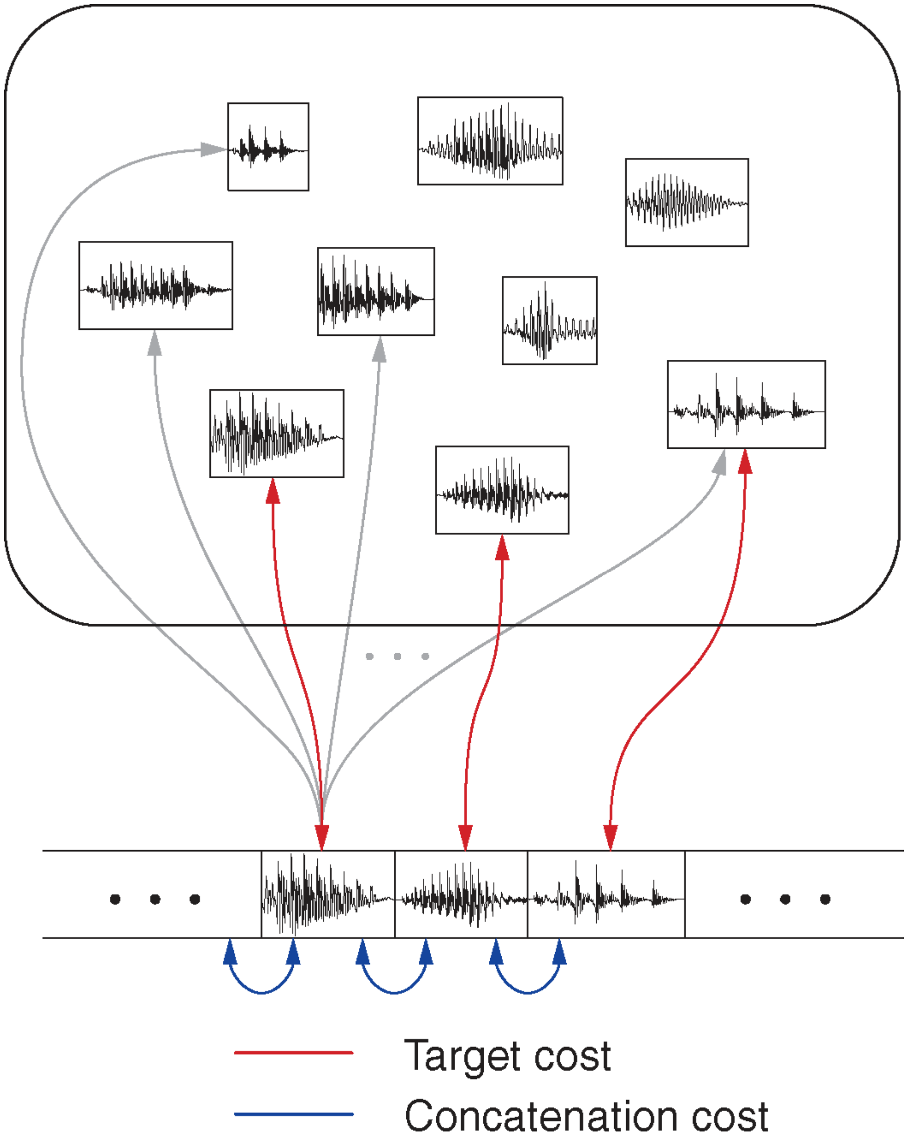
- Stream-dependent context clustering
- State output probability for $F0$ modeling
- State duration modeling and clustering

HMM-based Speech Synthesis System (HTS) (http://hts.ics.nitech.ac.jp)

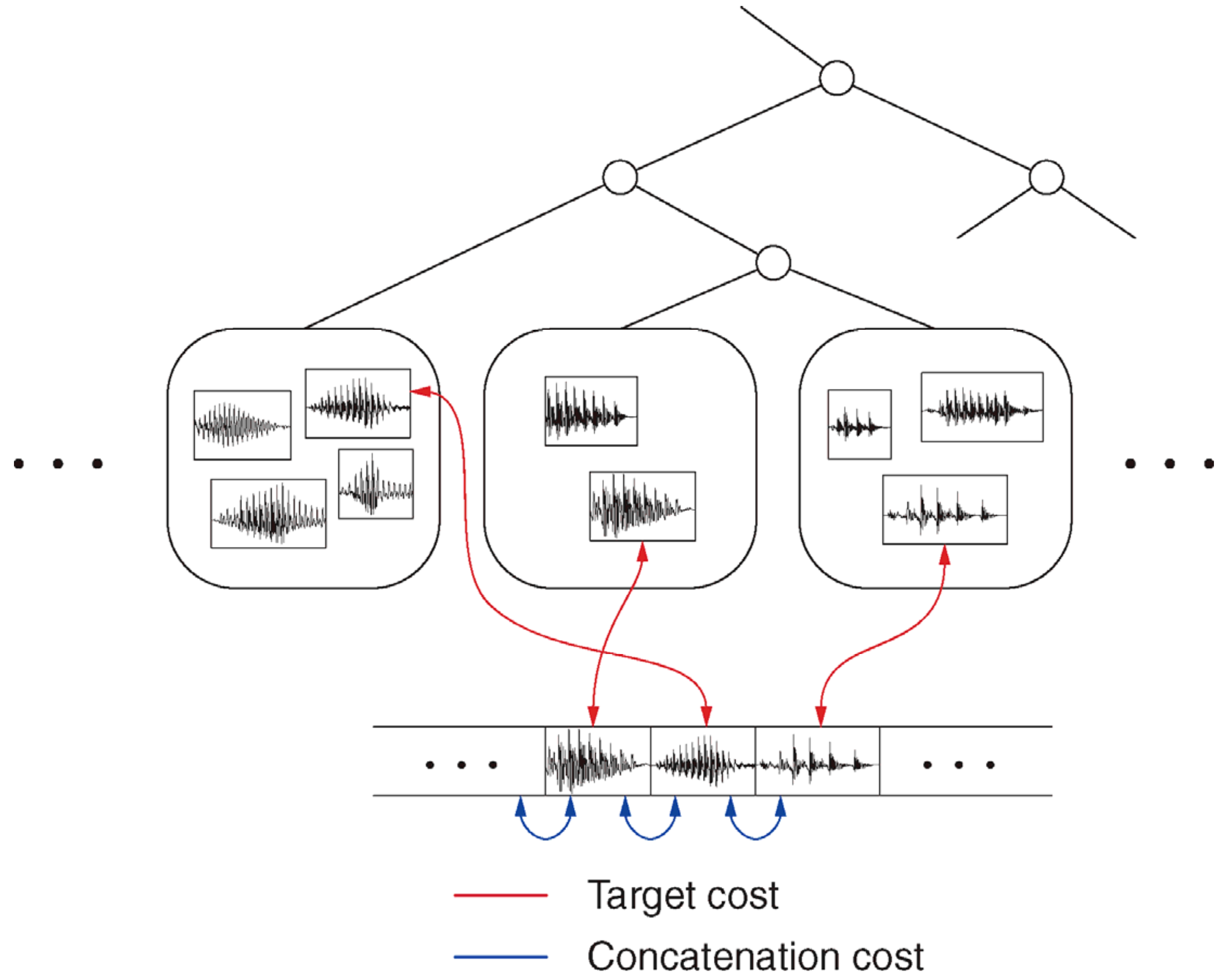
Overview of This Talk

- Basic Techniques
  - Vocoding technique
  - Speech Parameter generation algorithm
  - F0 pattern modeling
- Recent improvements and evaluation
- Relation to the unit selection approach
- Flexibility of the approach
  - Speaker adaptation (mimicking voices)
  - Speaker Interpolation (mixing voices)
  - Eigenvoices (producing voices), etc.

Unit selection



Unit Selection Based on Clustering



Comparison between Two Approaches

Unit selection	HMM-based
Clustering (possible use of HMM)	Clustering (use of HMM)
Multi-template of waveform	Statistics $\Rightarrow$ small footprint
Single tree for waveform (possible use of additional trees for prosody prediction)	Multiple tree for Spectrum, F0 duration
Advantage: • Waveform concatenation $\Rightarrow$ high quality speech Disadvantage: • Discontinuity • Hit or miss	Disadvantage: • Vocoder-based $\Rightarrow$ buzzy Advantage: • Smooth • Stable
• Fixed voice	• Various voices

Overview of This Talk

- Basic Techniques
  - Vocoding technique
  - Speech Parameter generation algorithm
  - F0 pattern modeling
- Recent improvements and evaluation
- Relation to the unit selection approach
- Flexibility of the approach
  - Speaker adaptation (mimicking voices)
  - Speaker Interpolation (mixing voices)
  - Eigenvoices (producing voices), etc.

What We Can Do?

- Emotional speech synthesis
- Speaker adaptation (mimicking voices)
- Speaker interpolation (mixing voices)
- Eigenvoices (producing voices)
- Multilingual speech synthesis
- Singing voice synthesis
- Audio-visual speech synthesis
- Human motion synthesis

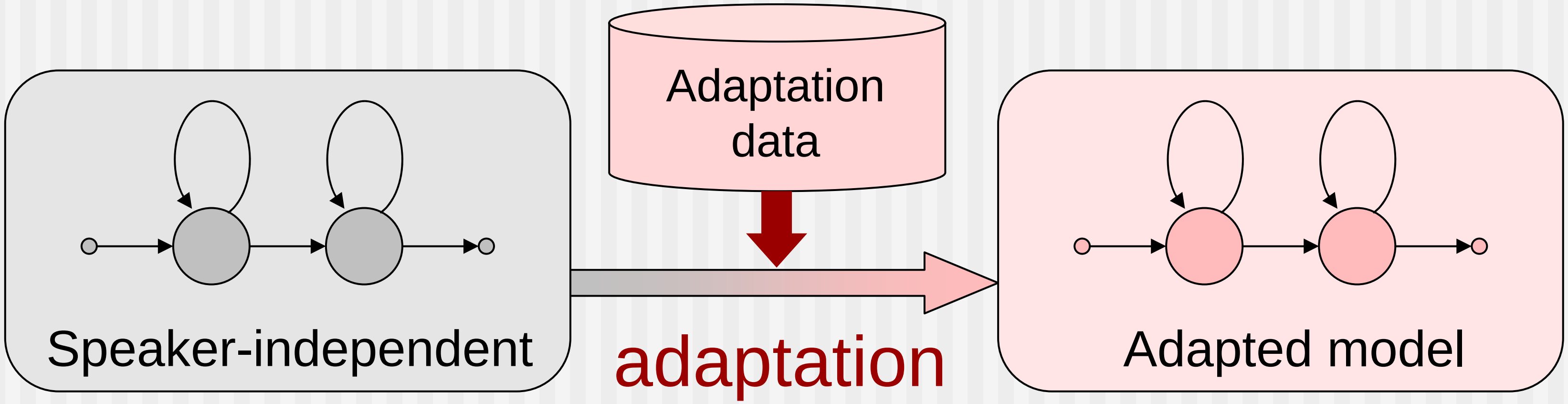
Emotional Speech Synthesis

text	neutral	angry
「授業中に携帯いじってんじゃねえよ！ 電源切っとけ！」 “Don’t touch your cell phone during a class! Turn off it!”		
「ミーティングには毎週参加しなさい！」 “You must attend the weekly meeting!”		

trained with 200 utterances

Speaker Adaptation (mimicking voices)

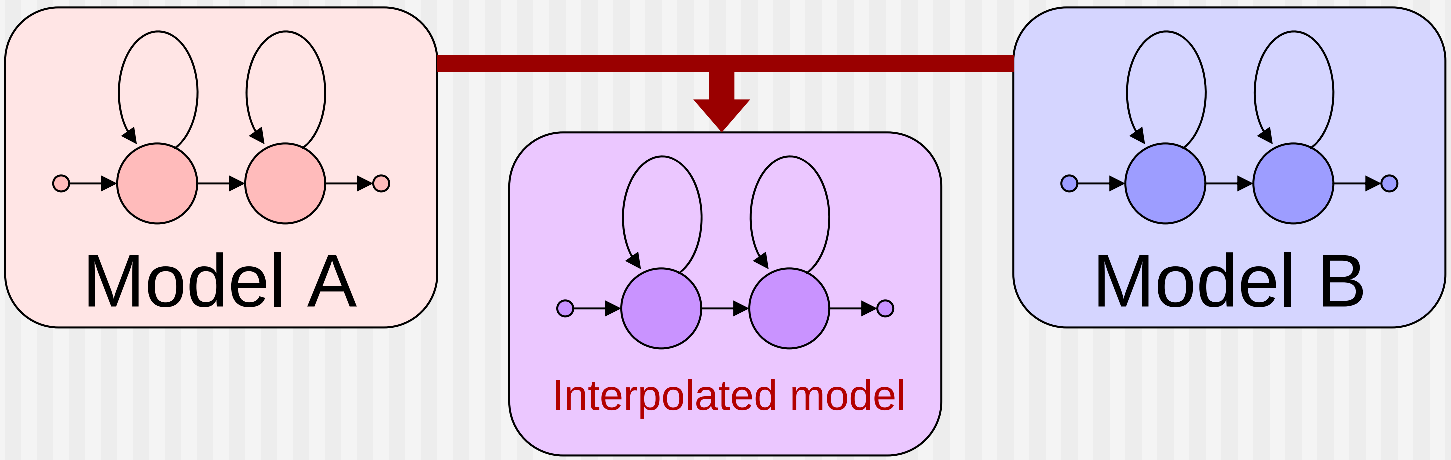
MLLR-based adaptation



- w/o adaptation (initial model)
- Adapted with 4 utterances
- Adapted with 50 utterances
- Speaker-dependent model

Speaker Interpolation (mixing voices)

Linear combination of two speaker-dependent models



A:	1.00	0.75	0.50	0.25	0.00

B:	0.00	0.25	0.50	0.75	1.00

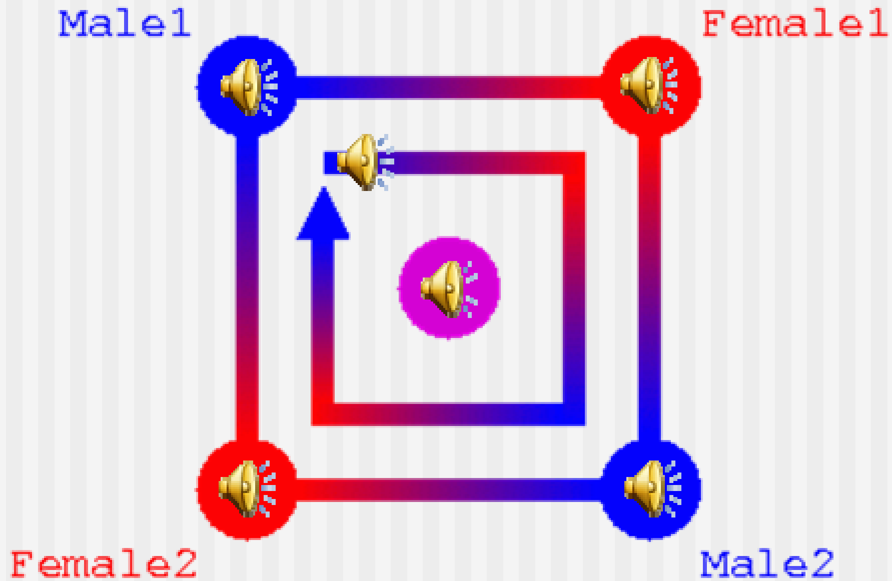
Voice Morphing

Two voices:

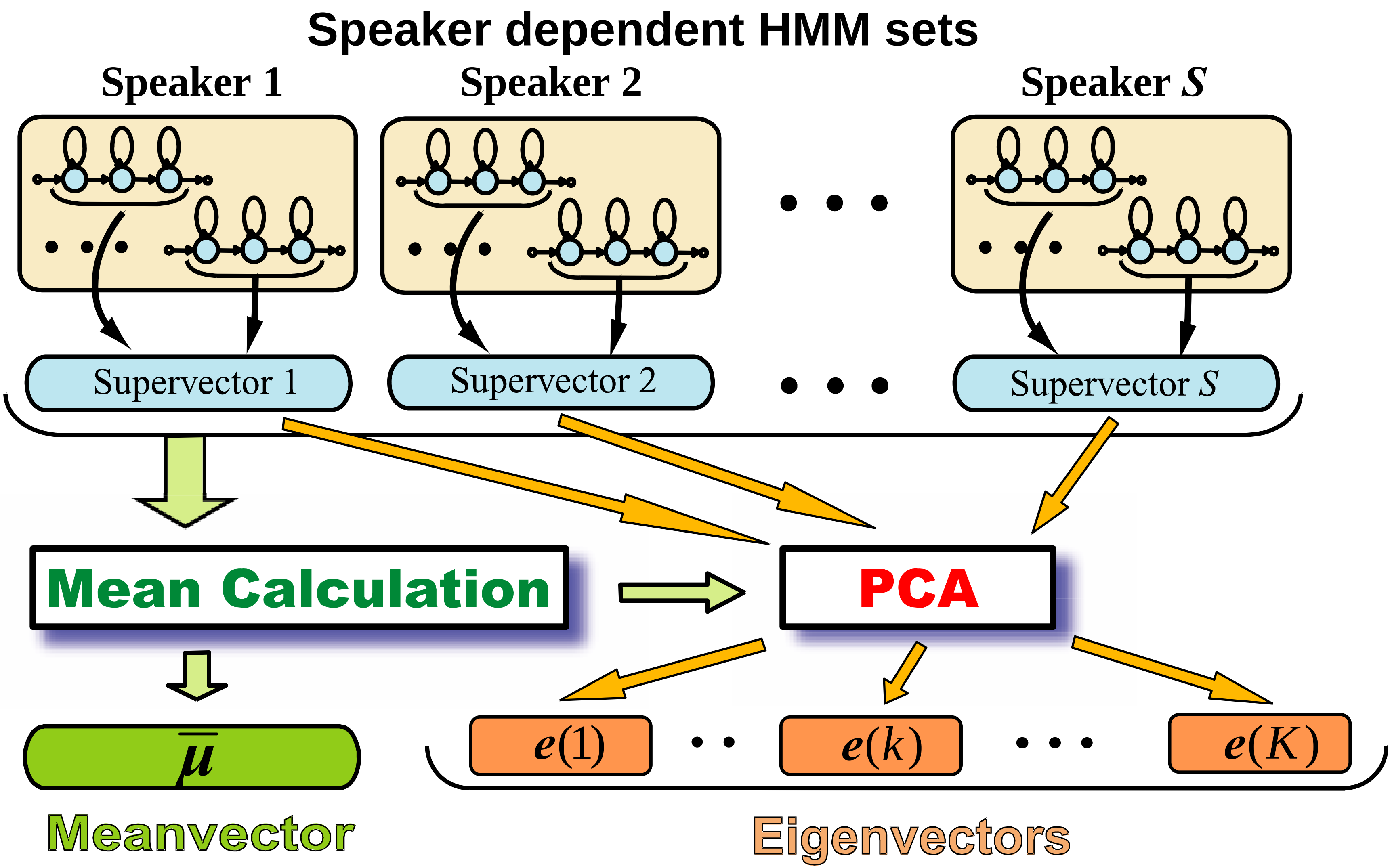
A ⇨ ⇨ ⇨ ⇨ ⇨ ⇨ ⇨ ⇨ ⇨ B

A ⇦ ⇦ ⇦ ⇦ ⇦ ⇦ ⇦ ⇦ ⇦ B

Four voices:



Eigenvoices (producing voices)



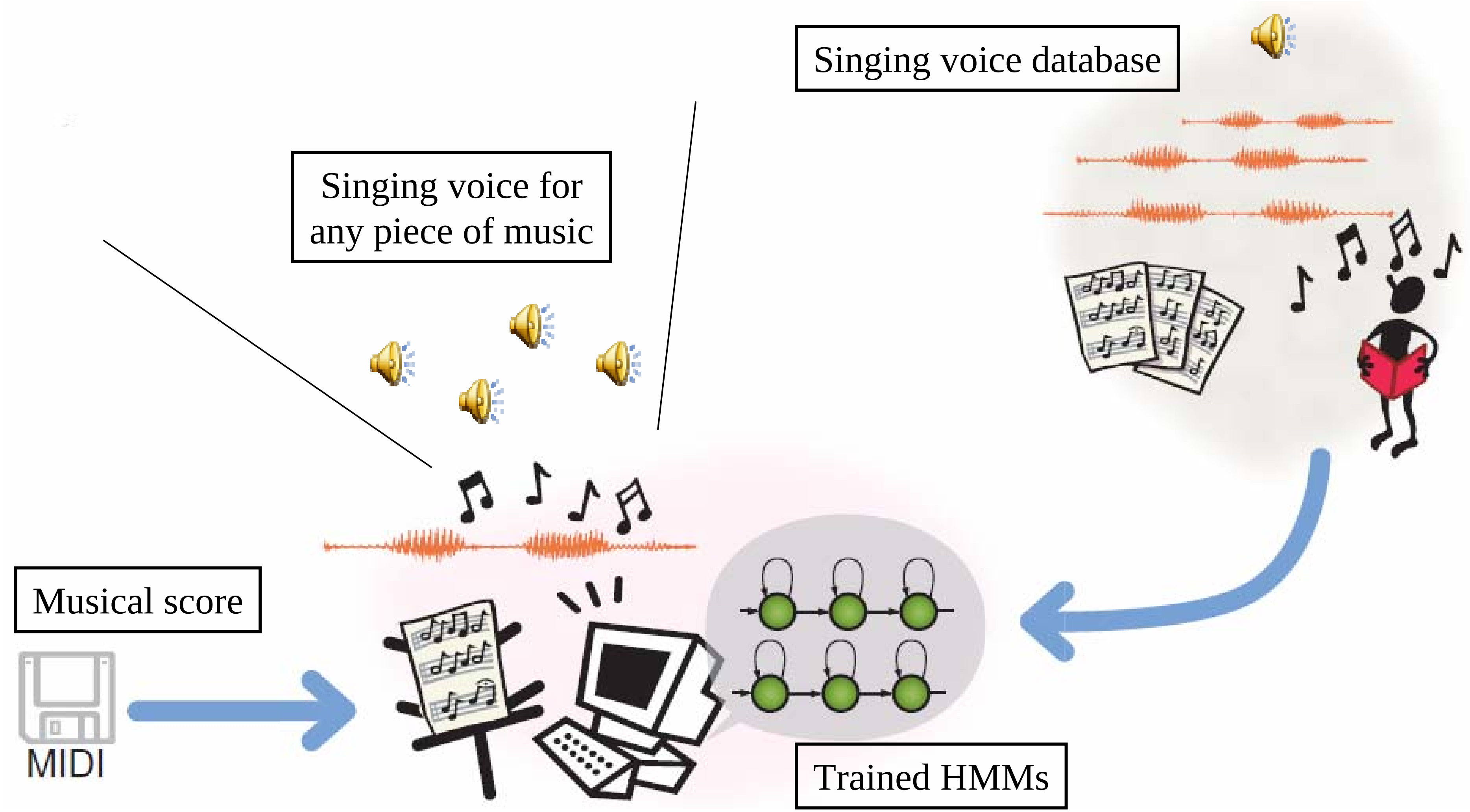
Click here for a demo

Multilingual Speech Synthesis

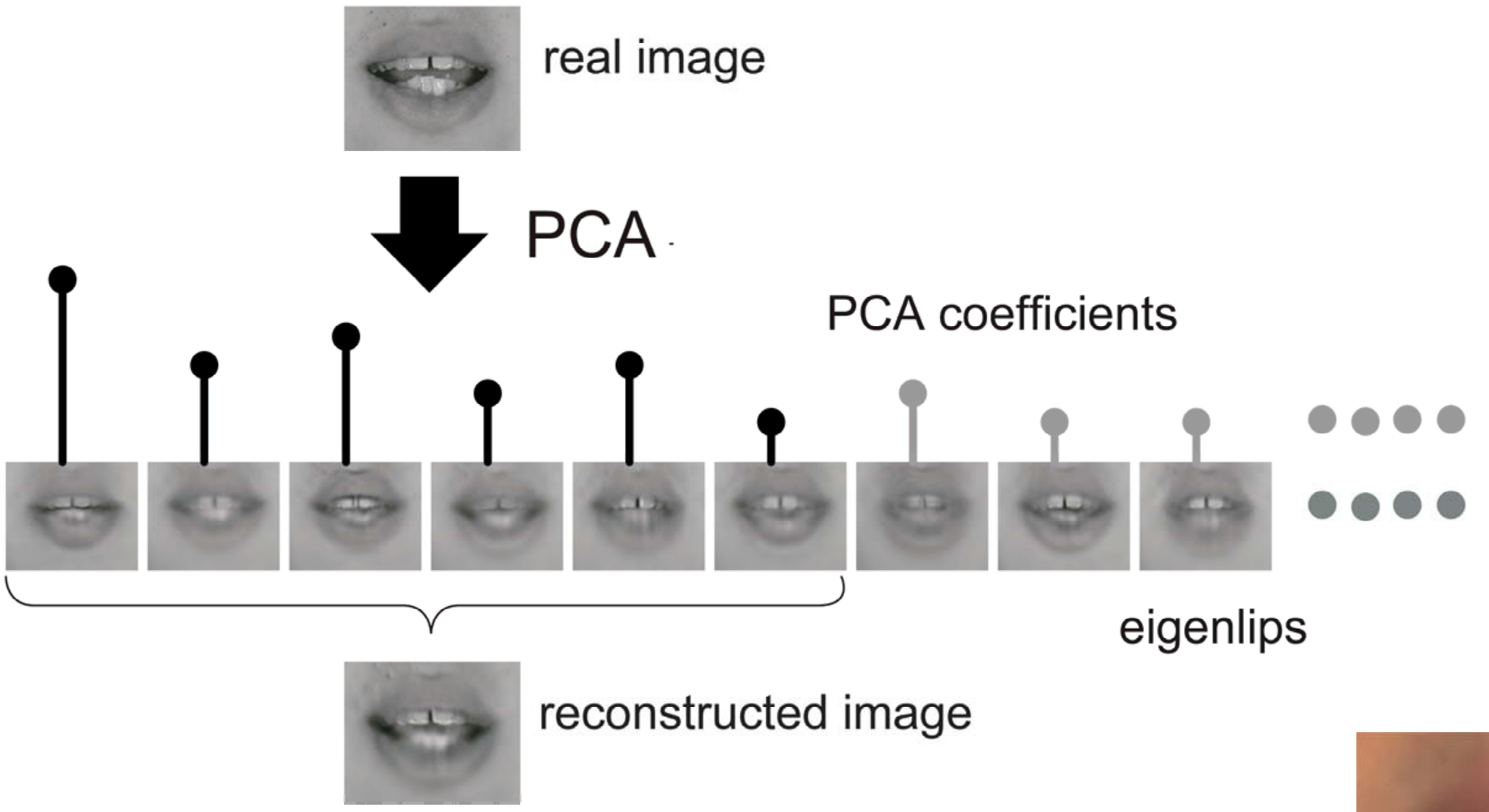
Latest system

- Japanese
- American English
- Chinese (Mandarin) (by ATR)
- Brazilian Portuguese (by Nitech, and UFRJ)
- European Portuguese
  (by Nitech, Univ of Porto, and UFRJ)
- Slovenian
  (by Bostjan Vesnicer, University of Ljubljana, Slovenia)
- Swedish
  (by Anders Lundgren, KTH, Sweden)
- German (by University of Bonn, and Nitech)
- Korean (by Sang-Jin Kim, ETRI, Korea)
- Polish, Slovak, Finnish, Arabic, Farsi, Polyglot, etc.

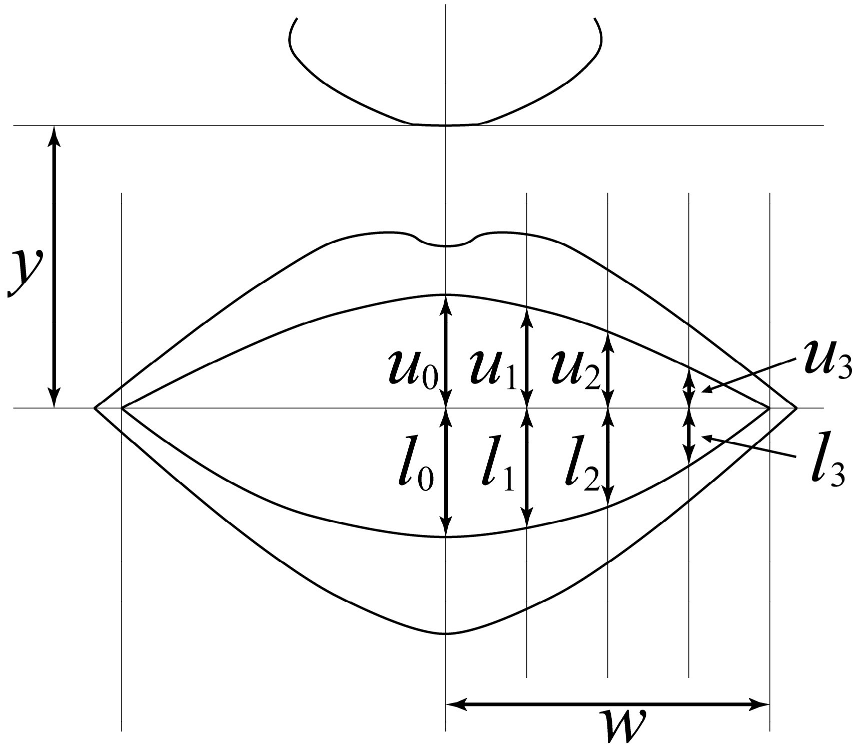
Singing Voice Synthesis



Audio-Visual Speech Synthesis (Pixel-based)



Audio-Visual Speech Synthesis (Model-based)



Click here for a demo
by Tamura, et al., Titech, Eurospeech99

Human Motion Synthesis and Others

Click here for various demos
by Prof Kobayashi's group at Titech

Small-Foot Print Synthesizer

- Acoustic model size < 100KB
- 0.1 Real Time

  - Sample 1
  - Sample 2
  - Sample 3
  - Sample 4
  - Sample 5

In A Dialog System

- **User**: 「バーカ!」"You Fool!"
- **Agent**: 「何よ！馬鹿って言う方が馬鹿なのよ!」
  "What? Who slanders others is a real fool!"

Summary

HMM-based Approach to Flexible Speech Synthesis

- Simultaneous modeling of spectrum, F0, and duration
- Provide flexibility: various voices, speaking styles, emotional expressions, etc.

A tool for constructing spoken dialogue systems