Fundamentals and recent advances in HMM-based speech synthesis

Keiichi TOKUDA

Nagoya Institute of Technology

Heiga ZEN

Toshiba Research Europe

TOSHIBA
Leading Innovation

Sept. 6, 2009

Time-line

14:15 ~ 15:45: First half

- Fundamentals
- Q&A (10min)

15:45 ~ 16:15: Break

16:15 ~ 17:45: Second half

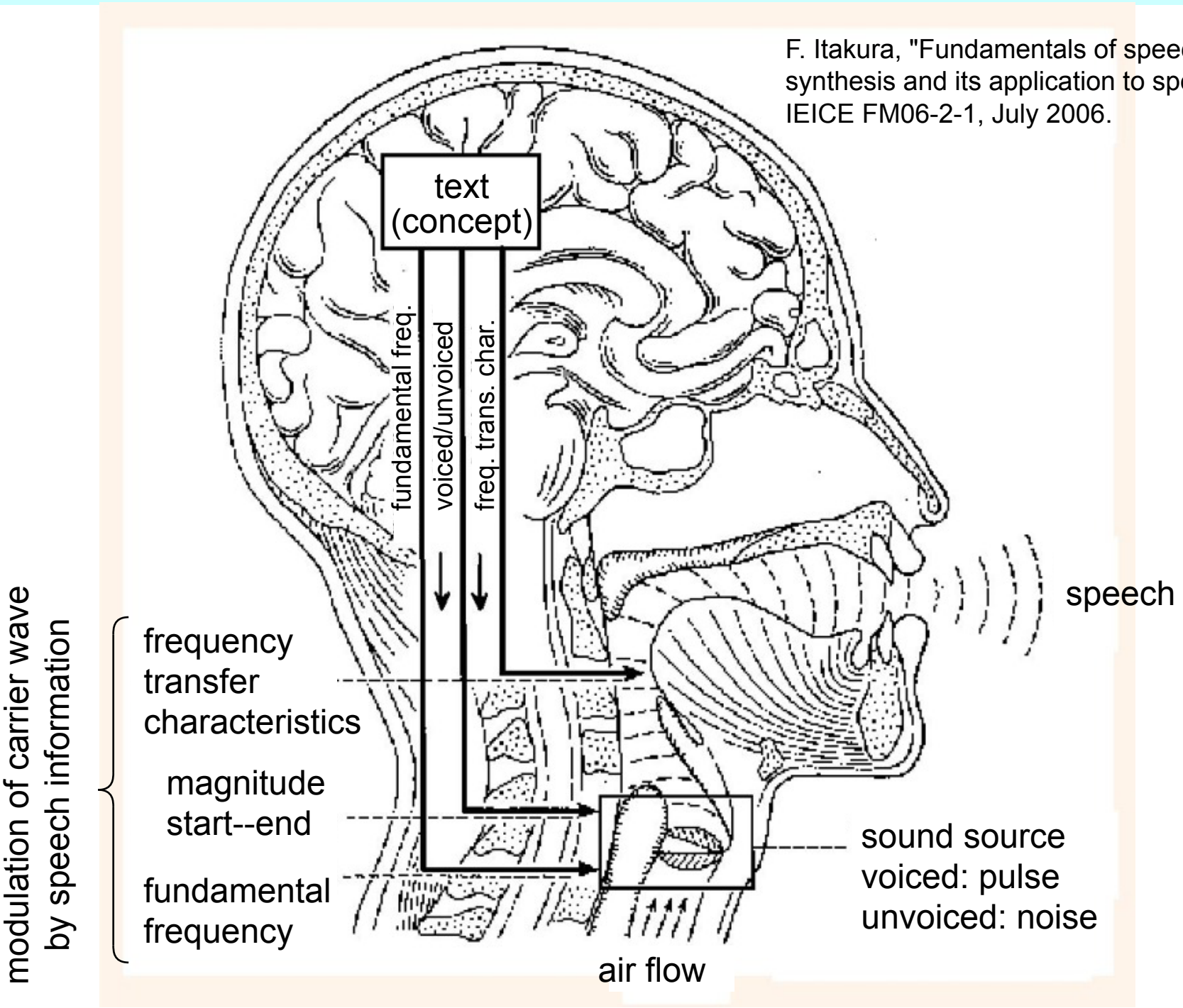
- Related topics
- Recent advances
- Applications
- Q&A (10min)

Time-line

14:15 ~ 15:45: First half

- Fundamentals
  - * Probabilistic formulation
  - * HTS framework
    - · Feature extraction
    - · HMM
    - · Parameter generation
    - · Waveform reconstruction
- Q&A (10min)

Speech production mechanism



Speech synthesis methods

- **Rule-based, *formant synthesis* (∼'90s)**
  - Hand-crafting each phonetic units by rules
  - Based on source-filter model
    * DECtalk [Klatt;'82]
- **Corpus-based, *concatenative synthesis* ('90s∼)**
  - Concatenate speech units (waveform) from a database
  - Large data + automatic learning

    ⇒ High-quality synthetic voices can be built automatically
  - Single inventory: diphone synthesis [Moulines;'90]
  - Multiple inventory: unit selection synthesis (USS)
    * ATR ν-Talk [Sagisaka;'92], CHATR [Black;'96]
    * AT&T Next-Gen TTS [Beutnagel;'99]

Speech synthesis methods

- **Corpus-based, *statistical parametric synthesis***
  - Proposed in mid-'90s, becomes popular since mid-'00s
  - Large data + automatic training
    ⇒ Automatic voice building
  - Source-filter model + statistical acoustic model
    ⇒ Flexible to change its voice characteristics
  - HMM as its statistical acoustic model
    ⇒ HMM-based speech synthesis (HTS) [Yoshimura;'99]

This tutorial focuses on HMM-based speech synthesis

Time-line

14:15 ~ 15:45: First half

- Fundamentals
  * Probabilistic formulation
  * HTS framework
    - Feature extraction
    - HMM
    - Parameter generation
    - Waveform reconstruction
- Q&A (10min)

Formulation of corpus-based synthesis

- **Speech synthesis based on the ML criterion**

  - Training

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} p(\boldsymbol{O} \mid \mathcal{W}, \boldsymbol{\lambda})$$

  - Synthesis

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid w, \hat{\boldsymbol{\lambda}})$$

$\boldsymbol{\lambda}$: model parameters

$\boldsymbol{O}$: training data

$\mathcal{W}$: transcriptions

$\boldsymbol{o}$: synthesized speech

w : input text

- **Using HMM as its acoustic model**

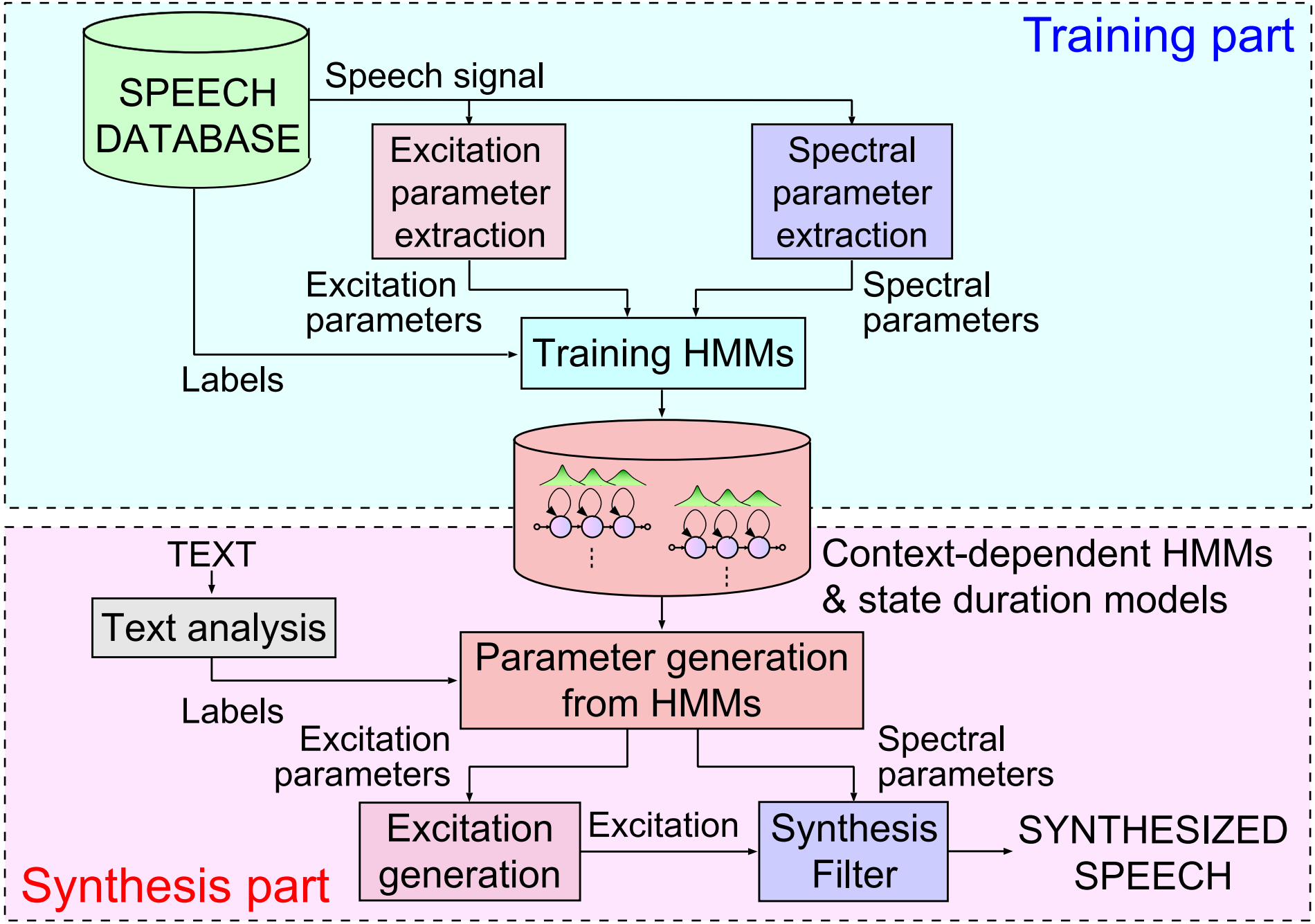
  $\Rightarrow$ HMM-based speech synthesis (HTS) [Yoshimura;'02]

HMM-based speech synthesis system (HTS)

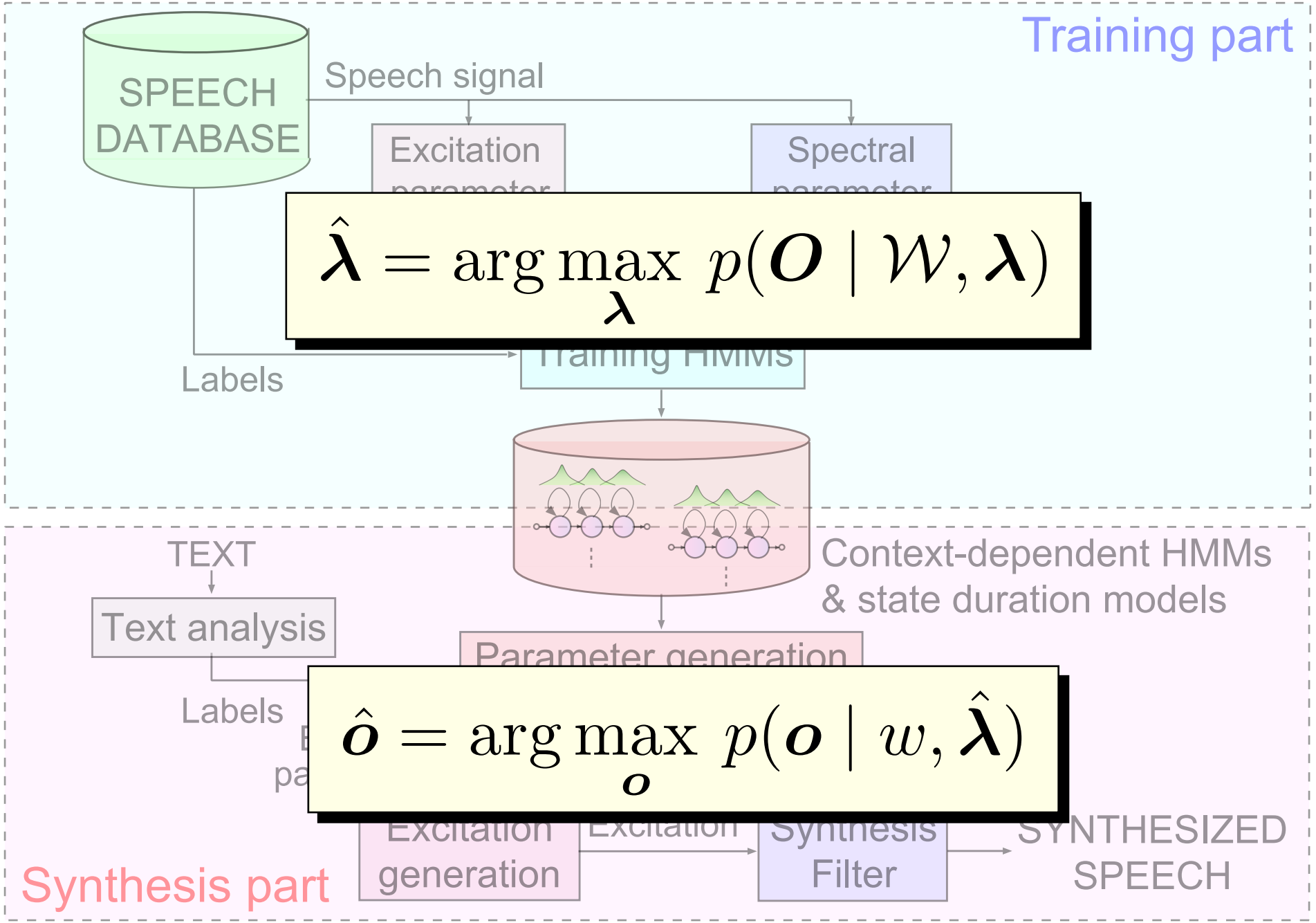
- **HTS: Toolkit for HMM-based speech synthesis**
  - Web: http://hts.sp.nitech.ac.jp/
  - Research platform for HMM-based speech synthesis
  - Released as a patch code for HTK
  - Latest verion: HTS-2.1 (July 2008)
  - Speaker dependent (SD) / adaptation (SA) demo scripts
    * SD: HTS-demo_CMU-ARCTIC-SLT
    * SA: HTS-demo_CMU-ARCTIC-ADAPT

We will explain the HTS framework in details based on the implementation of the SD demo script (HTS-2.0.1)

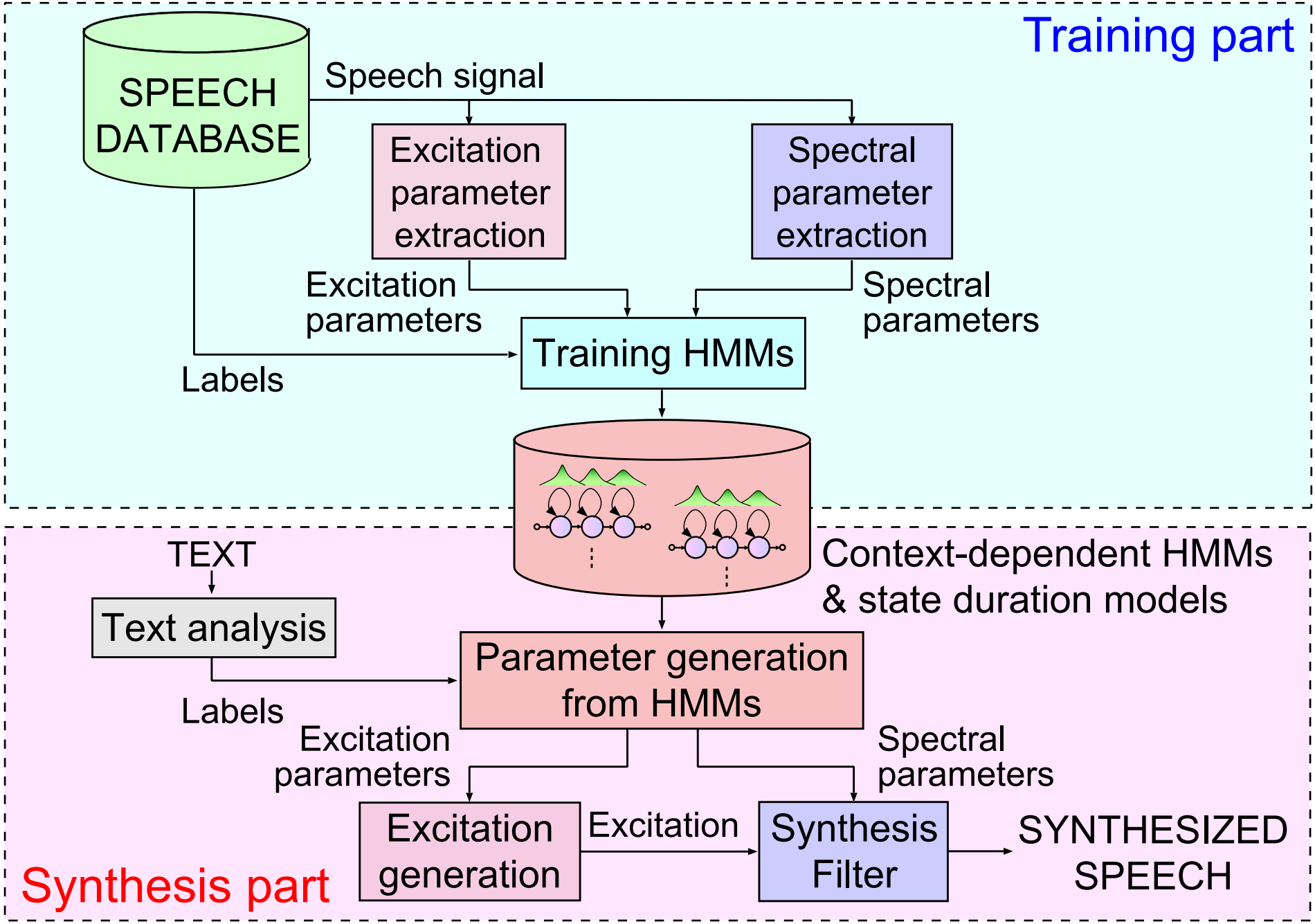
HMM-based speech synthesis system (HTS)



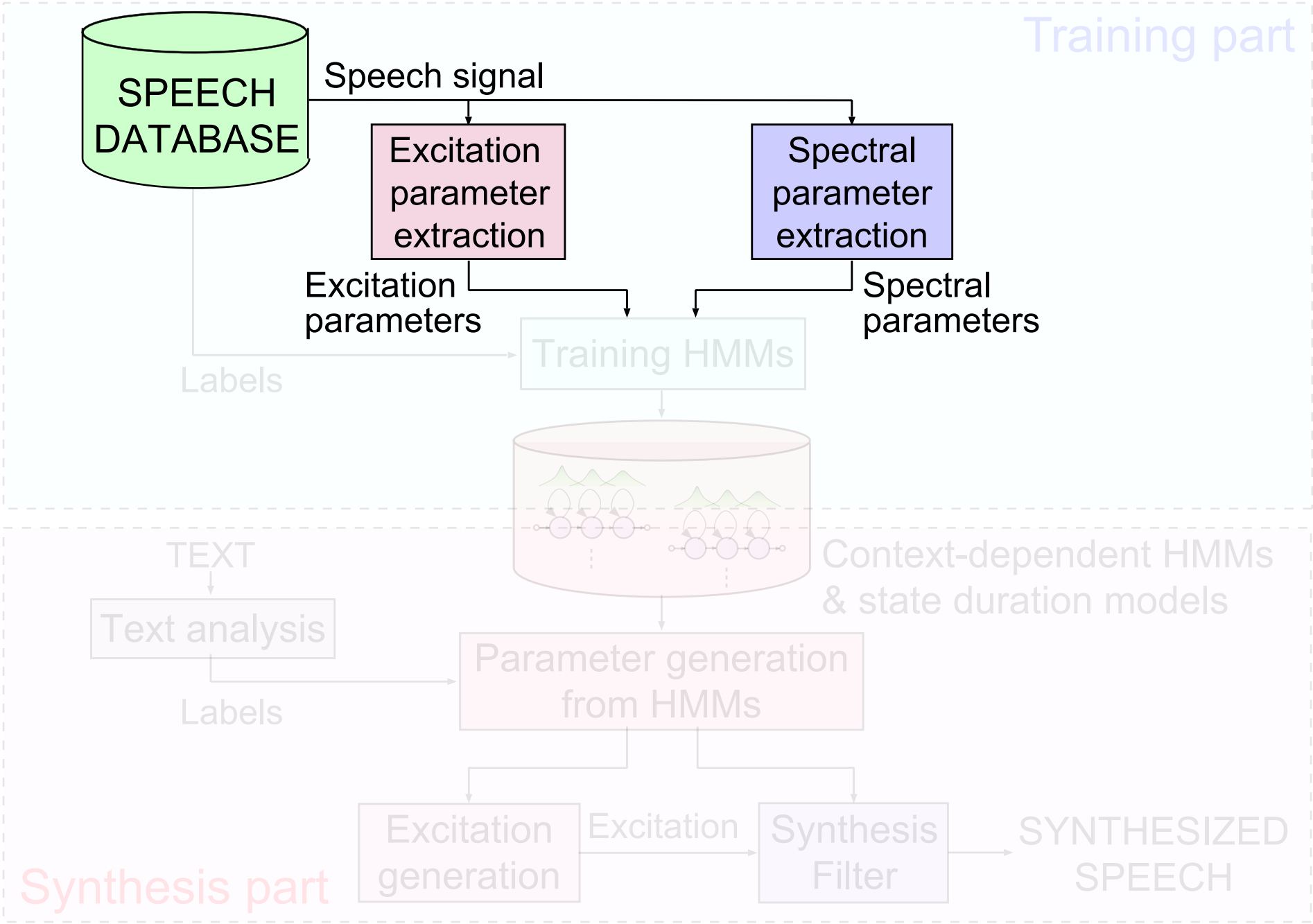
HMM-based speech synthesis system (HTS)



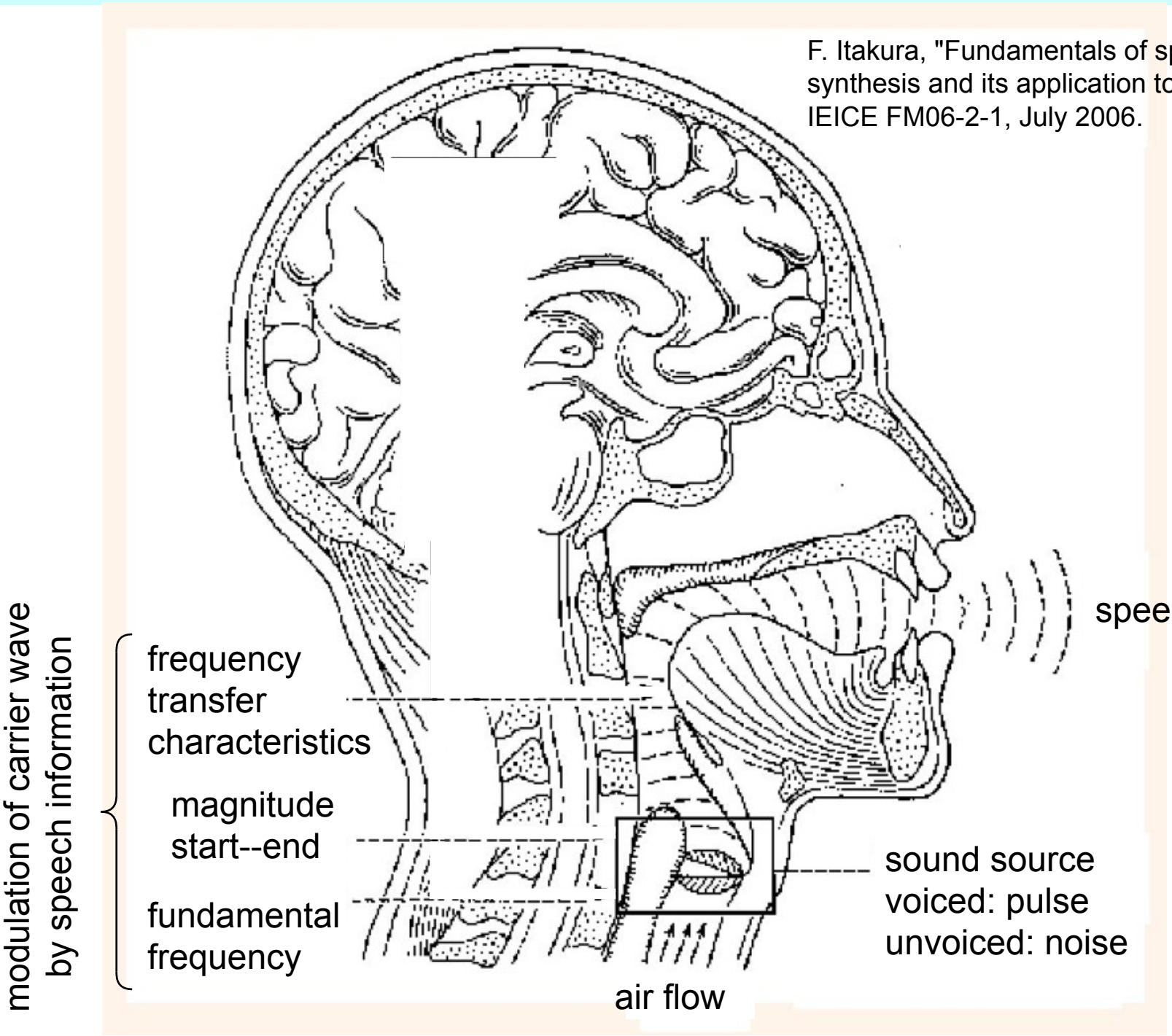
HMM-based speech synthesis system (HTS)



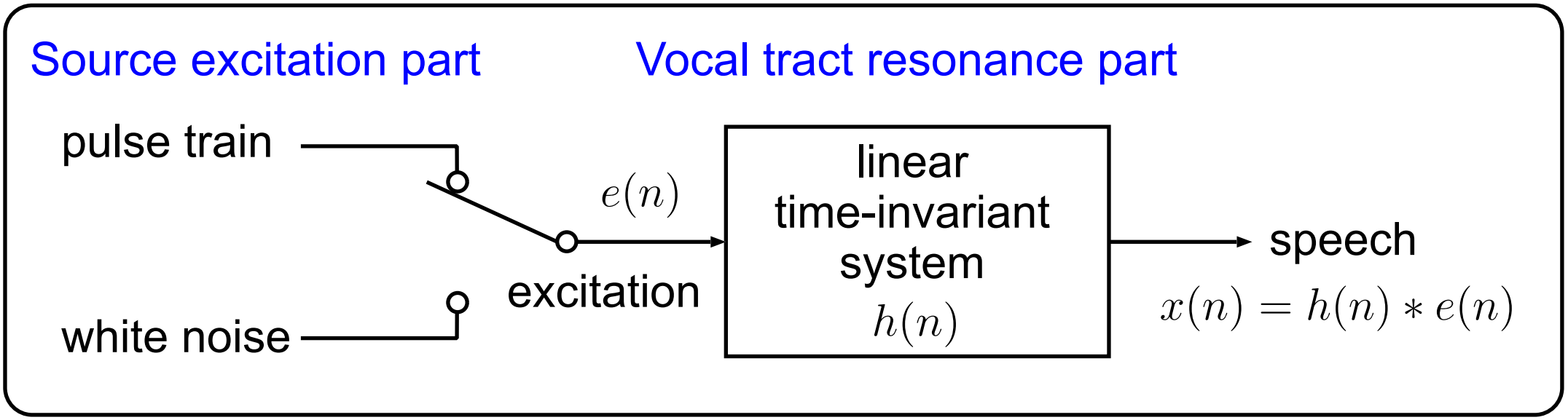
HMM-based speech synthesis system (HTS)



Speech production mechanism



Source-filter model



$$x(n) = h(n) * e(n)$$

$$\downarrow \text{Fourier transform}$$

$$X(e^{j\omega}) = H(e^{j\omega})E(e^{j\omega})$$

$H(e^{j\omega})$ should be defined by the state-output vector of HMMs,

e.g., mel-cepstral coefficients, LSP coefficients

Maximum-likelihood estimation of spectral model

Parametric models of speech spectrum

Autoregressive (AR) model

$$H(z) = K \Big/ \left\{ 1 - \sum_{m=0}^{M} c(m) z^{-m} \right\}$$

Exponential (EX) model

$$H(z) = \exp \sum_{m=0}^{M} c(m) z^{-m}$$

Estimate parameters of spectral model to maximize its likelihood

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

$p(\boldsymbol{x} \mid \boldsymbol{c})$: AR model $\Rightarrow$ Linear prediction (LP) [Itakura;'70]

$p(\boldsymbol{x} \mid \boldsymbol{c})$: EX model $\Rightarrow$ ML-based cepstral analysis

Generalized cepstral analysis [Kobayashi;'84]

Generalized cepstral coefficients: $c_\gamma(m)$

$$H(z) = s_\gamma^{-1}\left(\sum_{m=0}^{M} c_\gamma(m) z^{-m}\right)$$

$$= \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_\gamma(m) z^{-m}\right)^{1/\gamma}, & 0 < |\gamma| \leq 1 \\ \exp \sum_{m=0}^{M} c_\gamma(m) z^{-m}, & \gamma = 0 \end{cases}$$

Introduction of auditory frequency scale

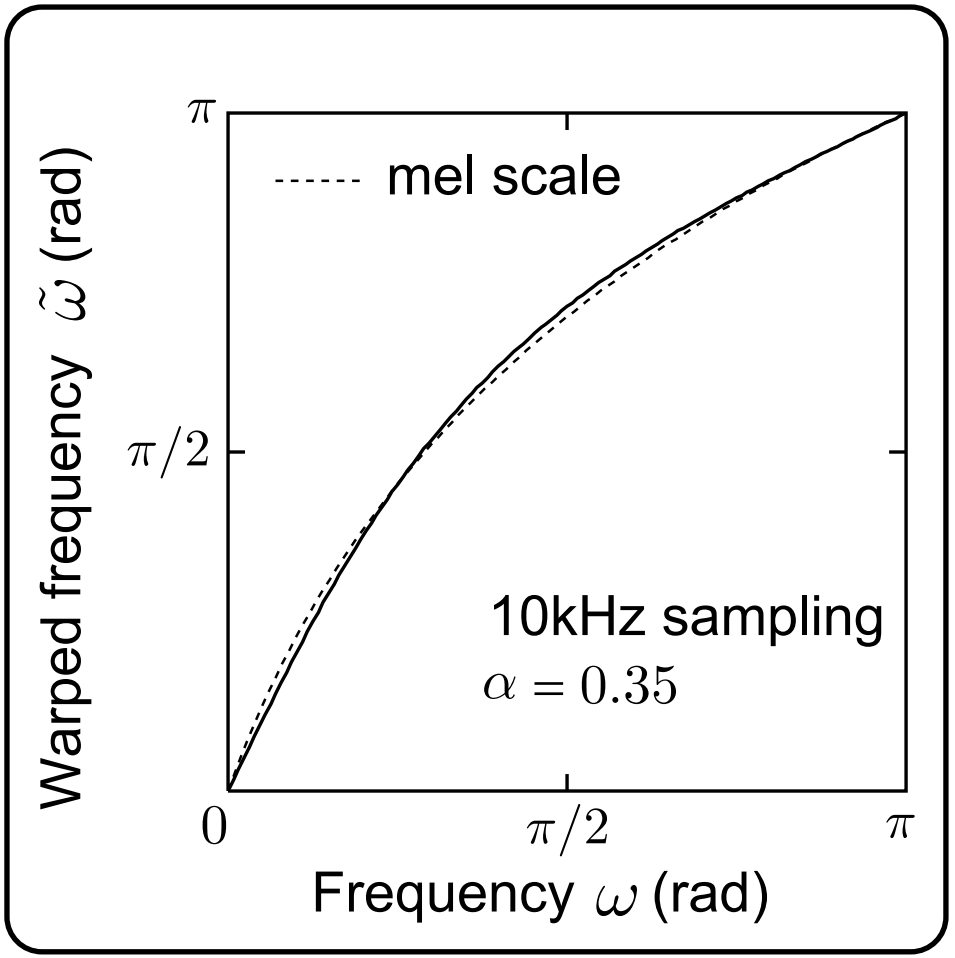
First-order all-pass function:

$$z_\alpha^{-1} = \Psi(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}$$

Phase chacteristics can be used for frequency transformation:

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}$$

where $\Psi(e^{j\omega}) = e^{-j\tilde{\omega}}$



Mel-generalized cepstral analysis [Tokuda;'94]

Mel-generalized cepstral coefficients: $c_{\alpha,\gamma}(m)$

$$H(z) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m}\right)^{1/\gamma}, & 0 < |\gamma| \leq 1 \\ \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m}, & \gamma = 0 \end{cases}$$

$$z_{\alpha}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}$$

Mel-generalized cepstral analysis

$(\alpha, \gamma) = (0, 0)$

$\Rightarrow$ Cepstral analysis

$$H(z) = \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z^{-m}$$

$(\alpha, \gamma) = (0, -1)$

$\Rightarrow$ LP analysis

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z^{-m}}$$

$(\alpha, \gamma) = (0.42\,, 0)$

$\Rightarrow$ Mel-cepstral analysis

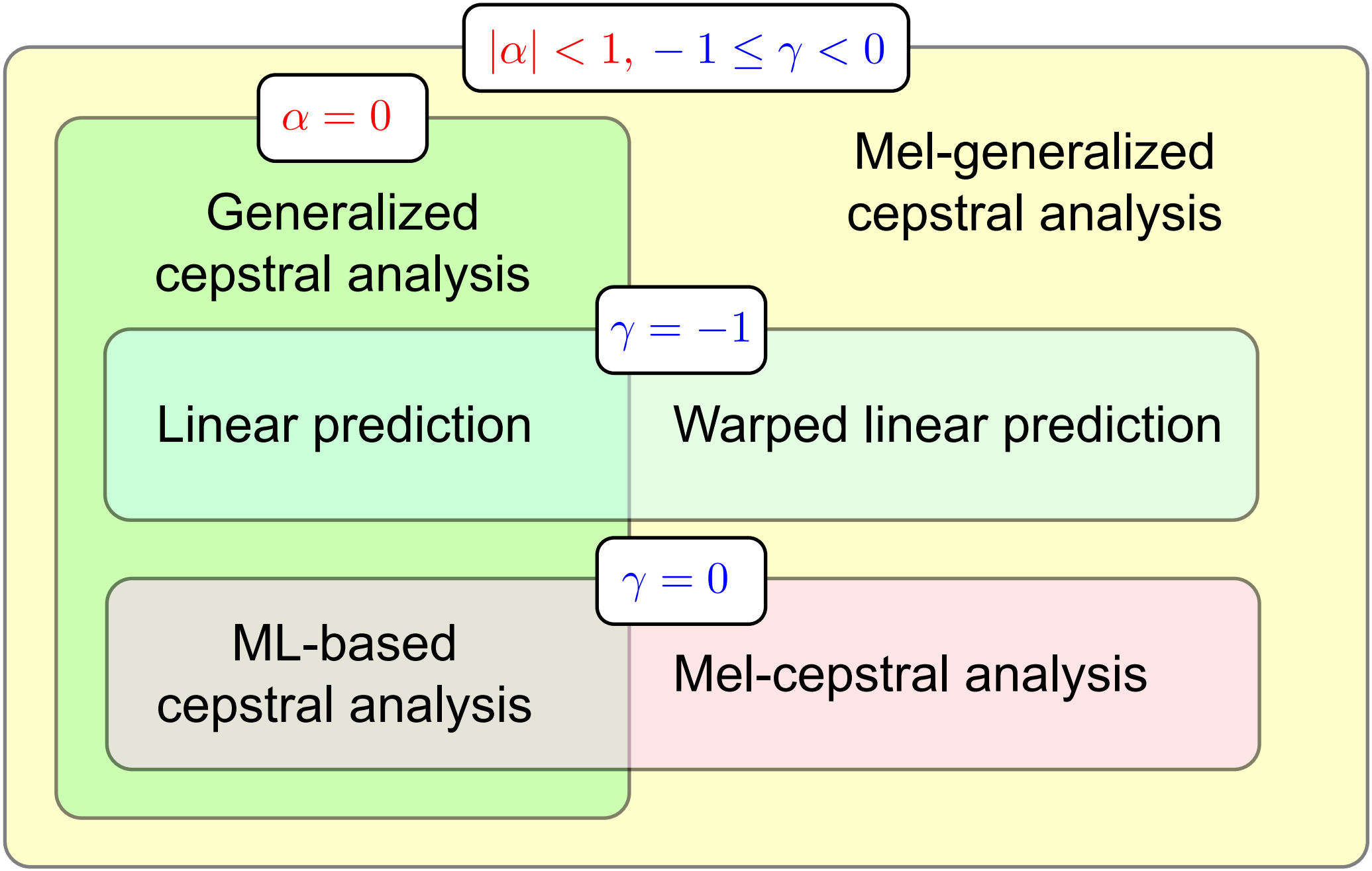
$$H(z) = \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m}$$

$(\alpha, \gamma) = (0.42\,, -1)$

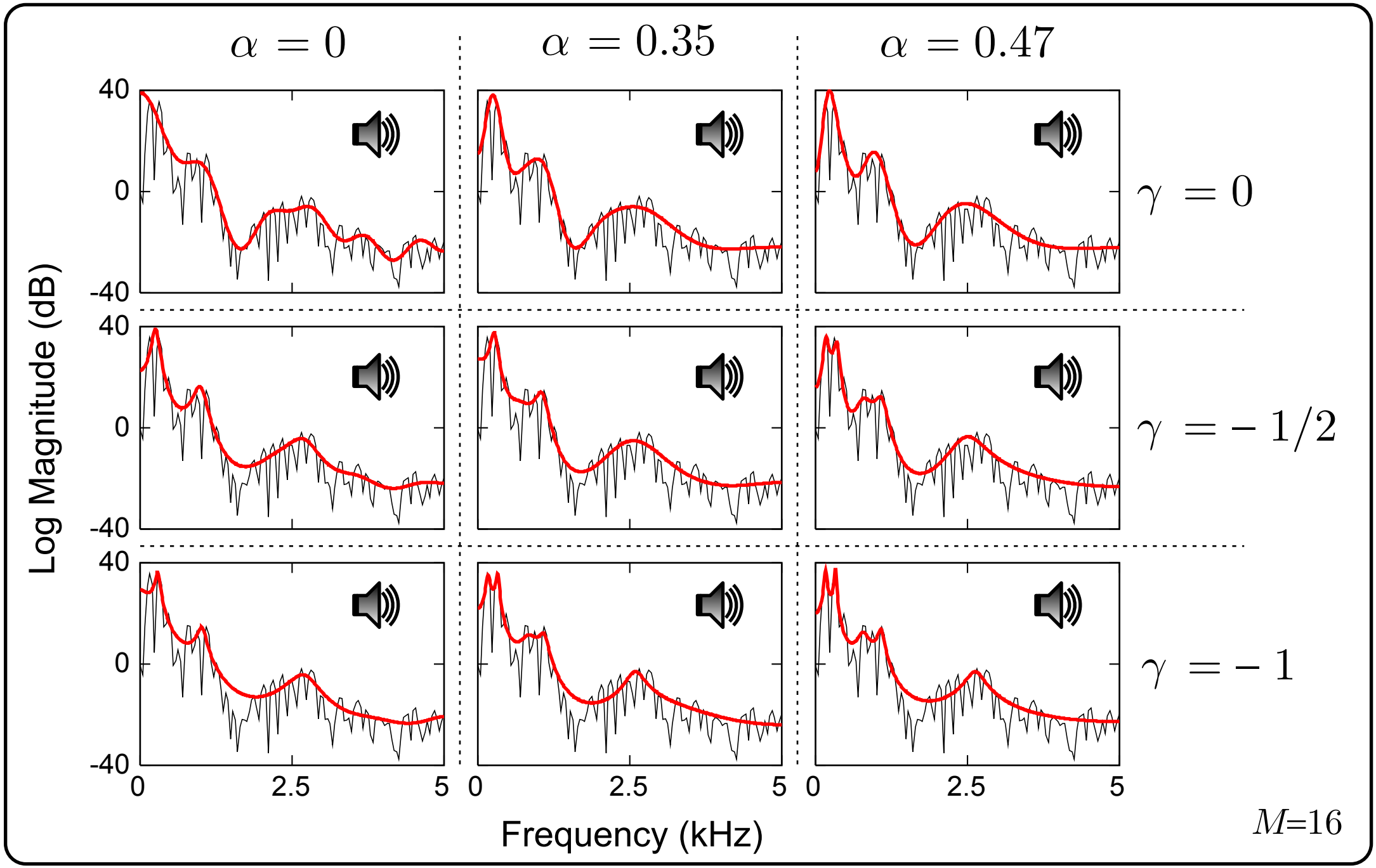
$\Rightarrow$ Warped LP analysis

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m}}$$

A unified approach to speech spectral estimation



Mel-generalized analysis of natural speech



Quantization of spectral parameters

- **Problems**
  - Keep synthesis filter stable after quantization
  - Quantize parameters as little distortion as possible
- $\gamma = 0$ **(cepstral model)**
  - Stability is always ensured for any $c_{\alpha,\gamma}(m)$
- $-1 \leq \gamma < 0$
  - Direct quantization of $c_{\alpha,\gamma}(m)$ may cause unstable synthesis filter
  - Not easy to check the filter stability

$\Rightarrow$ Transform $c_{\alpha,\gamma}(m)$ into other representations

Stability theorem

- Transformation should be one-to-one mapping
- Decomposition of polynomial [Itakura;'75]

$$H(z) = \left(1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m}\right)^{1/\gamma} = \frac{1}{\{C(z_{\alpha})\}^{n}}$$

$$(n = -1/\gamma)$$

$$C(z_{\alpha}) = \underbrace{C_p(z_{\alpha})}_{\text{symmetric}} + \underbrace{C_q(z_{\alpha})}_{\text{antisymmetric}}$$

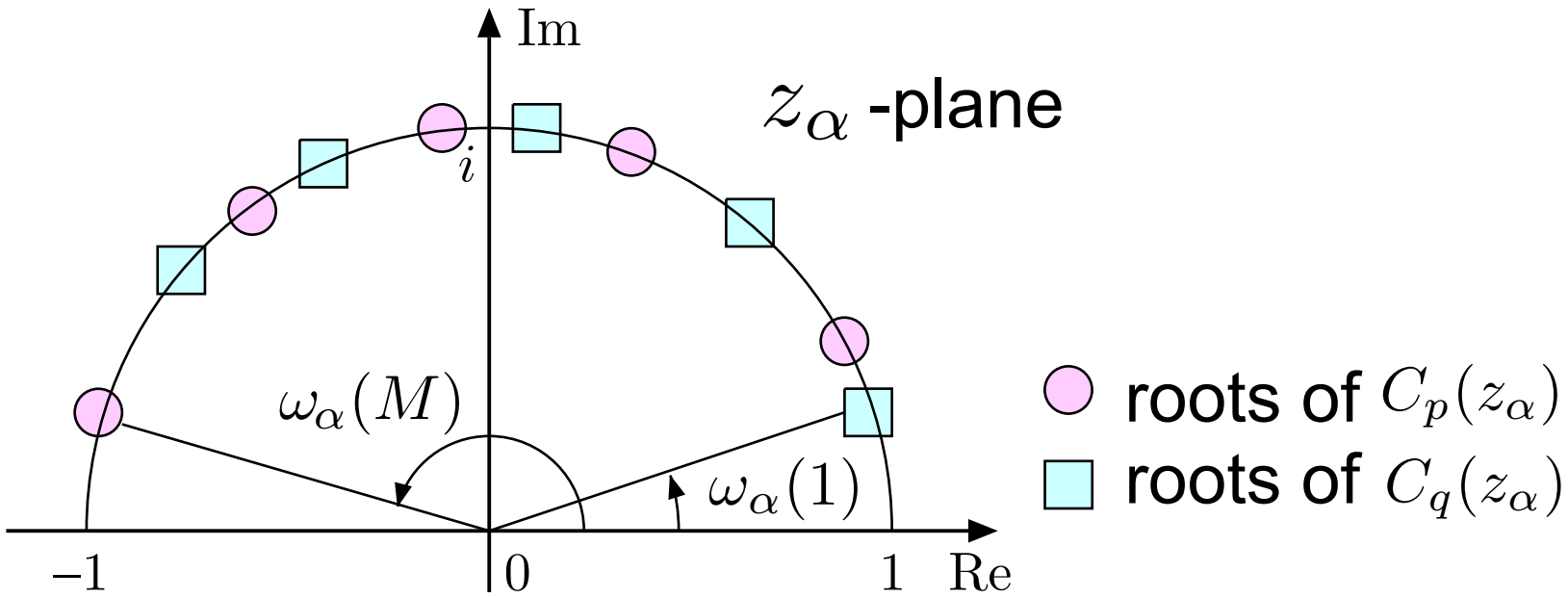
For all of the roots of $C(z_{\alpha})$ to be inside the unit circle:

- Zeros of $C_p(z_{\alpha})$ & $C_q(z_{\alpha})$ are located on the unit circle
- They are simple, & they separete each other

MGC-based line spectrum pairs (MGC-LSPs)

$$C_1(z_\alpha) = C_p(z_\alpha) + C_q(z_\alpha)$$
$$C_p(z_\alpha) = C_1(z_\alpha) + z_\alpha^{(M+1)} C_1(z_\alpha^{-1})$$
$$C_q(z_\alpha) = C_1(z_\alpha) - z_\alpha^{(M+1)} C_1(z_\alpha^{-1})$$

Root location of MGC-LSP $[\omega_\alpha(1), \ldots, \omega_\alpha(M)]$



The choise of α & γ for speech analysis

Analysis/synthesis speech with fixed α & γ

- Speech quality changes with γ
  - $\gamma \rightarrow -1$ (all pole) Clear
  - $\gamma \rightarrow 0$ (cepstral) Smooth
- 18-order LSP analysis gives almost the same quality as 24-order mel-cepstral analysis [Kim;'06]
- When the analysis order is high enough, the difference becomes small

Speech signal processing toolkit (SPTK)

- **SPTK: Toolkit for speech signal processing**
  - Web: http://sptk.sourceforge.net/
  - Open source, BSD license
  - Supports various basic operations
    * Speech analysis & synthesis
    * Speech parameter conversion
    * Filtering, transforms, VQ
    * Data manipulation, data conversion

HTS-demo scripts adopt SPTK for speech analysis

Speech signal processing toolkit (SPTK)

Analysis

LP analysis: lpc

ML-based cepstral analysis: uels

Generalized cepstral analysis: gcep

Mel-cepstral analysis: mcep

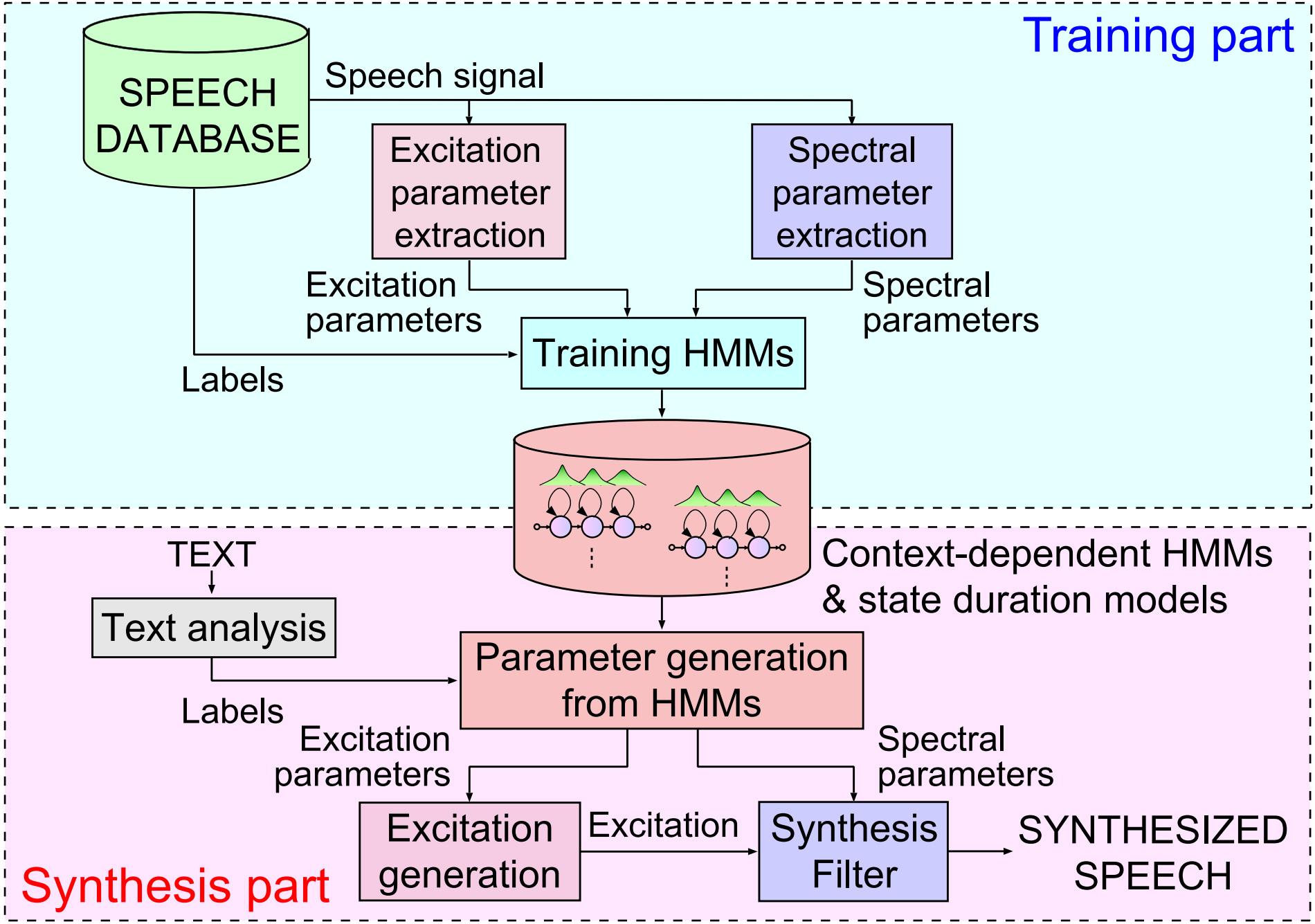
Mel-generalized cepstral analysis: mgcep

LPC ⇔ LSP conversion: lpc2lsp, lsp2lpc

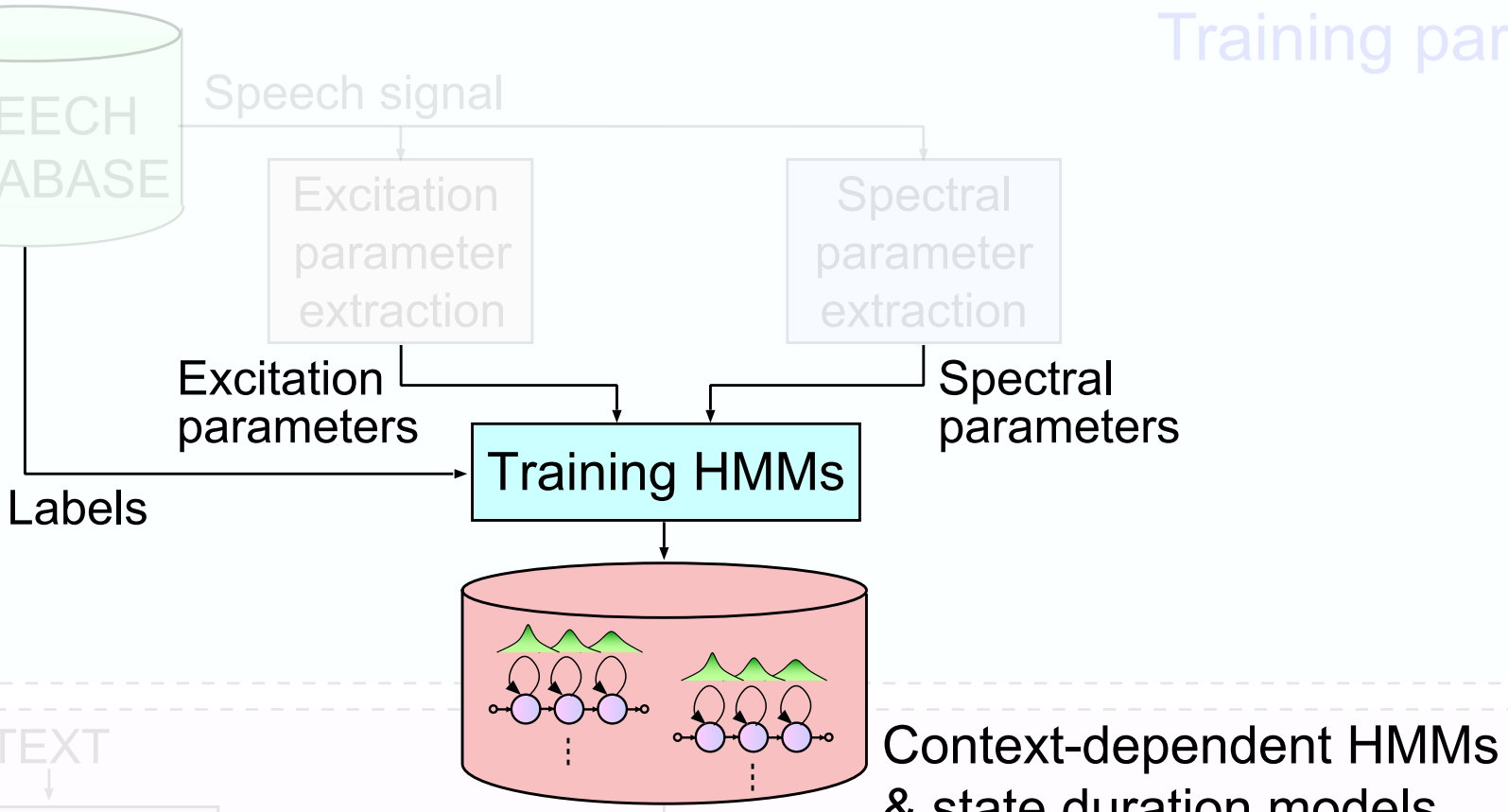
F0 detection: pitch

HTS-demo uses Snack get_f0 because of its better accuracy

HMM-based speech synthesis system (HTS)

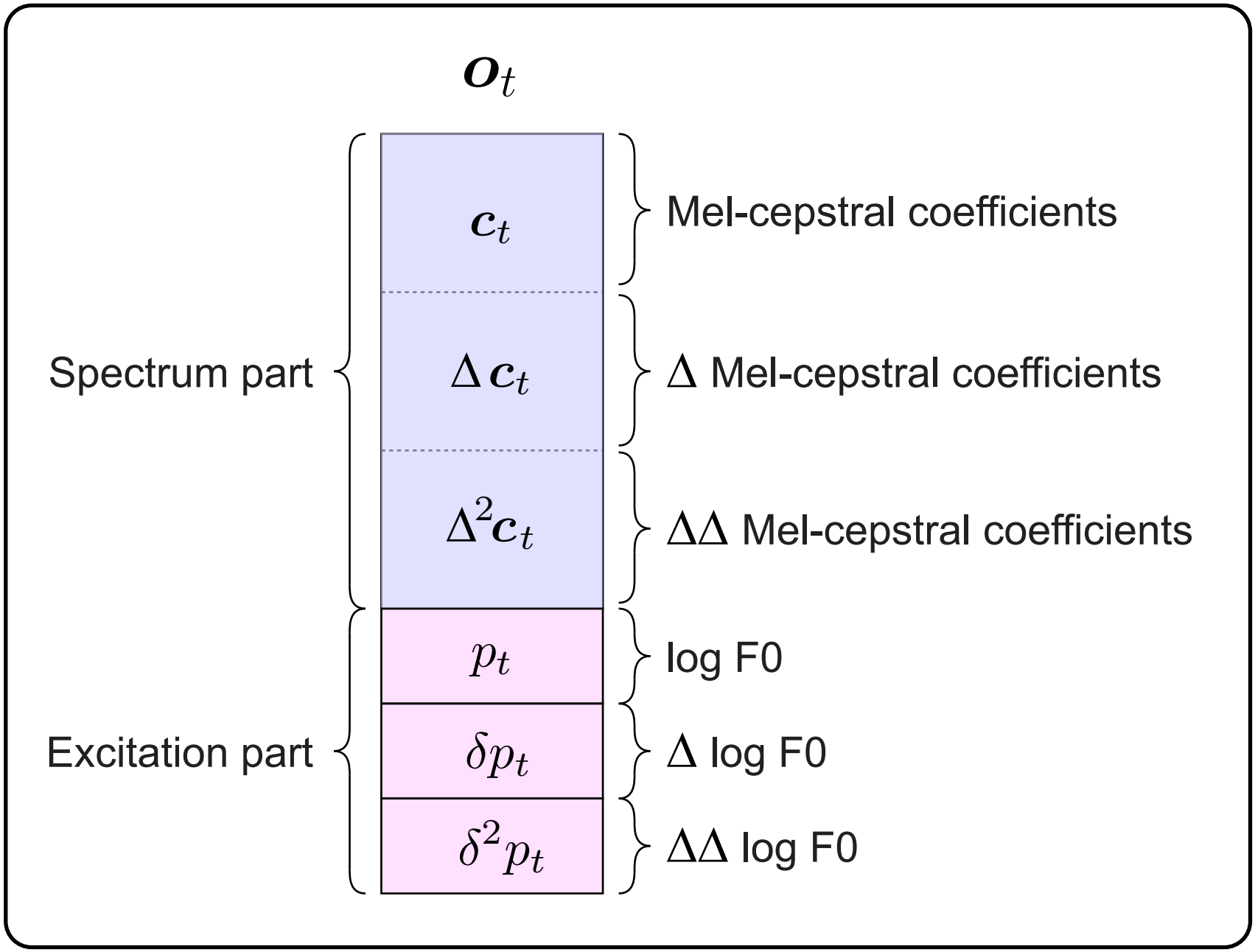


HMM-based speech synthesis system (HTS)



$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} p(\hat{\boldsymbol{O}} \mid \mathcal{W}, \boldsymbol{\lambda})$$

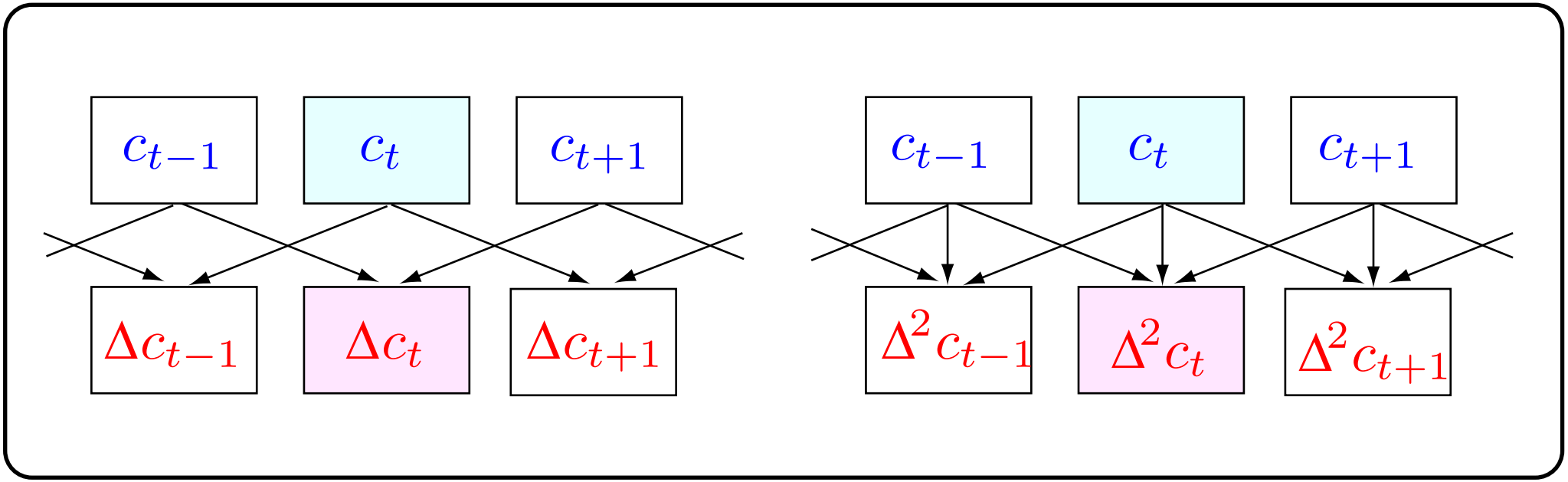
Structure of state-output (observation) vector



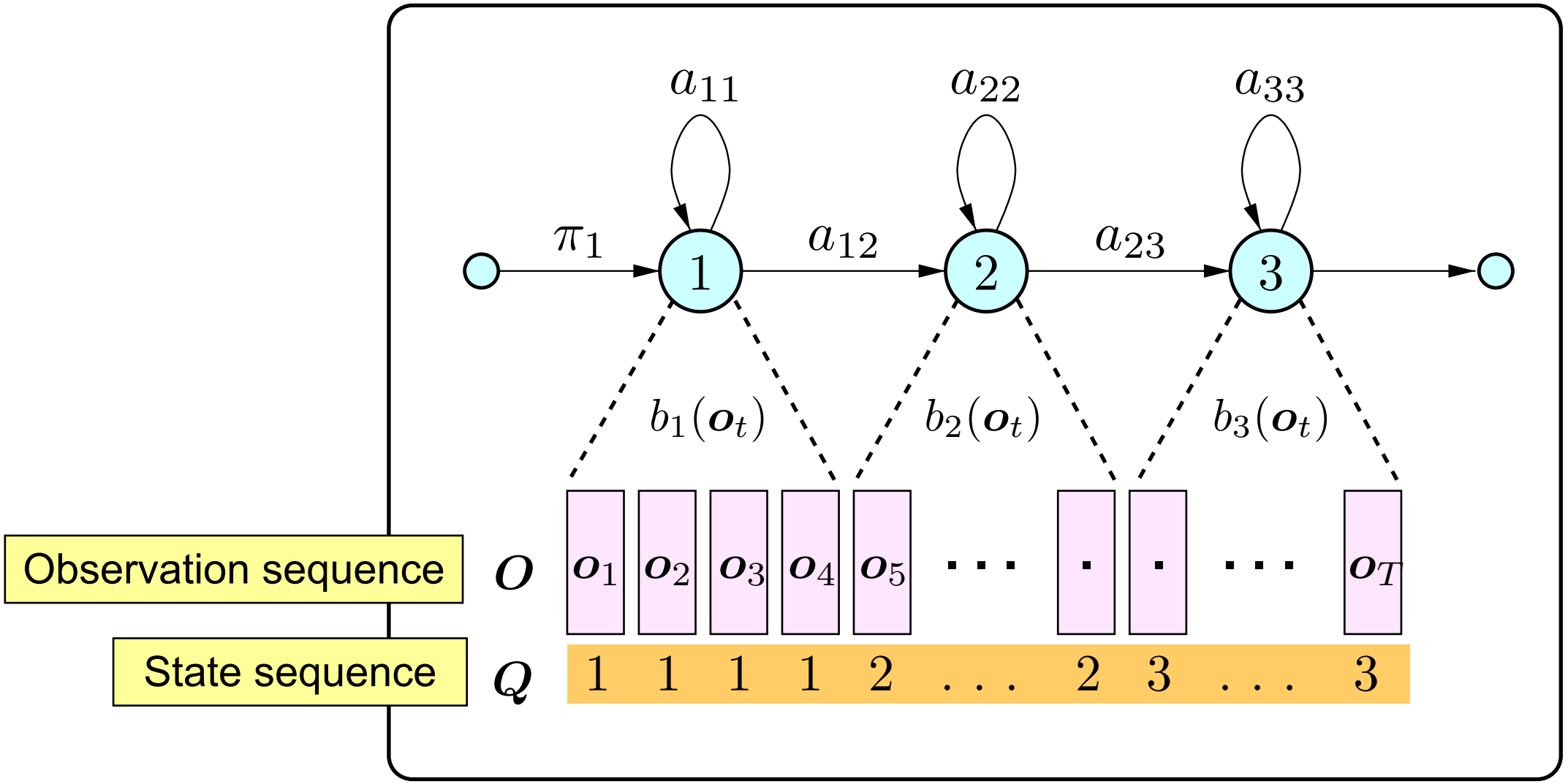
Dynamic features

$$\Delta c_t = \frac{\partial c_t}{\partial t} \approx 0.5(c_{t+1} - c_{t-1})$$

$$\Delta^2 c_t = \frac{\partial^2 c_t}{\partial t^2} \approx c_{t+1} - 2c_t + c_{t-1}$$

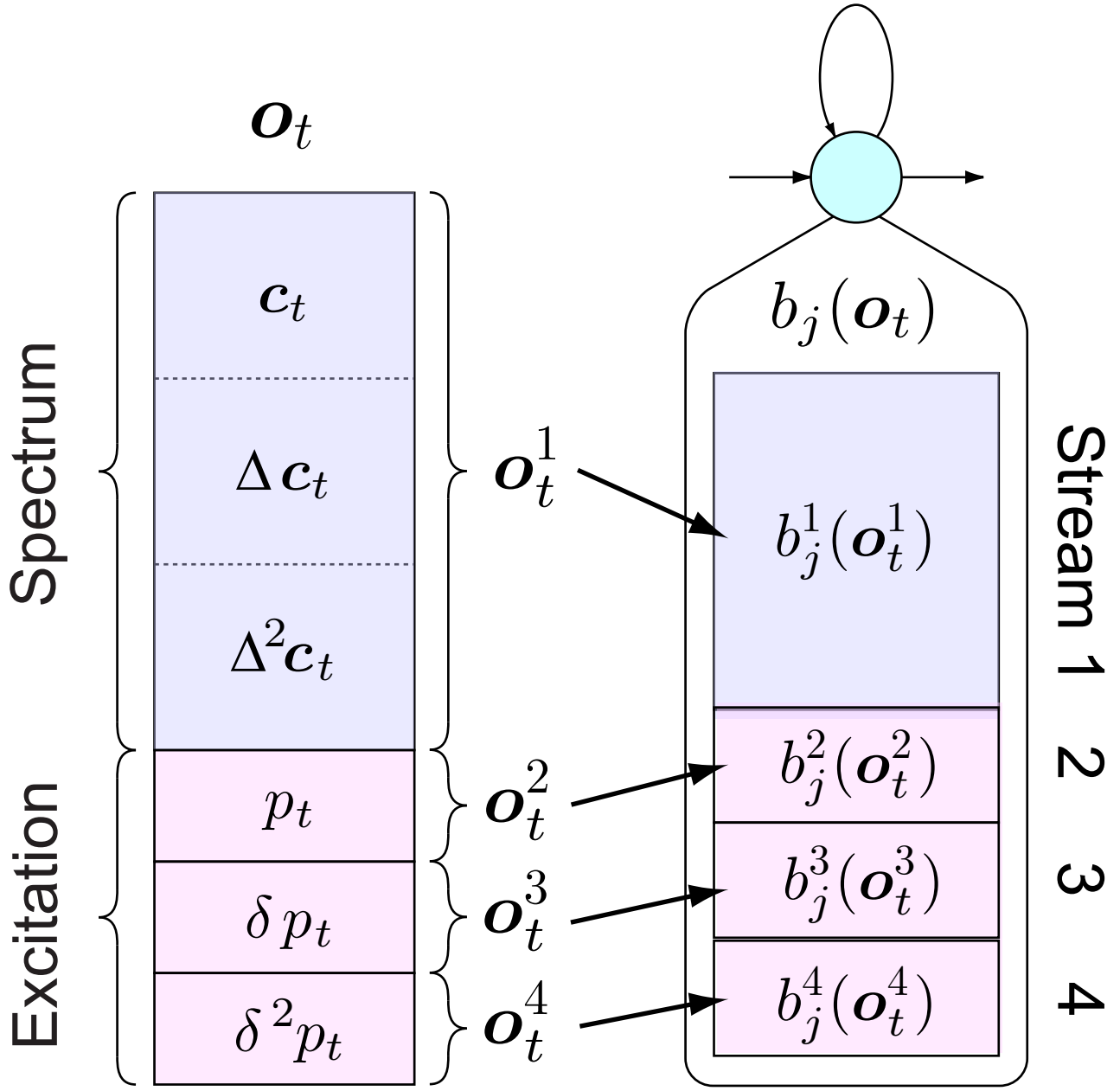


Hidden Markov model (HMM)

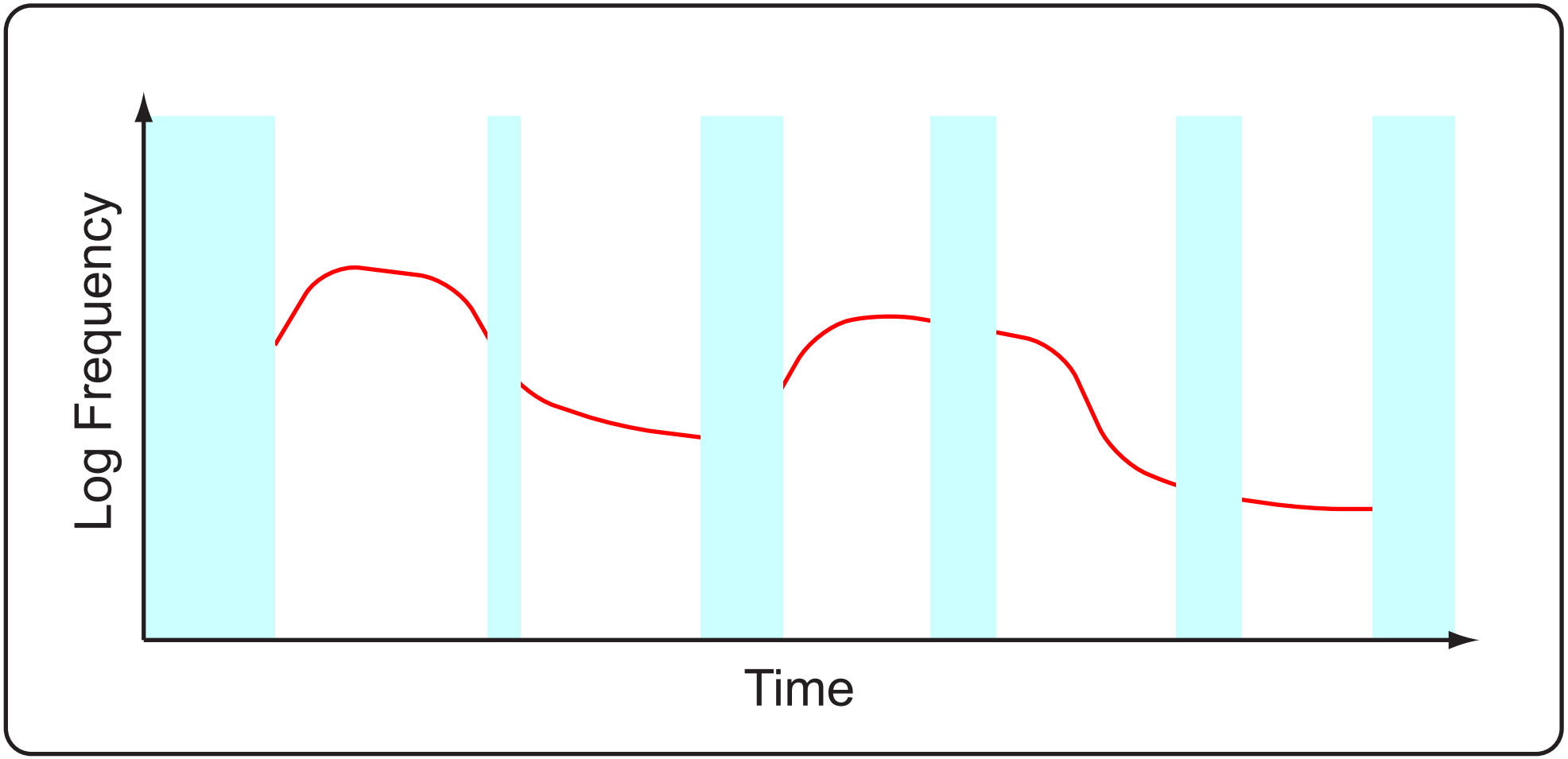


Multi-stream HMM structure

$$b_j(\boldsymbol{o}_t) = \prod_{s=1}^{S} \left(b_j^s(\boldsymbol{o}_t^s)\right)^{w_s}$$

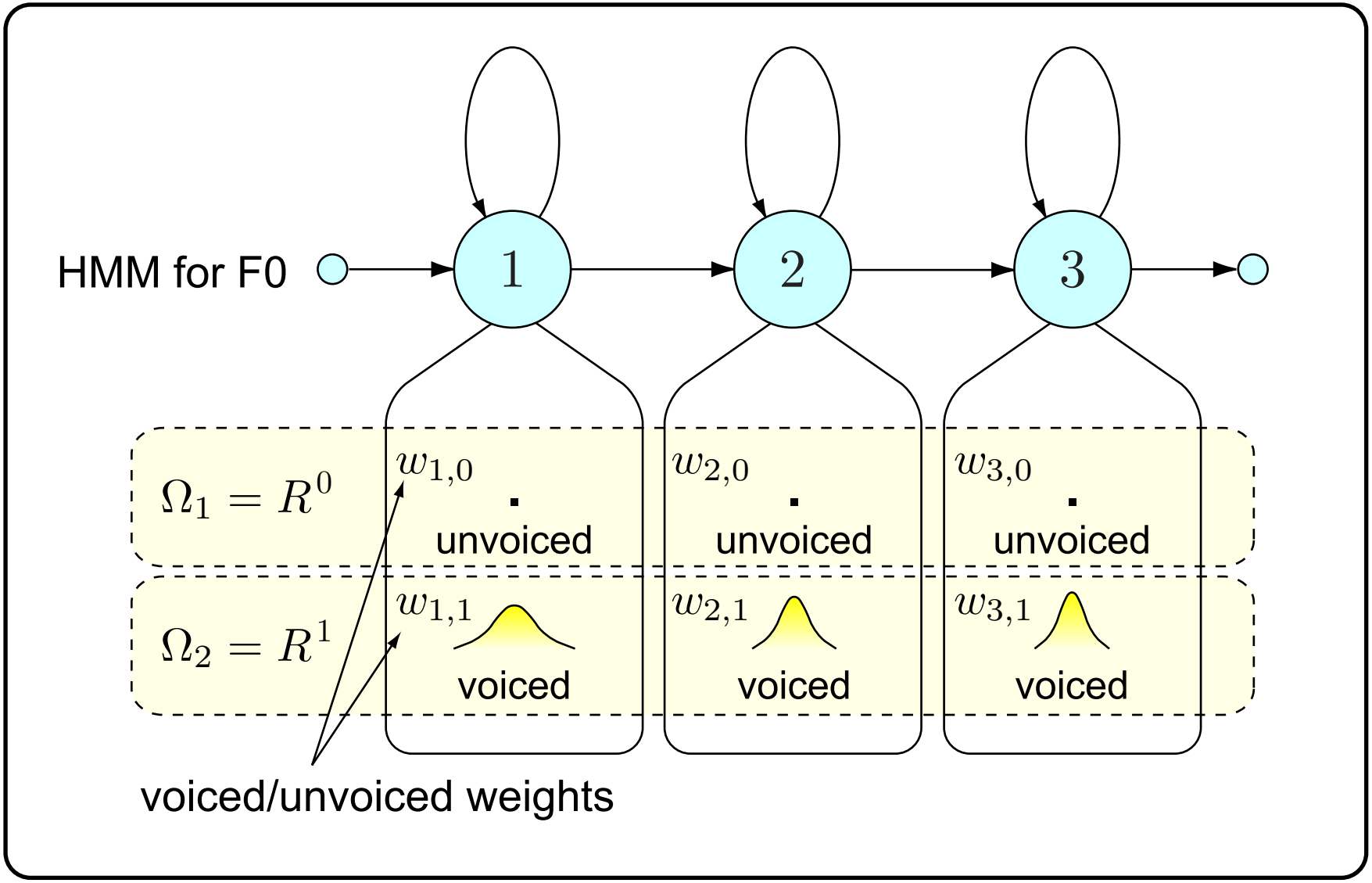


Observation of F0

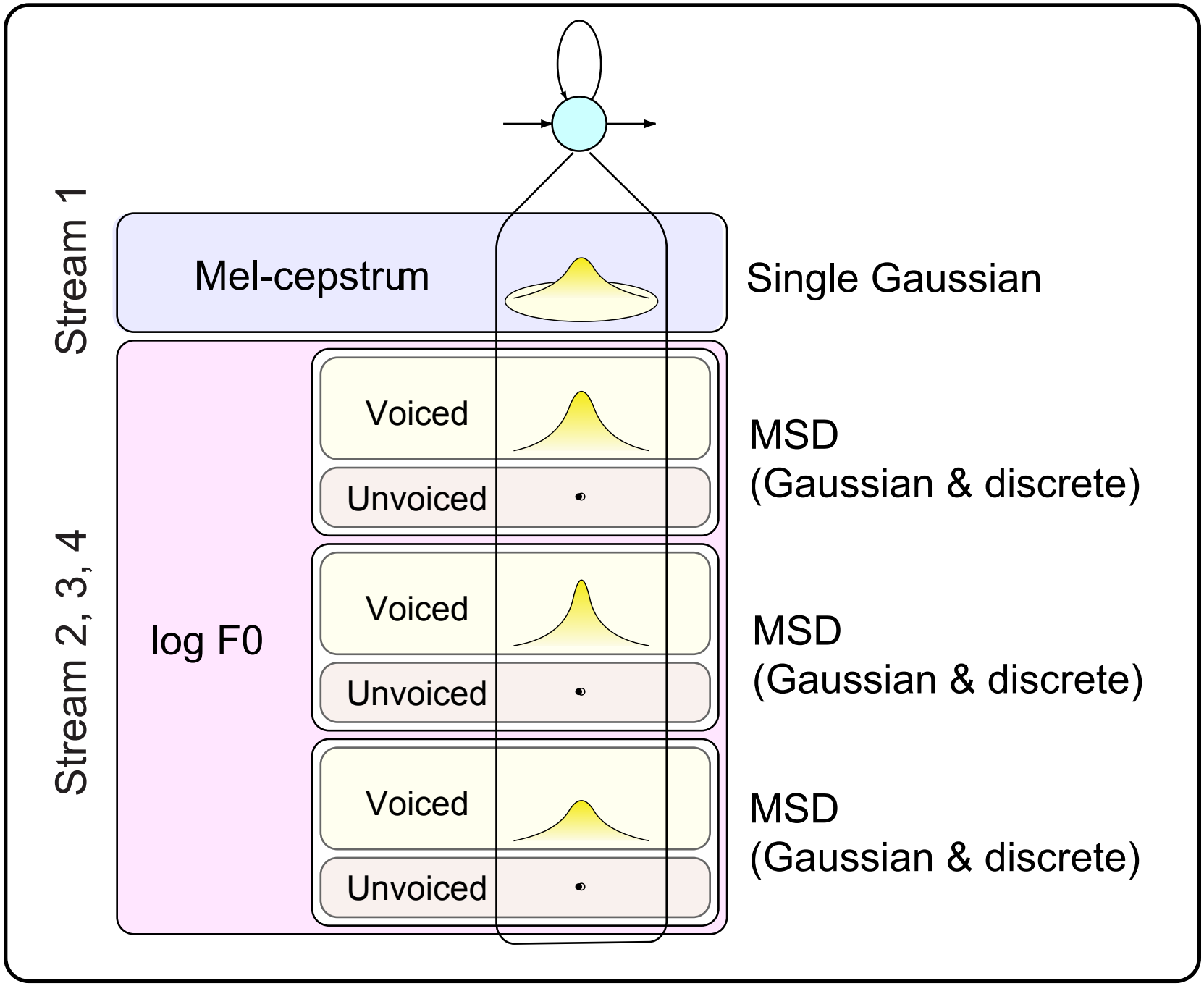


Unable to model by continuous or discrete distribution

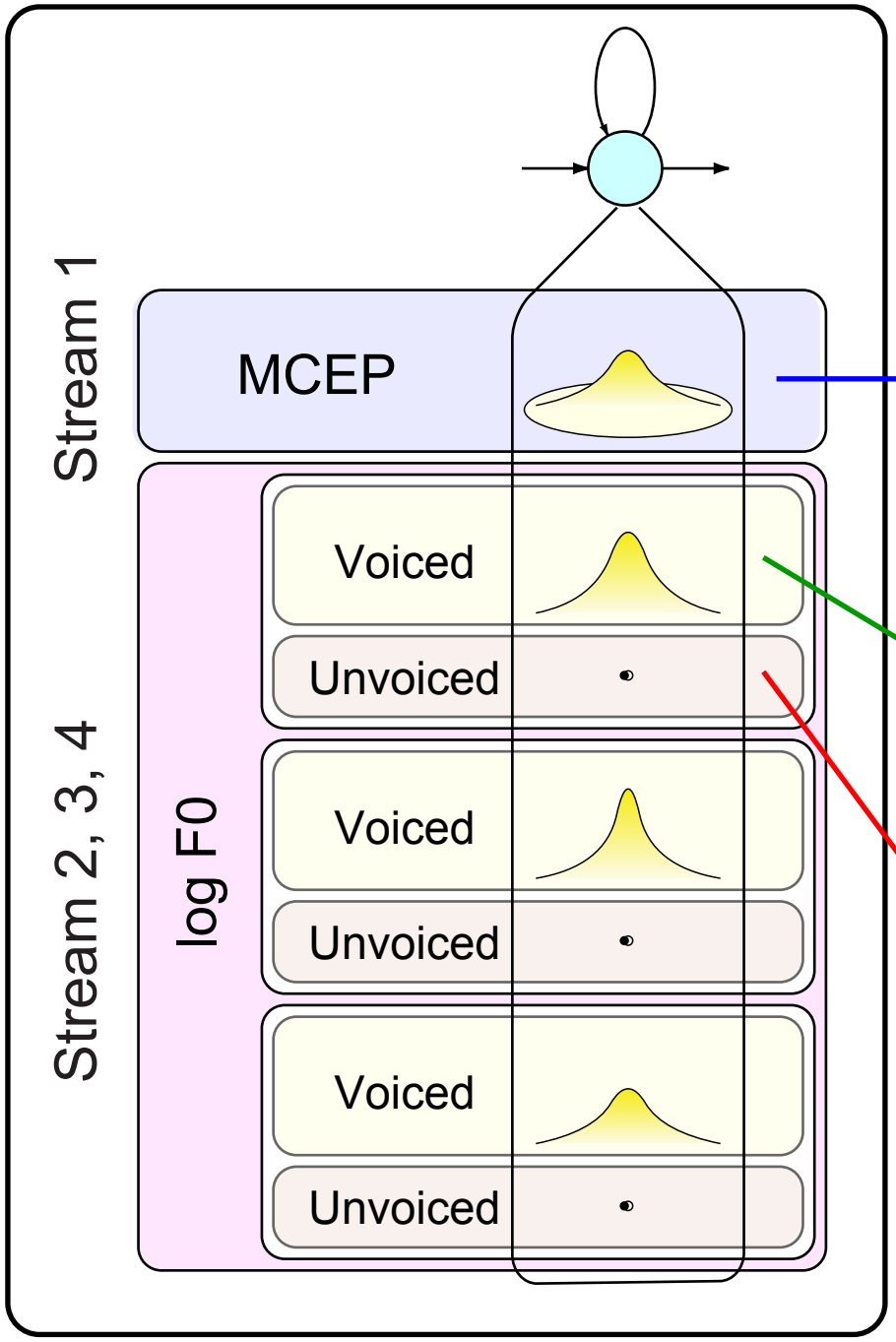
MSD-HMM for F0 modeling



Structure of state-output distributions

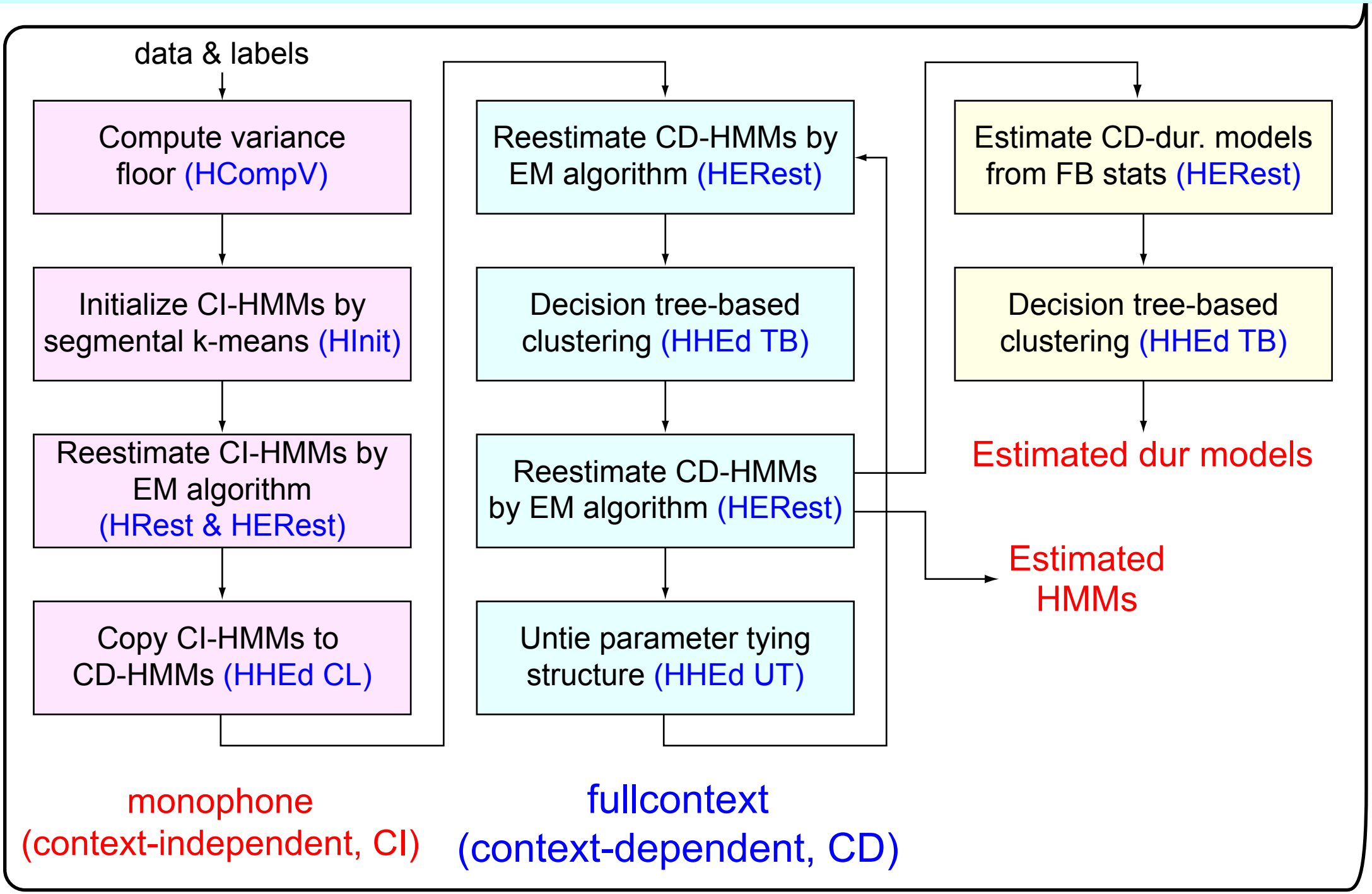


Proto-type definition (from HTS-demo)

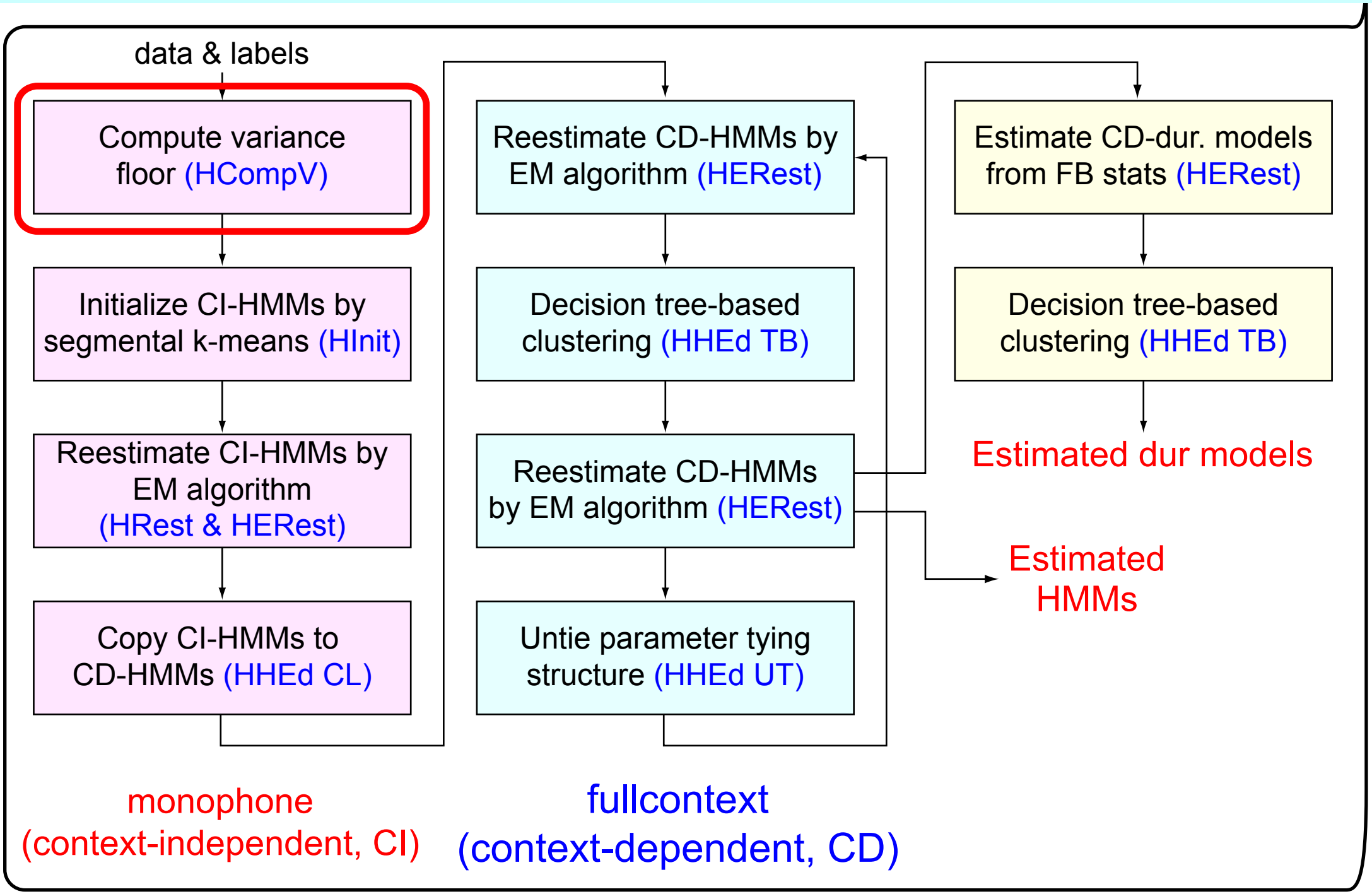



```
~o <VecSize> 78 <USER> <DIAGC>
<StreamInfo> 4 75 1 1 1 <MSDInfo> 4 0 1 1 1
<BeginHMM> <NumStates> 7 <State> 2
    <Stream> 1
     <Mean> 75
      0.0 0.0 0.0 0.0 0.0 ...
     <Variance> 75
      1.0 1.0 1.0 1.0 1.0 ...
    <Stream> 2
    <NumMixes> 2
    <Mixture> 1 0.5000
     <Mean> 1
      0.0
     <Variance> 1
      1.0
    <Mixture> 2 0.5000
     <Mean> 0
     <Variance> 0
    <Stream> 3 ...
```

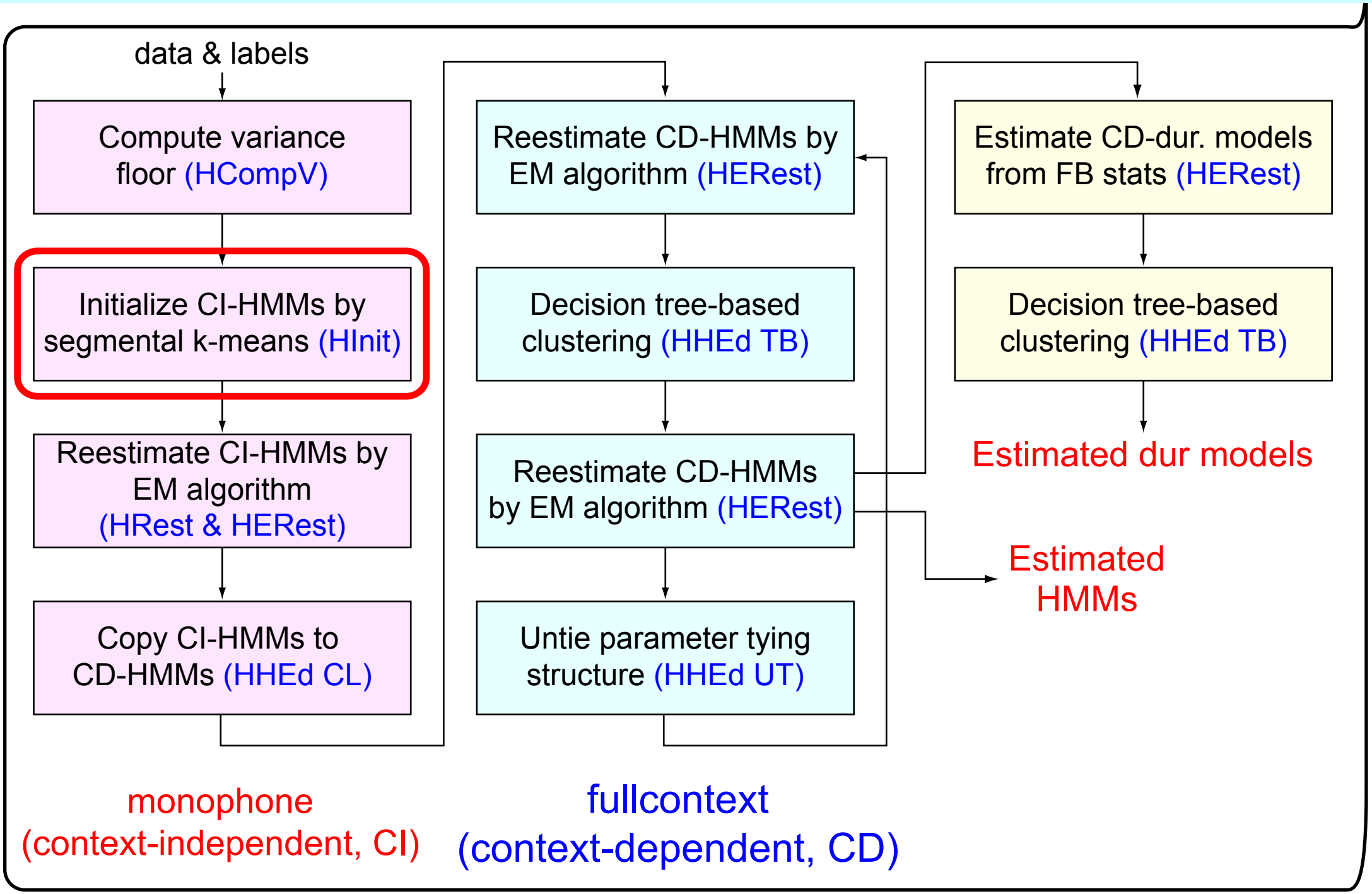
Training process (from HTS-demo)



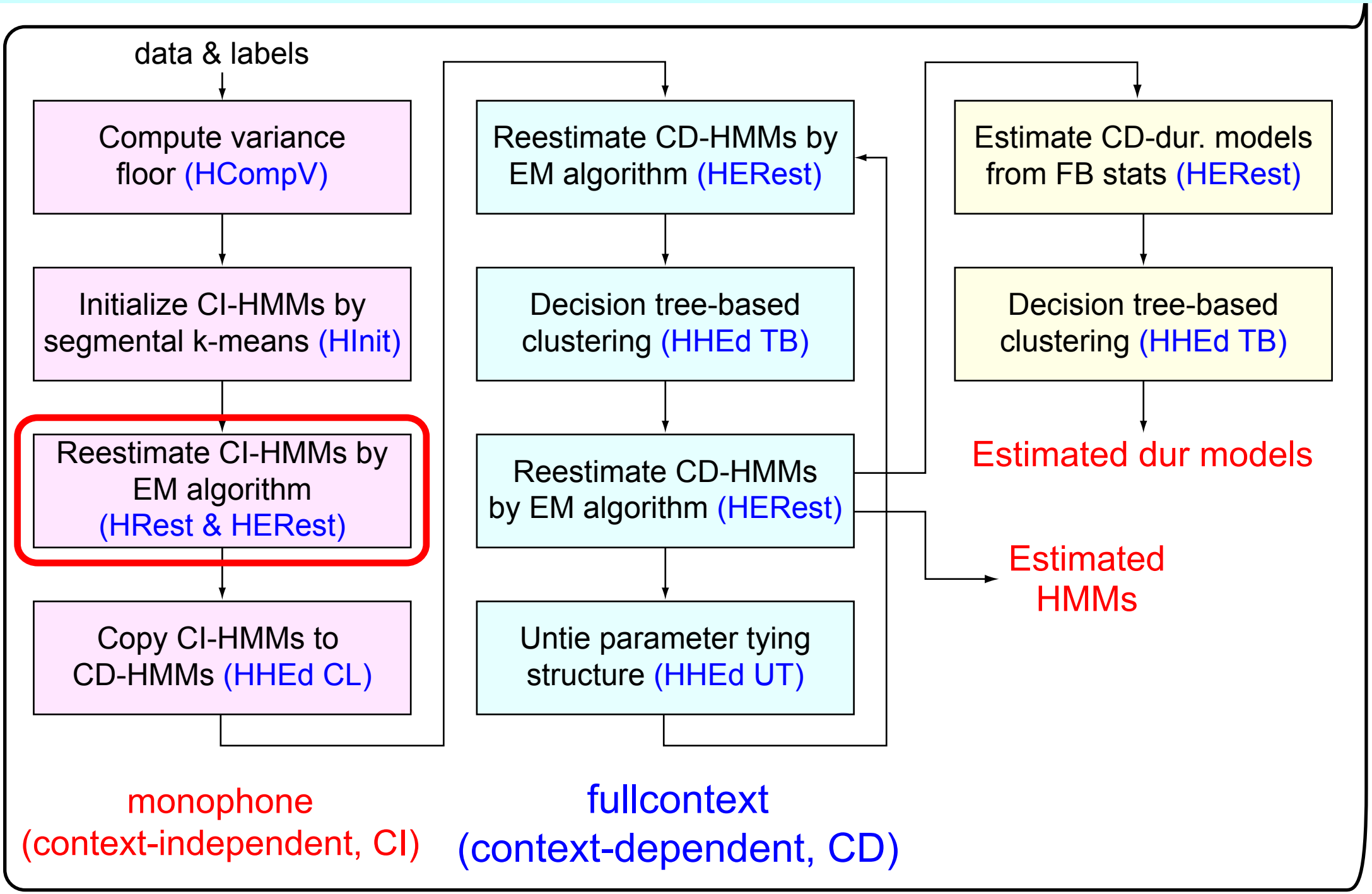
Training process (from HTS-demo)



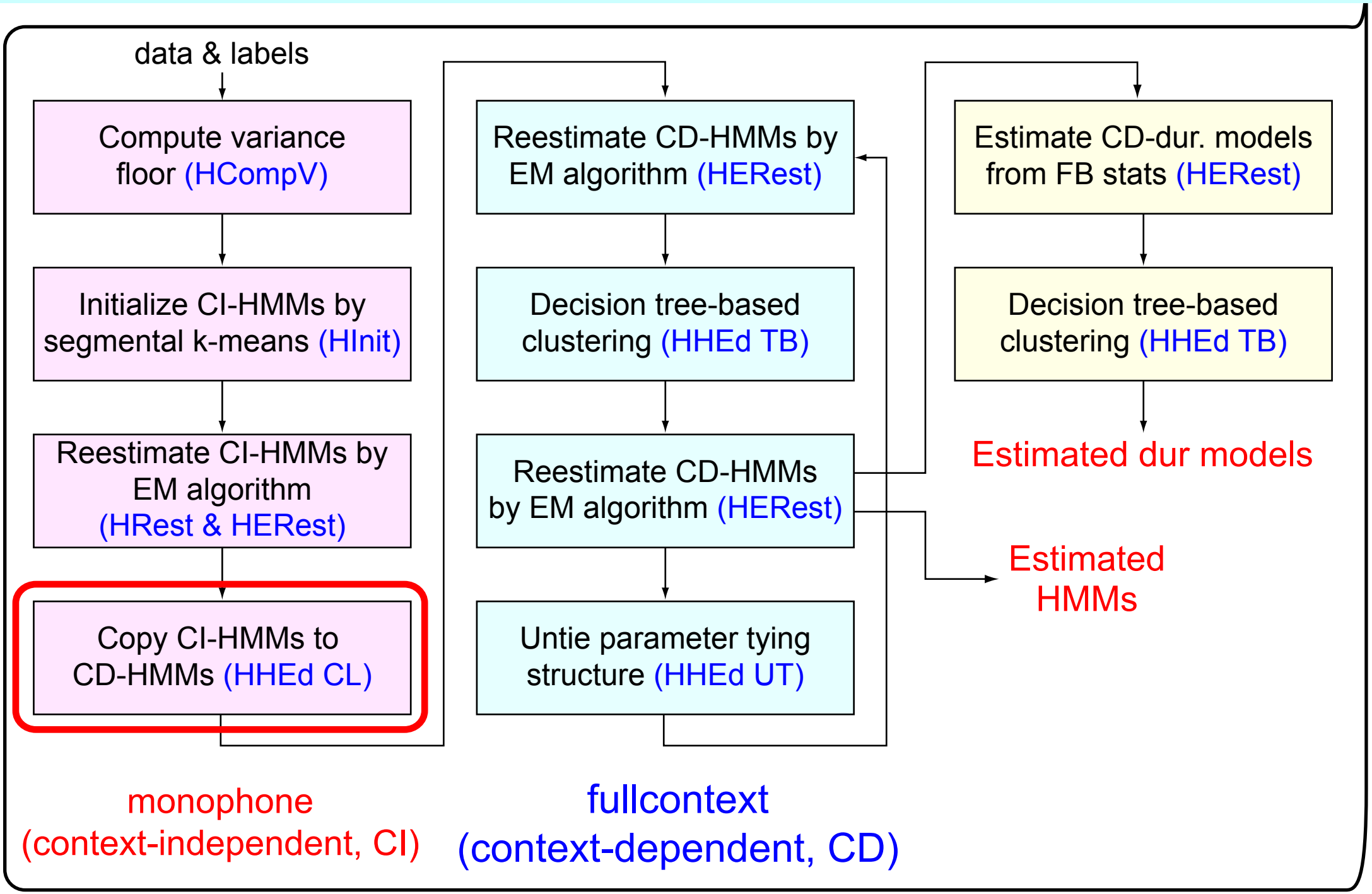
Training process (from HTS-demo)



Training process (from HTS-demo)



Training process (from HTS-demo)



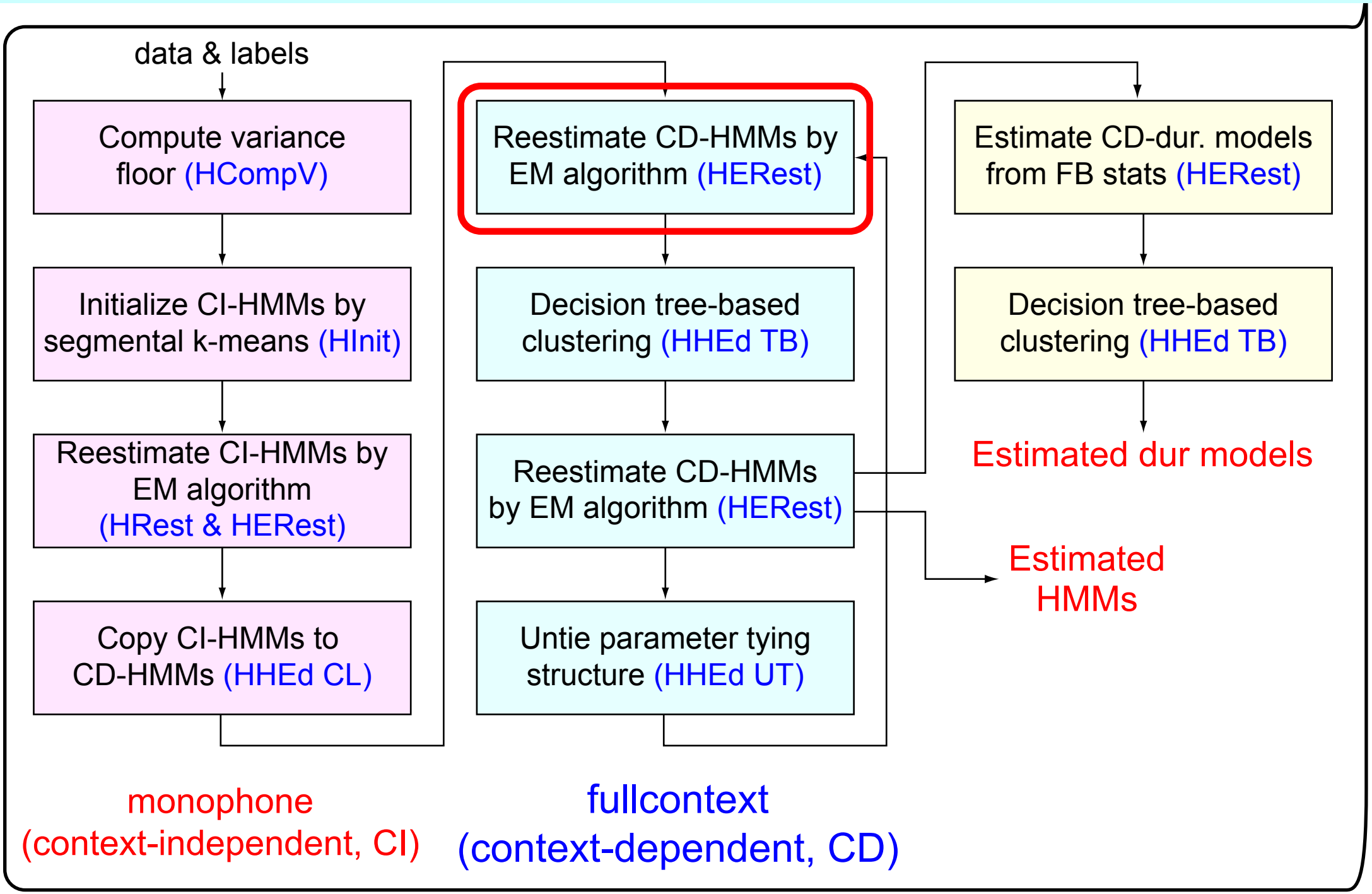
Context-dependent modeling

- Current phoneme
- {preceding, succeeding} two phonemes
- Position of current phoneme in current syllable
- # of phonemes at {preceding, current, succeeding} syllable
- {accent, stress} of {preceding, current, succeeding} syllable
- Position of current syllable in current word
- # of {preceding, succeeding} {stressed, accented} syllables in current phrase
- # of syllables {from previous, to next} {stressed, accented} syllable
- Vowel within current syllable
- Guess at part of speech of {preceding, current, succeeding} word
- # of syllables in {preceding, current, succeeding} word
- Position of current word in current phrase
- # of {preceding, succeeding} content words in current phrase
- # of words {from previous, to next} content word
- # of syllables in {preceding, current, succeeding} phrase

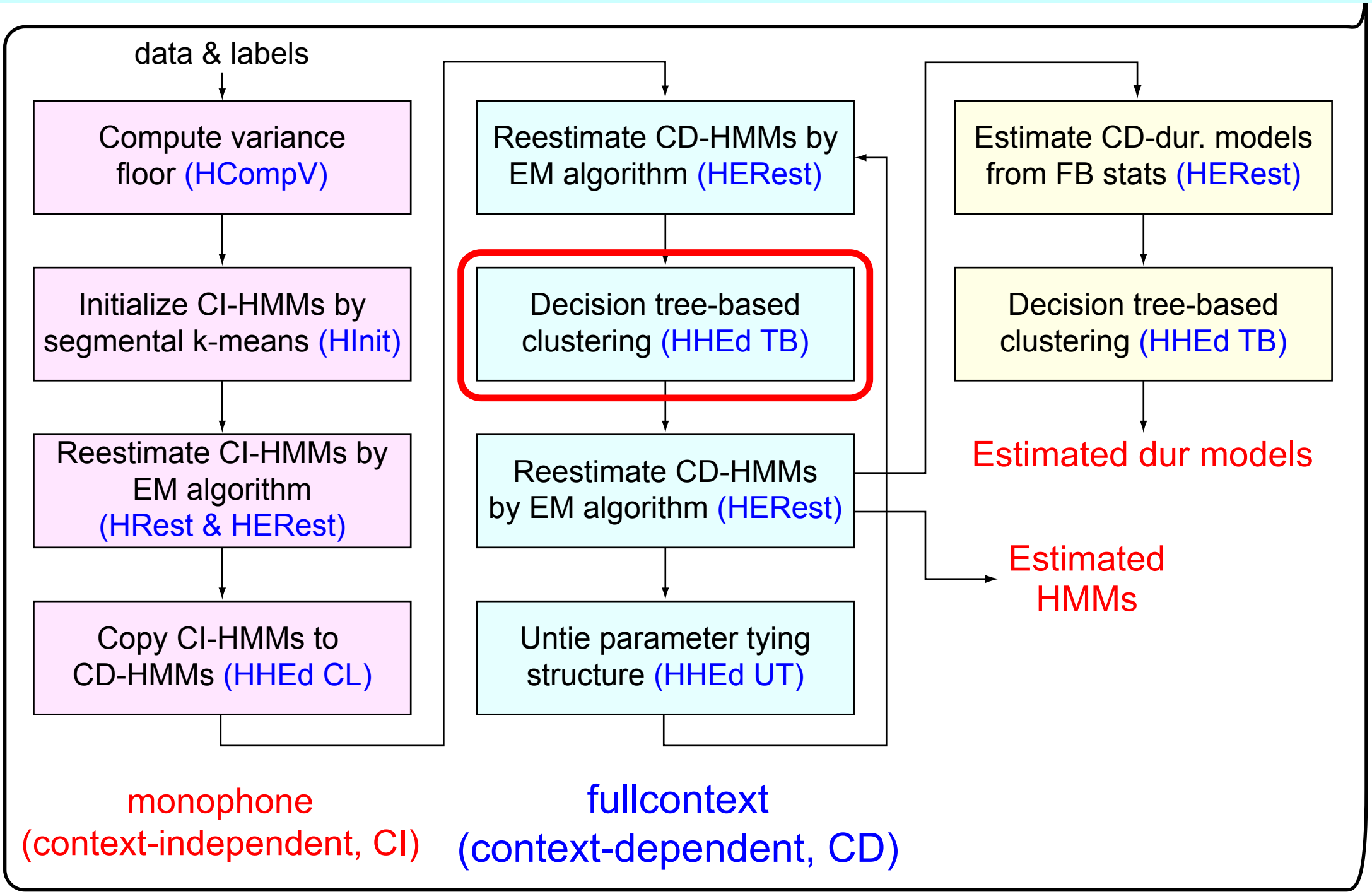
...

Vast # of combinations ⇒ Difficult to have all possible models

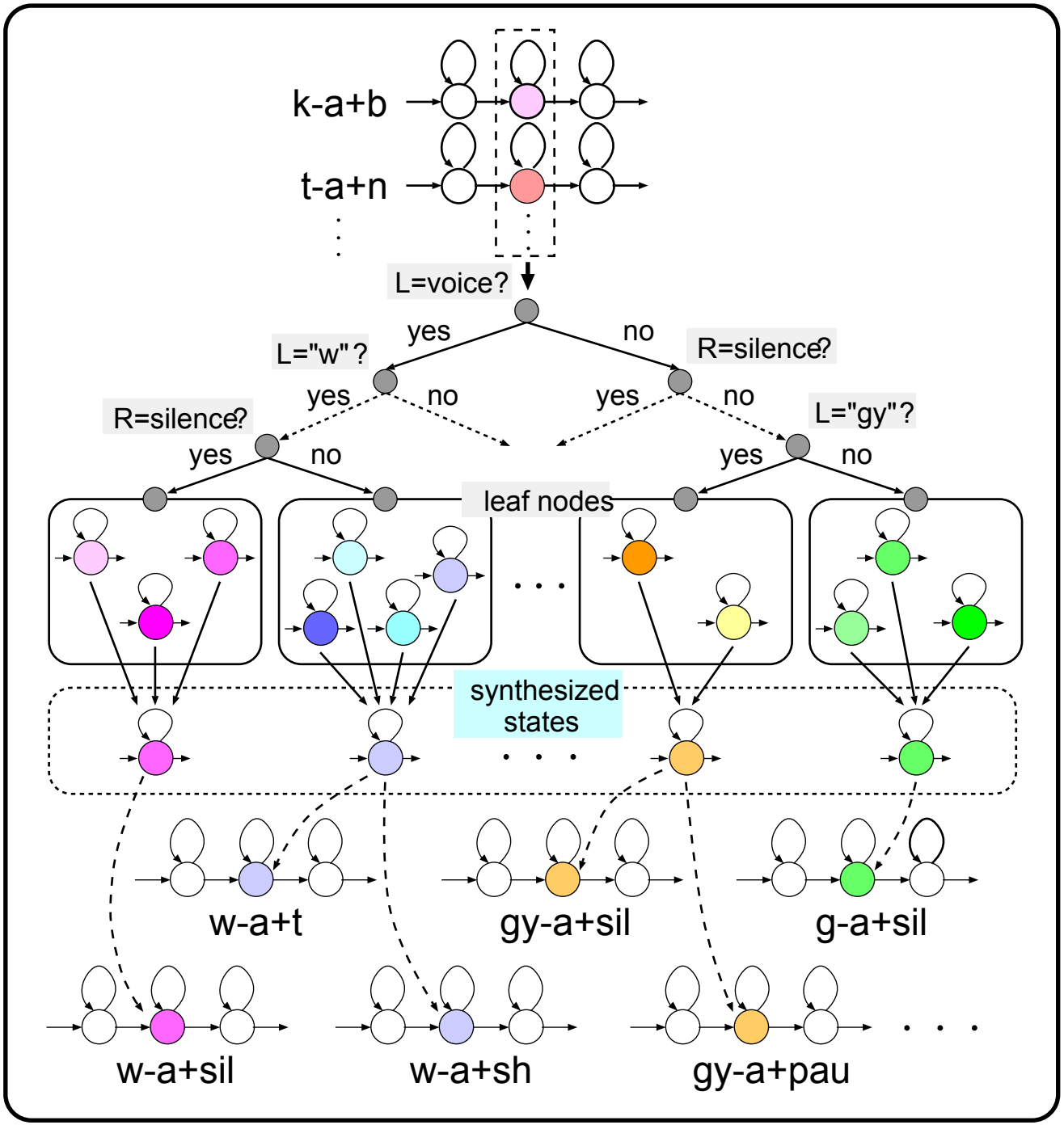
Training process (from HTS-demo)



Training process (from HTS-demo)



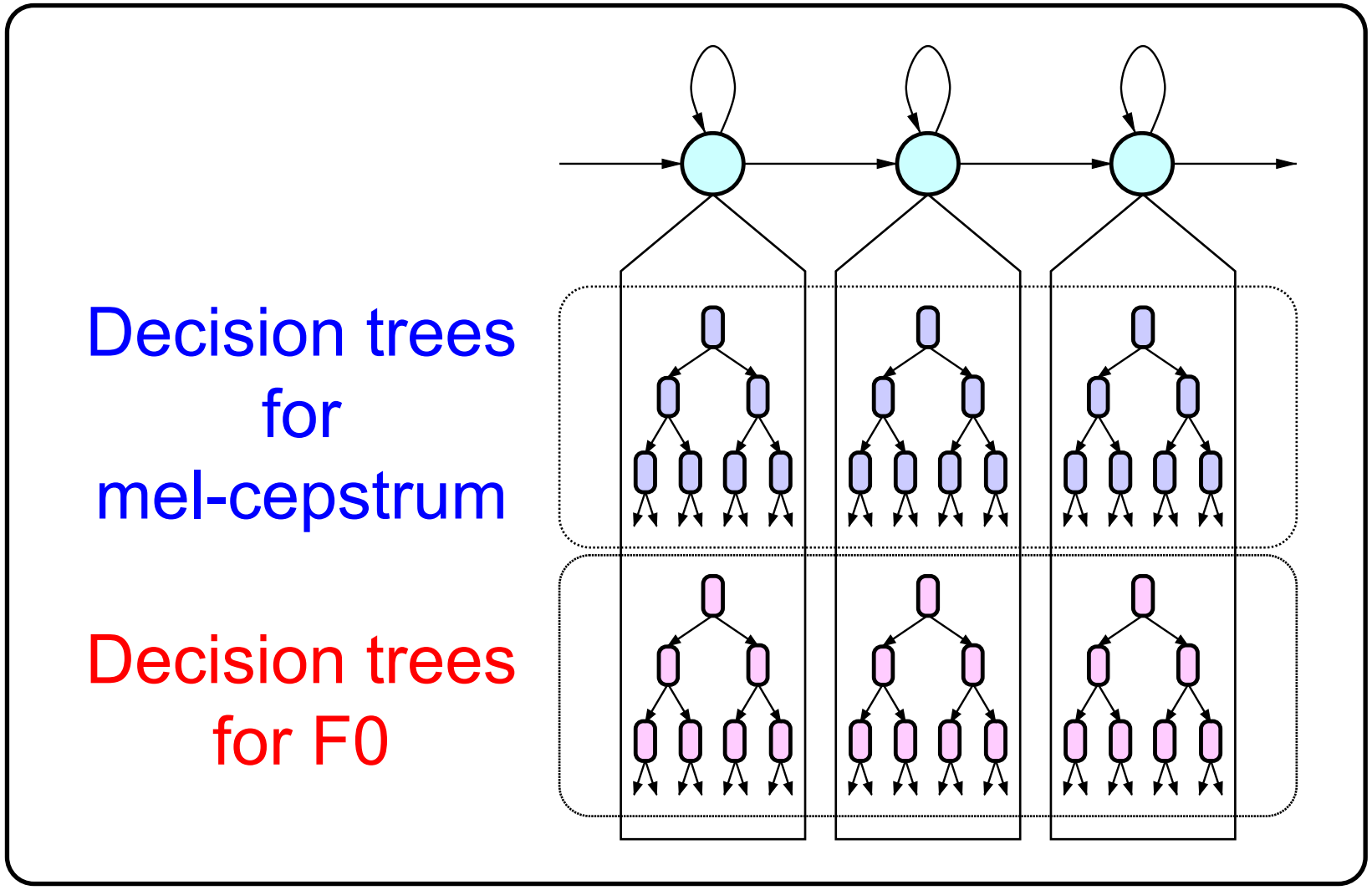
Decision tree-based state clustering [Odell;'95]



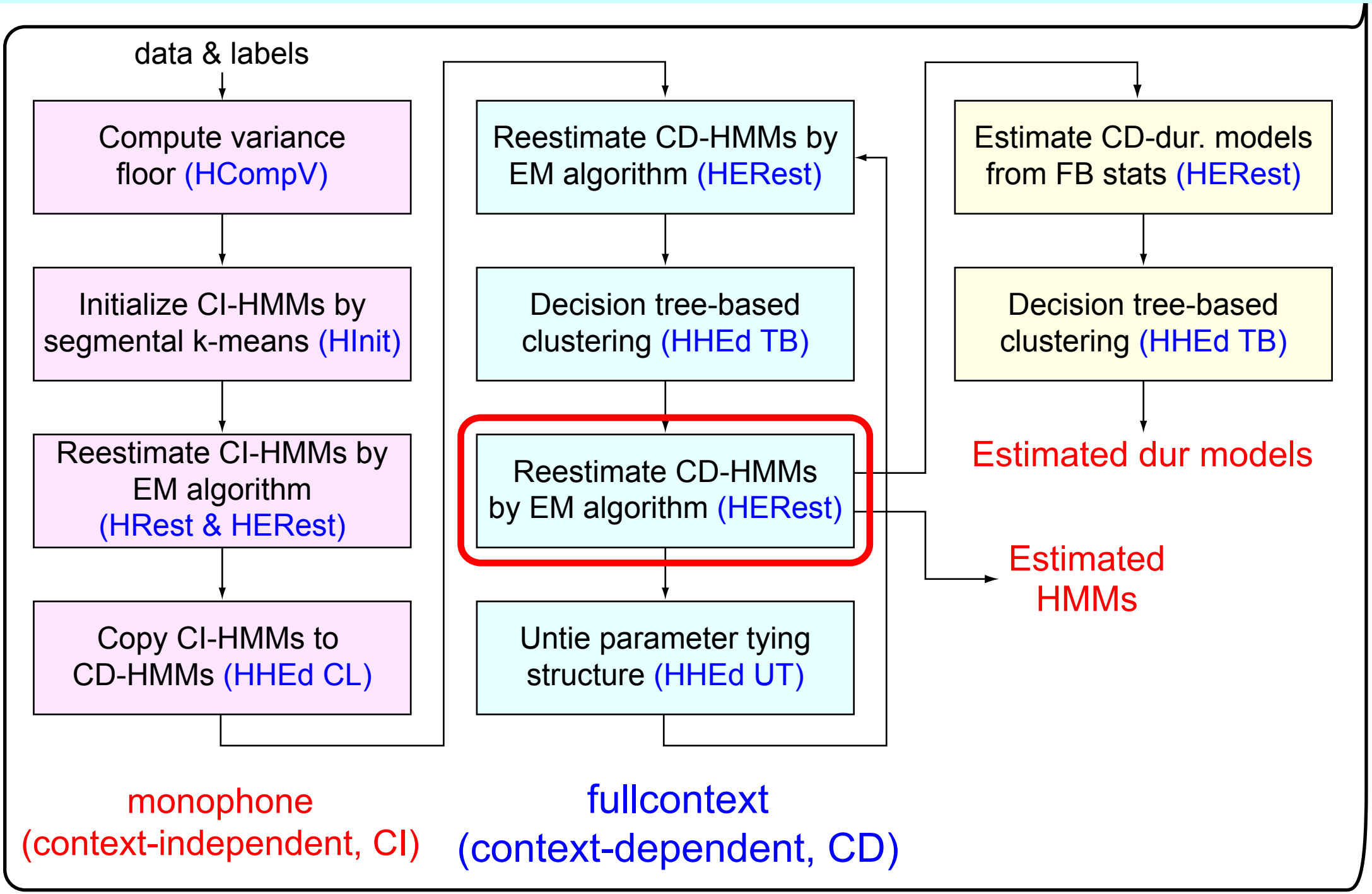
Stream-dependent tree-based clustering

Spectrum & excitation have different context dependency

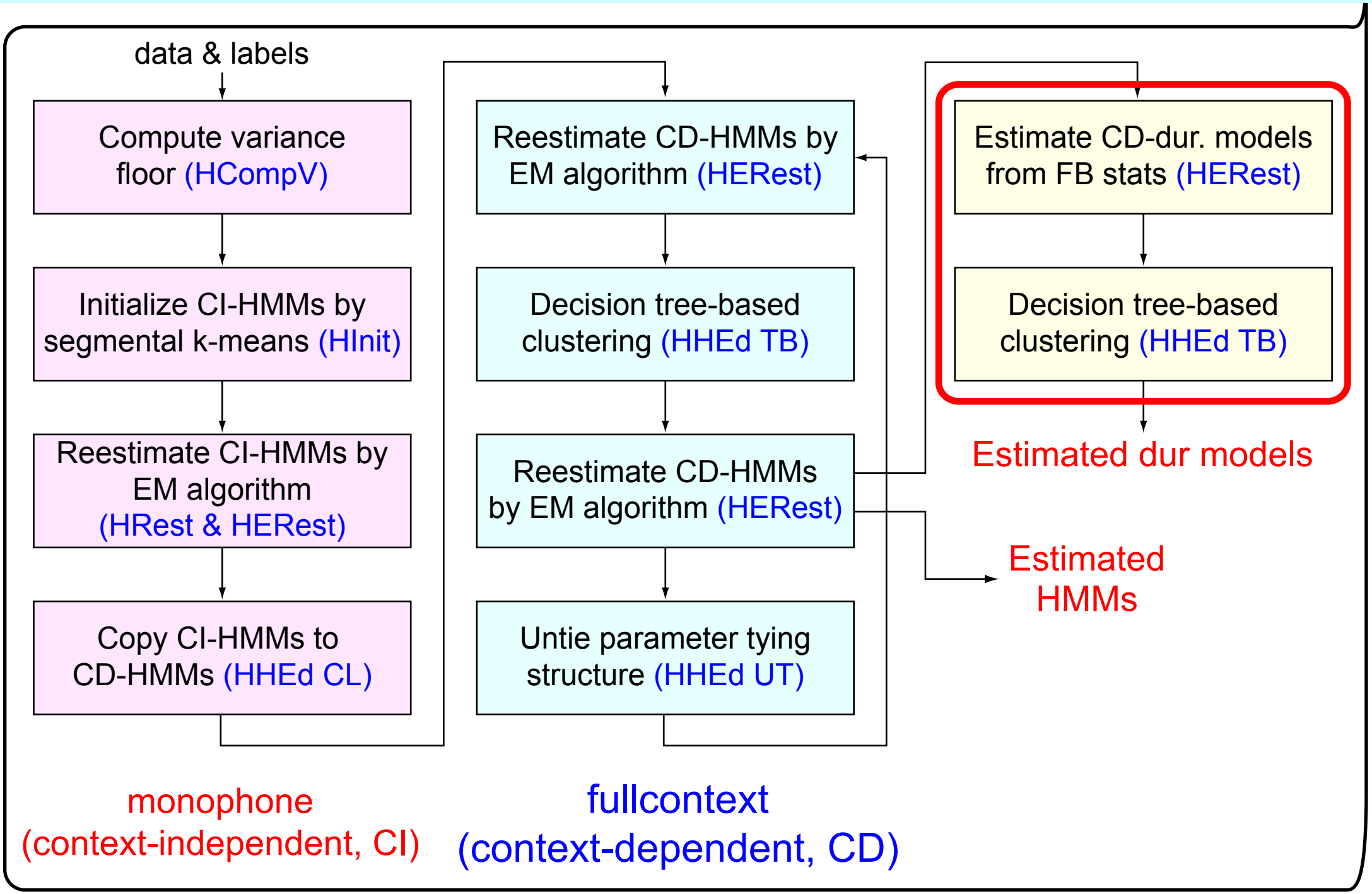
⇒ Build decision trees separately



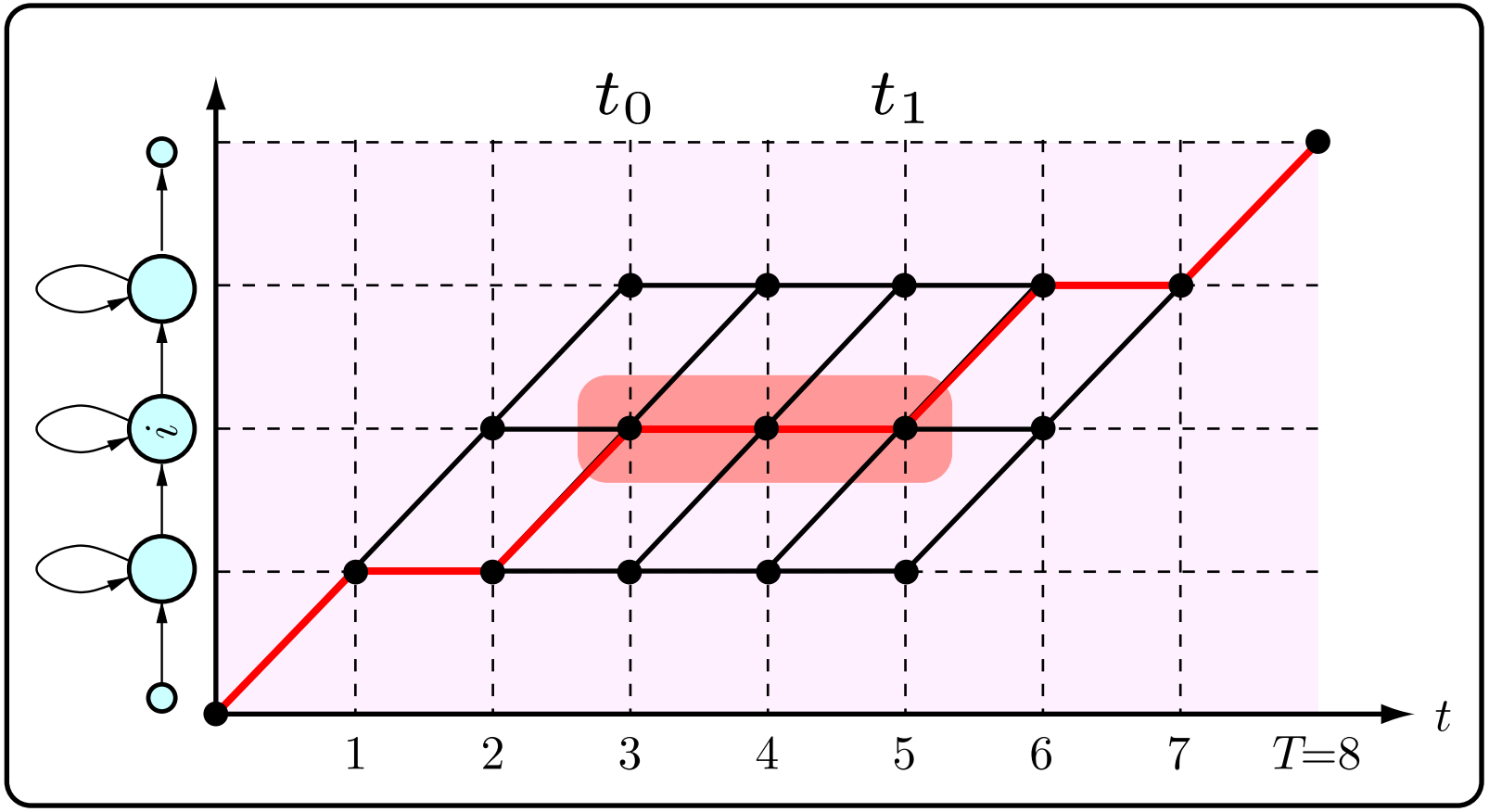
Training process (from HTS-demo)



Training process (from HTS-demo)



Estimation of state duration models [Yoshimura;'98]



$$\chi_{t_0,t_1}(i) \propto \sum_{j \neq i} \alpha_{t_0-1}(j) a_{ji} \cdot a_{ii}^{t_1-t_0} \cdot \prod_{t=t_0}^{t_1} b_i(\boldsymbol{o}_t)$$

$$\cdot \sum_{k \neq i} a_{ik} b_k(\boldsymbol{o}_{t_1+1}) \beta_{t_1+1}(k)$$

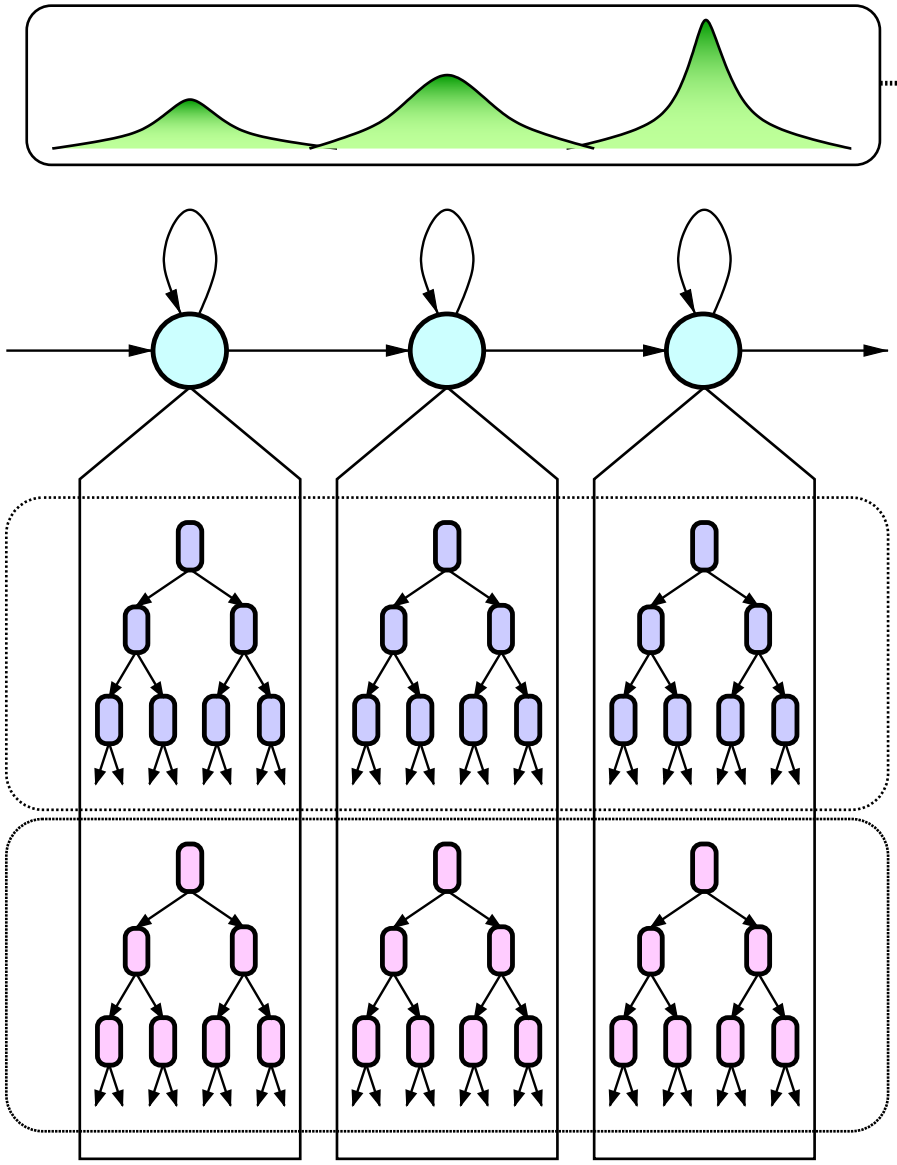
Stream-dependent tree-based clustering

State duration model

HMM

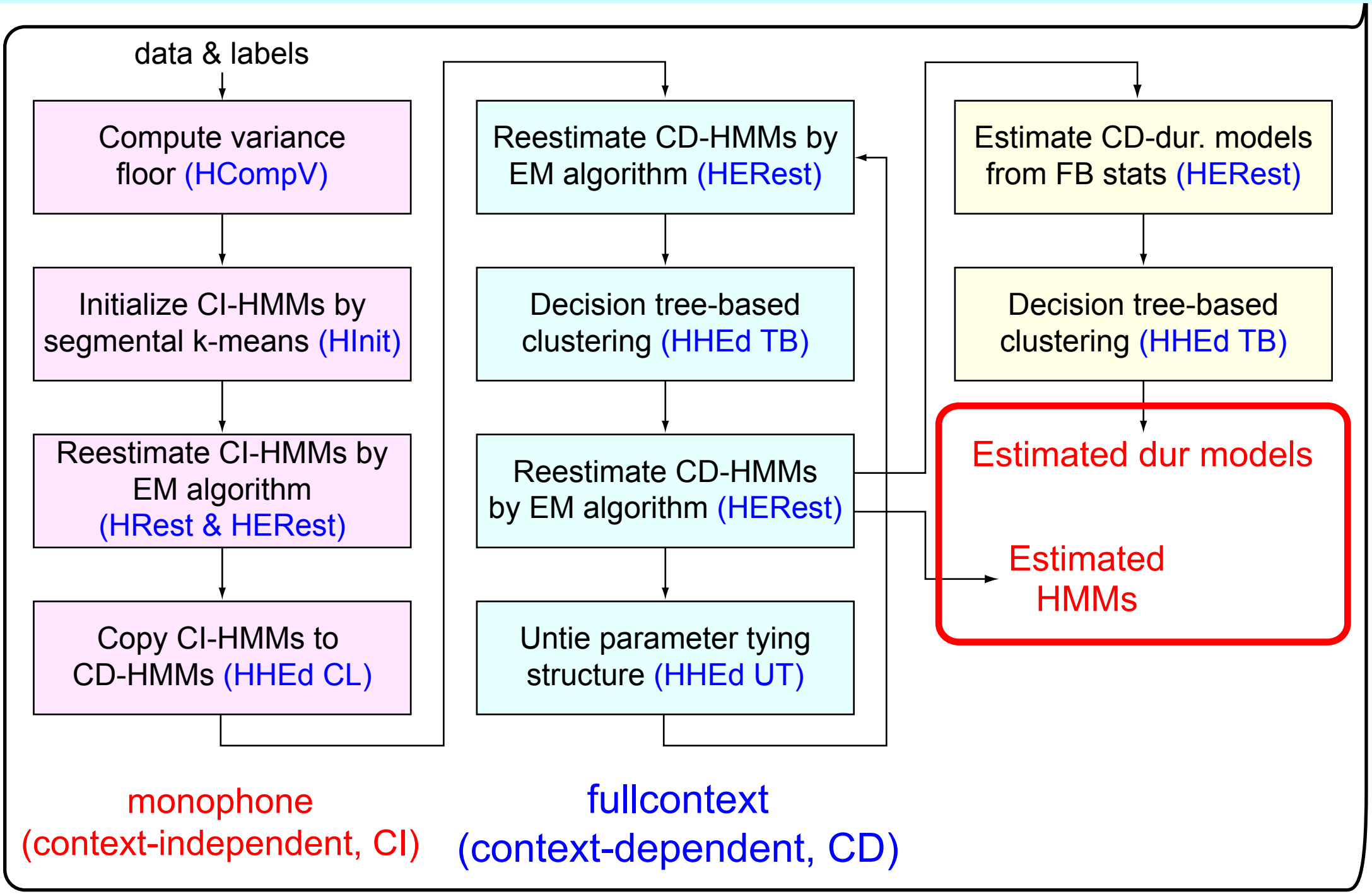
Decision trees for mel-cepstrum

Decision trees for F0

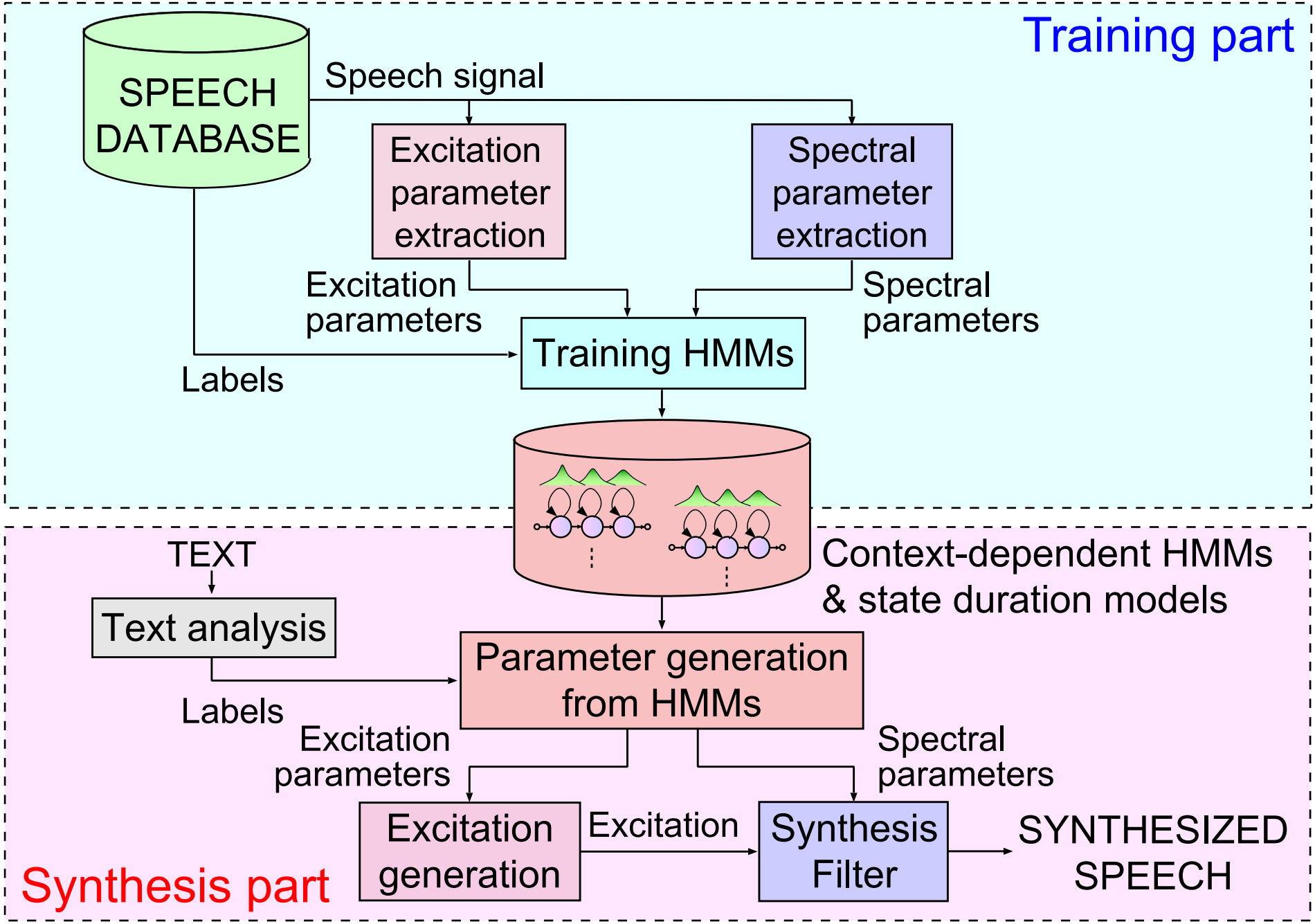


Decision tree for state dur. models

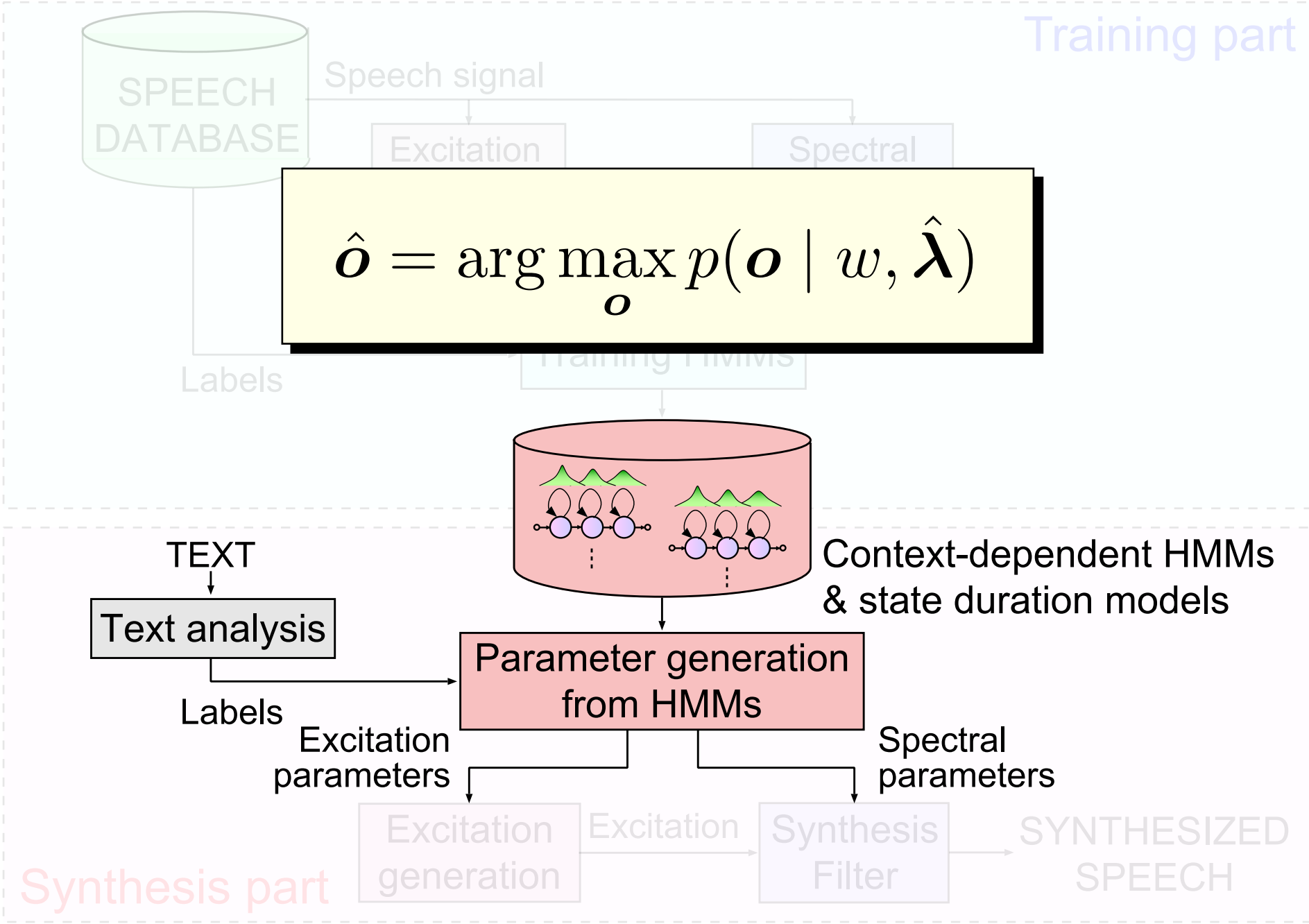
Training process (from HTS-demo)



HMM-based speech synthesis system (HTS)



HMM-based speech synthesis system (HTS)



Speech parameter generation algorithm [Tokuda;'00]

$$
\begin{aligned}
\hat{\boldsymbol{o}} &= \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid w, \hat{\boldsymbol{\lambda}}) \\
&= \arg\max_{\boldsymbol{o}} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}}) \\
&\approx \arg\max_{\boldsymbol{o}} \max_{\boldsymbol{q}} p(\boldsymbol{o} \mid \boldsymbol{q}, \hat{\boldsymbol{\lambda}})\, p(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}})
\end{aligned}
$$

$$
\begin{aligned}
\hat{\boldsymbol{q}} &= \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}}) \\
\hat{\boldsymbol{o}} &= \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \hat{\boldsymbol{\lambda}})
\end{aligned}
$$

Speech parameter generation algorithm [Tokuda;'00]

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid w, \hat{\boldsymbol{\lambda}})$$

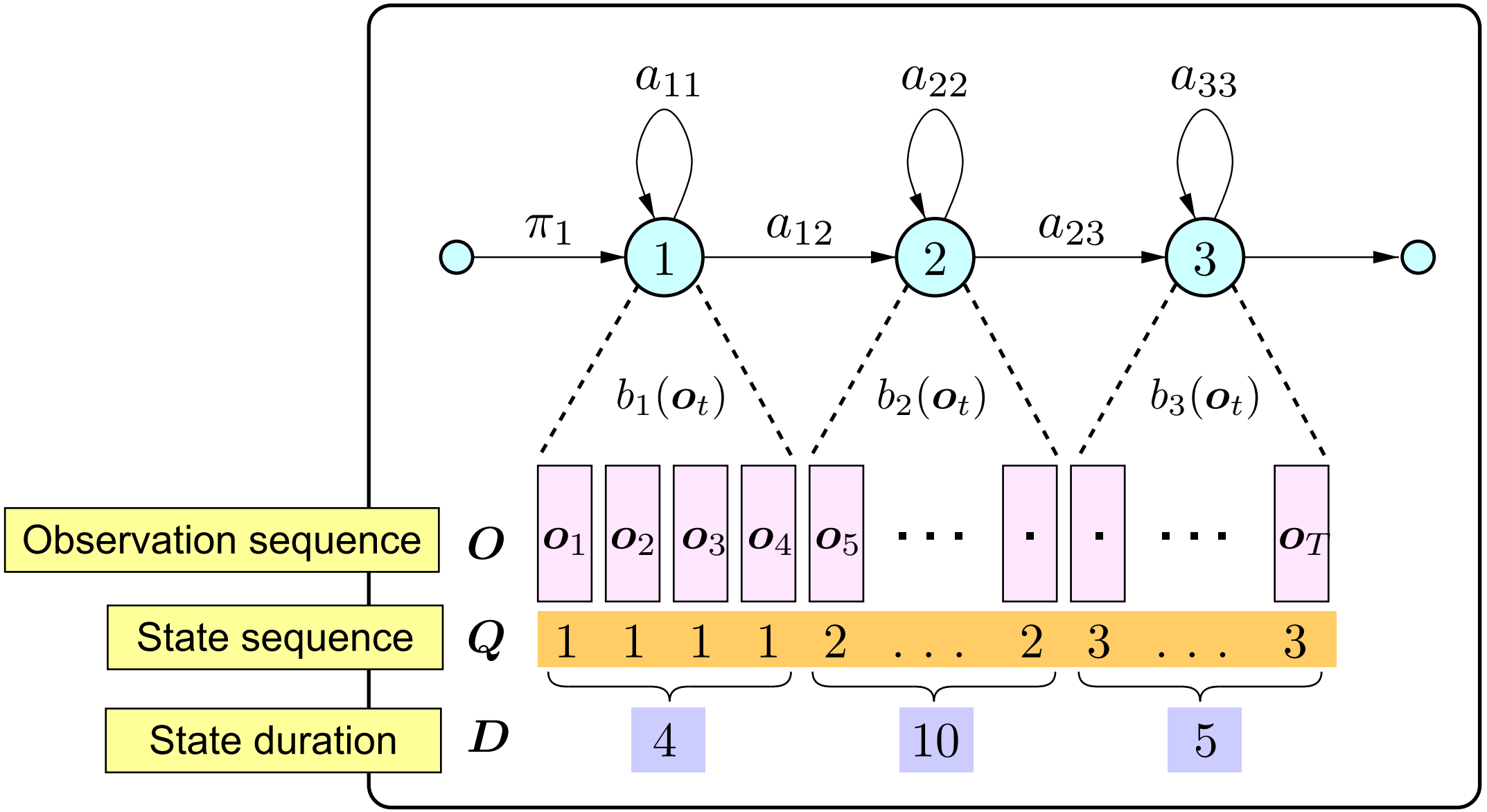
$$= \arg\max_{\boldsymbol{o}} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}})$$

$$\approx \arg\max_{\boldsymbol{o}} \max_{\boldsymbol{q}} p(\boldsymbol{o} \mid \boldsymbol{q}, \hat{\boldsymbol{\lambda}})\, p(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}})$$

$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}})$$

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \hat{\boldsymbol{\lambda}})$$

Determination of state sequence (1)



Determine state sequence via determining state durations

Determination of state sequence (2)

$$P(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}}) = \prod_{i=1}^{K} p_i(d_i)$$

$p_i(\cdot)$: state-duration distribution of i-th state

d_i : state duration of i-th state

K : # of states in a sentence HMM for w

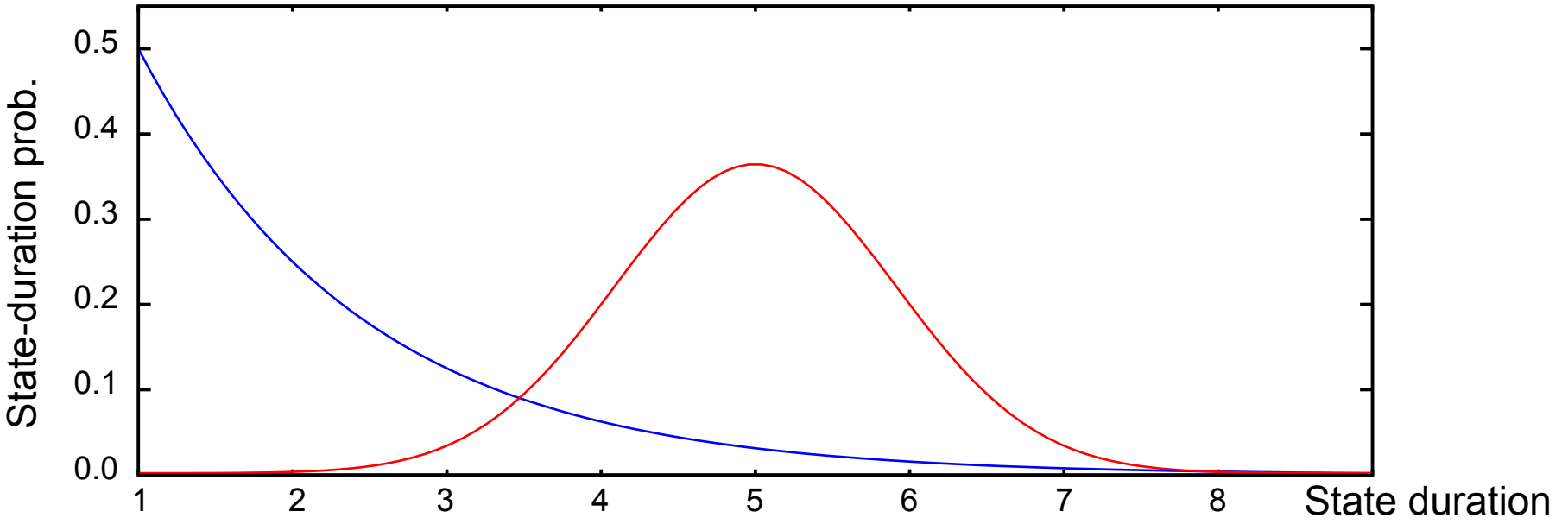
Determination of state sequence (3)

Geometric

$$p_i(d_i) = a_{ii}^{d_i - 1}(1 - a_{ii}) \;\Rightarrow\; \hat{d}_i = 1$$

Gaussian

$$p_i(d_i) = \mathcal{N}(d_i \,;\, m_i, \sigma_i^2) \;\Rightarrow\; \hat{d}_i = m_i$$



Speech parameter generation algorithm [Tokuda;'00]

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid w, \hat{\boldsymbol{\lambda}})$$

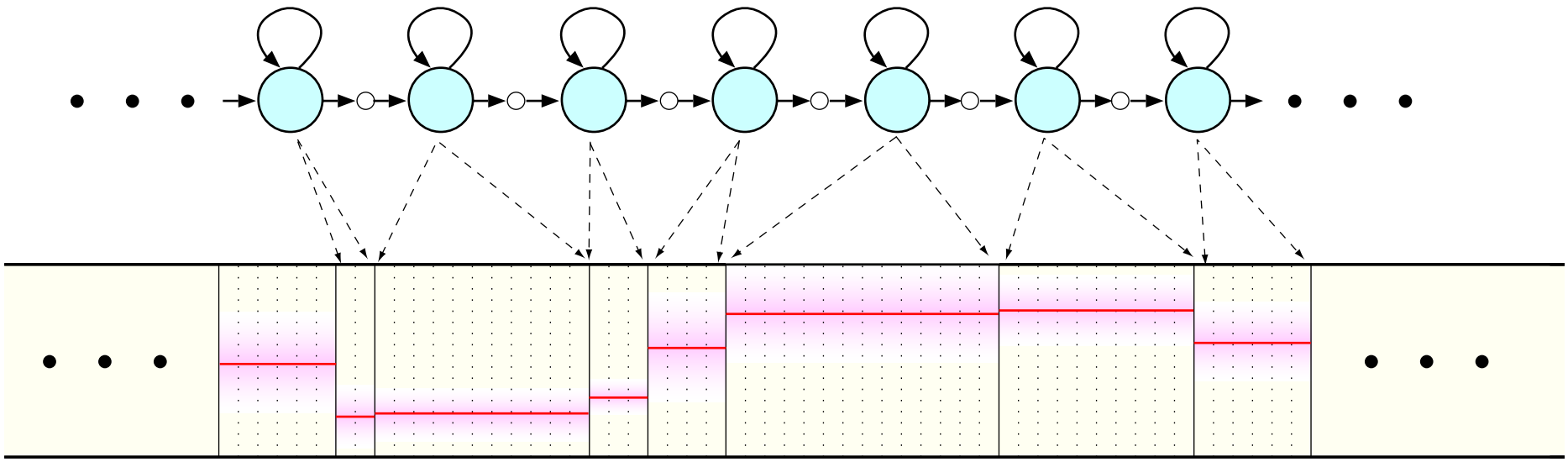
$$= \arg\max_{\boldsymbol{o}} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}})$$

$$\approx \arg\max_{\boldsymbol{o}} \max_{\boldsymbol{q}} p(\boldsymbol{o} \mid \boldsymbol{q}, \hat{\boldsymbol{\lambda}})\, p(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}})$$

$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid w, \hat{\boldsymbol{\lambda}})$$

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \hat{\boldsymbol{\lambda}})$$

Without dynamic features

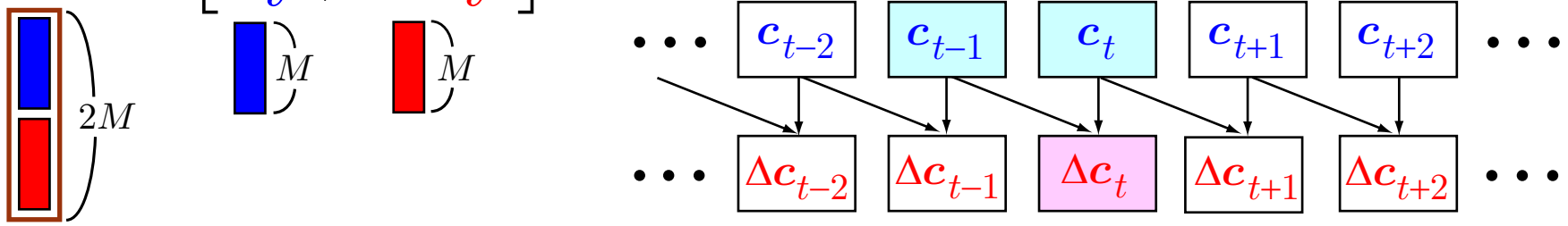


$\hat{o}$ becomes a sequence of mean vectors

Integration of dynamic features

Speech param. vec. $\boldsymbol{o}_t$ includes both static & dyn. feats.

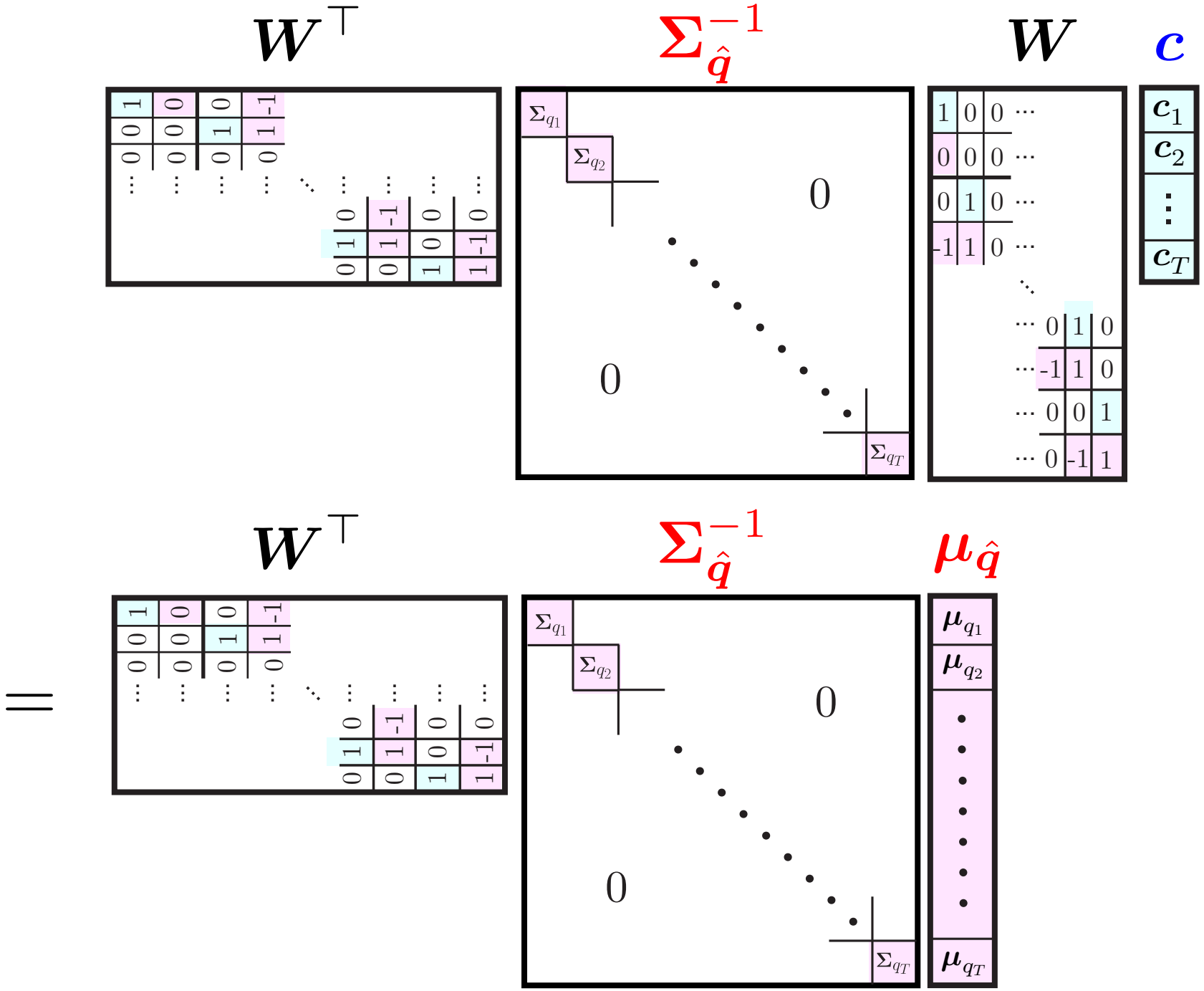
$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top\right]^\top \qquad \Delta\boldsymbol{c}_t = \boldsymbol{c}_t - \boldsymbol{c}_{t-1}$$



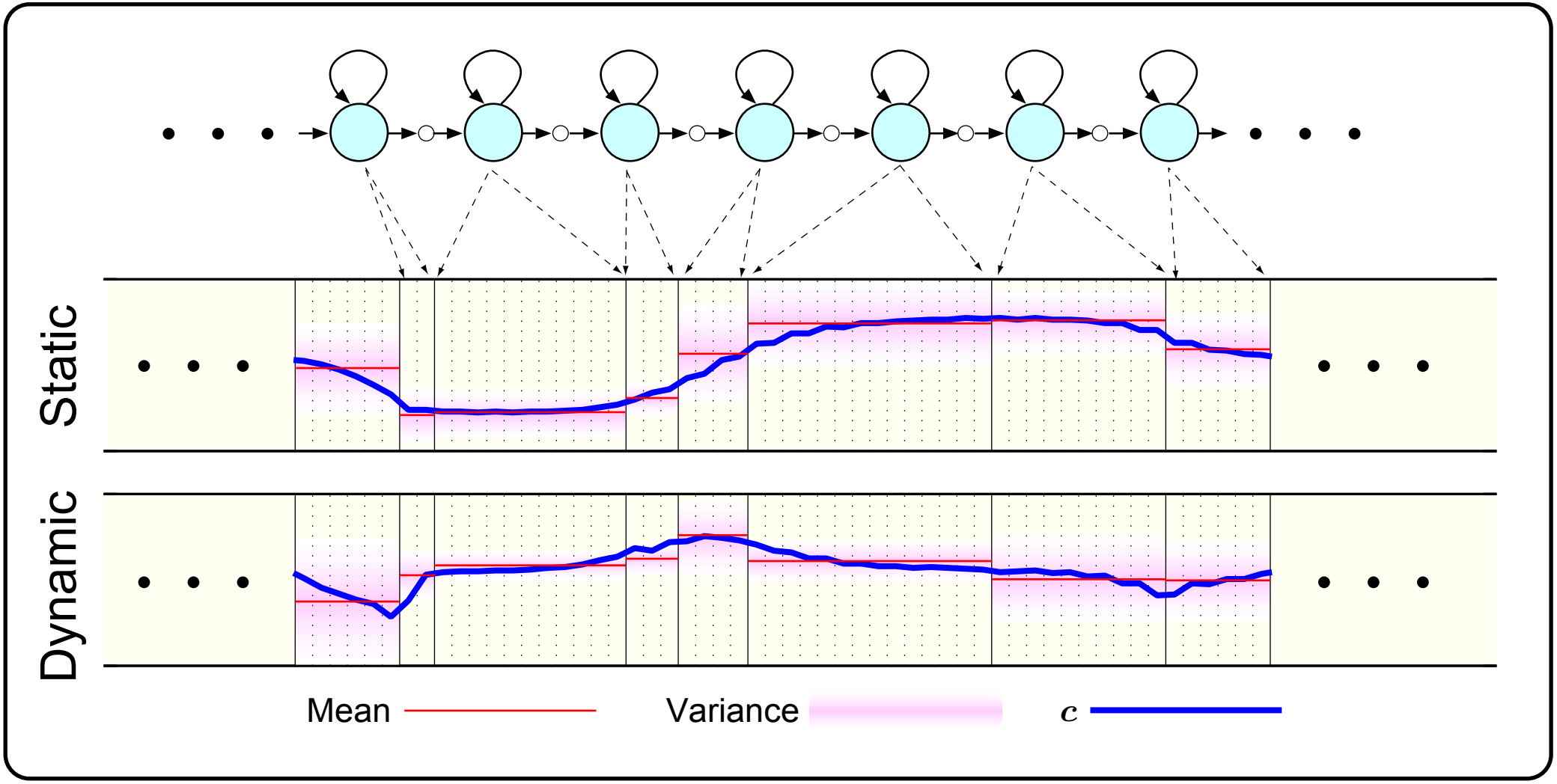
The relationship between $\boldsymbol{o}_t$ & $\boldsymbol{c}_t$ can be arranged as

$$\begin{array}{c} \boldsymbol{o} \\ \begin{bmatrix} \vdots \\ \boldsymbol{c}_{t-1} \\ \Delta\boldsymbol{c}_{t-1} \\ \boldsymbol{c}_t \\ \Delta\boldsymbol{c}_t \\ \boldsymbol{c}_{t+1} \\ \Delta\boldsymbol{c}_{t+1} \\ \vdots \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \boldsymbol{o}_{t-1} \\ \\ \boldsymbol{o}_t \\ \\ \boldsymbol{o}_{t+1} \\ \\ \\ \end{array} = \begin{array}{c} \boldsymbol{W} \\ \begin{bmatrix} \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & -\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{I} & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \end{bmatrix} \end{array} \begin{array}{c} \boldsymbol{c} \\ \begin{bmatrix} \vdots \\ \boldsymbol{c}_{t-2} \\ \boldsymbol{c}_{t-1} \\ \boldsymbol{c}_t \\ \boldsymbol{c}_{t+1} \\ \vdots \end{bmatrix} \end{array}$$

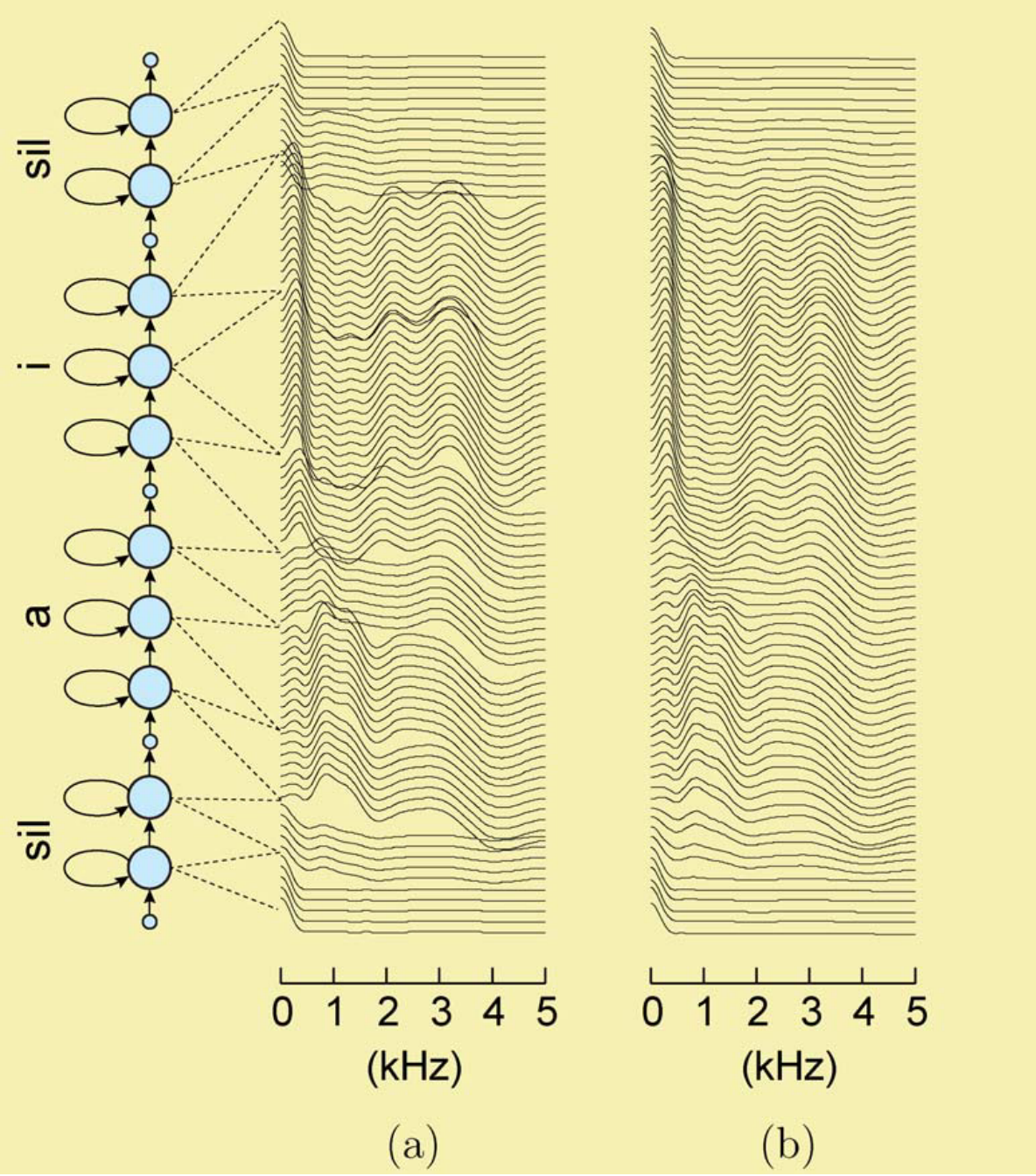
Solution for the problem



Generated speech parameter trajectory



Generated spectra



Synthesis tools in HTS

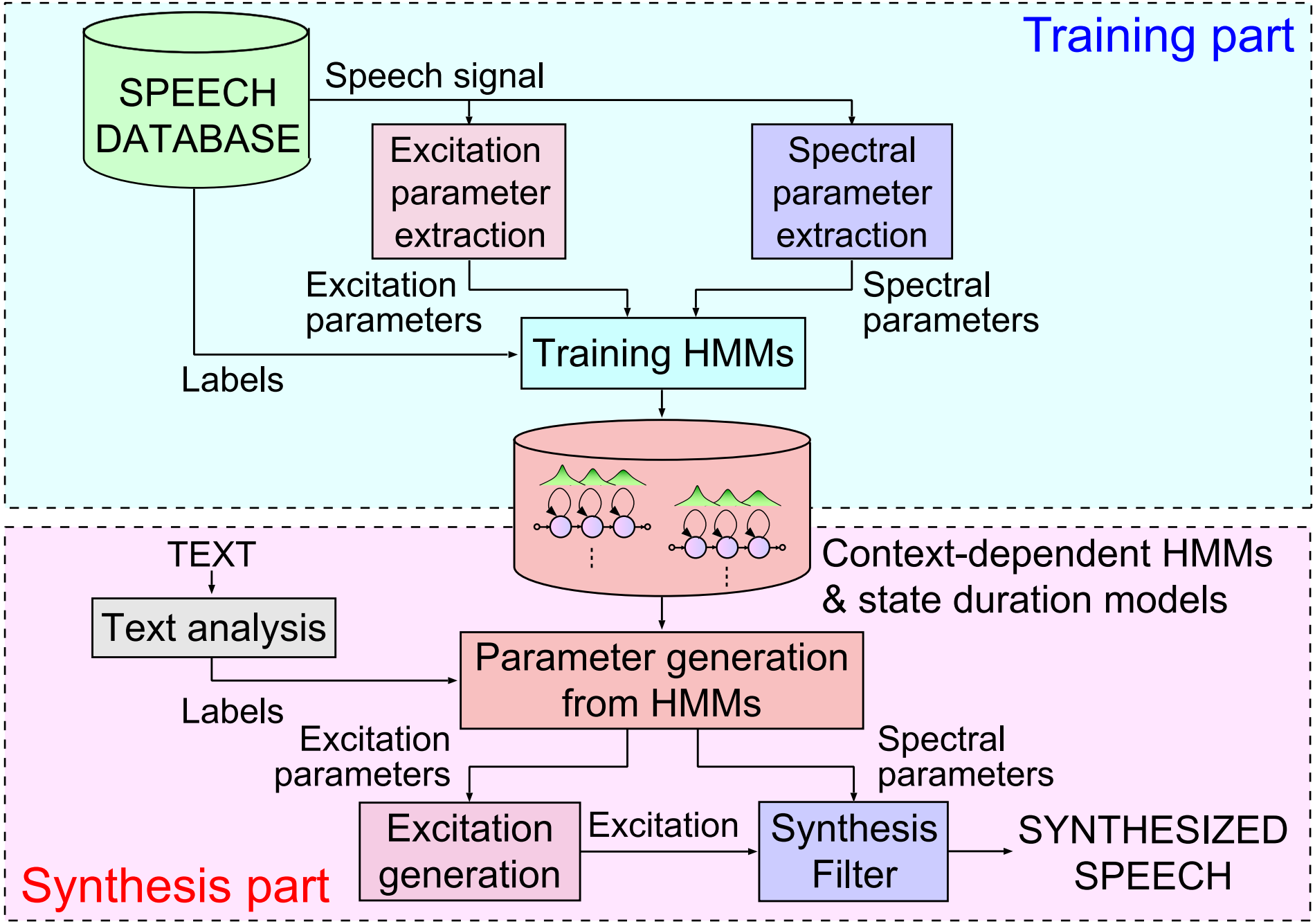
- **HMGenS**
  - Tool to generate speech parameters from HMMs
  - Requires HTK/HTS library modules
  - Many functionalities

    ⇒ **For research purpose**
- **hts_engine**
  - Web: http://hts-engine.sourceforge.net/
  - Small run-time synthesis engine including waveform synth.
  - API & stand-alone program
  - Works without HTK/HTS library

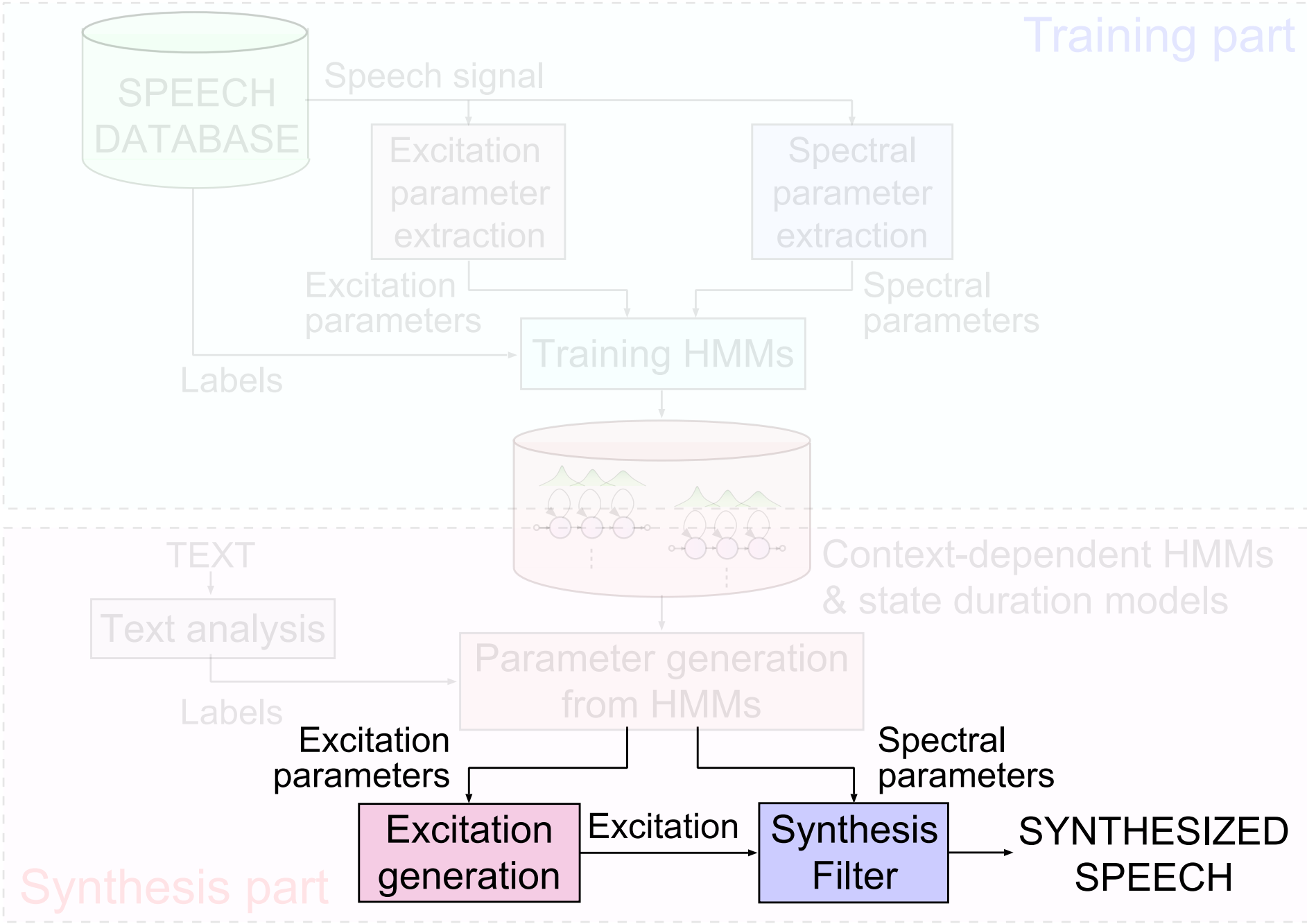
    ⇒ **For application development purpose**

HTS-demo demonstrates how to use both tools

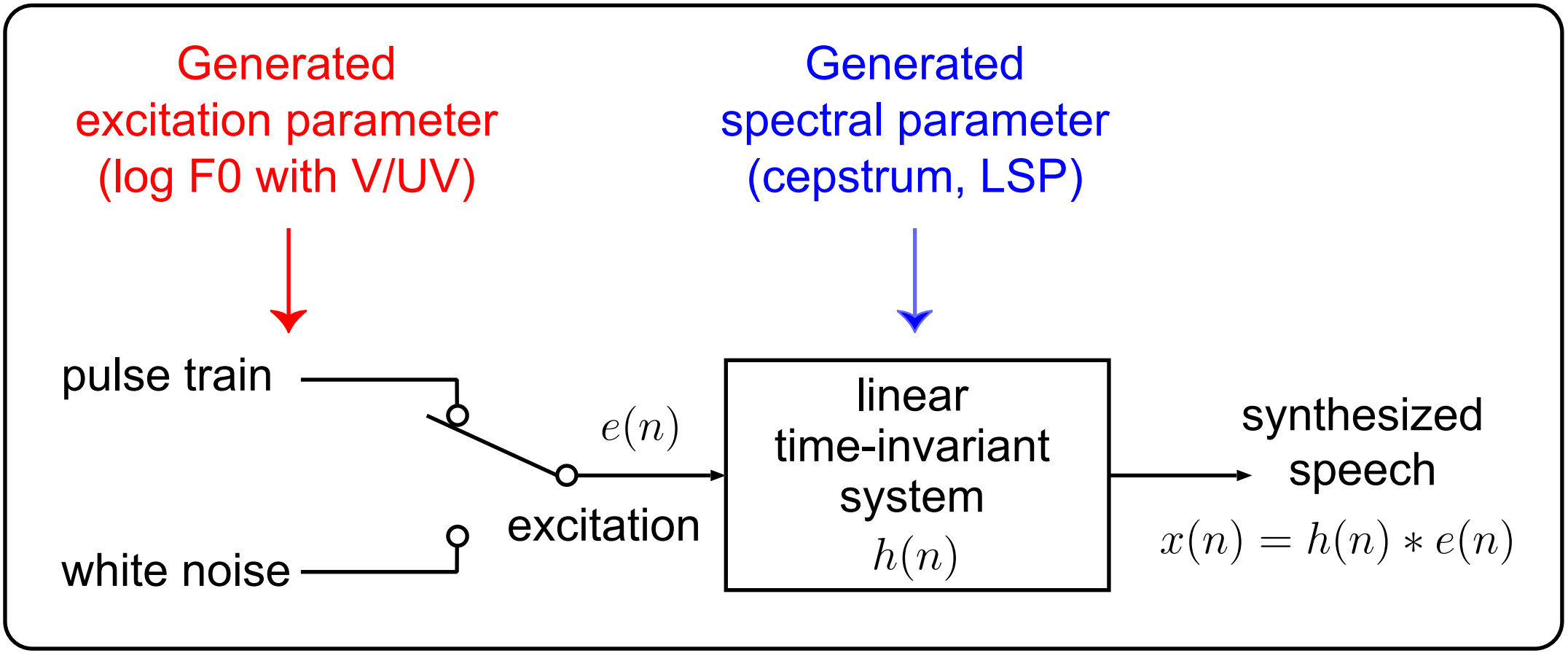
HMM-based speech synthesis system (HTS)



HMM-based speech synthesis system (HTS)



Source-filter model



Spectral representation & corresponding filter

cepstrum: LMA filter

generalized cepstrum: GLSA filter

mel-cepstrum: MLSA filter

mel-generalized cepstrum: MGLSA filter

LSP: LSP filter

PARCOR: all-pole lattice filter

LPC: all-pole filter

Structure of synthesis filter $H(z)$ $(\gamma = 0)$

$$H(z) = \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m}$$

Exponential function
$\Rightarrow$ digital filter 

$\downarrow$ Pade approximation

$$\approx R_L \left(\sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_{\alpha}^{-m} \right)$$

Rational function
$\Rightarrow$ digital filter 

- Sufficient accuracy: maximum spectral error 0.24dB
- $O(8M)$ multiply-add operations per sample
- Stable filter
- M multiply-add operations for filter coefficients calculation

Speech signal processing toolkit (SPTK)

Synthesis

Excitation signal generation: excite

LMA filter (cepstrum): lmadf

GLSA filter (generalized cepstrum): glsadf

MLSA filter (mel-cepstrum): mlsadf

MGLSA filter (mel-generalized cepstrum): mglsadf

LSP filter (LSP): lspdf

All-pole lattice filter (PARCOR): ltcdf

All-pole filter (LPC): poledf

FIR filter: zerodf

Speech samples

		Mel-cepstrum	
		w/ dyn.	w/o dyn.
log F0	w/ dyn.		
	w/o dyn.		

Summary (first half)

Fundamentals in HMM-based speech synthesis

- Probabilistic formulation of corpus-based synthesis
- HMM-based speech synthesis system
  - Spectral analysis
  - Acoustic modeling (HMM)
  - Parameter generation
  - Speech synthesis filter

Any questions?

Time-line

16:15 ~ 17:45: Second half

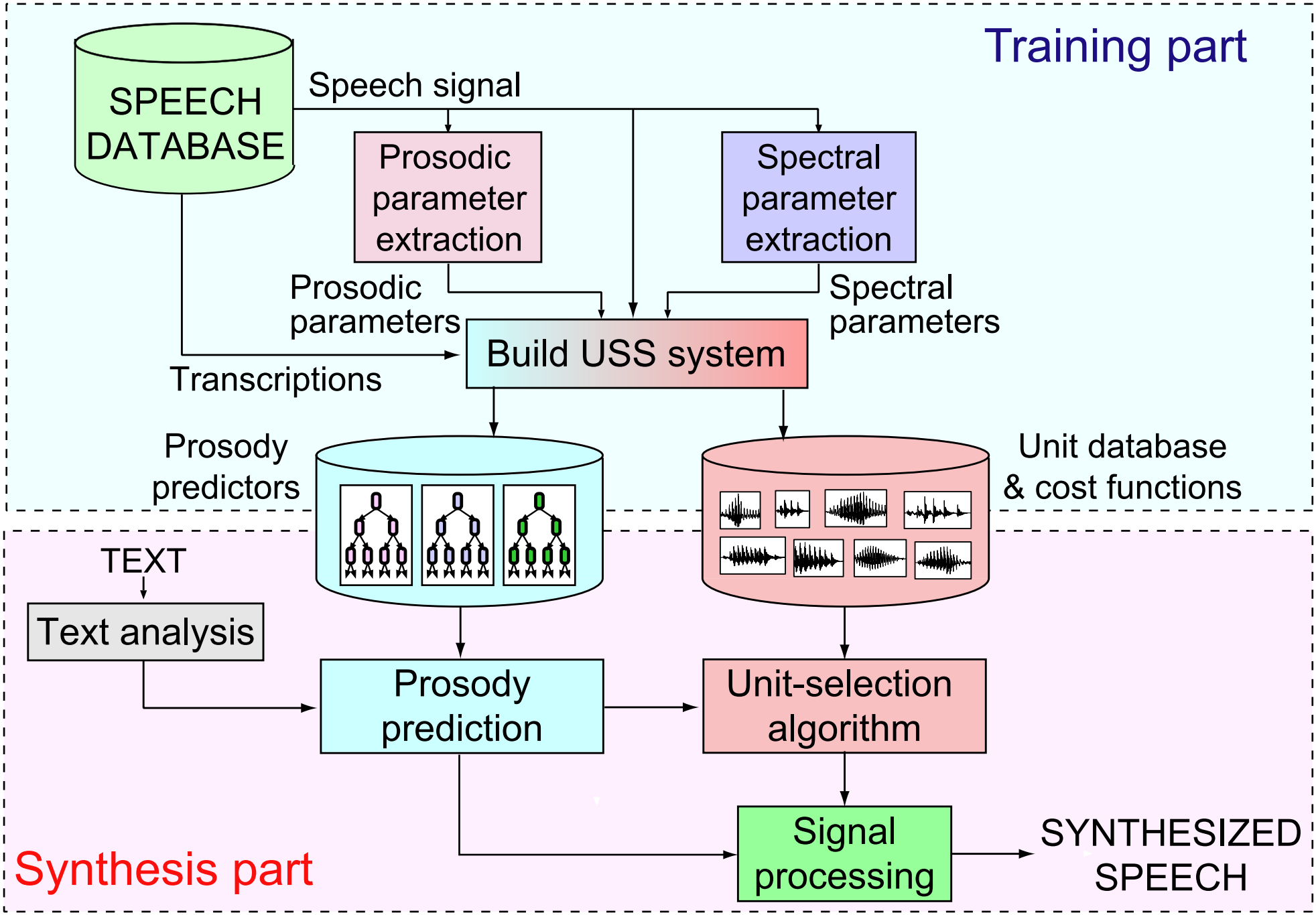
- Related topics
  - * Unit selection & hybrid
  - * Flexibility to control voice characteristics
    - · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  - * Vocoding
  - * Acoustic modeling
  - * Over-smoothing compensation
- Applications
- Q&A (10min)

Time-line

16:15 ~ 17:45: Second half

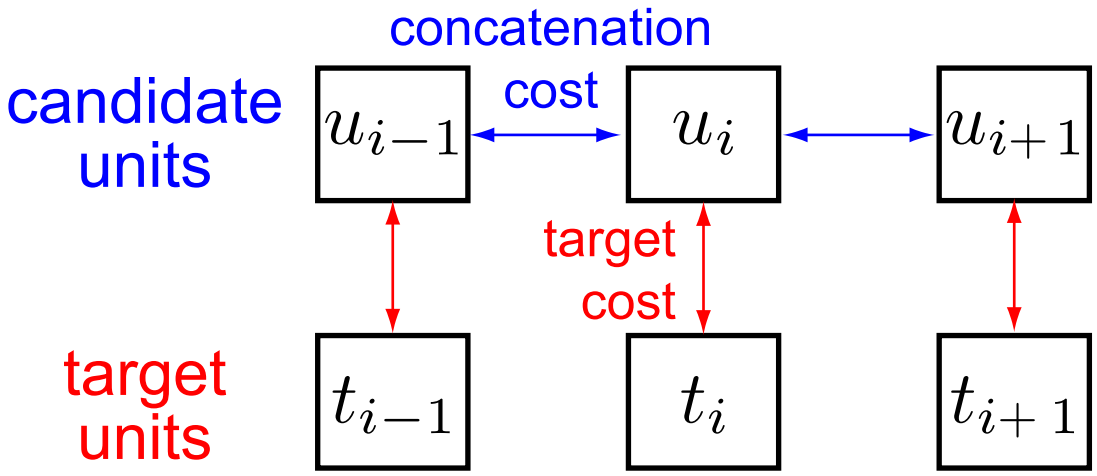
- Related topics
  * Unit selection & hybrid
  * Flexibility to control voice characteristics
    · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  * Vocoding
  * Acoustic modeling
  * Over-smoothing compensation
- Applications
- Q&A (10min)

Unit-selection synthesis (USS) (1)



Unit-selection synthesis (USS) (2)

Unit-selection algorithm [Hunt;'96]



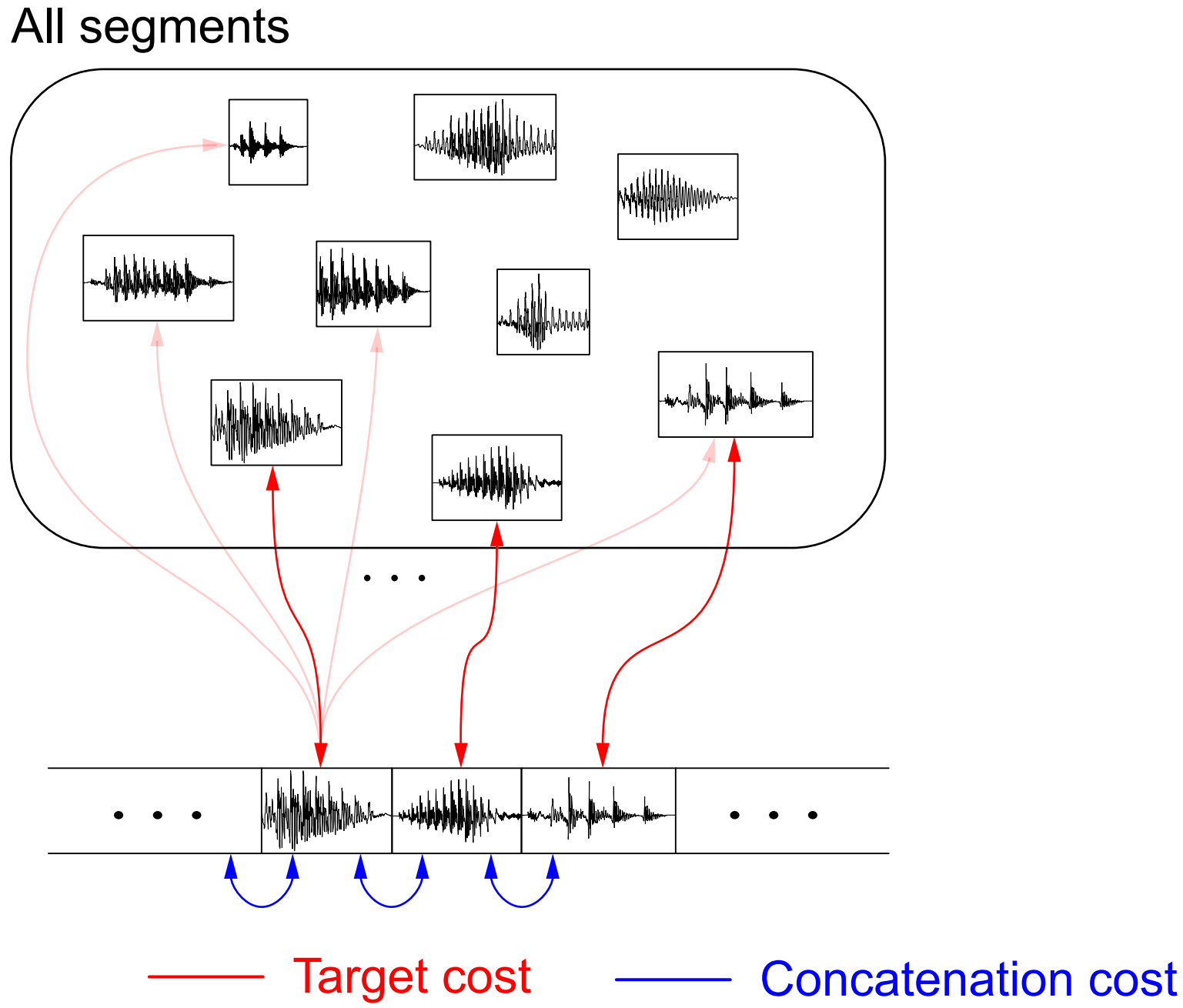
$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^c C_k^c(u_{i-1}, u_i)$$

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i)$$

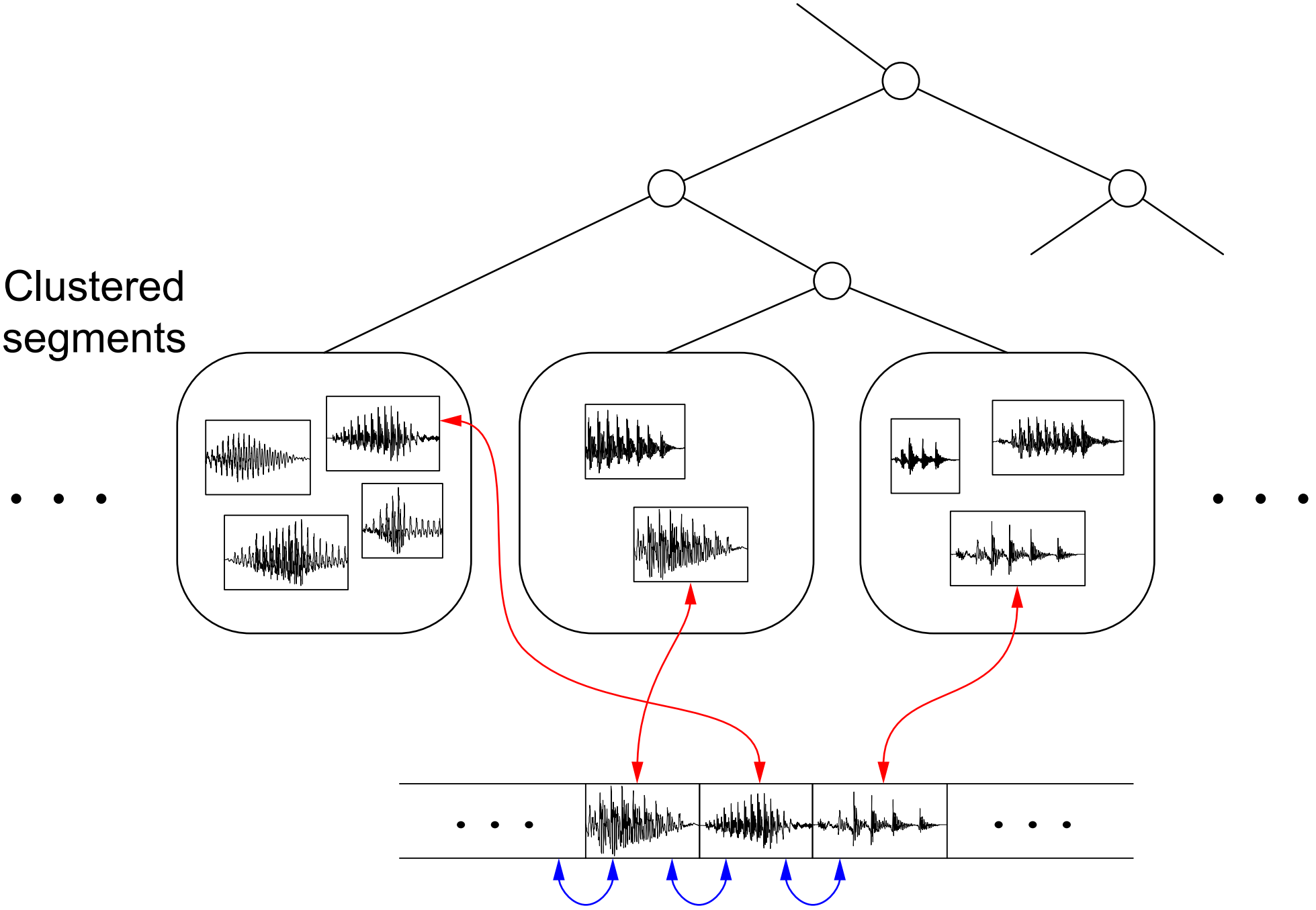
$$\hat{u}_1^n = \arg\min_{u_1^n} \{C(t_1^n, u_1^n)\} \qquad C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i)$$

- **Major techniques**
  - General selection technique [Hunt;'96]
  - Clustering-based technique [Donovan;'95]

General unit-selection synthesis scheme

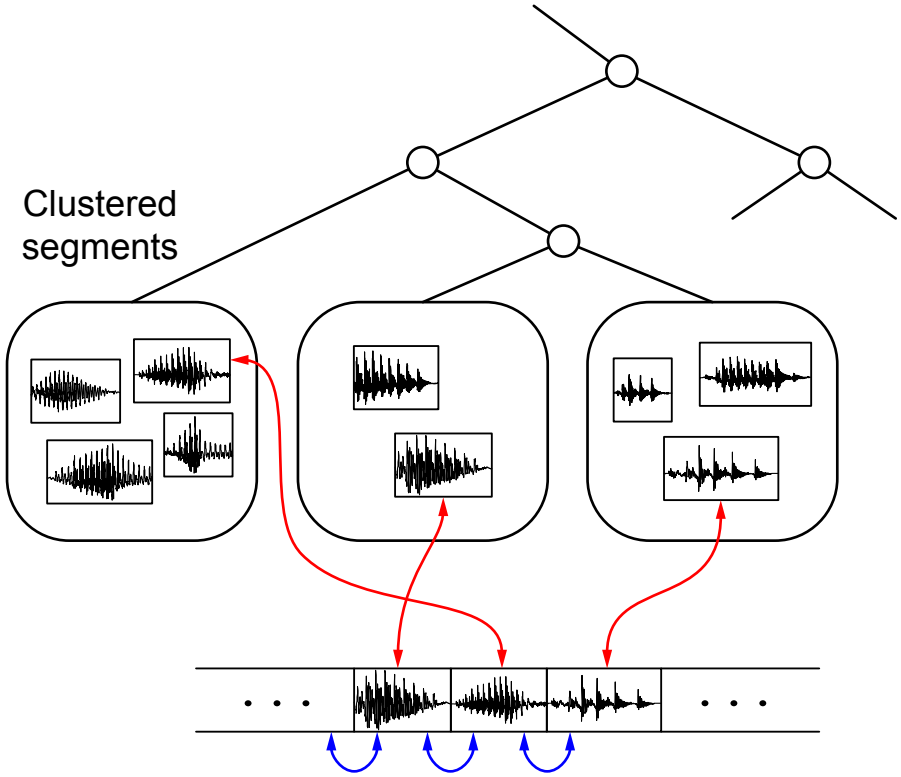


Clustering-based unit-selection synthesis scheme

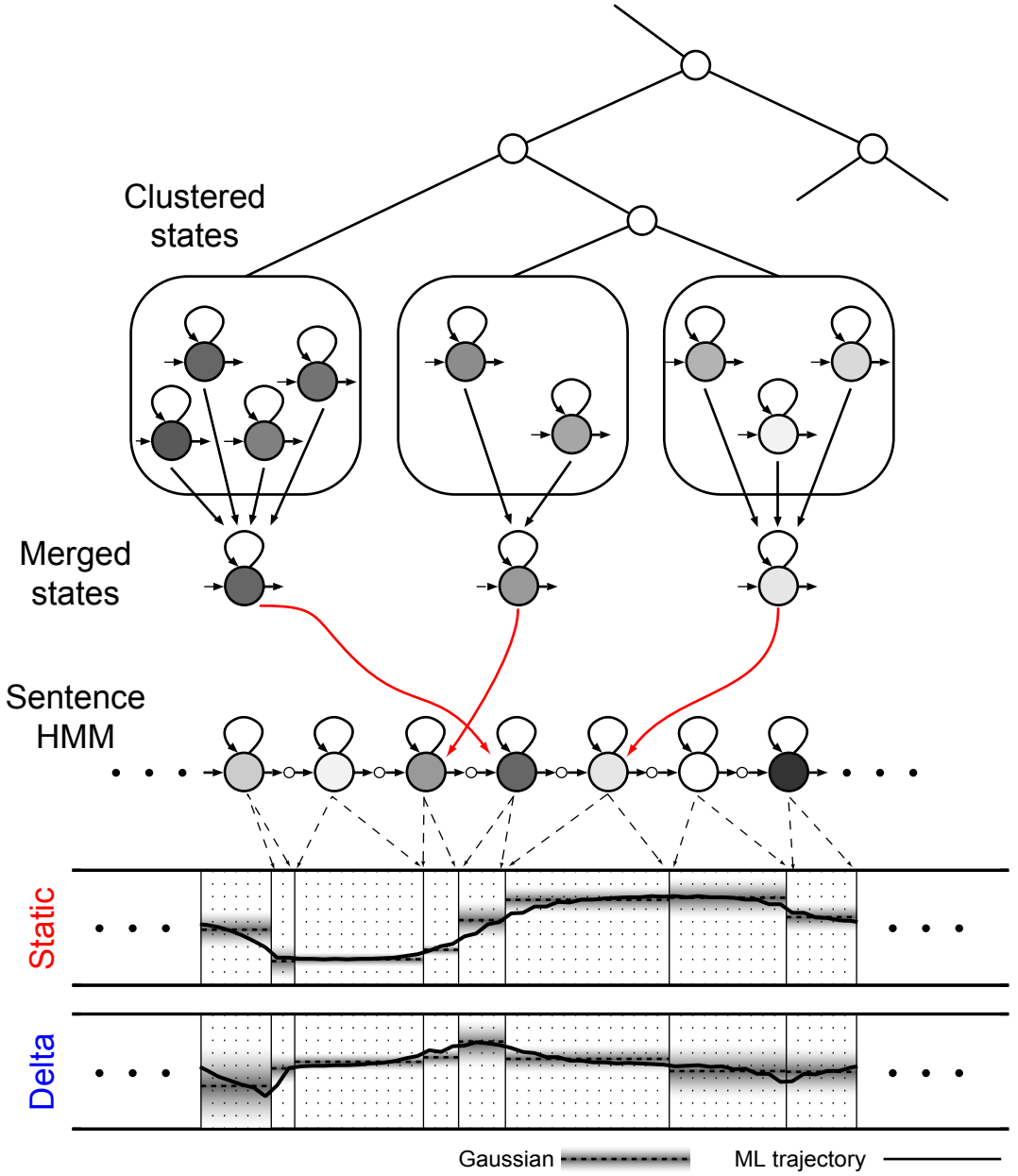


Relation between two approaches (1)

Unit selection

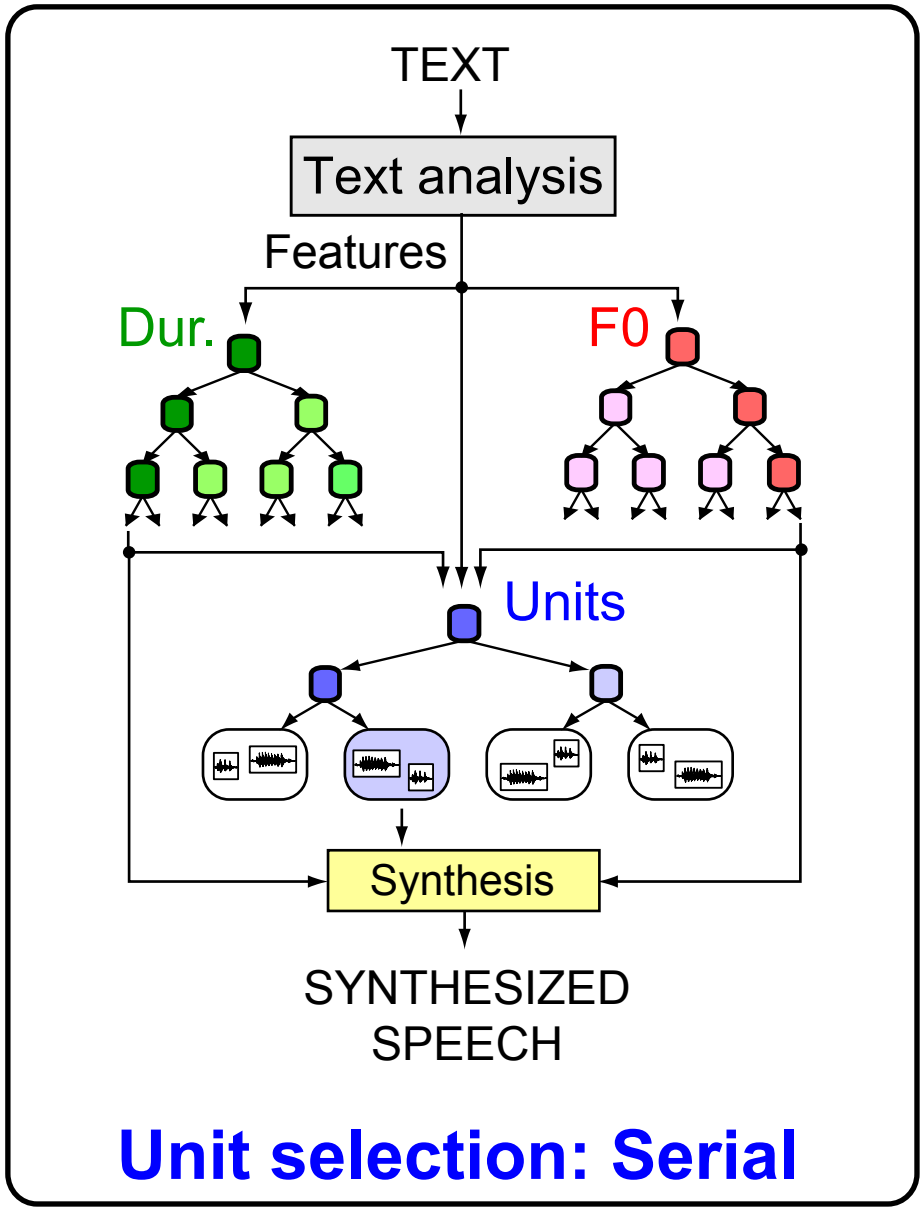


HMM-based synthesis



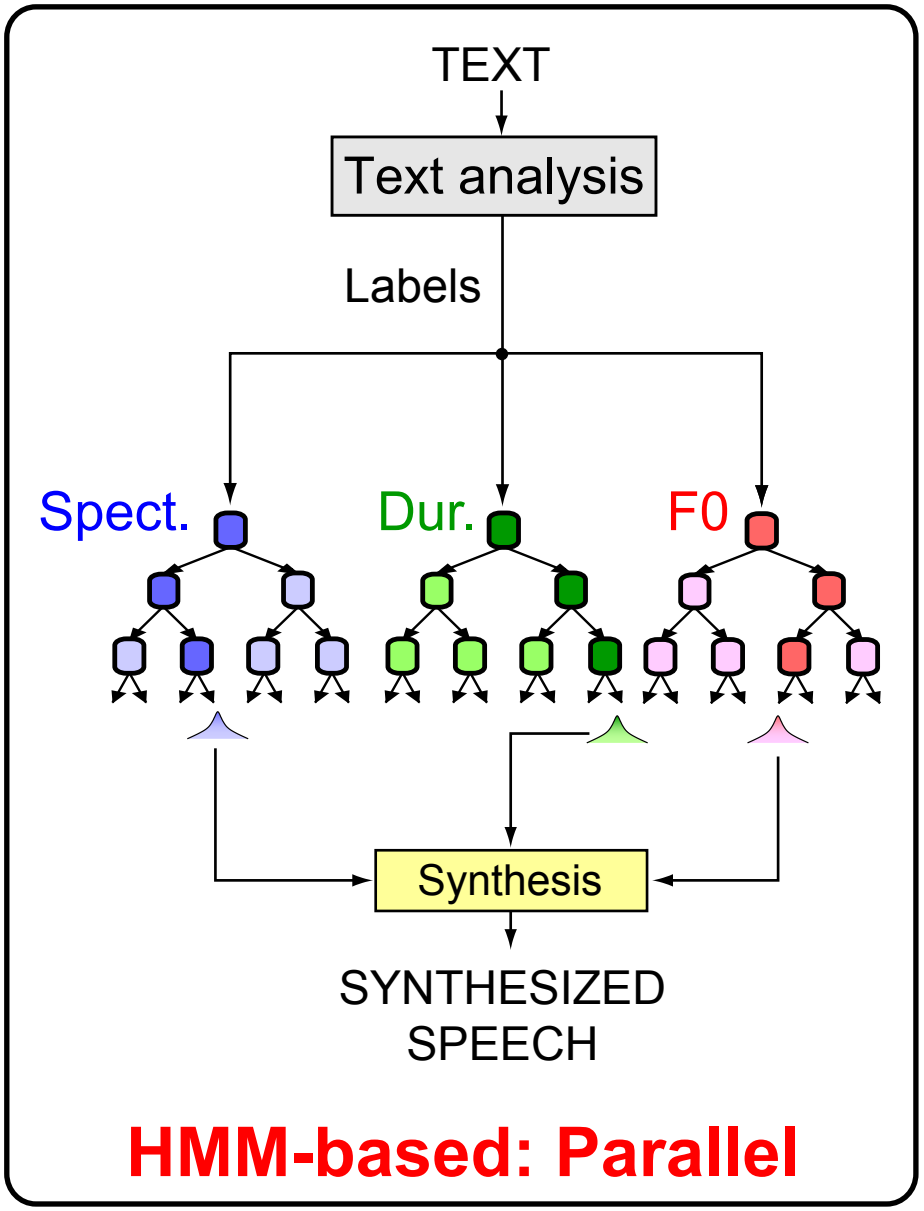
Relation between two approaches (2)

Unit selection



Unit selection: Serial

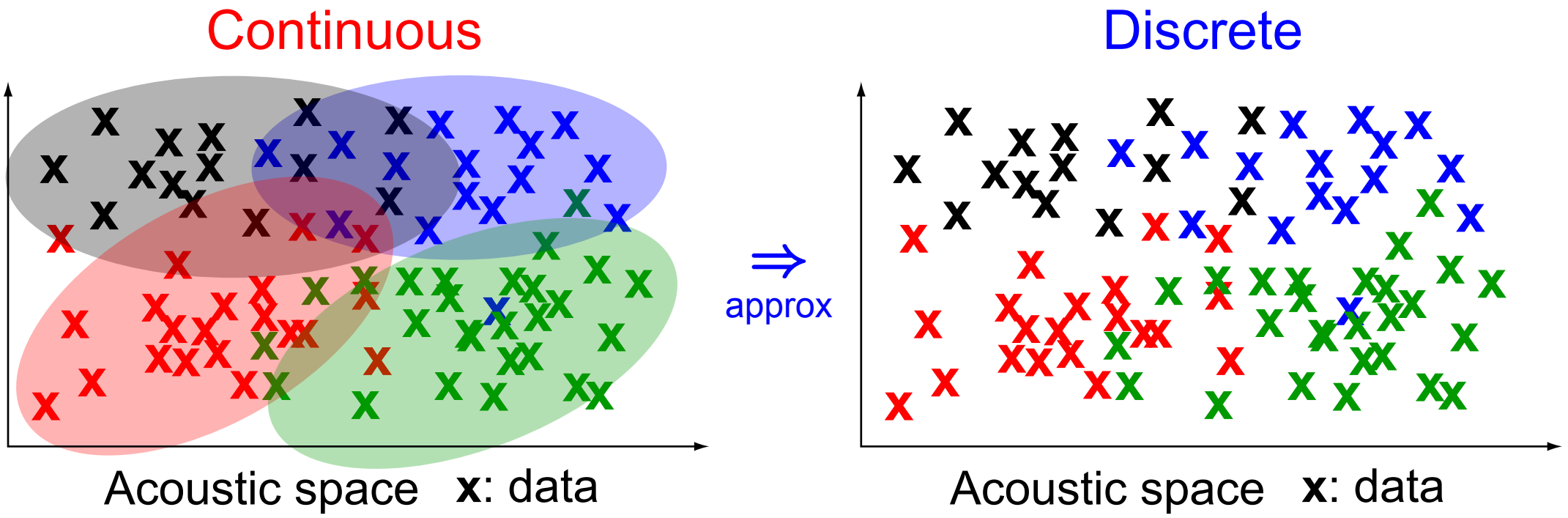
HMM-based synthesis



HMM-based: Parallel

Relation between two approaches (3)

Approximate state-output distribution by templates



ML-based param. generation with discrete distribution

⇒ Result in frame-wise DP search like unit selection

Parameter generation ⇒ "analogue" version of US

Hybrid approaches (1)

Hybrid between unit-selection & HMM-based synth.

- **Target prediction**
  - HMM outputs as targets
  - HMM probabilities as costs
- **Smoothing units**
  - Smooth discontinuities between units
- **Mixing units**
  - Mix natural & generated units

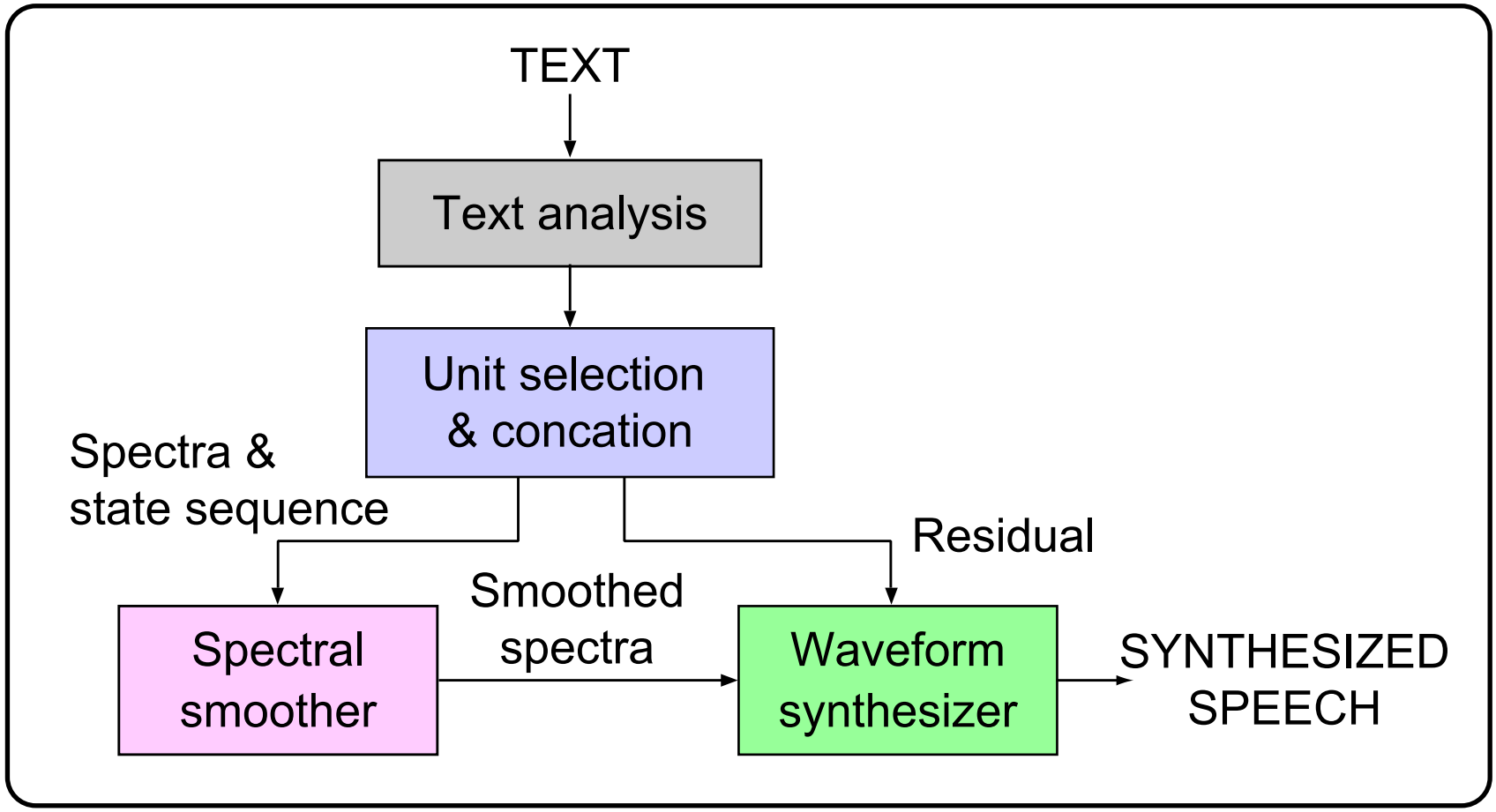
Hybrid approaches (2)

Target prediction

- **HMM outputs as targets**
  - MFCCs, F0s, durations from HMMs as target
    * ATR [Kawai;'04], Arcadia [Hirai;'04], France Telecom [Rouibia;'05]
- **HMM probabilities as costs**
  - Static features ⇒ Target costs
  - Dynamic features ⇒ Concatenation costs
    * Microsoft [Huang;'96], Nitech [Mizutani;'02], iFlyTek [Ling;'07]
- **Minimum unit-selection error training [Ling;'08]**
  - Loss func.: # of diff. units between selected & natural units
  - Optimize HMM params. to minimize the loss func. by GPD

Hybrid approaches (3)

Smoothing units

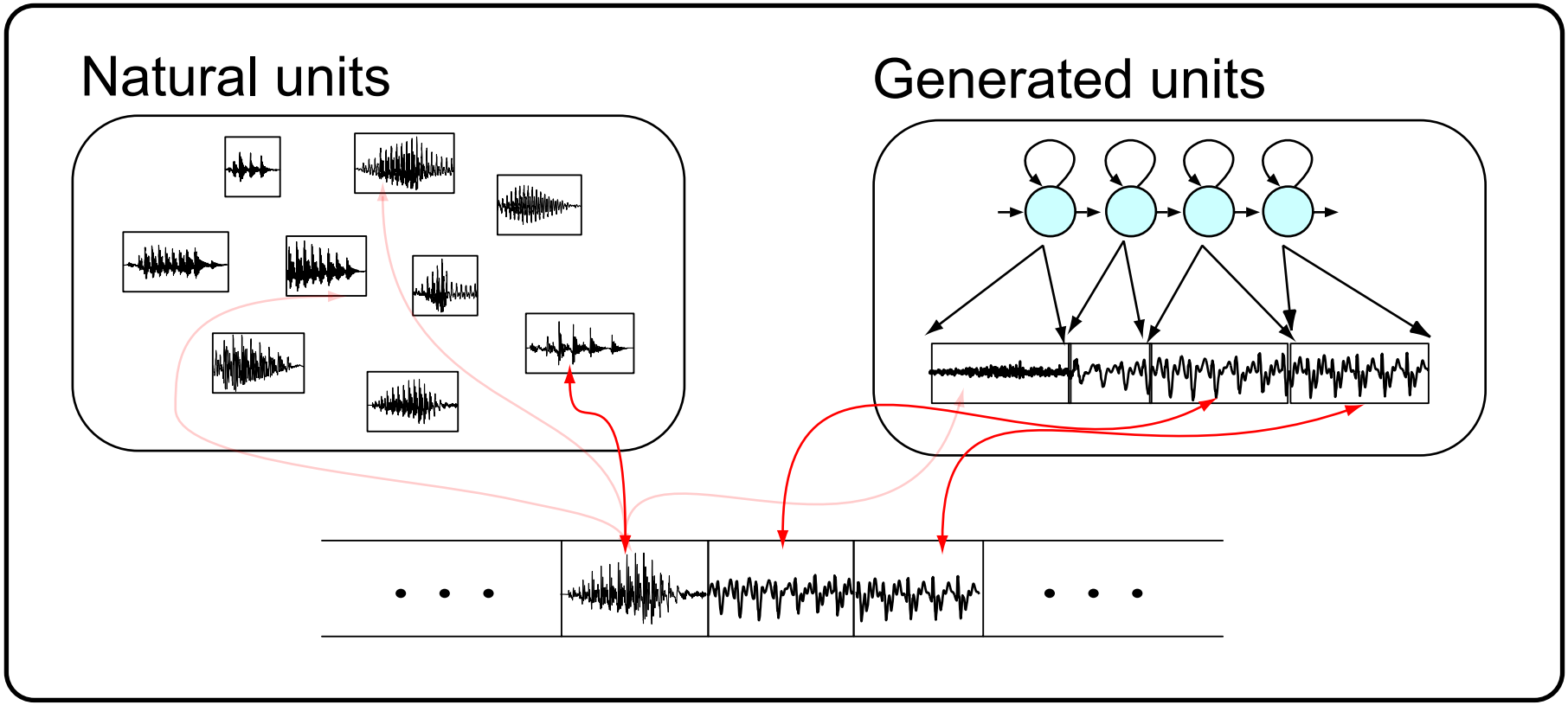


Smooth unit boundaries using HMM statistics

- Microsoft [Plumpe;'98], OGI [Wouters;'00]

Hybrid approaches (4)

Mixing units



Mix natural & generated units

⇒ Avoid data sparsity & produce large proportion of speech

- Waseda Univ. [Okubo;'06], Cereproc [Aylett;'08], Nuance [Pollet;'08]

Hybrid approaches (5)

Issues in hybrid approach

- **Non-parametric approach**
  - Store templates & statistics
    → Increase footprint
  - Generation + search
    → Increase computational cost
  - Speech is synthesized from **templates** & statistics
    → Lose **flexibility** to change its voice characteristics

Time-line

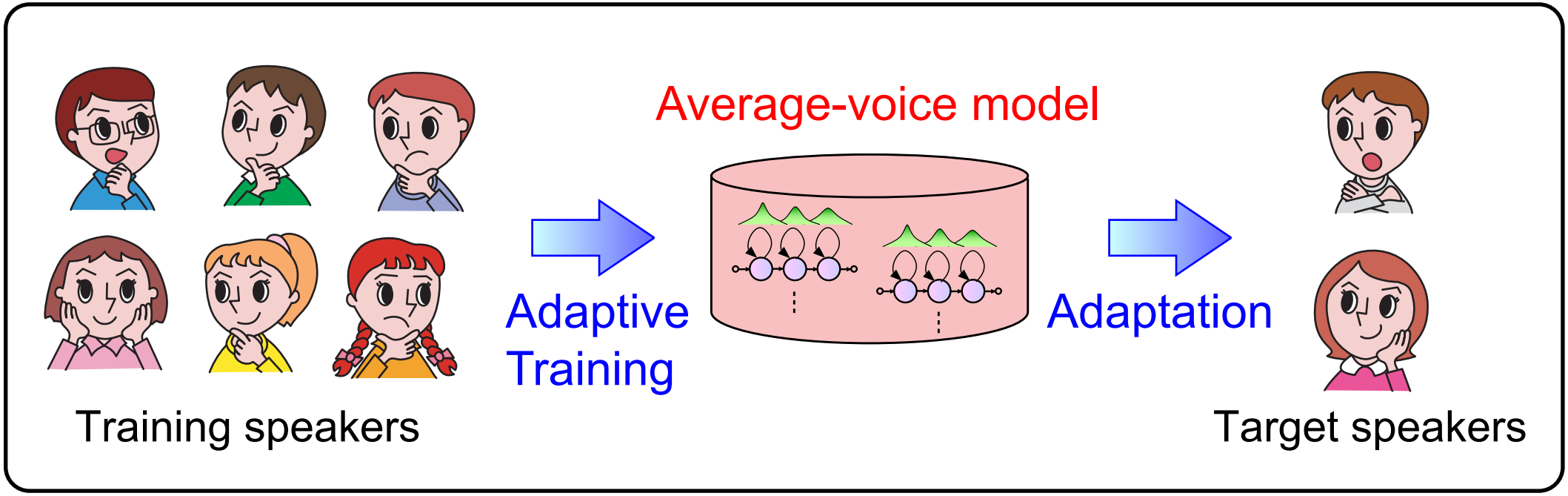
16:15 ~ 17:45: Second half

- Related topics
  - * Unit selection & hybrid
  - * Flexibility to control voice characteristics
    - · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  - * Vocoding
  - * Acoustic modeling
  - * Over-smoothing compensation
- Applications
- Q&A (10min)

Adaptation (Mimicking voices)

Adaptation/adaptive training of HMMs

- Originally developed in ASR, but works very well in TTS
- Average voice-based speech synthesis (AVSS) [Yamagishi;'06]



- Require small data of target speaker/speaking style

⇒ **Smal cost to create new voices**

Adaptation demo

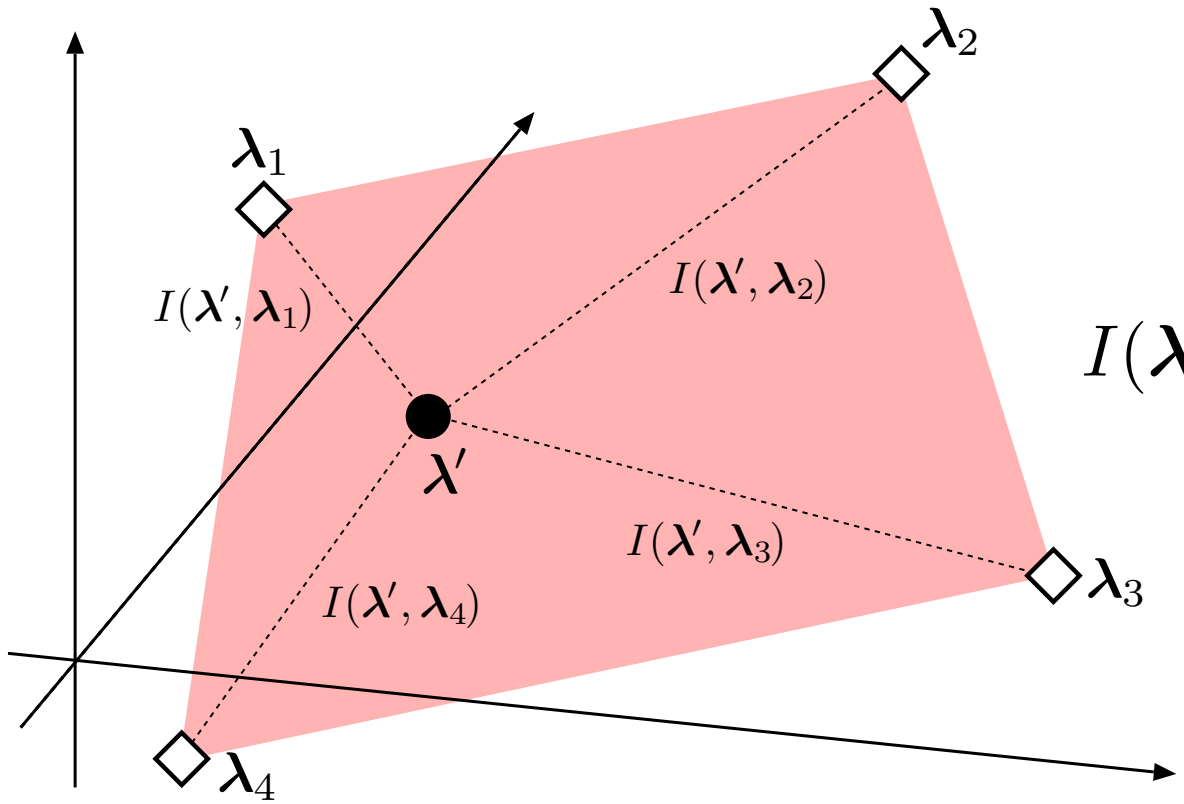
- **Speaker adaptation**
  - VIP voice: GWB BHO
  - Child voice:
- **Style adaptation (in Japanese)**
  - Joyful
  - Sad
  - Rough

From http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/demo.html

Interpolation (Mixing voices)

Interpolate parameters among representive HMM sets

- Can obtain new voices even no adaptation data is available
- Gradually change spkrs. & speaking styles [Yoshimura;'97,Tachibana;'05]

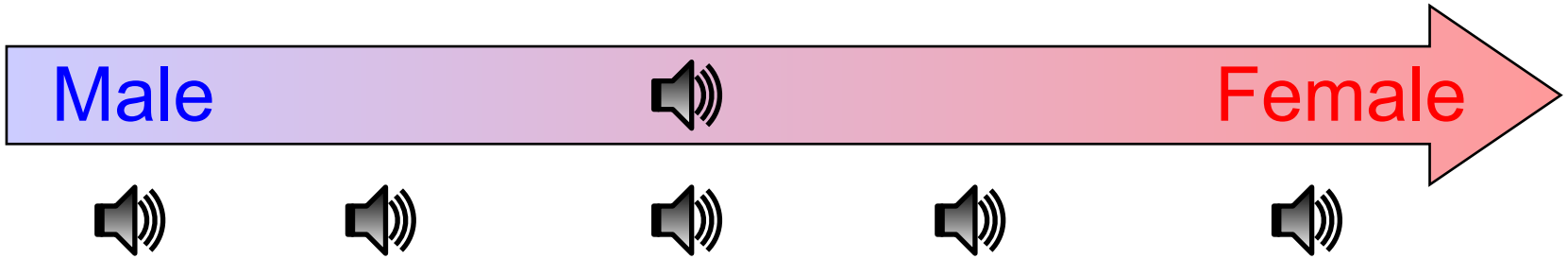


λ : HMM set

$I(\lambda', \lambda)$: Interpolation ratio

Interpolation demo

- **Speaker interpolation (in Japanese)**
  - Male & Female

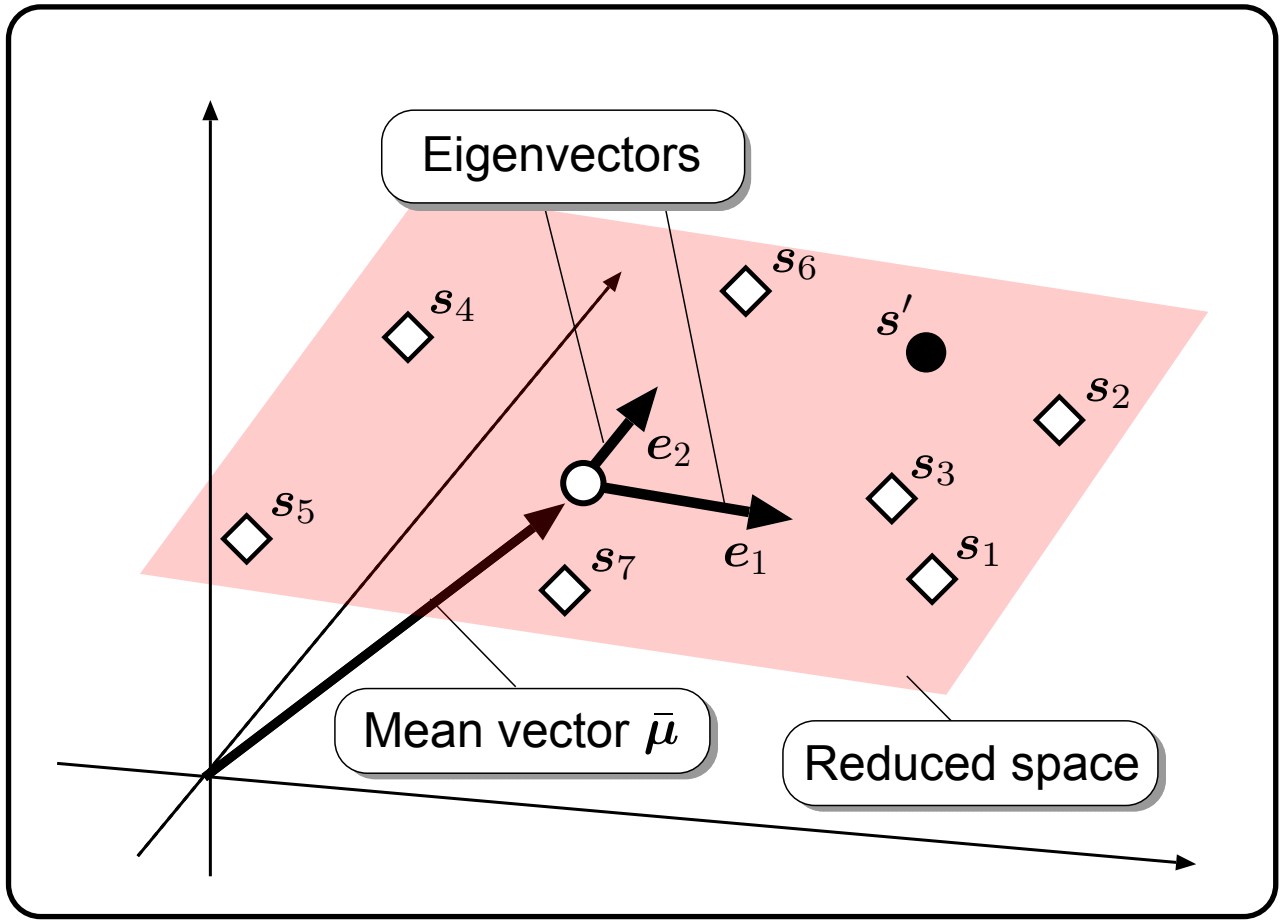


- **Style interpolation**
  - Neutral → Angry
  - Neutral → Happy

From http://www.sp.nitech.ac.jp/
& http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/demo.html

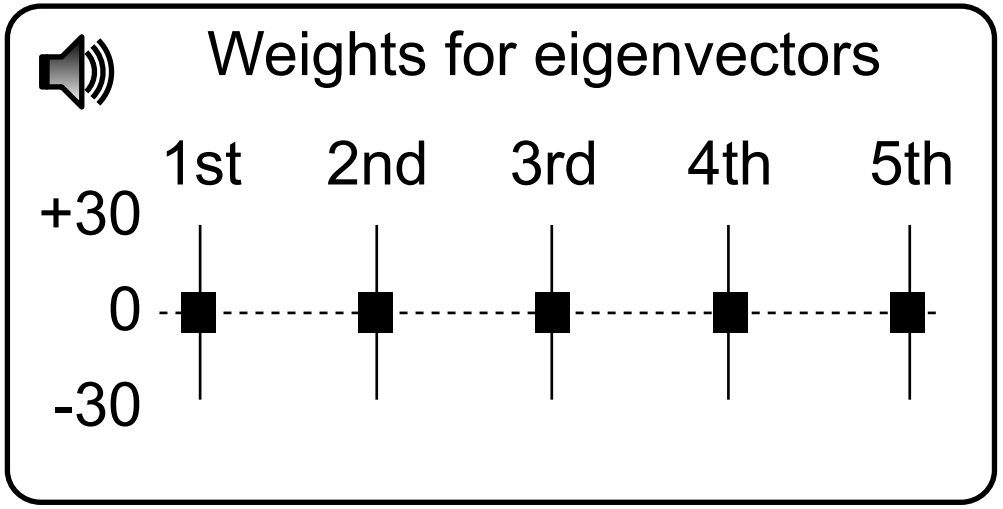
Eigenoice (Producing voices) [Shichiri;'02]

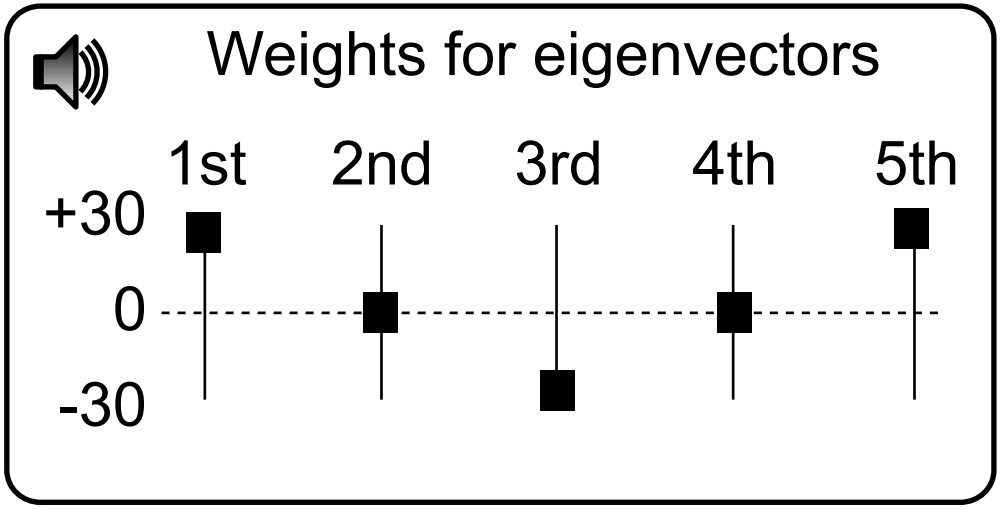
- **Increase # of representative HMM sets in interpolation**
  ⇒ Difficult to set interpolation ratio to obtain desired voice
- **Eigenvoice [Kuhn;'00]: Apply PCA to super-vectors**
  ⇒ Can reduce the dimensionality of speaker space

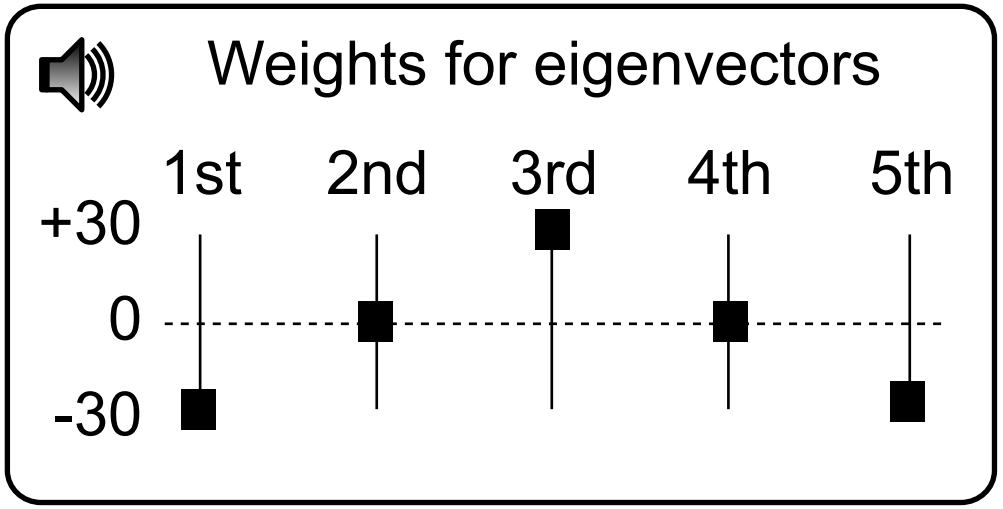


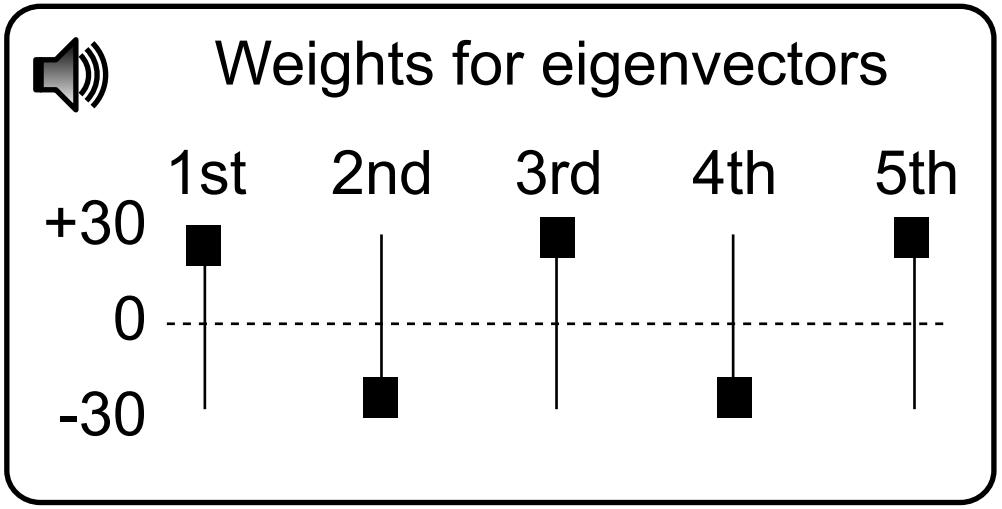
Eigenvoice demo

Speaker characteristics modification









From http://www.sp.nitech.ac.jp/~demo/synthesis_demo_2001/

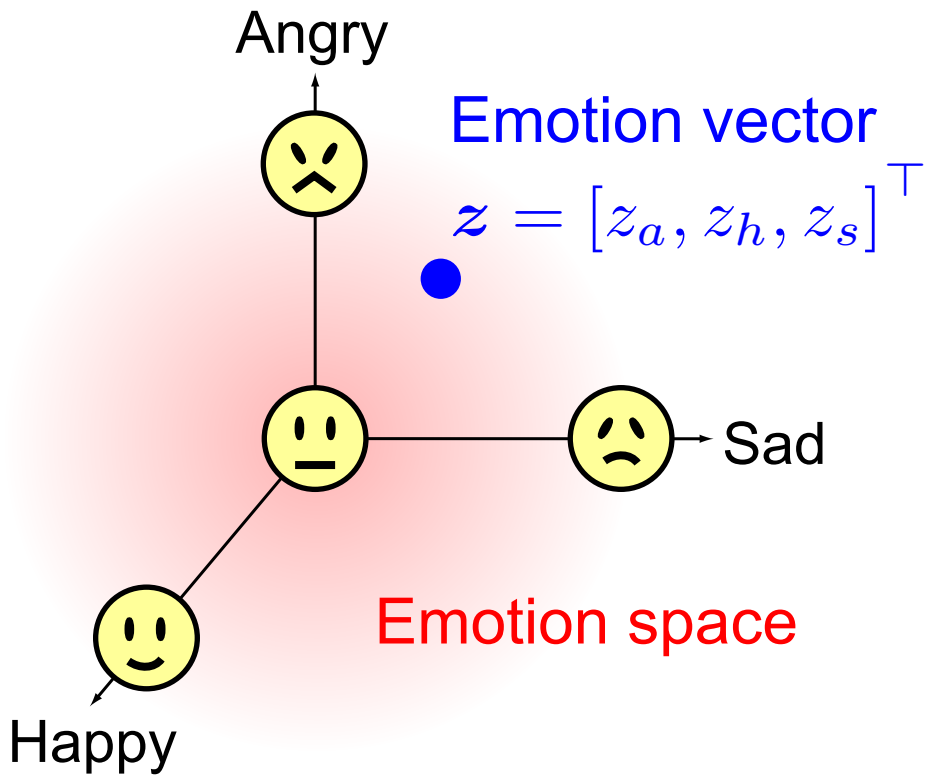
Multiple-regression (Controlling voices)

- **No eigen-vectors represents specific physical meaning**

  $\Rightarrow$ Difficult to control voice characteristics intutively

- **Multiple-regression HMMs** [Fujinaga;'01]

  - Assign intutive meaning to control synthetic voice [Nose;'07]



Emotion vector

$\boldsymbol{z} = [z_a, z_h, z_s]^\top$

$$\boldsymbol{\mu}_j = \boldsymbol{M}_j \boldsymbol{\xi}, \quad \boldsymbol{\xi} = \left[1, \boldsymbol{z}^\top\right]^\top$$

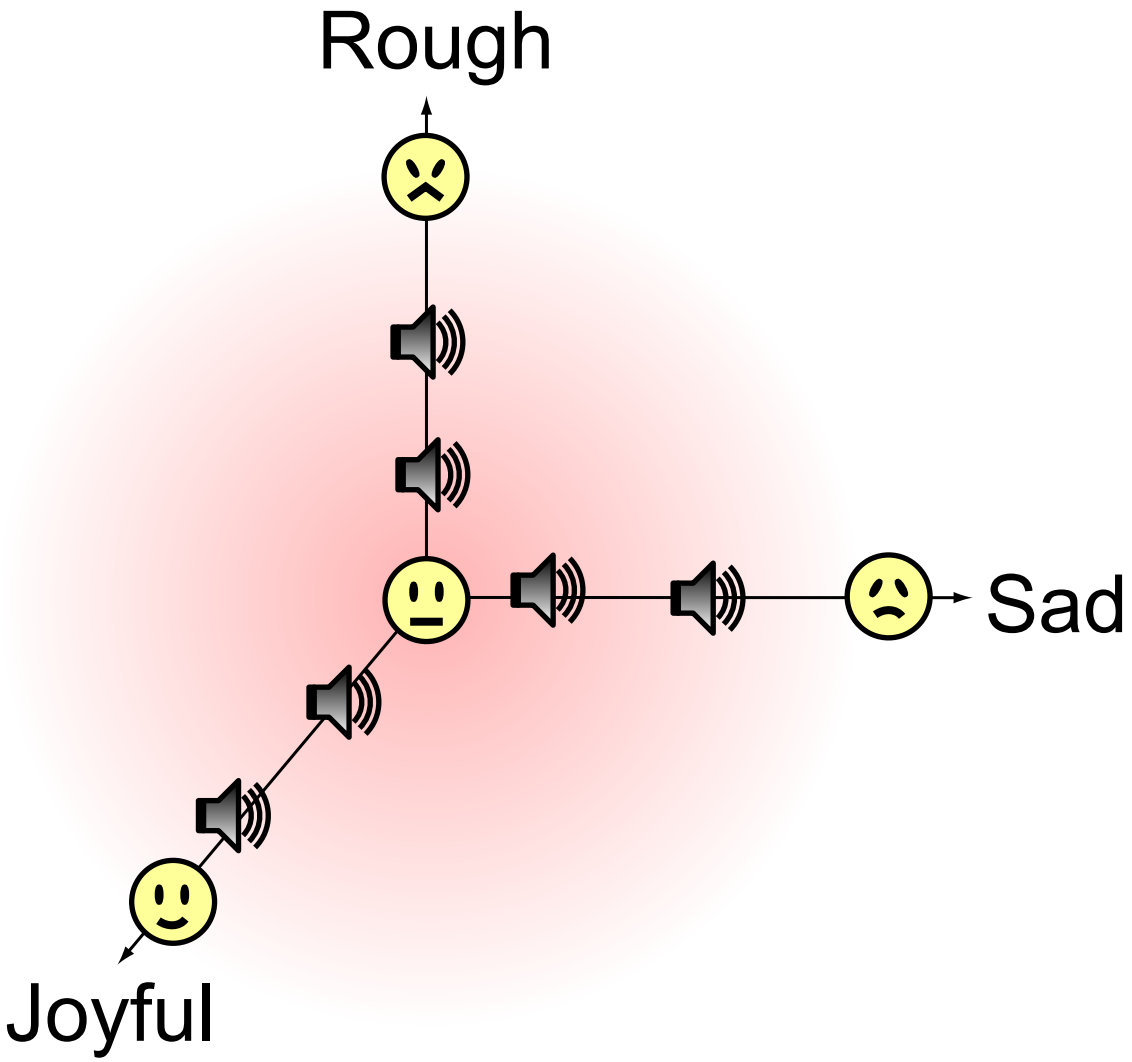
$\boldsymbol{\mu}_j$: mean vector of j-th state

$\boldsymbol{M}_j$: regression matrix

$\boldsymbol{\xi}$: extended control vector

Multiple-regression demo

Style-control



From http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/demo.html

Time-line

16:15 ~ 17:45: Second half

- Related topics
  * Unit selection & hybrid
  * Flexibility to control voice characteristics
    · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  * Vocoding
  * Acoustic modeling
  * Over-smoothing compensation
- Applications
- Q&A (10min)

Drawbacks of HMM-based speech synthesis

- **The biggest drawback of HTS is quality**
- **Three major factors**
  - Poor vocoding
    * How to *parameterize* speech waveform?
  - Inaccurate acoustic modeling
    * How to *model* extracted speech parameter trajectories?
  - Over-smoothing
    * How to *recover* generated speech parameter trajectories?

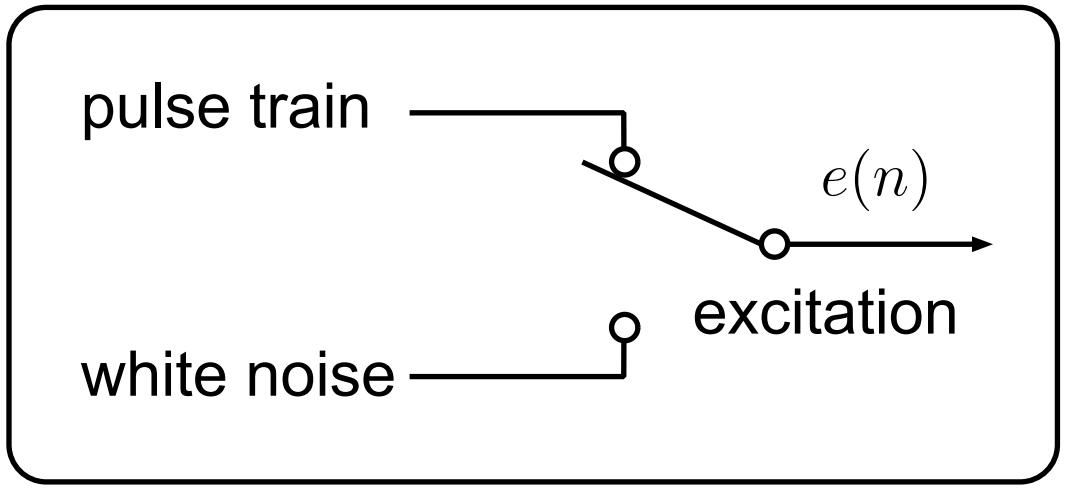
Recent improvements drastically have enhanced the quality of HMM-based speech synthesis

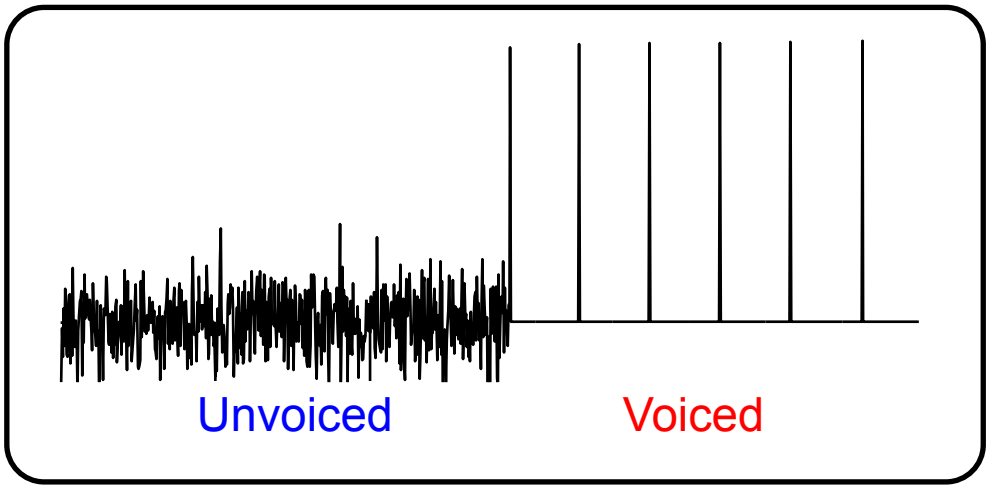
Time-line

16:15 ~ 17:45: Second half

- Related topics
  - * Unit selection & hybrid
  - * Flexibility to control voice characteristics
    - · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  - * Vocoding
  - * Acoustic modeling
  - * Over-smoothing compensation
- Applications
- Q&A (10min)

Excitation model



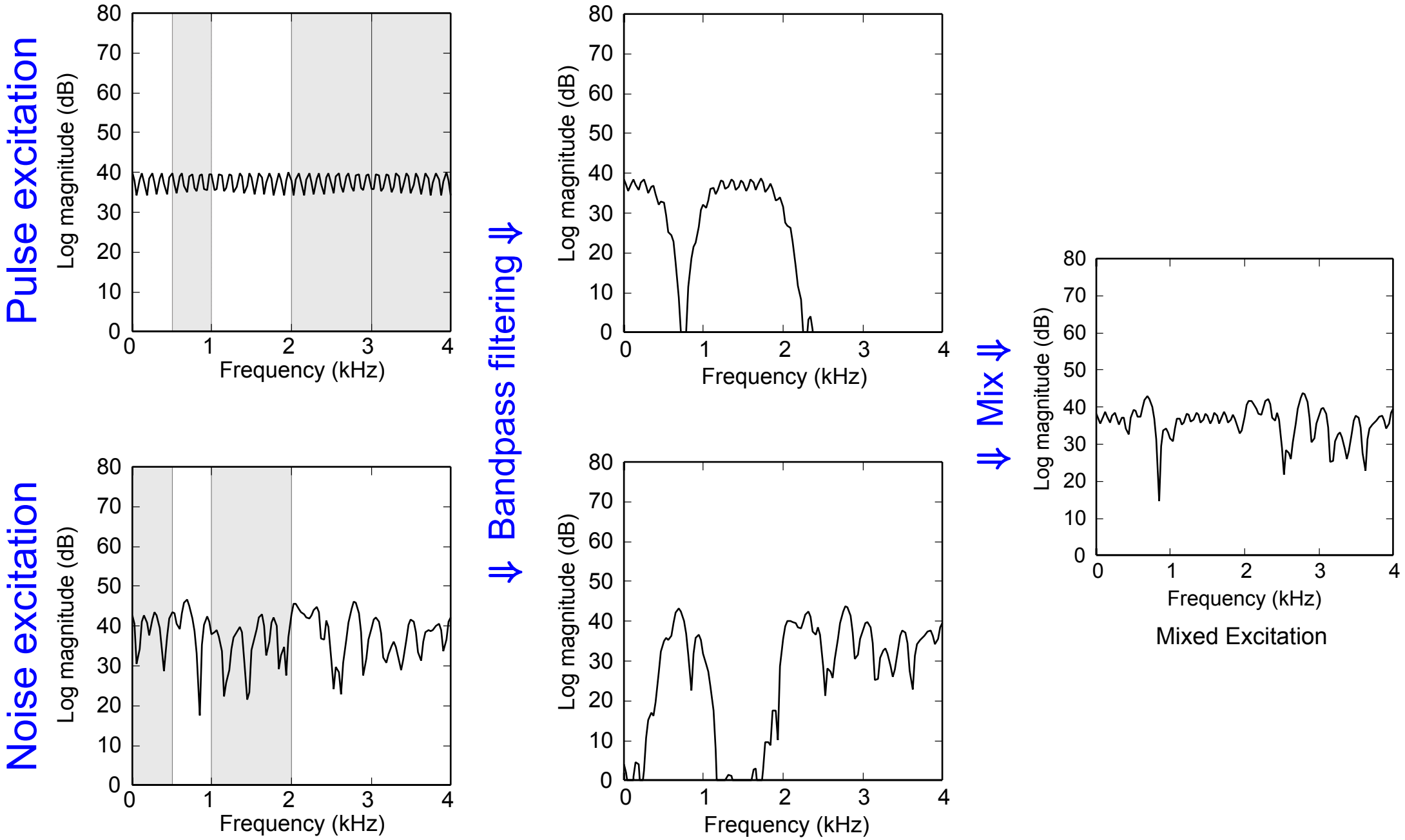


- **Excitation model: pulse/noise excitation**
  - voiced (periodic) → pulse trains
  - unvoiced (aperiodic) → white noise
- Difficult to model mix of V/UV excitations (e.g., V. fricatives)
- Synthesized speech sounds buzzy

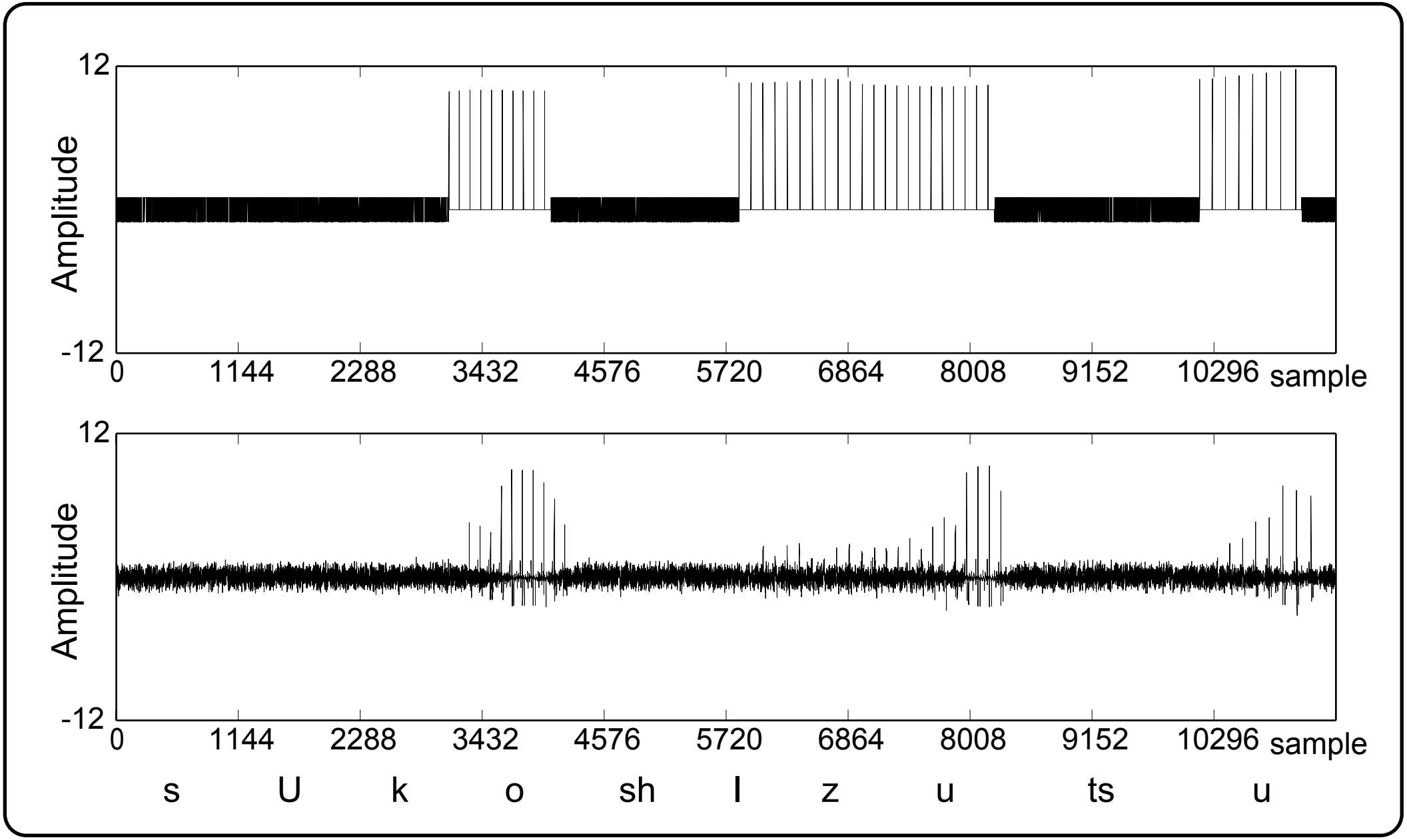
Advanced vocoders & excitation models

- MELP-style mixed excitation [Yoshimura;'01]
- STRAIGHT [Kawahara;'99, Zen;'07]
- Multi-band excitation [Abdel-Hamid;'06]
- Harmonic plus noise model (HNM) [Hemptinne;'06]
- Harmonic / stochastic model [Banos;'08]
- Glottal-flow derivative model (LF model) [Cabral;'07]
- ML excitation [Maia;'07]
- Glottal waveform [Raitio;'08]
- Residual codebook [Drugman;'09]

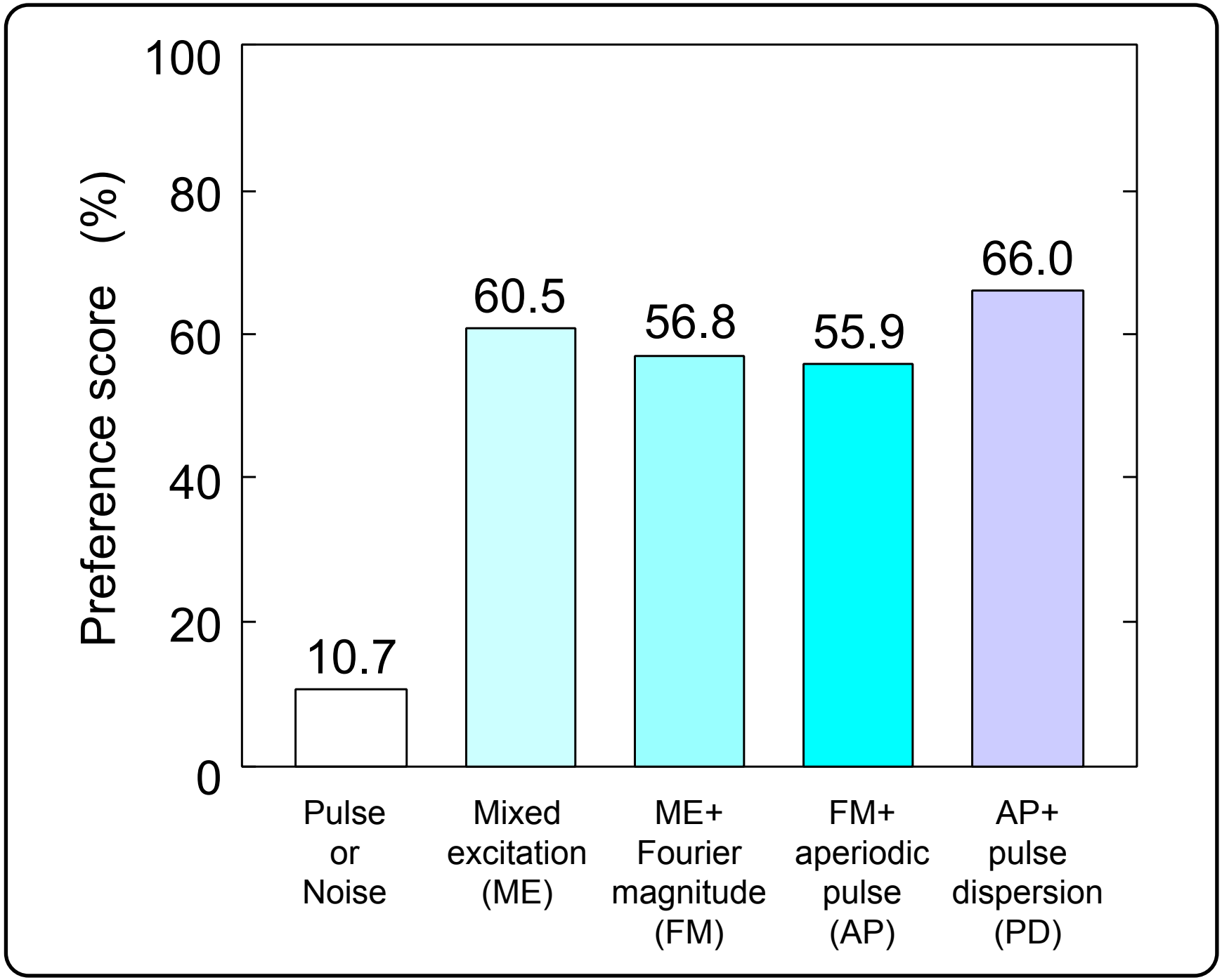
MELP-style mixed excitation (1)



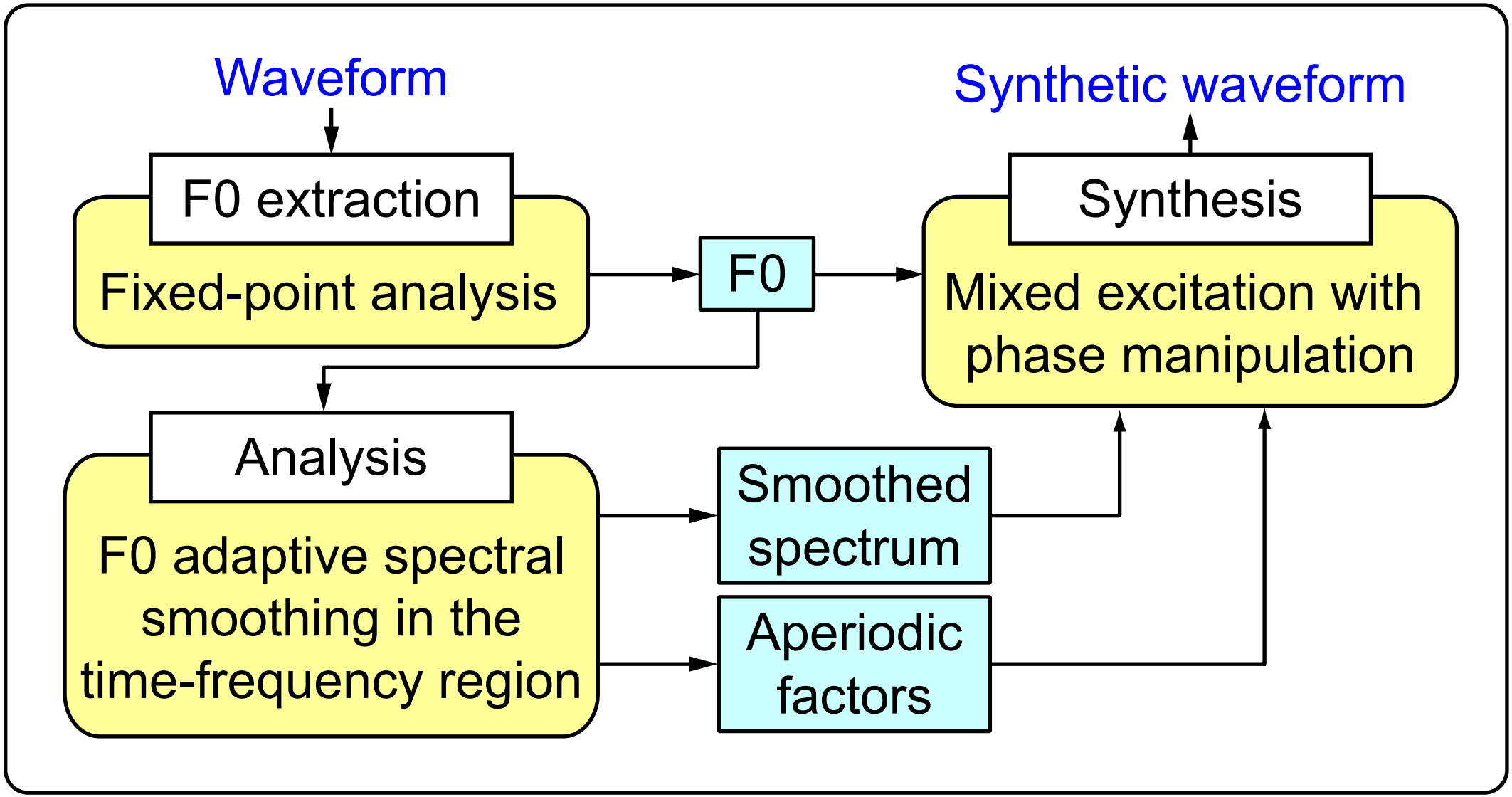
MELP-style mixed excitation (2)



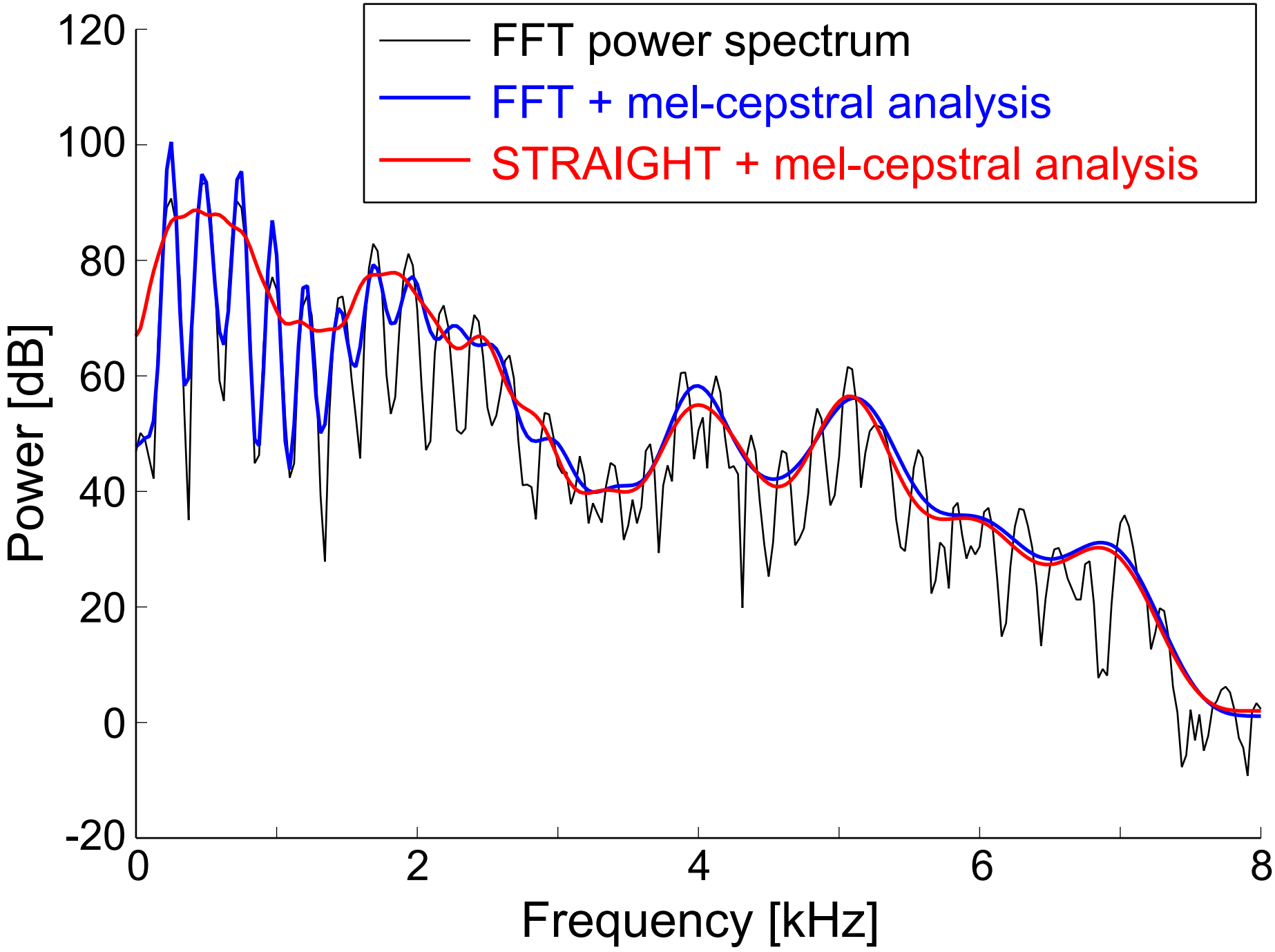
MELP-style mixed excitation (3)



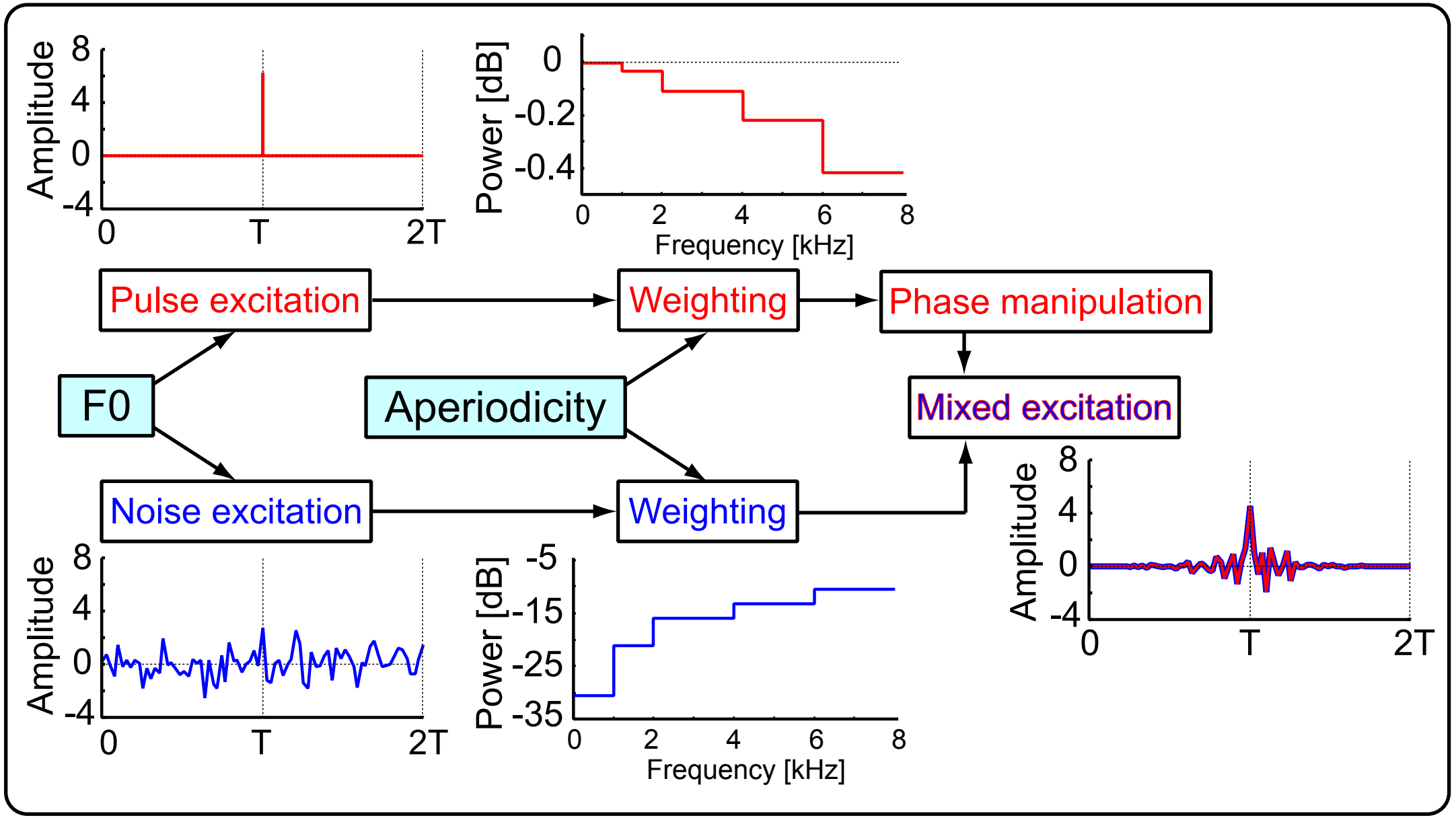
STRAIGHT


Waveform
F0 extraction
Fixed-point analysis
F0
Synthetic waveform
Synthesis
Mixed excitation with phase manipulation
Analysis
F0 adaptive spectral smoothing in the time-frequency region
Smoothed spectrum
Aperiodic factors

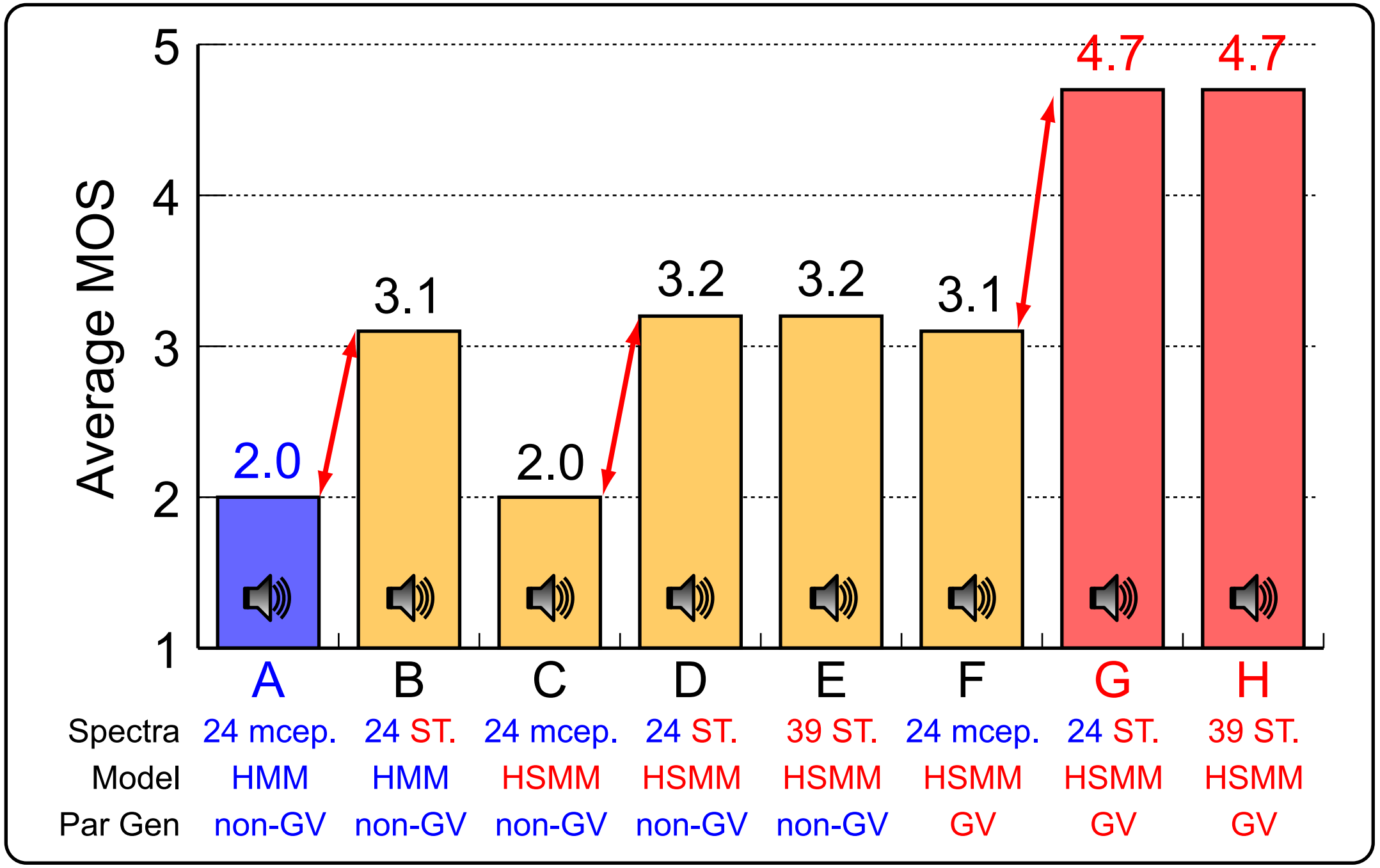
STRAIGHT + ML-based mel-cepstral analysis



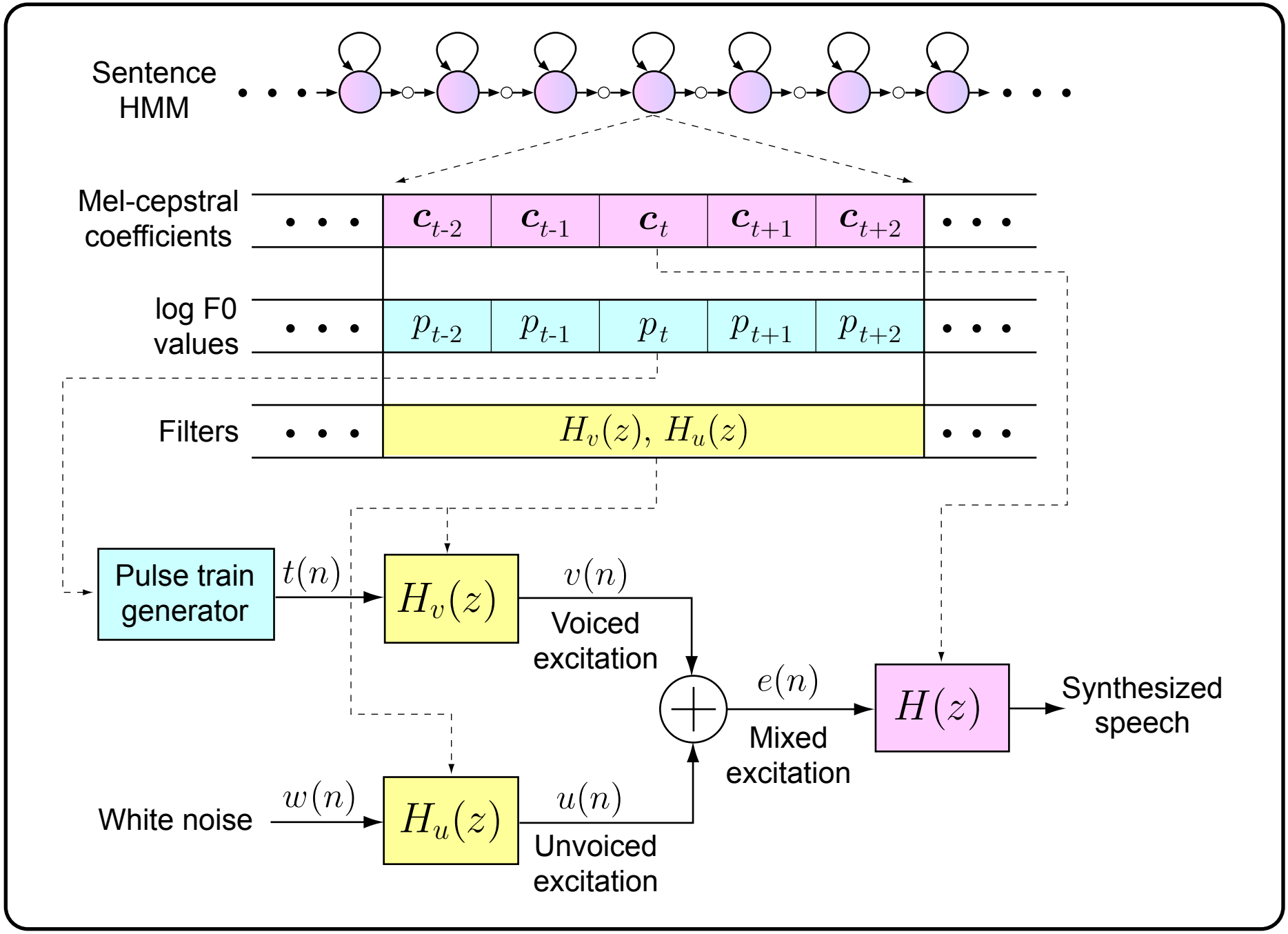
Excitation signal generation in STRAIGHT



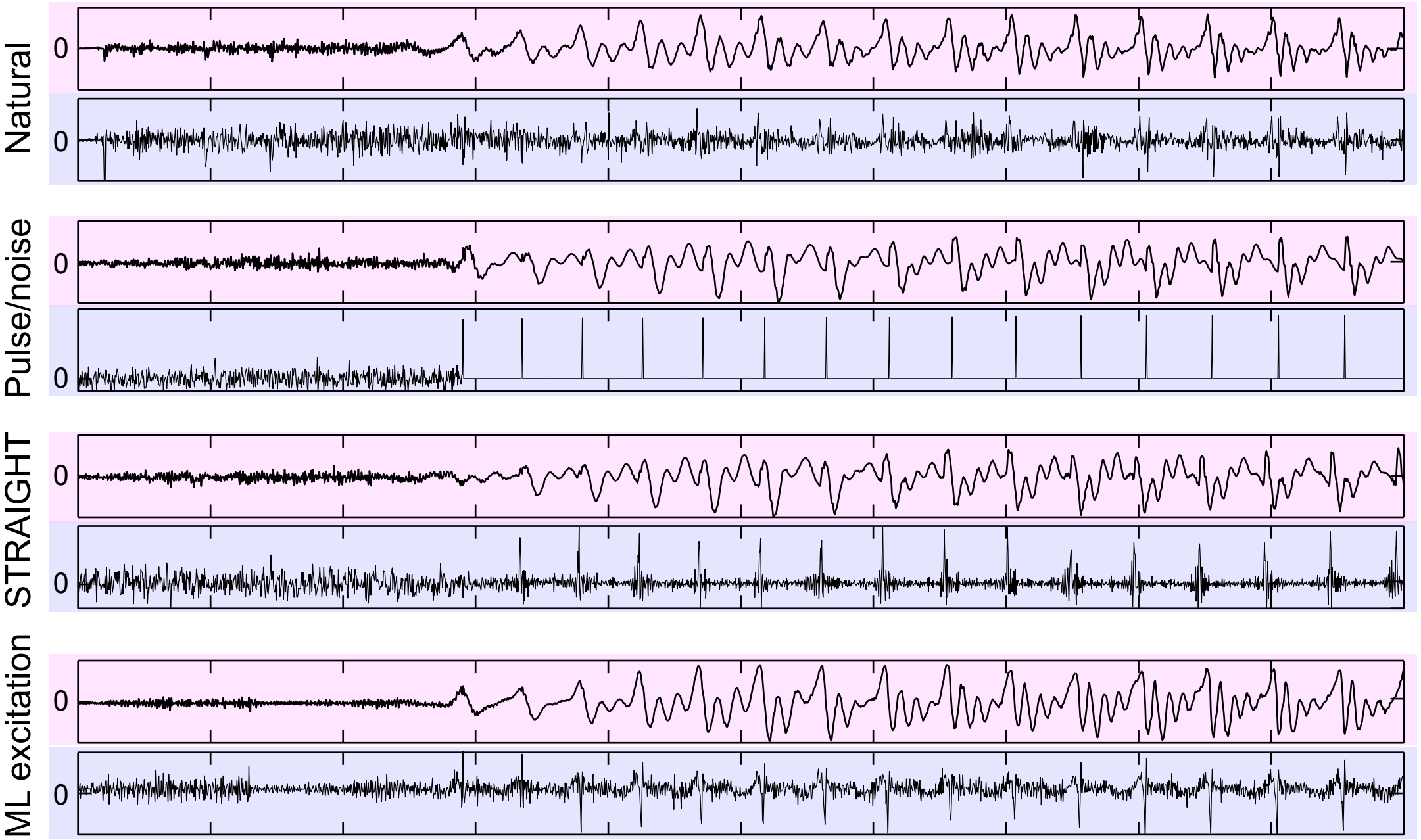
Effect of STRAIGHT [Zen;'07]



ML excitation [Maia;'07]



Examples of excitation signals



Time-line

16:15 ~ 17:45: Second half

- Related topics
  - * Unit selection & hybrid
  - * Flexibility to control voice characteristics
    - · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  - * Vocoding
  - * Acoustic modeling
  - * Over-smoothing compensation
- Applications
- Q&A (10min)

Acoustic modeling research

- **Improve accuracy of $p(O \mid \mathcal{W}, \boldsymbol{\lambda})$**
  - Trended HMM [Dines;'01]
  - Polynomial segment model [Sun;'09]
  - Buried Markov model (BMM) [Bulyko;'02]
  - AR-HMM [Shannon;'09]
  - HSMM [Zen;'04]
  - Trajectory HMMs [Zen;'04]
  - Minimum generation error (MGE) training [Wu;'06]
- **Relax approximations**
  - Bayesian approach [Hashimoto;'09]
  - MGE log-spectral distortion (MGE-LSD) [Wu;'08]
  - Statistical vocoder (STAVOCO) [Toda;'08]
  - Joint front-end/back-end training [Oura;'08]

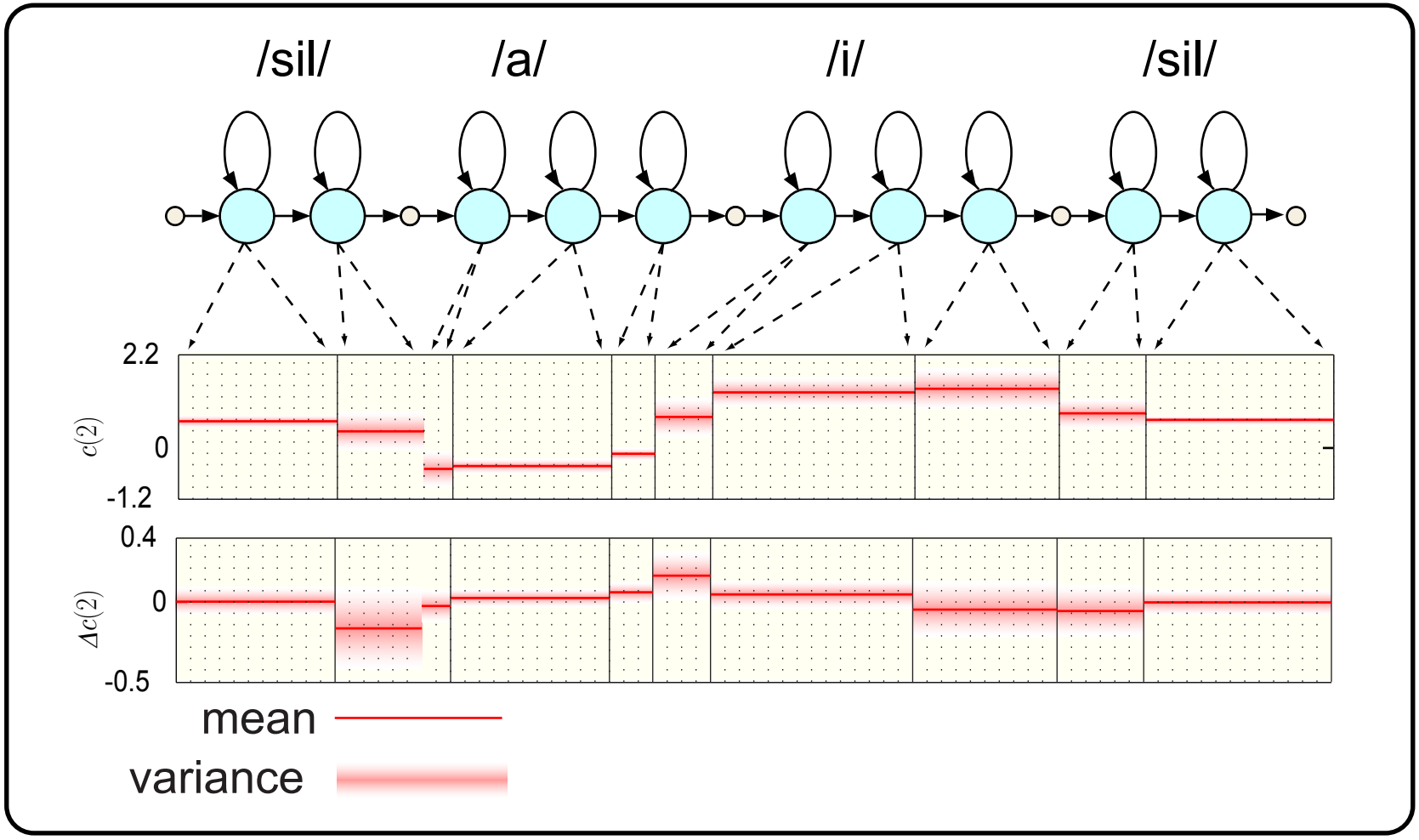
Acoustic modeling research

- **Improve accuracy of $p(O \mid \mathcal{W}, \boldsymbol{\lambda})$**
  - Trended HMM [Dines;'01]
  - Polynomial segment model [Sun;'09]
  - Buried Markov model (BMM) [Bulyko;'02]
  - AR-HMM [Shannon;'09]
  - HSMM [Zen;'04]
  - Trajectory HMMs [Zen;'04]
  - Minimum generation error (MGE) training [Wu;'06]
- **Relax approximations**
  - Bayesian approach [Hashimoto;'09]
  - MGE log-spectral distortion (MGE-LSD) [Wu;'08]
  - Statistical vocoder (STAVOCO) [Toda;'08]
  - Joint front-end/back-end training [Oura;'08]

Acoustic modeling

Limitations of HMMs for modeling speech

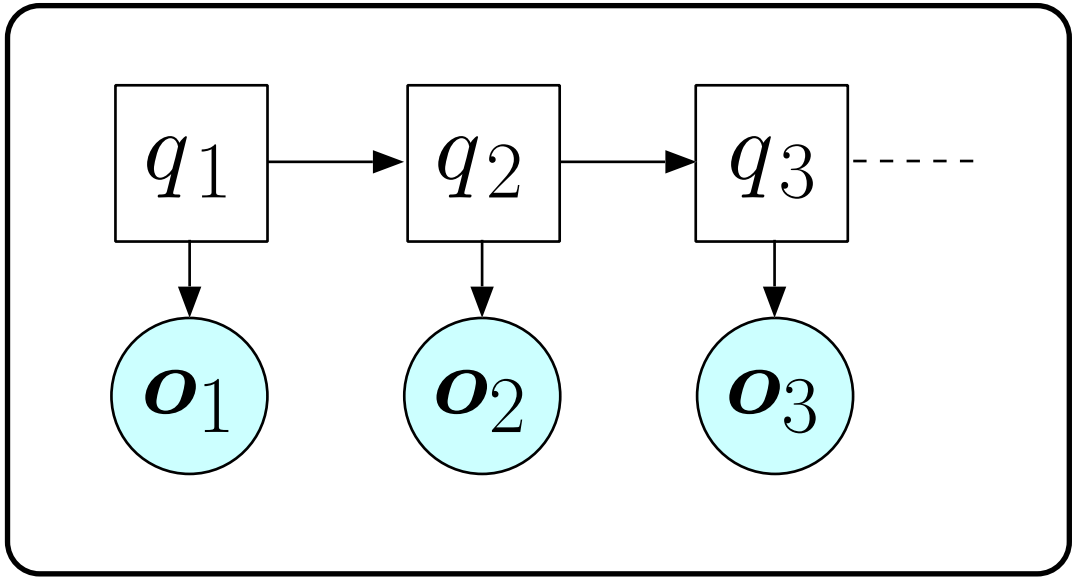
- **Piece-wise constant statistics**
  - Statistics do not vary within an HMM state



Acoustic modeling

Limitations of HMMs for modeling speech

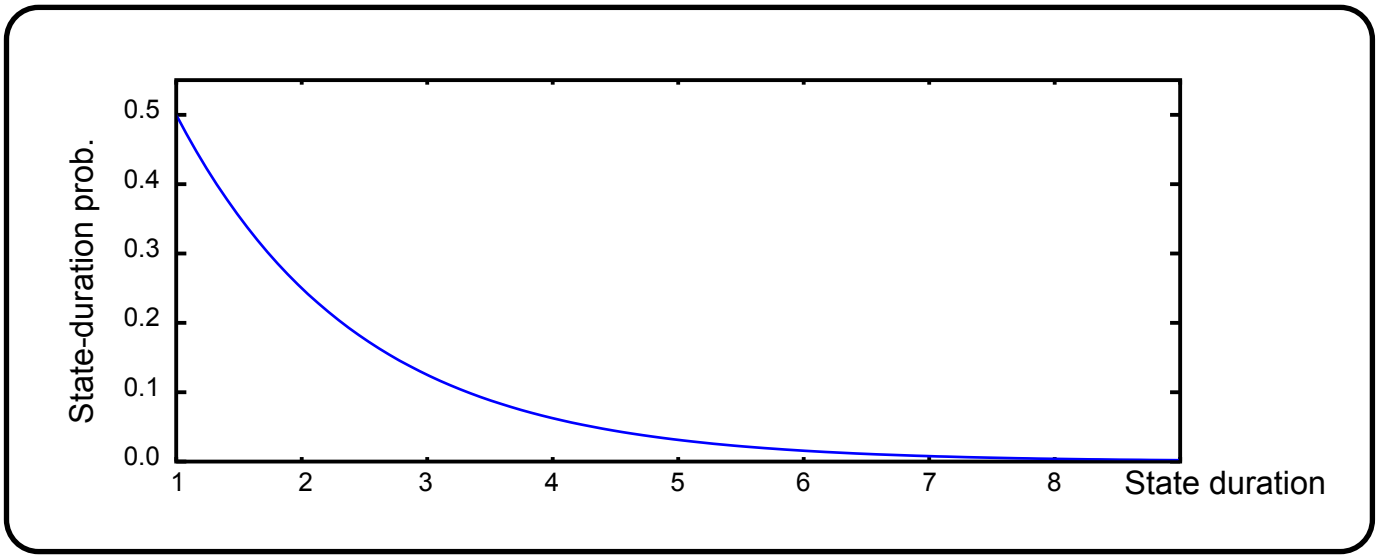
- **Piece-wise constant statistics**
  - Statistics do not vary within an HMM state
- **Frame-wise conditional independence assumption**
  - State output prob. depends only on the current state



Acoustic modeling

Limitations of HMMs for modeling speech

- **Piece-wise constant statistics**
  - Statistics do not vary within an HMM state
- **Frame-wise conditional independence assumption**
  - State output prob. depends only on the current state
- **Weak duration modeling**
  - State duration prob. decreases exponentially with time



Acoustic modeling

Limitations of HMMs for modeling speech

- **Piece-wise constant statistics**
  - Statistics do not vary within an HMM state
- **Frame-wise conditional independence assumption**
  - State output prob. depends only on the current state
- **Weak duration modeling**
  - State duration prob. decreases exponentially with time

None of them hold for real speech

Speech params. are generated from acoustic models

⇒ Better acoustic model may produce better speech

Better acoustic modeling (1)

Advanced acoustic model

- **Piece-wise constant statistics ⇒ Dynamical model**
  - Capture dynamics of speech parameter trajectory
    * Trended HMM [Dines;'01]
    * Polynomial segment model [Sun;'09]
- **Conditional indep. assumption ⇒ Graphical model**
  - Additional dependency between variables
    * Buried Markov model [Bulyko;'02]
    * Autoregressive HMM [Shannon;'09]
- **Weak duration modeling ⇒ Explicit duration model**
  - State duration PDFs are explicitly modeled
    * Hidden semi-Markov model [Zen;'07]

Better acoustic modeling (2)

Trajectory HMM [Zen;'06]

- Derived from HMM with dynamic feature constraints
- Underlying generative model of HMM-based synthesis

$$P(\boldsymbol{c} \mid \boldsymbol{\lambda}) = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q} \mid \boldsymbol{\lambda})$$

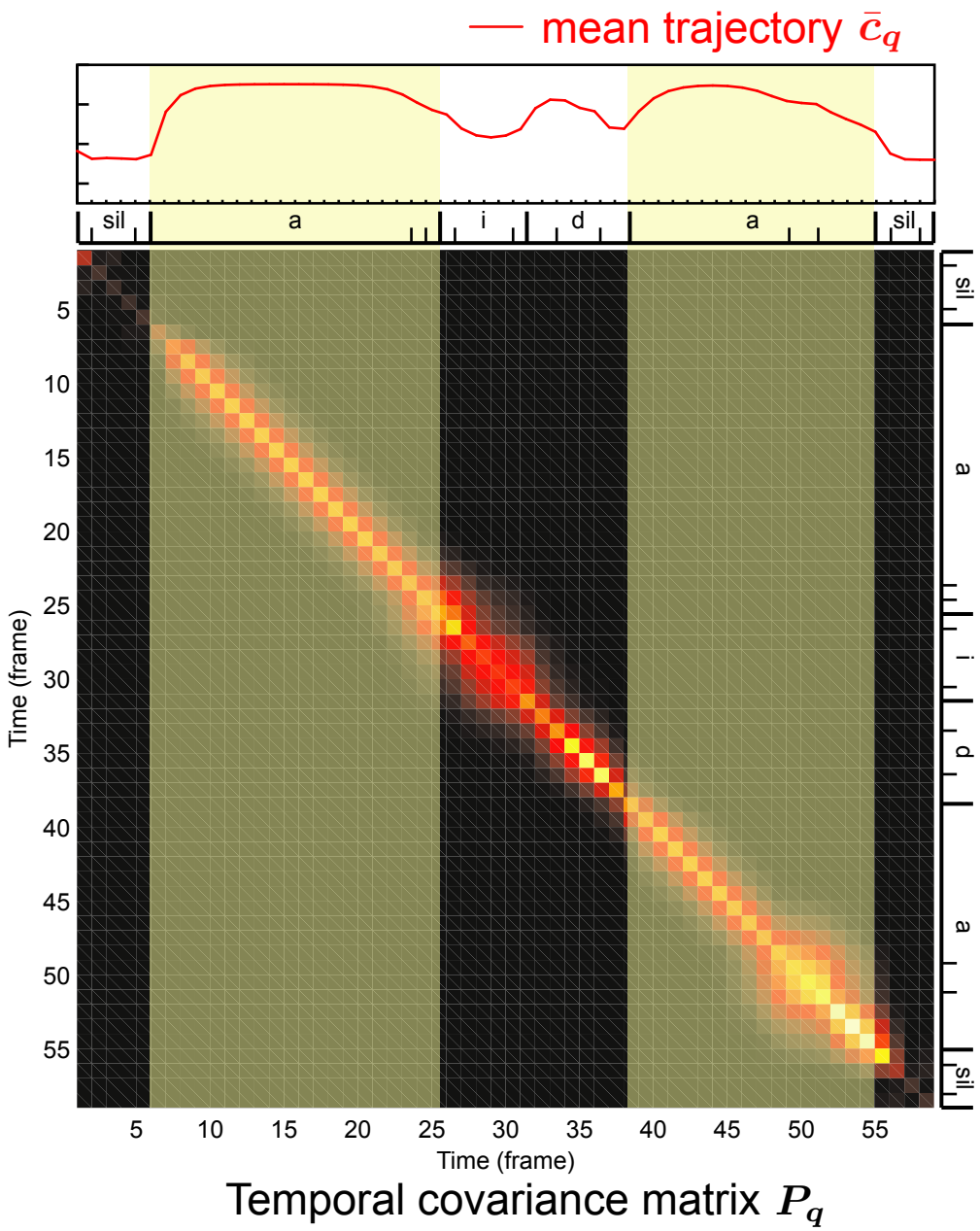
$$P(\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{c} \mid \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}})$$

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} = \boldsymbol{P}_{\boldsymbol{q}}^{-1}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}}$$

$$\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{P}_{\boldsymbol{q}} \boldsymbol{r}_{\boldsymbol{q}}$$

Better acoustic modeling (3)



Temporal covariance matrix P_q

$\bar{c}_q$ is given as a *smooth trajectory*

⇒ Statistics vary within a state

Covariance matrix P_q is *full*

⇒ All state output probabilities depend on each other

Limitations of HMMs can be avoided

Statistics vary according to their neighboring models

⇒ Coarticulation effect can naturally be captured

Better acoustic modeling (4)

Relation between HMM-based synth. & trajectory HMM

- Mean vector of trajectory HMM, $\bar{c}_q$

$$W^\top \Sigma_q^{-1} W \bar{c}_q = W^\top \Sigma_q^{-1} \mu_q$$

- ML trajectory by speech param. generation algorithm, c

$$W^\top \Sigma_q^{-1} W c = W^\top \Sigma_q^{-1} \mu_q$$

$\bar{c}_q$ and c are identical

Better acoustic modeling (4)

AM training of trajectory HMM

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \{p(\boldsymbol{C} \mid \mathcal{W}, \boldsymbol{\lambda})\} : \text{ML [Zen;'06]}$$

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \{\mathcal{E}(\boldsymbol{C}; \mathcal{W}, \boldsymbol{\lambda})\} : \text{MMSE (MGE) [Wu;'06]}$$

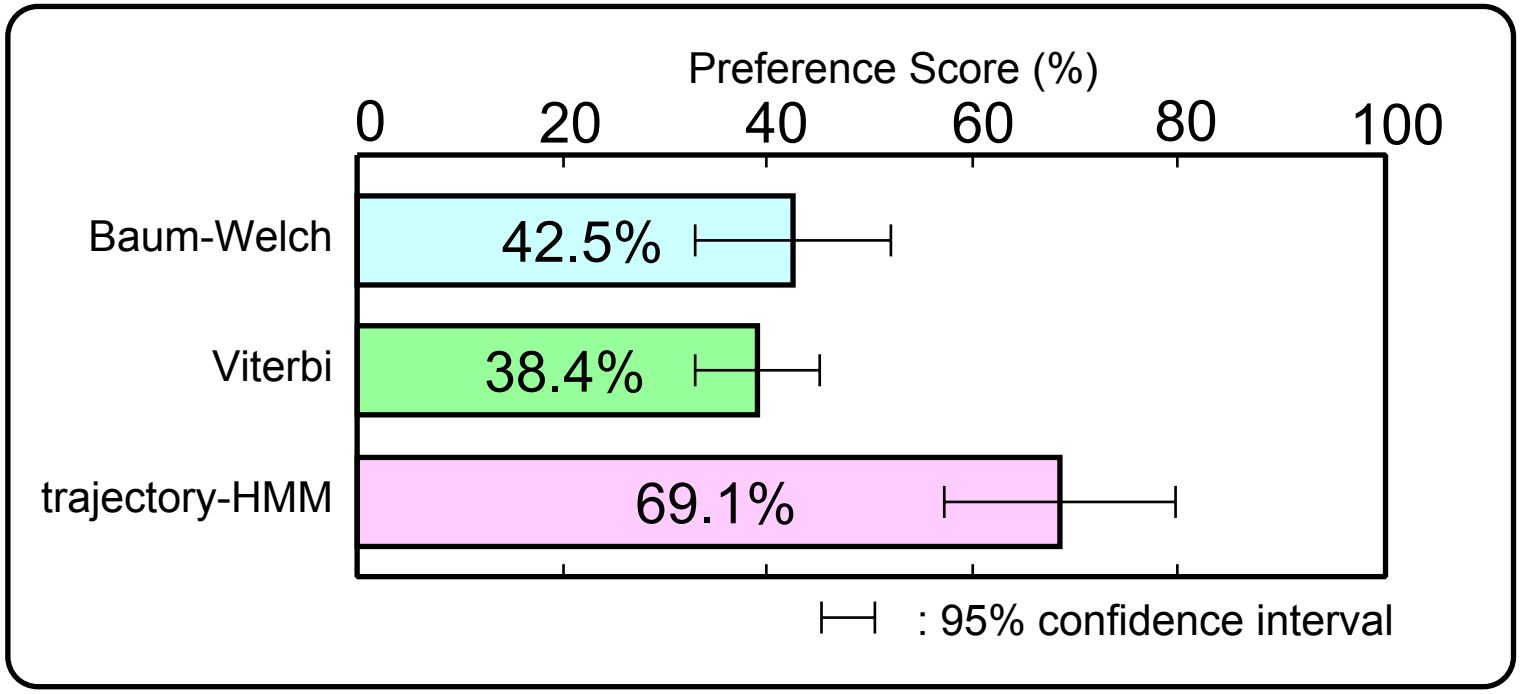
Minimize error betw. training & generated trajectories

	Conventional	trajectory HMM
Training	$p(\boldsymbol{O} \mid \mathcal{W}, \boldsymbol{\lambda})$	$p(\boldsymbol{C} \mid \mathcal{W}, \boldsymbol{\lambda})$
Synthesis	$p(\boldsymbol{c} \mid w, \hat{\boldsymbol{\lambda}})$	$p(\boldsymbol{c} \mid w, \hat{\boldsymbol{\lambda}})$

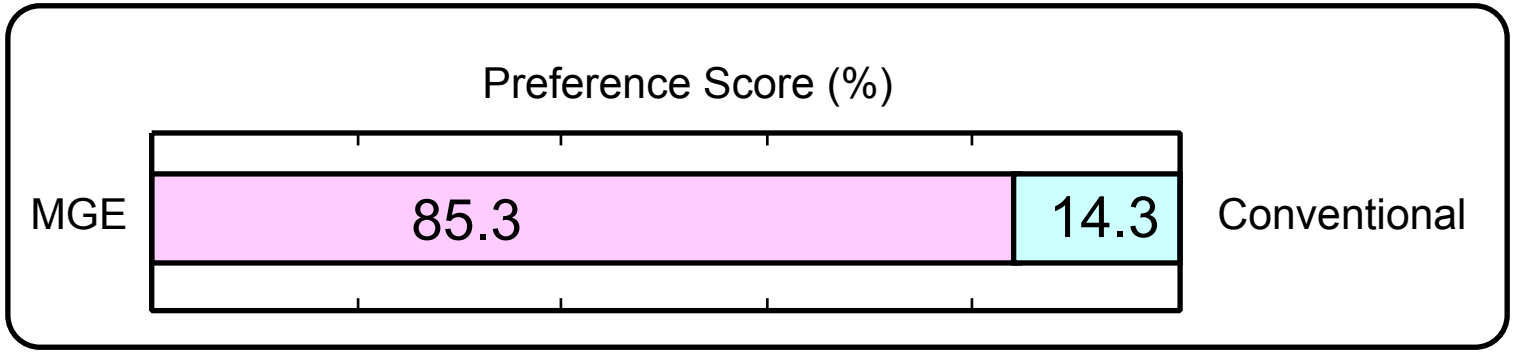
Solve inconsistency between training & synthesis

Better acoustic modeling (5)

Effect of trajectory HMM [Zen;'07]



Effect of MGE [Wu;'06]



Acoustic modeling research

- **Improve accuracy of $p(O \mid \mathcal{W}, \boldsymbol{\lambda})$**
  - Trended HMM [Dines;'01]
  - Polynomial segment model [Sun;'09]
  - Buried Markov model (BMM) [Bulyko;'02]
  - AR-HMM [Shannon;'09]
  - HSMM [Zen;'04]
  - Trajectory HMMs [Zen;'04]
  - Minimum generation error (MGE) training [Wu;'06]
- **Relax approximations**
  - Bayesian approach [Hashimoto;'09]
  - MGE log-spectral distortion (MGE-LSD) [Wu;'08]
  - Statistical vocoder (STAVOCO) [Toda;'08]
  - Joint front-end/back-end training [Oura;'08]

Bayesian speech synthesis [Hashimoto;'09]

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} p(\hat{\boldsymbol{O}} \mid \mathcal{W}, \boldsymbol{\lambda})$$

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid w, \hat{\boldsymbol{\lambda}})$$

$$\Downarrow$$

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} \int p(\boldsymbol{o} \mid w, \boldsymbol{\lambda}) p(\boldsymbol{O} \mid \mathcal{W}, \boldsymbol{\lambda}) p(\lambda) d\boldsymbol{\lambda}$$

- ML estimation (point estimate, $\hat{\boldsymbol{\lambda}}$)
  - → Bayesian estimation (distribution, $p(\boldsymbol{\lambda} \mid \boldsymbol{O}, \mathcal{W})$)
- Exact Bayesian is difficult due to integral over $\boldsymbol{\lambda}$
  - → Variational Bayes (VB) [Beal;'03]

STAVOCO & MGE-LSD

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \; p(\boldsymbol{X}, \boldsymbol{O} \mid \mathcal{W}, \boldsymbol{\lambda})$$

$$= \arg\max_{\boldsymbol{\lambda}} \; p(\boldsymbol{X} \mid \boldsymbol{O}) p(\boldsymbol{O} \mid \mathcal{W}, \boldsymbol{\lambda})$$

- **STAVOCO** [Toda;'08]
  - Developed for spectral analysis, but applicable to synthesis
  - O: (Mel) Cepstrum, X: harmonic components in log spectrum
  - $O \rightarrow X$: stochastic linear transform
- **MGE-LSD** [Wu;'08]
  - O: (Mel) LSP, X: harmonic components in log spectrum
  - $O \rightarrow X$: deterministic non-linear mapping

Joint front-end / back-end model training (1)

Introduce context-dependent labels l & L

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \sum_{\boldsymbol{L}} p(\boldsymbol{O}, \boldsymbol{L} \mid \mathcal{W}, \boldsymbol{\lambda})$$

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} \sum_{\boldsymbol{l}} p(\boldsymbol{o}, \boldsymbol{l} \mid w, \boldsymbol{\lambda})$$

Approximate summation by max

$$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} p(\boldsymbol{O}, \boldsymbol{L} \mid \mathcal{W}) \quad \rightarrow \quad \hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} p(\boldsymbol{O} \mid \hat{\boldsymbol{L}}, \boldsymbol{\lambda})$$

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} p(\boldsymbol{l} \mid w) \quad \rightarrow \quad \hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \boldsymbol{\lambda})$$

front-end (word→label) back-end (training/synthesis)

Joint front-end / back-end model training (2)

Introduce front-end model Λ to predict label from text

$$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} p(\boldsymbol{O}, \boldsymbol{L} \mid \mathcal{W})$$ manual/automatic annotation

$$\hat{\boldsymbol{\Lambda}} = \arg\max_{\boldsymbol{\Lambda}} p(\hat{\boldsymbol{L}} \mid \mathcal{W}, \boldsymbol{\Lambda})$$ front-end model training

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} p(\hat{\boldsymbol{l}} \mid w, \boldsymbol{\Lambda})$$ label prediction

1. Manually annotate prosodic labels of training data
2. Train prosodic label prediction models (e.g., decision trees)
3. Predict prosodic labels of input text to be synthesized

Joint front-end / back-end model training (3)

$$\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\Lambda}}\} = \arg\max_{\boldsymbol{\lambda},\boldsymbol{\Lambda}} \sum_{\boldsymbol{L}} p(\boldsymbol{O} \mid \boldsymbol{L}, \boldsymbol{\lambda}) p(\boldsymbol{L} \mid \mathcal{W}, \boldsymbol{\Lambda})$$

Joint front-end / back-end model training [Oura;08]

- Label sequence: regarded as latent variable & marginalized

  → Robust against labelling errors

- Front-end & back-end models are simulteneously estimated

  → Combine front-end/back-end models as a unified model

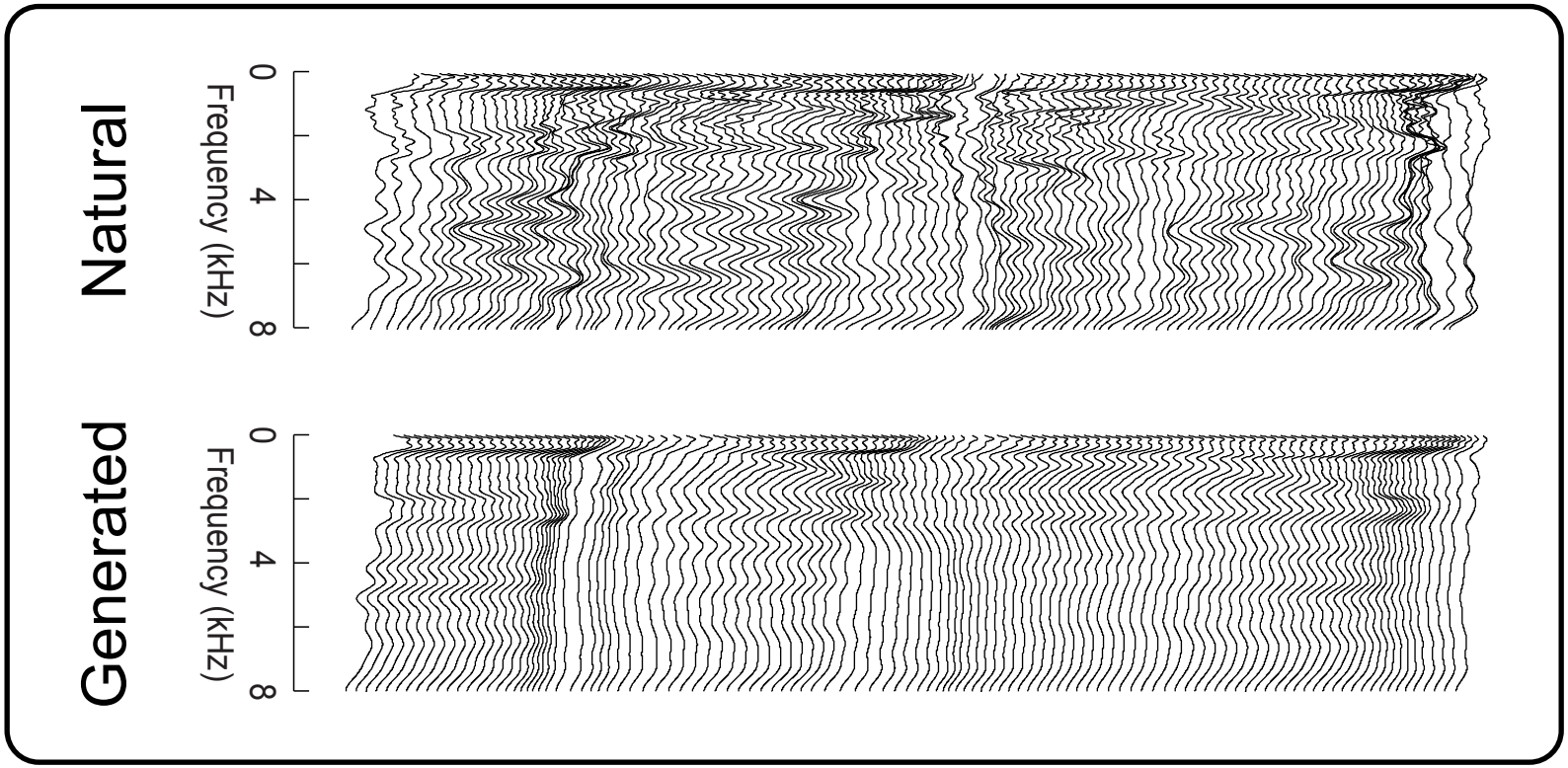
Time-line

16:15 ~ 17:45: Second half

- Related topics
  * Unit selection & hybrid
  * Flexibility to control voice characteristics
    · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  * Vocoding
  * Acoustic modeling
  * Over-smoothing compensation
- Applications
- Q&A (10min)

Over-smoothing problem

- **Speech parameter generation algorithm** [Tokuda;'00]
  - Dynamic feature constraints make generated params. smooth
  - Sometimes too smooth ⇒ Sounds *muffled*



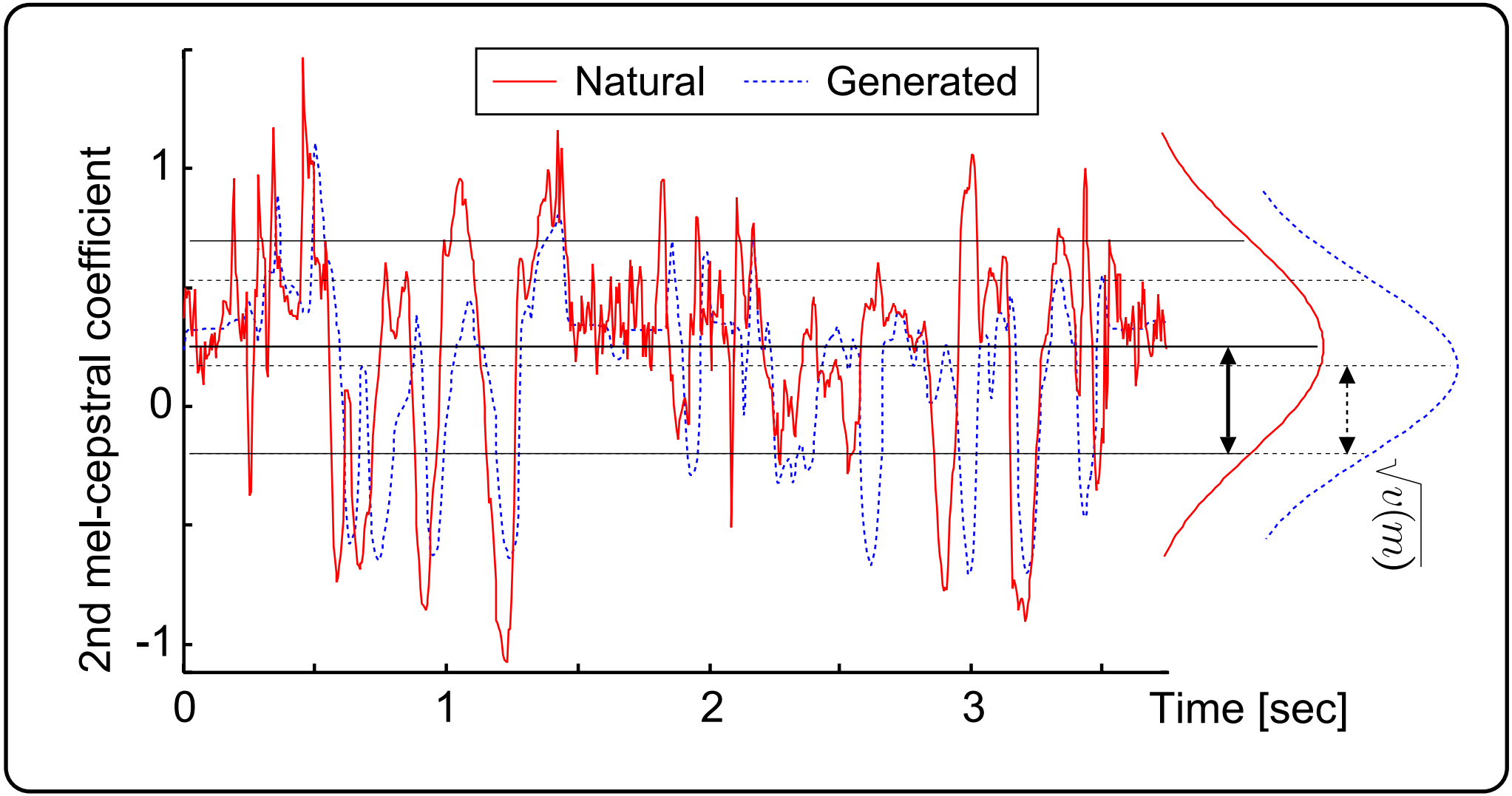
- **Why?**
  - Details of spectral (formant) structure are removed
  - Use of advanced AM relaxes this problem, but not enough

Compensate over smoothing (1)

- **Emphasize spectral structure by post-filtering**
  - Developed for speech coding
  - Reduce buzziness & muffling
- **Using training data explicitly in synthesis process**
  - Conditional parameter generation algorithm [Masuko;'03]
  - Discrete HMM-based speech synthesis [Yu;'07]
- **Combining multiple-level statistics**
  - Boosting-style additive F0 trees [Yao;'08]
  - DCT-based F0 models [Latorre;'08]
  - Intra-phoneme dynamics model [Tiomkin;'08]
  - Param. gen. algorithm considering global variance [Toda;'07]

Compensate over smoothing (2)

Speech param. gen. algorithm considering GV [Toda;'07]



Generated trajectory usually has smaller dynamic range

Compensate over smoothing (3)

Speech param. gen. algorithm considering GV [Toda;'07]

- Objective function of standard param. gen. algorithm

$$\mathcal{F}_{\mathrm{ML}}(\boldsymbol{c}\,;\boldsymbol{\lambda}) = \log \mathcal{N}(\boldsymbol{W}\boldsymbol{c}\,;\ \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}})$$

- Objective function of param. gen. algorithm considering GV

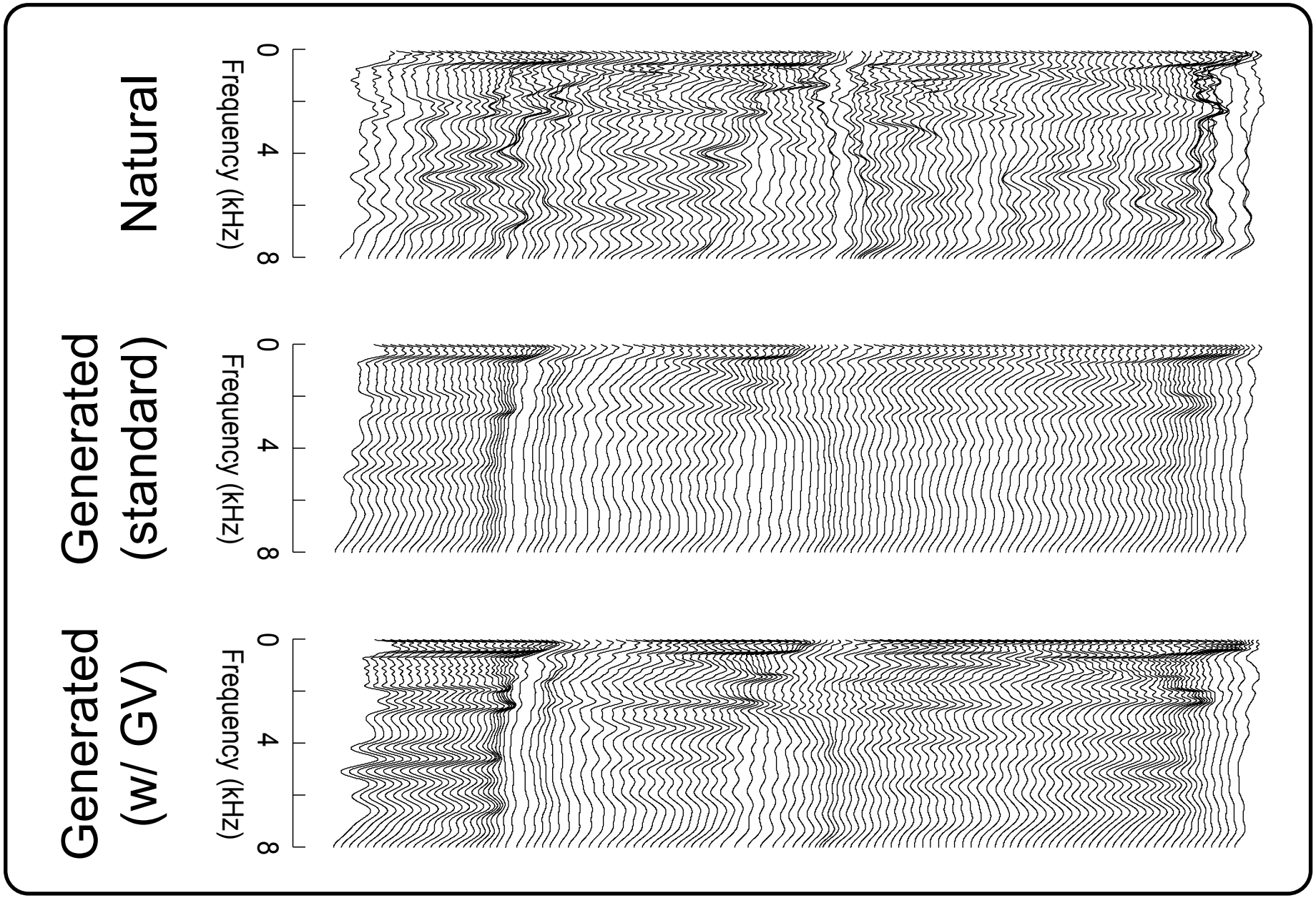
$$\mathcal{F}_{\mathrm{GV}}(\boldsymbol{c}\,;\boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{GV}}) = \omega \log \mathcal{N}(\boldsymbol{W}\boldsymbol{c}\,;\ \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{\Sigma}_{\boldsymbol{q}}) + \log \mathcal{N}(\boldsymbol{v}(\boldsymbol{c})\,;\ \boldsymbol{\mu}_{\mathrm{GV}}, \boldsymbol{\Sigma}_{\mathrm{GV}})$$

  - 1st term is the same as the above objective function
  - 2nd term works to keep dynamic range close to training data

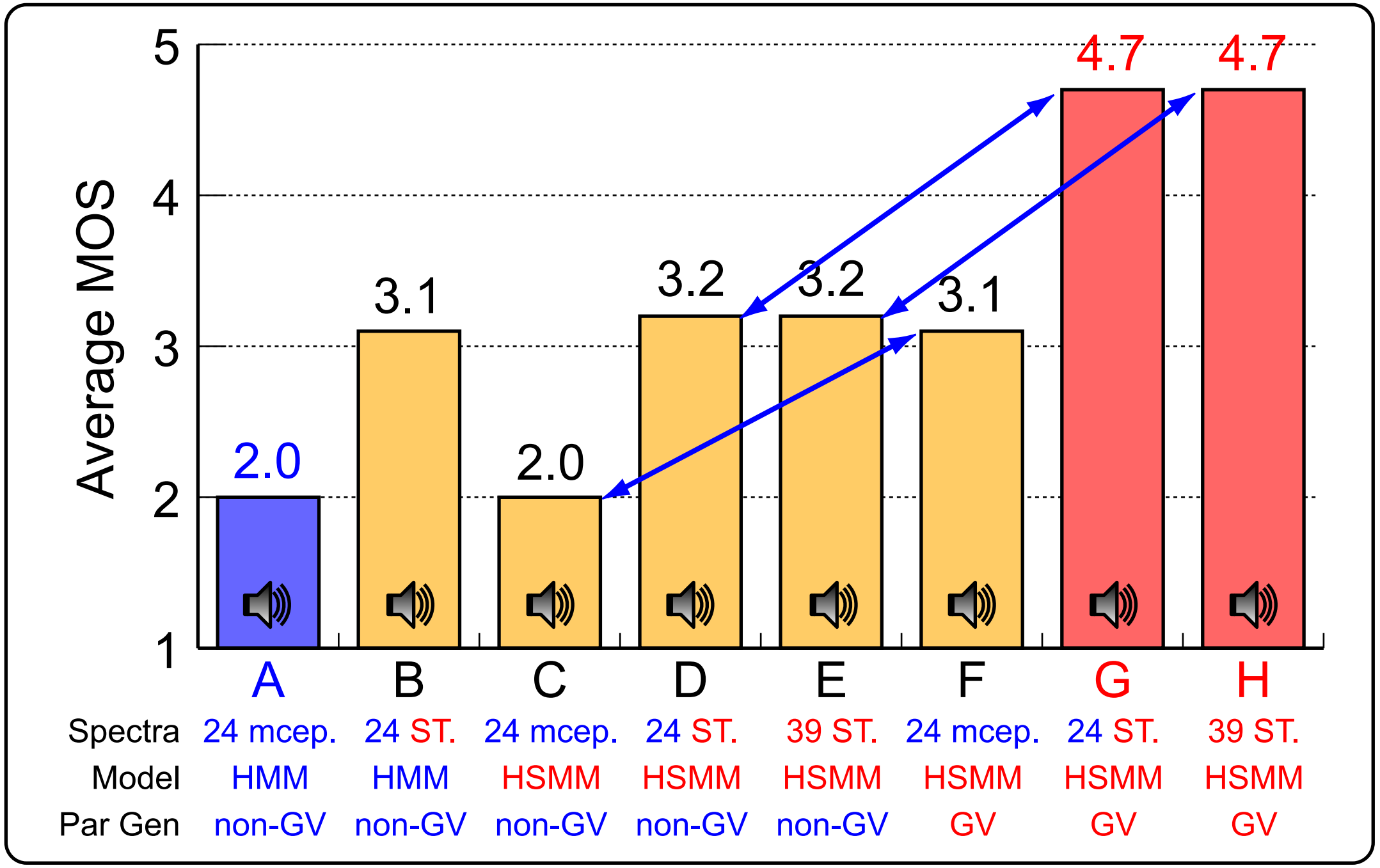
  $\Rightarrow$ Penalty for over-smoothing

Compensate over smoothing (4)

Effect of GV



Effect of GV [Zen;'07]



Compensate over smoothing (5)

Incorporate GV into training

- **MGE + GV** [Wu;07]

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left\{ \mathcal{E}\left(\boldsymbol{c}; \bar{\boldsymbol{c}}_{\boldsymbol{q}}\right) + \omega \mathcal{E}\left(v(\boldsymbol{c}); v(\bar{\boldsymbol{c}}_{\boldsymbol{q}})\right) \right\}$$

- **Trajectory HMM + GV** [Toda;09]

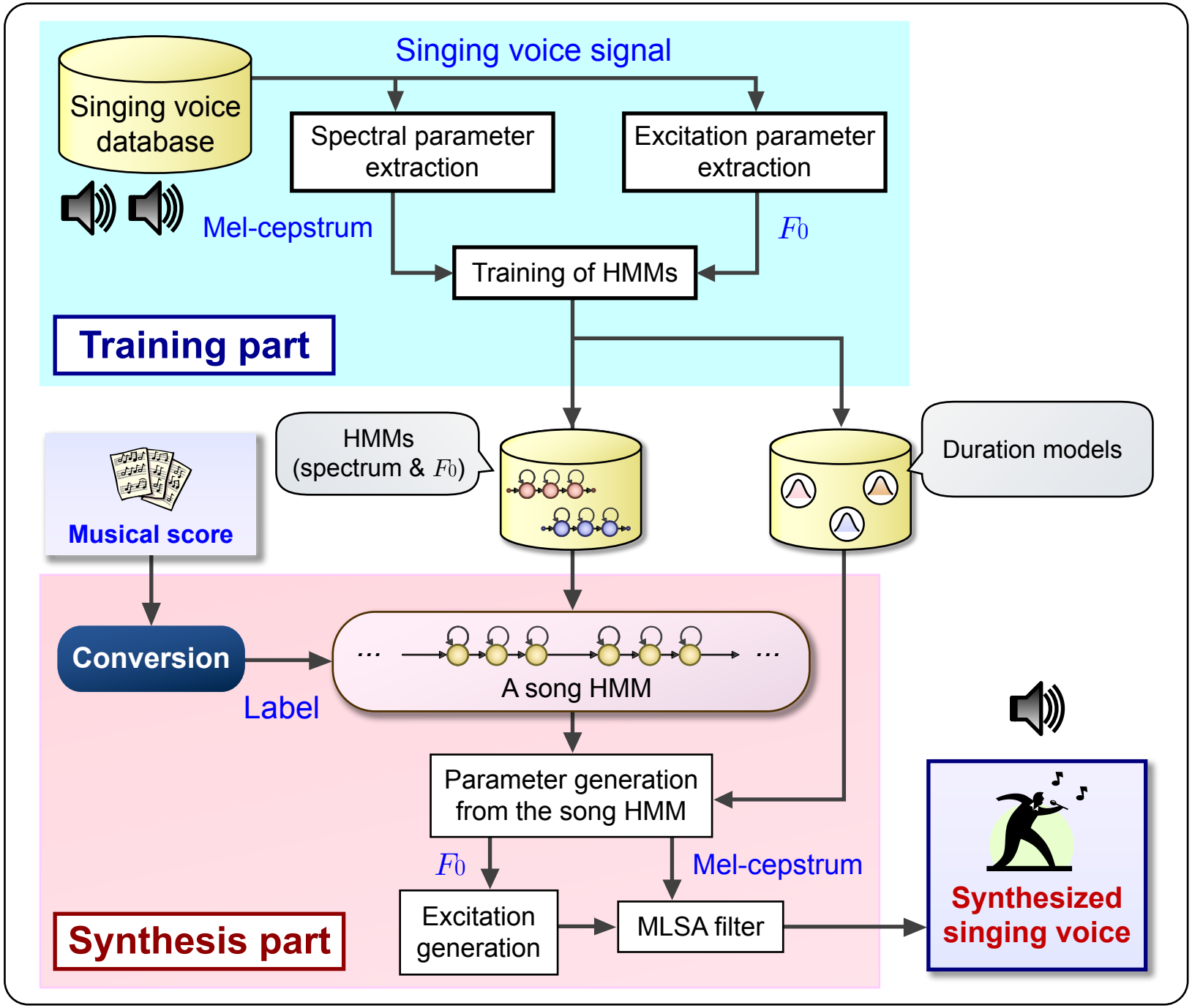
$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \left\{ \log \mathcal{N}\left(\boldsymbol{c}; \bar{\boldsymbol{c}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right) + \omega \log \mathcal{N}\left(v(\boldsymbol{c}); v(\bar{\boldsymbol{c}}_{\boldsymbol{q}}), \boldsymbol{\Sigma}_v\right) \right\}$$

Time-line

16:15 ~ 17:45: Second half

- Related topics
  - * Unit selection & hybrid
  - * Flexibility to control voice characteristics
    - · Adaptation, interpolation, eigenvoice, multiple regress.
- Recent advances
  - * Vocoding
  - * Acoustic modeling
  - * Over-smoothing compensation
- Applications
- Q&A (10min)

Singing voice synthesis [Saino;'06]



Emotional speech synthesis [Tsuzuki;04]

"Finally, I present a conclusion. We constructed emotional speech synthesizer based on HMM-based speech synthesis system. neutral

Furthermore, a modeling technique based on subjective evaluation was proposed. angry

Listening test results show that the proposed technique improves the expressiveness of emotions for synthesized speech, particularly, neutral and happiness. neutral

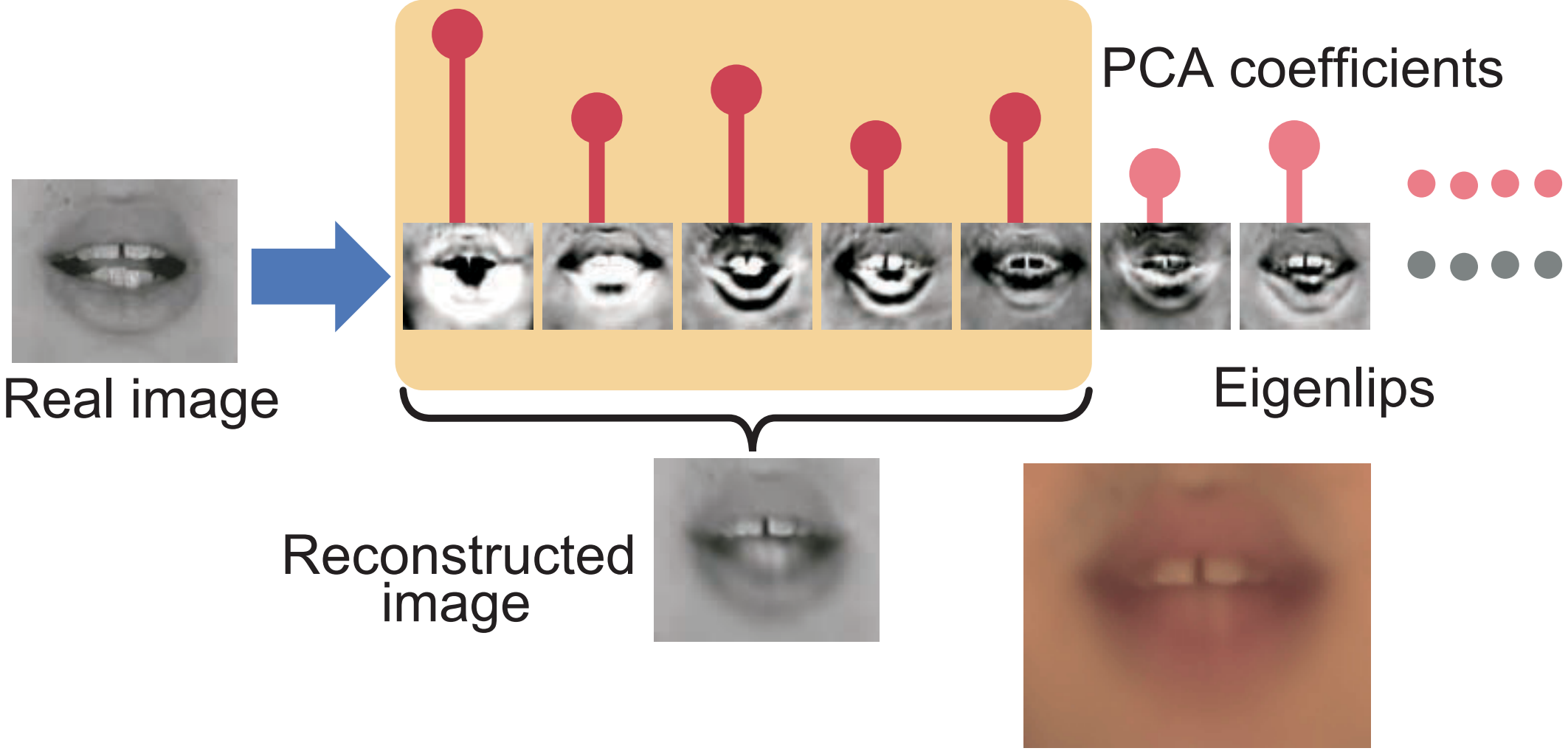
Future work includes constructing more effective contexts for emotional speech synthesis. angry

In addition, improvements in the quality of synthesized speech is also future work. That's all."

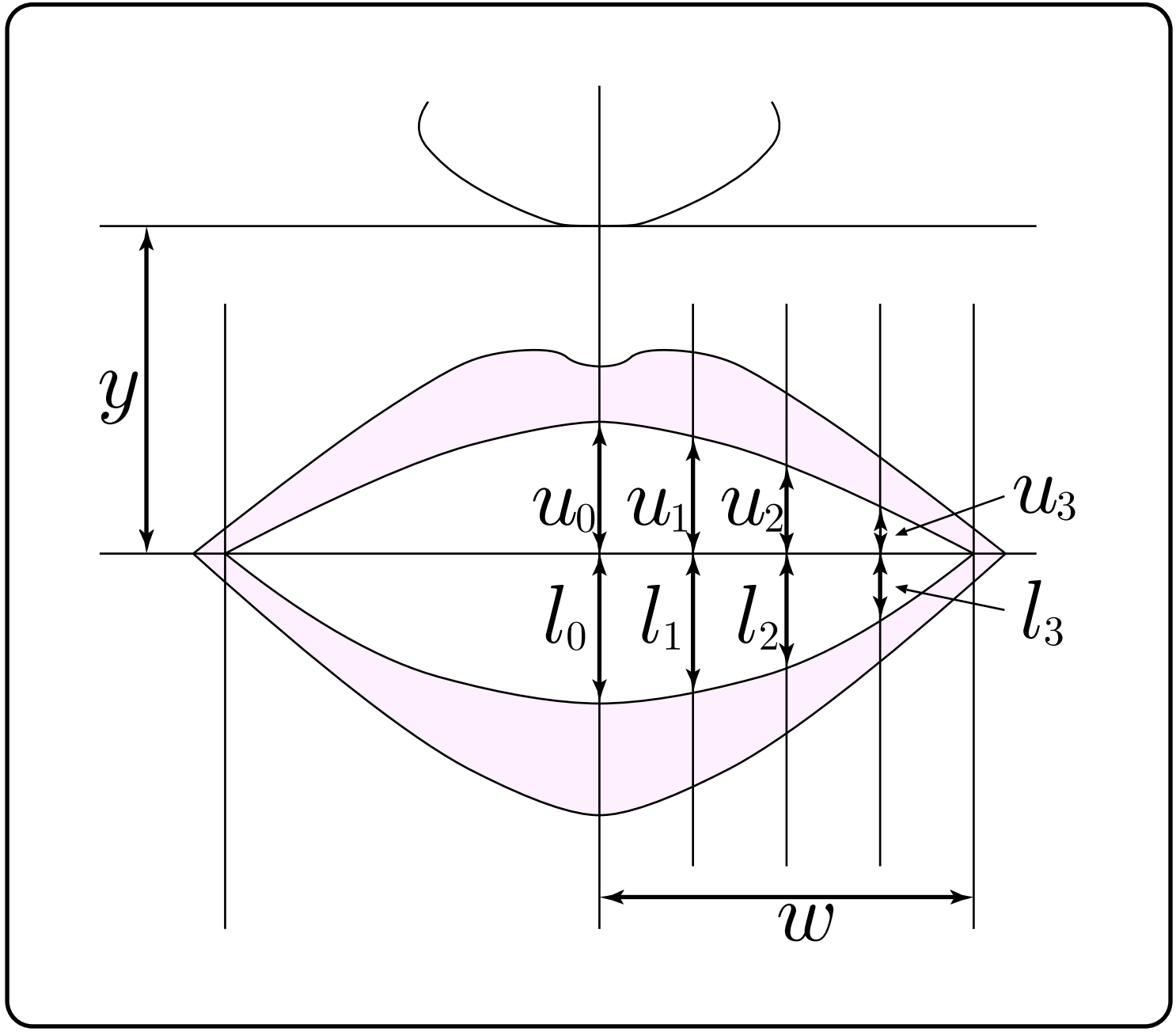
Audio-visual synthesis (pixel-based) [Sako;'00]

Pixel image: high dimensionality

⇒ Dimensionality reduction by PCA



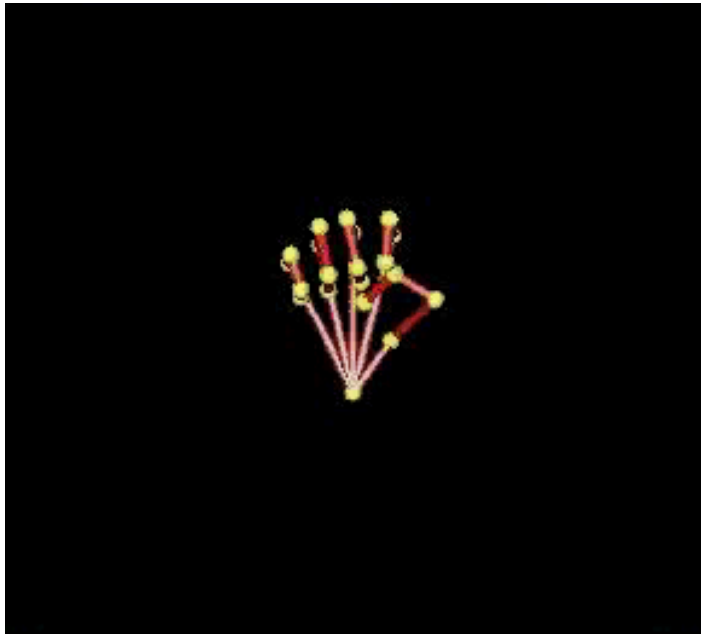
Audio-visual synthesis (model-based) [Tamura;'98]



Human motion synthesis

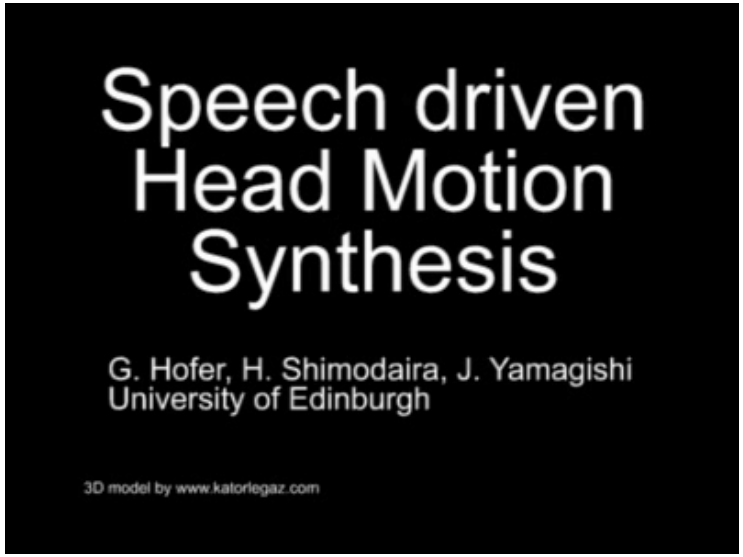
- Record human movements by a motion capture system
- Train HMMs
- Synthesize motions

Finger [Haoka;'02]

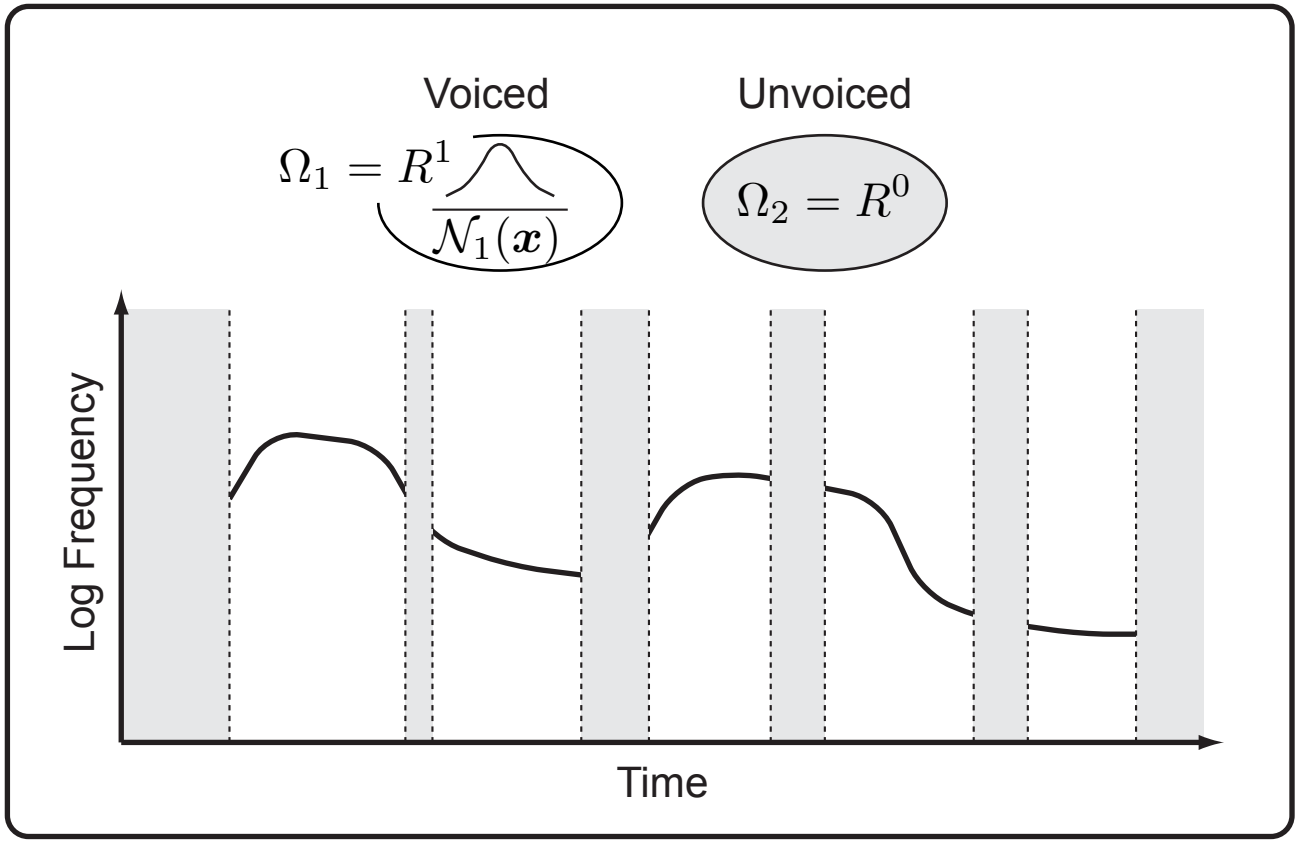


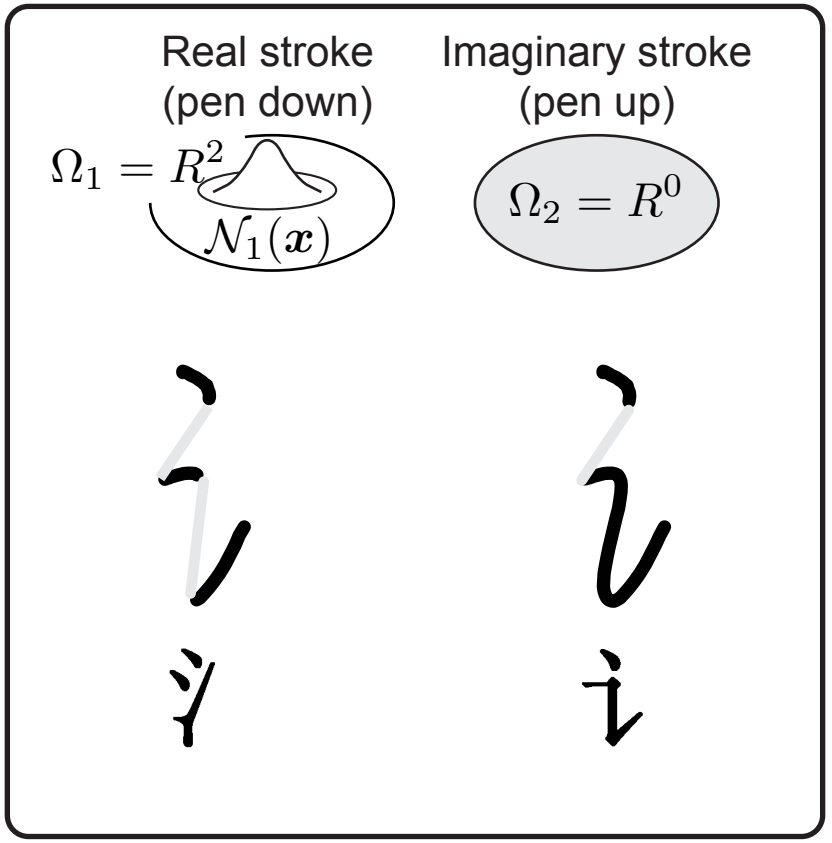
Walking [Niwase;05]

Head [Hofer;'07]



Online handwriting recognition [Ma;'07]





- Observation consists of continuous value or discrete symbol

  ⇒ Can be modeled by MSD-HMMs

- Properly handle imaginary strokes

  ⇒ Better discrimination between similar characters

Very small footprint synthesizer [Morioka;'04]

- Acoustic model size < 100KBytes
  - Tree size
  - Order of mel-cepstral analysis
  - Variance flooring scale
- 0.1 x Real Time on a old desktop PC

	Samples	File size*
Before shrinking		950 KB
After shrinking		100 KB

* In single precision floating point

Summary (second half)

Recent advances in HMM-based speech synthesis

- Related topics
  - Unit selection synthesis & hybrid
  - Flexibility to control voice characteristics
- Recent advances
  - Vocoding
  - Acoustic modeling
  - Over-smoothing compensation
- Applications

Further information

HTS website: http://hts.sp.nitech.ac.jp/

- **hts-users mailing list**
  - Over 500 posts per year
  - All posts are archived & searchable
  - Bug reports, Q&A, announce
- **Download page**
  - Patches/fixes after official releases
- **Publication page**
  - List of important papers
  - Tutorial slides
- **Demo page**
  - List of demonstrations

Acknowledgments / Special thanks!!

K. Tokuda; Overall concept & design
H. Zen; Throwing away all old codes of HTS & rewriting them as HTS ver. 1.0--2.1. Former maintainer of HTS
K. Oura; Implementation of hts_engine API. Current maintainer of HTS
J. Yamagishi; Testing out various adaptation algorithms. Implementation of CSMAPLR adaptation
T. Yoshimura; Development of the first version of HTS
M. Tamura; Implementation of various speaker adaptation algorithms for HMM-based speech synthesis
S. Sako; Applying HMM-based speech synthesis to audio-visual synthesis
N. Miyazaki; Implementation of the first version of MSD-HMM trainer
T. Yamada; Testing out various HMM training algorithms for HMM-based speech synthesis
T. Masuko; Discussions in detail & implementing speech parameter generation algorithm
K. Koishida; Discussions in detail & implementing speech coder based on mel-generalized cepstral analysis
Y. Nankaku; Discussions & supervising students working on HMM-based speech synthesis
T. Kobayashi; Discussions & supervising students working on HMM-based speech synthesis
T. Kitamura; Discussions & supervising students on HMM-based speech synthesis
S. Imai; Supervising students working on HMM-based speech synthesis
A. W. Black; Discussions about all aspects of speech synthesis, helping to use Festival
T. Toda; Discussions & implementation of speech parameter generation algorithm considering GV
Y.-J. Wu; Discussions & implementation of minimum generation error training
T. Nose; Implementing multiple-regression based voice characteristics control techniques
M. Tachibana; Testing out various speaking style modification techniques
S. J. Young, P. C. Woodland, M. J. F. Gales, G. Evermann, & other HTK developers; Development of HTK
& others,...

References (1)

Sagisaka;'92 - "ATR nu-TALK speech synthesis system," ICSLP, '92.

Black;'96 - "Automatically clustering similar units...," Eurospeech, '97.

Beutnagel;'99 - "The AT&T Next-Gen TTS system," Joint ASA, EAA, & DAEA meeting, '99.

Yoshimura;'99 - "Simultaneous modeling of spectrum ...," Eurospeech, '99.

Itakura;'70 - "A statistical method for estimation of speech spectral density...," Trans. IEICE, J53-A, '70.

Imai;'88 - "Unbiased estimator of log spectrum and its application to speech signal...," EURASIP, '88.

Kobayashi;'84 - "Spectral analysis using generalized cepstrum," IEEE Trans. ASSP, 32, '84.

Tokuda;'94 - "Mel-generalized cepstral analysis -- A unified approach to speech spectral...," ICSLP, '94.

Imai;'83 - "Cepstral analysis synthesis on the mel frequency scale," ICASSP, '83.

Fukada;'92 - "An adaptive algorithm for mel-cepstral analysis of speech," ICASSP, '92.

Itakura;'75 - "Line spectrum representation of linear predictive coefficients of speech...," J. ASA (57), '75.

Tokuda;'02 - "Multi-space probability distribution HMM," IEICE Trans. E85-D(3), '02.

Odell;'95 - "The use of context in large vocaburary...," PhD thesis, University of Cambridge, '95.

Shinoda;'00 - "MDL-based context-dependent subword modeling...," Journal of ASJ(E) 21(2), '00.

Yoshimura;'98 - "Duration modeling for HMM-based speech synthesis," ICSLP, '98.

Tokuda;'00 - "Speech parameter generation algorithms for HMM-based speech synthesis," ICASSP, '00.

Kobayashi;'85 - "Mel generalized-log spectrum approximation...," IEICE Trans. J68-A (6), '85.

Hunt;'96 - "Unit selection in a concatenative speech synthesis system using...," ICASSP, '96.

Donovan;'95 - "Improvements in an HMM-based speech synthesiser," Eurospeech, '95.

Kawai;'04 - "XIMERA: A new TTS from ATR based on corpus-based technologies," ISCA SSW5, '04.

Hirai;'04 - "Using 5 ms segments in concatenative speech synthesis," Proc. ISCA SSW5, '04.

References (2)

Rouibia;'05 - "Unit selection for speech synthesis based on a new acoustic target cost," Interspeech, '05.
Huang;'96 - "Whistler: A trainable text-to-speech system," ICSLP, '96.
Mizutani;'02 - "Concatenative speech synthesis based on HMM," ASJ autumn meeting, '02.
Ling;'07 - "The USTC and iFlytek speech synthesis systems...," Blizzard Challenge workshop, 07.
Ling;'08 - "Minimum unit selection error training for HMM-based unit selection...," ICASSP, 08.
Plumpe;'98 - "HMM-based smoothing for concatenative speech synthesis," ICSLP, '98.
Wouters;'00 - "Unit fusion for concatenative speech synthesis," ICSLP, '00.
Okubo;'06 - "Hybrid voice conversion of unit selection and generation...," IEICE Trans. E89-D(11), '06.
Aylett;'08 - "Combining statistical parametric speech synthesis and unit selection..." LangTech, '08.
Pollet;'08 - "Synthesis by generation and concatenation of multiform segments," Interspeech, '08.
Yamagishi;'06 - "Average-voice-based speech synthesis," PhD thesis, Tokyo Inst. of Tech., '06.
Yoshimura;'97 - "Speaker interpolation in HMM-based speech synthesis system," Eurospeech, '97.
Tachibana;'05 - "Speech synthesis with various emotional expressions...," IEICE Trans. E88-D(11), '05.
Kuhn;'00 - "Rapid speaker adaptation in eigenvoice space," IEEE Trans. SAP 8(6), '00.
Shichiri;'02 - "Eigenvoices for HMM-based speech synthesis," ICSLP, '02.
Fujinaga;'01 - "Multiple-regression hidden Markov model," ICASSP, '01.
Nose;'07 - "A style control technique for HMM-based expressive speech...," IEICE Trans. E90-D(9), '07.
Yoshimura;'01 - "Mixed excitation for HMM-based speech synthesis," Eurospeech, '01.
Kawahara;'97 - "Restructuring speech representations using a ...", Speech Communication, 27(3), '97.
Zen;'07 - "Details of the Nitech HMM-based speech synthesis system...", IEICE Trans. E90-D(1), '07.
Abdl-Hamid;'06 - "Improving Arabic HMM-based speech synthesis quality," Interspeech, '06.

References (3)

Hemptinne;'06 - "Integration of the harmonic plus noise model into the...," Master thesis, IDIAP, '06.

Banos;'08 - "Flexible harmonic/stochastic modeling...," V. Jornadas en Tecnologias del Habla, '08.

Cabral;'07 - "Towards an improved modeling of the glottal source in...," ISCA SSW6, '07.

Maia;'07 - "An excitation model for HMM-based speech synthesis based on ...," ISCA SSW6, '07.

Ratio;'08 - "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering," Interspeech, '08.

Drugman;'09 - "Using a pitch-synchronous residual codebook for hybrid HMM/frame...", ICASSP, '09.

Dines;'01 - "Trainable speech synthesis with trendedn hidden Markov models," ICASSP, '01.

Sun;'09 - "Polynomial segment model based statistical parametric speech synthesis...," ICASSP, '09.

Bulyko;'02 - "Robust splicing costs and efficient search with BMM models for...," ICASSP, '02.

Shannon;'09 - "Autoregressive HMMs for speech synthesis," Interspeech, '09.

Zen;'06 - "Reformulating the HMM as a trajectory model...", Computer Speech & Language, 21(1), '06.

Wu;'06 - "Minimum generation error training for HMM-based speech synthesis," ICASSP, '06.

Hashimoto;'09 - "A Bayesian approach to HMM-based speech synthesis," ICASSP, '09.

Wu;'08 - "Minimum generation error training with log spectral distortion for...," Interspeech, '08.

Toda;'08 - "Statistical approach to vocal tract transfer function estimation based on...," ICASSP, '08.

Oura;'08 - "Simultaneous acoustic, prosodic, and phrasing model training for TTS...," ISCSLP, '08.

Ferguson;'80 - "Variable duration models...," Symposium on the application of HMM to text speech, '80.

Levinson;'86 - "Continuously variable duration hidden...," Computer Speech & Language, 1(1), '86.

Beal;'03 - "Variational algorithms for approximate Bayesian inference," PhD thesis, Univ. of London, '03.

Masuko;'03 - "A study on conditional parameter generation from HMM...," Autumn meeging of ASJ, '03.

Yu;'07 - "A novel HMM-based TTS system using both continuous HMMs and discrete...," ICASSP, '07.

References (4)

Qian;'08 - "Generating natural F0 trajectory with additive trees," Interspeech, '08.

Latorre;'08 - "Multilevel parametric-base F0 model for speech synthesis," Interspeech, '08.

Tiomkin;'08 - "Statistical text-to-speech synthesis with improved dynamics," Interspeech, '08.

Toda;'07 - "A speech parameter generation algorithm considering global...," IEICE Trans. E90-D(5), '07.

Wu;'08 - "Minimum generation error criterion considering global/local variance...," ICASSP, '08.

Toda;'09 - "Trajectory training considering global variance for HMM-based speech...," ICASSP, '09.

Saino;'06 - "An HMM-based singing voice synthesis system," Interspeech, '06.

Tsuzuki;'04 - "Constructing emotional speech synthesizers with limited speech...," Interspeech, '04.

Sako;'00 - "HMM-based text-to-audio-visual speech synthesis," ICSLP, '00.

Tamura;'98 - "Text-to-audio-visual speech synthesis based on parameter generation...," ICASSP, '98.

Haoka;'02 - "HMM-based synthesis of hand-gesture animation," IEICE Commun. syst.,102(517), '02.

Niwase;'05 - "Human walking motion synthesis with desired pace and...," IEICE Trans. E88-D(11), '05.

Hofer;'07 - "Speech driven head motion synthesis based on a trajectory model," SIGGRAPH, '07.

Ma;'07 - "A MSD-HMM approach to pen trajectory modeling for online handwriting...," ICDAR, '07.

Morioka;'04 - "Miniaturization of HMM-based speech synthesis," Autumn meeting of ASJ, '04.

Kim;'06 - "HMM-Based Korean speech synthesis system for...," IEEE Trans. Comsum. Elec., 52(4), '06.

Klatt;'82 - "The Klatt-Talk text-to-speech system," ICASSP, '82.