基于HMM的语音合成
HMM based Speech Synthesis

清华大学深圳研究生院
吴志勇
zywu@sz.tsinghua.edu.cn

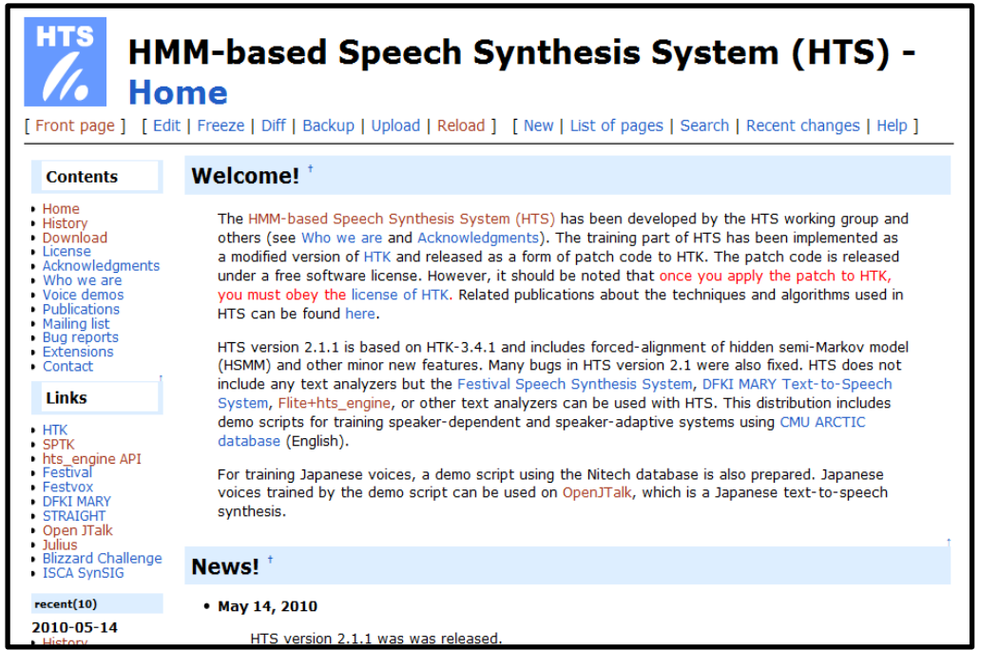
MEDIA JOINT Research Center

Adapted from HTS Slides

HTS Slides
released by HTS Working Group
http://hts.sp.nitech.ac.jp/

HTS: HMM-base Speech Synthesis System

- **HMM-based Speech Synthesis System (HTS)**
  - Released as a form of patch code to HTK
    - Under the New and Simplified BSD license
      - Once you apply the patch to HTK, you must obey the license of HTK
  - HTS-users mailing list
    - Over 500 posts per year
    - All posts are archived & searchable
    - Bug reports, Q&A, announce
  - Becoming a research platform
    - Using by various organizations (e. g., Microsoft, IBM)



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

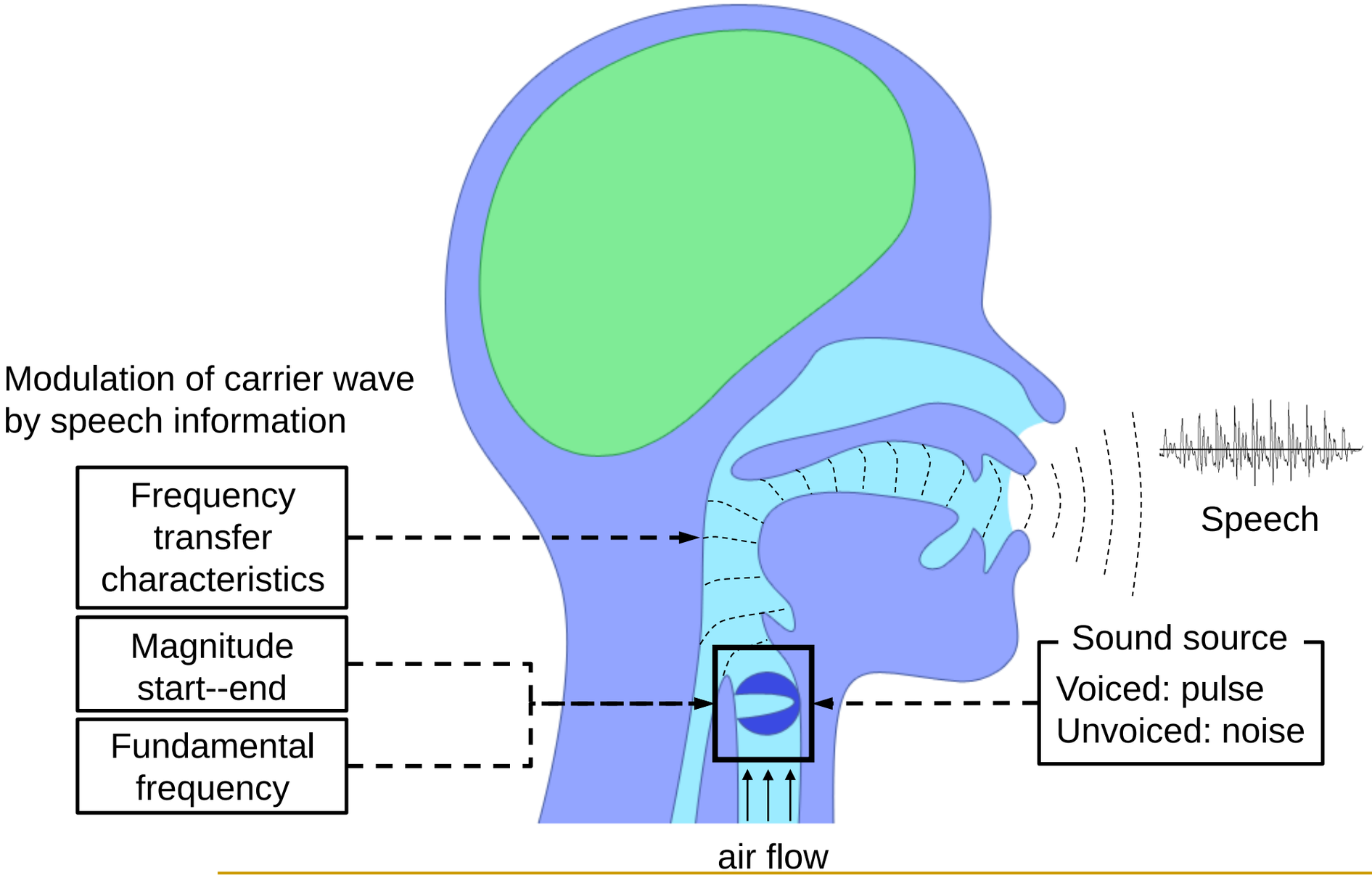
Outlines

- **Speech vocoding: Source-filter model**
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - State duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- **Voice character controlling**
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- **Application**
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

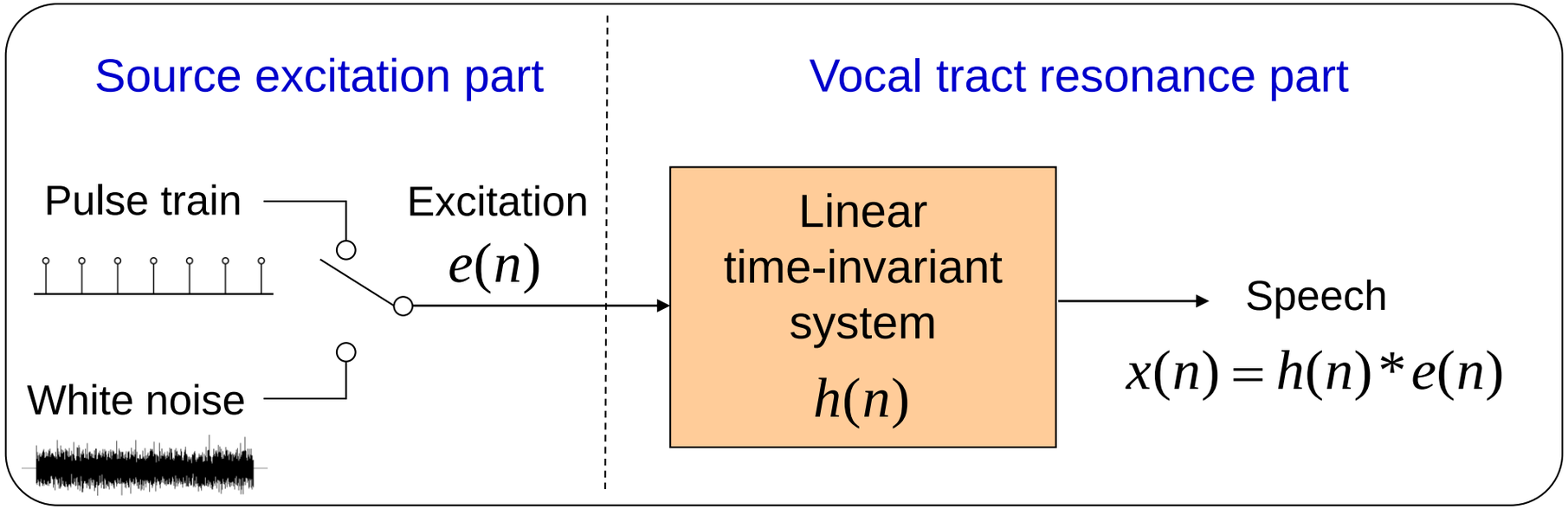
Outlines

- **Speech vocoding: Source-filter model**
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - State duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- **Voice character controlling**
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- **Application**
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

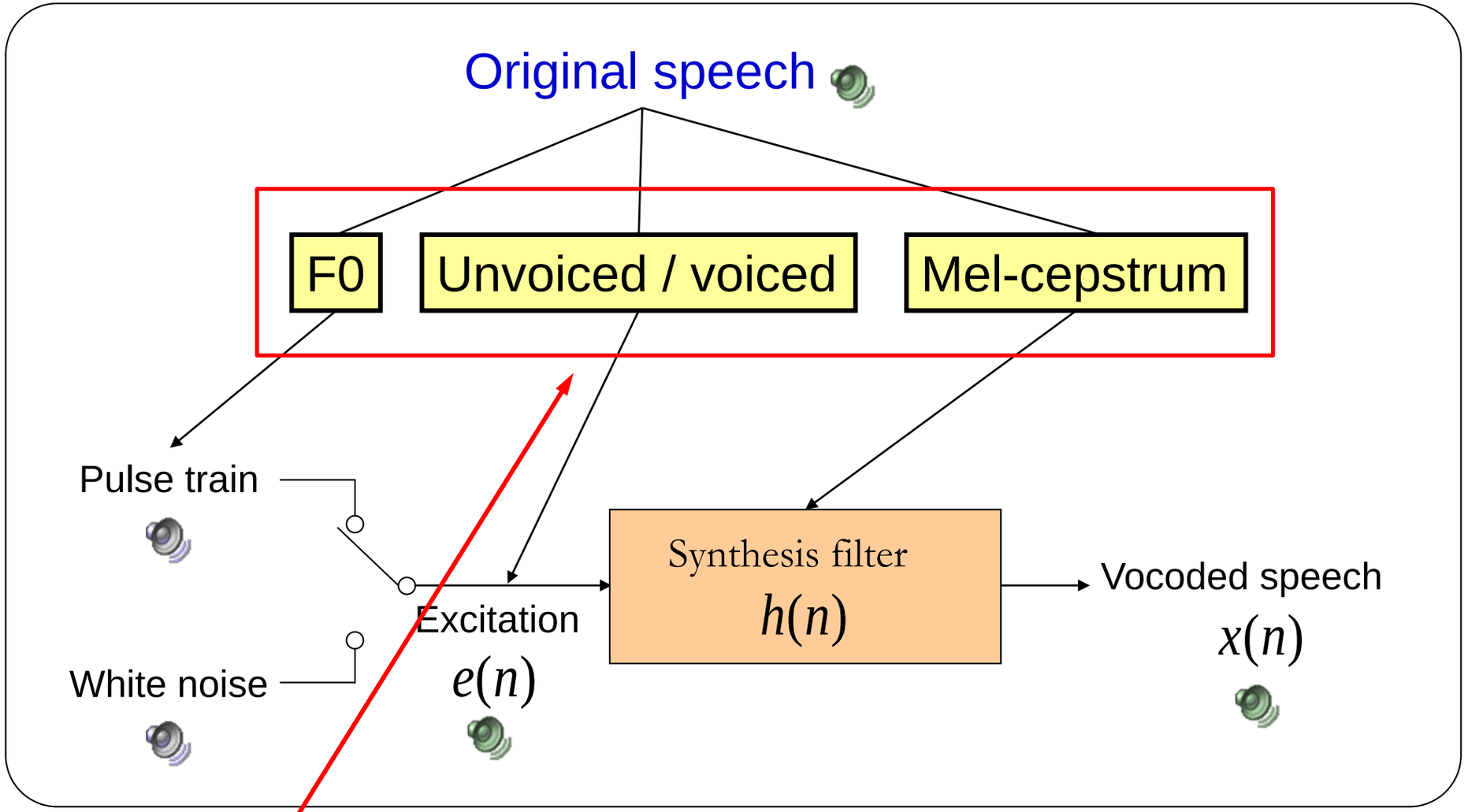
Speech Production Mechanism



Source-Filter Model

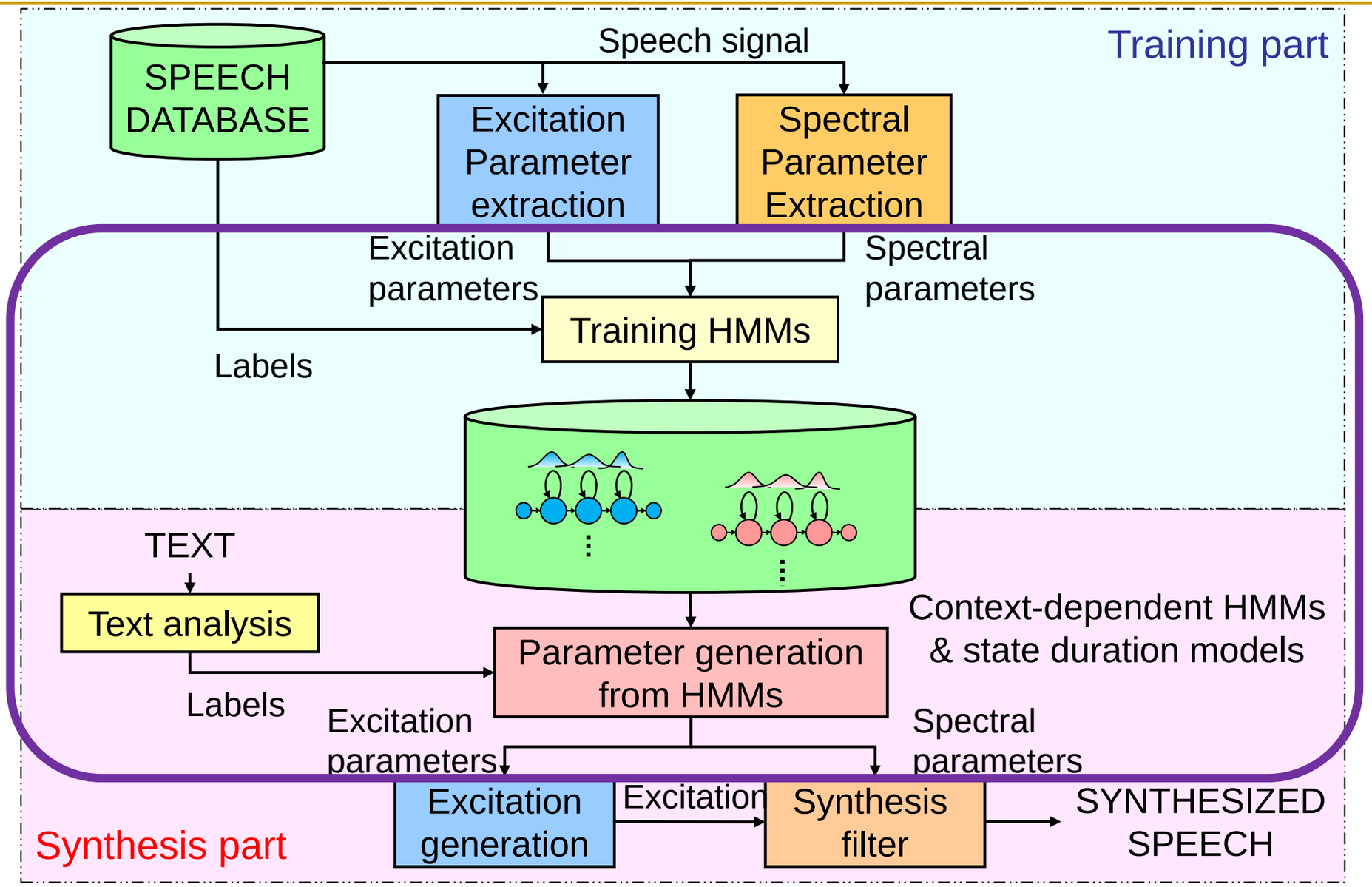


Overview of Speech Vocoding



These speech parameters are to be modeled by HMM

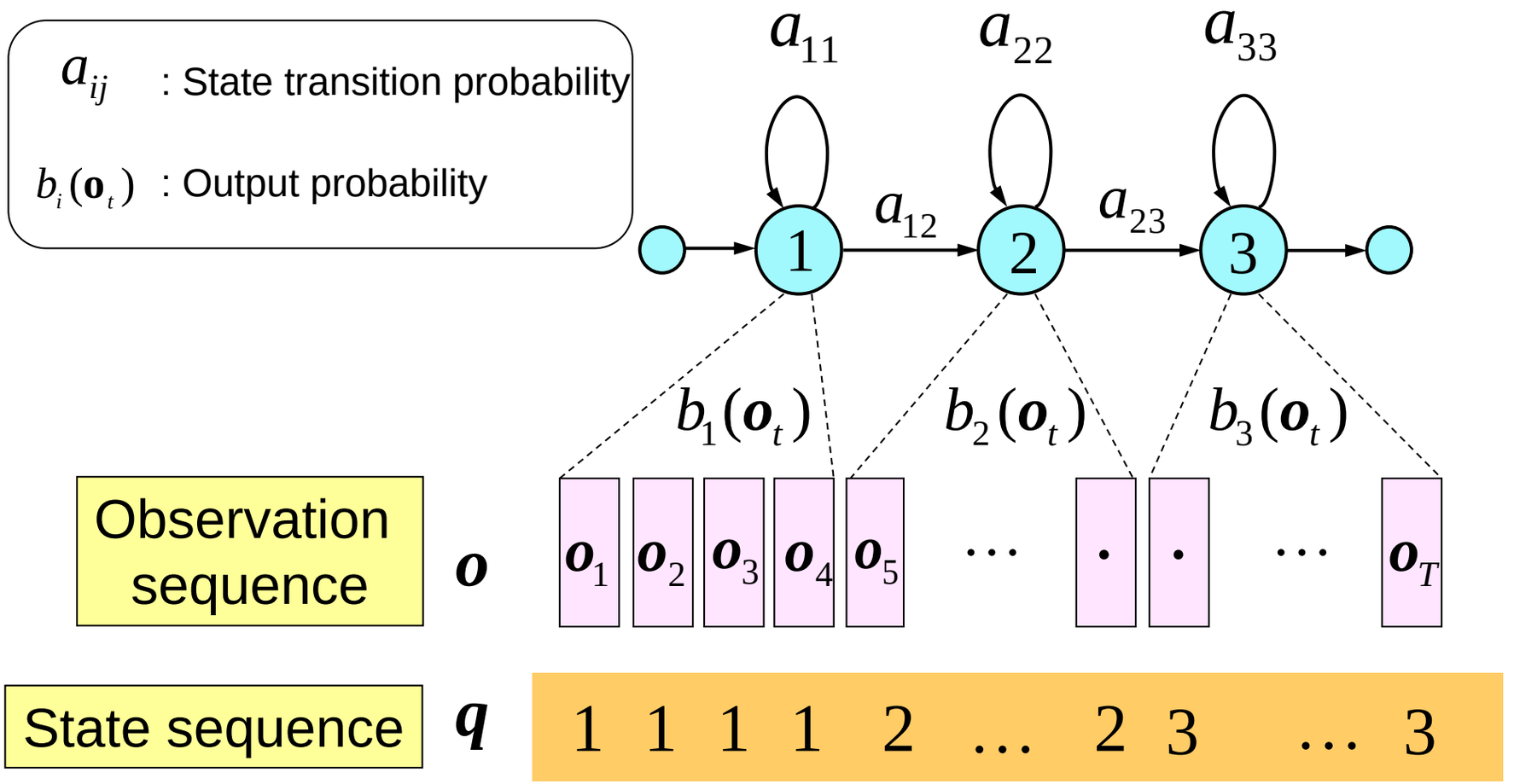
HMM-based Speech Synthesis System



Outlines

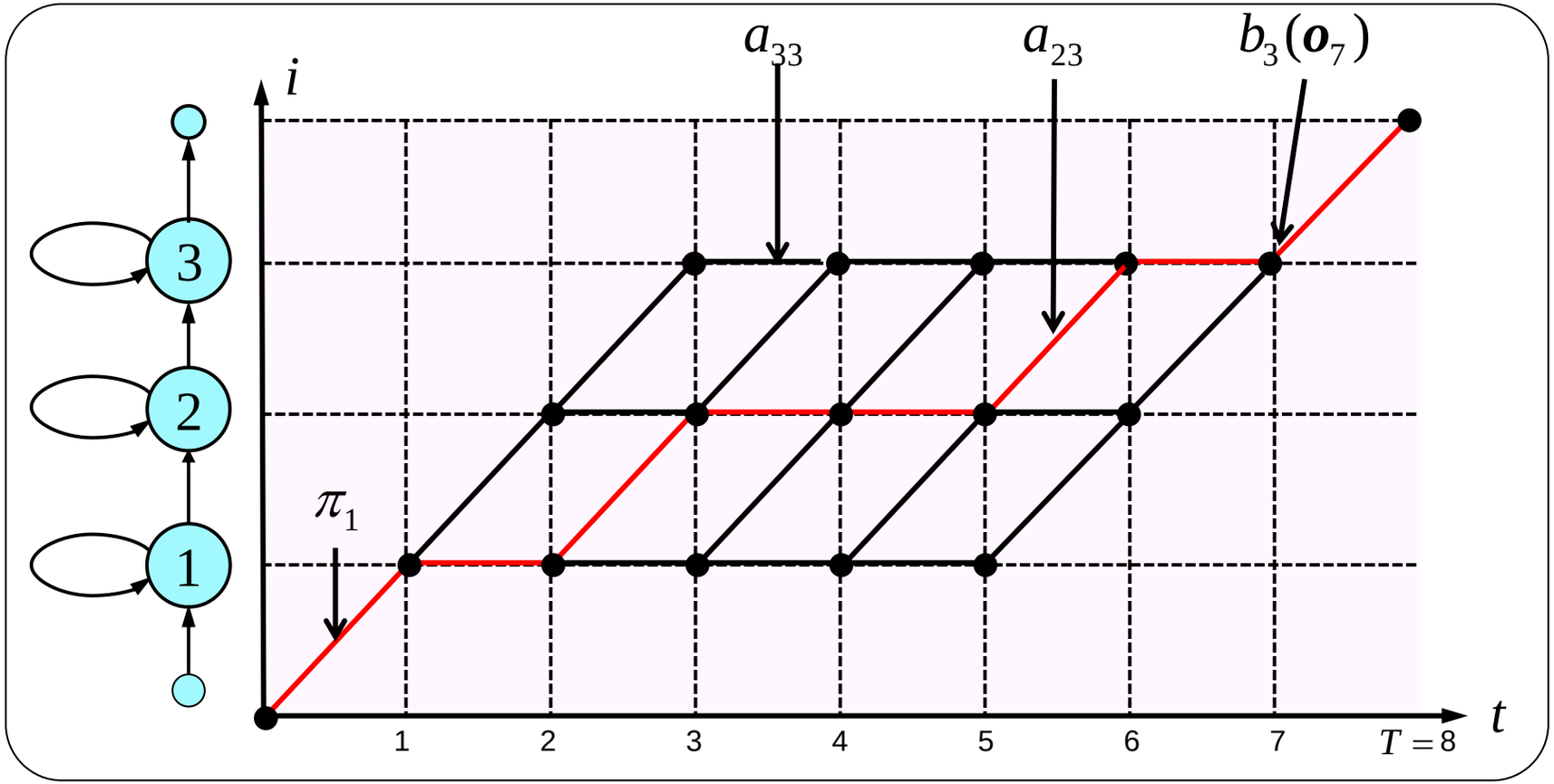
- Speech vocoding: Source-filter model
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - State duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- Voice character controlling
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- Application
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

Hidden Markov Model (HMM)



The Markov chain whose state sequence is unknown
⇒ Estimating state sequence by the observation

Output Probability of HMM



Likelihood function

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

Speech Parameter Generation Algorithm [Tokuda; '00]

For a given HMM $\boldsymbol{\lambda}$, determine a speech parameter vector sequence $\boldsymbol{o} = [\boldsymbol{o}_1^{\mathsf{T}}, \boldsymbol{o}_2^{\mathsf{T}}, \ldots, \boldsymbol{o}_T^{\mathsf{T}}]^{\mathsf{T}}$ which maximizes

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

$$\approx \arg\max_{\mathbf{q},\mathbf{o}} P(\mathbf{o} \mid \mathbf{q}, \boldsymbol{\lambda}) P(\mathbf{q} \mid \boldsymbol{\lambda})$$

$$\Downarrow$$

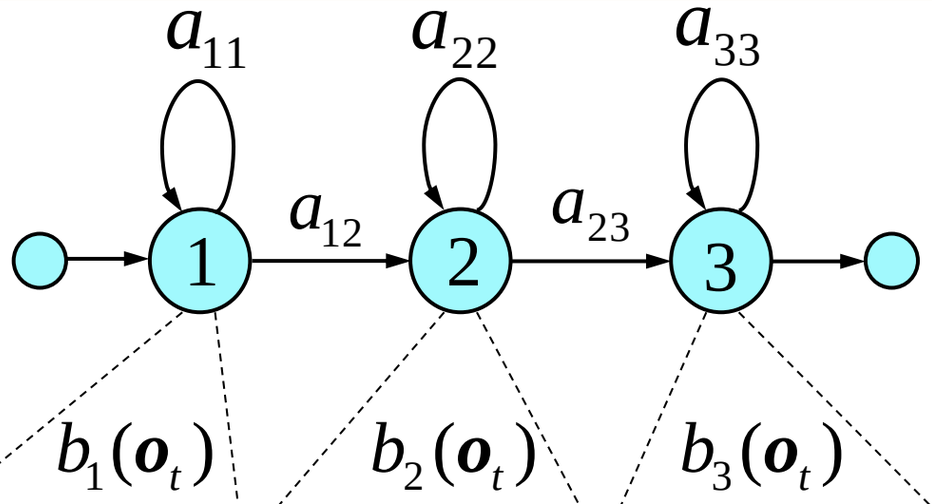
$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid w, \boldsymbol{\lambda})$$

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

Determination of State Sequence

a_{ij} : State transition probability

$b_i(\mathbf{o}_t)$: Output probability



Observation sequence

$\boldsymbol{o}$ $\boldsymbol{o}_1$ $\boldsymbol{o}_2$ $\boldsymbol{o}_3$ $\boldsymbol{o}_4$ $\boldsymbol{o}_5$ … . . … $\boldsymbol{o}_T$

State sequence

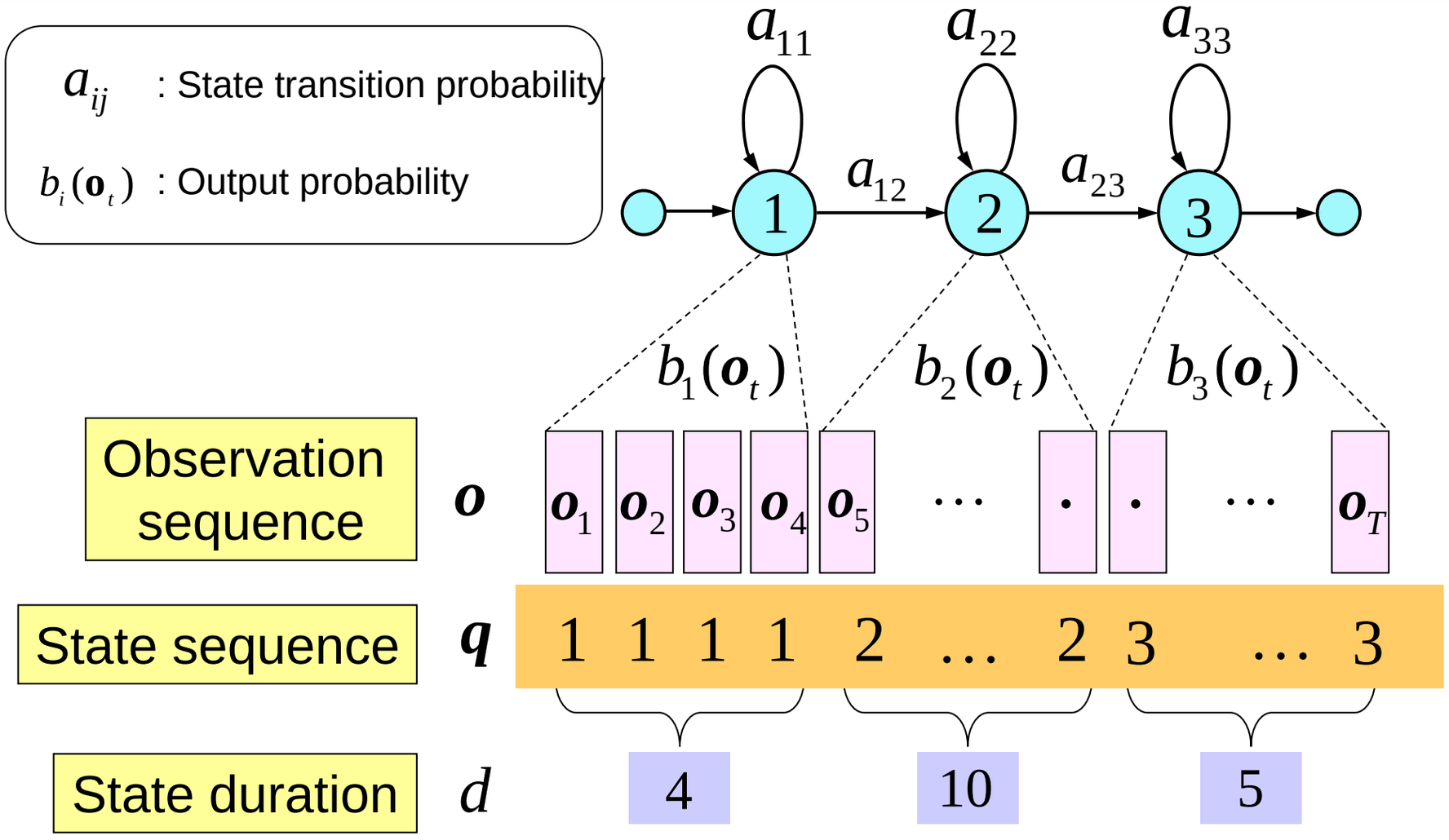
$\boldsymbol{q}$ 1 1 1 1 2 … 2 3 … 3

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

Outlines

- Speech vocoding: Source-filter model
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - State duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- Voice character controlling
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- Application
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

Determination of State Sequence



Determine state sequence via determining state durations

Determination of State Sequence

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

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

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

d_i : state duration of i-th state

K : # of states in a sentence HMM for w

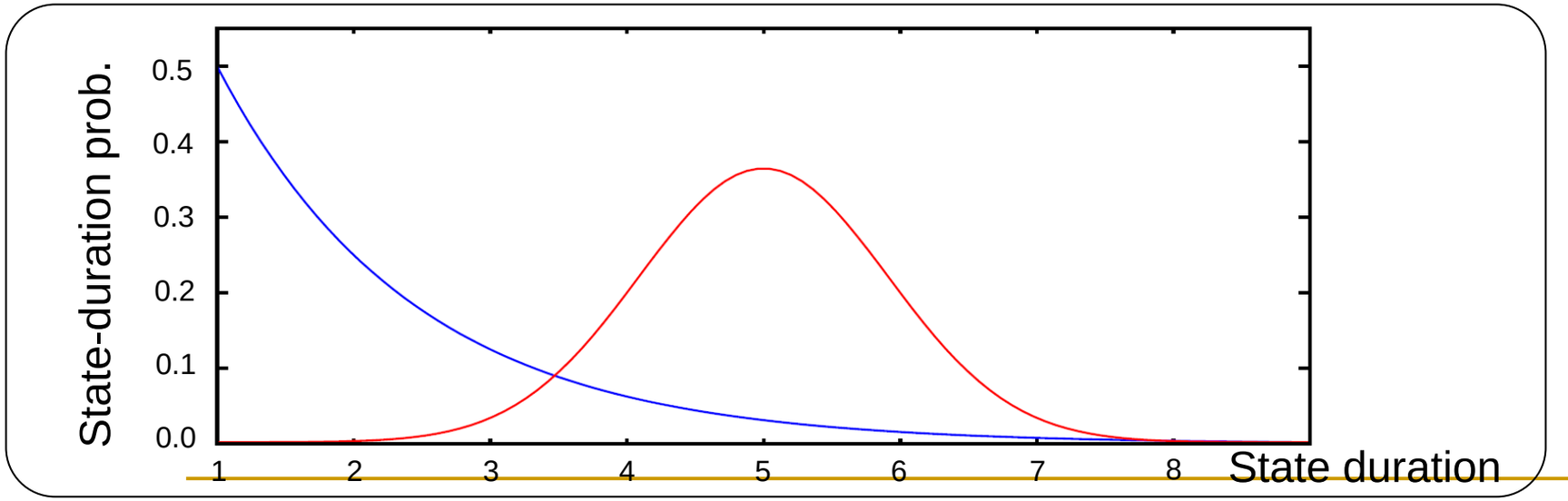
State Duration Modeling

Geometric

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

Gaussian

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



Outlines

- Speech vocoding: Source-filter model
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - Duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- Voice character controlling
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- Application
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

Speech Parameter Generation Algorithm [Tokuda; '00]

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

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

$$\approx \max_{\boldsymbol{q}} P(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q} \mid \boldsymbol{\lambda})$$

$$\Downarrow$$

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

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

Determination of State Output

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

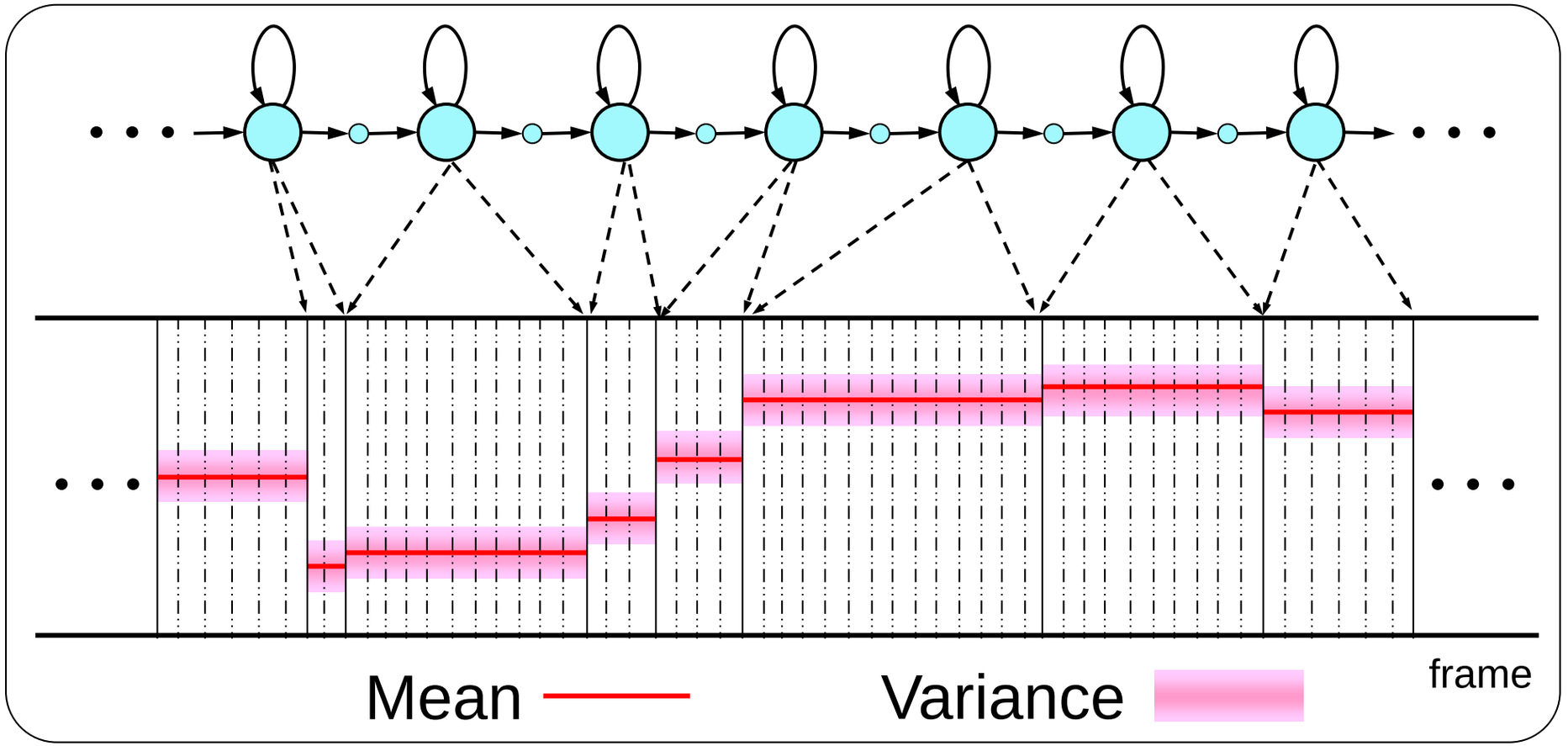
$$P(\mathbf{o} \mid \hat{\mathbf{q}}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} b_{q_t}(\mathbf{o}_t)$$

Gaussian distribution

$$b_j(\mathbf{o}_t) = N(\mathbf{o}_t \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

$$\hat{\mathbf{o}}_t = \boldsymbol{\mu}_j$$

Generated Feature Sequence



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

⇒ discontinuous outputs between states

To Solve the Problem …

Constrained with dynamic features!

Let the speech parameter vector $\mathbf{o}_t$ at frame t consists of the static feature vector $\mathbf{c}_t$ and the dynamic feature vector $\Delta\mathbf{c}_t$:

$$\mathbf{o}_t = \{\mathbf{c}_t, \Delta\mathbf{c}_t\}$$

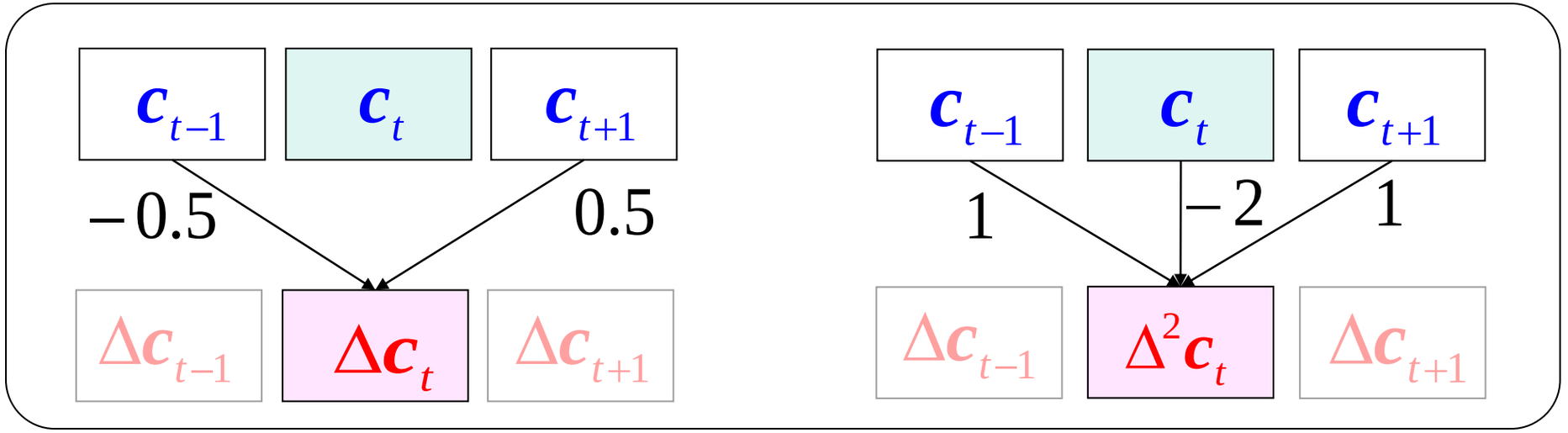
Assume $\mathbf{c}_t$ and $\Delta\mathbf{c}_t$ are statistically independent, then

$$P(\mathbf{o} \mid \hat{\mathbf{q}}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} b_{q_t}(\mathbf{o}_t) = \prod_{t=1}^{T} b_{q_t}(\mathbf{c}_t) b_{q_t}(\Delta\mathbf{c}_t)$$

Dynamic Features

$$\Delta \boldsymbol{c}_t = \frac{\partial \boldsymbol{c}_t}{\partial t} \approx 0.5(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1})$$

$$\Delta^2 \boldsymbol{c}_t = \frac{\partial^2 \boldsymbol{c}_t}{\partial t^2} \approx \boldsymbol{c}_{t+1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-1}$$



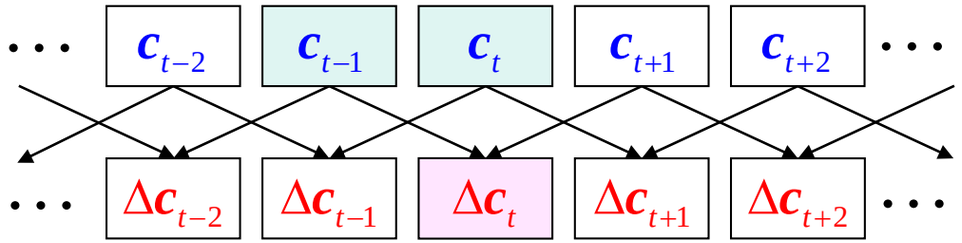
Integration of Dynamic Features

Speech parameter vector $\boldsymbol{o}_t$ includes both static & dynamic features

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^{\mathsf{T}}, \Delta\boldsymbol{c}_t^{\mathsf{T}} \right]^{\mathsf{T}}$$

$2M$ = M + M

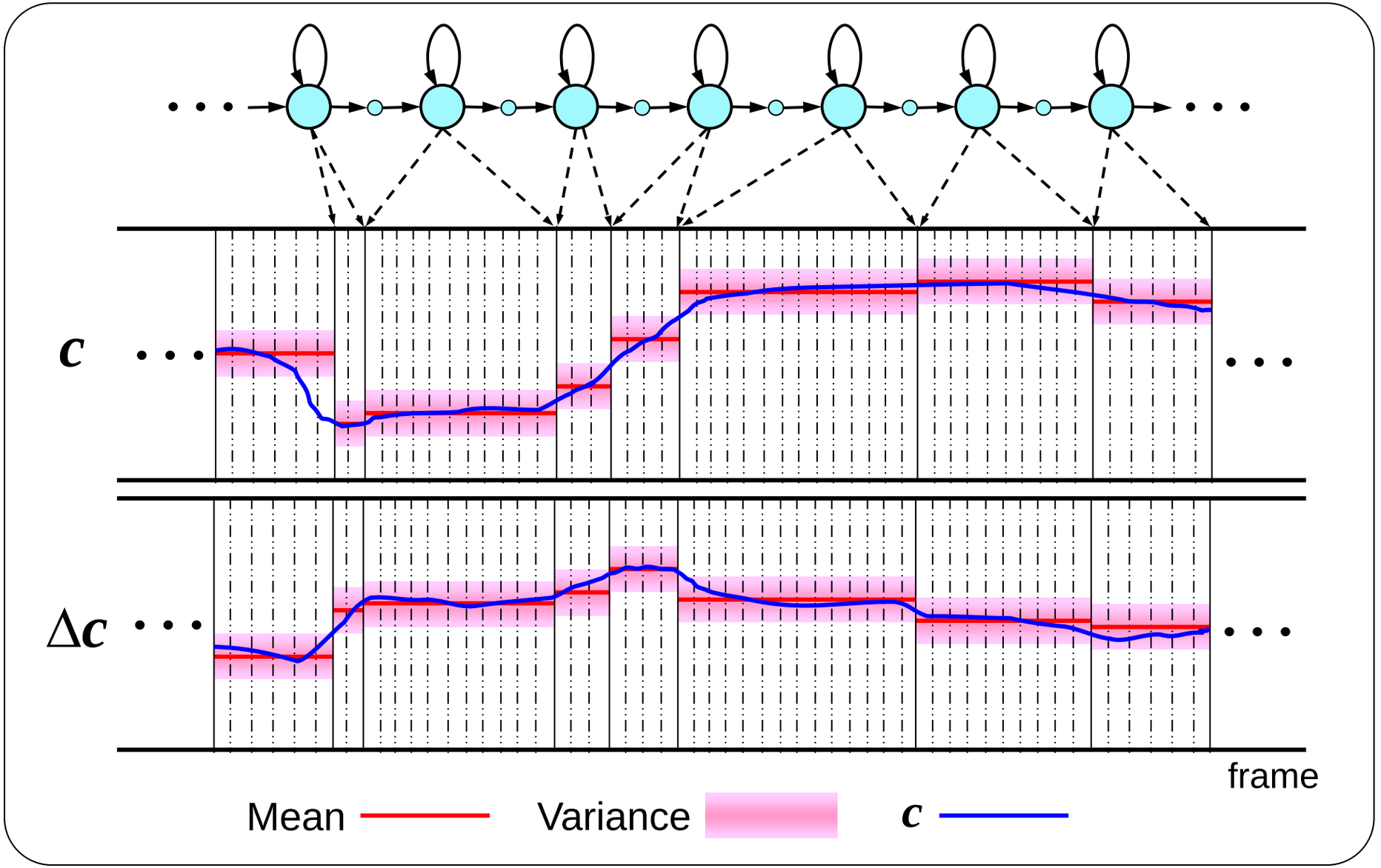
$$\Delta\boldsymbol{c}_t = 0.5(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1})$$



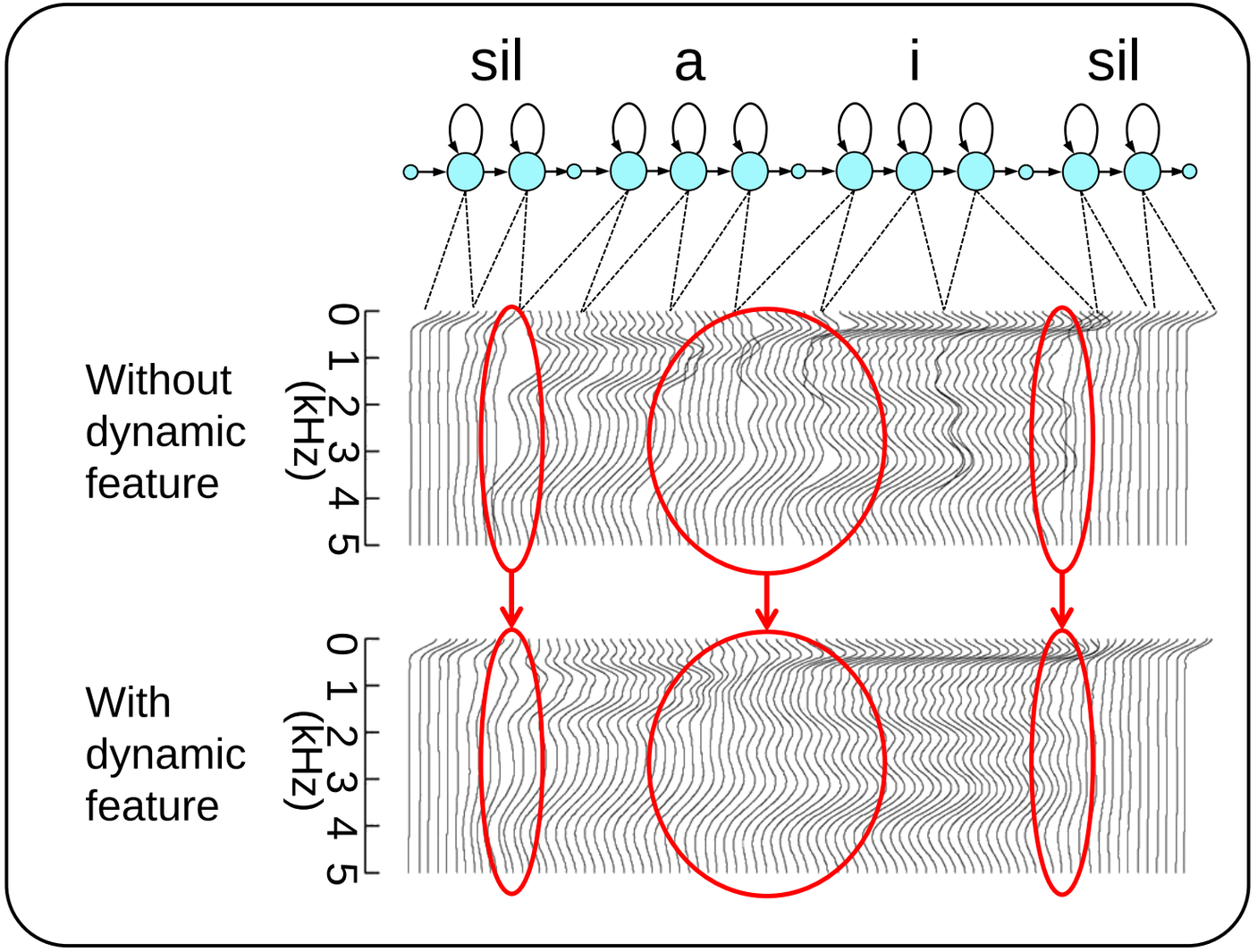
The relationship between $\boldsymbol{o}_t$ & $\boldsymbol{c}_t$ can be arranged as

$$\begin{bmatrix} \vdots \\ \boldsymbol{c}_{t-1} \\ \Delta\boldsymbol{c}_{t-1} \\ \boldsymbol{c}_t \\ \Delta\boldsymbol{c}_t \\ \boldsymbol{c}_{t+1} \\ \Delta\boldsymbol{c}_{t+1} \\ \vdots \end{bmatrix} \begin{matrix} \\ \boldsymbol{o}_{t-1} \\ \\ \boldsymbol{o}_t \\ \\ \boldsymbol{o}_{t+1} \\ \\ \end{matrix} = \begin{bmatrix} \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & -1/2\,\boldsymbol{I} & \boldsymbol{0} & 1/2\,\boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & -1/2\,\boldsymbol{I} & \boldsymbol{0} & 1/2\,\boldsymbol{I} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & -1/2\,\boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \end{bmatrix} \begin{matrix} \\ \\ \boldsymbol{c}_{t-2} \\ \boldsymbol{c}_{t-1} \\ \boldsymbol{c}_t \\ \boldsymbol{c}_{t+1} \\ \\ \end{matrix} \begin{matrix} \boldsymbol{o} \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \quad \boldsymbol{W}\ (M \times N) \quad \boldsymbol{c} \quad N$$

Generated Speech Parameter Trajectory



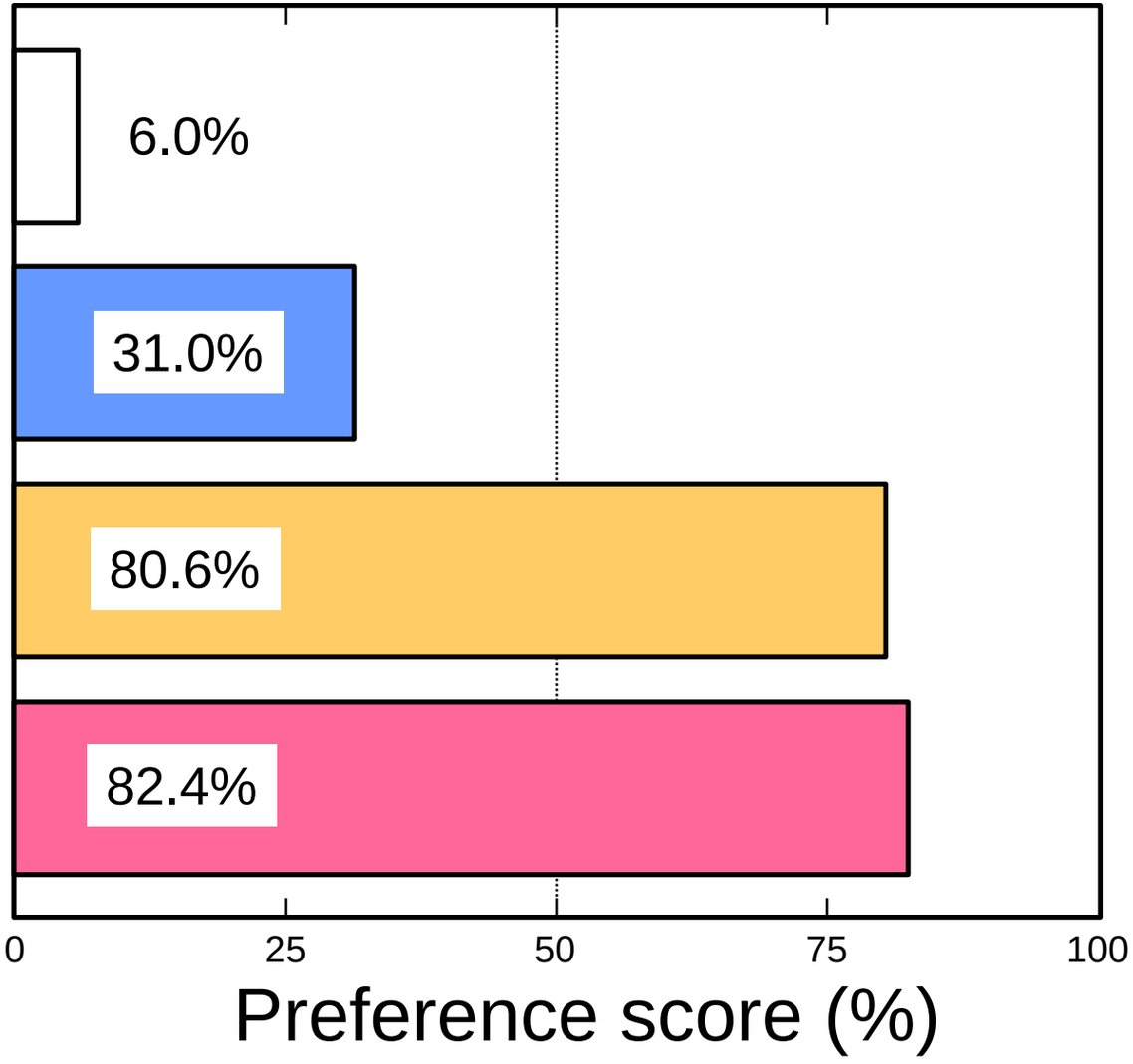
Generated Spectra



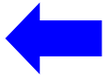
Spectra changing smoothly between phonemes

Effect of Dynamic Features (Japanese)

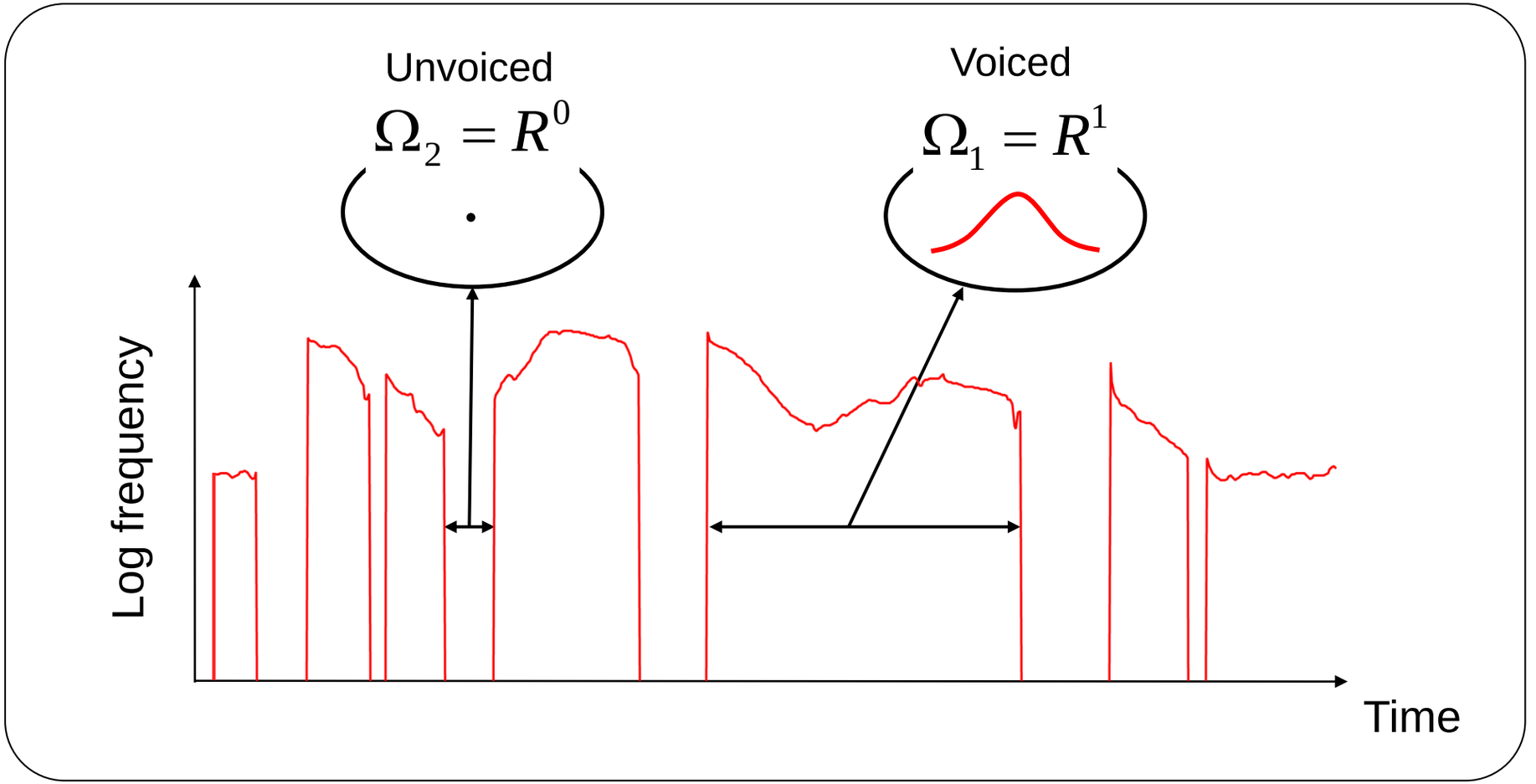
- static
- liner interpolation
- static + Δ
- static + $\Delta + \Delta^2$



Outlines

- Speech vocoding: Source-filter model
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - Duration modeling
  - Spectrum modeling
  - F0 modeling 
  - Context clustering
- Voice character controlling
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- Application
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

Observation of F0

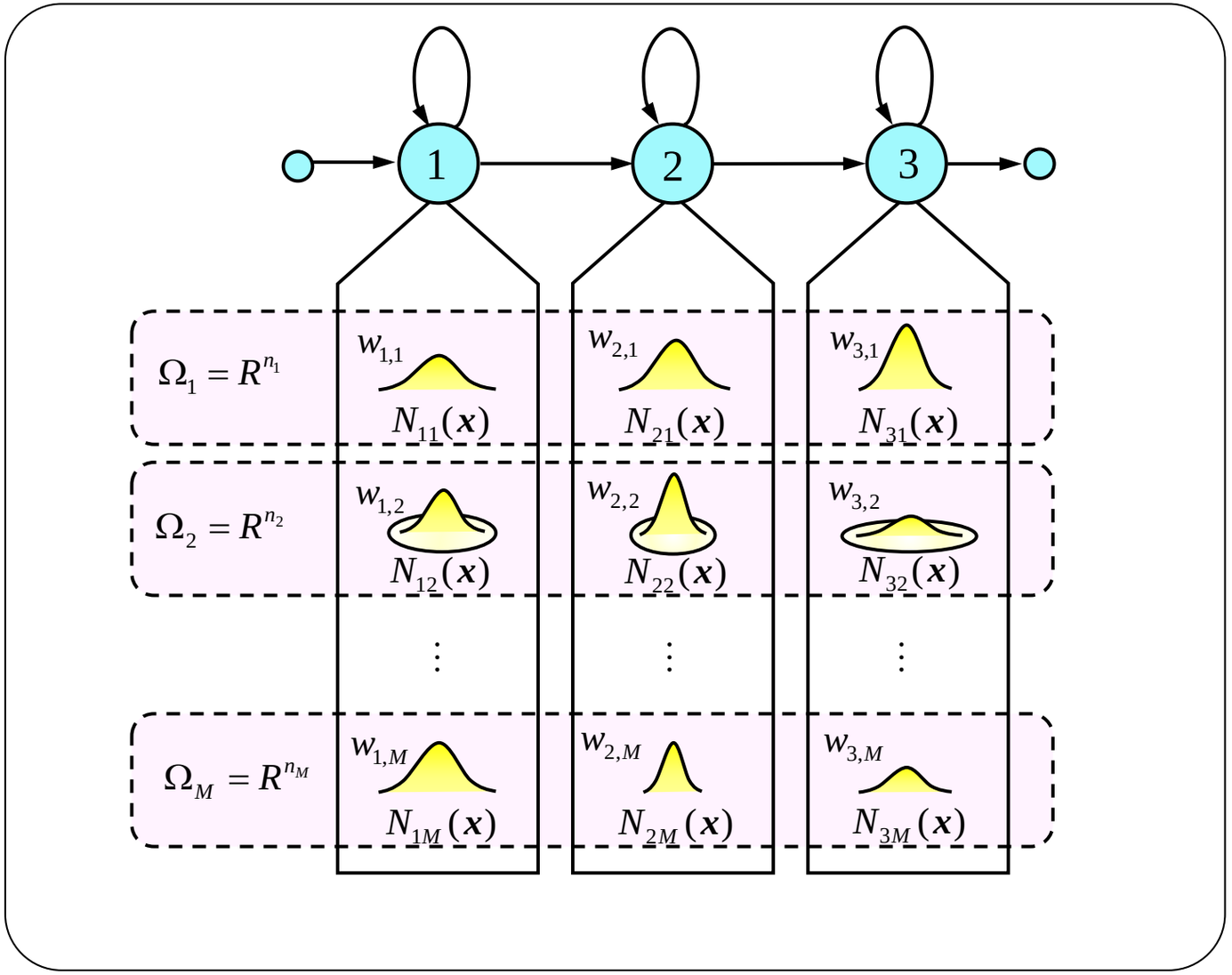


Unable to model by continuous or discrete distribution

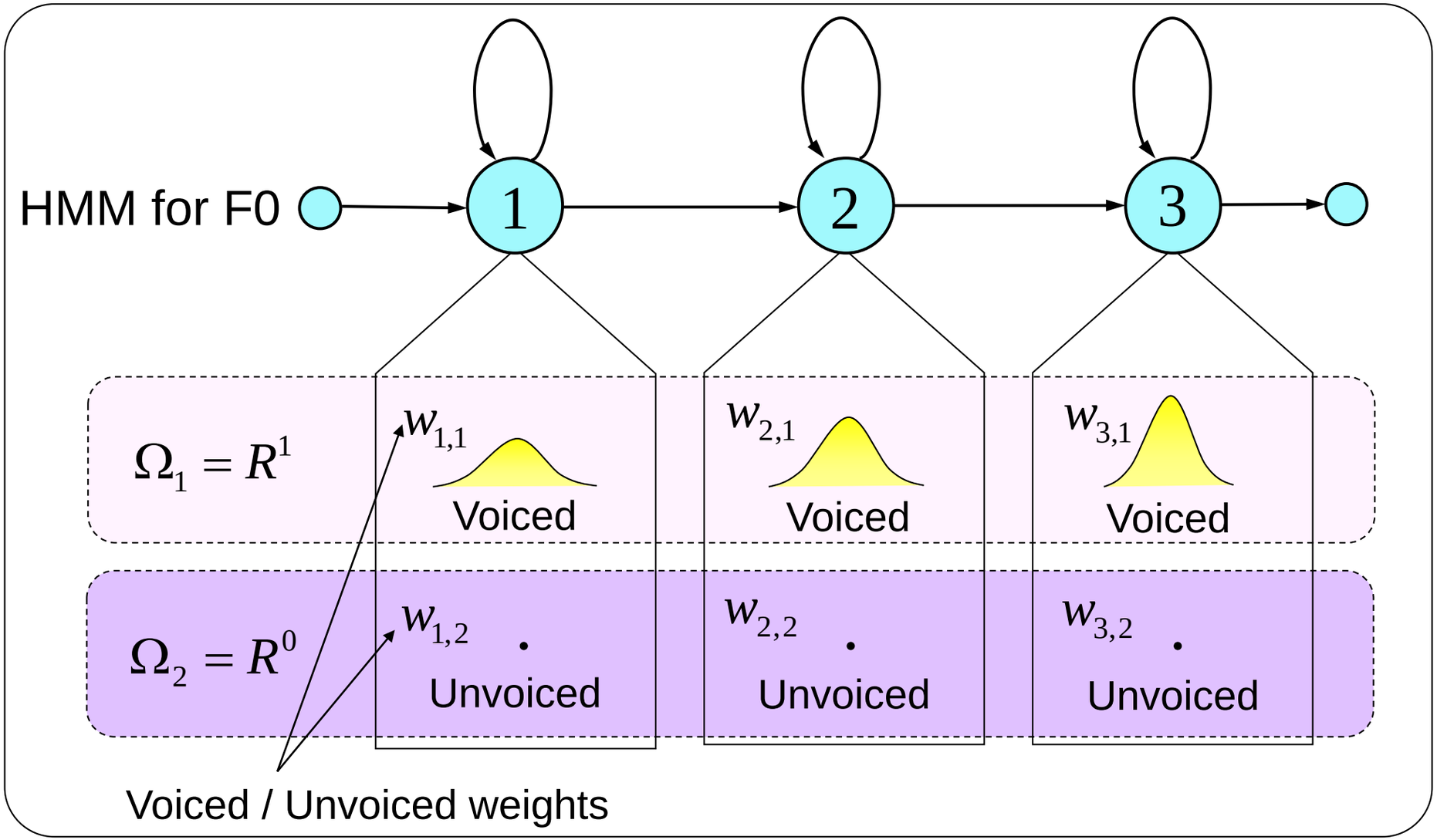
⇒ Multi-space probability distribution HMM (MSD-HMM)

Structure of MSD-HMM

Each state has two or more distributions

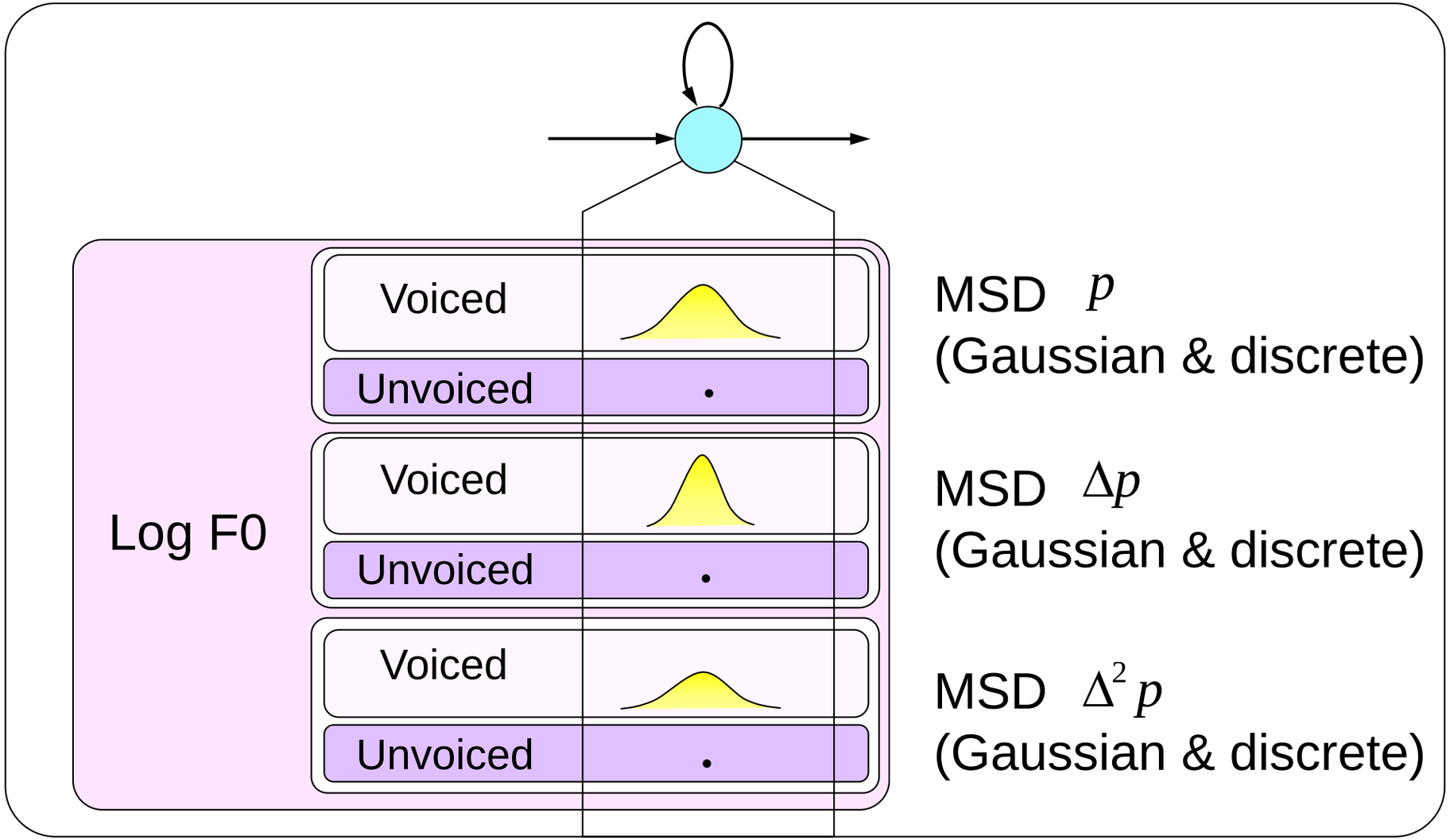


MSD-HMM for F0 Modeling

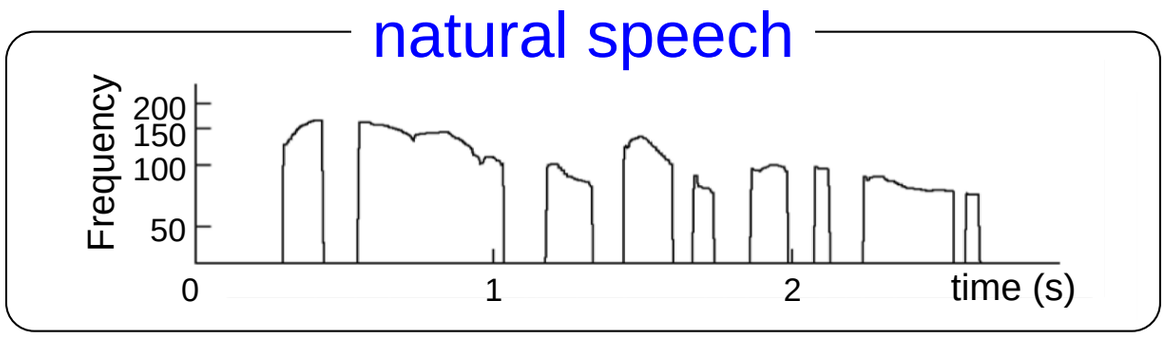


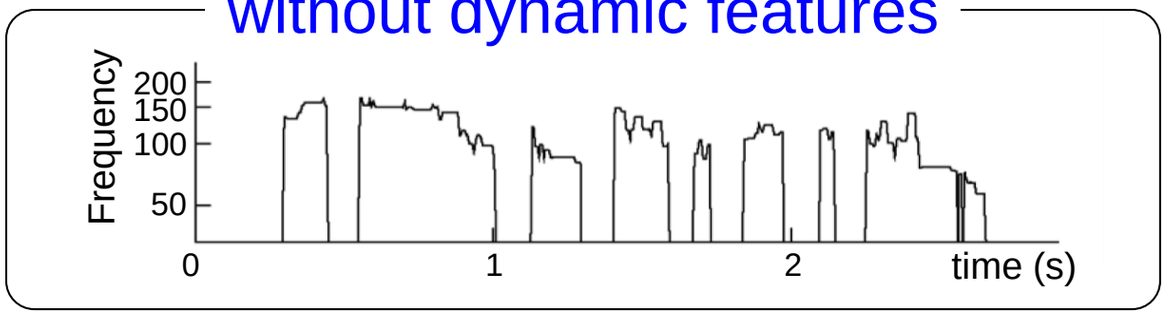
To Model Dynamic F0 Features

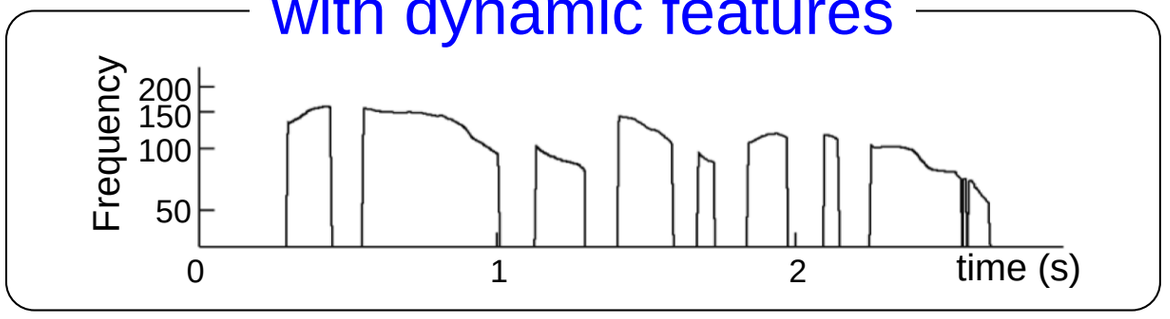
- **Structure of F0 state-output distributions**



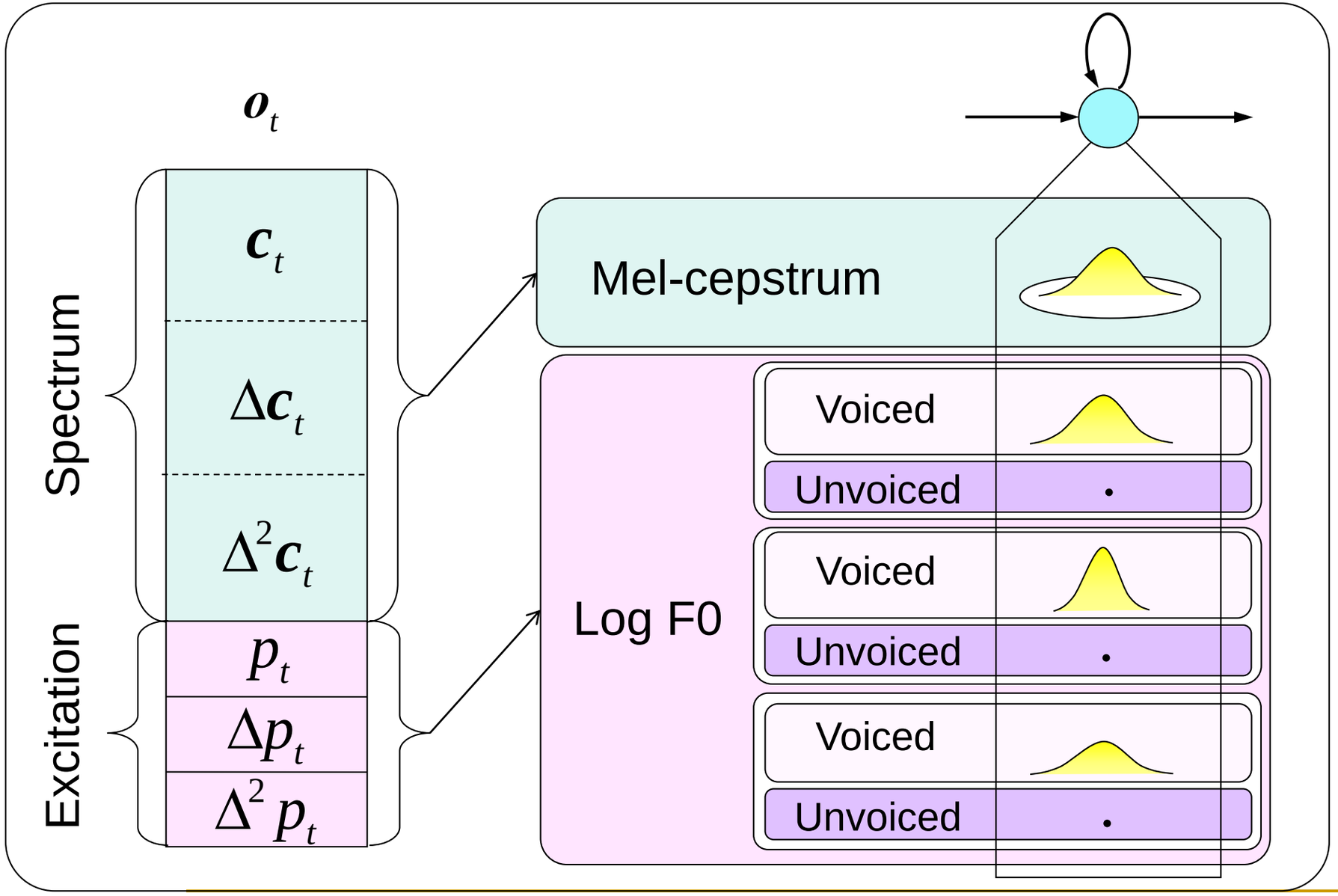
Generated F0 Trajectory







To Summarize: Structure of State-Output Distribution



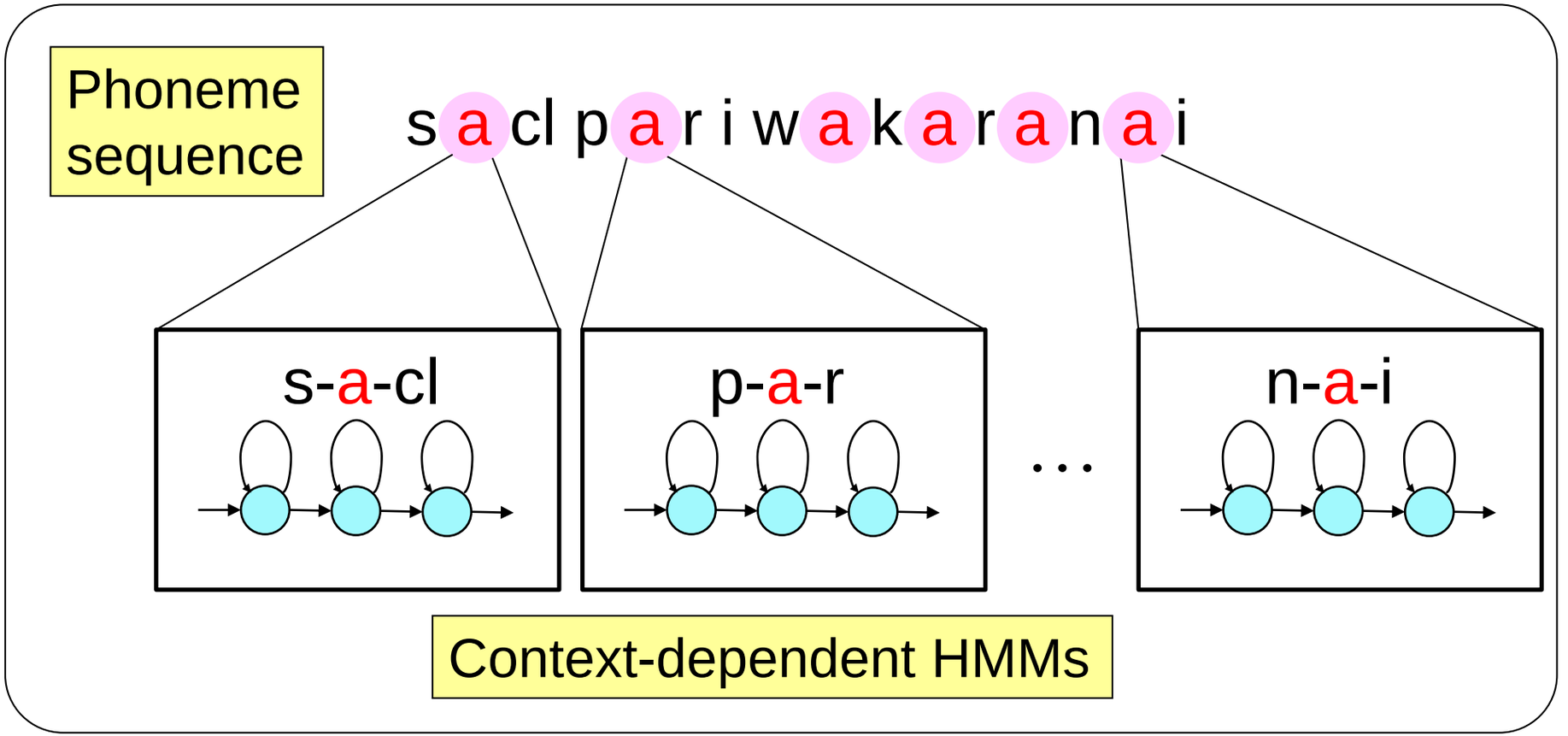
To Summarize: Speech Samples (Japanese)

		Mel-cepstrum		
		static	static + Δ	static+ $\Delta + \Delta^2$
log F0	static			
	static + Δ			
	static + $\Delta + \Delta^2$			

Outlines

- Speech vocoding: Source-filter model
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - Duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- Voice character controlling
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- Application
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

Context-Dependent Model



- **Considering relations between phonemes**
  - Context ⇒ factor of speech variations
  - Improving model accuracy

Context-Dependent Modeling

MEDIA JOINT Research Center

Phoneme

- {preceding, succeeding} two phonemes
- current phoneme

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} {accented, stressed} syllable in current phrase
- # of syllables {from previous, to next} {accented, stressed} syllable
- Vowel within current syllable

Word

- 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

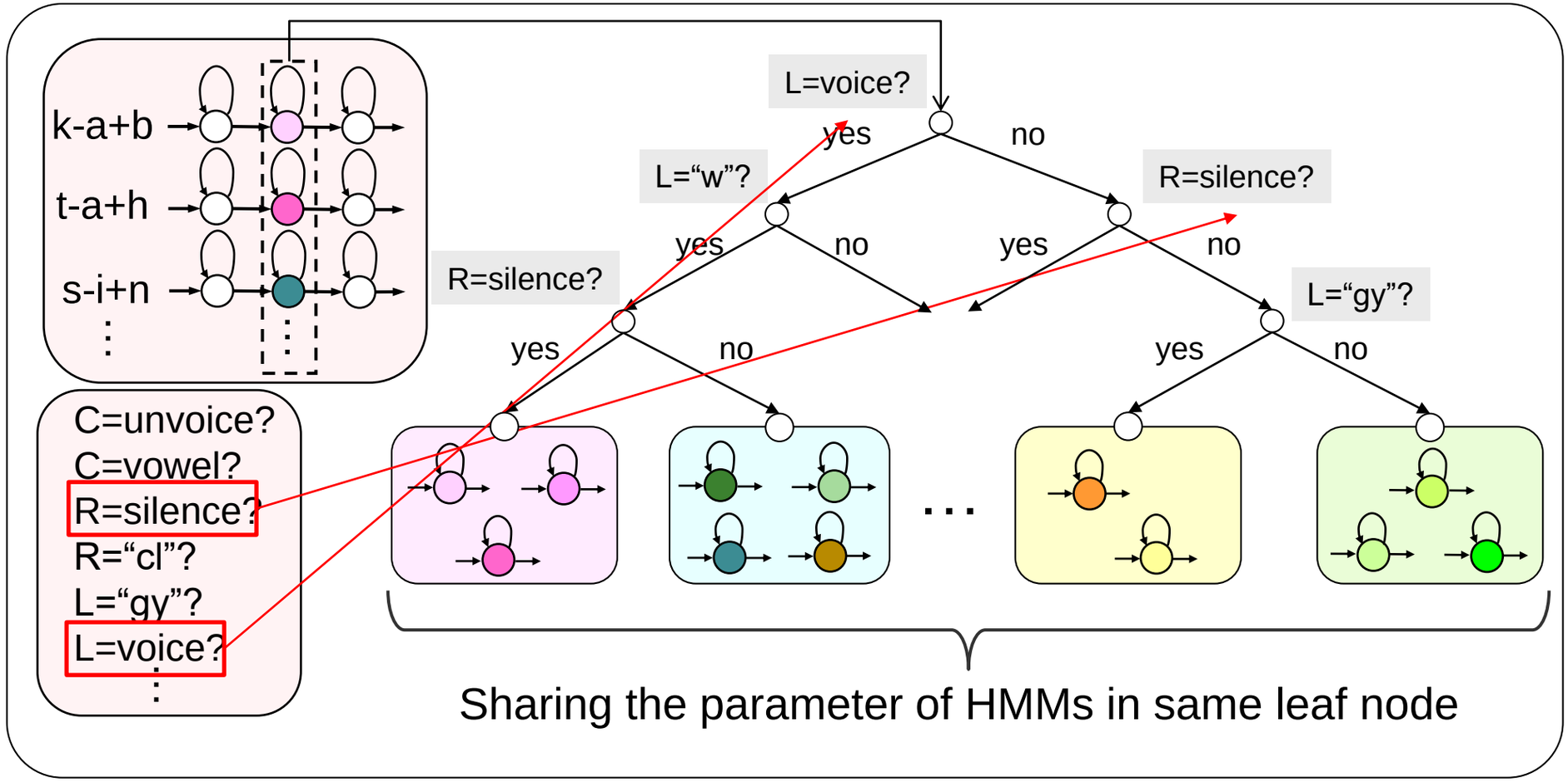
Phrase

- # of syllables in {preceding, current, succeeding} phrase

…..

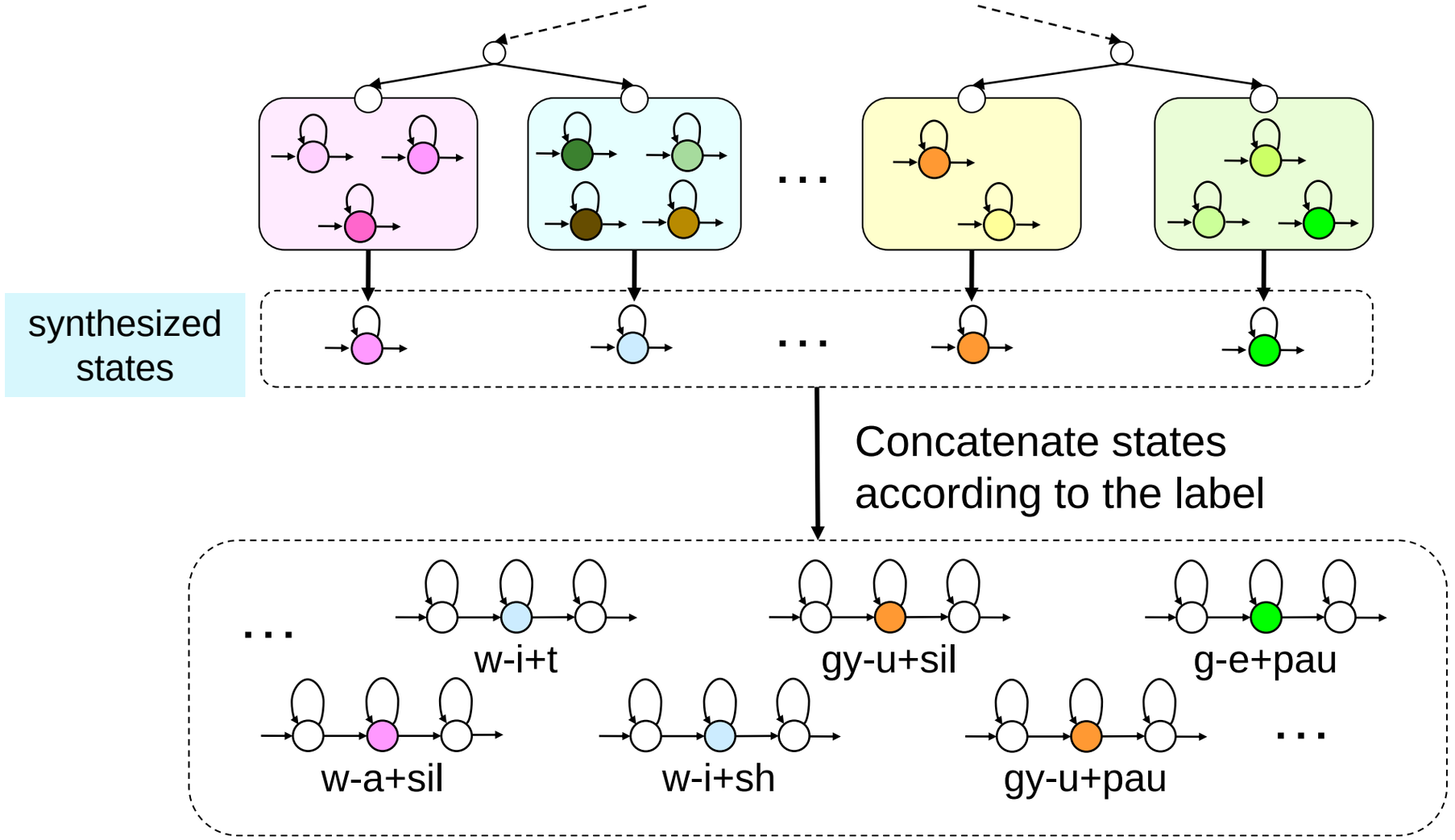
Huge # of combinations ⇒ Difficult to have all possible models

Decision Tree-based State Clustering [Odell; '95]

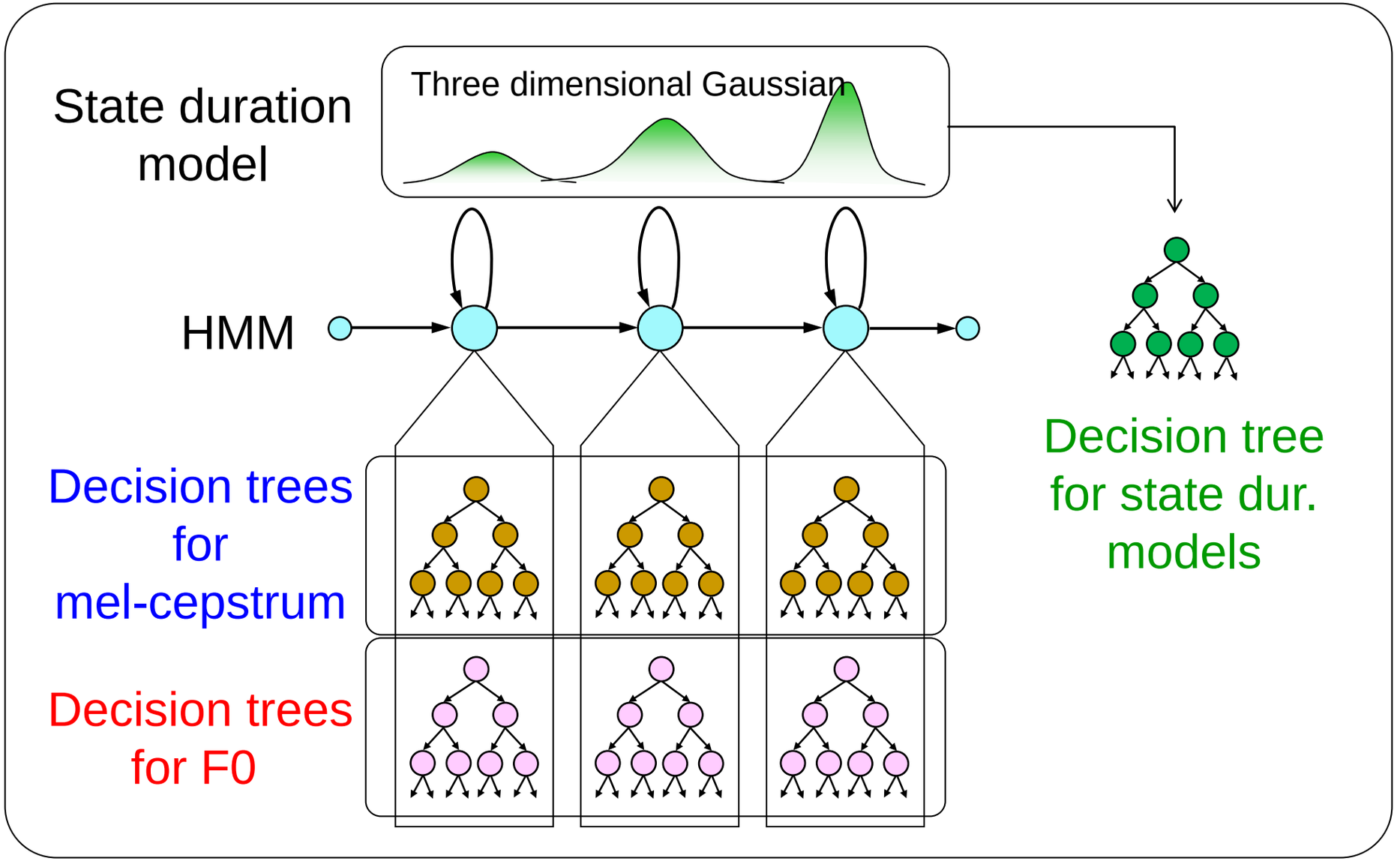


– Each state separated automatically by the optimum question
– The optimum question determined for increasing likelihood

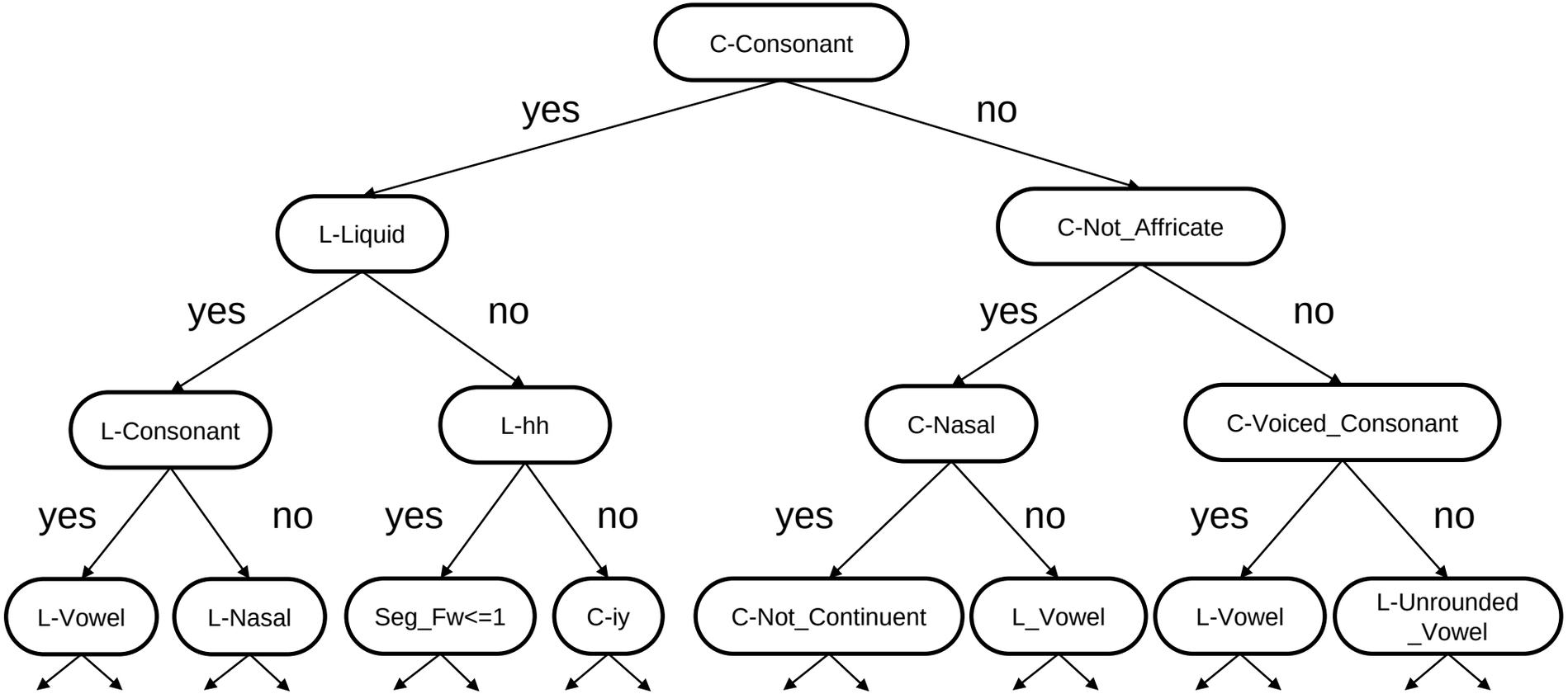
Synthesize From Leaf Nodes



Stream-dependent Tree-based Clustering

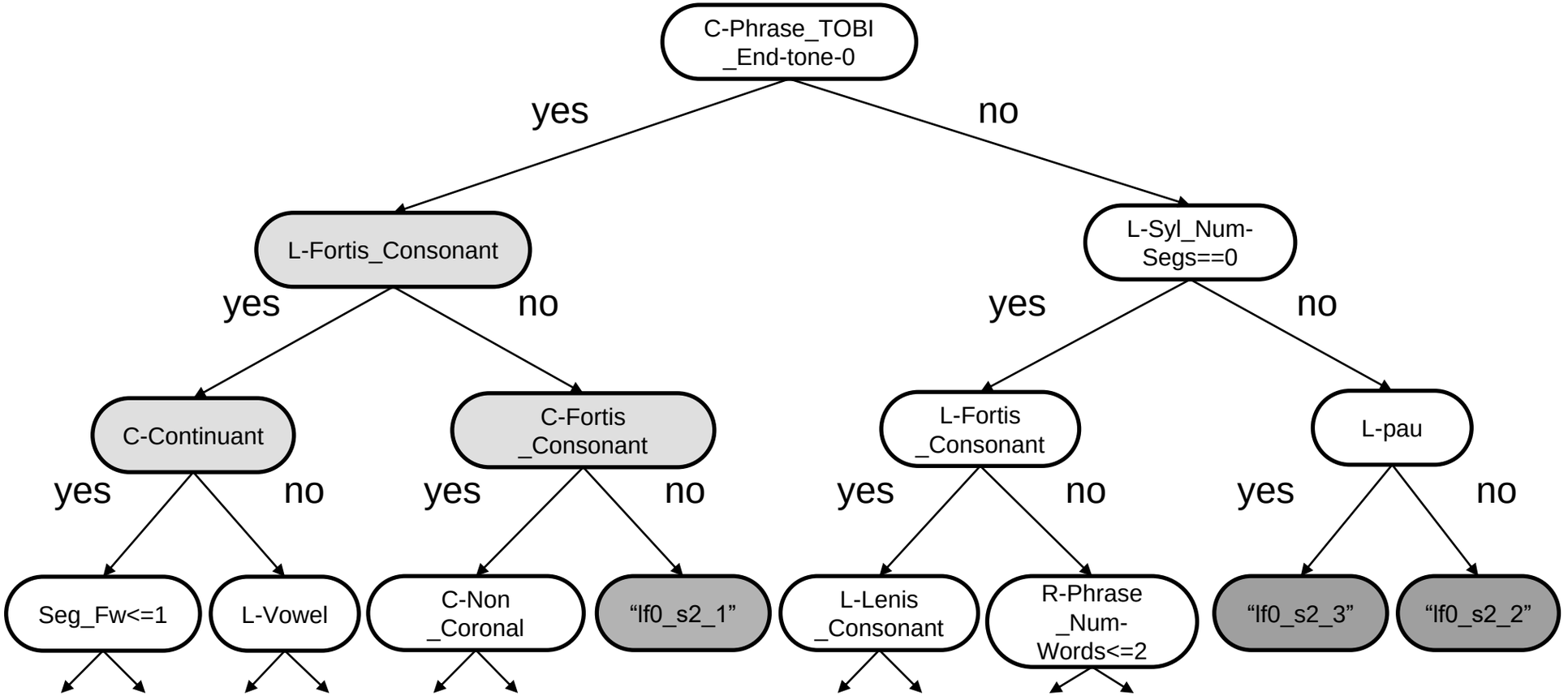


Tree for Spectrum (1st state)



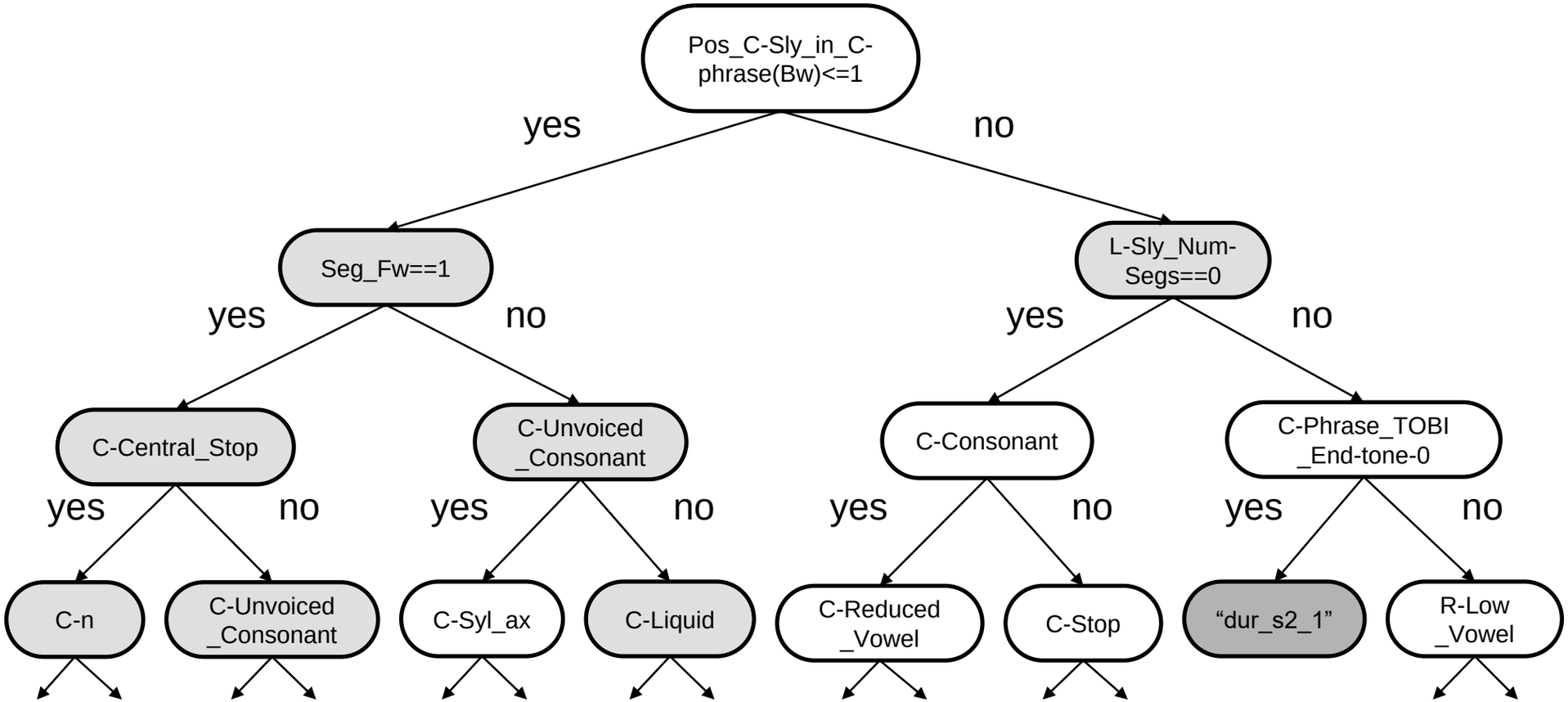
- Questions about phonetic attributes

Tree for F0 (1st state)



– Questions about linguistic attributes

Tree for State Duration

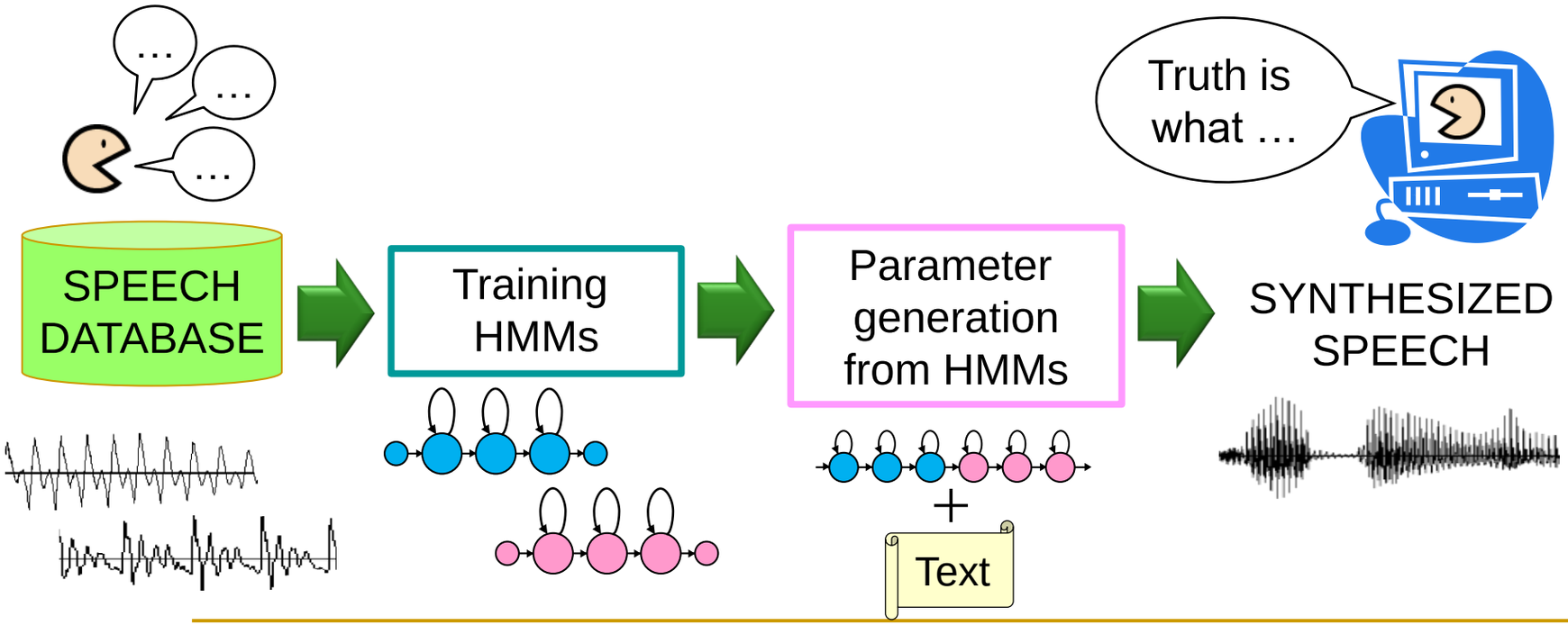


- Linguistic questions for pause
- Phonetic questions for speech

Speech Synthesis from HMM

- **Speech parameters are generated from HMMs**
  - Spectrum parameters
  - Excitation parameters (F0)
- **Vocoding parameters to synthesize speech**

  ⇒ Obtain high-quality synthesized speech

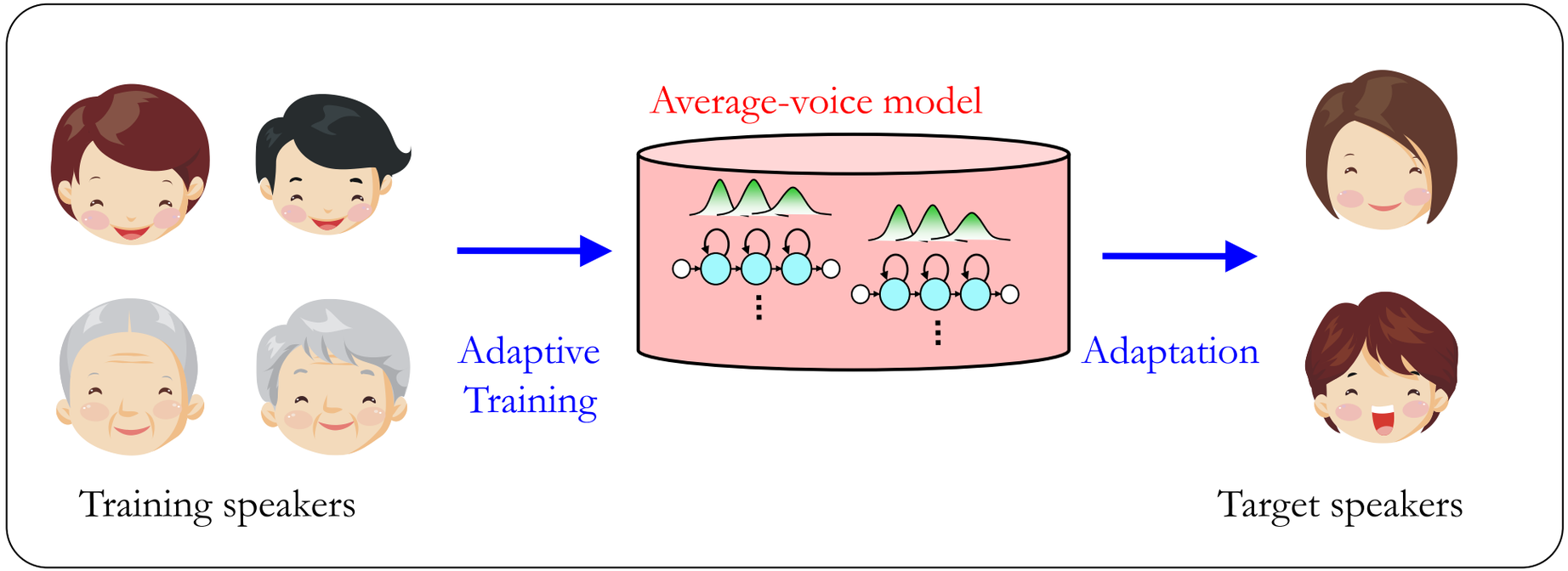


Outlines

- **Speech vocoding: Source-filter model**
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - State duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- **Voice character controlling**
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- **Application**
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

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

⇒ small cost to create new voices

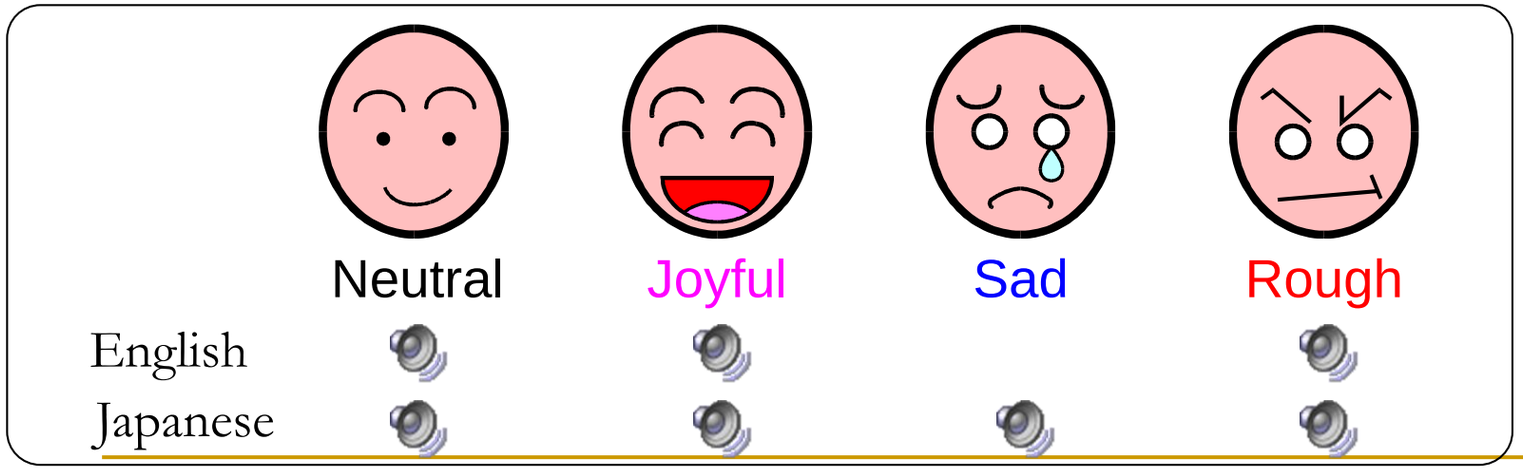
Adaptation Demo

- **Speaker adaptation**

  Original voice:

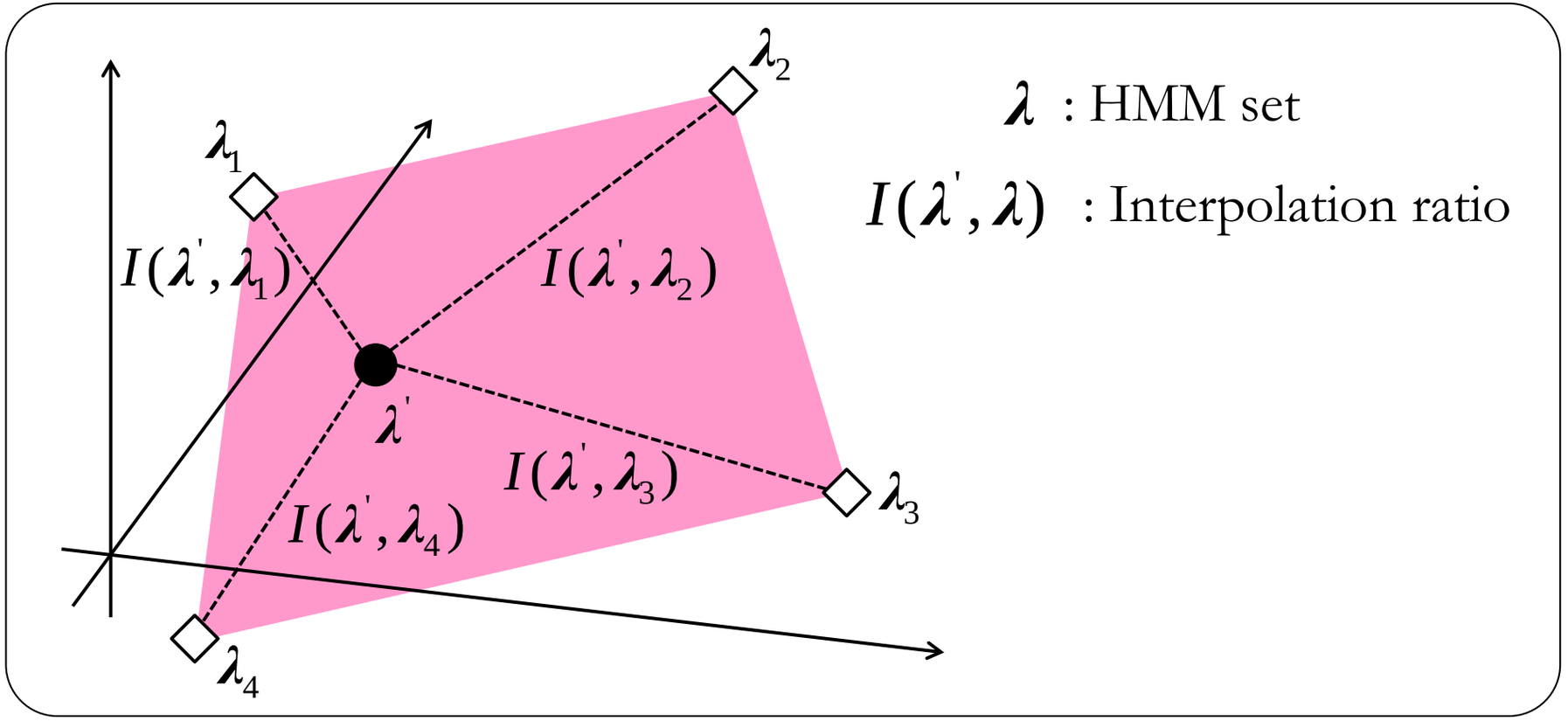
Average voice model	Number of adaptation sentences			
	10 sentences	100 sentences	500 sentences	1132 sentences

- **Style adaptation**



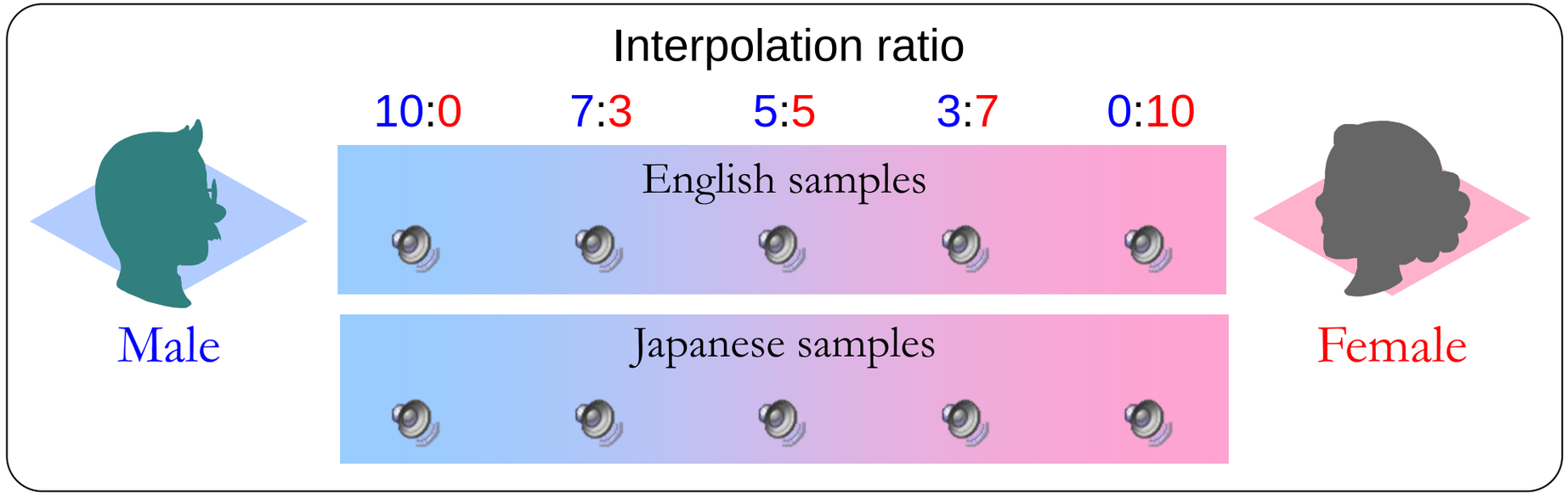
Interpolation (Mixing Voices)

- **Interpolate parameters among representative HMM sets**
  - Create new voices even if no adaptation data is available
  - Gradually change speaker & speaking styles [Yoshimura; ’97, Tachibana; ’05]

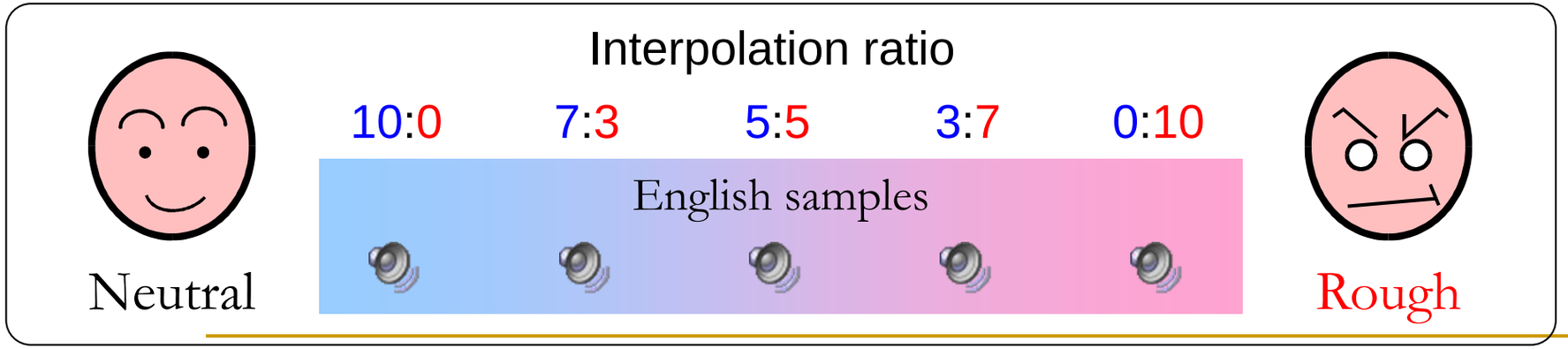


Interpolation Demo

- **Speaker interpolation**



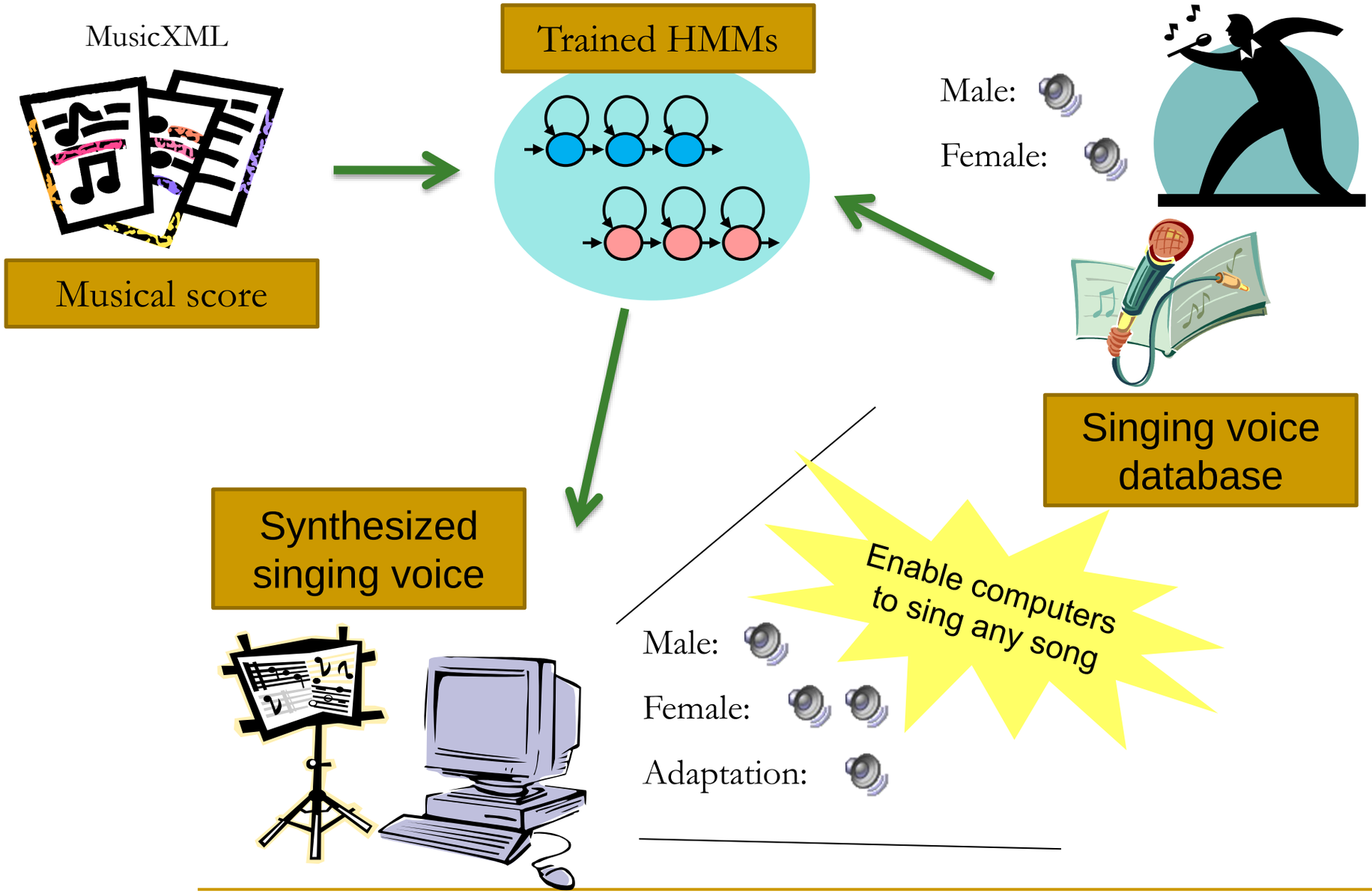
- **Style interpolation**



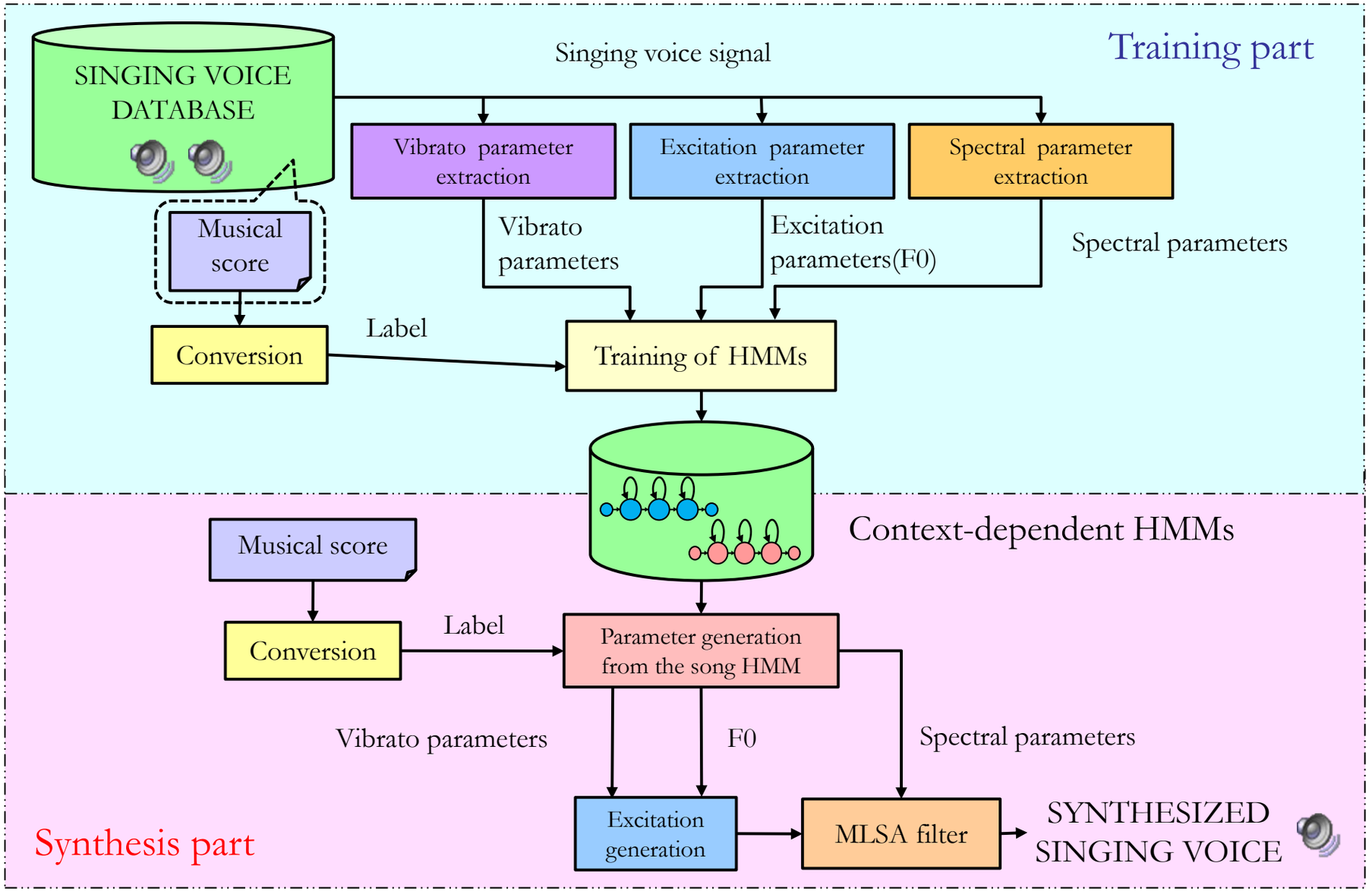
Outlines

- **Speech vocoding: Source-filter model**
- **Speech parameter modeling and generation with HMM**
  - Overview of HMM framework
  - State duration modeling
  - Spectrum modeling
  - F0 modeling
  - Context clustering
- **Voice character controlling**
  - Adaptation (mimicking voices)
  - Interpolation (mixing voices)
- **Application**
  - Singing voice
  - Emotional voice
  - Audio-visual speech synthesis

Singing Voice Synthesis [Oura; '10]

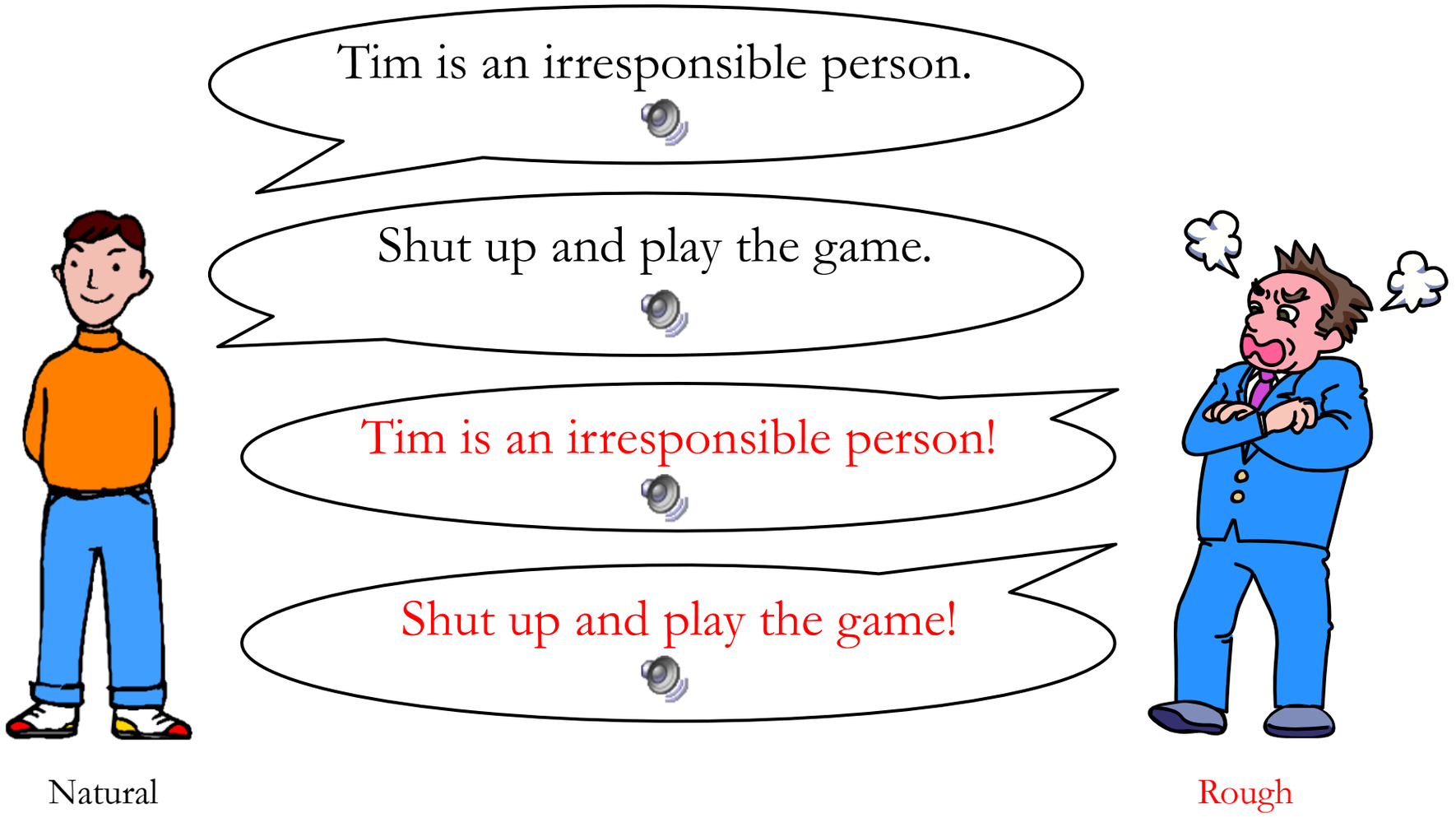


Singing Voice Synthesis



Emotional Speech Synthesis [Tsuzuki; '04]

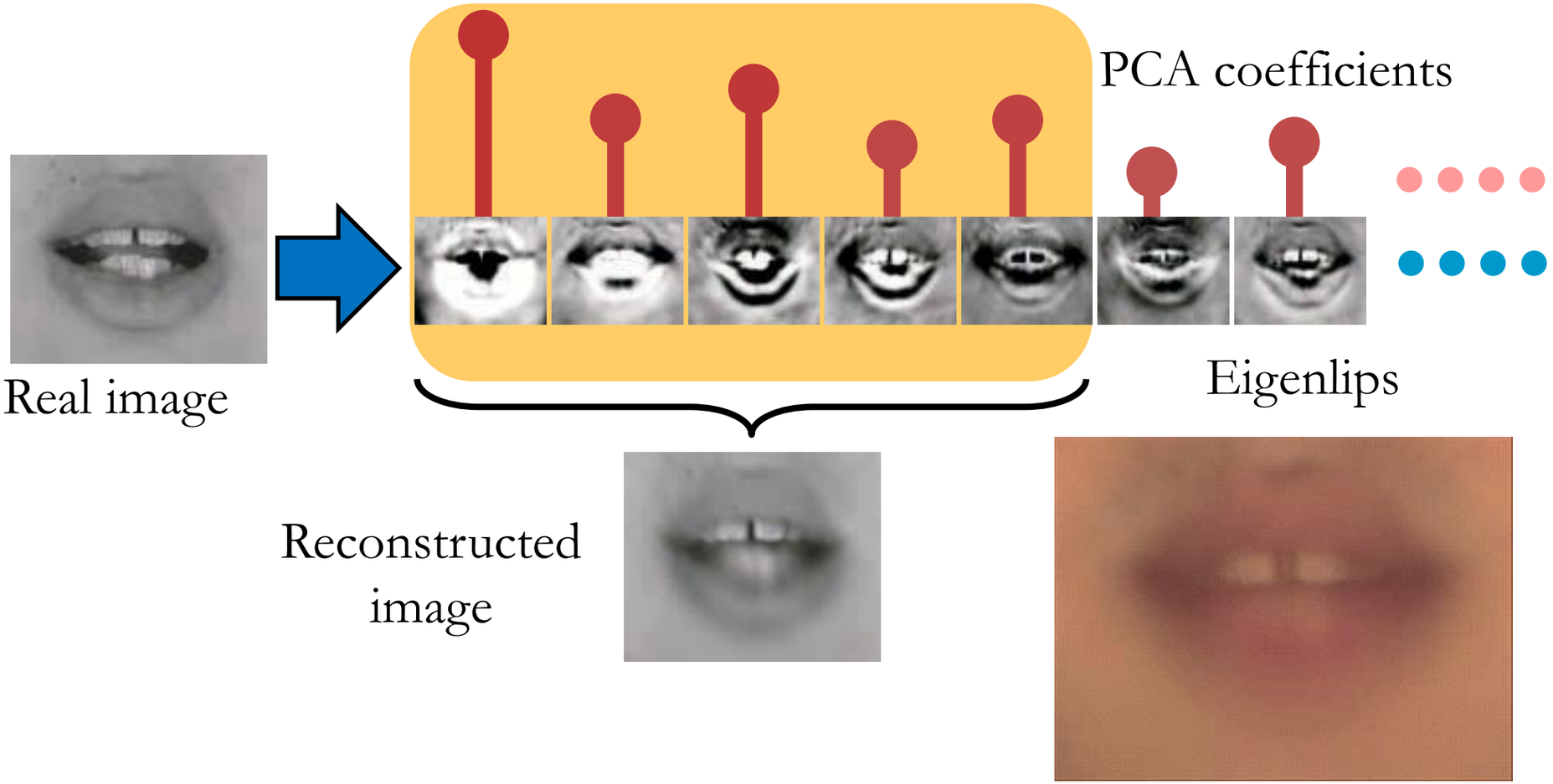
Using emotional voices for training data



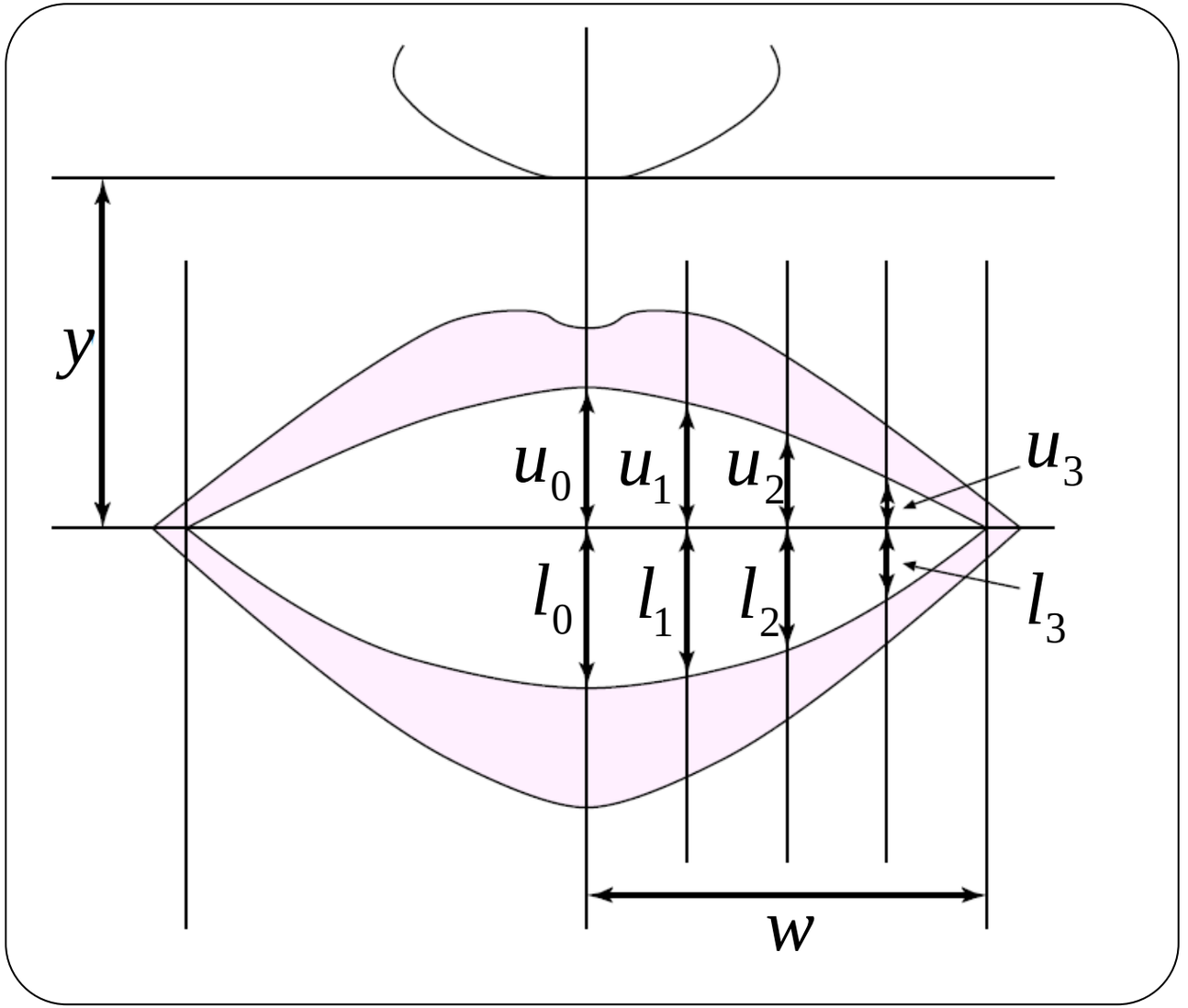
Audio-Visual Synthesis (Pixel-based) [Sako; '00]

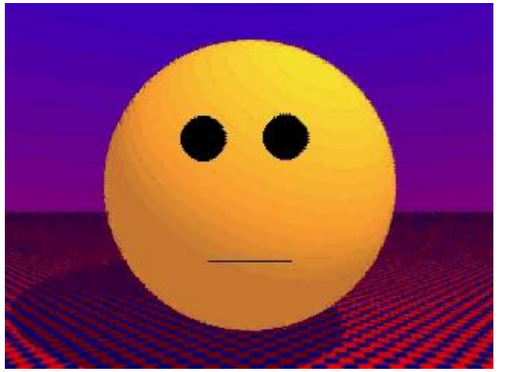
- **Pixel image: high dimensionality**

  ⇒Dimensionality reduction by PCA



Audio-Visual Synthesis (Model-based) [Tamura; '98]





References (1/4)

- Sagisaka;'92 - "ATR nu-TALK speech synthesis system," ICSLP, '92.
- Black;'96 - "Automatically clustering similar units...," Euro speech, '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 (2/4)

- 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 (3/4)

- 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 trended 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 meeting of ASJ, '03.
- Yu;'07 - "A novel HMM-based TTS system using both continuous HMMs and discrete...," ICASSP, '07.

References (4/4)

- 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 Technical report,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. Consumer Elec., 52(4), '06.
- Klatt;'82 - "The Klatt-Talk text-to-speech system," ICASSP, '82.