TOSHIBA
Leading Innovation >>>

Statistical Speech Synthesis

Heiga ZEN

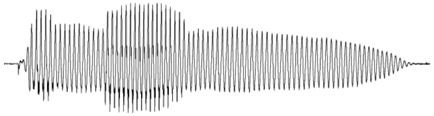
Toshiba Research Europe Ltd.
Cambridge Research Laboratory

Speech Synthesis Seminar Series @ CUED, Cambridge, UK
January 11th, 2011

Text-to-speech as a mapping problem

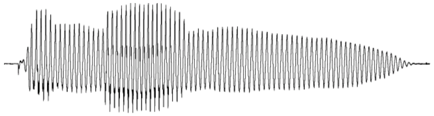
Text-to-speech synthesis (TTS)

Text (seq of discrete symbols) → Speech (continuous time series)

Good morning → 

Automatic speech recognition (ASR)

Speech (continuous time series) → Text (seq of discrete symbols)

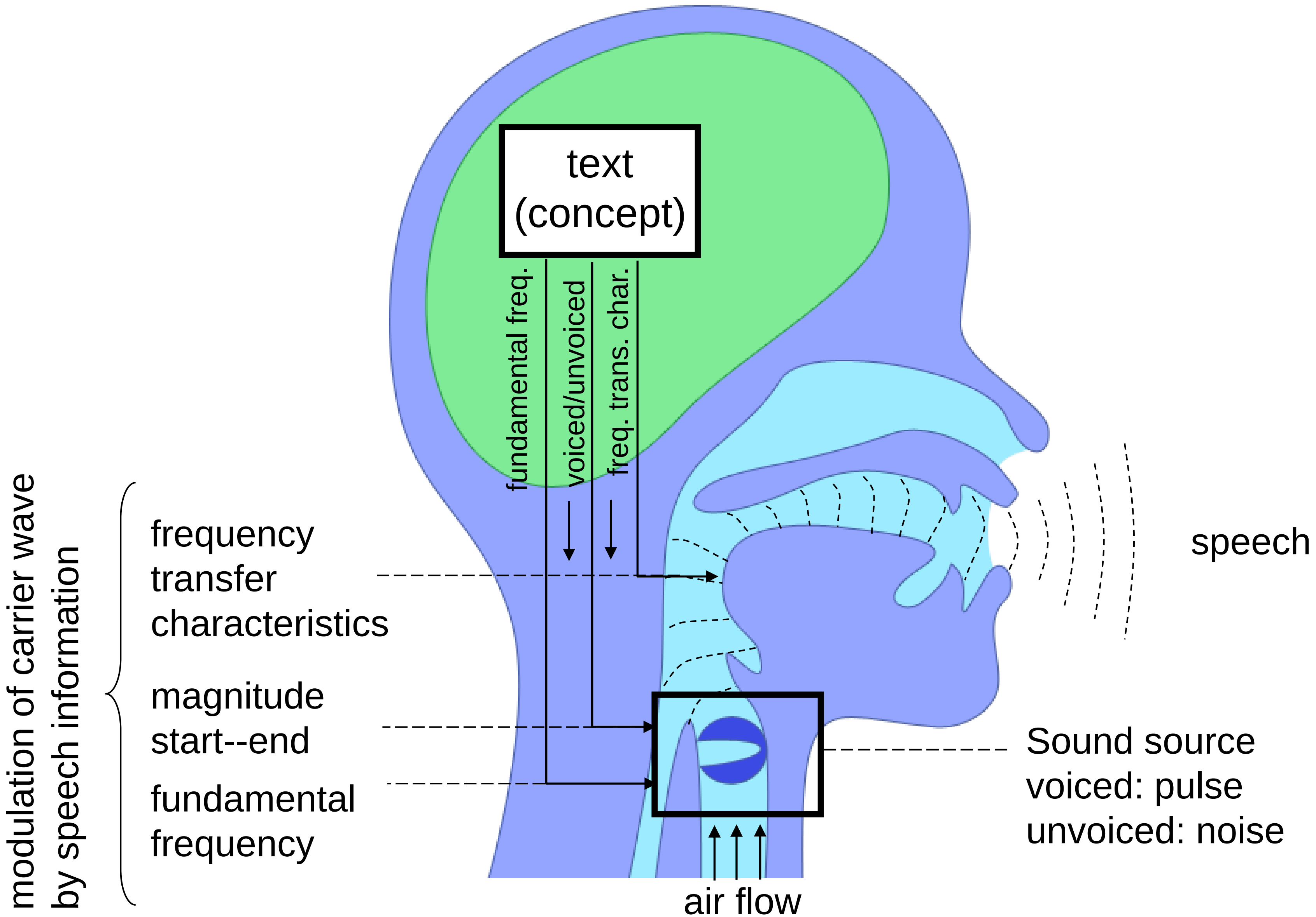
 **→ Good morning**

Machine Translation (MT)

Text (seq of discrete symbols) → Text (seq of discrete symbols)

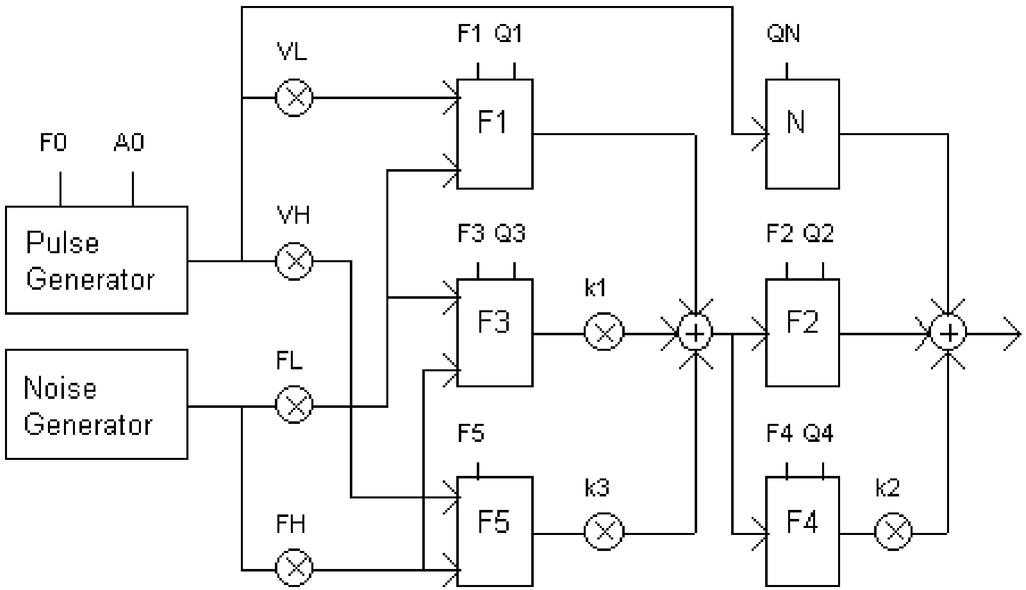
Dobré ráno → Good morning

Speech production process



Speech synthesis methods (1)

Rule-based, *formant synthesis* (~'90s)



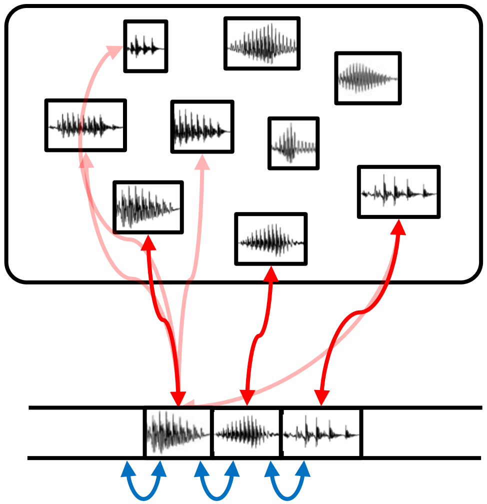
Block diagram of KlattTalk

- Based on parametric representation of speech
- Hand-crafted rules to control phonetic unit

DECtalk (or KlattTalk / MITTalk) [Klatt;'82]

Speech synthesis methods (2)

Corpus-based, *concatenative synthesis* ('90s~)



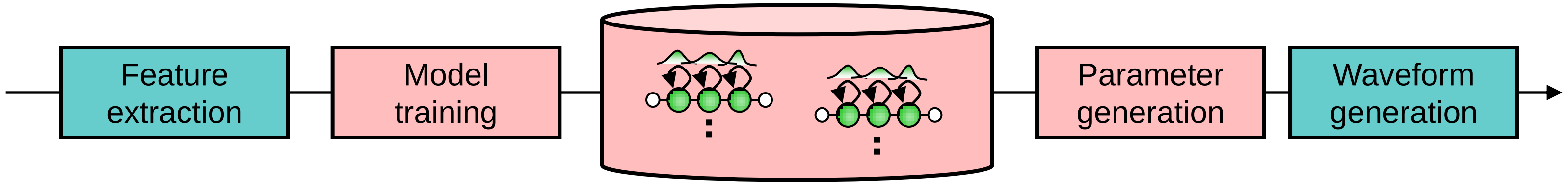
- Concatenate small speech units (e.g., phone) from a database
- Large data + automatic learning → High-quality synthetic voices

Single inventory; diphone synthesis [Moullnes;'90]

Multiple inventory; unit selection synthesis [Sagisaka;'92, Black;'96]

Speech synthesis methods (3)

Corpus-based, *statistical parametric synthesis* (mid ’90s~)



- Large data + automatic training

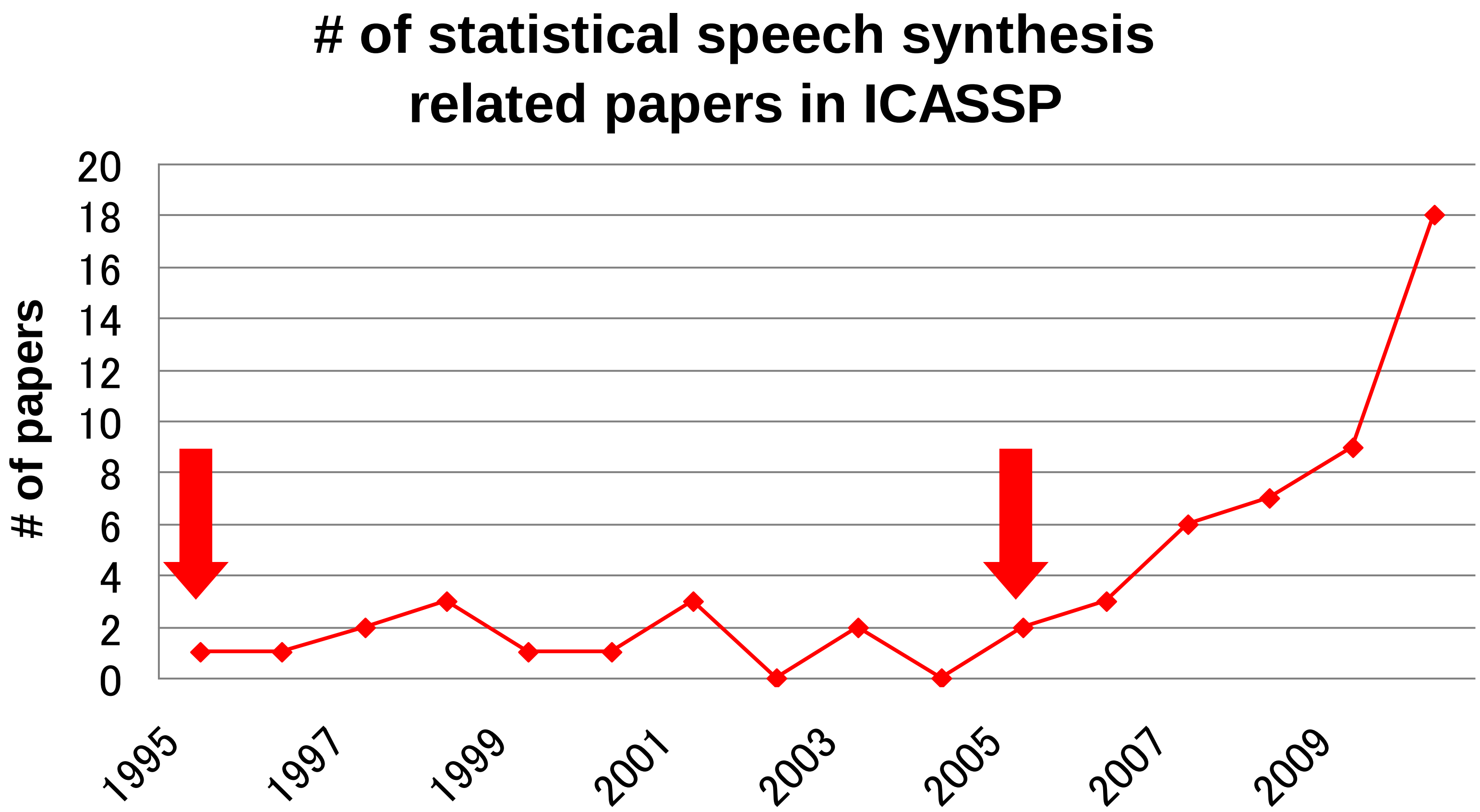
  → Automatic voice building
- Source-filter model + statistical modeling

  → Flexible to change its voice characteristics

Hidden Markov models (HMMs) as its statistical acoustic model

→ **HMM-based speech synthesis (HTS) [Yoshimura;‘02]**

Popularity of statistical speech synthesis



Aim of this talk

Statistical speech synthesis is getting popular, but…

not many researchers fully understand how it works

Formulate & understand the whole corpus-based speech synthesis process in a unified statistical framework

Outline

HMM-based speech synthesis

- Overview
- Implementation of individual components

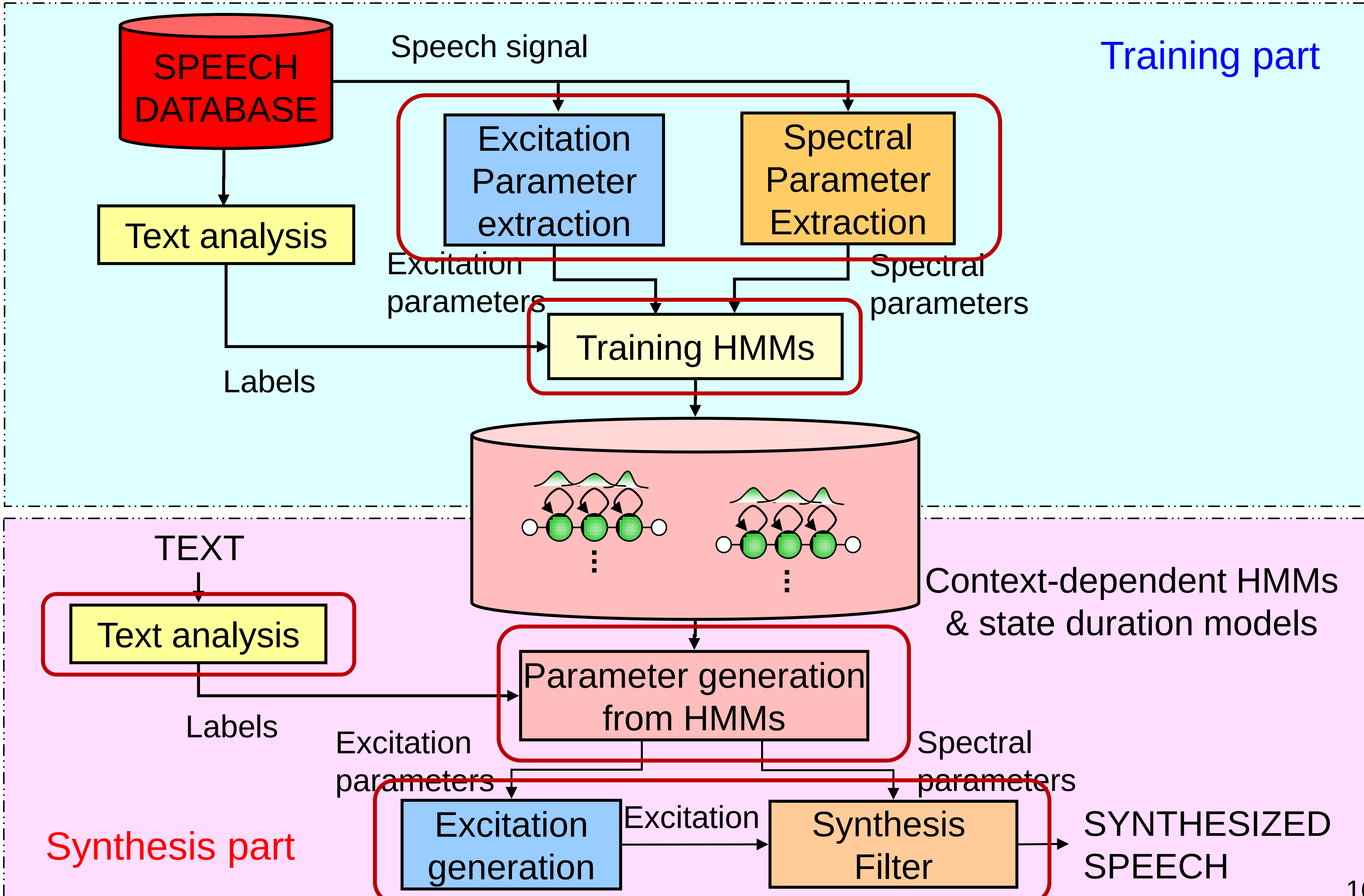
Bayesian framework for speech synthesis

- Formulation
- Realizations in HMM-based speech synthesis
- Recent works

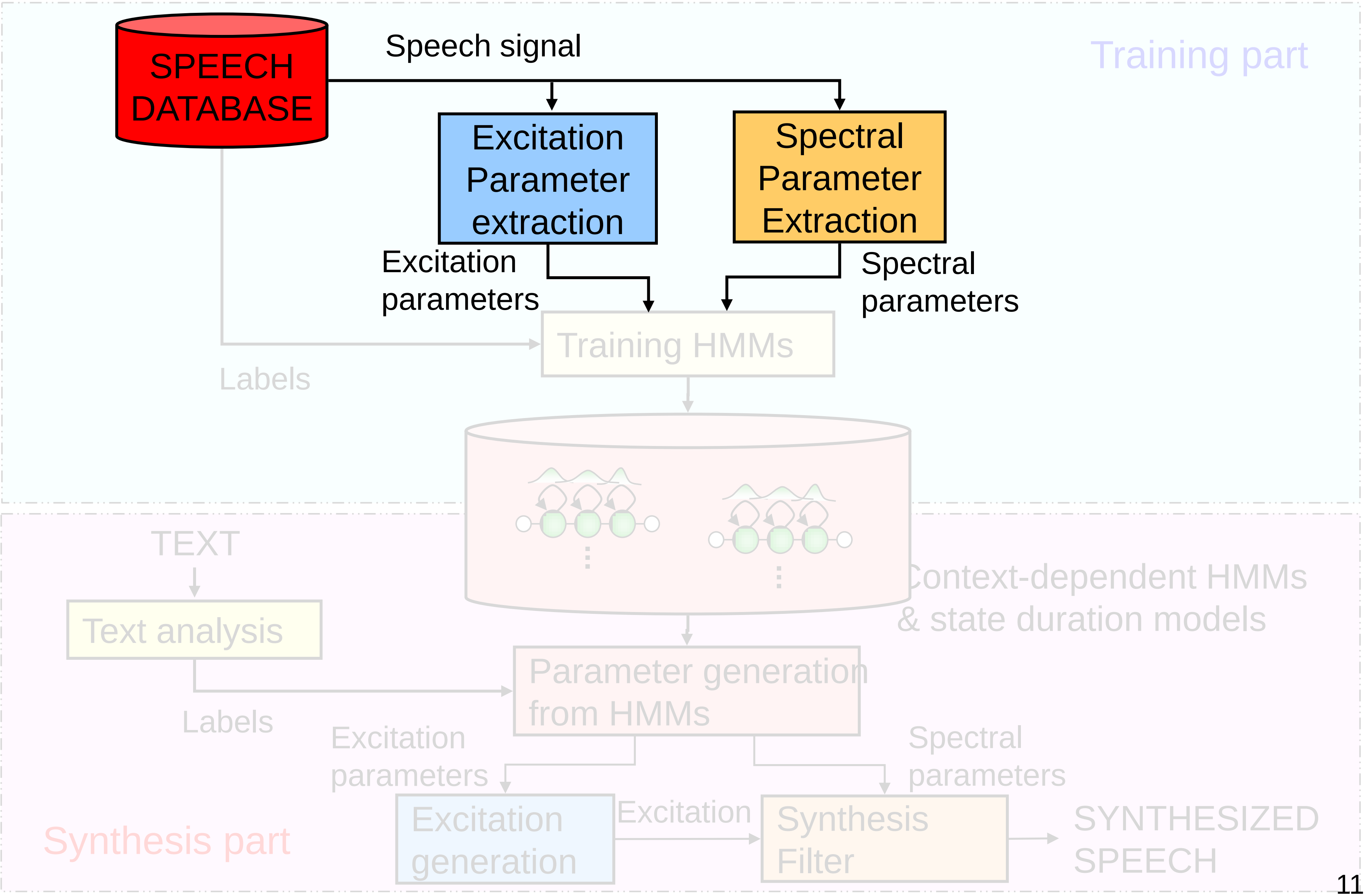
Conclusions

- Summary
- Future research topics

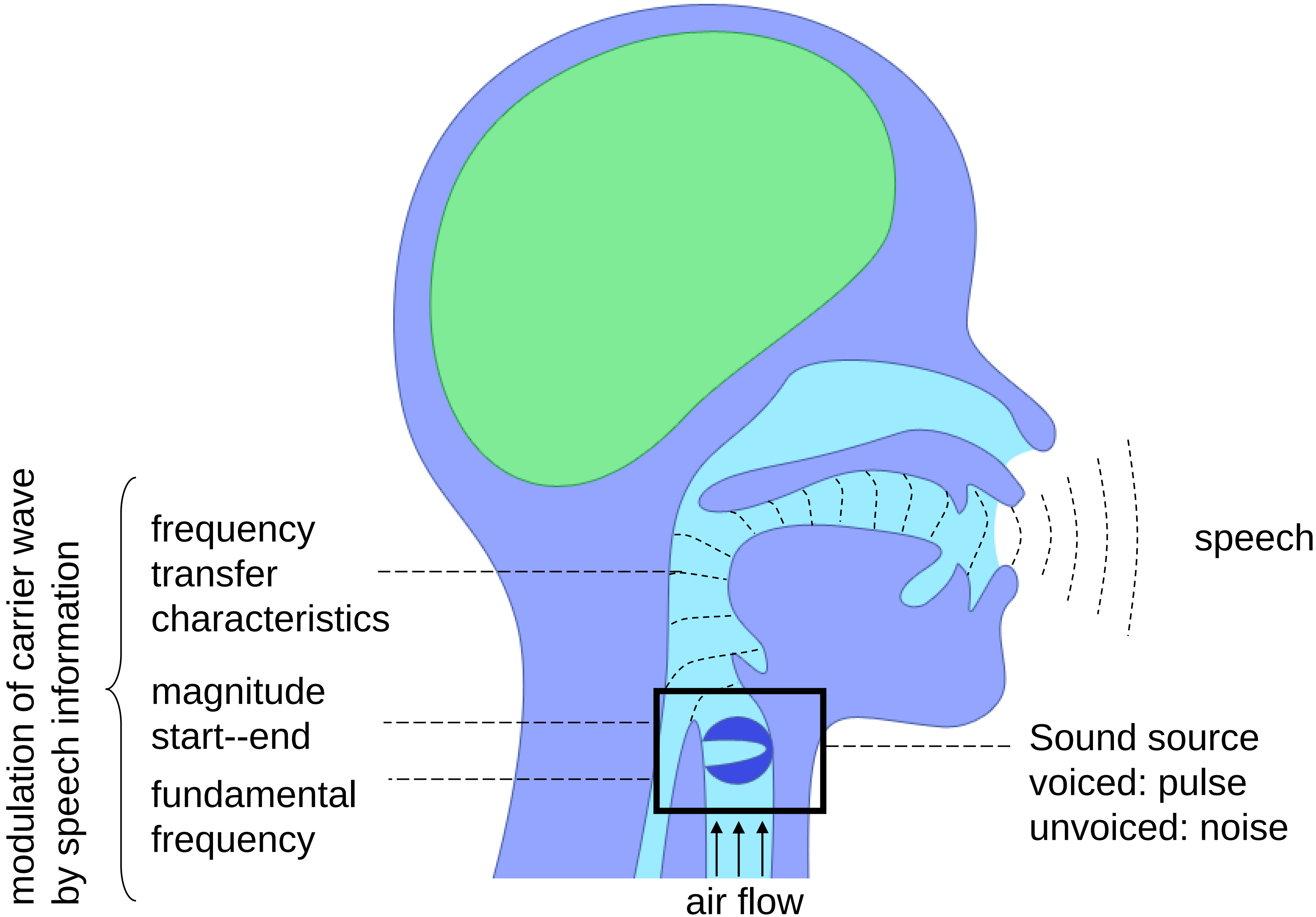
HMM-based speech synthesis system (HTS)



HMM-based speech synthesis system (HTS)



Speech production process

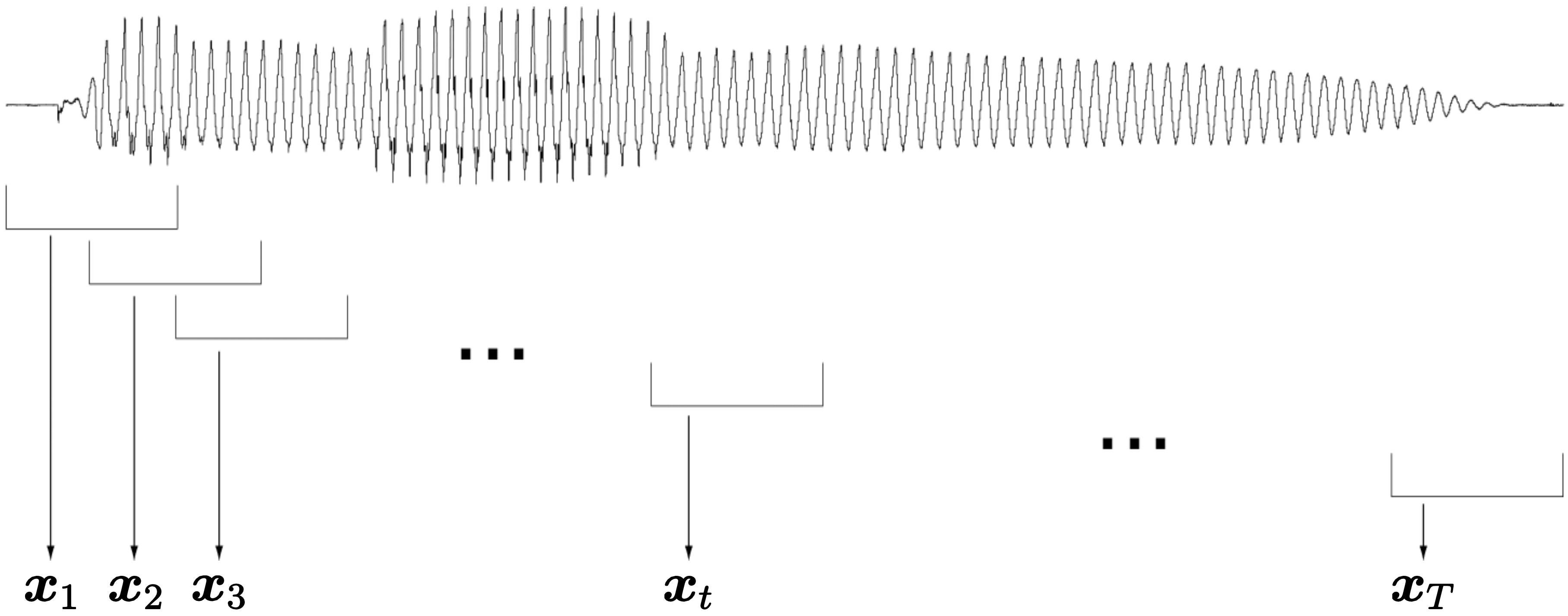


Divide speech into frames

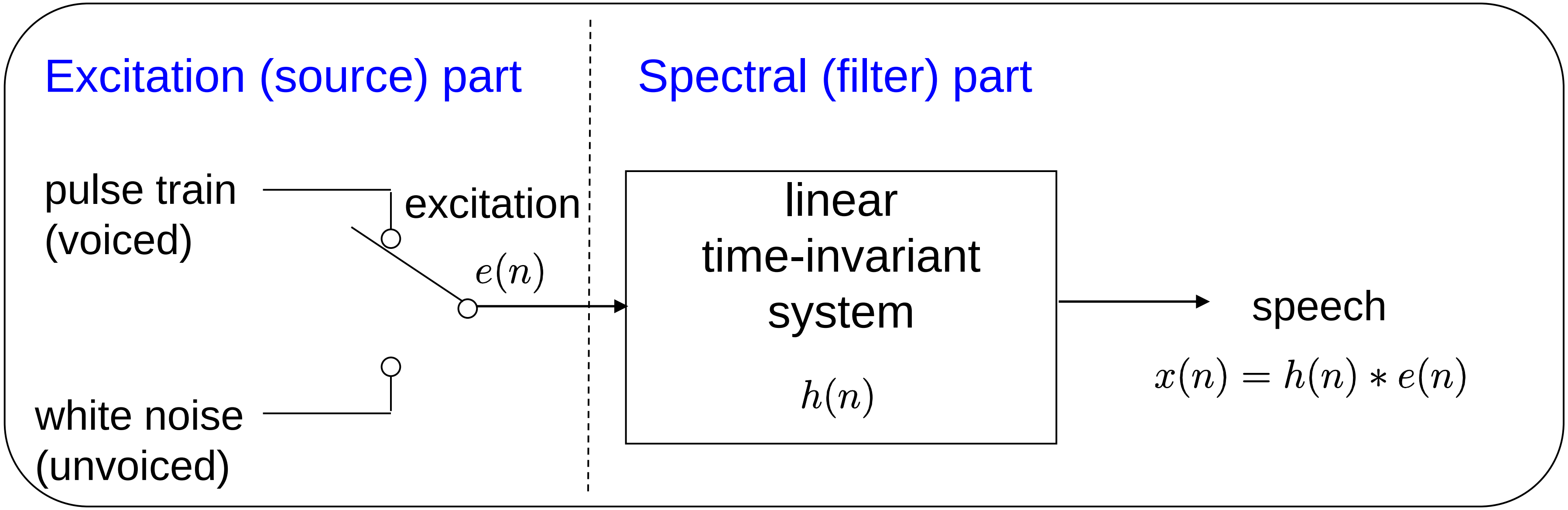
Speech is a non-stationary signal

… but can be assumed to be quasi-stationary

→ Divide speech into short-time frames (e.g., 5ms shift, 25ms length)



Source-filter model



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

$\downarrow$ Fourier transform

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

Spectral (filter) model

Parametric models speech spectrum

Autoregressive (AR) model

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

Exponential (EX) model

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

ML estimation of spectral model parameters

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

$\boldsymbol{c}_t = [c_t(0), \ldots, c_t(M)]^\top$:

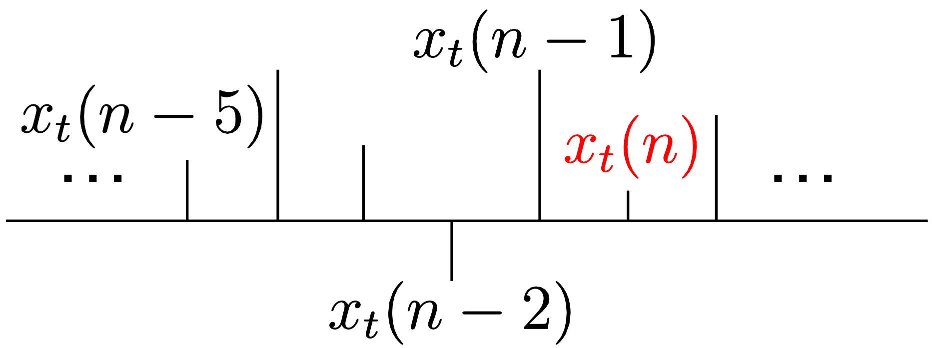
spectral model parameters

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

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

LP analysis (1)

$\boldsymbol{x}_t = [x_t(1), x_t(2), \ldots, x_t(N)]^\top$ short-time windowed speech waveform



LP analysis assumes that $\boldsymbol{x}_t$ is a sample from M-th order AR process

$$x_t(n) = \sum_{m=1}^{M} c_t(m) x_t(n-m) + \epsilon_t(n)$$ Linear AR process

$$\boldsymbol{c}_t = [c_t(0), c_t(1), \ldots, c_t(M)]^\top$$ M-th order LP coefficients

$$\epsilon_t(n) \sim \mathcal{N}(0, c_t(0))$$ Gaussian noise

LP analysis (2)

If we set $\boldsymbol{\Psi}$ as

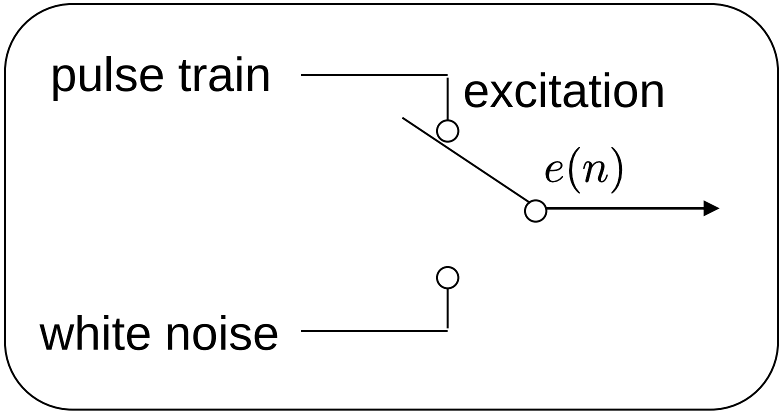
$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & & & & & 0 \\ -c_t(1) & \ddots & & & & \\ \vdots & \ddots & \ddots & & & \\ -c_t(M) & & \ddots & \ddots & & \\ & \ddots & & \ddots & \ddots & \\ 0 & & -c_t(M) & \dots & -c_t(1) & 1 \end{bmatrix}$$

then

$$p(\boldsymbol{x}_t \mid \boldsymbol{c}_t) = \mathcal{N}\left(\boldsymbol{x}_t \, ; \, \mathbf{0}, c_t(0) \left(\boldsymbol{\Psi}^\top \boldsymbol{\Psi}\right)^{-1}\right)$$

$$\hat{\boldsymbol{c}}_t = \arg\max_{\boldsymbol{c}_t} p(\boldsymbol{x}_t \mid \boldsymbol{c}_t) \quad \Rightarrow \text{LP analysis}$$

Excitation (source) model



Excitation model: pulse/noise excitation

– Voiced (periodic) → pulse trains
– Unvoiced (aperiodic) → white noise

Excitation model parameters

– V/UV decision
– V → fundamental frequency (F0): p_t

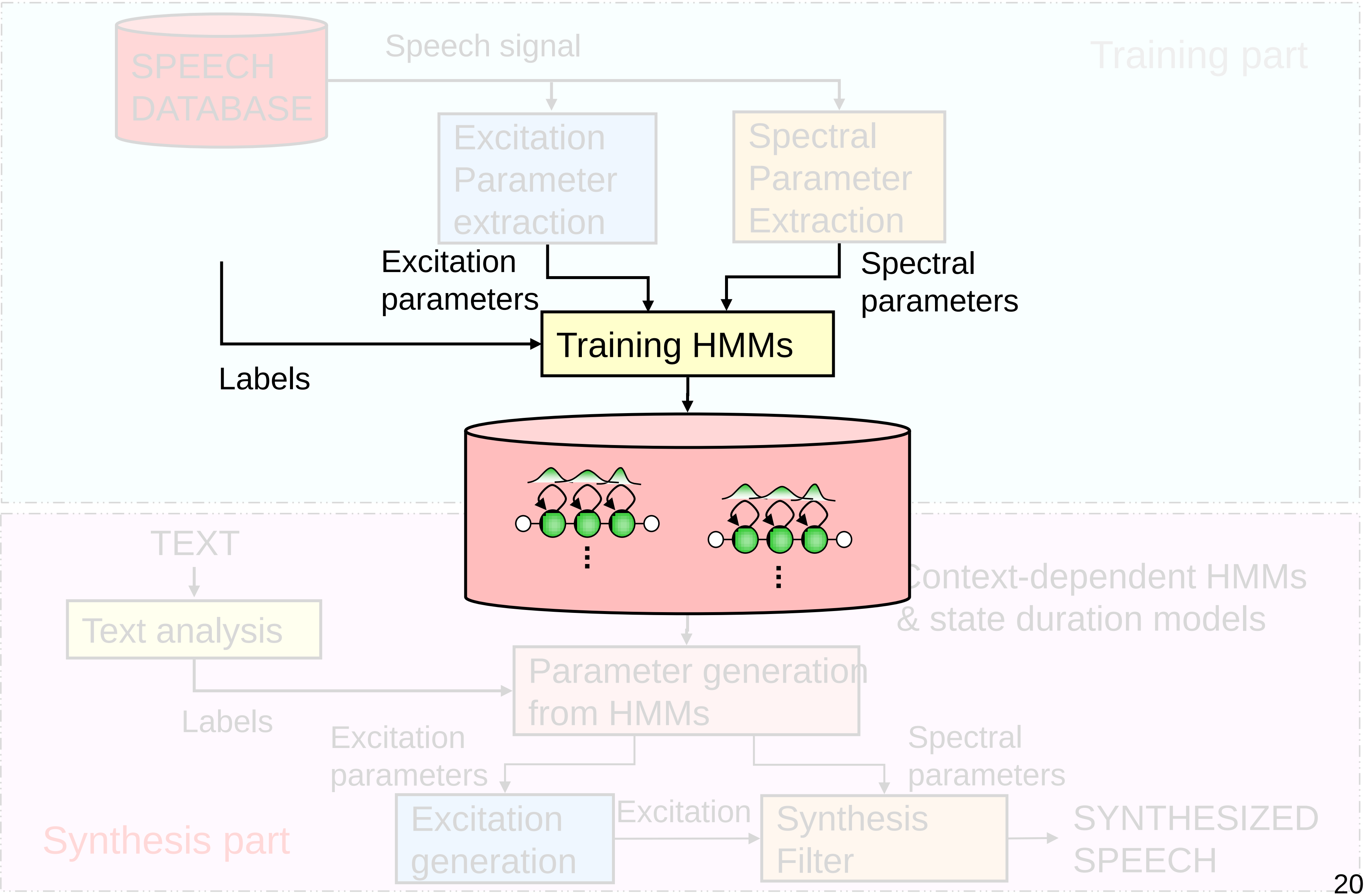
Speech samples

Natural speech

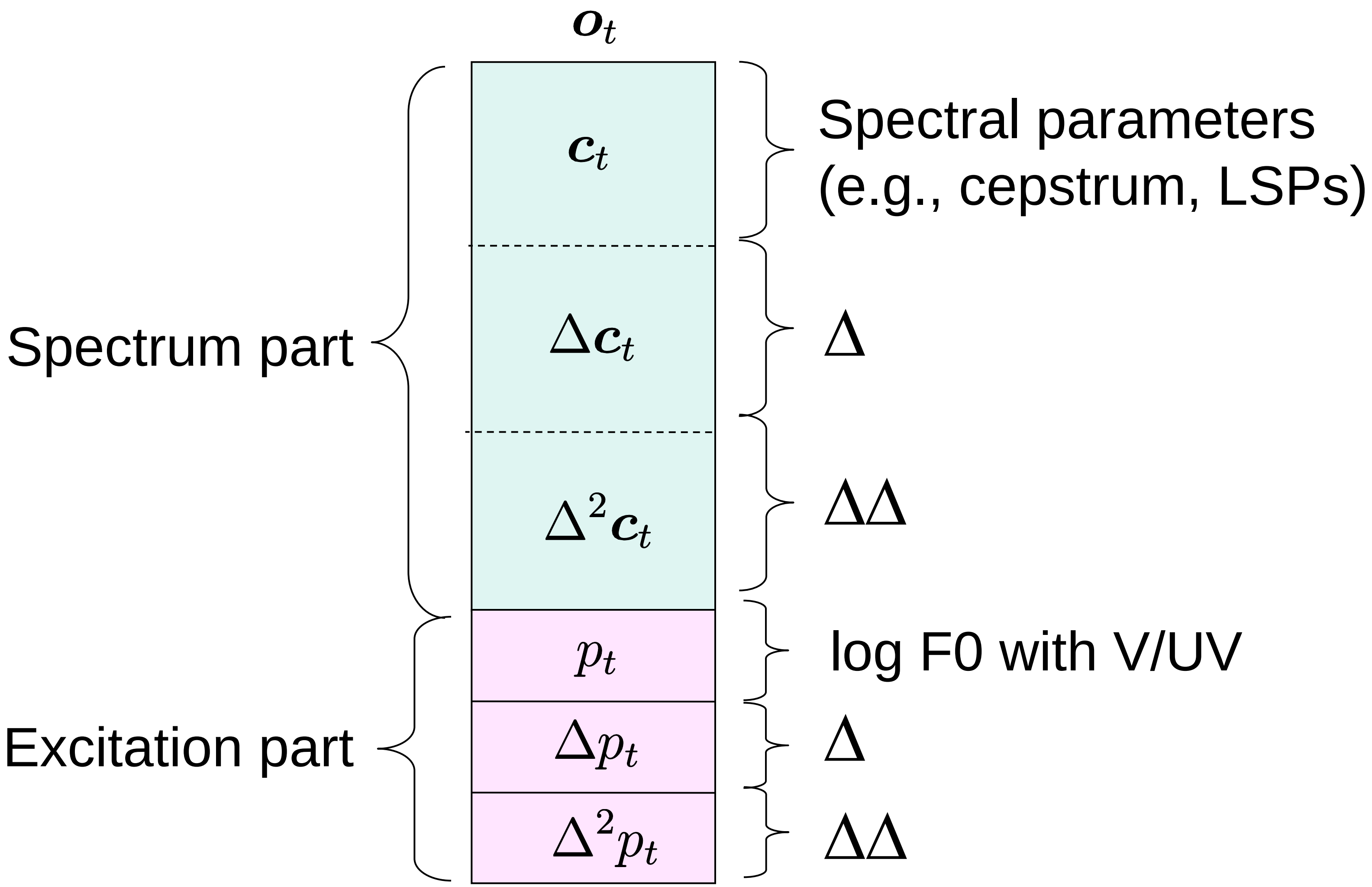
Reconstructed speech from extracted parameters (cepstral coefficients & F0 with V/UV decisions)

Quality degrades, but main characteristics are preserved

HMM-based speech synthesis system (HTS)



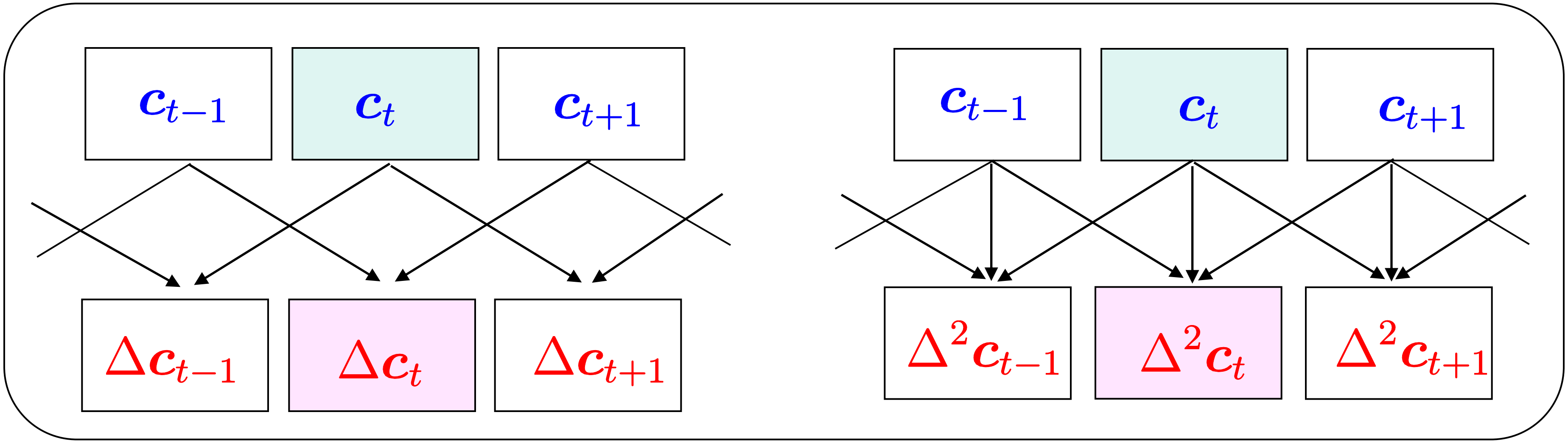
Structure of state-output (observation) vector



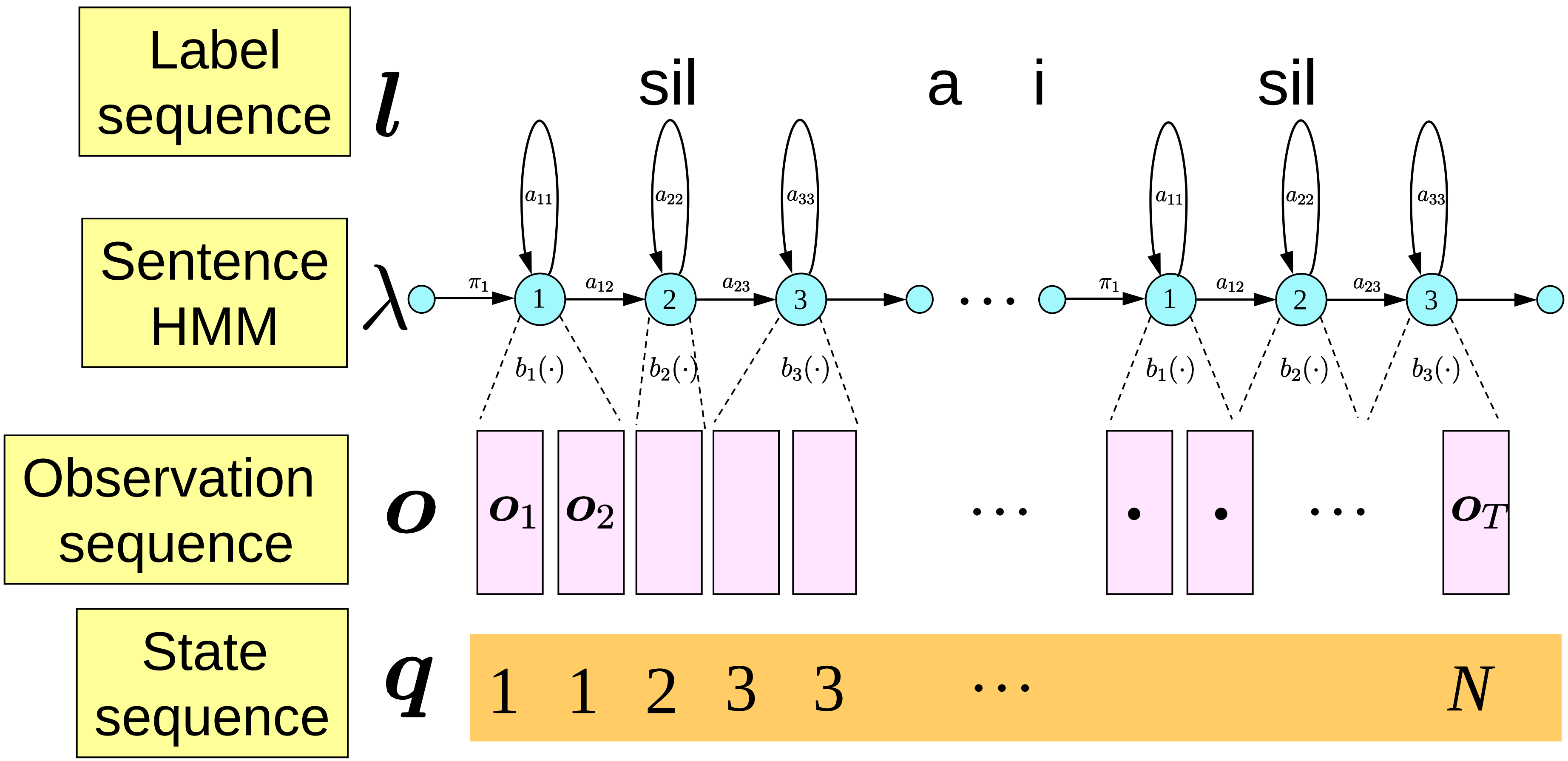
Dynamic features

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

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



HMM-based modeling

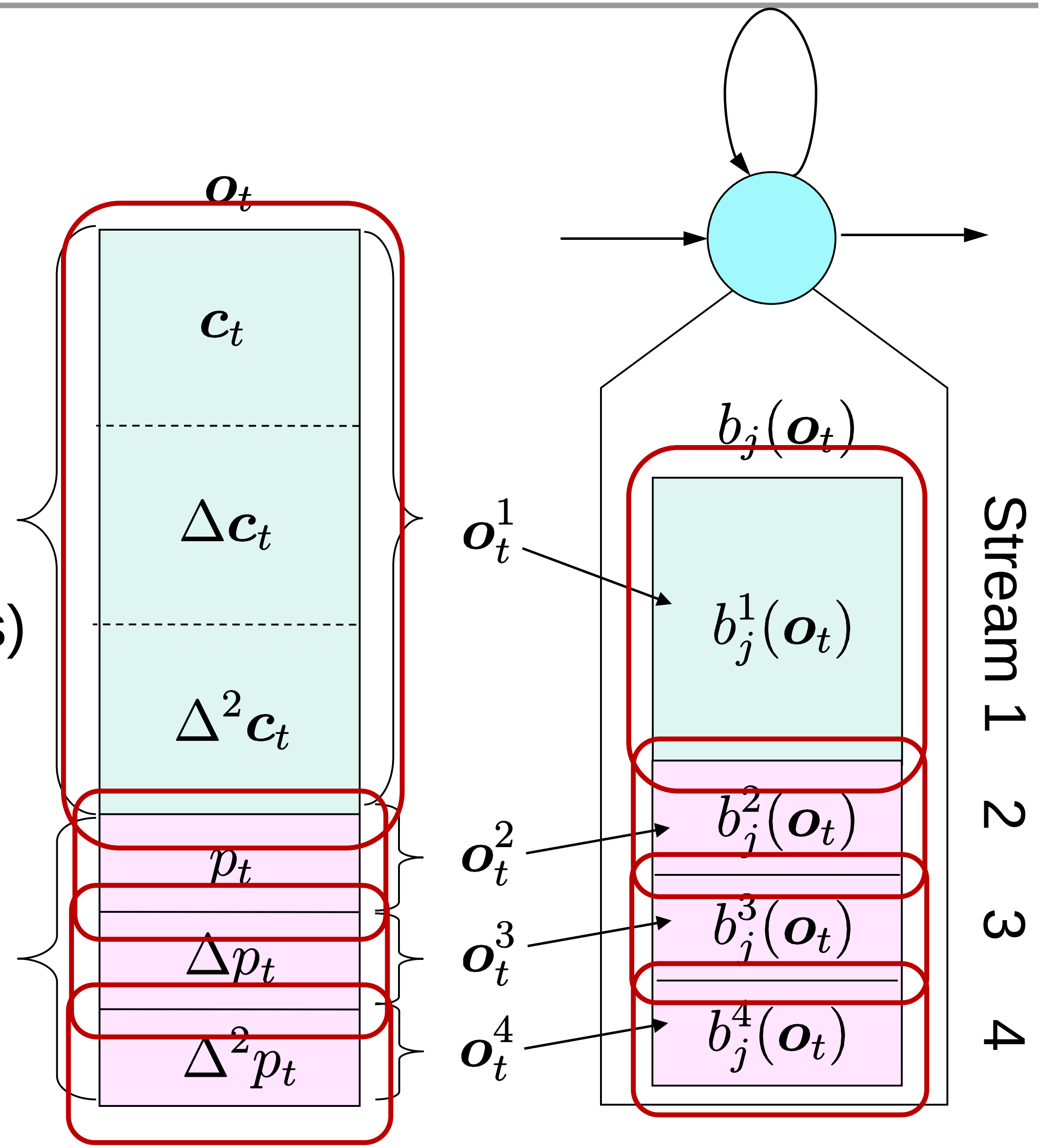


Multi-stream HMM structure

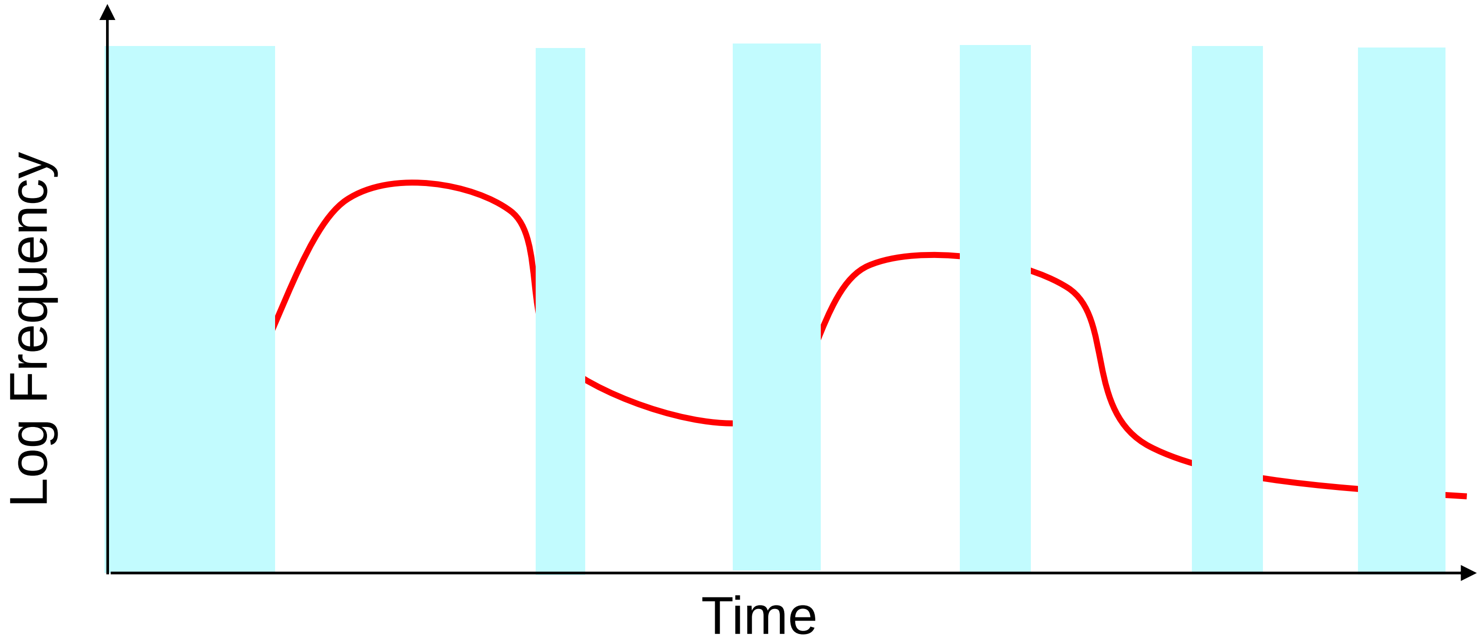
$$b_j(\boldsymbol{o}_t) = \prod_{s=1}^{S} \left\{ b_j^s(\boldsymbol{o}_t^s) \right\}^{w_s}$$

Spectrum
(cepstrum or LSP,
& dynamic features)

Excitation
(log F0
& dynamic features)

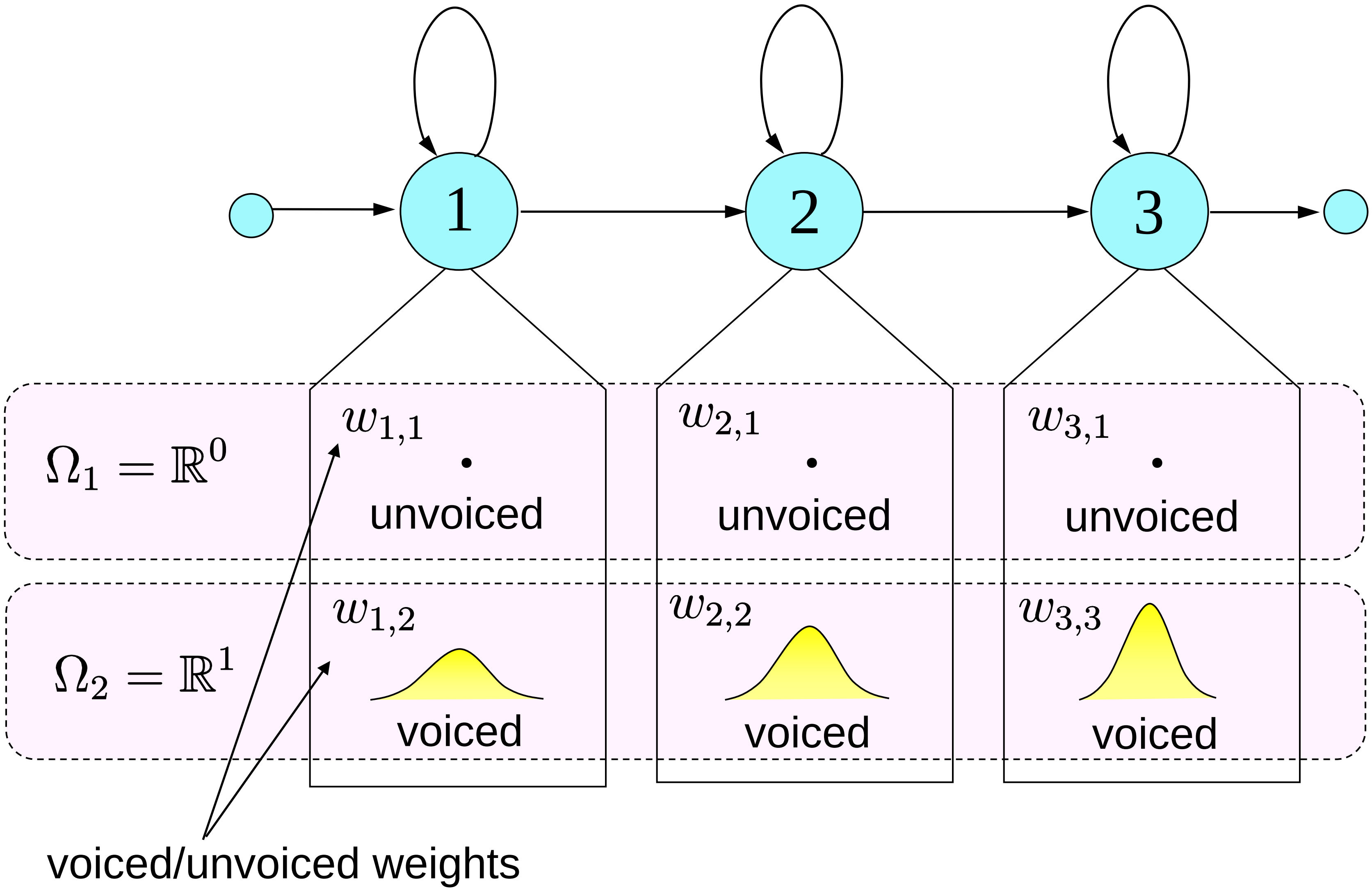


Observation of F0

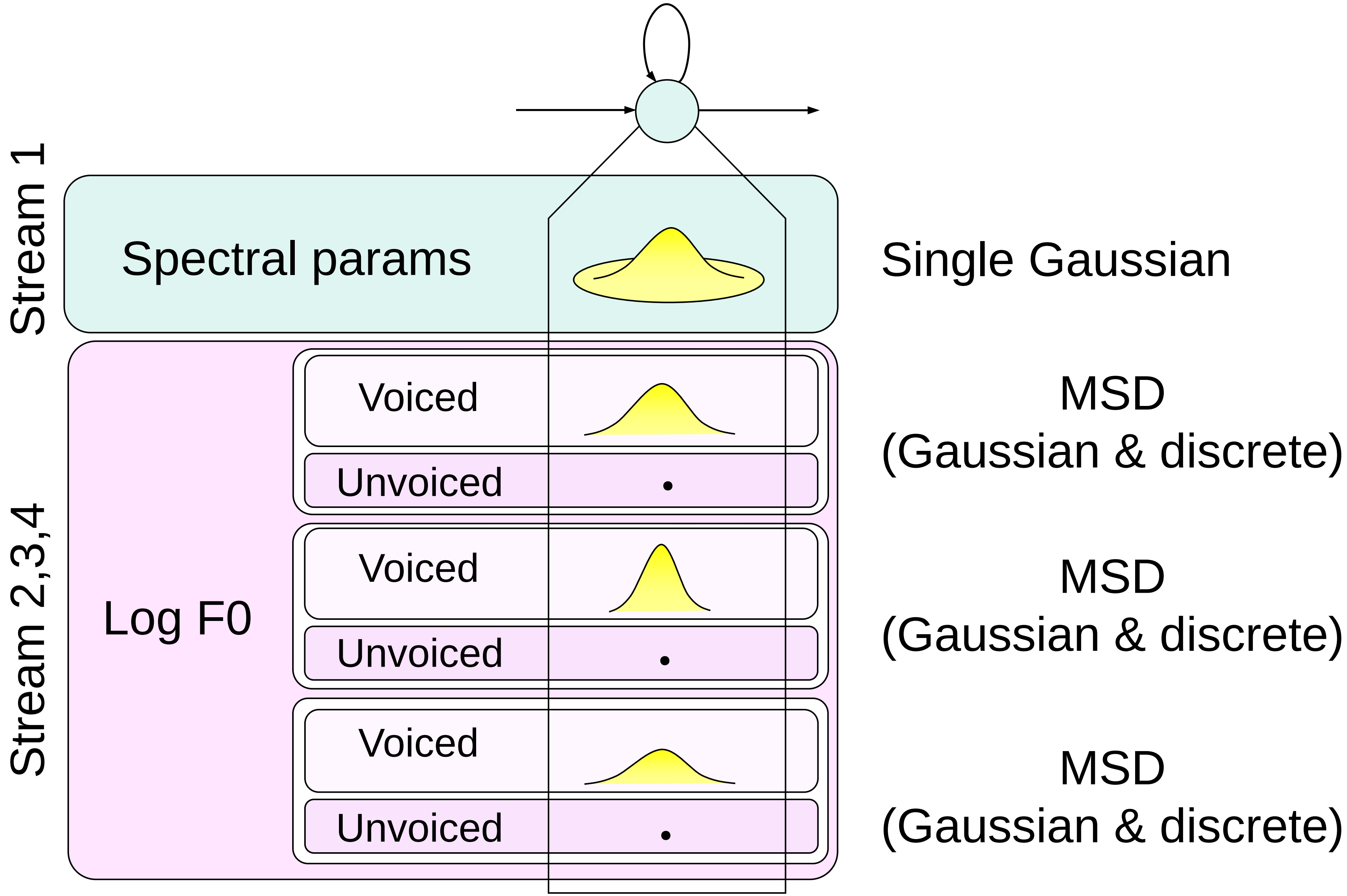


Unable to model by continuous or discrete distribution

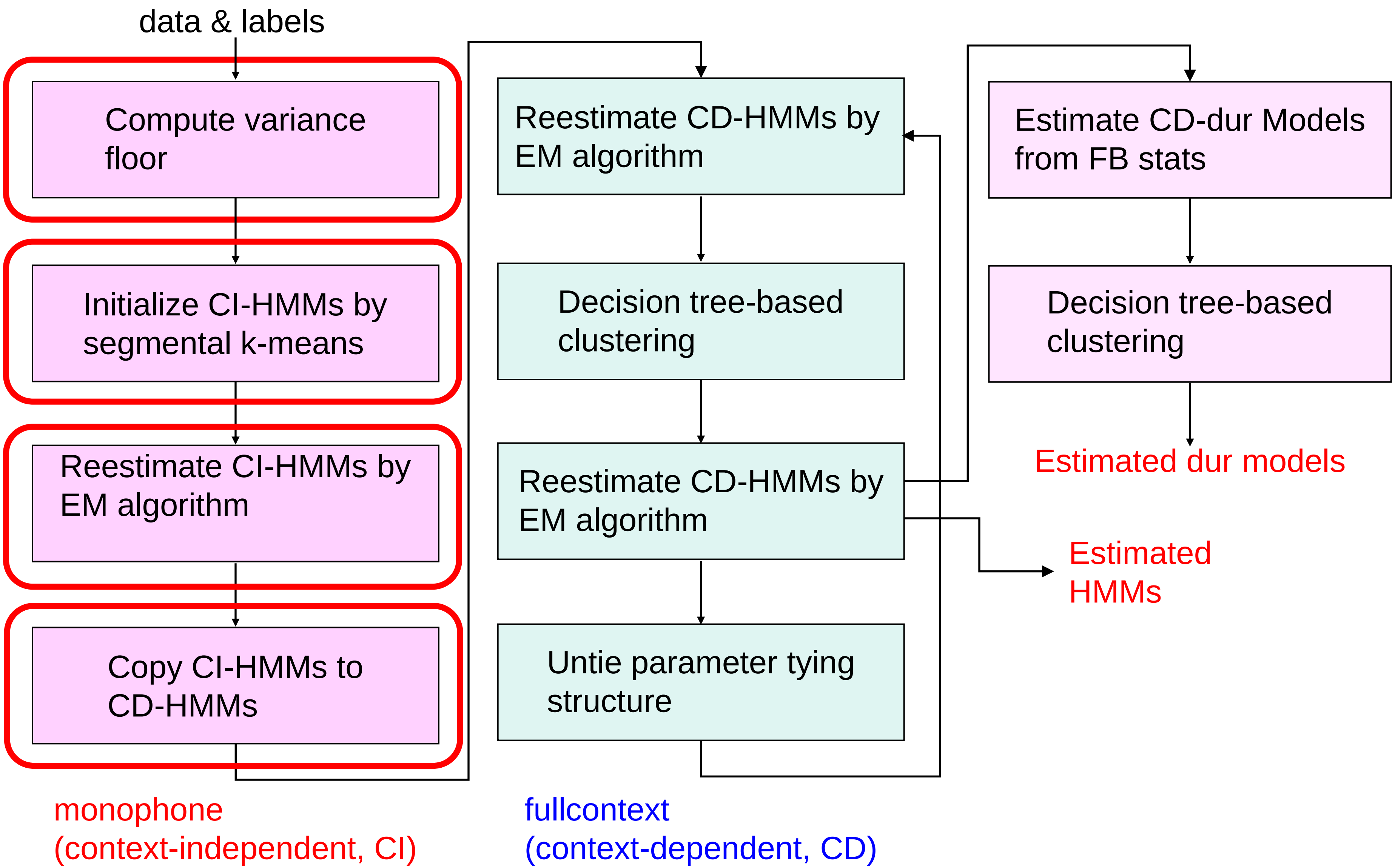
Multi-space probability distribution (MSD)



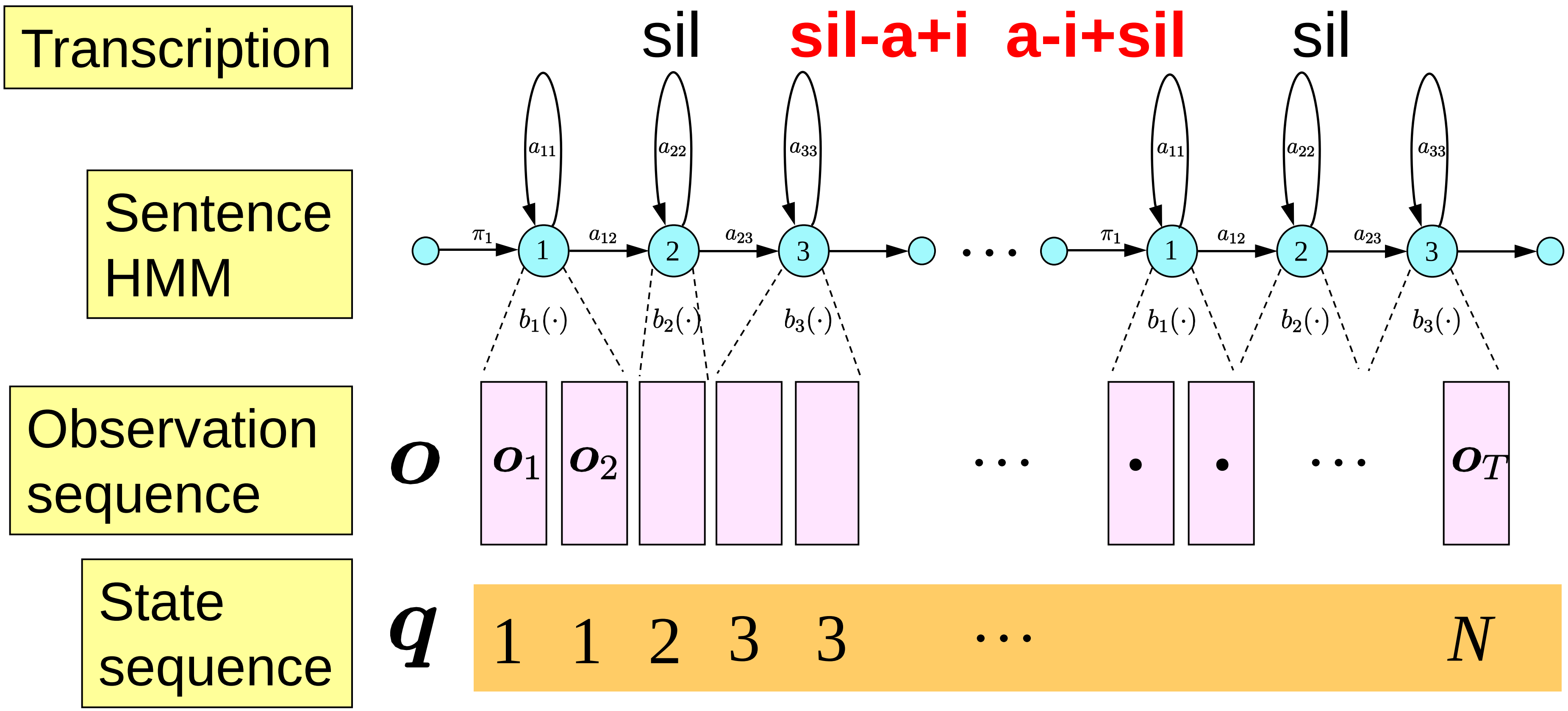
Structure of state-output distributions



Training process



HMM-based modeling



Context-dependent modeling

Phoneme

- **current phoneme**
- **{preceding, succeeding} two phonemes**

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

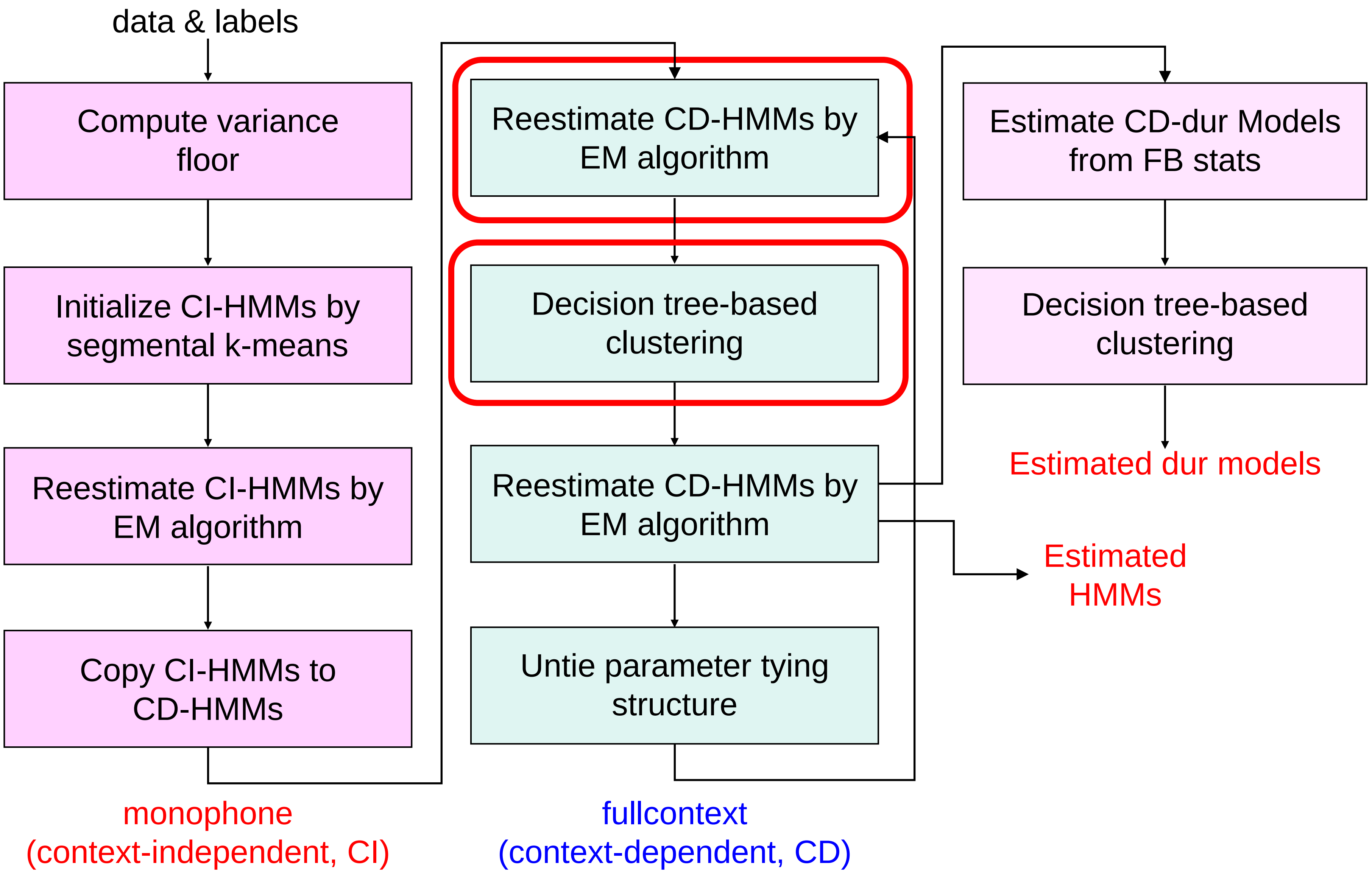
Phrase

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

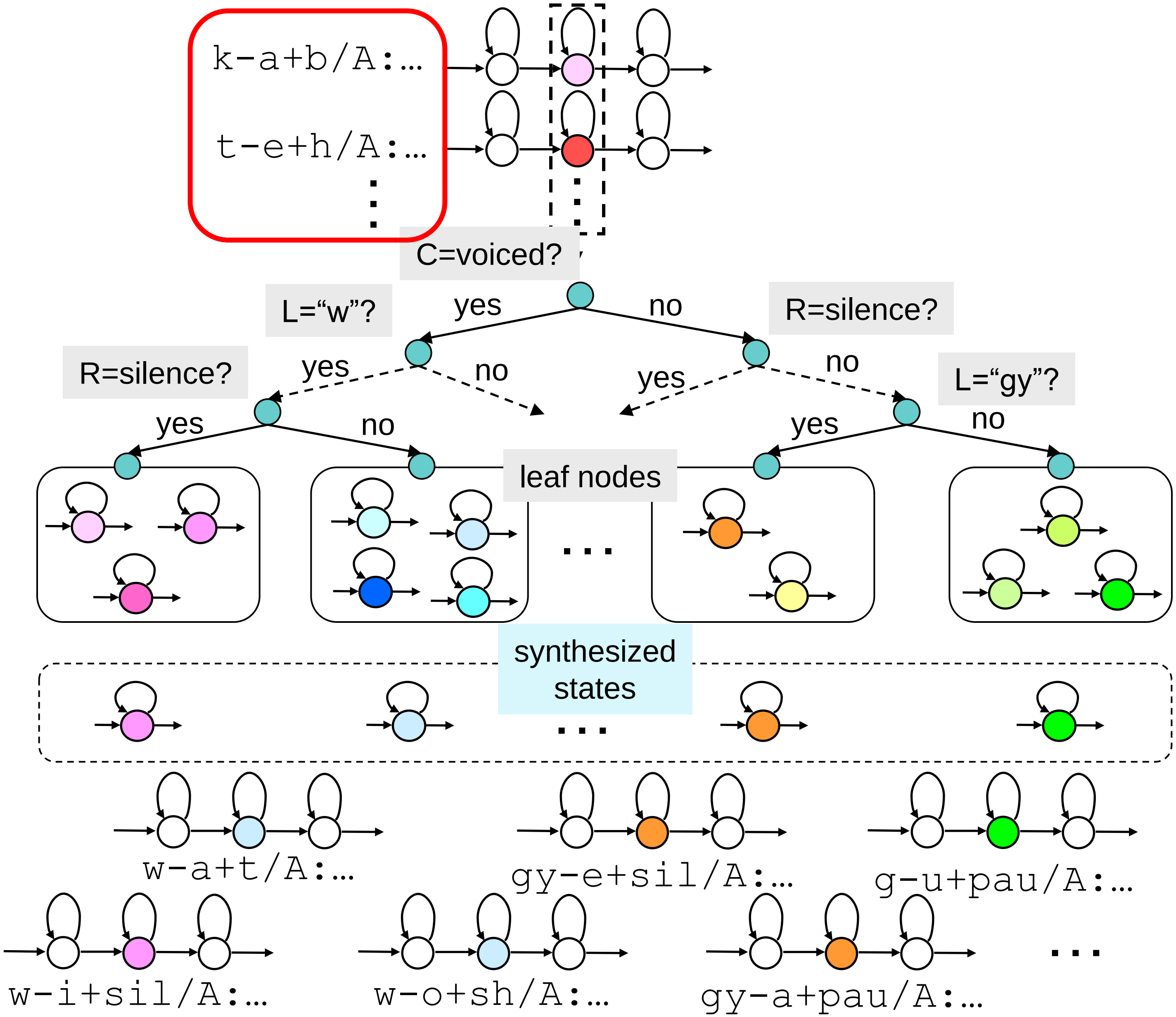
.....

Huge # of combinations ⇒ Difficult to have all possible models

Training process



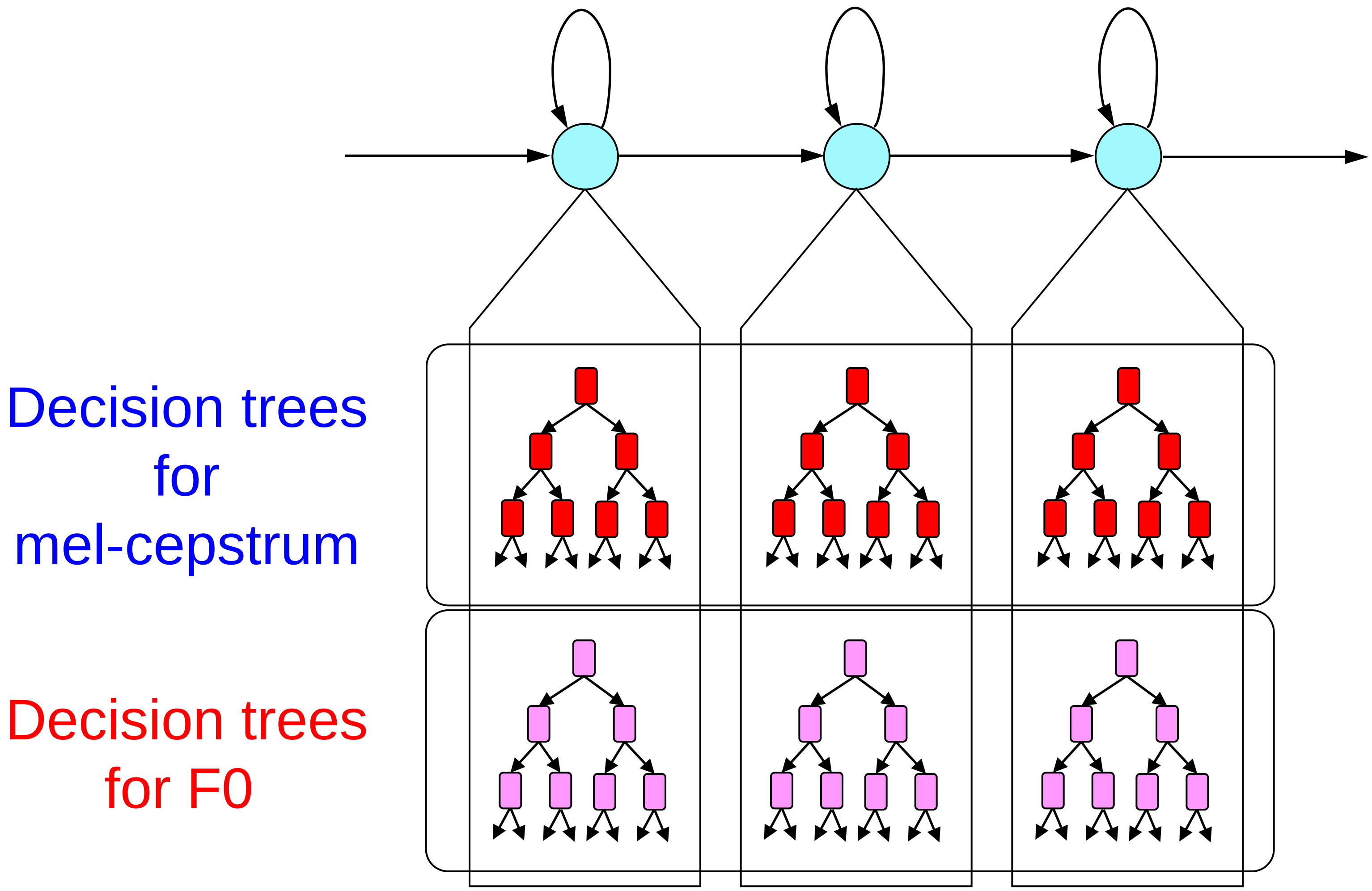
Decision tree-based context clustering [Odell;'95]



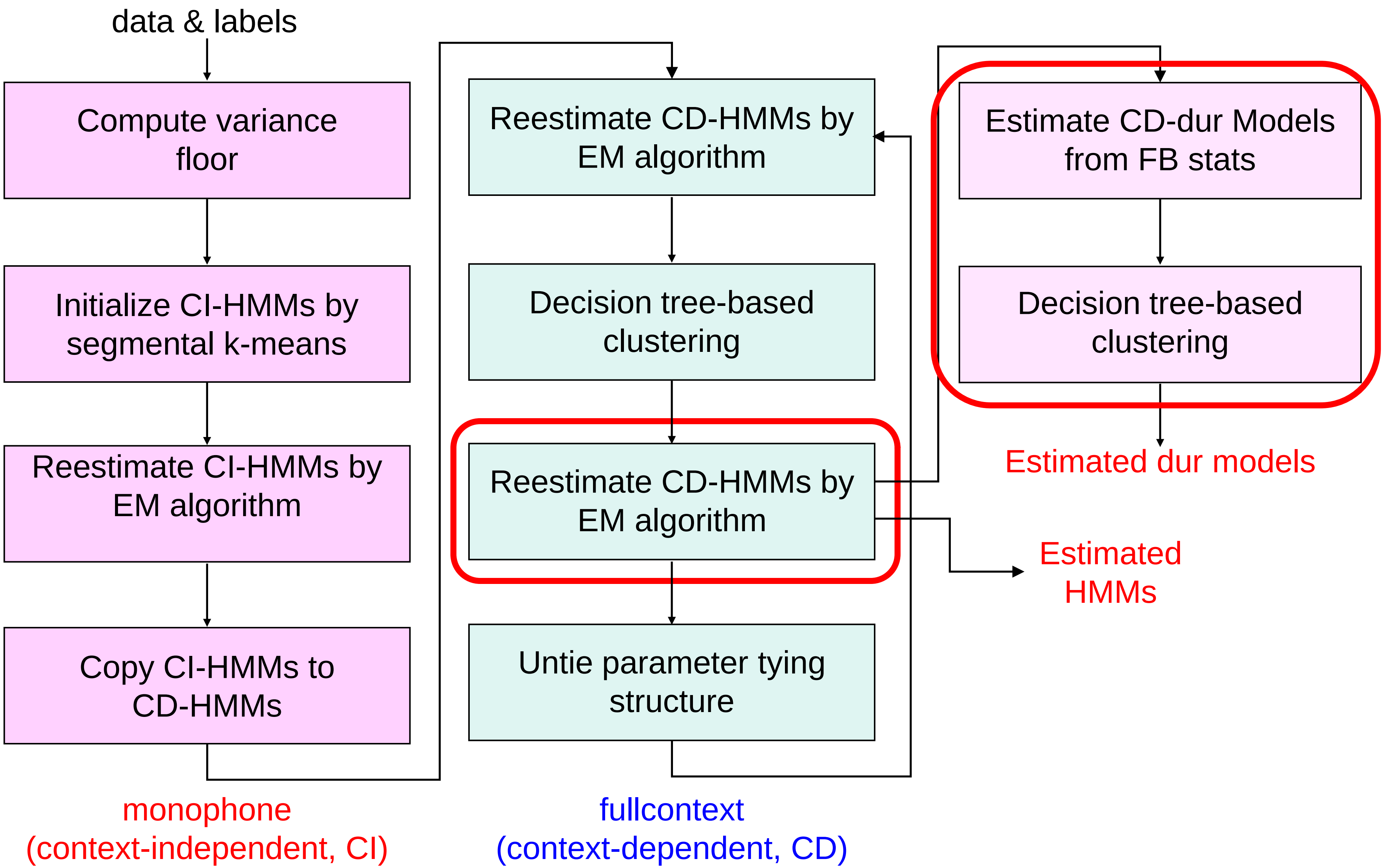
Stream-dependent clustering

Spectrum & excitation have different context dependency

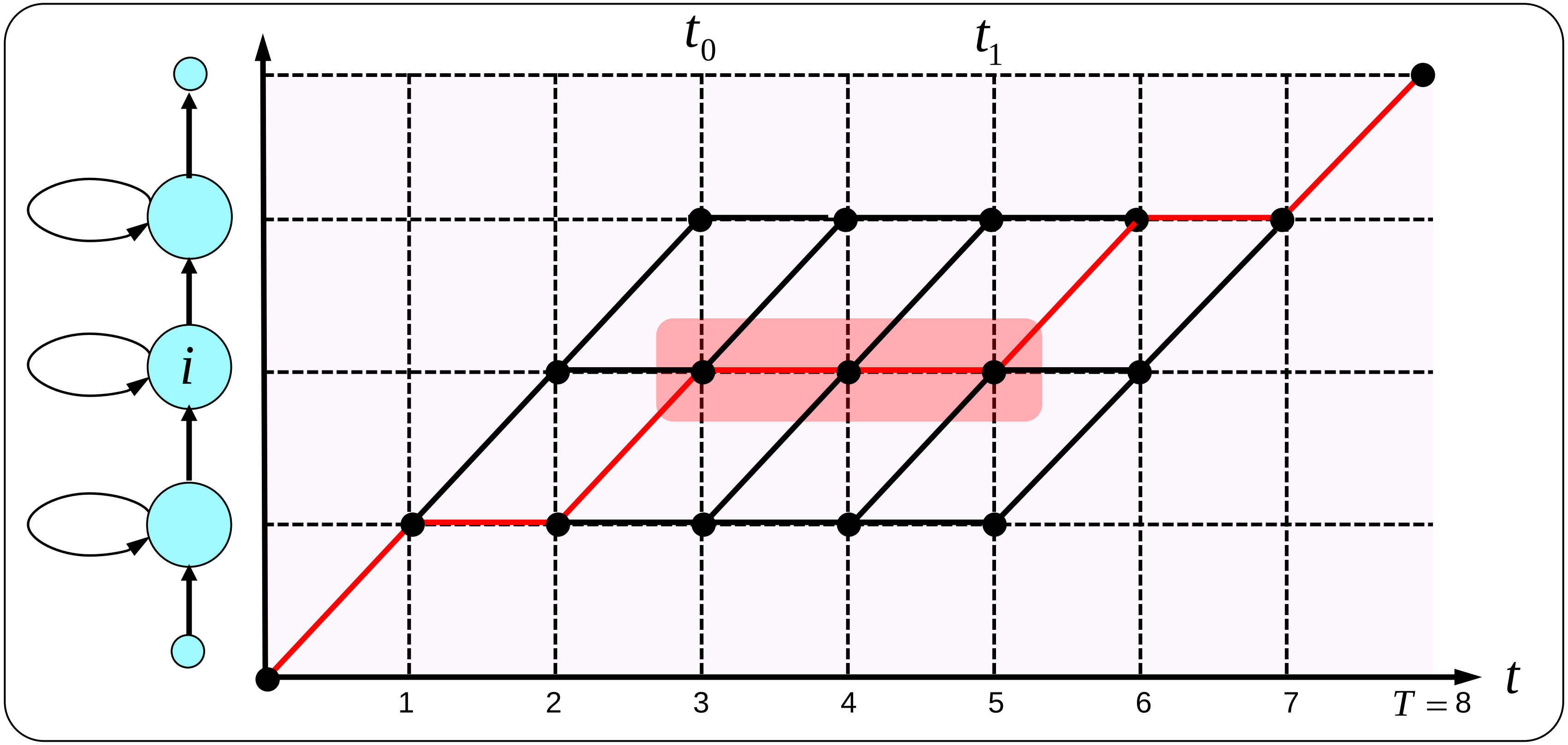
→ Build decision trees separately



Training process

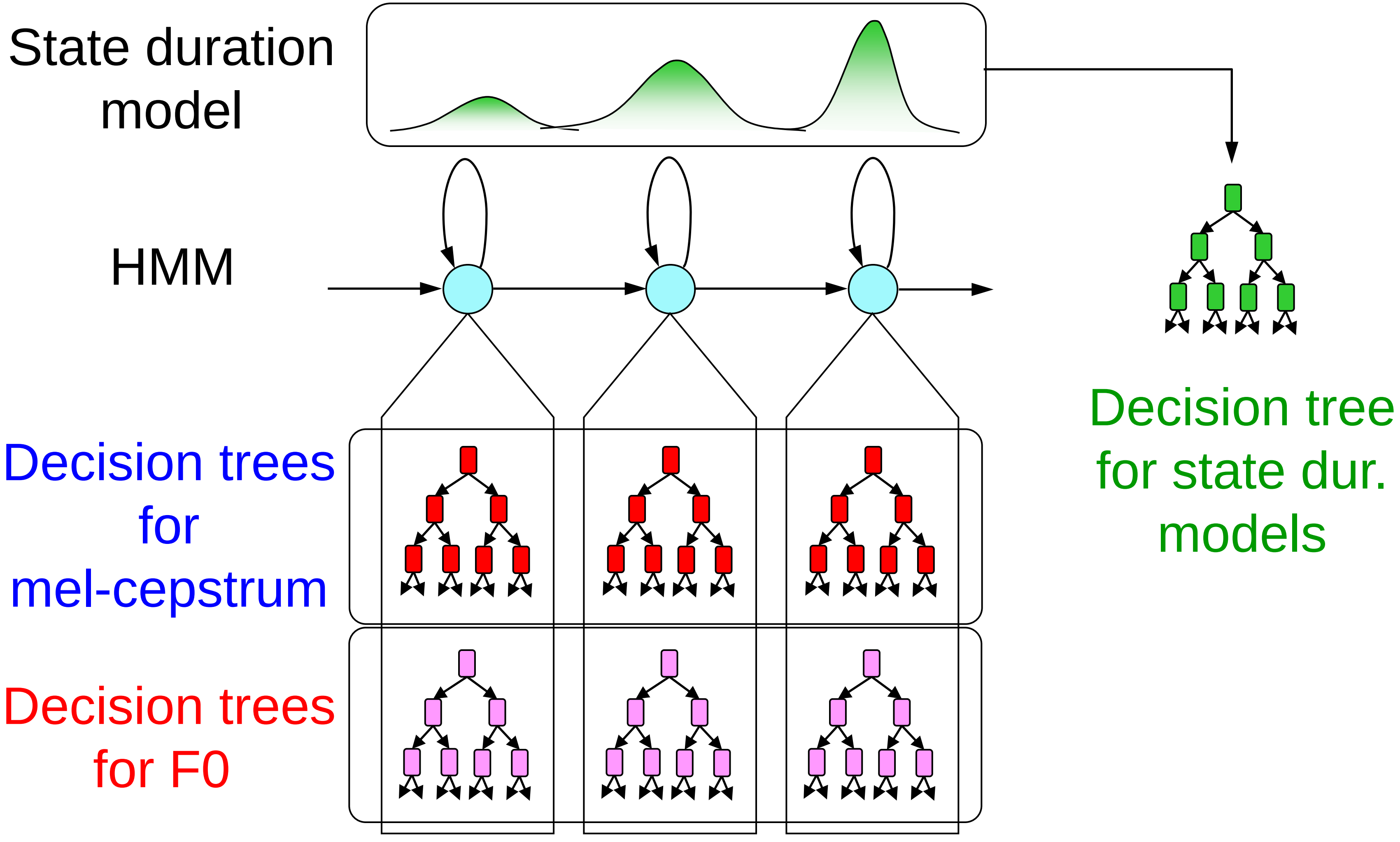


Estimation of state duration models [Yoshimura;'98]

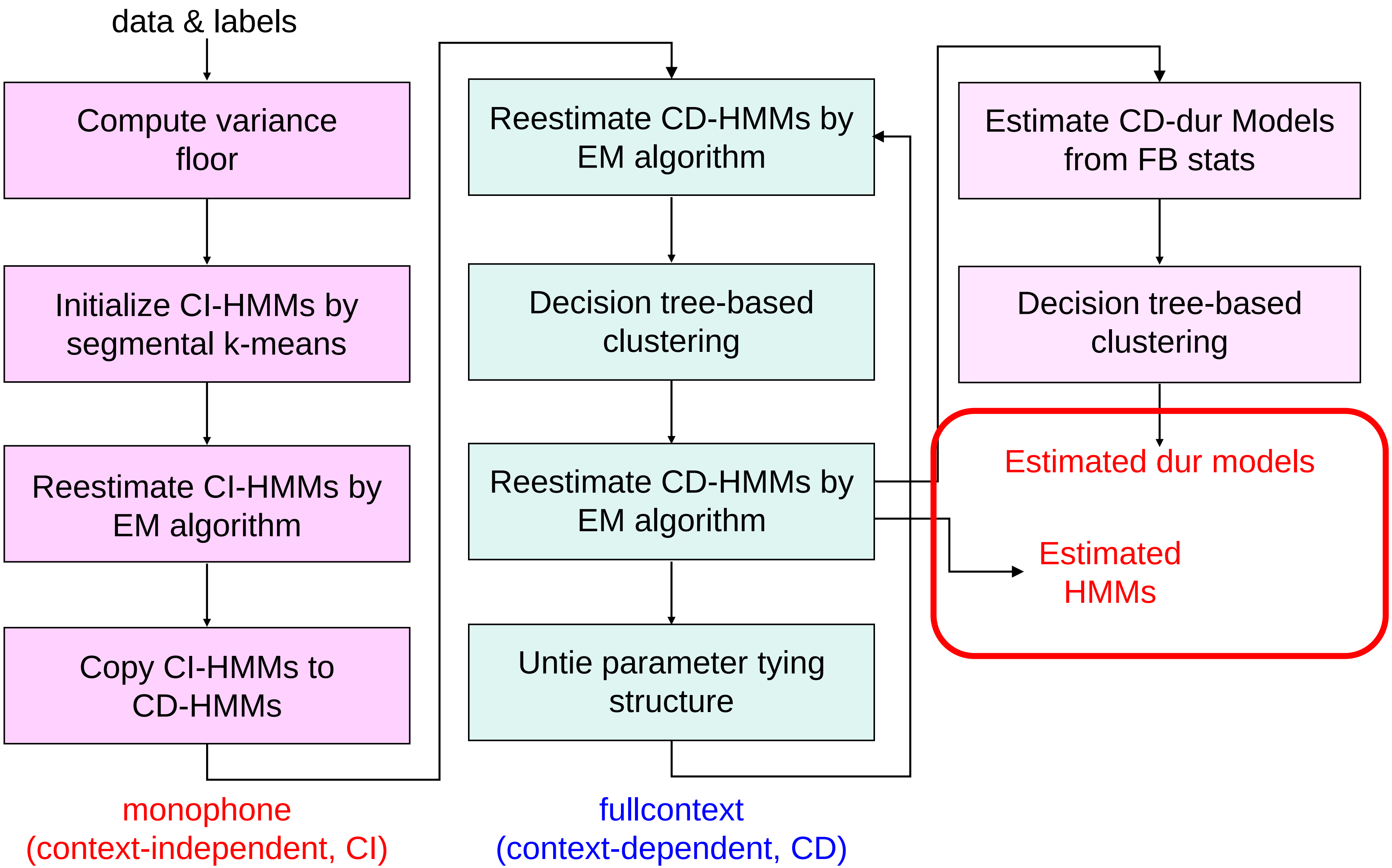


$$\chi_{t_0,t_1}(i) \propto \sum_{j \neq i} \alpha_{t_0-1}(j) a_{ij} a_{ii}^{t_1-t_0} \prod_{t=t_0}^{t_1} b_i(\boldsymbol{o}_t)$$
$$\cdot \sum_{k \neq i} a_{ik} b_k(\boldsymbol{o}_{t_1+1}) \beta_{t_1+1}(k)$$

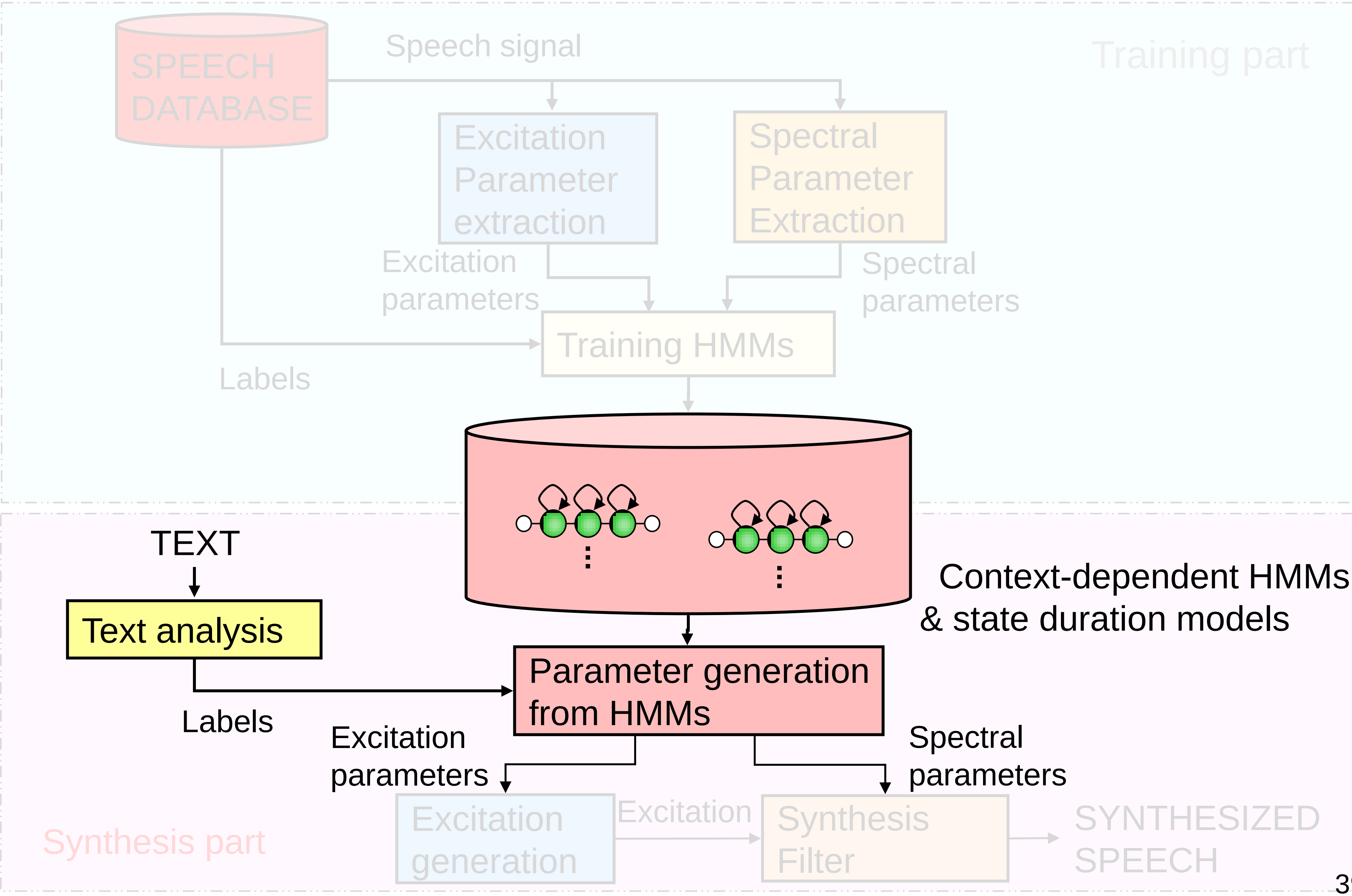
Stream-dependent clustering



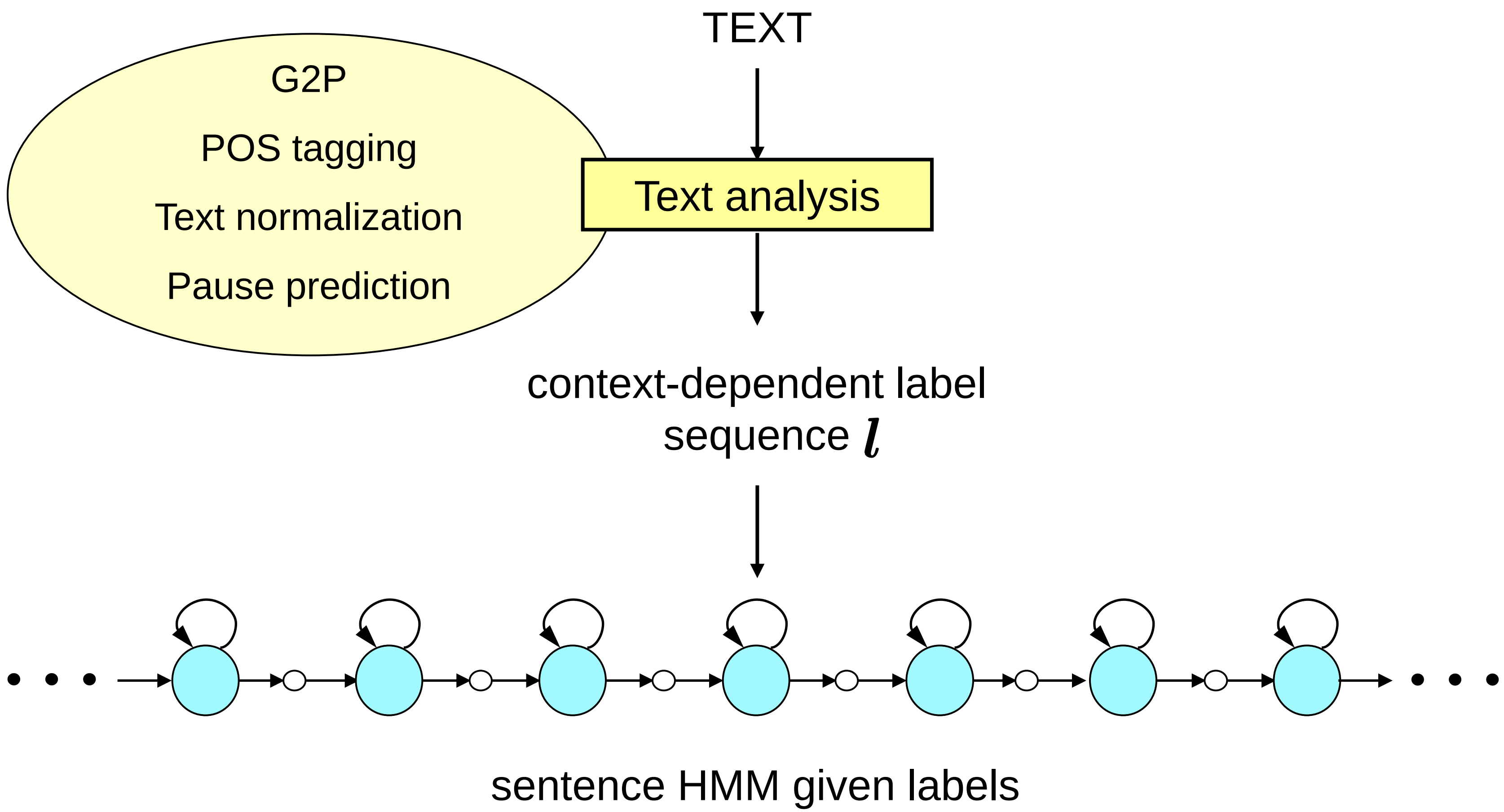
Training process



HMM-based speech synthesis system (HTS)



Composition of sentence HMM for given text



Speech parameter generation algorithm

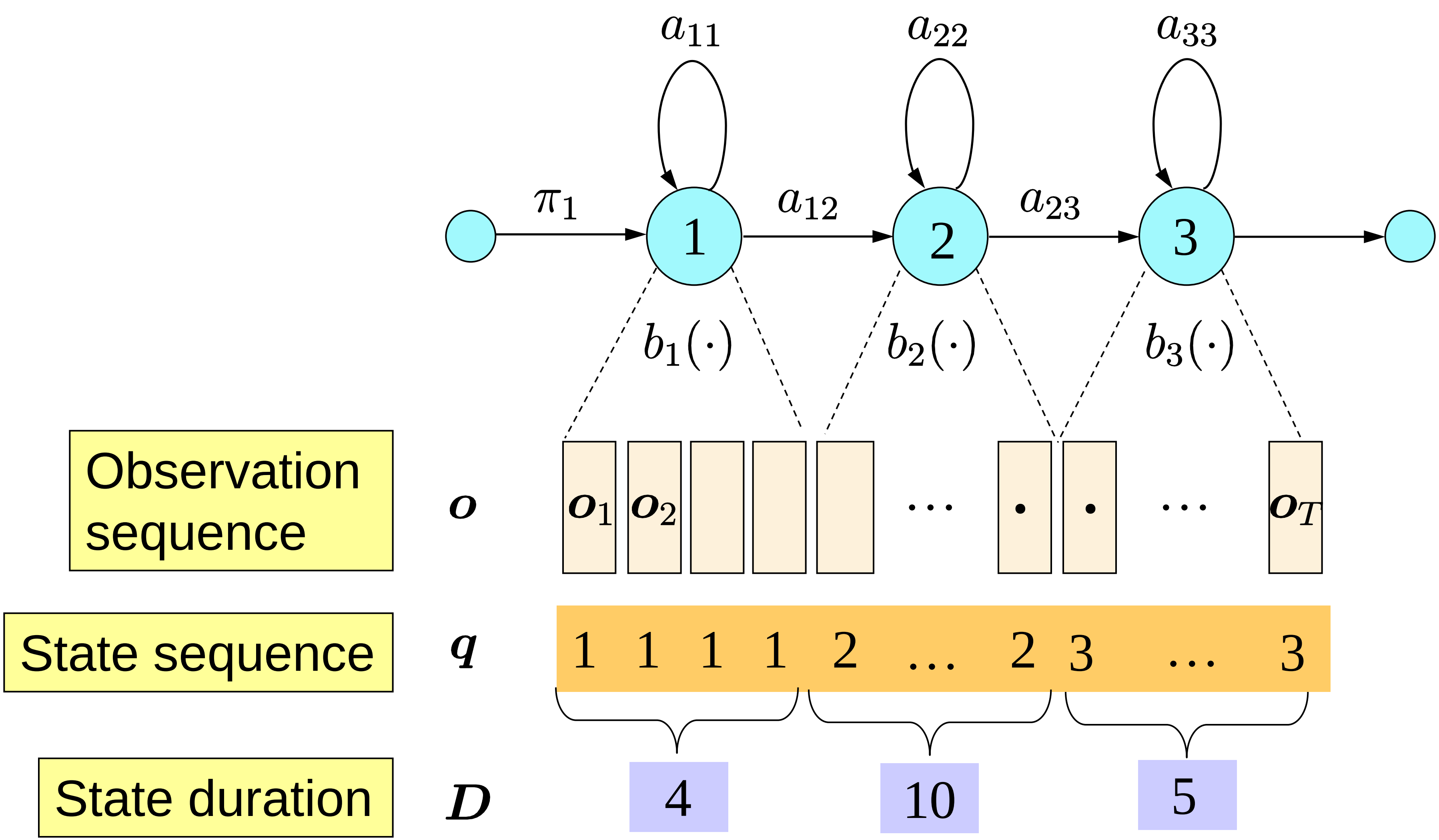
$$\begin{aligned}\hat{\boldsymbol{o}} &= \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \boldsymbol{l}, \hat{\lambda}) \\ &= \arg\max_{\boldsymbol{o}} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{l}, \hat{\lambda}) \\ &\approx \arg\max_{\boldsymbol{o}, \boldsymbol{q}}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

$$\Downarrow$$

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

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

Determination of state sequence (1)



Determine state sequence via determining state durations

Determination of state sequence (2)

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

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

d_i : state duration of i-th state

K : number of states in a sentence HMM for w

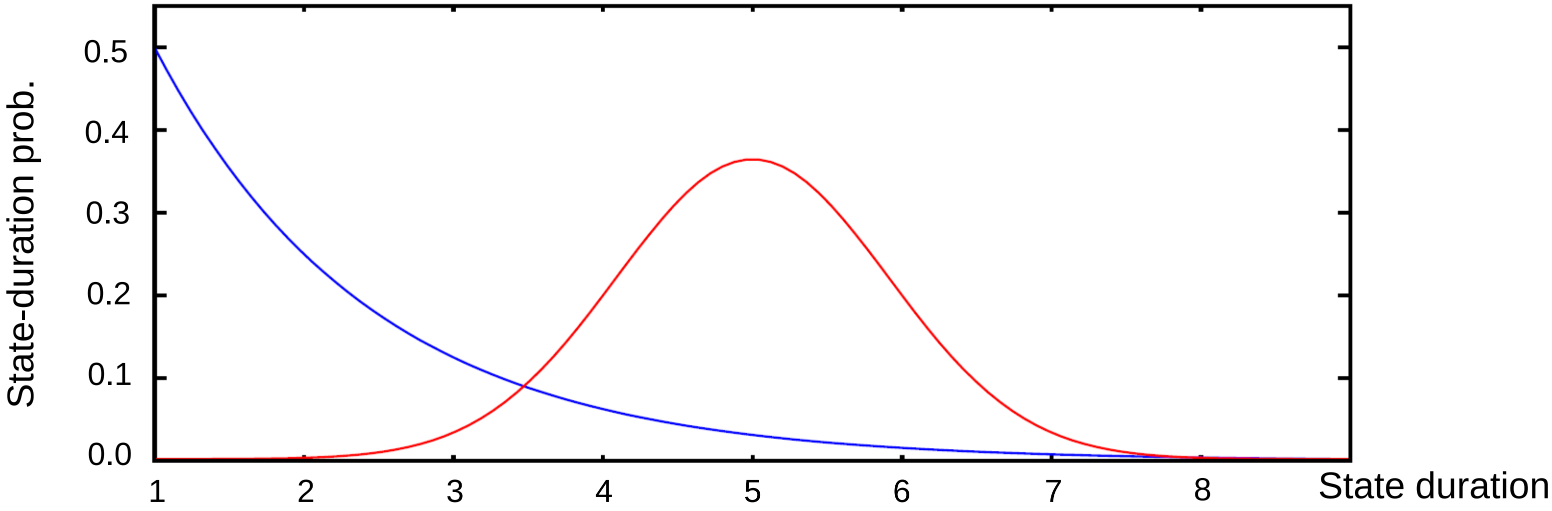
Determination of state sequence (3)

Geometric

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

Gaussian

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



Speech parameter generation algorithm

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

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

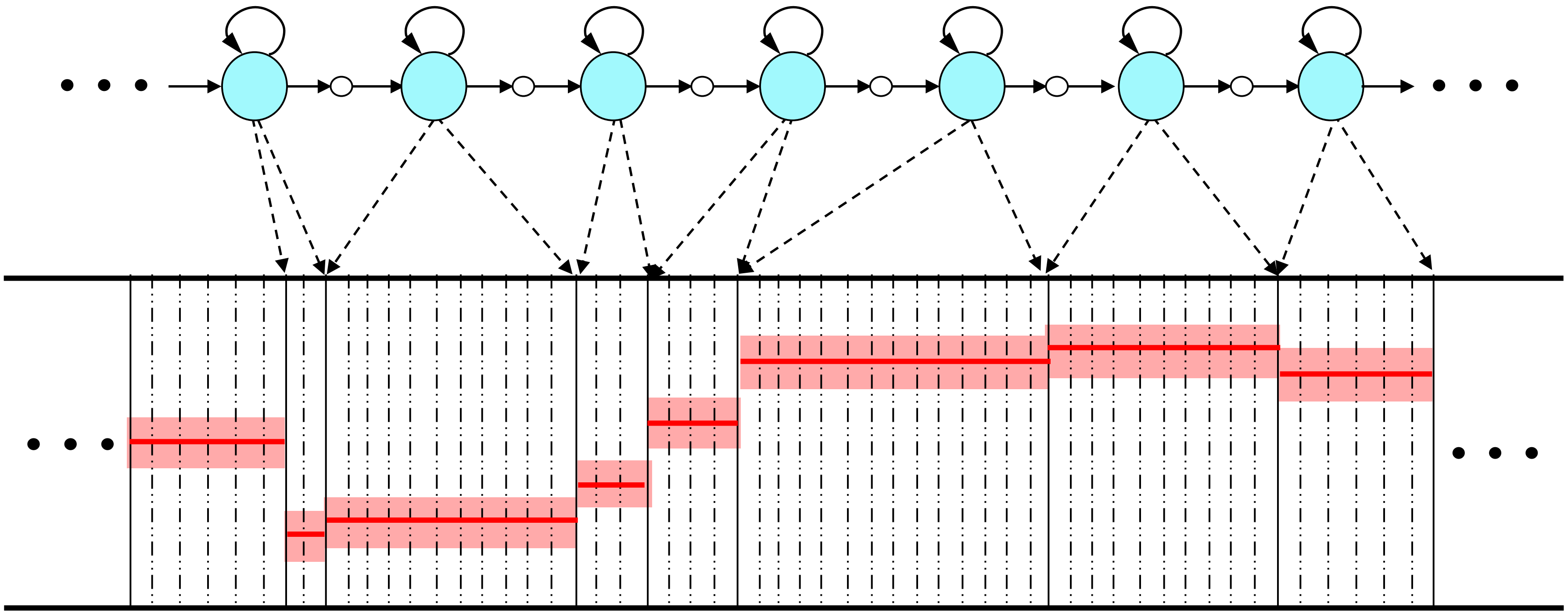
$$\approx \arg\max_{\boldsymbol{o},\boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{l}, \hat{\lambda})$$

↓

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

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

Without dynamic features



Mean

Variance

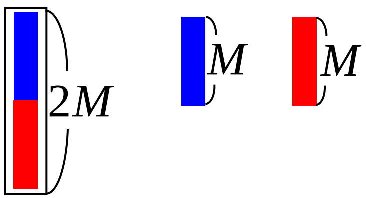
$\downarrow$

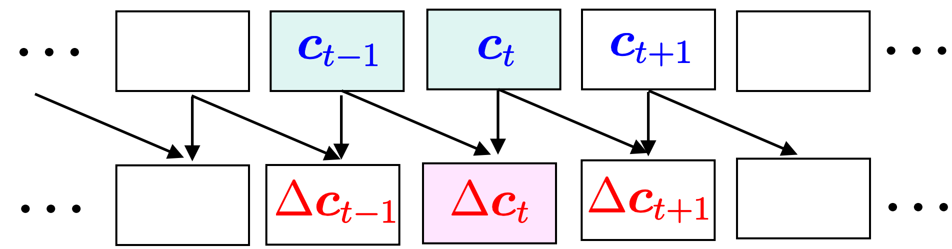
$\hat{o} \rightarrow$ step-wise, mean values

Integration of dynamic features

Speech param. vectors 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

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

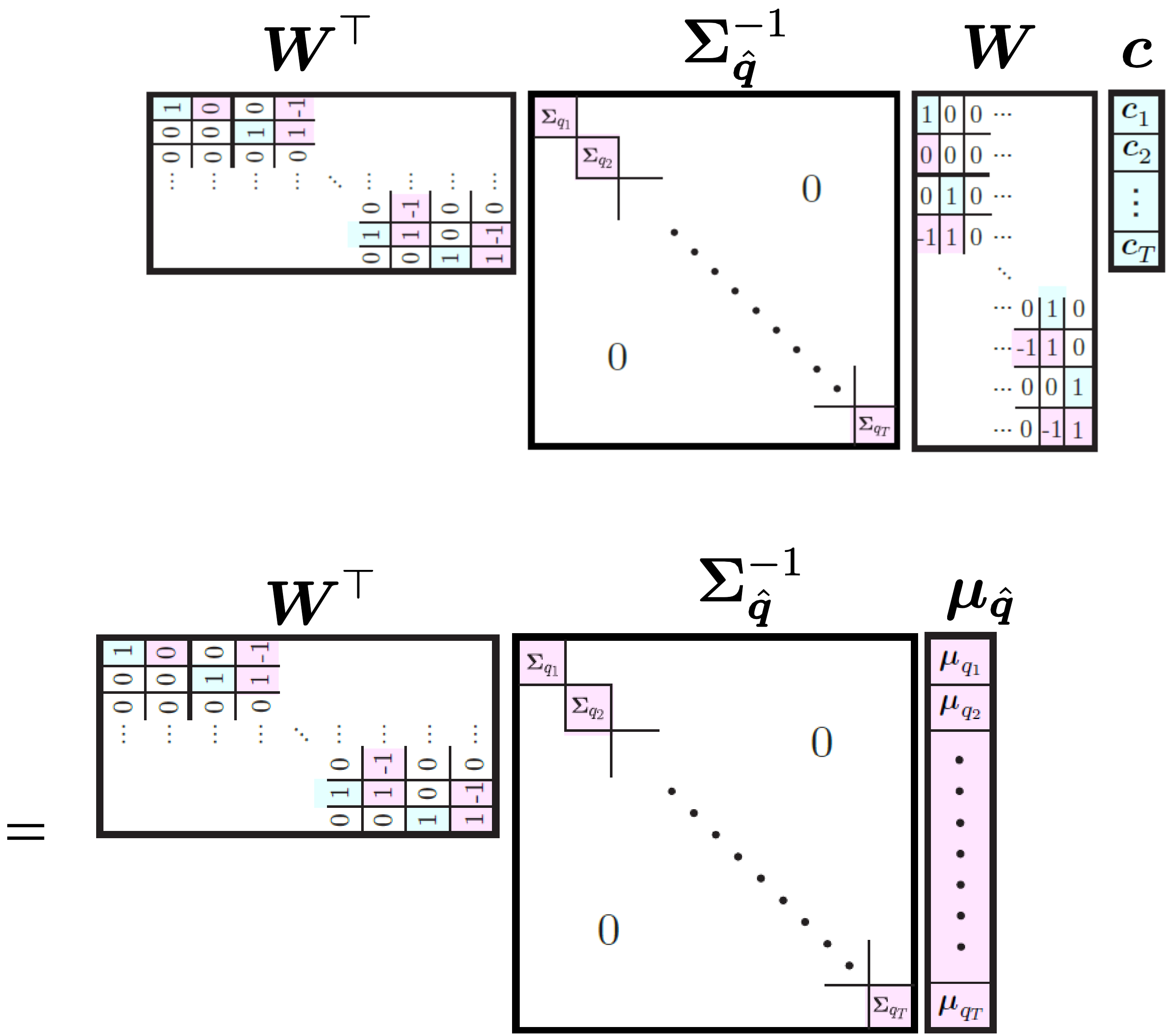
Speech parameter generation algorithm

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \hat{\lambda})\Big|_{\boldsymbol{o}=\boldsymbol{W}\boldsymbol{c}}$$

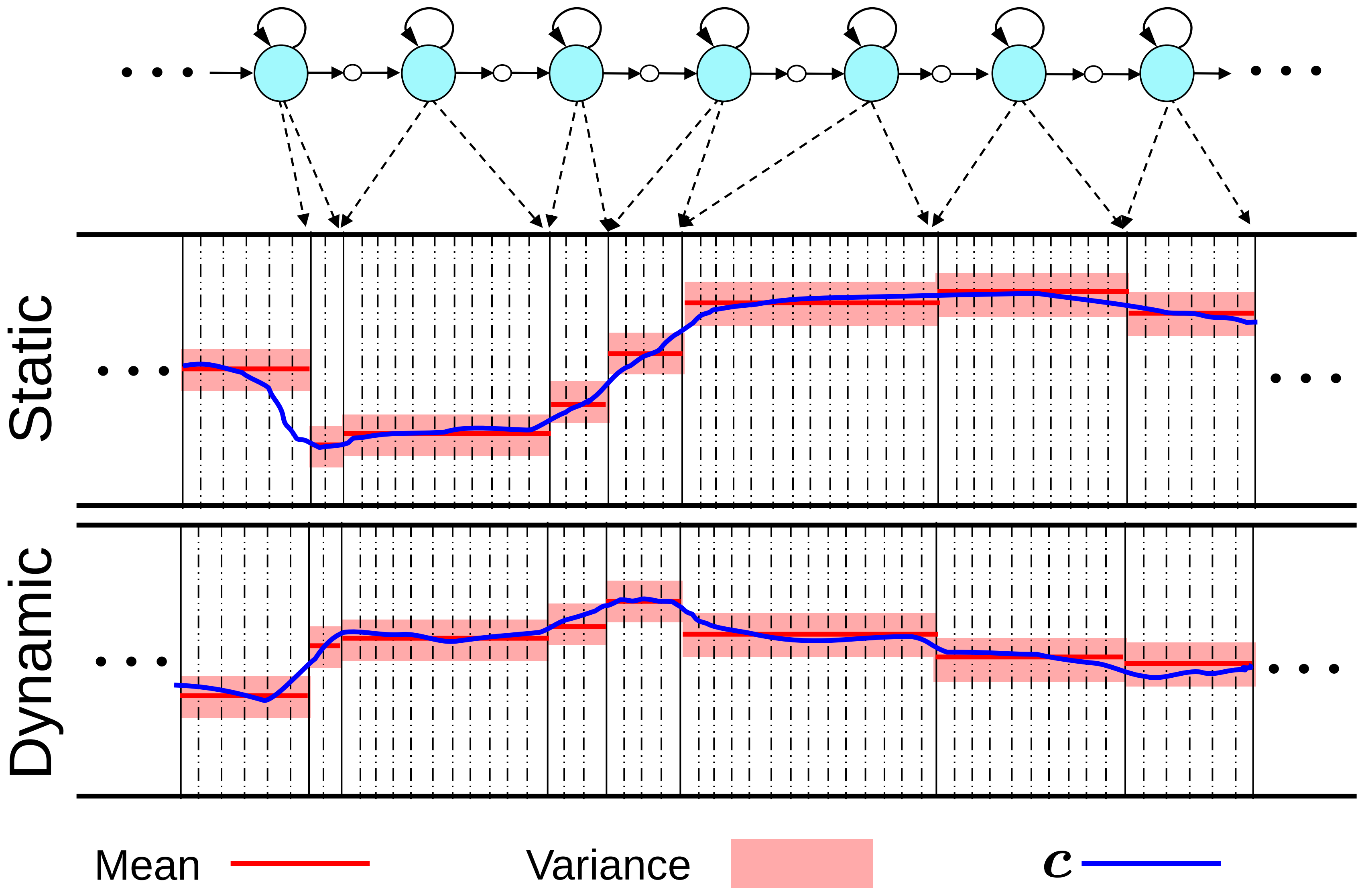
$$\downarrow$$

$$\begin{aligned}\hat{\boldsymbol{c}} &= \arg\max_{\boldsymbol{c}} p(\boldsymbol{W}\boldsymbol{c} \mid \hat{\boldsymbol{q}}, \hat{\lambda}) \\ &= \arg\max_{\boldsymbol{c}} \mathcal{N}(\boldsymbol{W}\boldsymbol{c}\,;\; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}})\end{aligned}$$

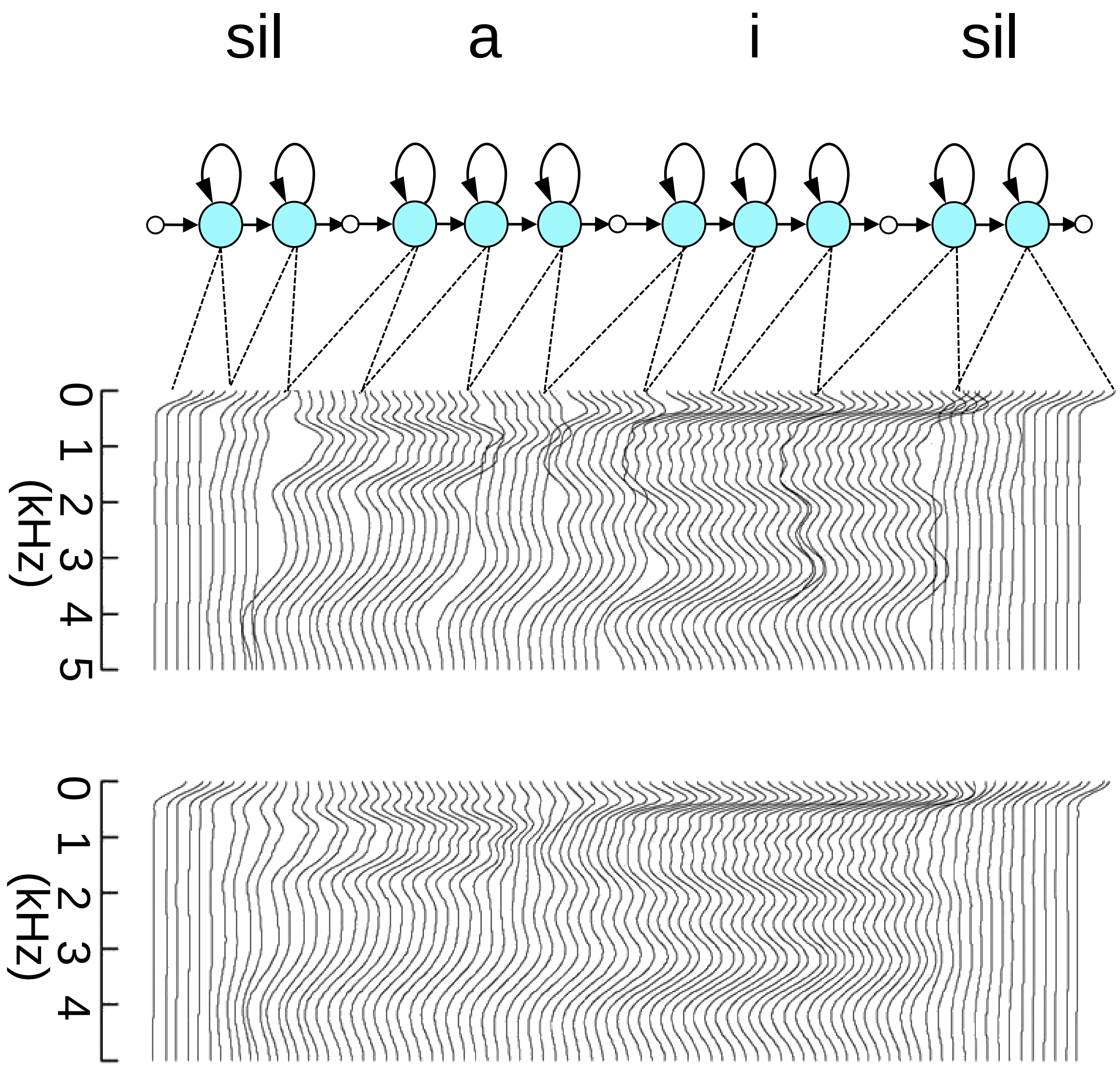
Solution



Generated speech parameter trajectory



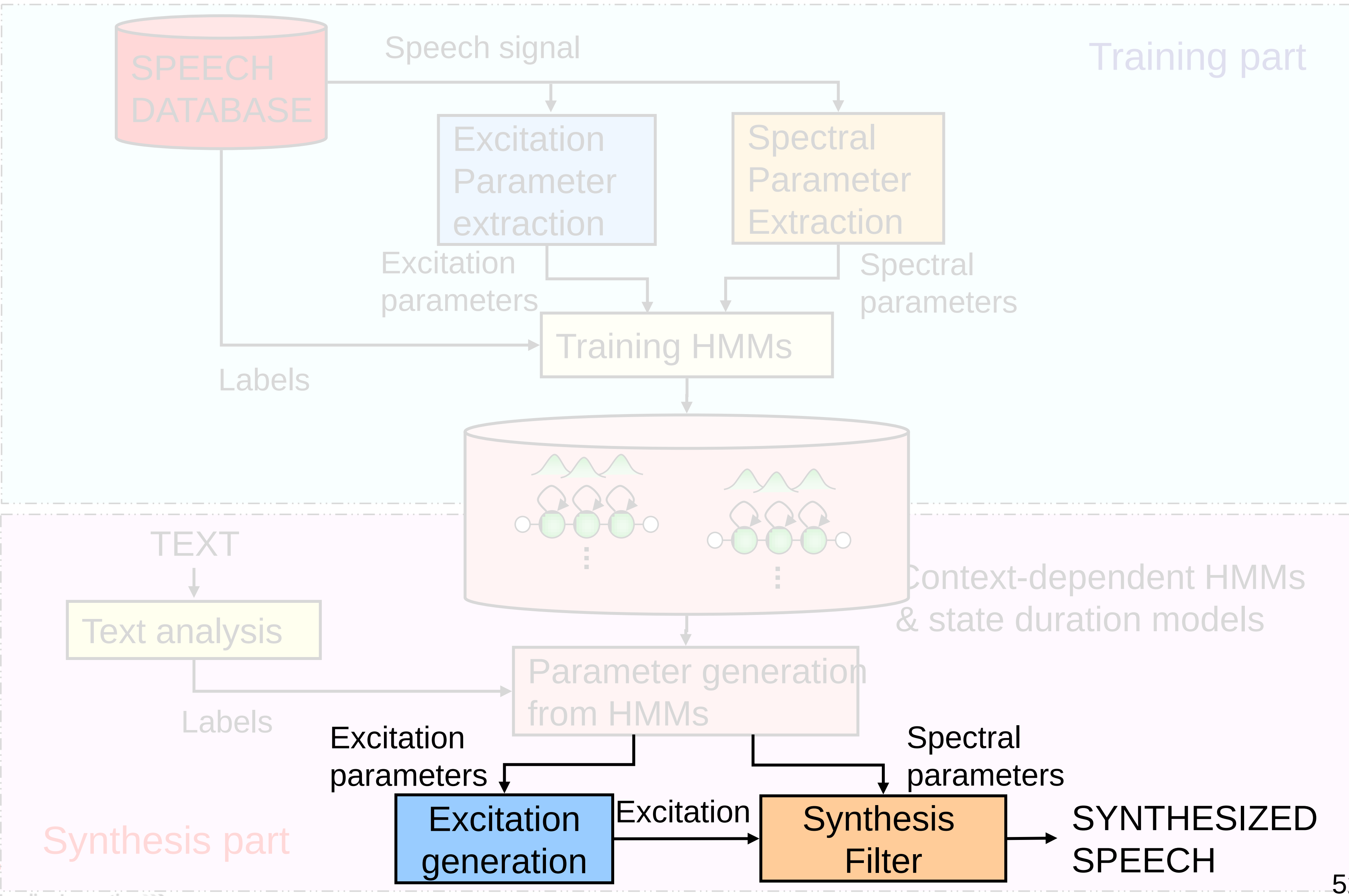
Generated spectra



w/o dynamic features

w/ dynamic features

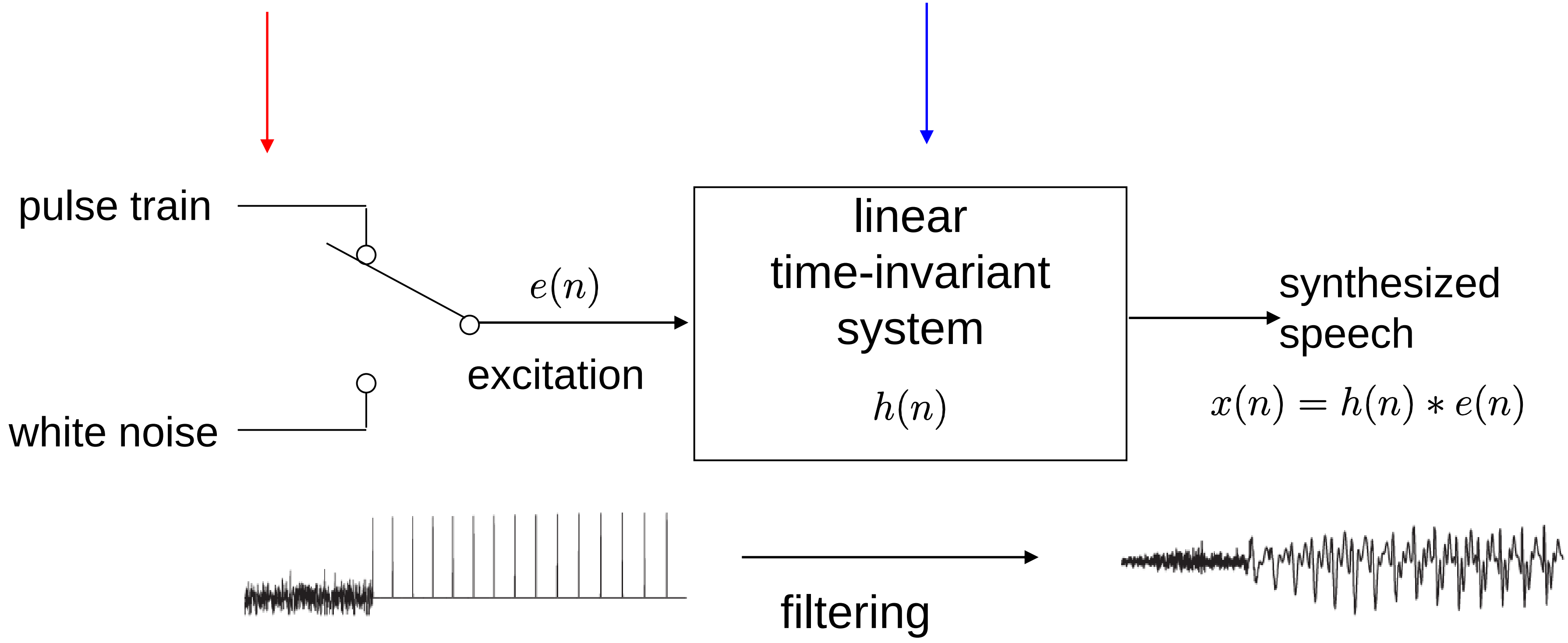
HMM-based speech synthesis system (HTS)



Source-filter model

Generated
excitation parameter
(log F0 with V/UV)

Generated
spectral parameter
(cepstrum, LSP)



Unvoiced frames & LP spectral coefficients

white noise

$$e_t(n) \longrightarrow \boxed{\dfrac{c_t(0)}{1 - \sum_{m=1}^{M} c_t(m) z^{-m}}} \longrightarrow x_t(n)$$

Synthesized speech

$$x_t(n) = \sum_{m=1}^{M} c_t(m) x_t(n-m) + e_t(n), \qquad e_t(n) \sim \mathcal{N}(0, c_t(0))$$

Drive linear filter using white noise

→ Equivalent to sampling from Gaussian distribution

$$\tilde{\boldsymbol{x}}_t \sim \mathcal{N}\left(\mathbf{0}, c_t(0) \left(\boldsymbol{\Psi}^\top \boldsymbol{\Psi}\right)^{-1}\right)$$

Speech samples

w/o dynamic features

w/ dynamic features

Use of dynamic features can reduce discontinuity

Outline

HMM-based speech synthesis

- Overview
- Implementation of individual components

Bayesian framework for speech synthesis

- Formulation
- Realizations in HMM-based speech synthesis
- Recent works

Conclusions

- Summary
- Future research topics

Statistical framework for speech synthesis (1)

We have a speech database, i.e., a set of texts & corresponding speech waveforms.

Given a text to be synthesized, what is the speech waveform corresponding to the text?

$\boldsymbol{W}$: set of texts

$\boldsymbol{X}$: speech waveforms

database

$\boldsymbol{w}$: text to be synthesized

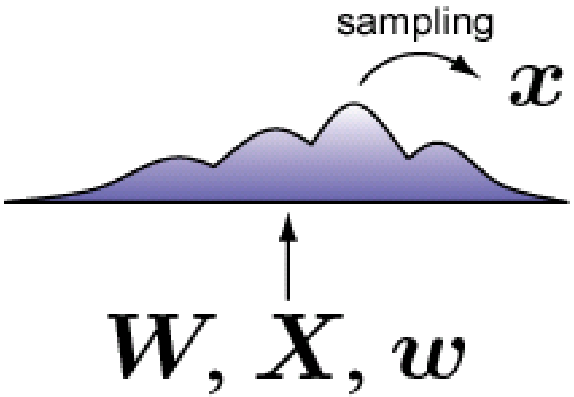
Given

$\boldsymbol{x}$: speech waveform $\longleftarrow$ unknown

Bayesian framework for speech synthesis (2)

Bayesian framework for prediction

Draw $\tilde{\boldsymbol{x}}$ from $p(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W})$



$\boldsymbol{W}$: set of texts

$\boldsymbol{X}$: speech waveforms

database

Given

$\boldsymbol{w}$: text to be synthesized

$\boldsymbol{x}$: speech waveform ← unknown

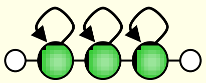
1. Estimate predictive distribution given variables
2. Draw sample from the distribution

Bayesian framework for speech synthesis (3)

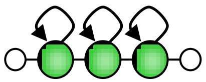
1. Estimating predictive distribution is hard ☹

→ Introduce acoustic model parameters

$$p(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W})$$

$\Downarrow$ introduce acoustic model λ 

$$= \int p(\boldsymbol{x}, \lambda \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{X}) d\lambda = \int p(\boldsymbol{x} \mid \boldsymbol{w}, \lambda) p(\lambda \mid \boldsymbol{W}, \boldsymbol{X}) d\lambda$$

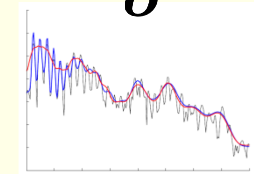
λ : acoustic model (e.g. HMM )

Bayesian framework for speech synthesis (4)

2. Using speech waveform directly is difficult ☹

→ Introduce parametric its representation

$$
\begin{aligned}
&p(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W}) \\
&= \int p(\boldsymbol{x} \mid \boldsymbol{w}, \lambda) p(\lambda \mid \boldsymbol{X}, \boldsymbol{W}) d\lambda
\end{aligned}
$$



⇓ introduce parametric representation of speech $\boldsymbol{o}$

$$
= \iint p(\boldsymbol{x} \mid \boldsymbol{o}) p(\boldsymbol{o} \mid \boldsymbol{w}, \lambda) p(\lambda \mid \boldsymbol{X}, \boldsymbol{W}) d\lambda d\boldsymbol{o}
$$

$\boldsymbol{o}$: parametric representation of speech waveform $\boldsymbol{x}$
(e.g., cepstrum, LPC, LSP, F0, aperiodicity)

Bayesian framework for speech synthesis (5)

3. Same texts can have multiple pronunciations, POS, etc. ☹

→ Introduce labels

$$p(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W})$$

$$= \iint p(\boldsymbol{x} \mid \boldsymbol{o}) p(\boldsymbol{o} \mid \boldsymbol{w}, \lambda) p(\lambda \mid \boldsymbol{X}, \boldsymbol{W}) d\lambda d\boldsymbol{o}$$

⇓ introduce labels derived from texts, $\boldsymbol{l}$ & $\boldsymbol{L}$

$$= \iint \sum_{\forall \boldsymbol{l}} p(\boldsymbol{x} \mid \boldsymbol{o}) p(\boldsymbol{o} \mid \boldsymbol{l}, \lambda) P(\boldsymbol{l} \mid \boldsymbol{w}) p(\lambda \mid \boldsymbol{X}, \boldsymbol{W}) d\lambda d\boldsymbol{o}$$

$\boldsymbol{l}$: labels derived from text $\boldsymbol{w}$

(e.g. prons, POS, lexical stress, grammar, pause)

Bayesian framework for speech synthesis (6)

4. Difficult to perform integral & sum over auxiliary variables ☹
→ Approximated by joint max

$$
\begin{aligned}
& p(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W}) \\
& = \iint \sum_{\forall \boldsymbol{l}} p(\boldsymbol{x} \mid \boldsymbol{o}) p(\boldsymbol{o} \mid \boldsymbol{l}, \lambda) P(\boldsymbol{l} \mid \boldsymbol{w}) p(\lambda \mid \boldsymbol{X}, \boldsymbol{W}) d\lambda d\boldsymbol{o}
\end{aligned}
$$

$\Downarrow$ approximate integral & sum by joint max

$$
\approx p(\boldsymbol{x} \mid \hat{\boldsymbol{o}}) p(\hat{\boldsymbol{o}} \mid \hat{\boldsymbol{l}}, \hat{\lambda}) P(\hat{\boldsymbol{l}} \mid \boldsymbol{w}) p(\hat{\lambda} \mid \boldsymbol{X}, \boldsymbol{W})
$$

where

$$
\left\{ \hat{\boldsymbol{o}}, \hat{\boldsymbol{l}}, \hat{\lambda} \right\} = \arg \max_{\boldsymbol{o}, \boldsymbol{l}, \lambda} p(\boldsymbol{x} \mid \boldsymbol{o}) p(\boldsymbol{o} \mid \boldsymbol{l}, \lambda) P(\boldsymbol{l} \mid \boldsymbol{w}) p(\lambda \mid \boldsymbol{X}, \boldsymbol{W})
$$

Bayesian framework for speech synthesis (7)

5. Joint maximization is hard ☹

→ Approximated by step-by-step maximizations

$$\left\{\hat{\boldsymbol{o}}, \hat{\boldsymbol{l}}, \hat{\lambda}\right\} = \arg\max_{\boldsymbol{o},\boldsymbol{l},\lambda} p(\boldsymbol{x} \mid \boldsymbol{o}) p(\boldsymbol{o} \mid \boldsymbol{l}, \lambda) P(\boldsymbol{l} \mid \boldsymbol{w}) p(\lambda \mid \boldsymbol{X}, \boldsymbol{W})$$

⇓ approx joint max by step-by-step max

$$\hat{\lambda} = \arg\max_{\lambda} p(\lambda \mid \boldsymbol{X}, \boldsymbol{W})$$ ⇐ training

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} P(\boldsymbol{l} \mid \boldsymbol{w})$$ ⇐ text analysis

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$ ⇐ speech parameter generation

Bayesian framework for speech synthesis (8)

6. Training also requires parametric form of wav & labels ☹

→ Introduce them & approx by step-by-step maximizations

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

$$\Downarrow$$

$$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} P(\boldsymbol{L} \mid \boldsymbol{W})$$ ⇐ labeling

$$\hat{\boldsymbol{O}} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{X} \mid \boldsymbol{O})$$ ⇐ feature extraction

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\boldsymbol{O}} \mid \hat{\boldsymbol{L}}, \lambda) p(\lambda)$$ ⇐ acoustic model training

$\boldsymbol{O}$: parametric representation of speech waveforms $\boldsymbol{X}$

$\boldsymbol{L}$: labels derived from texts $\boldsymbol{W}$

Bayesian framework for speech synthesis (9)

Draw $\tilde{\boldsymbol{x}}$ from $p(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W})$

$$\hat{\boldsymbol{O}} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{X} \mid \boldsymbol{O})$$ ⇐ feature extraction

$$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} P(\boldsymbol{L} \mid \boldsymbol{W})$$ ⇐ labeling

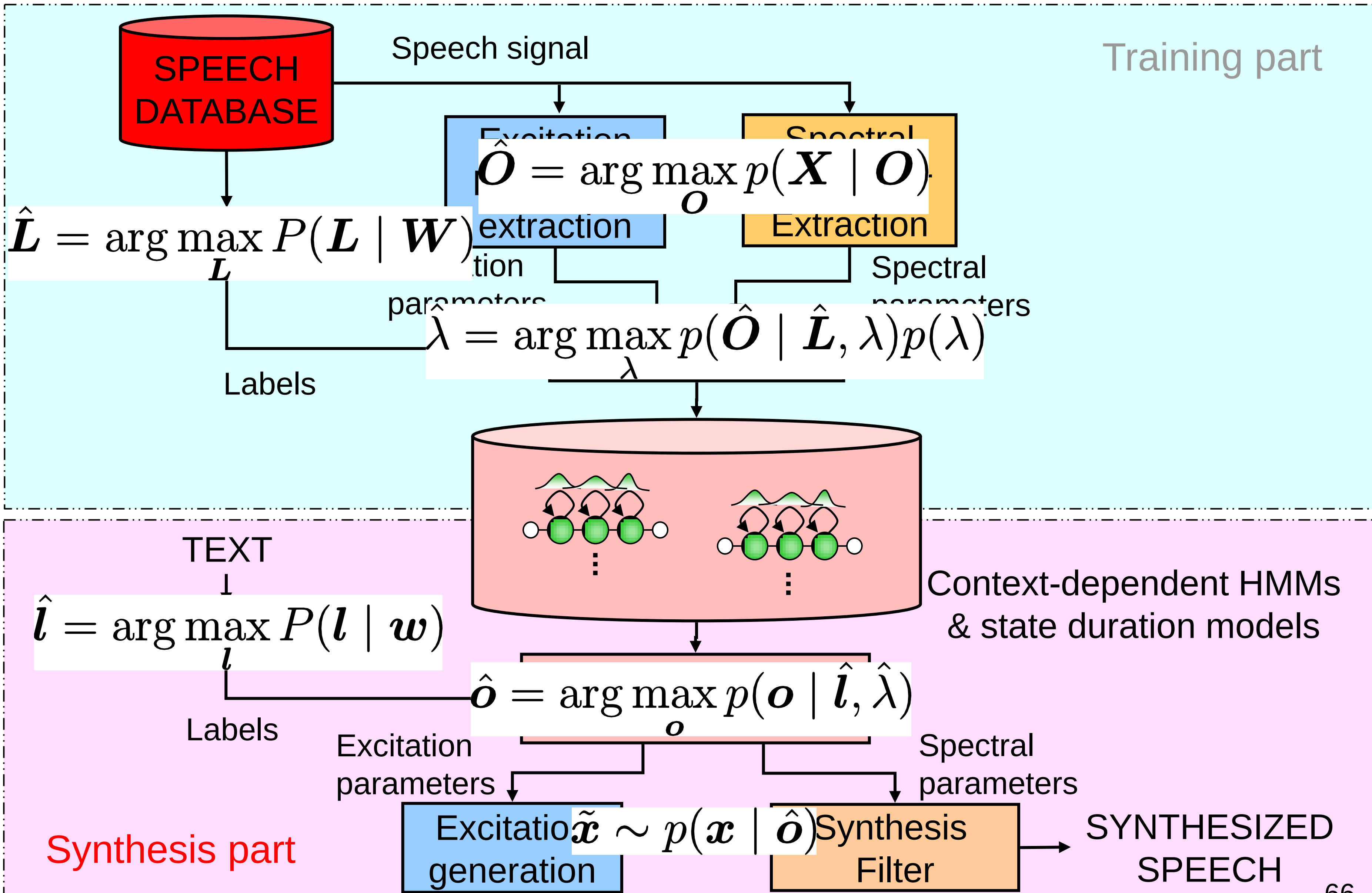
$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\boldsymbol{O}} \mid \hat{\boldsymbol{L}}, \lambda) p(\lambda)$$ ⇐ acoustic model training

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} P(\boldsymbol{l} \mid \boldsymbol{w})$$ ⇐ text analysis

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$ ⇐ speech parameter generation

$\tilde{\boldsymbol{x}}$ from $p(\boldsymbol{x} \mid \hat{\boldsymbol{o}})$ ⇐ waveform reconstruction

HMM-based speech synthesis system (HTS)



Problems

Many approximations

- Integral & sum ≈ max
- Joint max ≈ step-by-step max
  - → Poor approximation

Recent works to relax approximations

- Max → Integral & sum
  - ✓ Bayesian acoustic modeling
  - ✓ Multiple labels
- Step-wise max → Joint max
  - ✓ Statistical vocoding

Bayesian acoustic modeling (1)

ML-based approach (point estimate of λ)

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

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

Bayesian approach (posterior probability of λ)

$$\begin{aligned}
\hat{\boldsymbol{o}} &= \arg\max_{\boldsymbol{o}} \int p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\boldsymbol{O}}, \hat{\boldsymbol{L}}) d\lambda \\
&= \arg\max_{\boldsymbol{o}} \int p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \lambda) p(\lambda \mid \hat{\boldsymbol{O}}, \hat{\boldsymbol{L}}) d\lambda \\
&= \arg\max_{\boldsymbol{o}} \int p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \lambda) p(\hat{\boldsymbol{O}} \mid \hat{\boldsymbol{L}}, \lambda) p(\lambda) d\lambda
\end{aligned}$$

Bayesian acoustic modeling (2)

Bayesian approach

– Parameters are hidden variables & marginalized out

– Bayesian approach with hidden variables → intractable

→ Variational Bayes [Attias;'99]

$$\begin{aligned}
&\log P(\boldsymbol{o}, \hat{\boldsymbol{O}} \mid \hat{\boldsymbol{l}}, \hat{\boldsymbol{L}}) \\
&\quad = \log \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int Q(\boldsymbol{q}, \boldsymbol{Q}, \lambda) \frac{P(\boldsymbol{o}, \boldsymbol{q}, \hat{\boldsymbol{O}}, \hat{\boldsymbol{Q}}, \lambda \mid \hat{\boldsymbol{l}}, \hat{\boldsymbol{L}})}{Q(\boldsymbol{q}, \boldsymbol{Q}, \lambda)} d\lambda \\
&\quad \geq \left\langle \log \frac{P(\boldsymbol{o}, \boldsymbol{q}, \hat{\boldsymbol{O}}, \hat{\boldsymbol{Q}}, \lambda \mid \hat{\boldsymbol{l}}, \hat{\boldsymbol{L}})}{Q(\boldsymbol{q}, \boldsymbol{Q}, \lambda)} d\lambda \right\rangle_{Q(\boldsymbol{q}, \boldsymbol{Q}, \lambda)} \\
&\quad = \mathcal{F}
\end{aligned}$$

← **Jensen's inequality**

Bayesian acoustic modeling (3)

Variational Bayesian acoustic modeling for speech synthesis [Nankaku;'03]

- Fully VB-based speech synthesis
  - ✓ Training posterior distribution of model parameters
  - ✓ Parameter generation from predictive distribution
- Automatic model selection
  - ✓ Bayesian approach provides posterior probability of model structure
- Setting priors
  - ✓ Evidence maximization [Hashimoto;'06]
  - ✓ Cross validation [Hashimoto;'09]
- VB approach works better than ML one when
  - ✓ Data is small
  - ✓ Model is large

Multiple labels (1)

Conventional

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

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

Incorporate multiple possible labels

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

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

Label sequence is regarded as hidden variable & marginalized

Multiple labels (2)

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

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

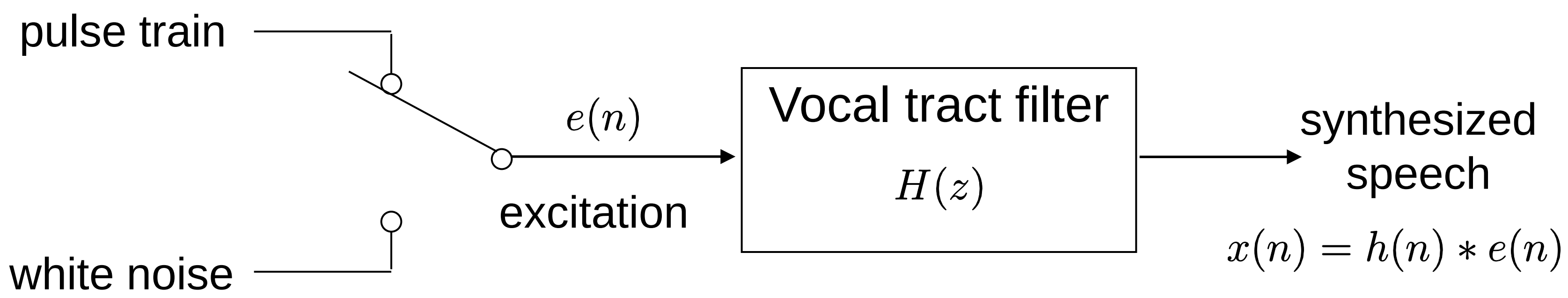
- Labels = regarded as hidden variable & marginalized

  → Robust against label errors
- Front- & back-end models are trained simultaneously

  → Combine text analysis & acoustic models as a unified model

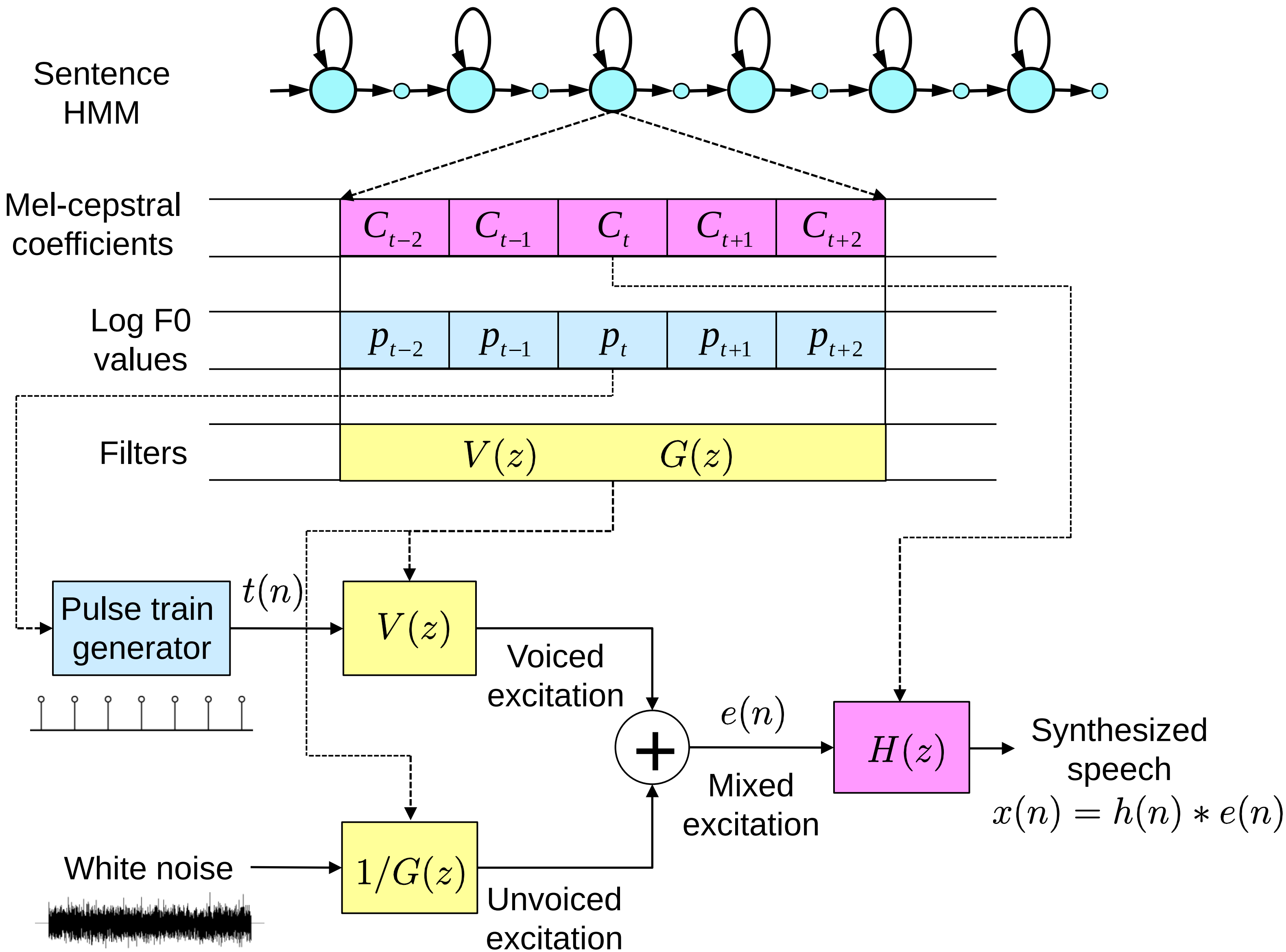
Simple pulse/noise vocoding

Basic pulse/noise vocoder

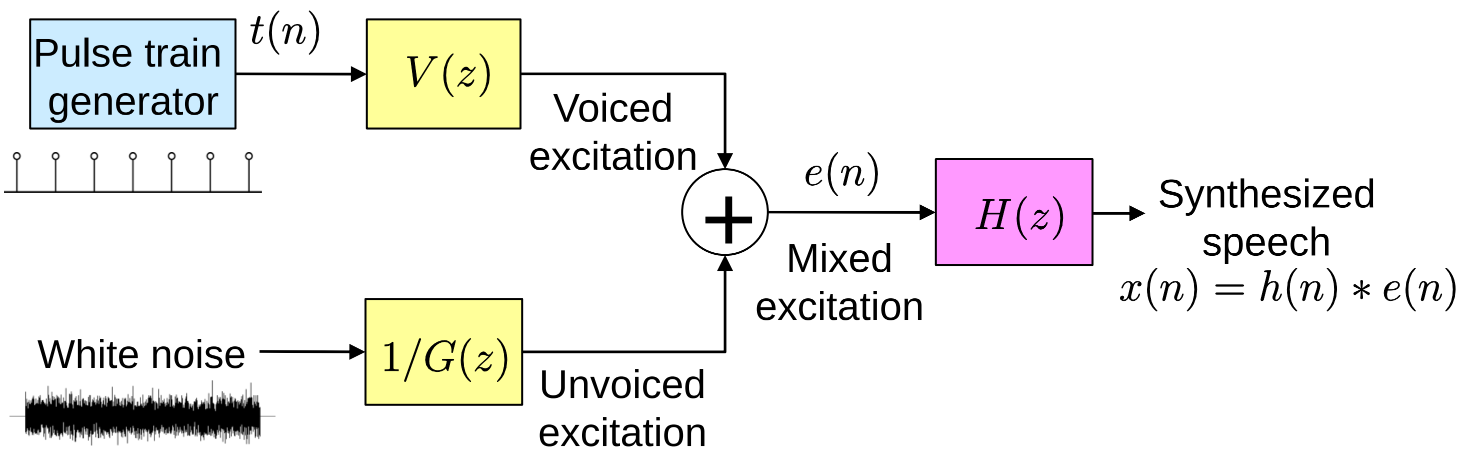


- Binary switching between voiced & unvoiced excitations
  → Difficult to represent mix of voiced & unvoiced sounds
- Excitations signals of human speech are not pulse or noise
  → Colored voiced/unvoiced excitations

State-dependent filtering [Maia;'07]



Waveform-level statistical model (1) [Maia;'10]



$$p(\boldsymbol{x} \mid \boldsymbol{q}, \boldsymbol{c}, \lambda_e) = |\boldsymbol{H_q}|^{-1} \mathcal{N}\left(\boldsymbol{H_q}^{-1}\boldsymbol{x} \; ; \; \boldsymbol{V_q}\boldsymbol{t}, \left(\boldsymbol{G_q}^\top \boldsymbol{G_q}\right)^{-1}\right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{l}, \lambda) = \sum_{\forall \boldsymbol{q}} \int p(\boldsymbol{x} \mid \boldsymbol{q}, \boldsymbol{c}, \lambda_e) p(\boldsymbol{c} \mid \boldsymbol{q}, \lambda_c) p(\boldsymbol{q} \mid \boldsymbol{l}, \lambda_c) d\boldsymbol{c}$$

$$p(\boldsymbol{x} \mid \boldsymbol{w}, \lambda) = \sum_{\forall \boldsymbol{l}} p(\boldsymbol{x} \mid \boldsymbol{l}, \lambda) p(\boldsymbol{l} \mid \boldsymbol{w})$$ $\Leftarrow$ waveform-level statistical model

$\boldsymbol{H_q}, \boldsymbol{V_q}, \boldsymbol{G_q}$ matrices representing impulse responses of $H(z)$, $V(z)$, & $G(z)$

$\lambda = \{\lambda_e, \lambda_c\}$ set of acoustic (λ_c) & excitation (λ_e) model parameters

Waveform-level statistical model (2) [Maia;'10]

Integral & sum are intractable ☹

→ Approx integral & sum by joint max

$$p(\boldsymbol{x} \mid \boldsymbol{l}, \lambda) = \sum_{\forall \boldsymbol{q}} \int p(\boldsymbol{x} \mid \boldsymbol{q}, \boldsymbol{c}, \lambda) p(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) p(\boldsymbol{q} \mid \boldsymbol{l}, \lambda) d\boldsymbol{c}$$

$$\approx p(\boldsymbol{x} \mid \hat{\boldsymbol{q}}, \hat{\boldsymbol{c}}, \lambda) p(\hat{\boldsymbol{c}} \mid \hat{\boldsymbol{q}}, \lambda) p(\hat{\boldsymbol{q}} \mid \boldsymbol{l}, \lambda) = p(\boldsymbol{x}, \hat{\boldsymbol{q}}, \hat{\boldsymbol{c}} \mid \boldsymbol{l}, \lambda)$$

iteratively optimize λ & $\boldsymbol{C}$

$$\hat{\boldsymbol{C}} = \arg\max_{\boldsymbol{C}} p(\boldsymbol{X} \mid \hat{\boldsymbol{Q}}, \boldsymbol{C}, \hat{\lambda}) p(\boldsymbol{C} \mid \hat{\boldsymbol{Q}}, \hat{\lambda}) p(\hat{\boldsymbol{Q}} \mid \hat{\boldsymbol{L}}, \hat{\lambda})$$ $\Leftarrow$ estimate $\boldsymbol{C}$ given $\hat{\lambda}$

$$\hat{\lambda} = \arg\max_{\lambda} p(\boldsymbol{X} \mid \hat{\boldsymbol{Q}}, \hat{\boldsymbol{C}}, \lambda) p(\hat{\boldsymbol{C}} \mid \hat{\boldsymbol{Q}}, \lambda) p(\hat{\boldsymbol{Q}} \mid \hat{\boldsymbol{L}}, \lambda)$$ $\Leftarrow$ estimate λ given $\hat{\boldsymbol{C}}$

Conventional → step-by-step maximization

Proposed → iterative joint maximization

Outline

HMM-based speech synthesis

- Overview
- Implementation of individual components

Bayesian framework for speech synthesis

- Formulation
- Realizations in HMM-based speech synthesis
- Recent works

Conclusions

- Summary
- Future research topics

Summary

HMM-based speech synthesis

- Statistical parametric speech synthesis approach
- Source-filter representation of speech + statistical acoustic modeling
- Getting popular

Bayesian framework for speech synthesis

- Formulation
- Decomposition to sub-problems
- Correspondence between sub-problems & modules in HMM-based speech synthesis system
- Recent works to relax approximations

Drawbacks of HMM-based speech synthesis

Quality of synthesized speech

- Buzzy
- Flat
- Muffled

Three major factors degrade the quality

- Poor vocoding
  → how to parameterize speech?
- Inaccurate acoustic modeling
  → how to model extracted speech parameter trajectories?
- Over-smoothing
  → how to recover generated speech parameter trajectories?

Still need a lot of works to improve the quality

Future challenging topics in speech synthesis

Keynote speech by Simon King in ISCA SSW7 last year

Speech synthesis is easy, if ...

- voice is built offline & carefully checked for errors
- speech is recorded in clean conditions
- word transcriptions are correct
- accurate phonetic labels are available or can be obtained
- speech is in the required language & speaking style
- speech is from a suitable speaker
- a native speaker is available, preferably a linguist

Speech synthesis is not easy if we don’t have right data

Future challenging topics in speech synthesis

Non-professional speakers

- AVM + adaptation (CSTR)

Too little speech data

- VTLN-based rapid speaker adaptation (Titech, IDIAP)

Noisy recordings

- Spectral subtraction & AVM + adaptation (CSTR)

No labels

- Un- / Semi-supervised voice building (CSTR, NICT, CMU, Toshiba)

Insufficient knowledge of the language or accent

- Letter (grapheme)-based synthesis (CSTR)
- No prosodic contexts (CSTR, Titech)

Wrong language

- Cross-lingual speaker adaptation (MSRA, EMIME)
- Speaker & language adaptive training (Toshiba)

Thanks!