Speech Synthesis as A Machine Learning Problem

Keiichi Tokuda

Nagoya Institute of Technology

Introduction

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

– Hand-crafting each phonetic units by rules

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

– Concatenate speech units (waveform) from a database

- Single inventory: diphone synthesis
- Multiple inventory: unit selection synthesis

Corpus-based, *statistical parametric synthesis*

– Source-filter model + statistical acoustic model

- Hidden Markov model: HMM-based synthesis

How we can formulate and understand the whole corpus-based speech synthesis process in a unified statistical framework?

Problem of speech synthesis (1/2)

We have a speech database, i.e., a set of texts and corresponding speech waveforms.
Given a text to be synthesized, what is the speech waveform corresponding to the text?

$\boldsymbol{W}$: texts
$\boldsymbol{X}$: speech waveforms
} database

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

$\boldsymbol{x}$: speech waveform ← ?

Problem of speech synthesis (2/2)

Problem of statistical parametric speech synthesis:

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{l}, \boldsymbol{O}, \boldsymbol{L})$$
$$= \arg\max_{\boldsymbol{o}} \int \underline{P(\boldsymbol{o} \mid \boldsymbol{l}, \boldsymbol{\lambda})} \underline{P(\boldsymbol{\lambda} \mid \boldsymbol{O}, \boldsymbol{L})} d\boldsymbol{\lambda}$$

$\boldsymbol{L}$: Label sequence for training data
(pronunciation, stress, POS, pause position, etc.)

$\boldsymbol{O}$: Training data (speech parameter sequence)

$\boldsymbol{l}$: Label sequence for synthesis data

$\boldsymbol{o}$: Synthesis data (speech parameter sequence)

$\boldsymbol{\lambda}$: Model parameters (HMM parameters)

Approximation

Maximum A Posterior (MAP) & ML estimation

$$\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{\lambda} \mid \boldsymbol{O}, \boldsymbol{L})$$

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

MAP estimation

$$\hat{\boldsymbol{\lambda}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{O} \mid \boldsymbol{L}, \boldsymbol{\lambda})$$

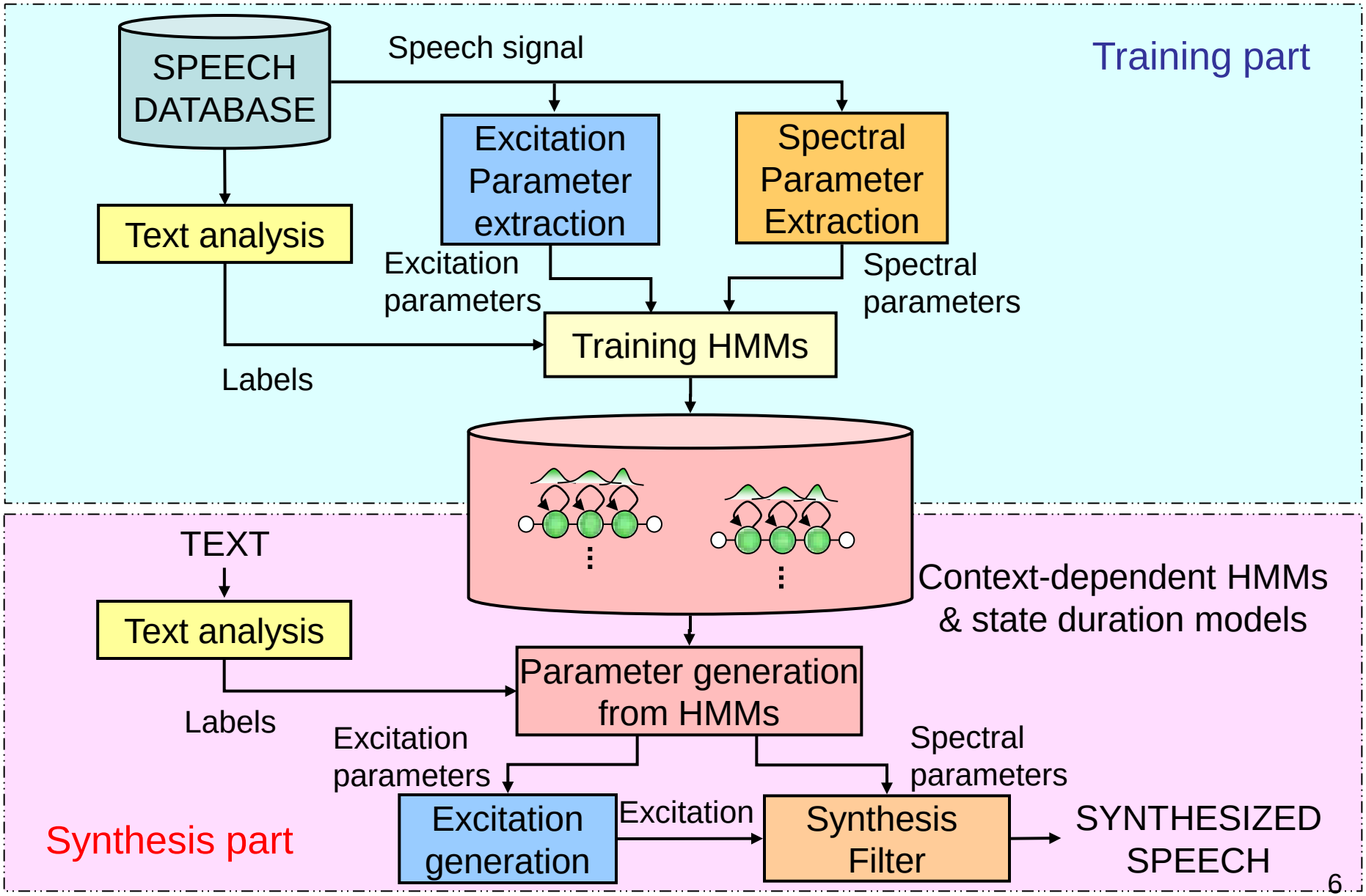
ML estimation

Speech parameter generation using estimated parameters

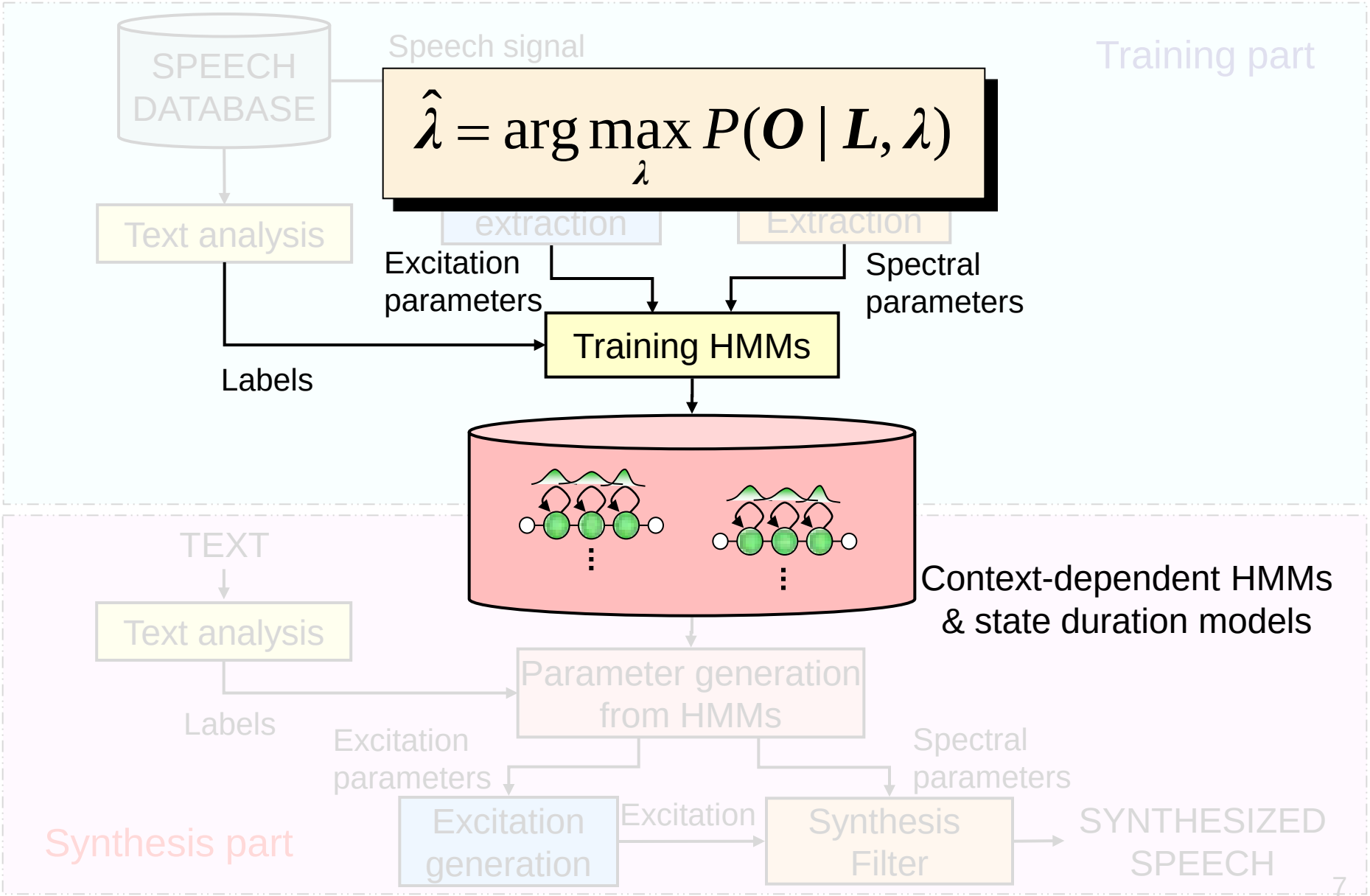
$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{l}, \hat{\boldsymbol{\lambda}})$$

Speech parameter generation

HMM-based speech synthesis system



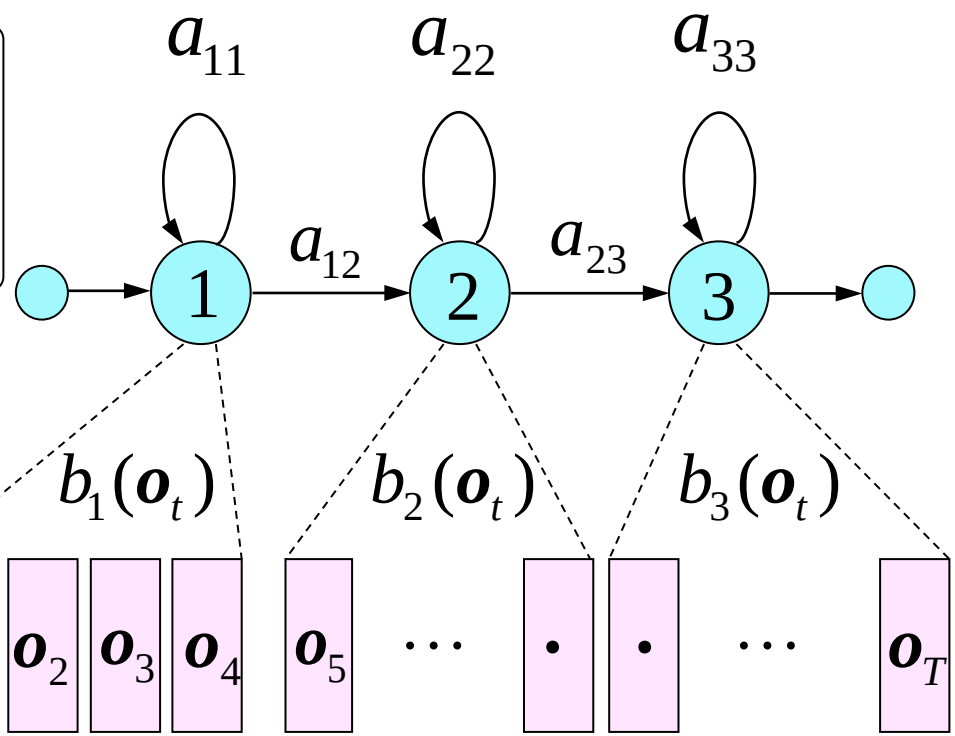
HMM-based speech synthesis system



Hidden Markov model (HMM)

a_{ij} : state transition probability

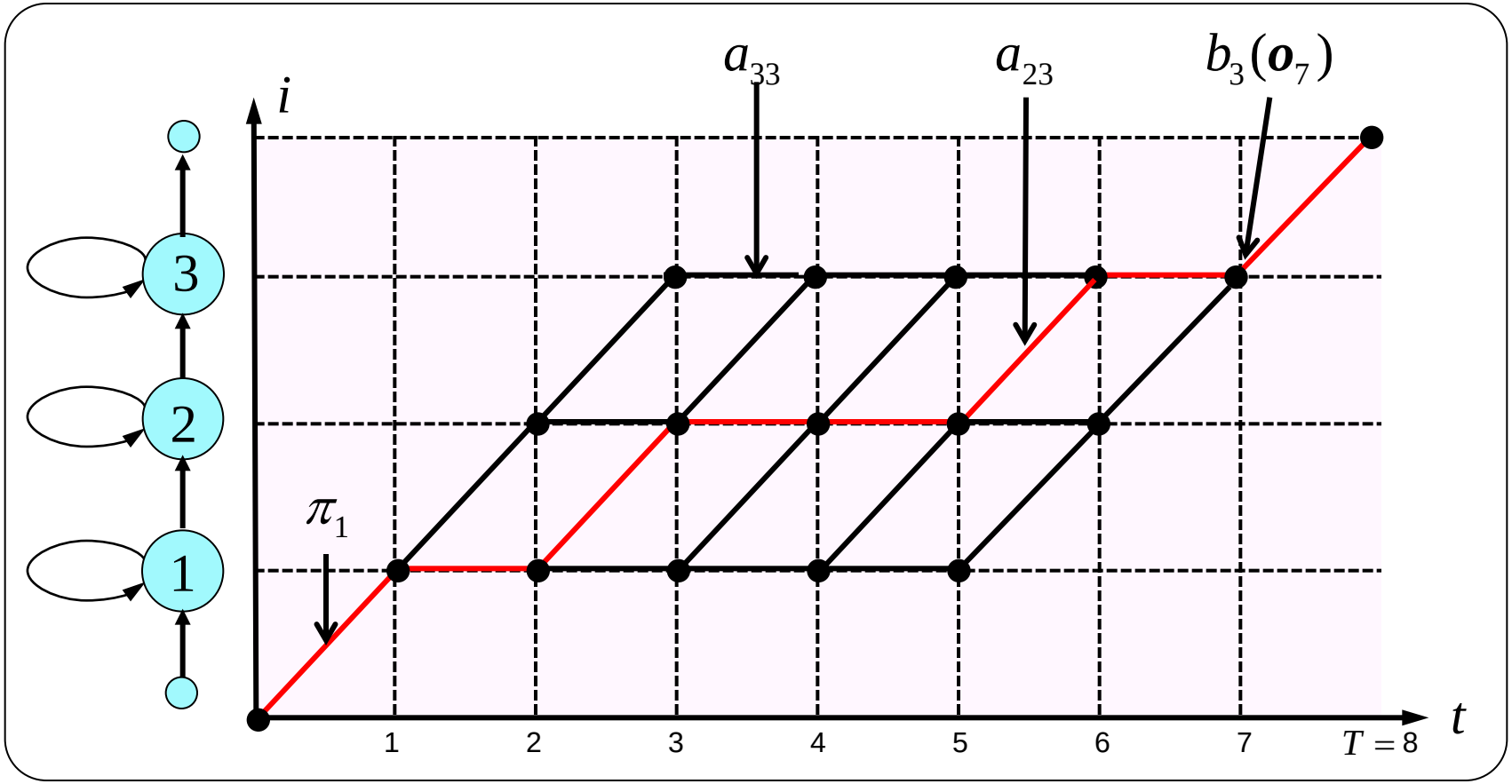
$b_q(\boldsymbol{o}_t)$: output probability



Observation sequence $\boldsymbol{o}$

State sequence $\boldsymbol{q}$ 1 1 1 1 2 … 2 3 … 3

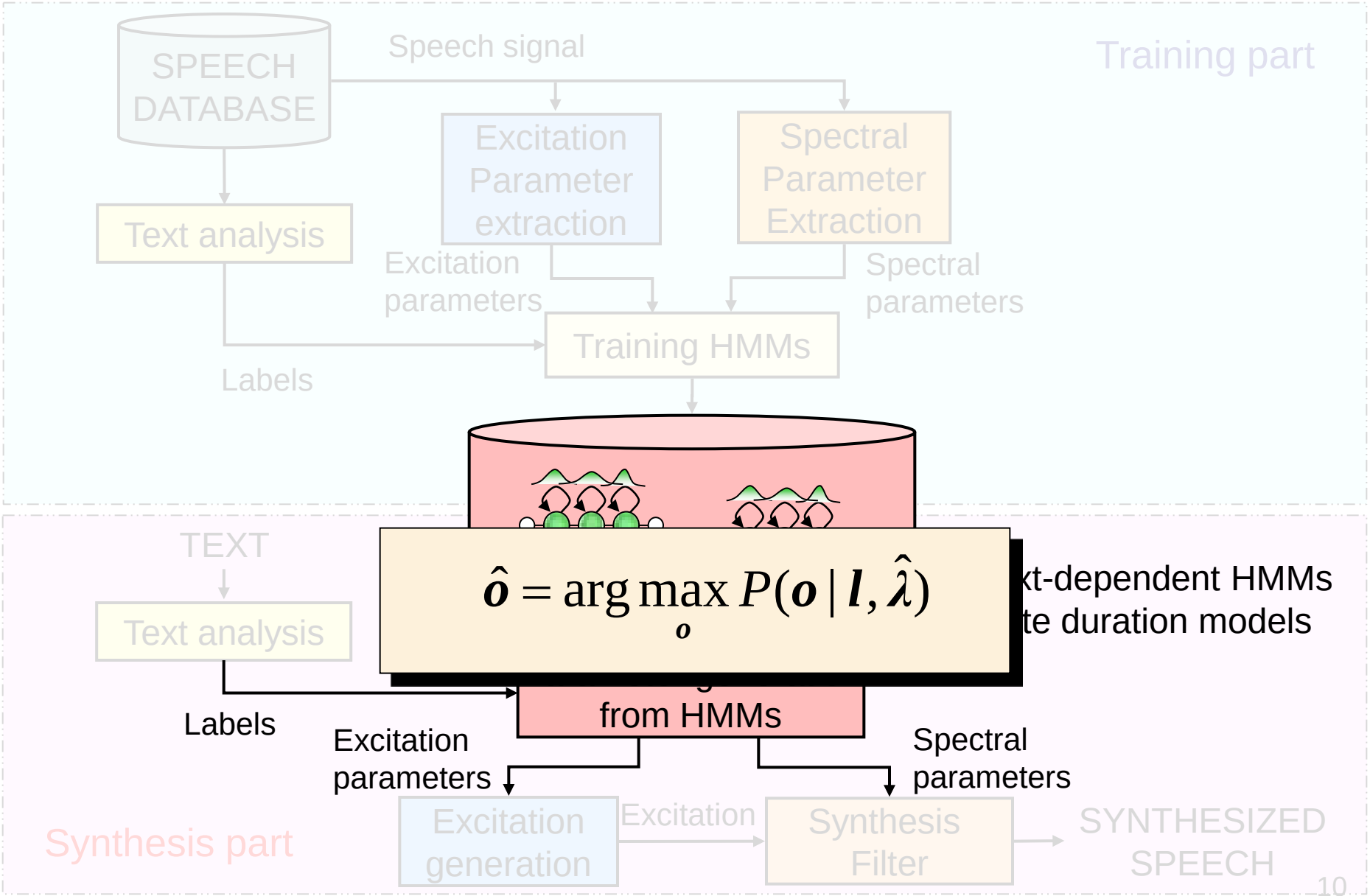
Output probability of HMM



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

Model parameters can be estimated by an EM algorithm

HMM-based speech synthesis system



Speech parameter generation algorithm

For 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

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

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

$$\Downarrow$$

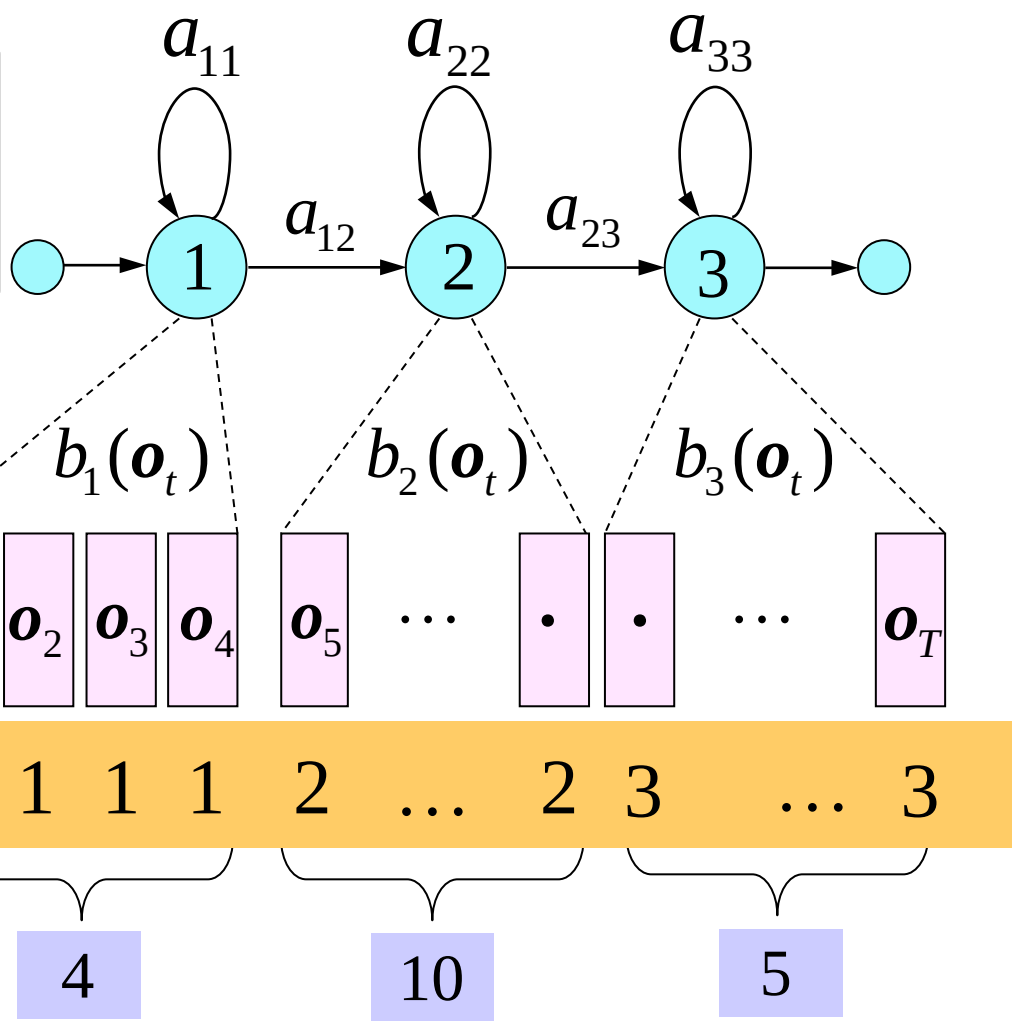
$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} p(\boldsymbol{q} \mid \boldsymbol{l}, \boldsymbol{\lambda})$$

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

Determination of state sequence

a_{ij} : state transition probability

$b_q(\boldsymbol{o}_t)$: output probability



The state sequence can be determined by state durations

Hidden Semi Markov Model

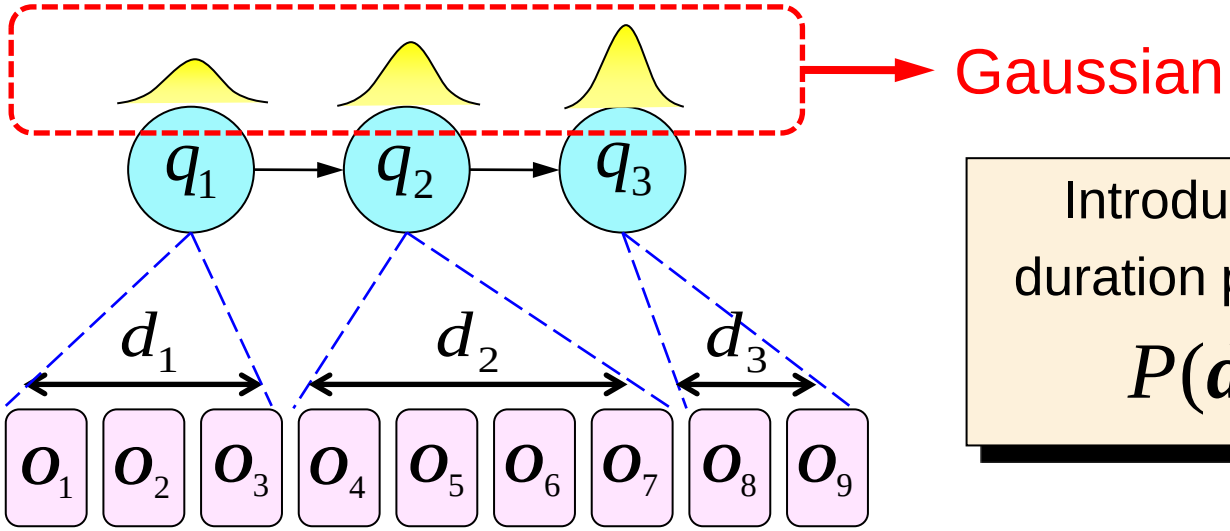
HMM (Hidden Markov Model)

– State duration prob. depends only on transition prob.

– State duration probability exponentially decreases

HSMM (Hidden Semi Markov Model)

– HMM + explicit duration model ⇒ HSMM



Introduce state duration probability

$$P(\boldsymbol{d} \mid \boldsymbol{q})$$

State durations are given by means of Gaussians.

Speech parameter generation algorithm

For 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

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

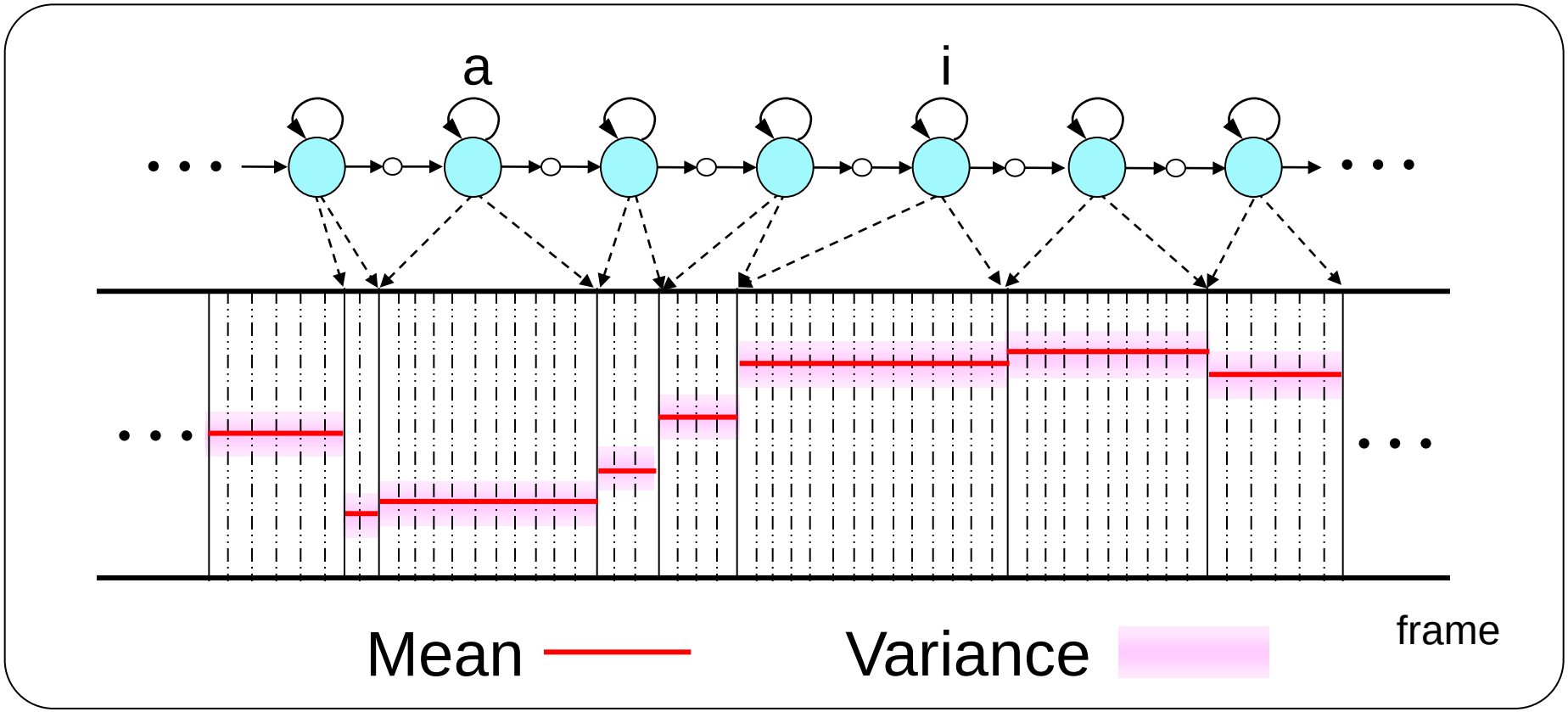
$$\approx \max_{\boldsymbol{q}} P(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q} \mid \boldsymbol{\lambda})$$

$$\Downarrow$$

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

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

Generated feature sequence



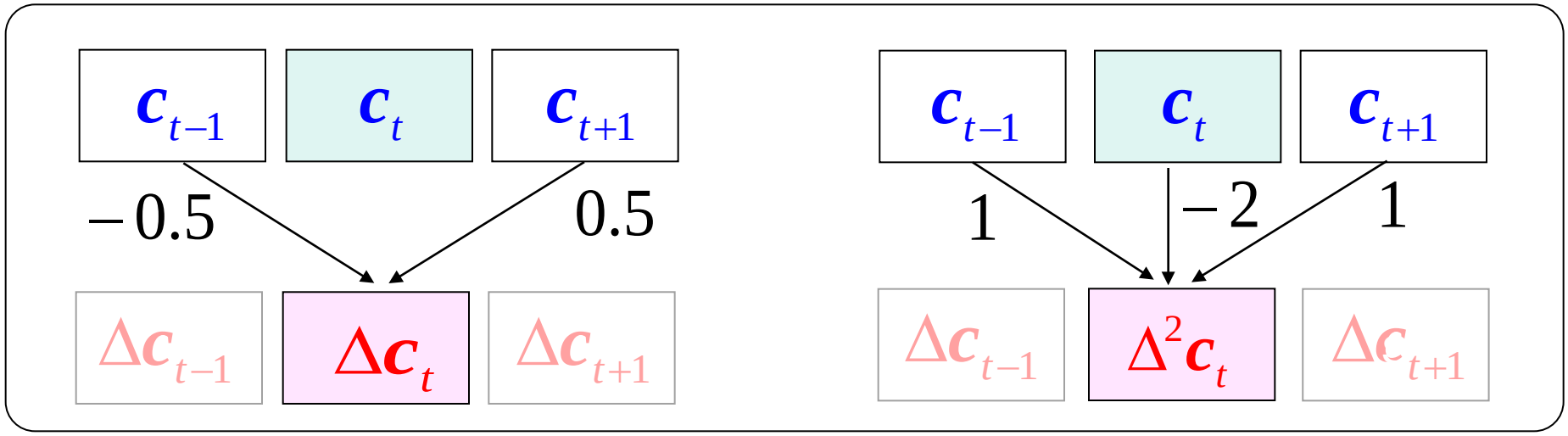
$\hat{\boldsymbol{o}}$ becomes a sequence of mean vectors

⇒ discontinuous outputs between states

Dynamic features

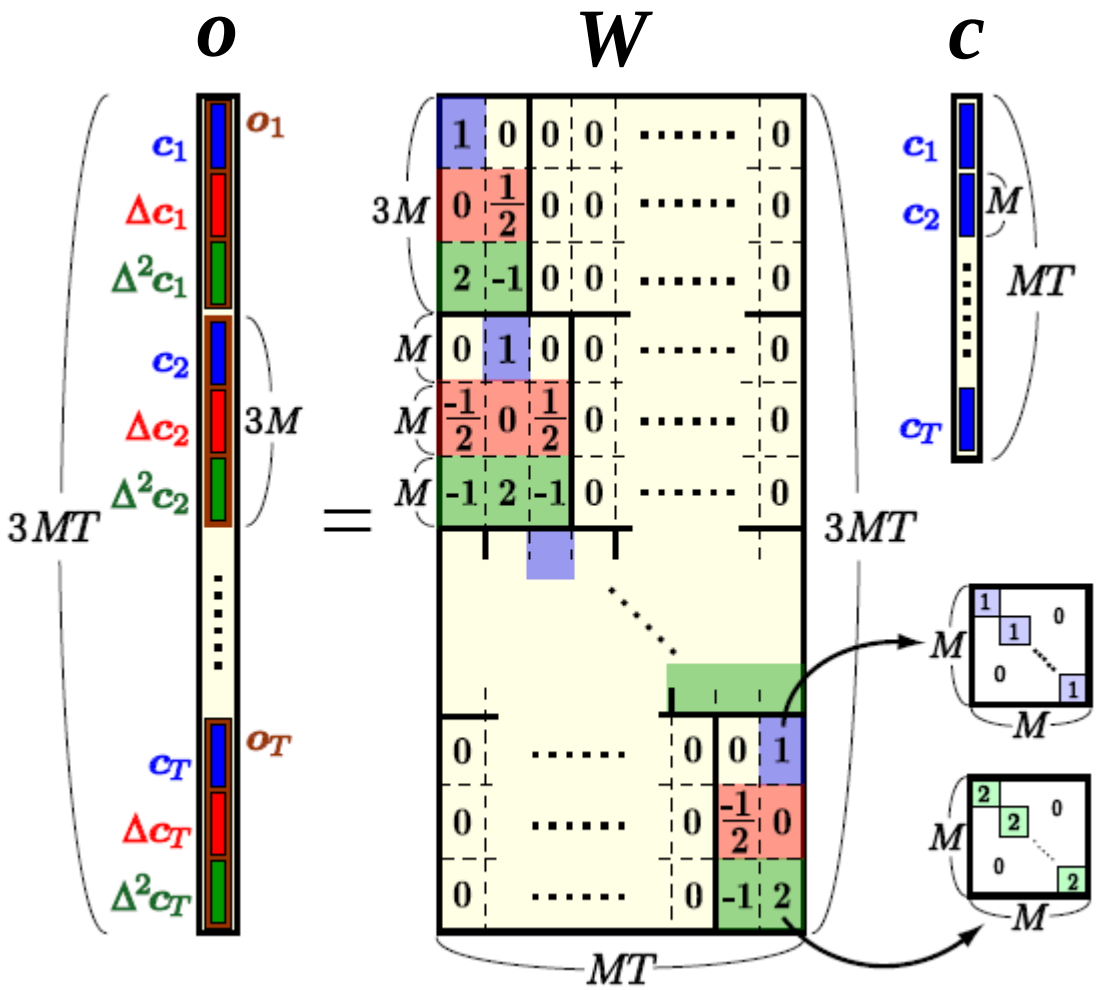
$$\Delta \boldsymbol{c}_t = \frac{\partial \boldsymbol{c}_t}{\partial \boldsymbol{t}} \approx 0.5(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1})$$

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



Integration of dynamic features

Relationship between speech parm. vec. & static feat



Solution for the problem (1/2)

By setting

$$\frac{\partial \log P(\boldsymbol{Wc} \mid \hat{\boldsymbol{q}}, \lambda)}{\partial \boldsymbol{c}} = \boldsymbol{0}$$

we obtain

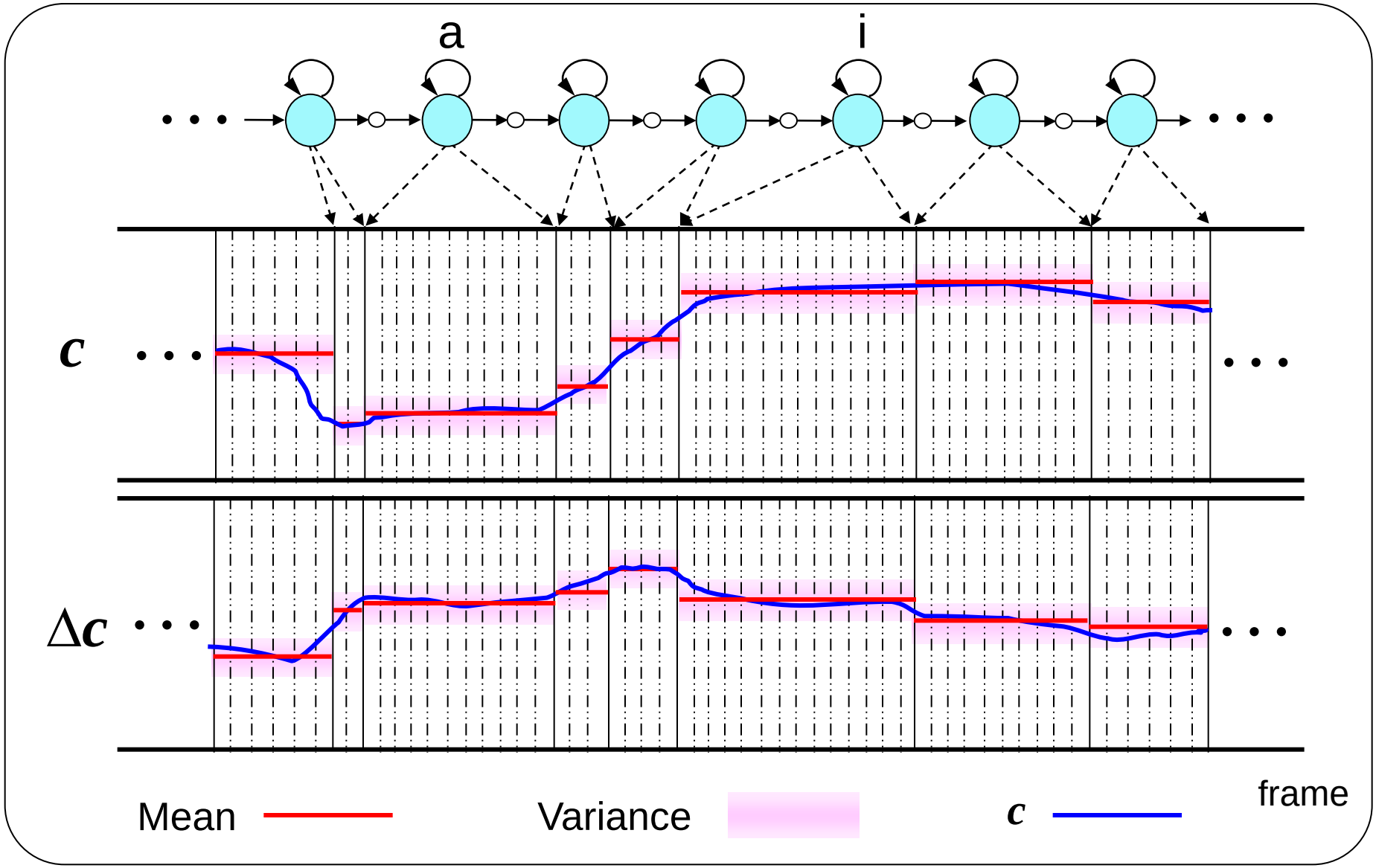
$$\boldsymbol{W}^{\top} \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{Wc} = \boldsymbol{W}^{\top} \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}$$

where

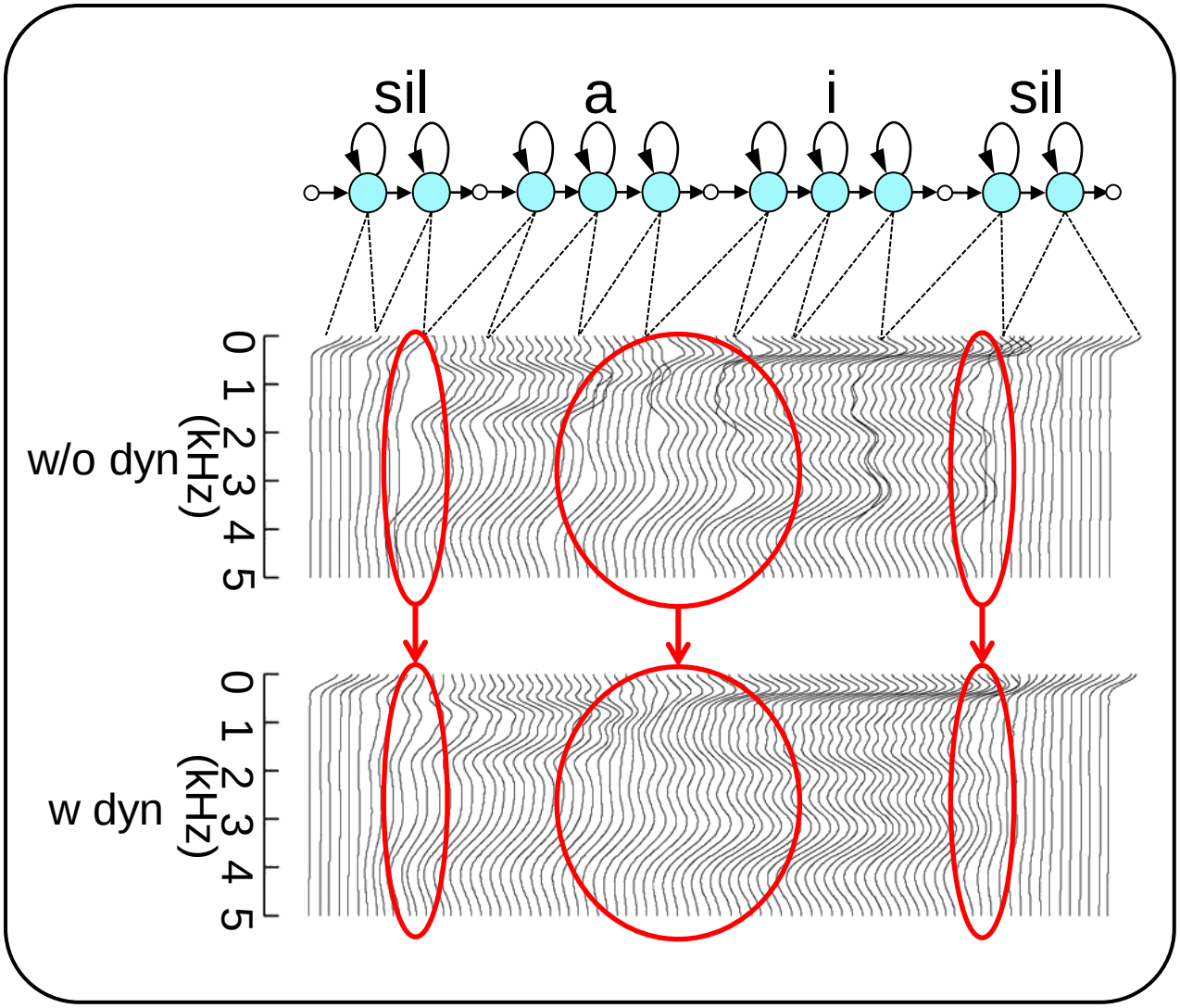
$$\boldsymbol{c} = [\boldsymbol{c}_1^{\top}, \boldsymbol{c}_2^{\top}, \ldots, \boldsymbol{c}_T^{\top}]^{\top}$$

$$\boldsymbol{\mu}_{\hat{\boldsymbol{q}}} = [\boldsymbol{\mu}_{\hat{\boldsymbol{q}}_1}^{\top}, \boldsymbol{\mu}_{\hat{\boldsymbol{q}}_2}^{\top}, \ldots, \boldsymbol{\mu}_{\hat{\boldsymbol{q}}_T}^{\top}]^{\top}$$

Generated speech parameter trajectory



Generated spectra



Spectra changing smoothly between phonemes

Trajectory HMM

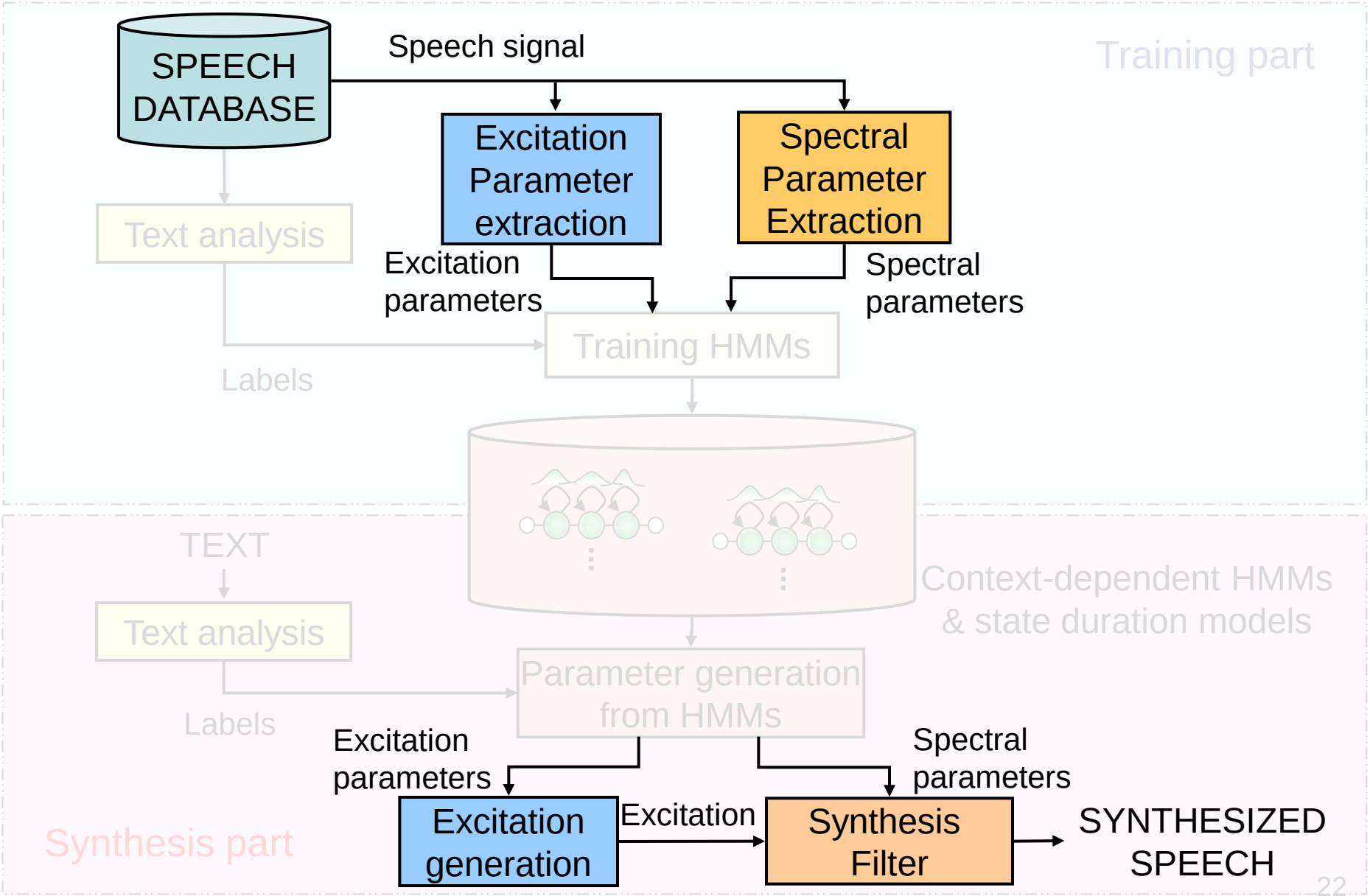
$P(\boldsymbol{o} \mid \boldsymbol{l}, \lambda) = P(\boldsymbol{Wc} \mid \boldsymbol{l}, \lambda)$ is not a proper distribution of $\boldsymbol{c}$

	Conventional HMM	Trajectory HMM
Training	$P(\boldsymbol{O} \mid \boldsymbol{L}, \boldsymbol{\lambda})$	$P(\boldsymbol{C} \mid \boldsymbol{L}, \boldsymbol{\lambda})$
Synthesis	$P(\boldsymbol{o} \mid \boldsymbol{l}, \hat{\boldsymbol{\lambda}}) \mid_{\boldsymbol{o}=\boldsymbol{Wc}}$ $\Longleftrightarrow$	$P(\boldsymbol{c} \mid \boldsymbol{l}, \hat{\boldsymbol{\lambda}})$

Solve inconsistency between training & synthesis

$\Rightarrow$ improving the model accuracy

HMM-based speech synthesis system

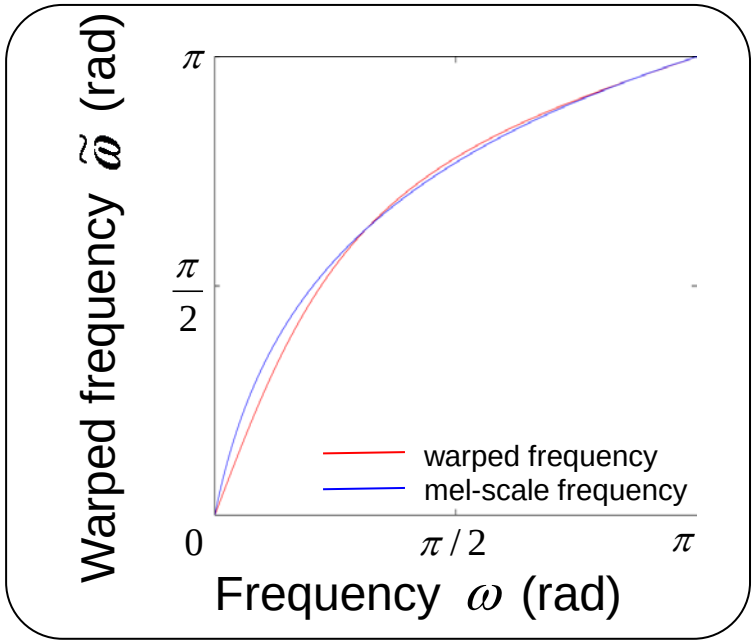


ML estimation of spectral parameter

Mel-cepstral representation of speech spectra

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

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} = e^{-j\tilde{\omega}}$$



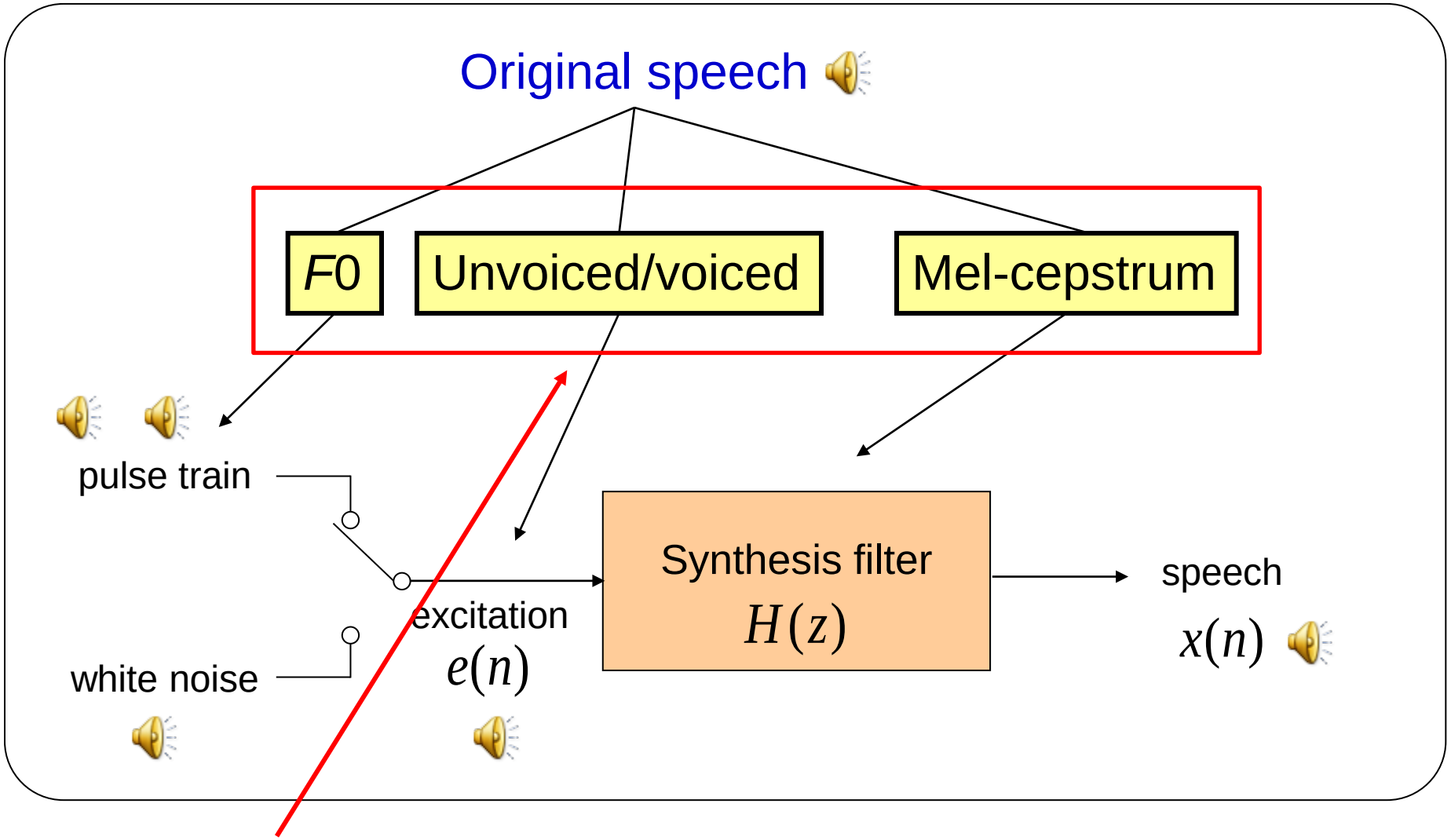
ML-estimation of mel-cepstrum

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

$\boldsymbol{x}$: speech waveform (Gaussian process)

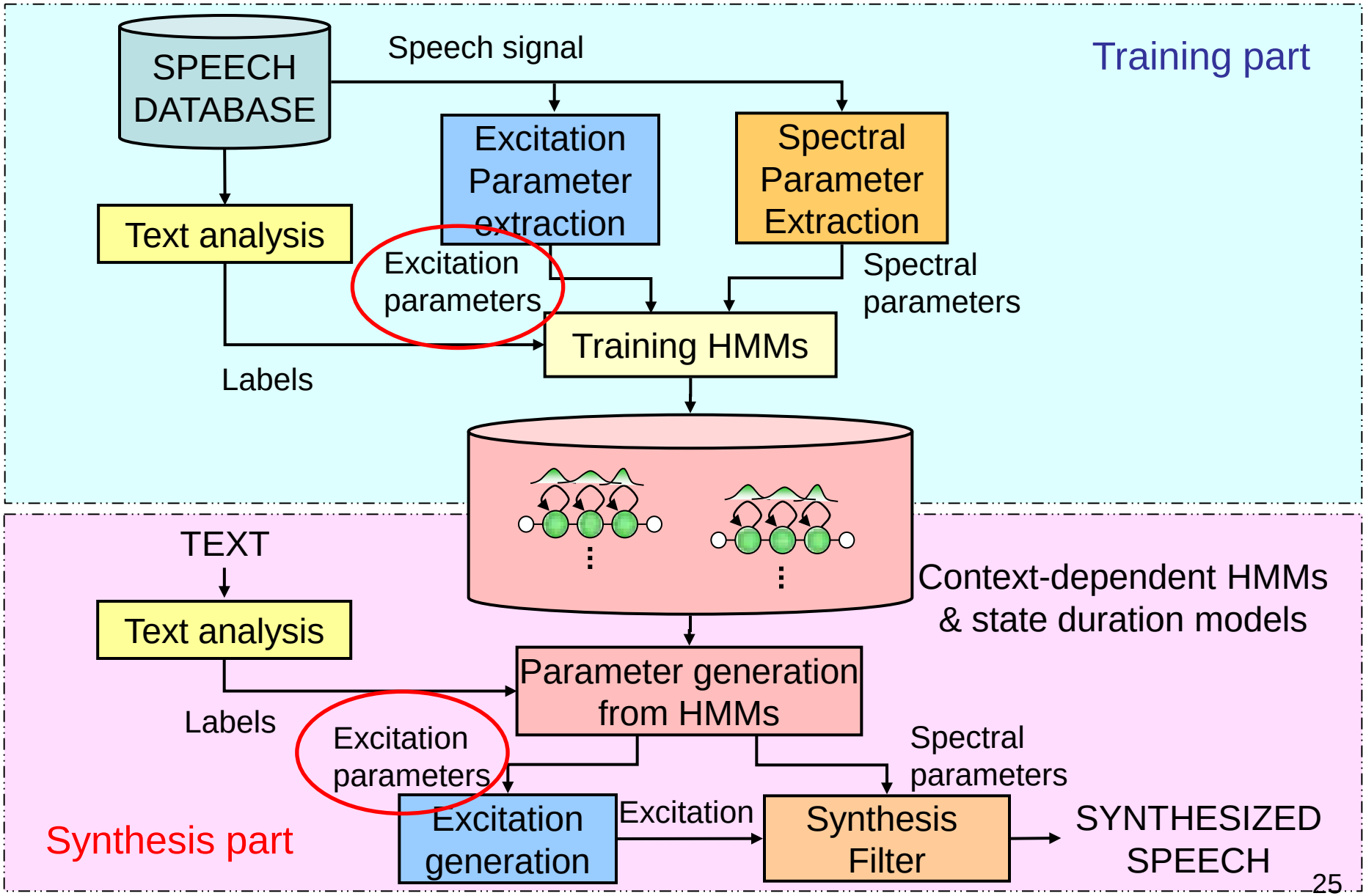
$\boldsymbol{c}$: mel-cepstrum

Overview of speech vocoding

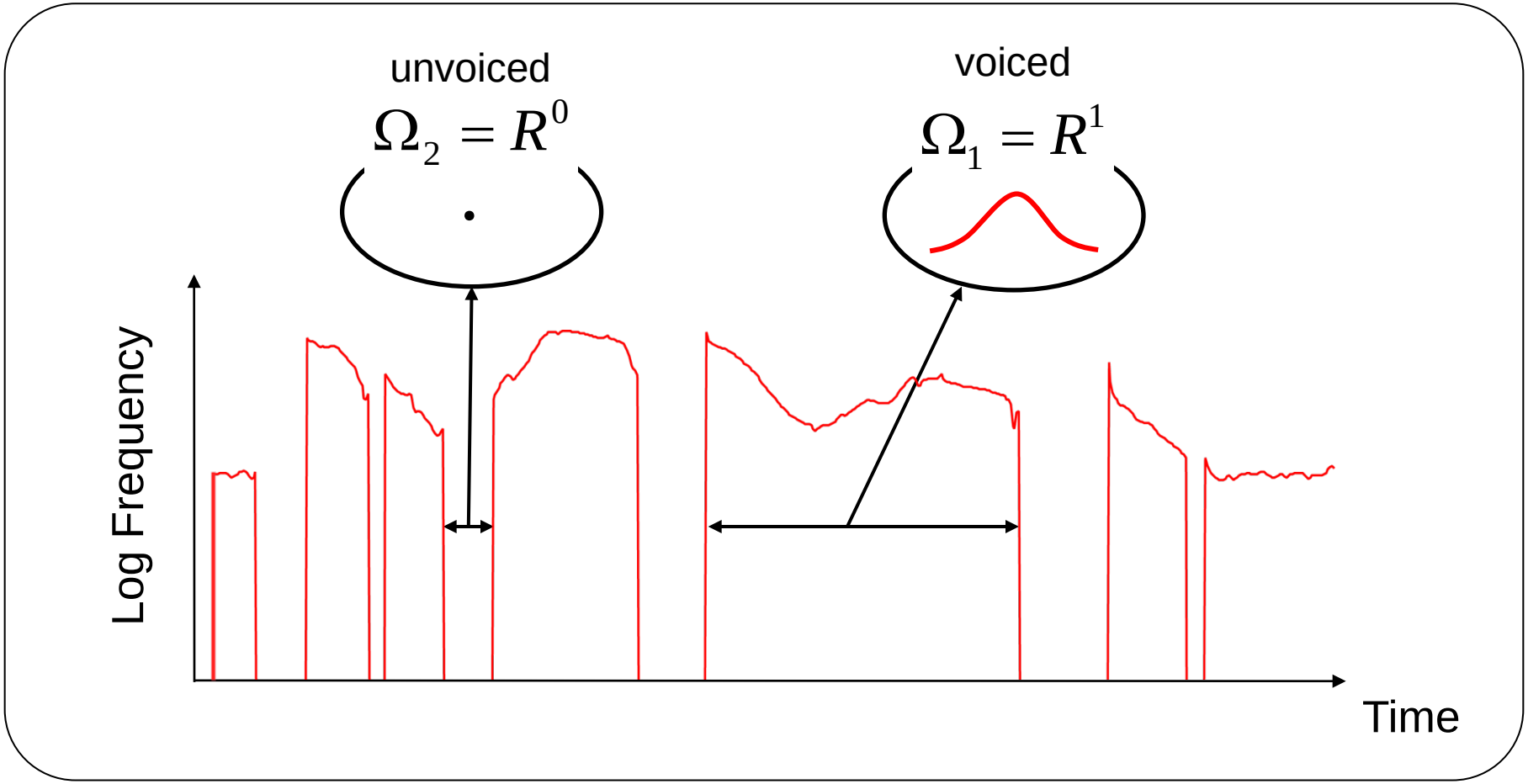


These speech parameter modeled by HMM

HMM-based speech synthesis system



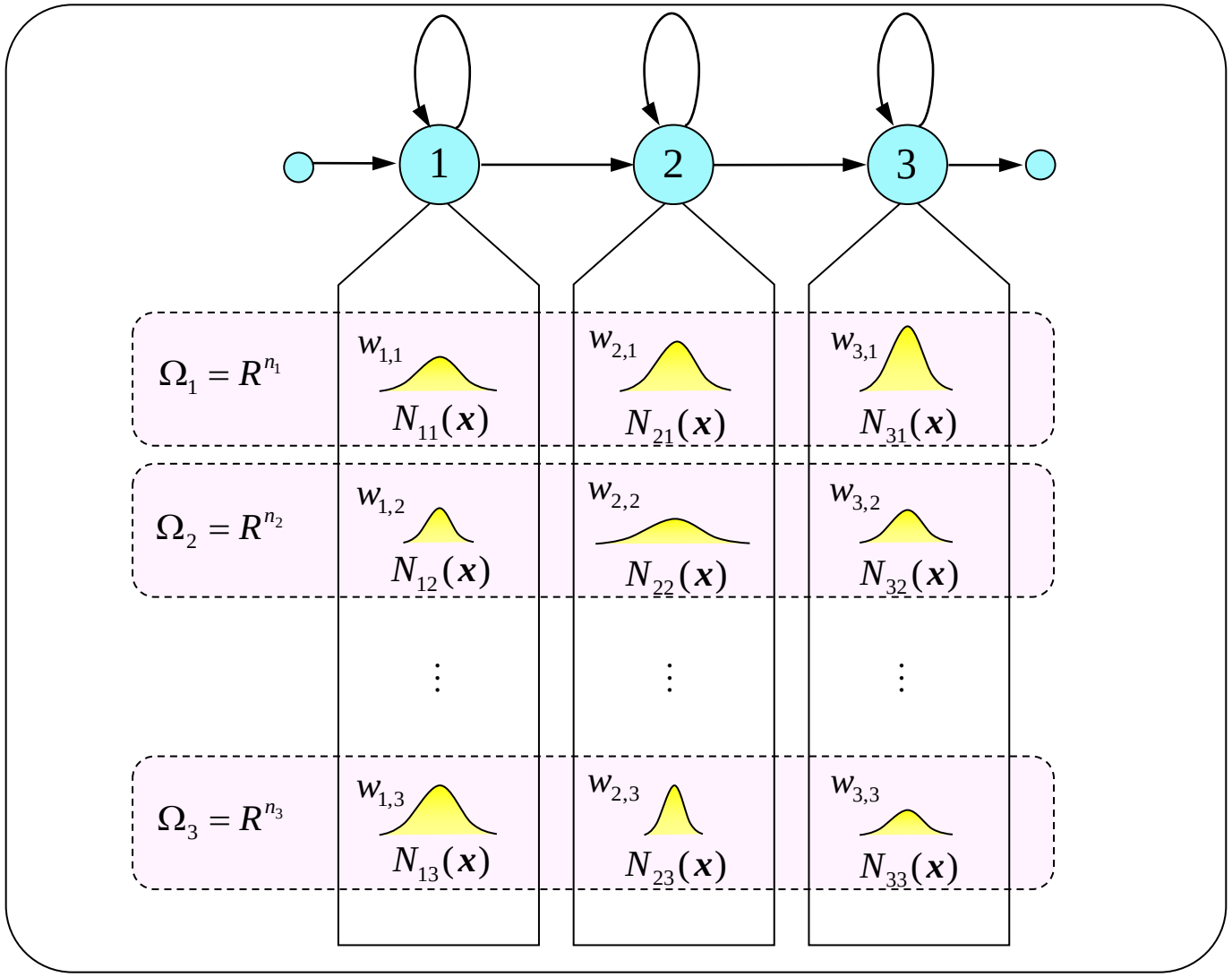
Observation of F0



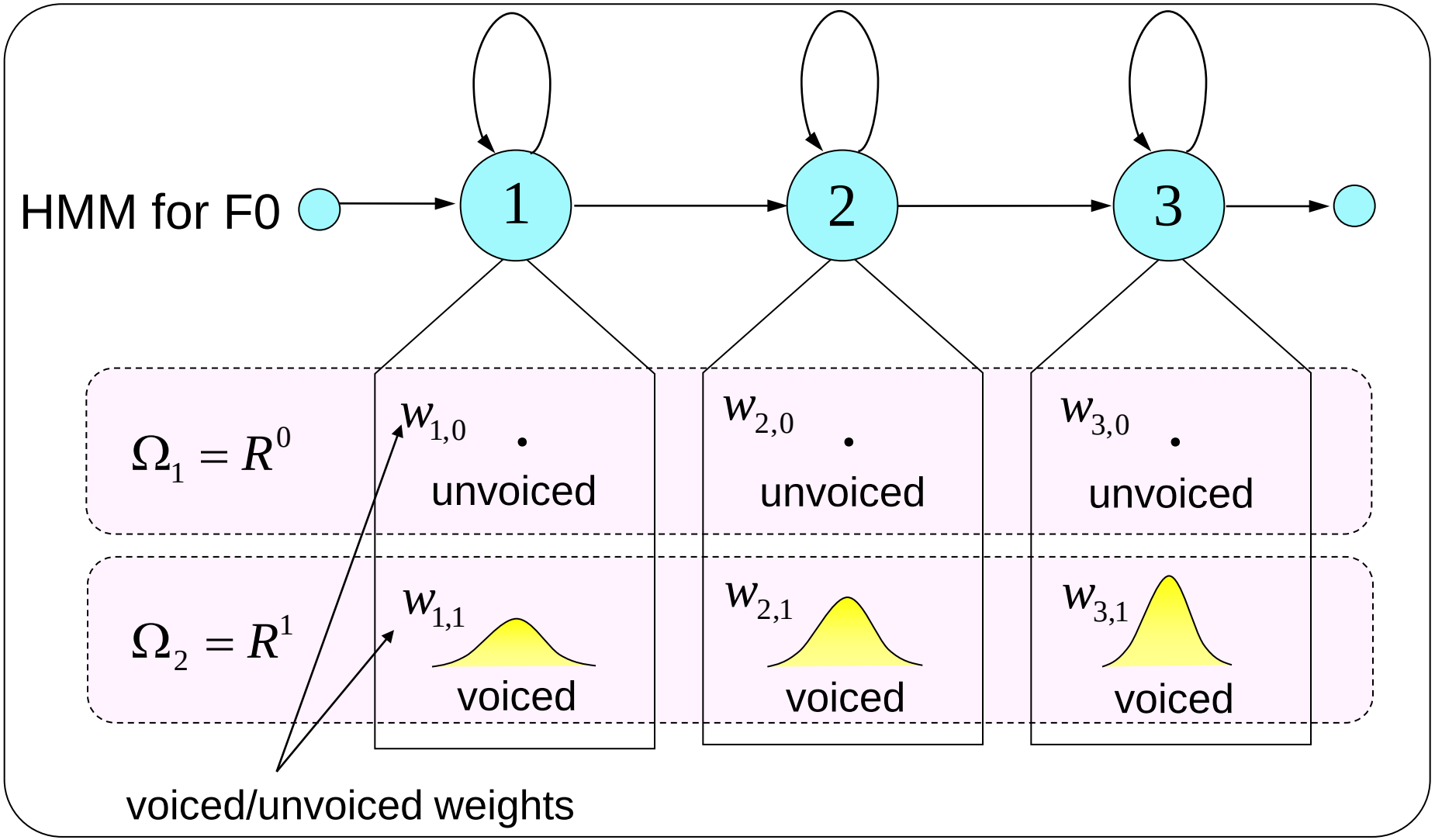
Unable to model by continuous or discrete distributions

⇒ Multi-space distribution HMM (MSD-HMM)

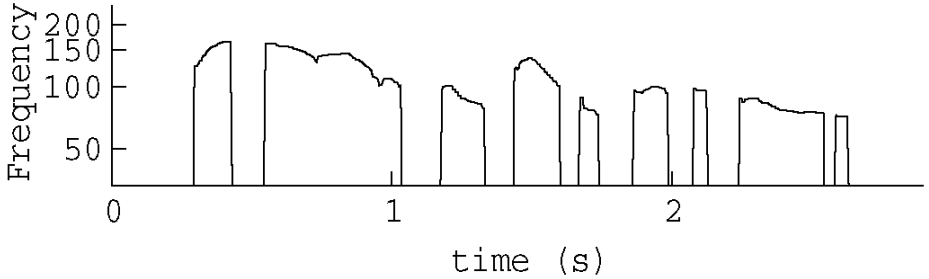
Structure of MSD-HMM



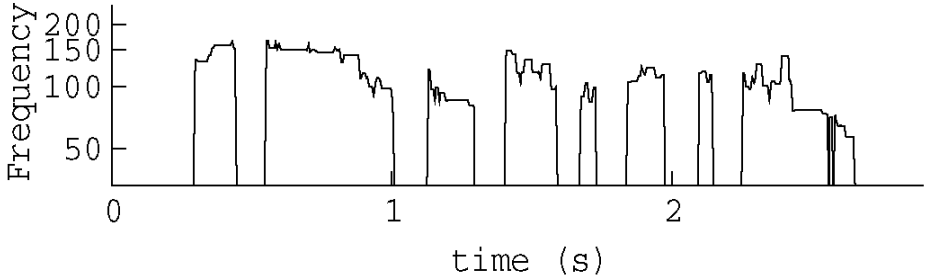
MSD-HMM for F0 modeling



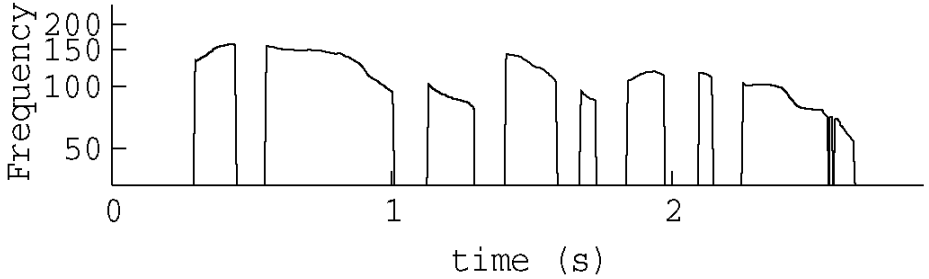
Generated F0



natural speech



without dynamic features



with dynamic features (Δ+Δ^2)

Speech samples

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

Inclusion of speech analysis & waveform reconstruction

Problem of statistical parametric speech synthesis

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{x} \mid \boldsymbol{l}, \boldsymbol{X}, \boldsymbol{L})$$

$$= \arg\max_{\boldsymbol{x}} \sum_{\boldsymbol{c}} \sum_{\boldsymbol{C}} \int \underbrace{P(\boldsymbol{x} \mid \boldsymbol{c})}_{\text{Waveform reconstruction}} \underbrace{P(\boldsymbol{c} \mid \boldsymbol{l}, \boldsymbol{\lambda})}_{\text{Speech parameter generation}}$$

$$\times \underbrace{P(\boldsymbol{\lambda} \mid \boldsymbol{C}, \boldsymbol{L})}_{\text{HMM parameter estimation}} \underbrace{P(\boldsymbol{C} \mid \boldsymbol{X})}_{\text{Speech analysis}} d\boldsymbol{\lambda}$$

$\boldsymbol{c}$ and $\boldsymbol{C}$ consist of mel-cepstrum & F0 parameters

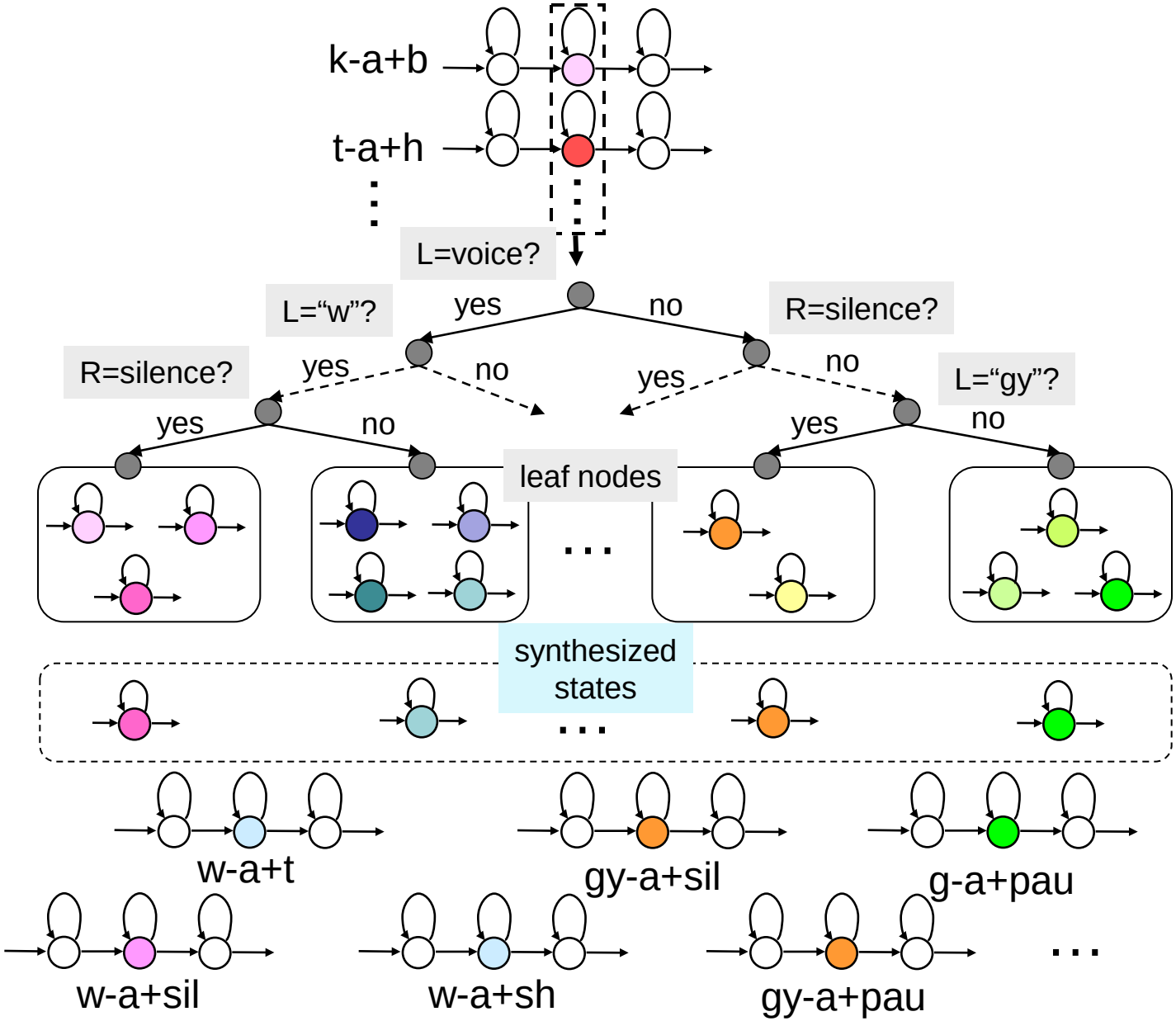
Context clustering factors

- {preceding, succeeding} two phonemes
- Current phoneme
- 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} {accented, stressed} syllable in current phrase
- # of syllables {from previous, to next} {accented, stressed} 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

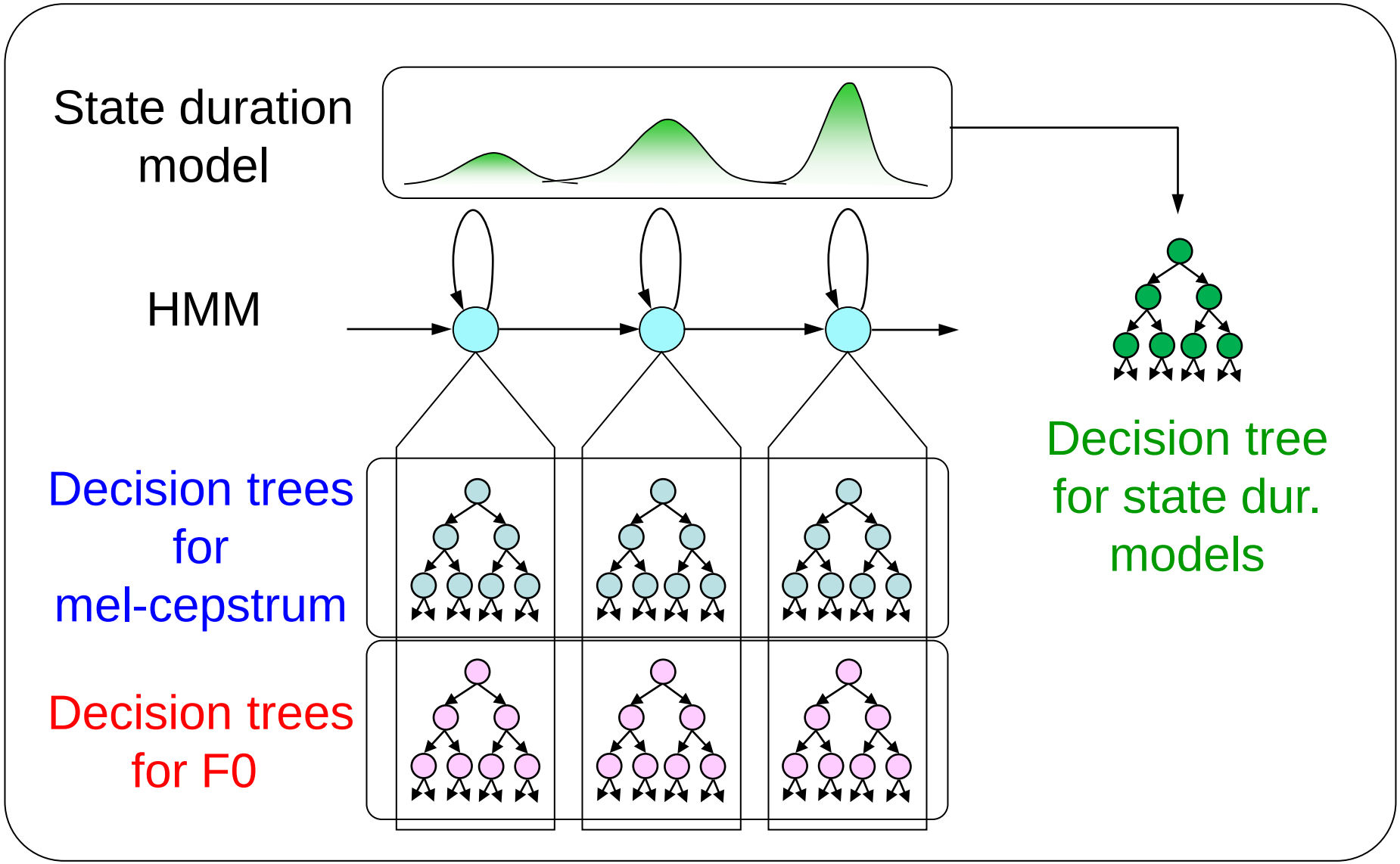
…

Vast # of combinations ⇒ Difficult to have possible models

Decision tree-based state clustering



Stream-dependent tree-based clustering



Inclusion of model structure parameter

Problem of statistical parametric speech synthesis

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{l}, \boldsymbol{O}, \boldsymbol{L})$$

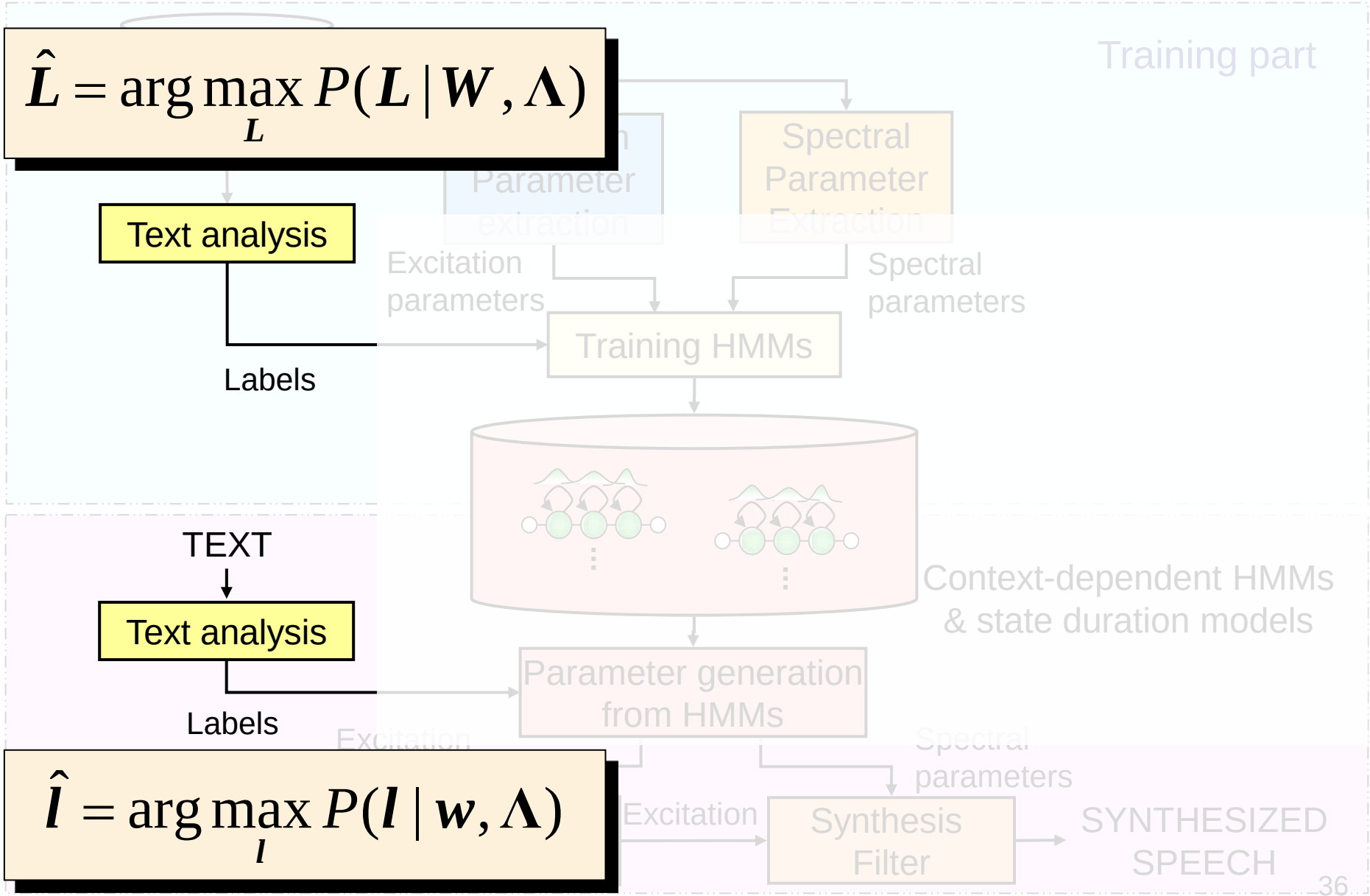
$$= \arg\max_{\boldsymbol{o}} \sum_{m} \int \underbrace{P(\boldsymbol{o} \mid \boldsymbol{l}, \boldsymbol{\lambda}, m)}_{\text{Generate coefficient}}$$

$$\times \underbrace{P(\boldsymbol{\lambda} \mid m, \boldsymbol{O}, \boldsymbol{L})}_{\text{Posterior of model parameters}} \underbrace{P(m \mid \boldsymbol{O}, \boldsymbol{L})}_{\text{Posterior of model structure}} d\boldsymbol{\lambda}$$

Usually fixed m is used

HMM-based speech synthesis system

$$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} P(\boldsymbol{L} \mid \boldsymbol{W}, \boldsymbol{\Lambda})$$



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

Inclusion of text analysis

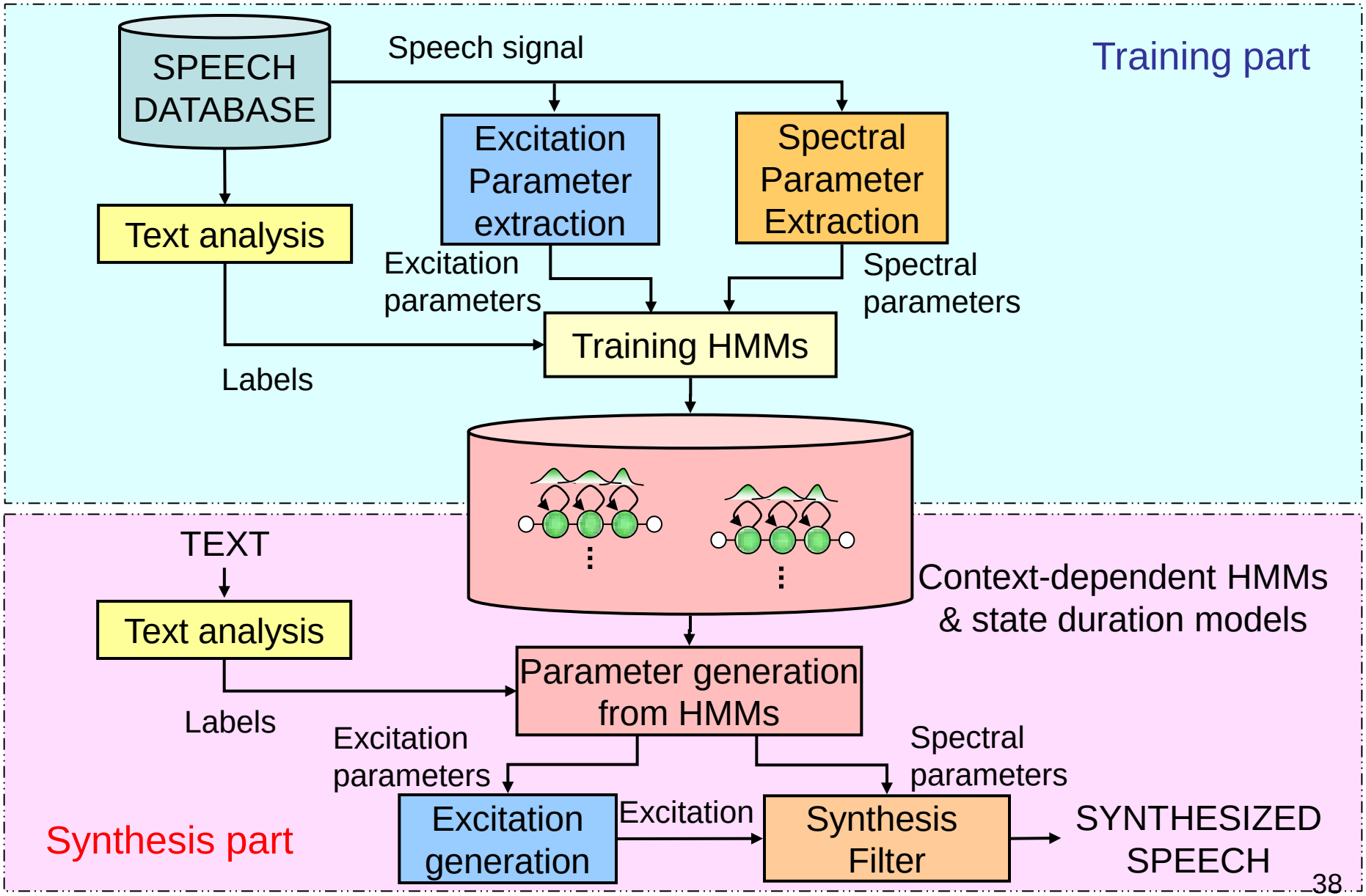
Problem of statistical parametric speech synthesis

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \boldsymbol{w}, \boldsymbol{O}, \boldsymbol{W})$$

$$= \arg\max_{\boldsymbol{o}} \sum_{\boldsymbol{l}} \sum_{\boldsymbol{L}} \int \int \underbrace{P(\boldsymbol{o} \mid \boldsymbol{l}, \boldsymbol{\lambda})}_{\text{Speech parameter generation}} \underbrace{P(\boldsymbol{l} \mid \boldsymbol{w}, \boldsymbol{\Lambda})}_{\text{Text processing}}$$

$$\times \underbrace{P(\boldsymbol{\lambda} \mid \boldsymbol{O}, \boldsymbol{L})}_{\text{HMM parameter estimation}} \underbrace{P(\boldsymbol{L} \mid \boldsymbol{W}, \boldsymbol{\Lambda})}_{\text{Text processing}} \underbrace{P(\boldsymbol{\Lambda})}_{\text{Prior}} d\boldsymbol{\lambda} d\boldsymbol{\Lambda}$$

HMM-based speech synthesis system



Inclusion of all components

Problem of statistical parametric speech synthesis

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

$$= \arg\max_{\boldsymbol{x}} \sum_{\boldsymbol{c},\boldsymbol{C}} \sum_{\boldsymbol{l},\boldsymbol{L}} \sum_{m} \int\int \underbrace{P\left(\boldsymbol{x} \mid \boldsymbol{c}\right)}_{\text{Waveform generation}} \underbrace{P\left(\boldsymbol{c} \mid \boldsymbol{l}, \boldsymbol{\lambda}, m\right)}_{\text{Parameter generation}} \underbrace{P\left(\boldsymbol{l} \mid \boldsymbol{w}, \boldsymbol{\Lambda}\right)}_{\text{Text processing}}$$

$$\times \underbrace{P\left(\boldsymbol{\lambda} \mid m, \boldsymbol{C}, \boldsymbol{L}\right)}_{\text{Posterior of model parameter}} \underbrace{P\left(m \mid \boldsymbol{C}, \boldsymbol{L}\right)}_{\text{Posterior of model structure}}$$

$$\times \underbrace{P\left(\boldsymbol{L} \mid \boldsymbol{W}, \boldsymbol{\Lambda}\right)}_{\text{Text processing}} \underbrace{P\left(\boldsymbol{C} \mid \boldsymbol{X}\right)}_{\text{Speech analysis}} \underbrace{P\left(\boldsymbol{\Lambda}\right)}_{\text{Prior}} d\boldsymbol{\lambda} d\boldsymbol{\Lambda}$$

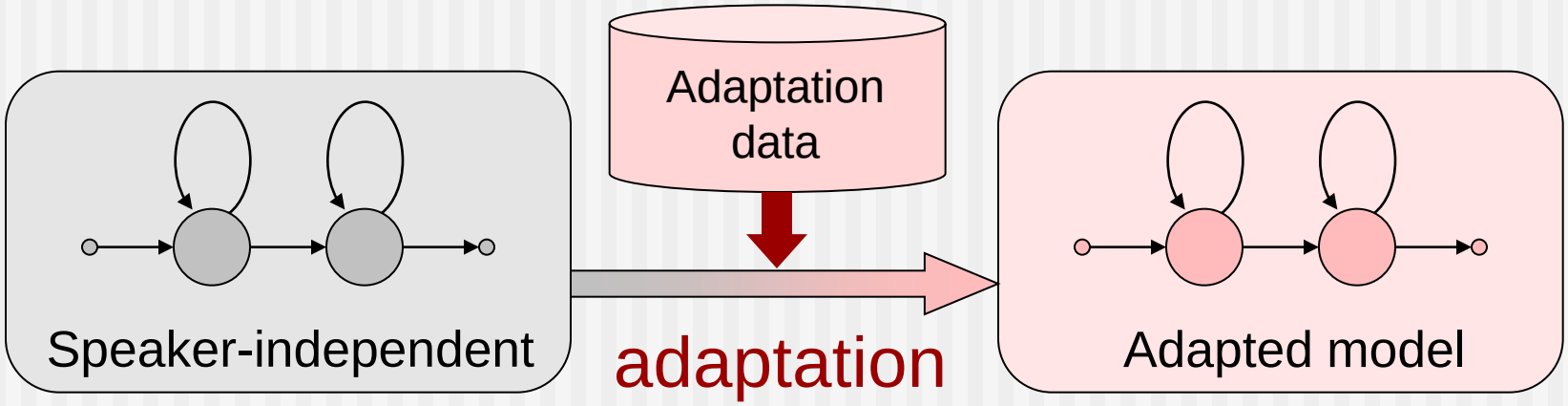
Emotional Speech Synthesis

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

trained with 200 utterances

Speaker Adaptation (mimicking voices)

MLLR-based adaptation

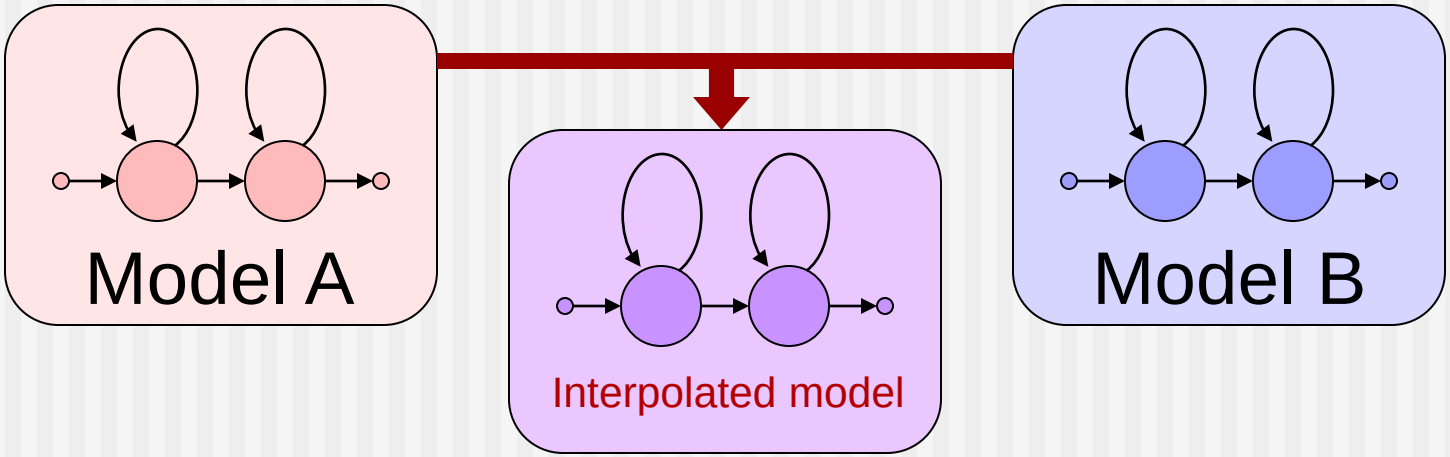


- w/o adaptation (initial model)
- Adapted with 4 utterances
- Adapted with 50 utterances
- Speaker-dependent model

?

Speaker Interpolation (mixing voices)

Linear combination of two speaker-dependent models



A:	1.00	0.75	0.50	0.25	0.00
B:	0.00	0.25	0.50	0.75	1.00

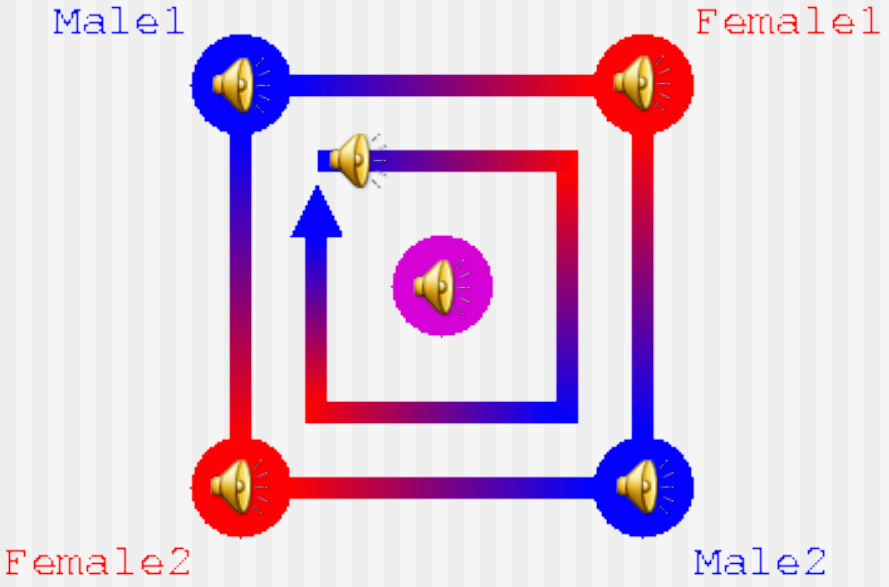
Voice Morphing

Two voices:

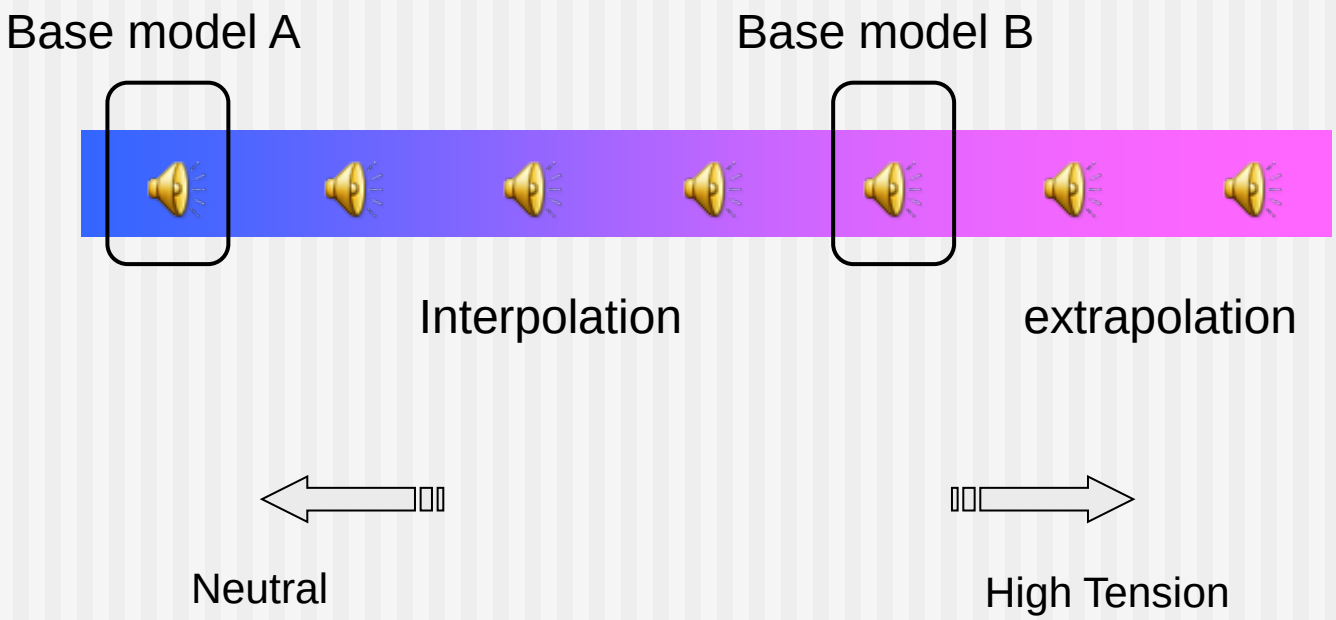
A ⇨⇨⇨⇨⇨⇨⇨⇨⇨ B

A ⇦⇦⇦⇦⇦⇦⇦⇦⇦ B

Four voices:



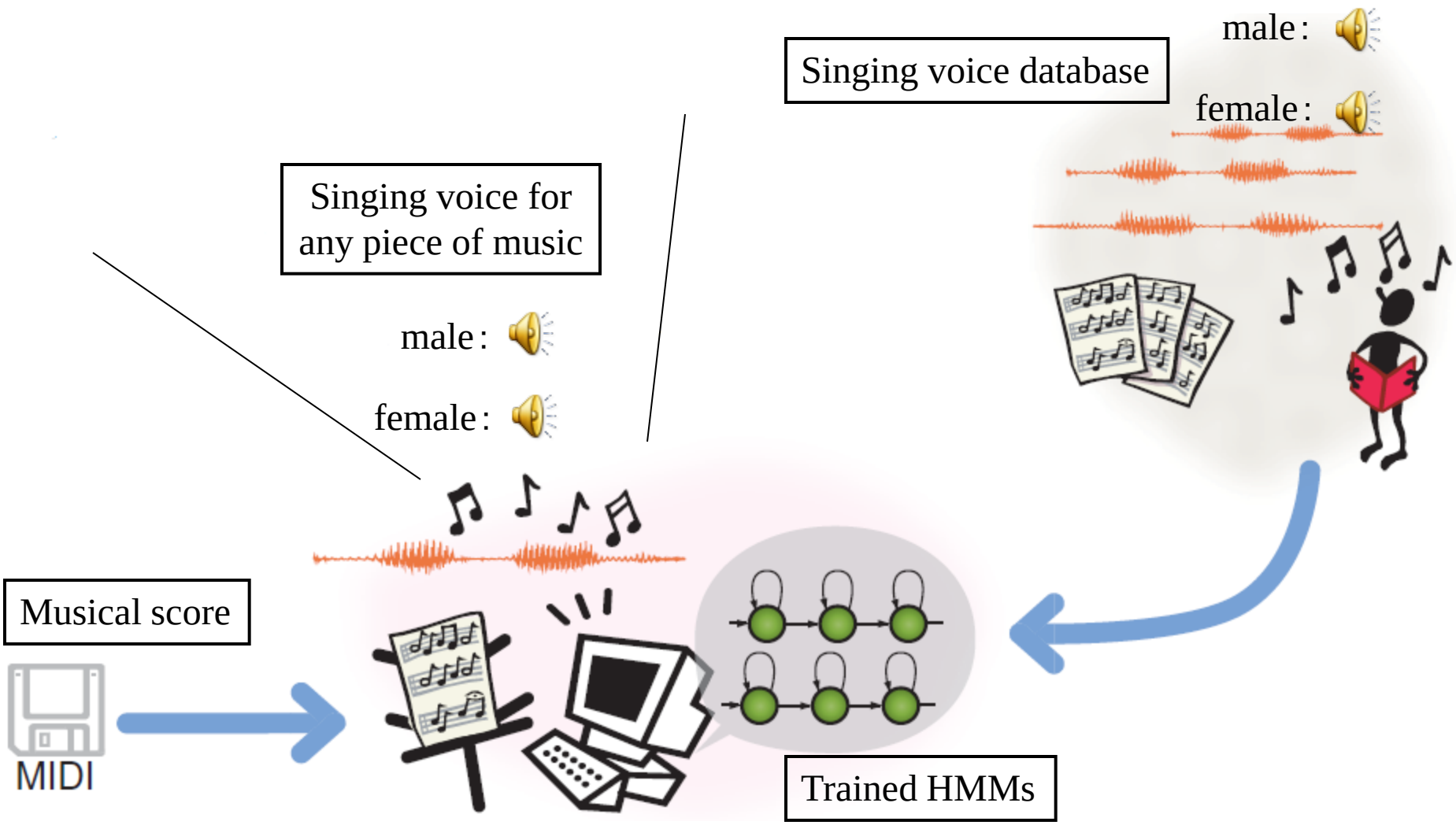
Interpolation of Speaking Styles



Multilingual Speech Synthesis

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

Singing Voice Synthesis



Conclusion

Statistical approach to speech synthesis

- Whole speech synthesis process is described as a statistical framework
- It gives a unified view and reveals what is correct and what is wrong
- Importance of the database

Future work

- Still we have many problems should be solved:
  - Speech waveform modeling
  - Combination with text processing part