Speech Synthesis as A Statistical Machine Learning Problem

Keiichi Tokuda
Nagoya Institute of Technology

ASRU2011, Hawaii
December 14, 2011

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

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?

W : texts

X : speech waveforms

database

Given

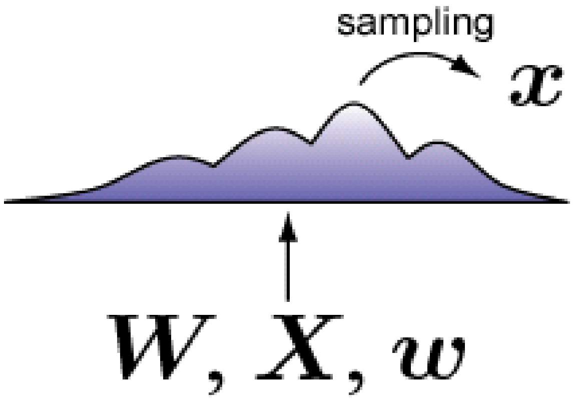
w : text to be synthesized

x : speech waveform $\longleftarrow$?

Statistical formulation of speech synthesis (1/8)

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

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

Given

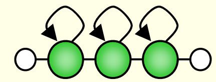
$\boldsymbol{x}$: speech waveform ← unknown

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

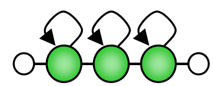
Statistical formulation of speech synthesis (2/8)

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 

$$= \int p(\boldsymbol{x}, \lambda \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{X}) d\lambda = \int \underbrace{p(\boldsymbol{x} \mid \boldsymbol{w}, \lambda)}_{\text{generation}} \underbrace{p(\lambda \mid \boldsymbol{W}, \boldsymbol{X})}_{\text{training}} d\lambda$$

λ : acoustic model (e.g. HMM )

Statistical formulation of speech synthesis (3/8)

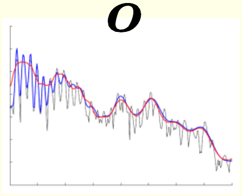
2. Using speech waveform directly is difficult ☹
→ Introduce parametric its representation

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

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

$\boldsymbol{x}$

$\boldsymbol{o}$



$\Downarrow$ 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)

Statistical formulation of speech synthesis (4/8)

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

→ Introduce labels

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

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

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

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

$\boldsymbol{l}$: labels derived from text $\boldsymbol{w}$
(e.g. prons, POS, lexical stress, grammar, pause)

Statistical formulation of speech synthesis (5/8)

4. Difficult to perform integral & sum over auxiliary variables ☹

→ Approximated by joint max

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

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

⇓ approximate integral & sum by joint max

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

where

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

Statistical formulation of speech synthesis (6/8)

5. Joint maximization is hard ☹

→ Approximated by step-by-step maximizations

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

⇓ approx joint max by step-by-step max

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

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

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

Statistical formulation of speech synthesis (7/8)

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

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

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

$$\Downarrow$$

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

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

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

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

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

Statistical formulation of speech synthesis (8/8)

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

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

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

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

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

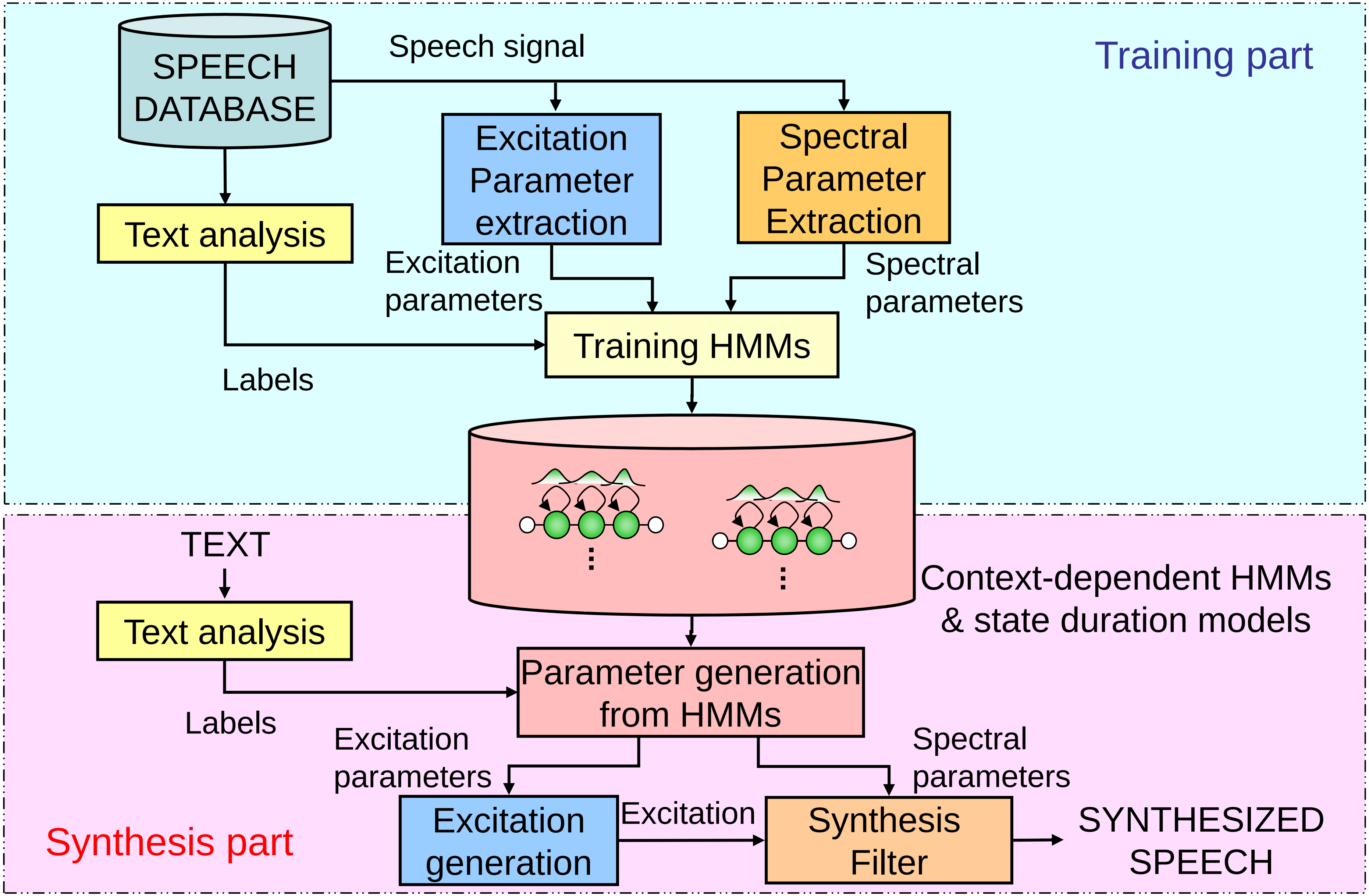
$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$ $\Leftarrow$ speech parameter generation

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

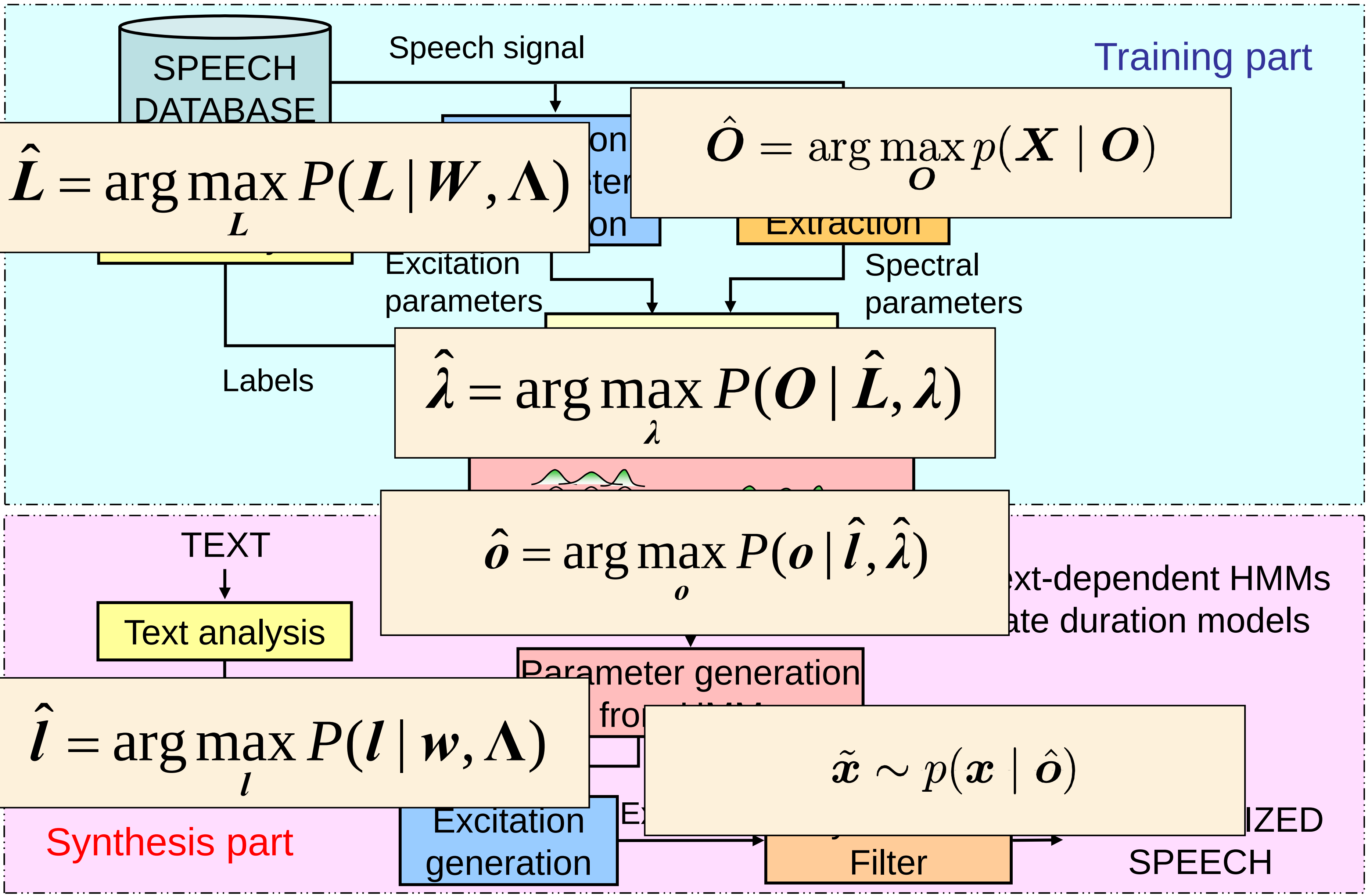
Overview of this talk

1. Mathematical formulation
2. Implementation of individual components
3. Examples demonstrating its flexibility
4. Discussion and conclusion

HMM-based speech synthesis system



HMM-based speech synthesis system



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

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

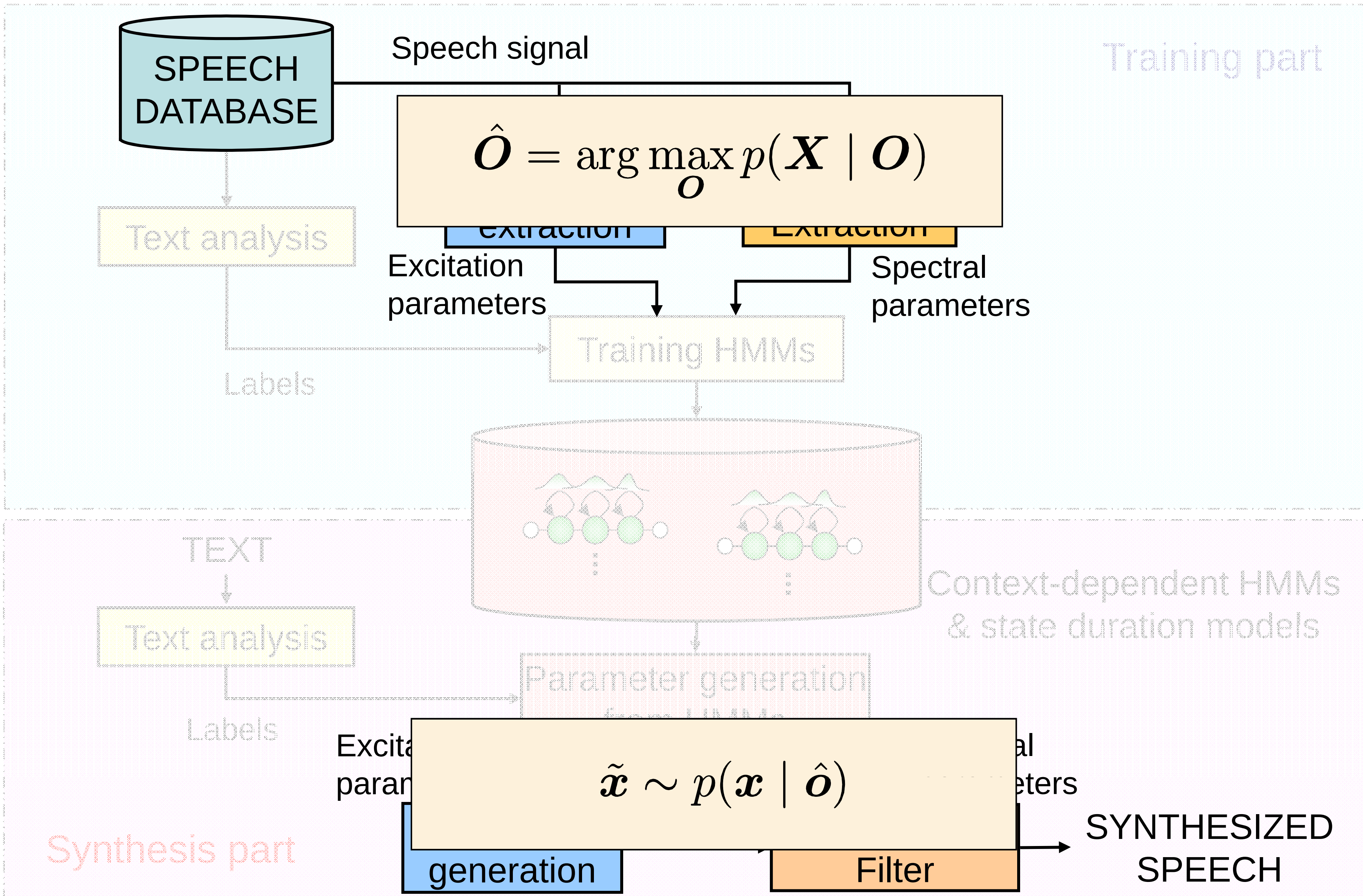
$$\hat{\lambda} = \arg\max_{\lambda} P(\boldsymbol{O} \mid \hat{\boldsymbol{L}}, \lambda)$$

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

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

$$\tilde{\boldsymbol{x}} \sim p(\boldsymbol{x} \mid \hat{\boldsymbol{o}})$$

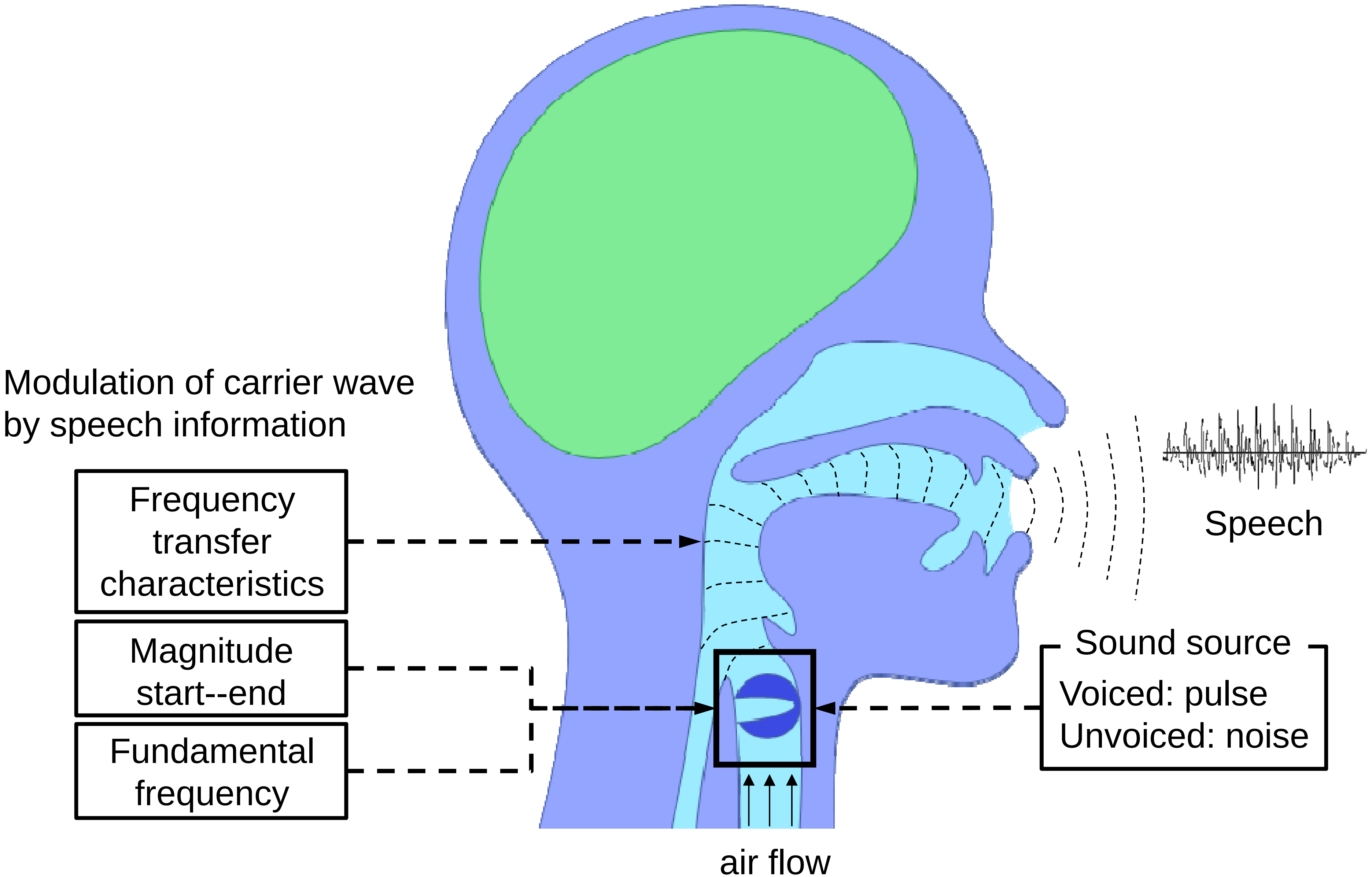
HMM-based speech synthesis system



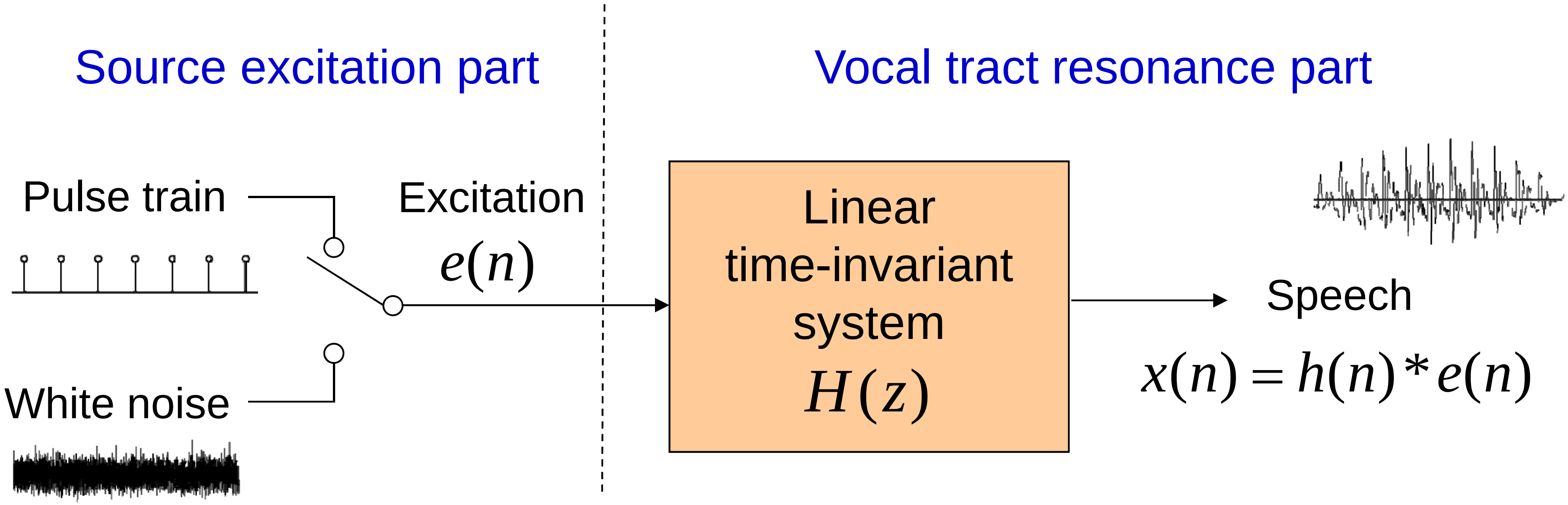
$$\hat{O} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{X} \mid \boldsymbol{O})$$

$$\tilde{\boldsymbol{x}} \sim p(\boldsymbol{x} \mid \hat{\boldsymbol{o}})$$

Human speech production



Source-filter model

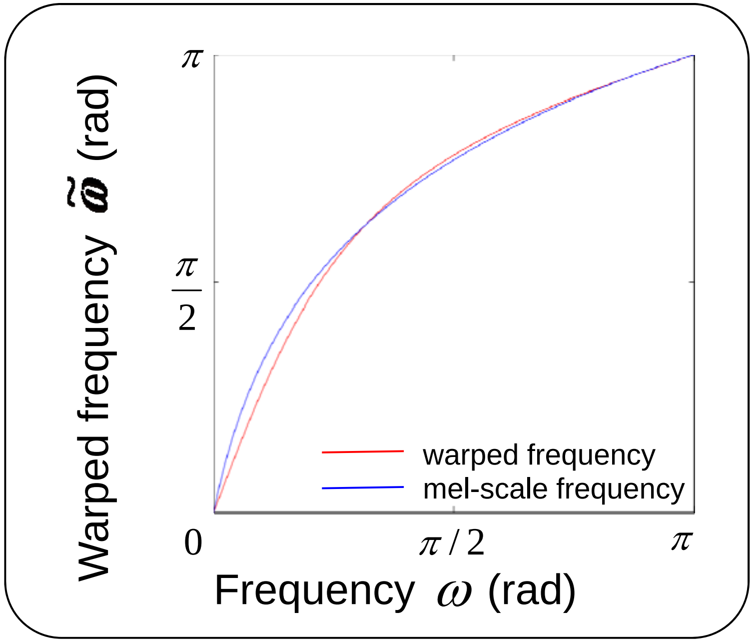


ML estimation of spectral parameter

Mel-cepstral representation of speech spectra

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

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



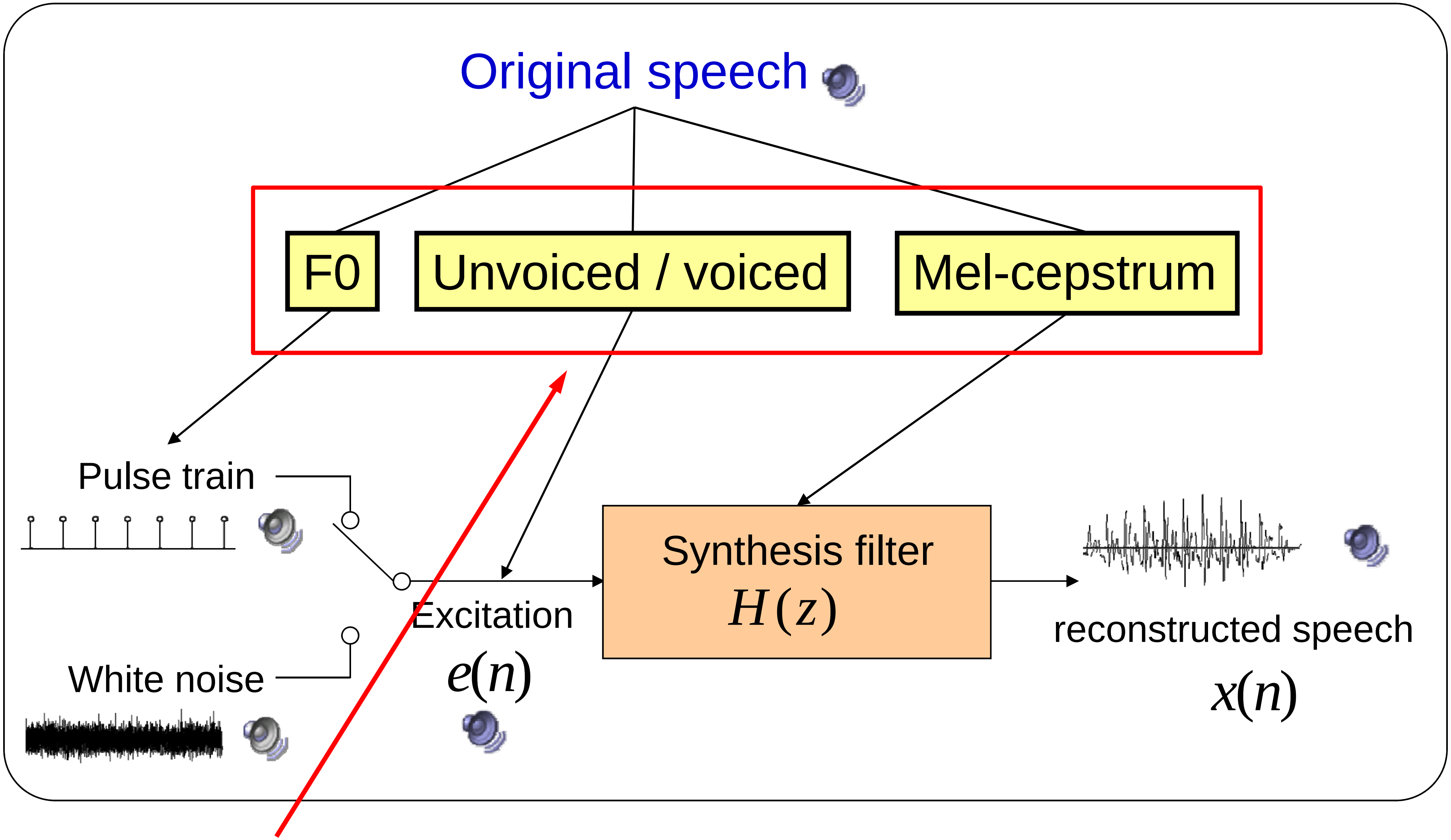
ML-estimation of mel-cepstrum

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

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

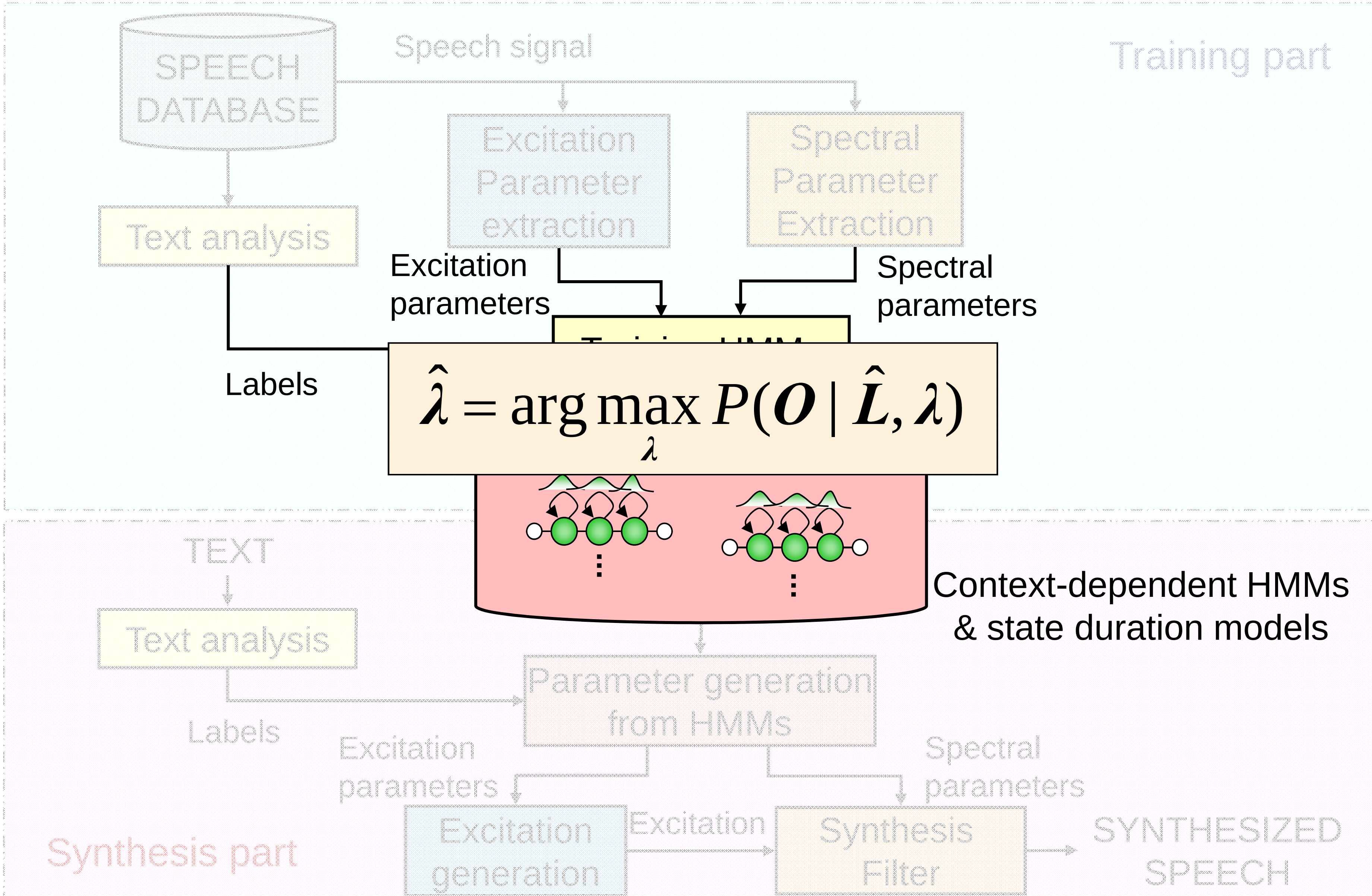
$\boldsymbol{c}$: mel-cepstrum

Waveform reconstruction

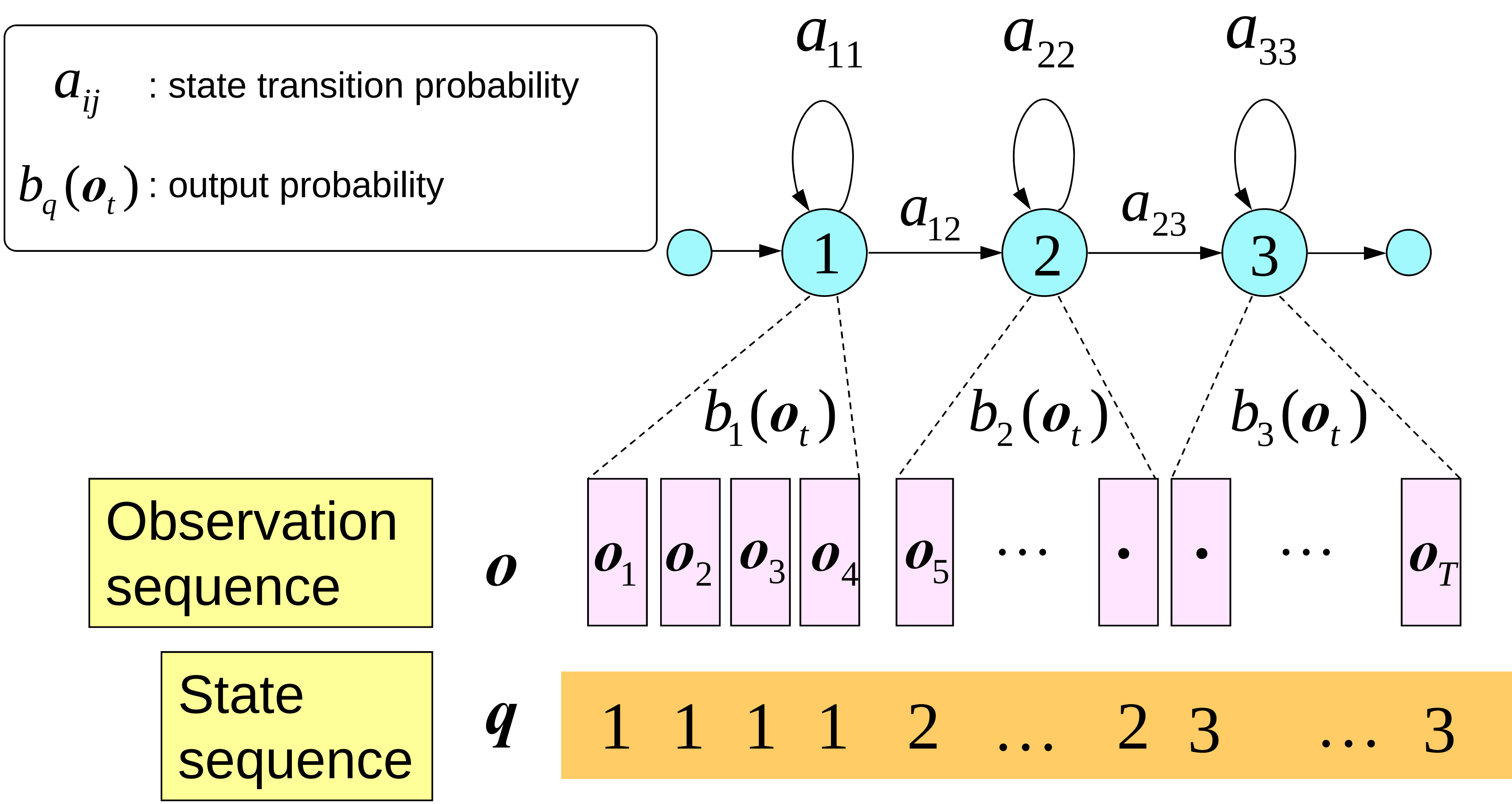


These speech parameters should be modeled by HMM

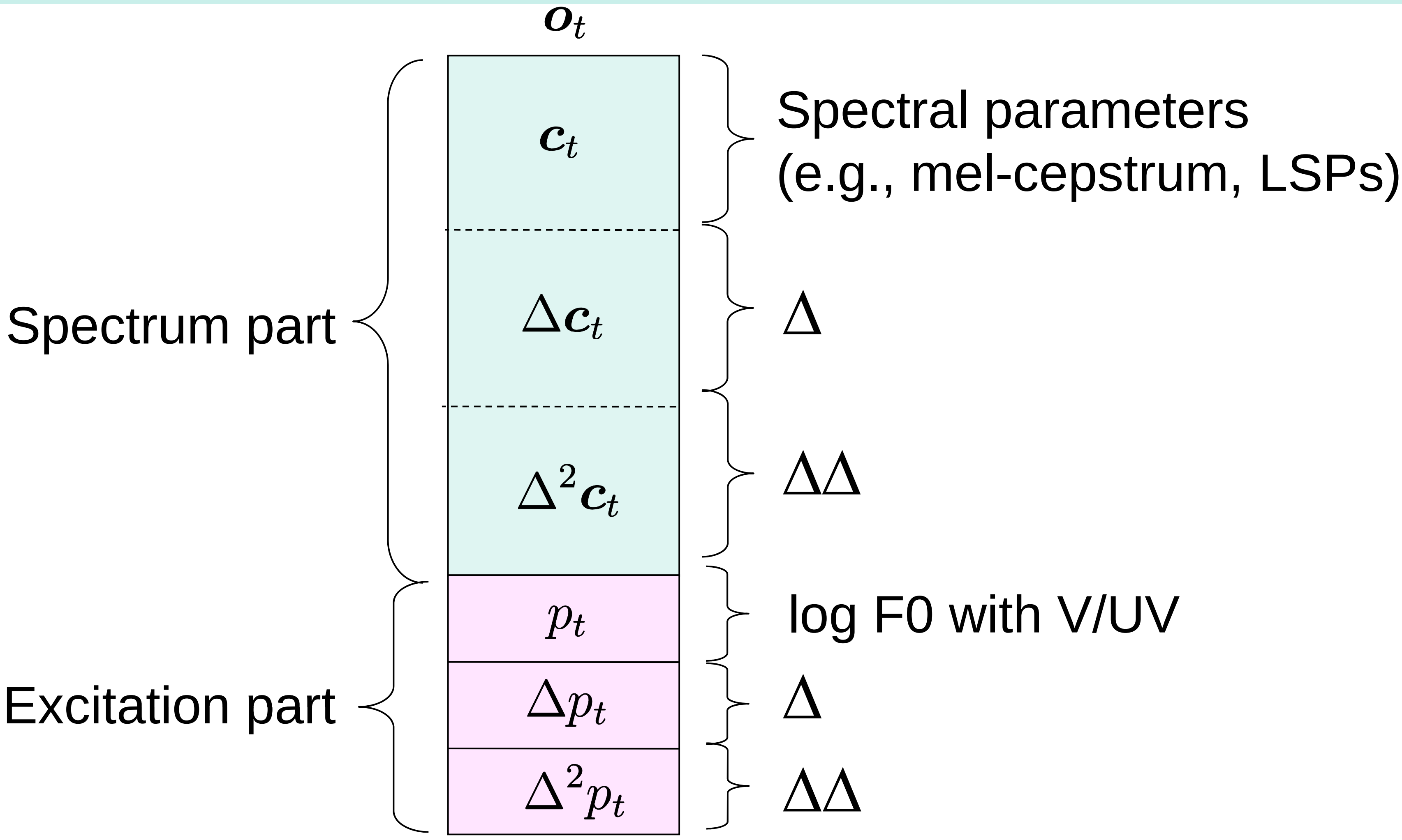
HMM-based speech synthesis system



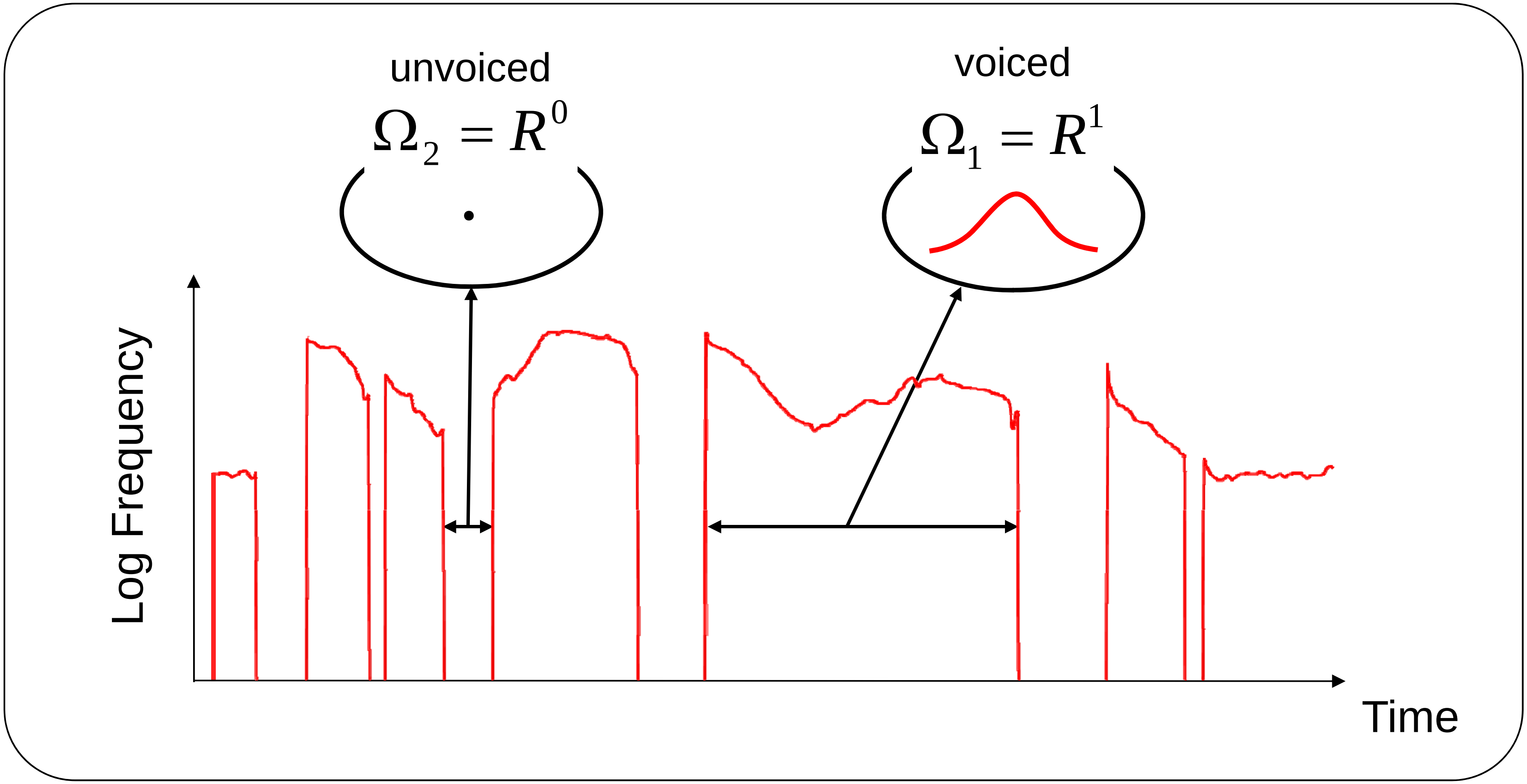
Hidden Markov model (HMM)



Structure of state output (observation) vector

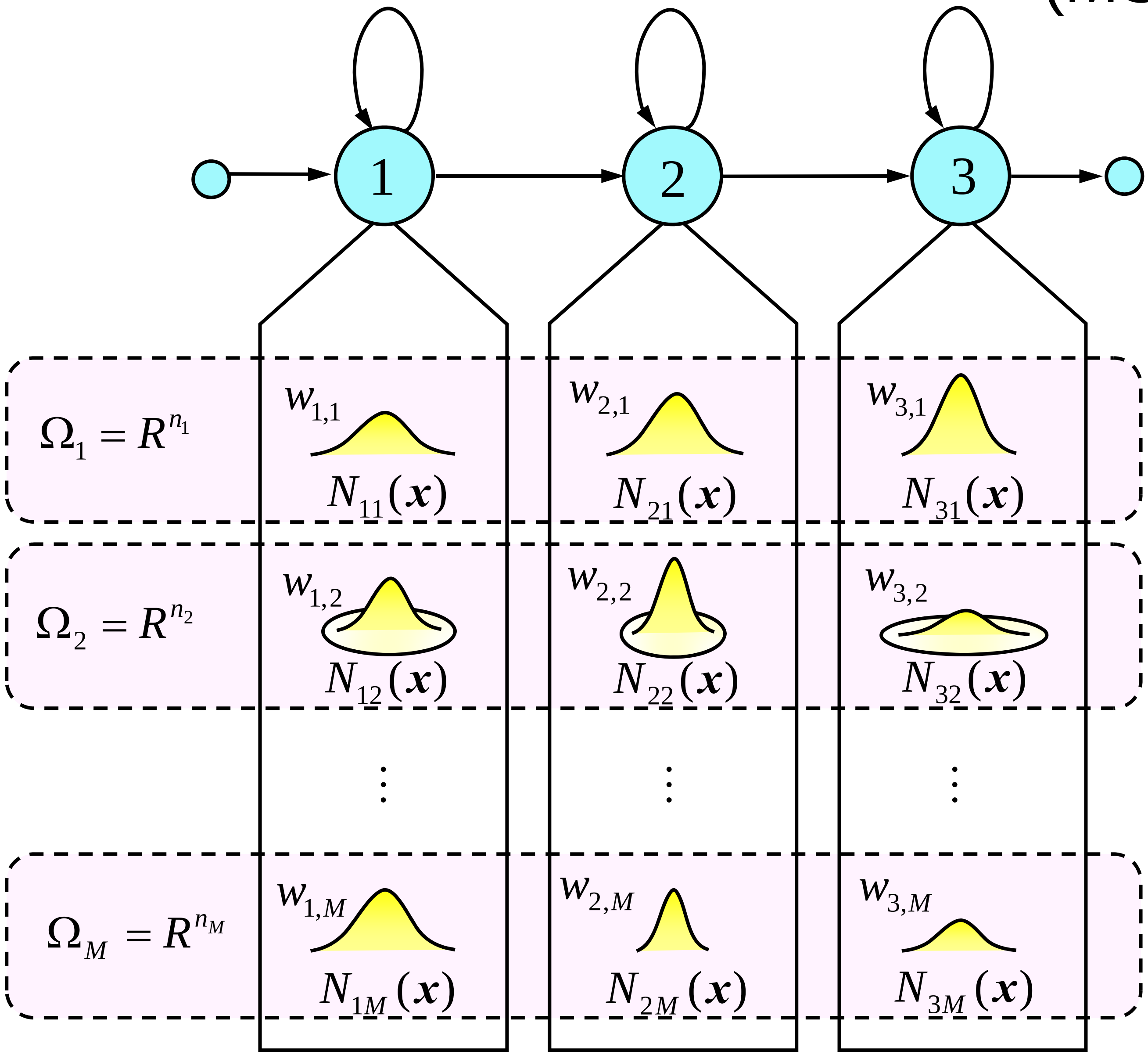


Observation of F0

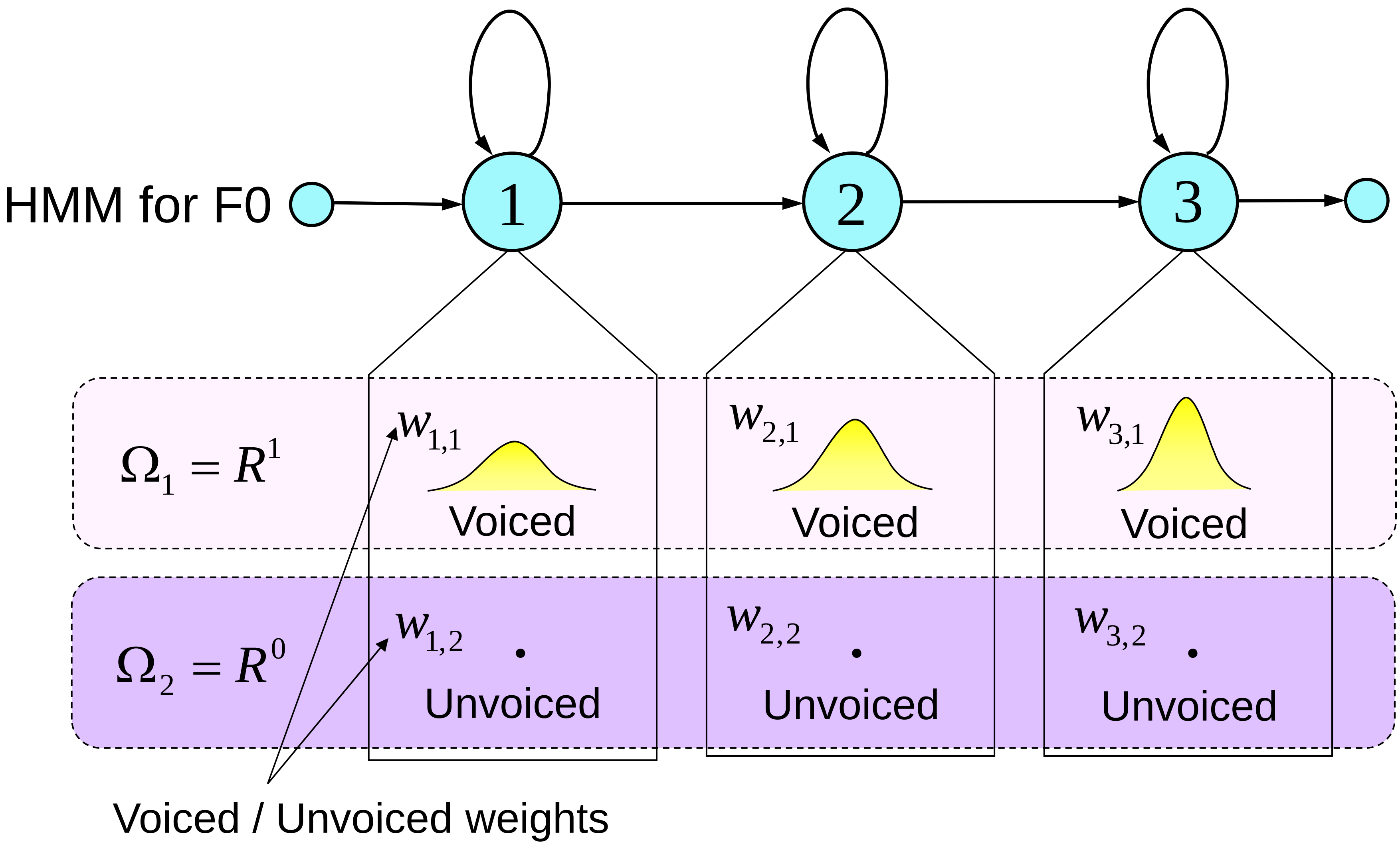


Unable to model by continuous or discrete distributions
⇒ Multi-space distribution HMM (MSD-HMM)

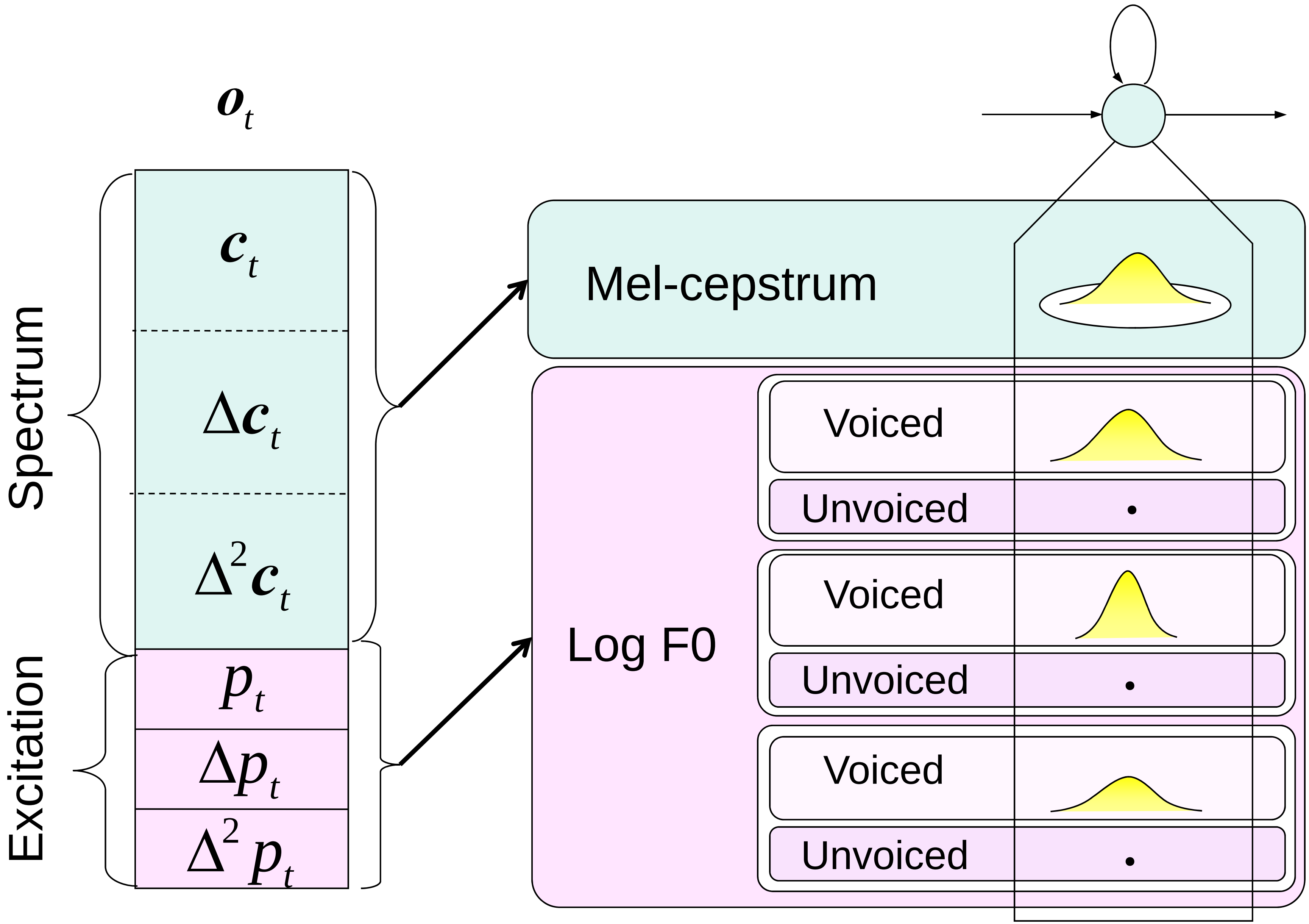
Multi-space probablity distribution HMM (MSD-HMM)



MSD-HMM for F0 modeling



Structure of state-output distributions



Contextual factors

Phoneme

- {preceding, succeeding} two phonemes
- current phoneme

Syllable

- # of phonemes in {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

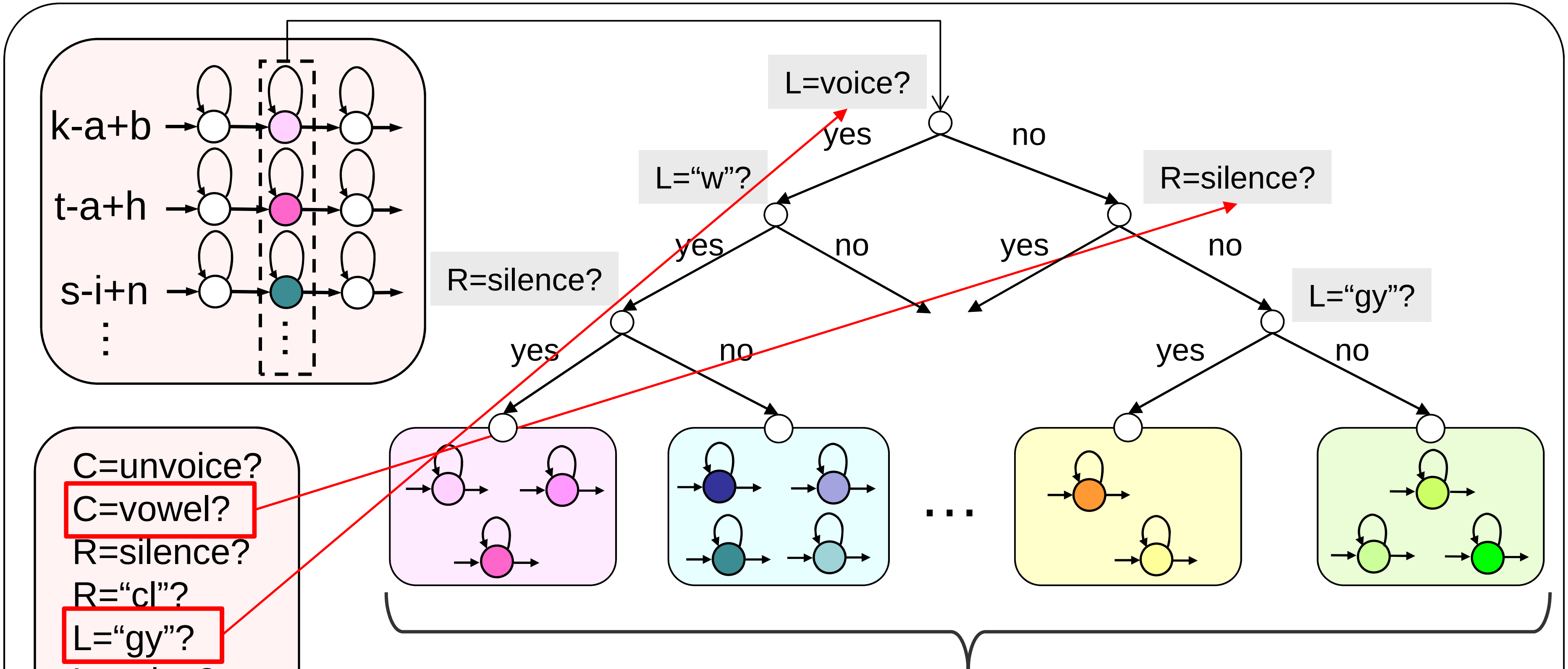
Phrase

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

…..

Huge # of combinations ⇒ Difficult to have all possible models

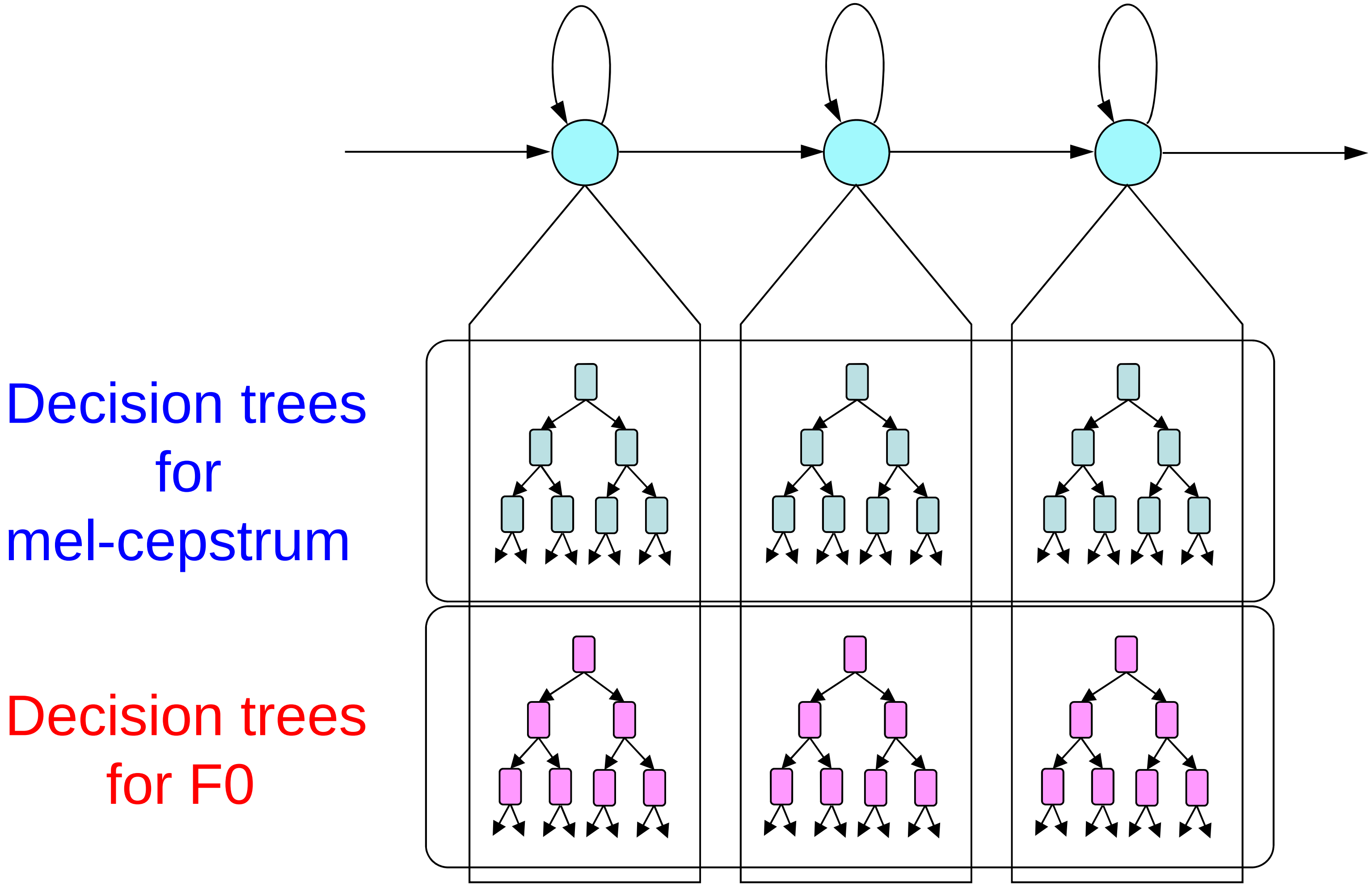
Decision tree-based state clustering [Odell; '95]



Sharing the parameter of HMMs in same leaf node

Stream-dependent tree-based clustering (1)

Spectrum & excitation have different context dependency → Build decision trees separately



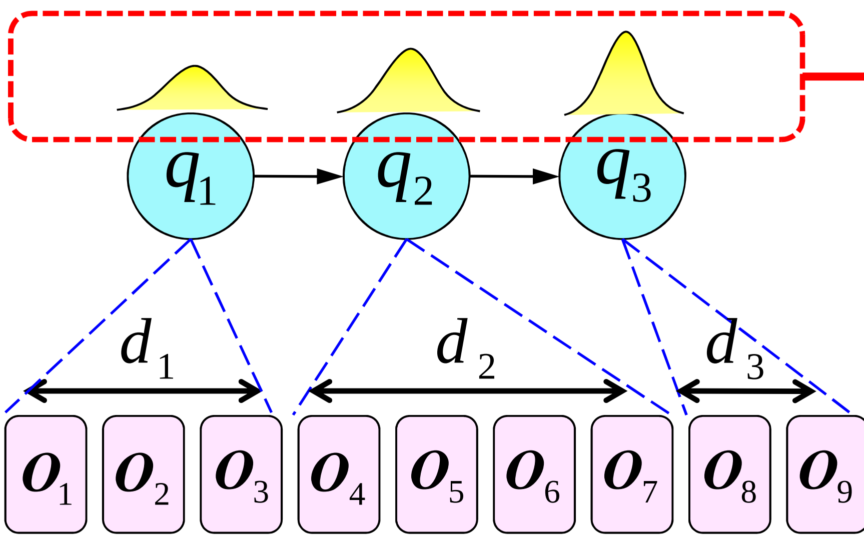
State duration modeling

HMM (Hidden Markov Model)

- State duration prob. depends only on transition prob.
- State duration probability exponentially decreases

HSMM (Hidden Semi Markov Model)

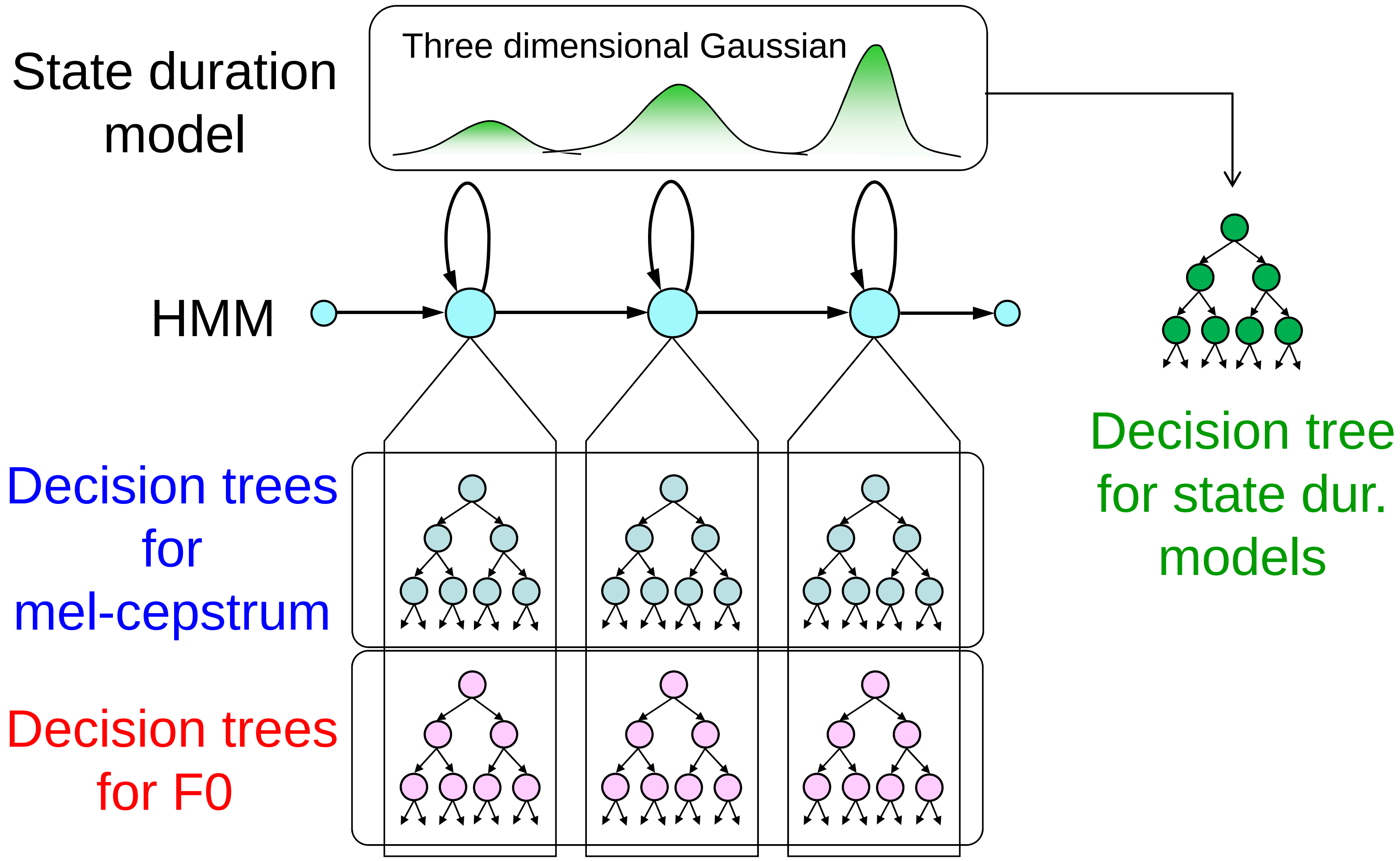
- HMM + explicit duration model ⇒ HSMM



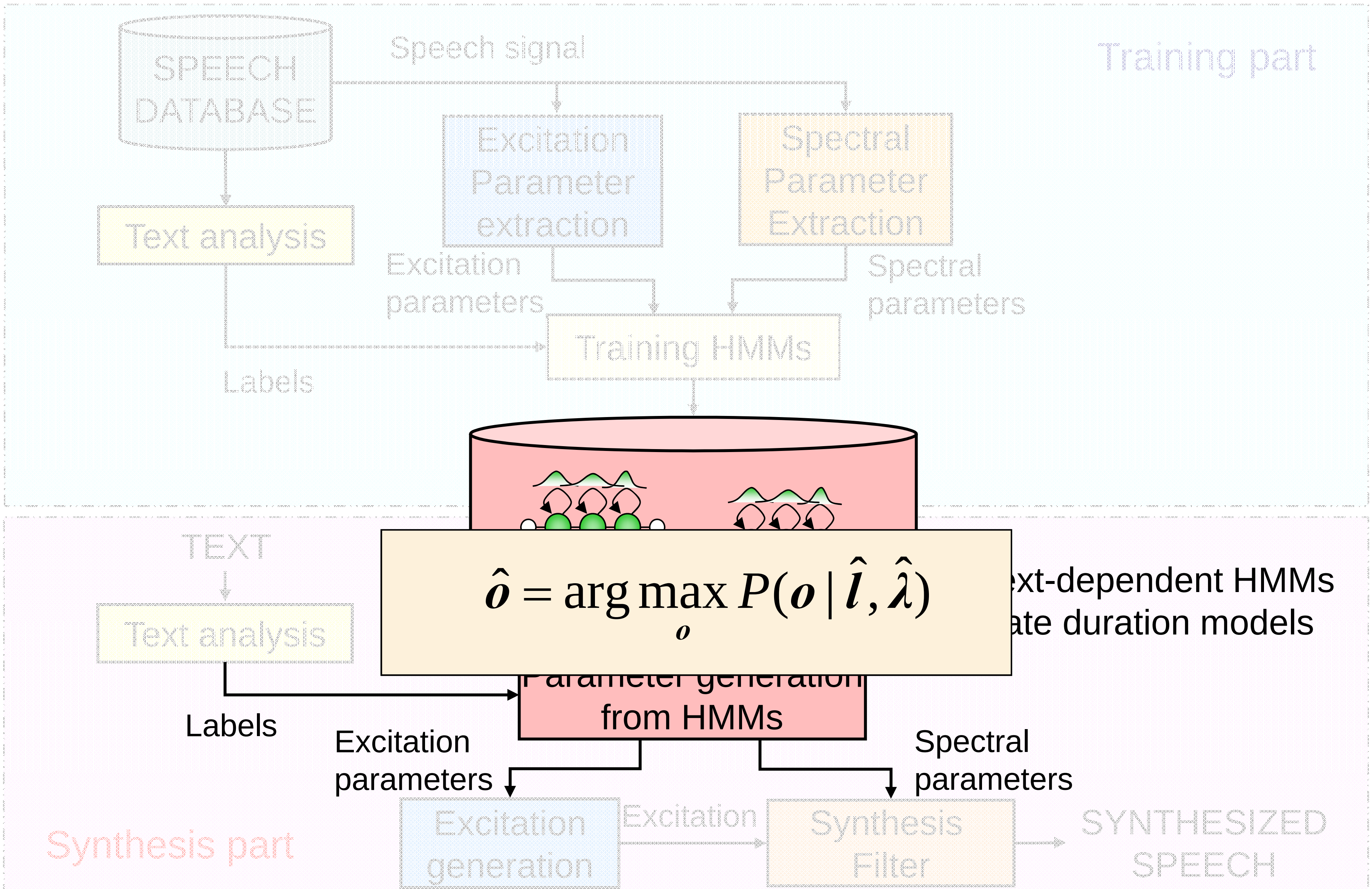
3-demensional Gaussian

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

Stream-dependent tree-based clustering (2)

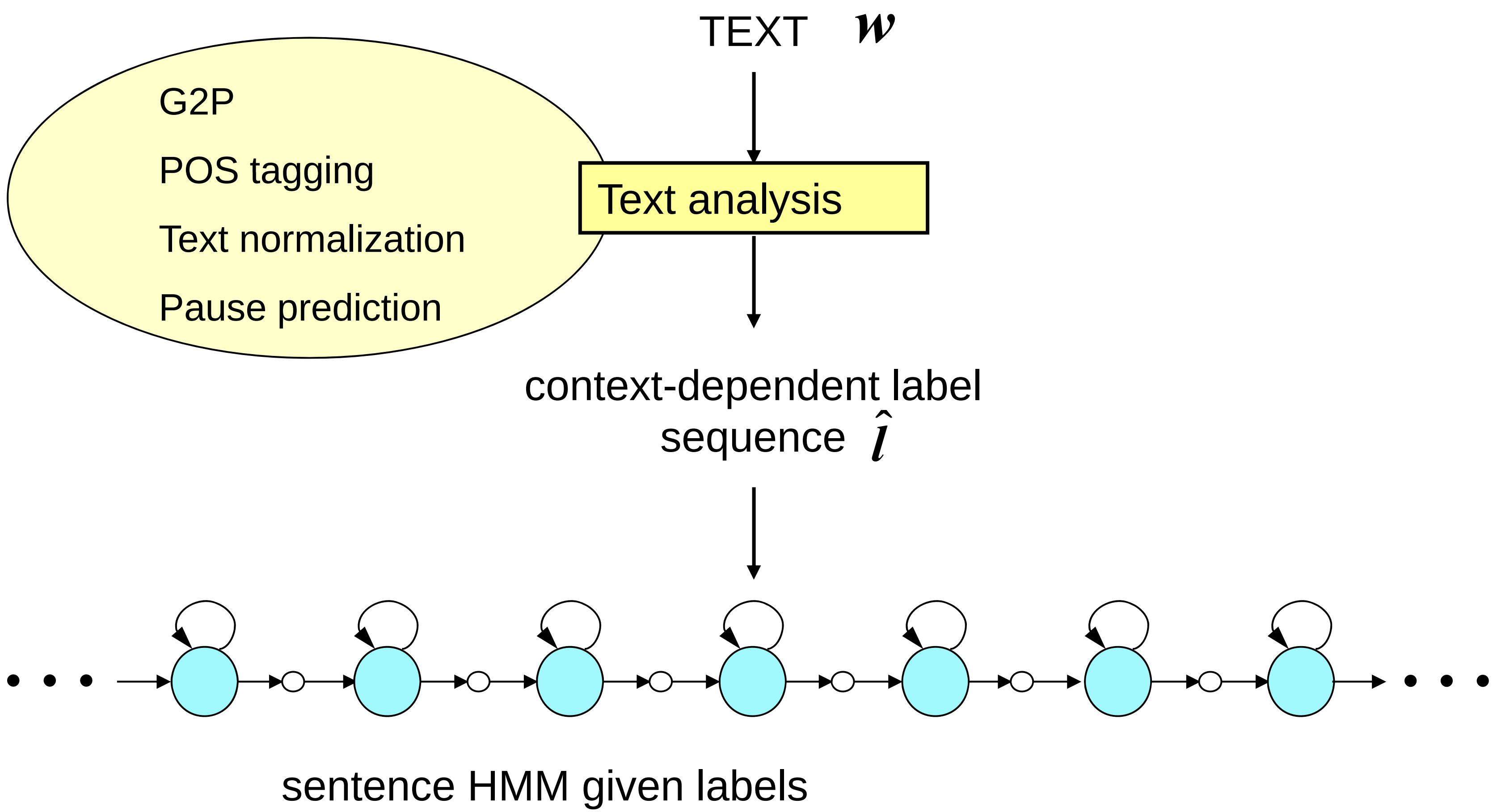


HMM-based speech synthesis system



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

Composition of sentence HMM for given text



This sentence HMM gives $p(o \mid \hat{l}, \hat{\lambda})$

Speech parameter generation algorithm [Tokuda; '00]

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

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

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

$$\Downarrow$$

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

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

Determination of state sequence

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

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

d_i : state duration of i-th state

K : # of states in a sentence HMM for $\hat{\boldsymbol{l}}$

Gaussian

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

Speech parameter generation algorithm

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

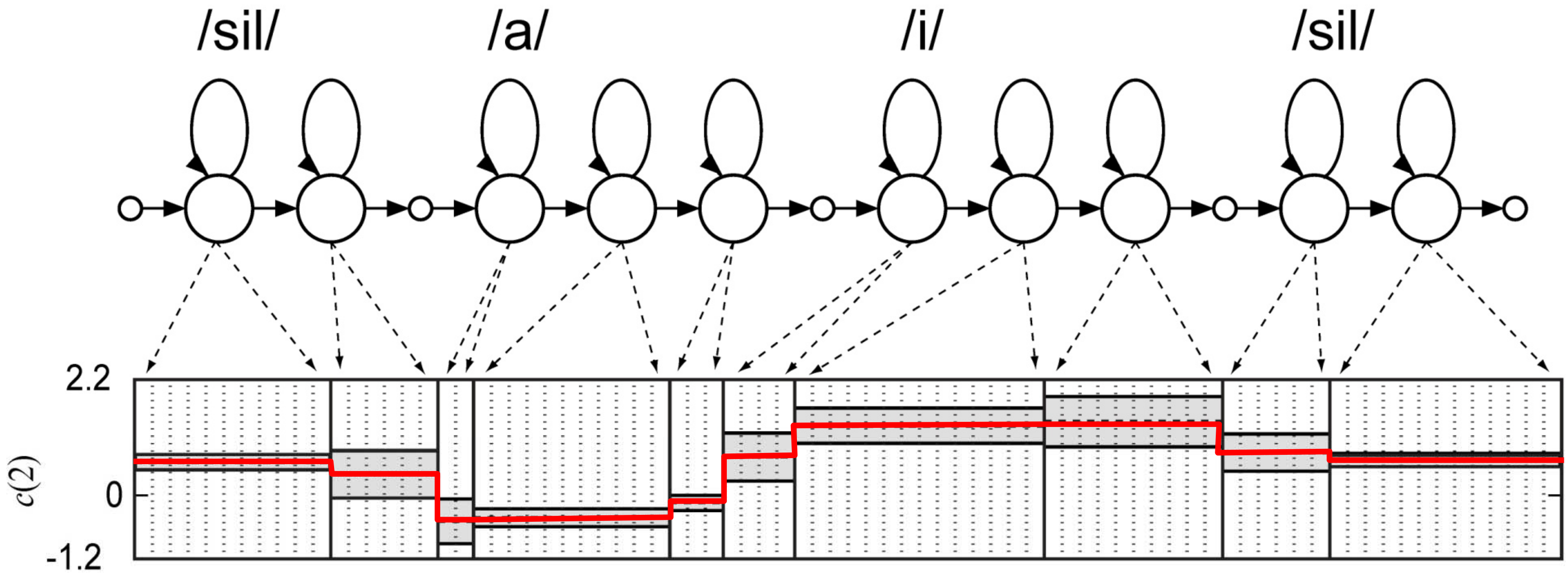
$$
\begin{aligned}
P(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda}) &= \sum_{\boldsymbol{q}} P(\boldsymbol{o} \mid \boldsymbol{q}, \hat{\lambda}) P(\boldsymbol{q} \mid \hat{\boldsymbol{l}}, \hat{\lambda}) \\
&\approx \max_{\boldsymbol{q}} P(\boldsymbol{o} \mid \boldsymbol{q}, \hat{\lambda})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

$$\Downarrow$$

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

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

Without dynamic feature

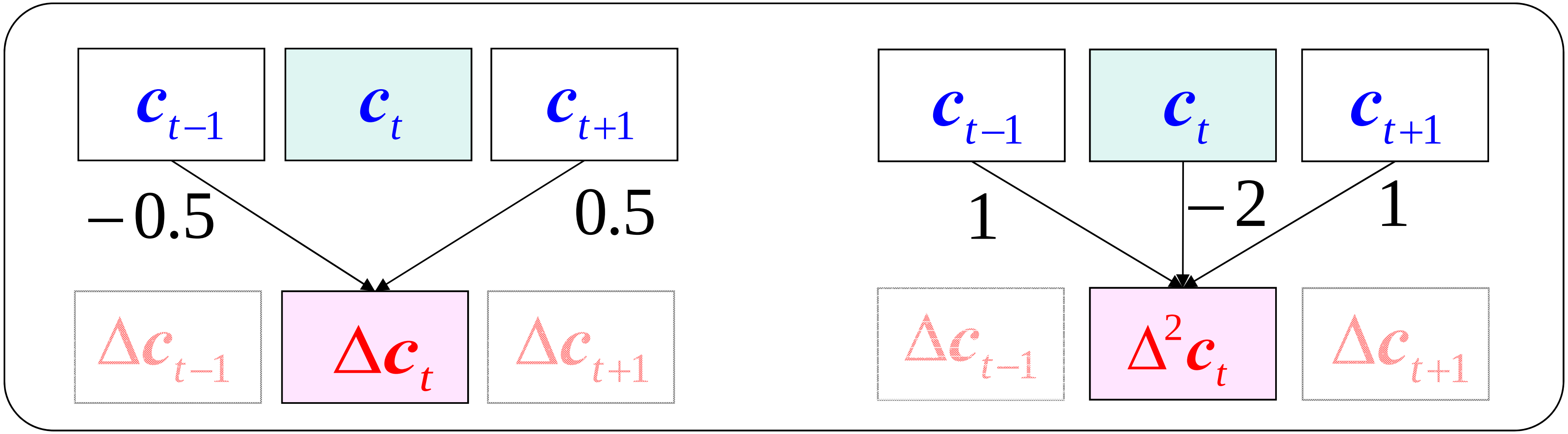


becomes a sequence of mean vectors
⇒ discontinuous outputs between states

Dynamic features

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

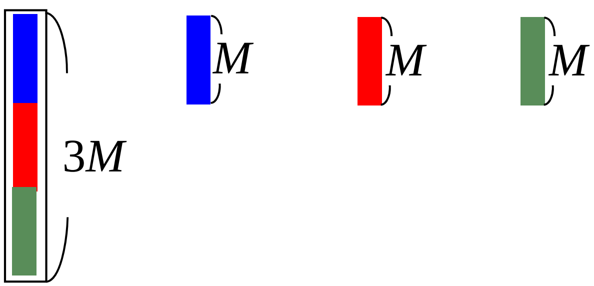
$$\Delta^2 \boldsymbol{c}_t = \frac{\partial^2 \boldsymbol{c}_t}{\partial t^2} \approx \boldsymbol{c}_{t+1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-1}$$

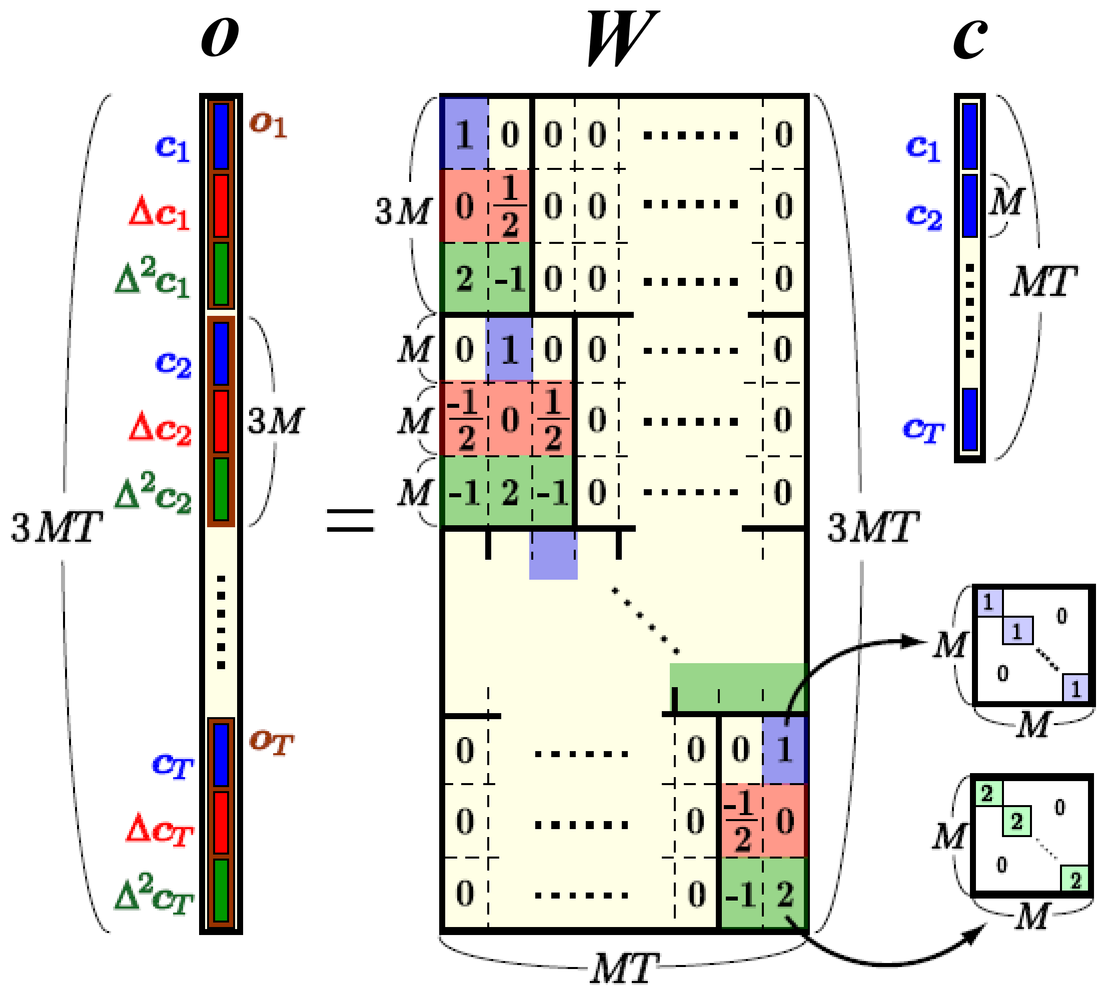


Integration of dynamic features

Relationship between speech parameter vectors & static feature vectors

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^{\mathsf{T}}, \Delta \boldsymbol{c}_t^{\mathsf{T}}, \Delta^2 \boldsymbol{c}_t^{\mathsf{T}}\right]^{\mathsf{T}}$$





Solution for the problem (1/2)

By setting

$$\frac{\partial \log P(\overbrace{\boldsymbol{W}\boldsymbol{c}}^{\boldsymbol{O}} \mid \hat{\boldsymbol{q}}, \lambda)}{\partial \boldsymbol{c}} = \boldsymbol{O},$$

we obtain

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

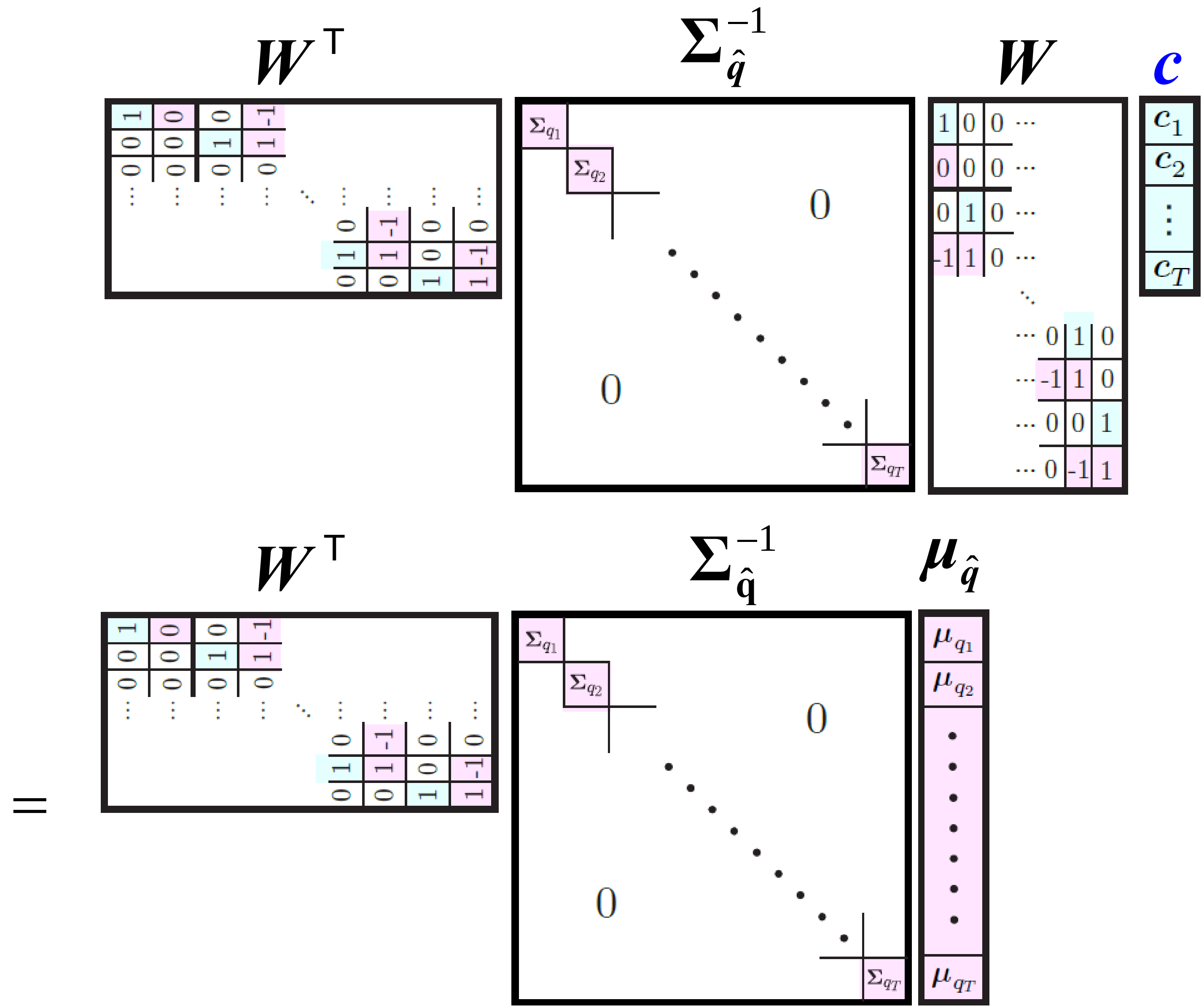
where

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

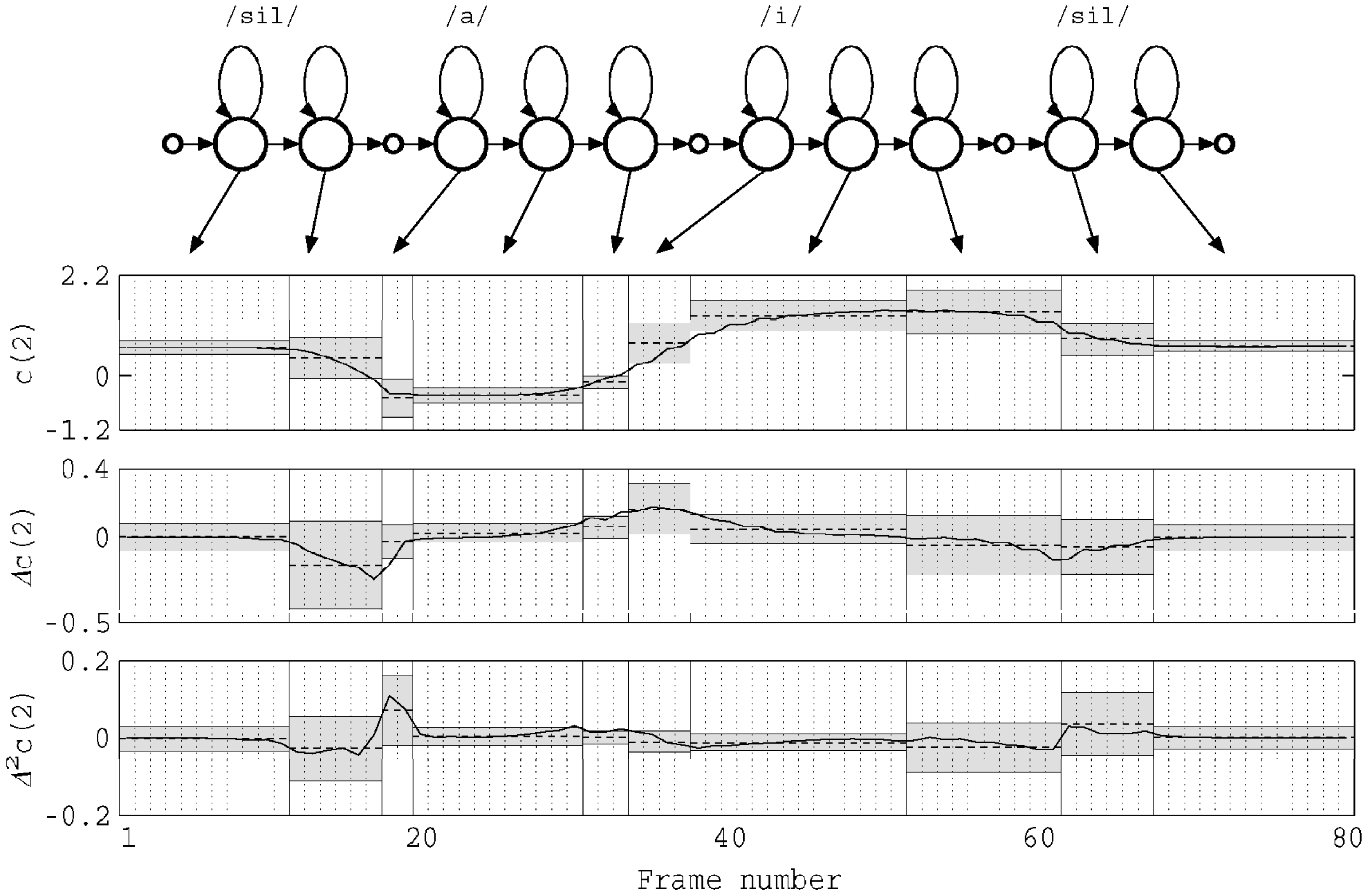
$$\boldsymbol{\mu}_{\hat{\boldsymbol{q}}} = [\boldsymbol{\mu}_{\hat{q}_1}^\top, \boldsymbol{\mu}_{\hat{q}_2}^\top, \ldots, \boldsymbol{\mu}_{\hat{q}_T}^\top]^\top$$

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

Solution for the problem (2/2)



Generated speech parameter trajectory



Trajectory HMM

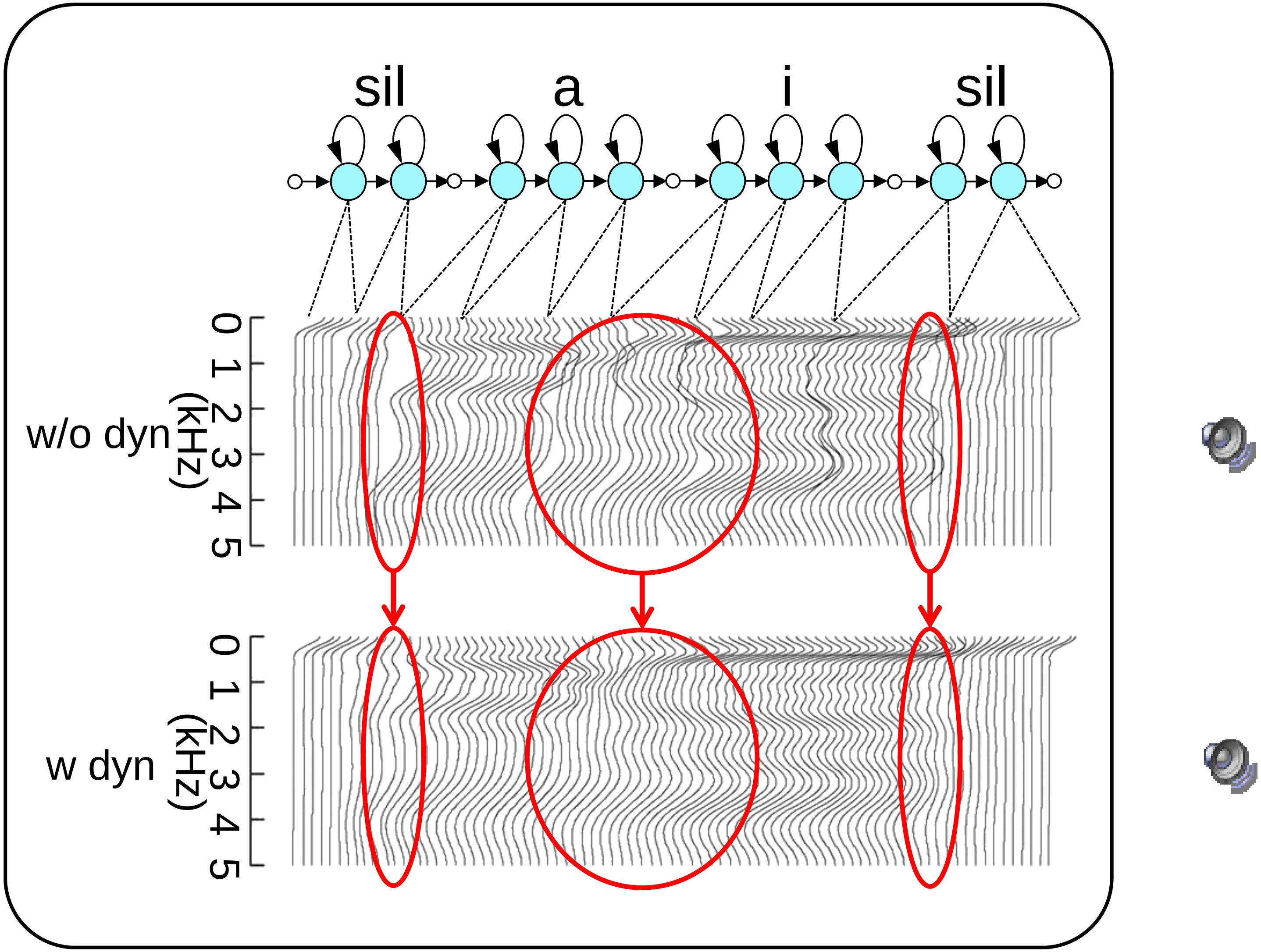
$P(\boldsymbol{o} \mid \boldsymbol{l}, \lambda) = P(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{l}, \lambda)$ is not a proper distribution of $\boldsymbol{c}$

	Conventional HMM	Trajectory HMM
Training	$\arg\max_{\lambda} P(\boldsymbol{O} \mid \hat{\boldsymbol{L}}, \lambda)$	$\arg\max_{\lambda} P(\boldsymbol{C} \mid \hat{\boldsymbol{L}}, \lambda)$
Synthesis	$\arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})\rvert_{\boldsymbol{o}=\boldsymbol{W}\boldsymbol{c}}$ $= \arg\max_{\boldsymbol{c}} P(\boldsymbol{c} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$ ⟺	$\arg\max_{\hat{\lambda}} P(\boldsymbol{c} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$

Solve inconsistency between training & synthesis

$\Rightarrow$ improving the model accuracy

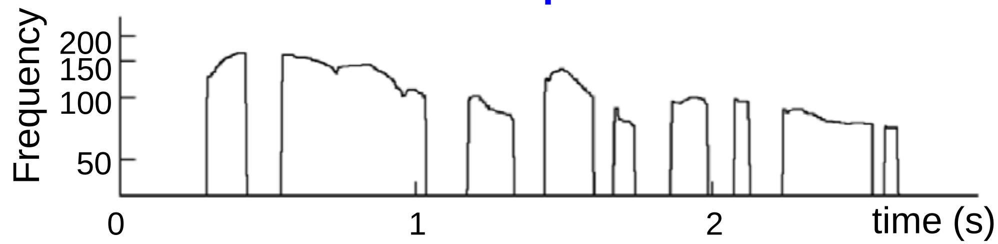
Generated spectra



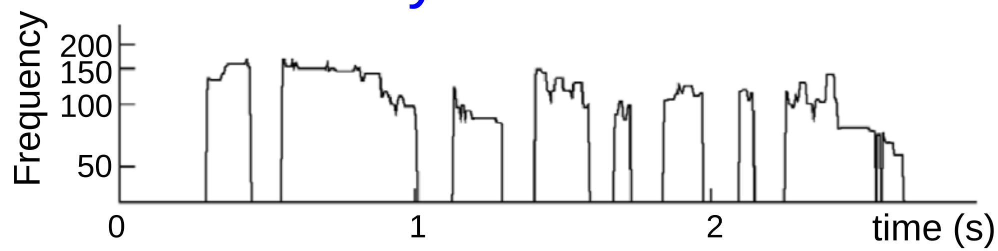
Spectra changing smoothly between phonemes

Generated F0

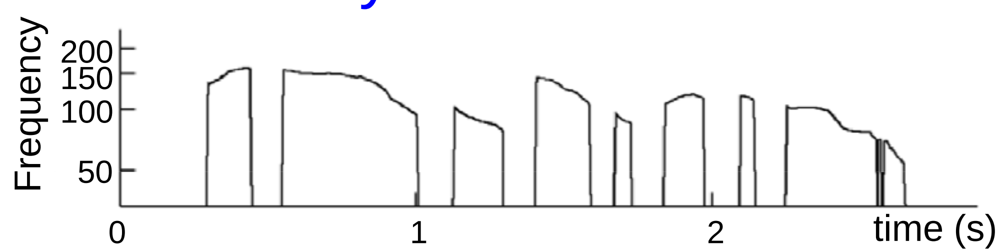
natural speech



without dynamic features



with dynamic features



Effect of dynamic features

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

Overview of this talk

1. Mathematical formulation
2. Implementation of individual components
3. Examples demonstrating its flexibility
4. Discussion and conclusion

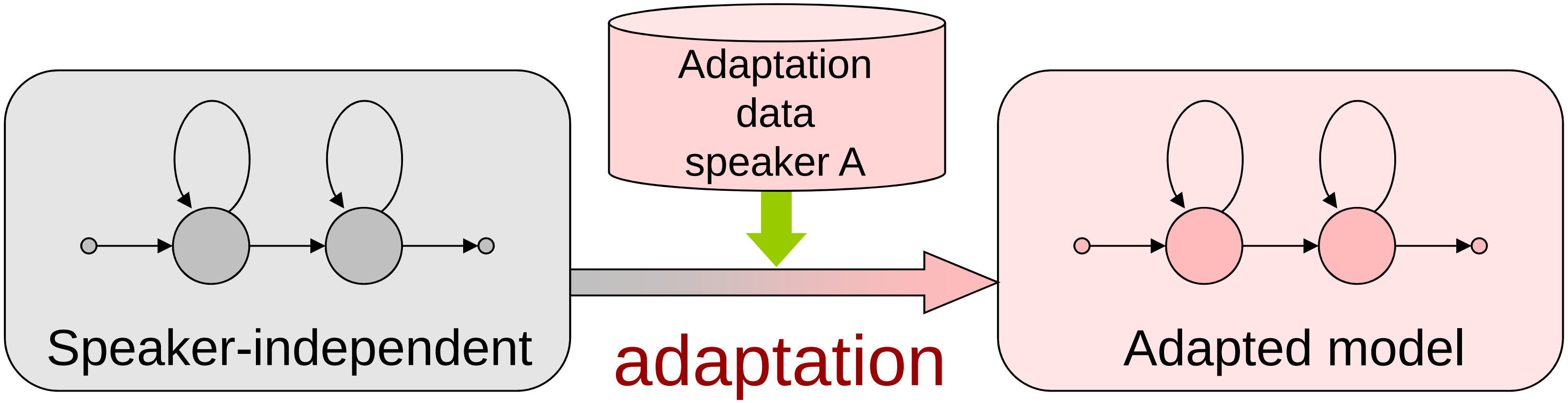
Emotional speech synthesis

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

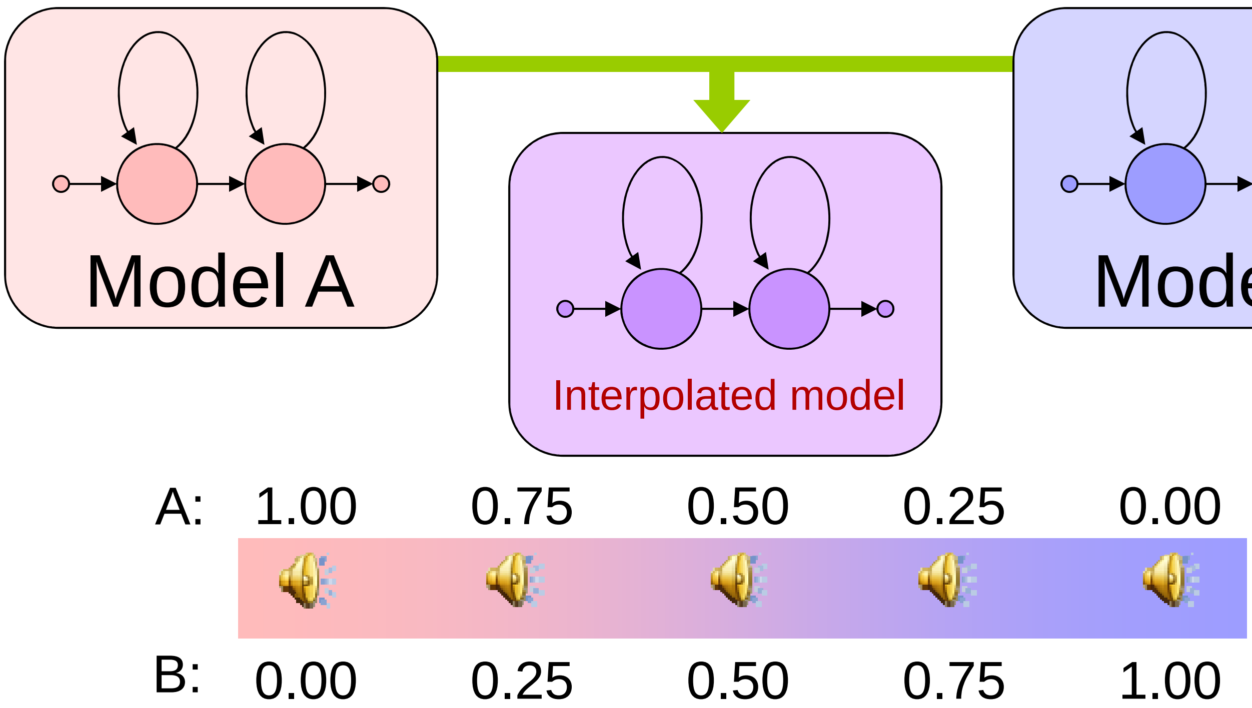
Adapted with 50 utterances

Speaker A’s speaker-dependent system

?

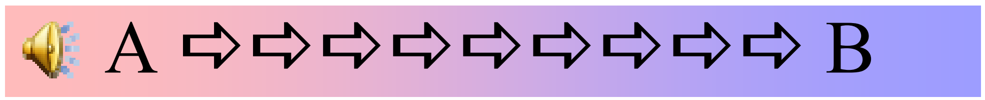
Speaker interpolation (mixing voices)

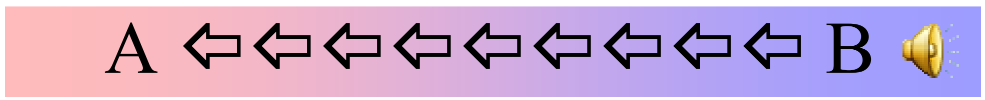
Linear combination of two speaker-dependent models



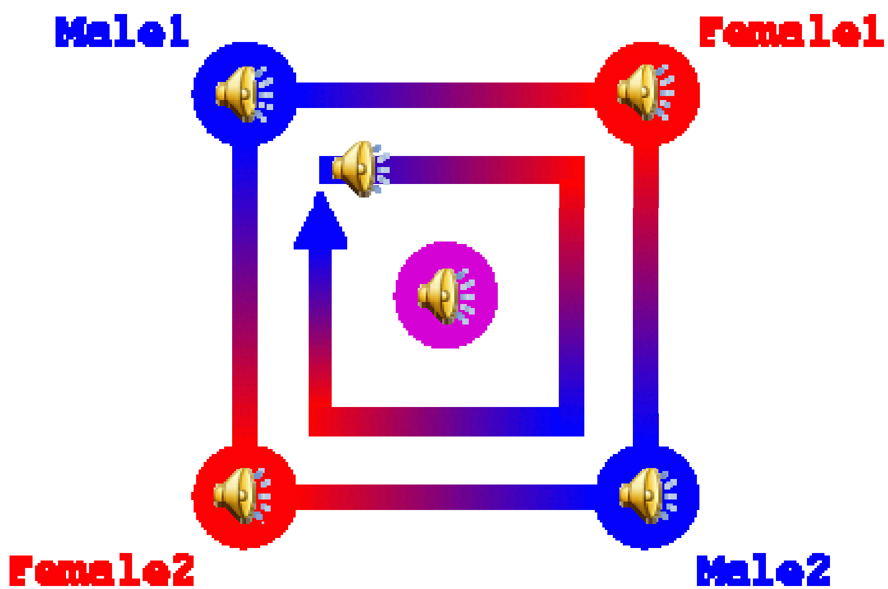
Voice morphing

Two voices:

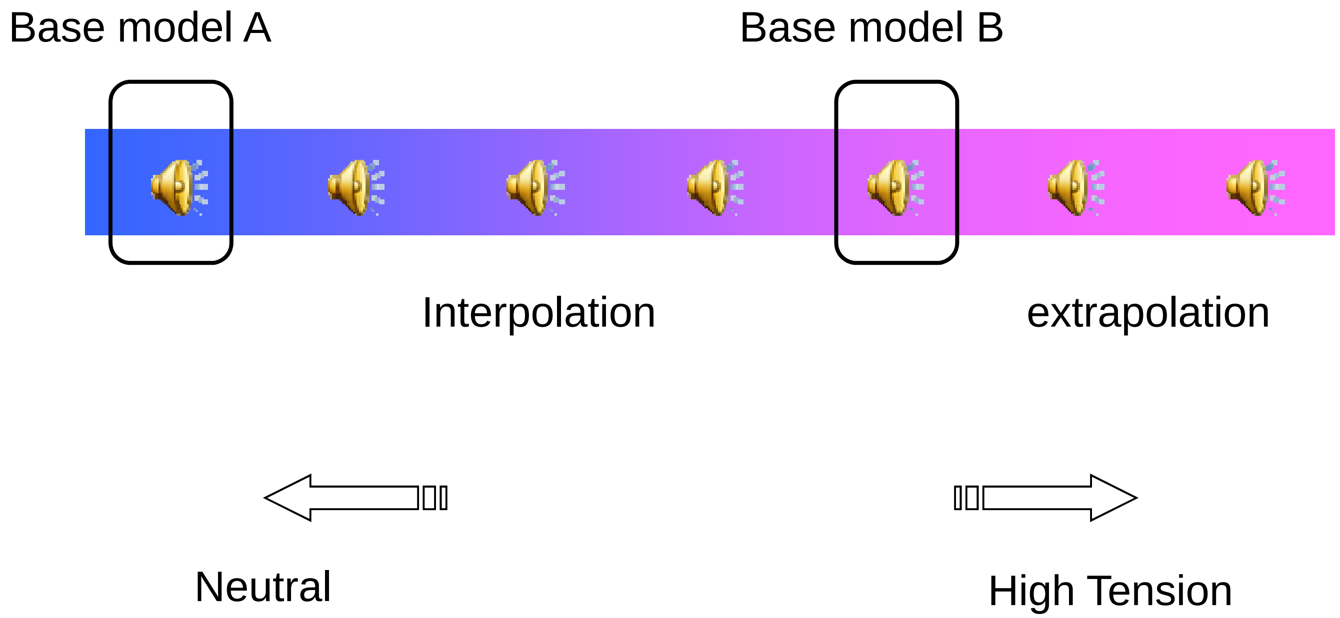




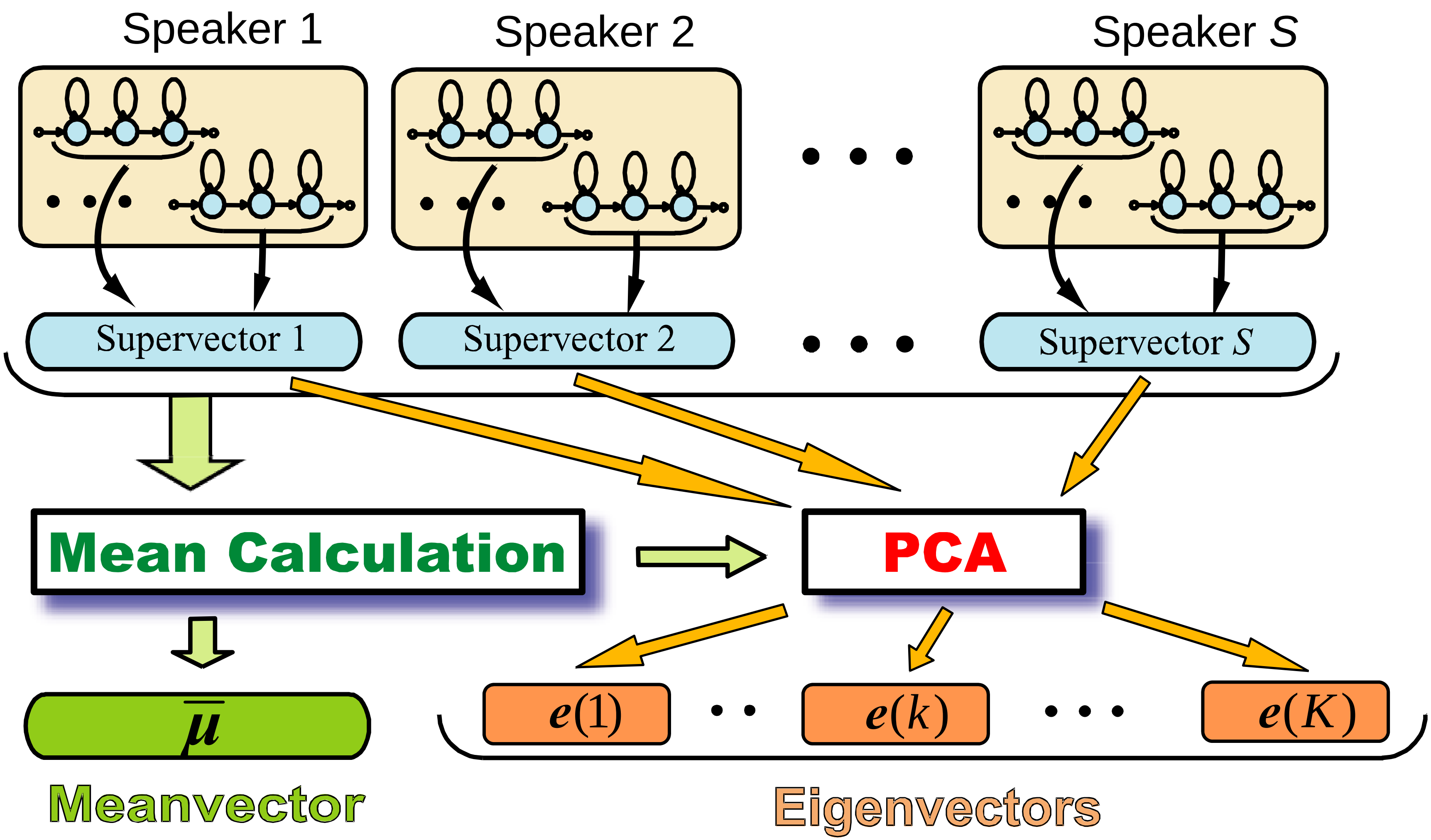
Four voices:



Interpolation of speaking styles



Eigenvoice (producing voices) [Shichiri; '02]



Click here for a demo

Multilingual speech synthesis

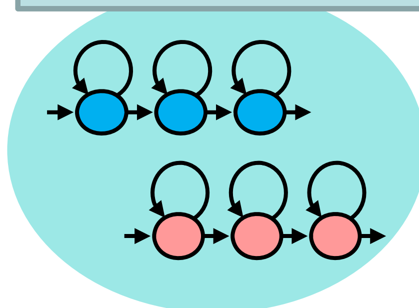
- Japanese
- American English
- Chinese (Mandarin) (by ATR)
- Brazilian Portuguese (by Nitech, and UFRJ)
- European Portuguese (by Nitech, Univ of Porto, and UFRJ)
- Slovenian
  (by Bostjan Vesnicer, University of Ljubljana, Slovenia)
- Swedish (by Anders Lundgren, KTH, Sweden)
- German (by University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

Singing voice synthesis [Oura; ’10] (1/2)

MusicXML

Musical score

Trained HMMs

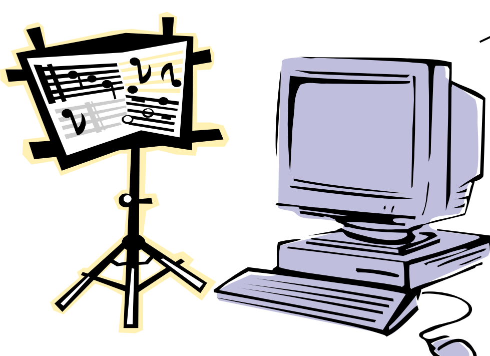


Male:

Female:

Singing voice database

Synthesized singing voice



Male:

Female:

Adaptation:

Enable computers to sing any song

Overview of this talk

1. Mathematical formulation
2. Implementation of individual components
3. Examples demonstrating its flexibility
4. Discussion and conclusion

Inclusion of all components

Problem of statistical parametric speech synthesis

Draw $\hat{\boldsymbol{x}}$ from $P(\boldsymbol{x} \mid \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W})$

$$= \sum_{\boldsymbol{l},\boldsymbol{L}} \iint \underbrace{P(\boldsymbol{x} \mid \boldsymbol{c})}_{\text{Waveform generation}} \; \underbrace{P(\boldsymbol{c} \mid \boldsymbol{l}, \lambda)}_{\text{Parameter generation}} \; \underbrace{P(\boldsymbol{l} \mid \boldsymbol{w}, \Lambda)}_{\text{Text processing}}$$

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

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

Relaxing approximations

Marginalizing model parameters

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

Marginalizing labels

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

Inclusion of waveform generation part

➔Waveform-level statistical model [Maia;’10]

Summary

Statistical approach to speech synthesis

- Whole speech synthesis process is described in 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, etc.

Final message

Is speech synthesis a messy problem?

No!

<u>Let us join speech synthesis research!</u>

Thanks!

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.

Acknowledgement

Keiichi Tokuda would like to thank HTS working group members, including Heiga Zen, Keiichiro Oura, Junichi Yamagichi, Tomoki Toda, Yoshihiko Nankaku, Kei Hahimoto, and Sayaka Shiota for their help.

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