Statistical approach to speech synthesis ---past, present, and future

Keiichi Tokuda
Nagoya Institute of Technology

INTERSPEECH 2019

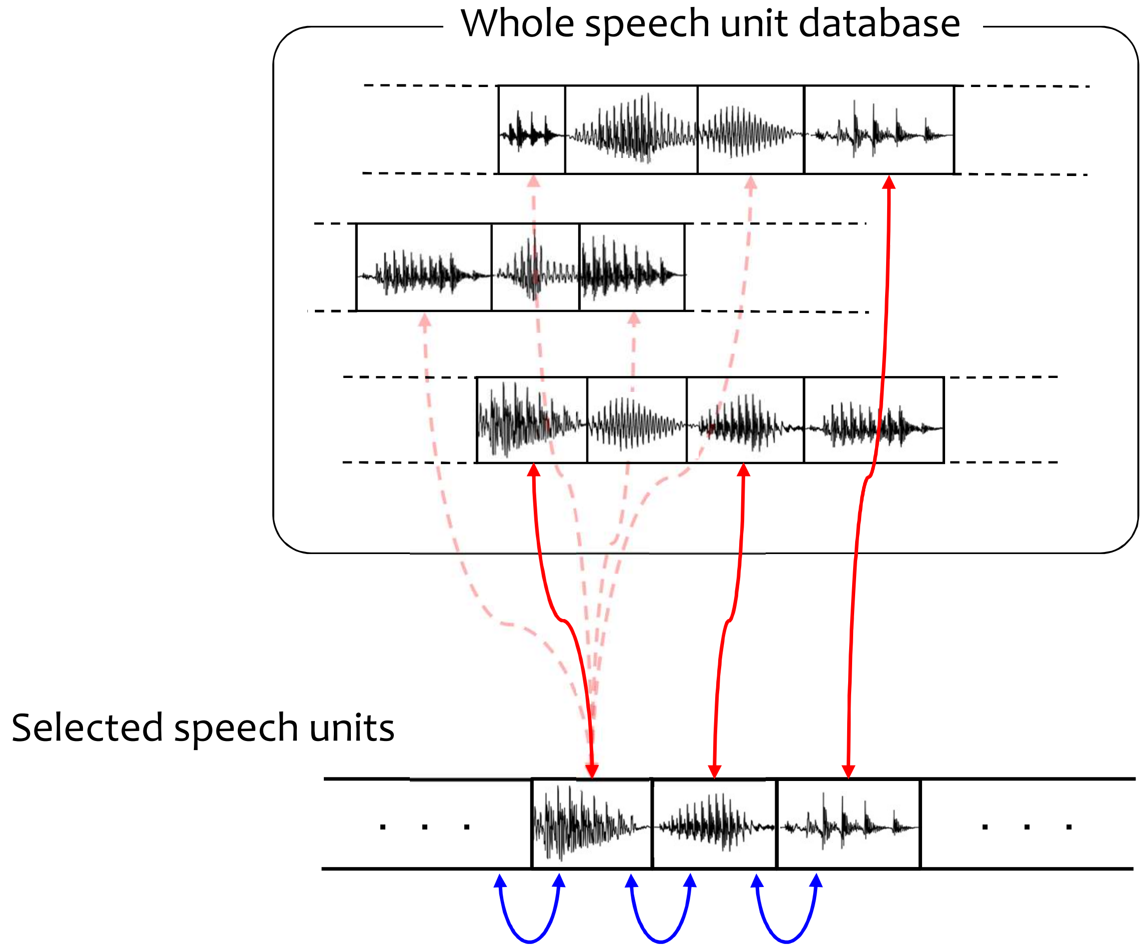
Speech synthesis approaches

- Rule-based, formant synthesis (~'90s)
  Phonetic units are built by hand-crafted rules

- Corpus-based, concatenative synthesis ('90s~)
  Concatenate speech units (in acoustic feature or waveform) from a database
  - Single inventory: diphone synthesis
  - Multiple inventory: unit selection synthesis →

- Corpus-based, statistical synthesis (late '90s~) ← We were working on this
  Source-filter model + statistical acoustic model
  - HMM (hidden Markov model) (1995~) >>
  - DNN (deep neural networks) (2013~)
  - WaveNet (2016~)

Unit-selection synthesis



When we are lucky:

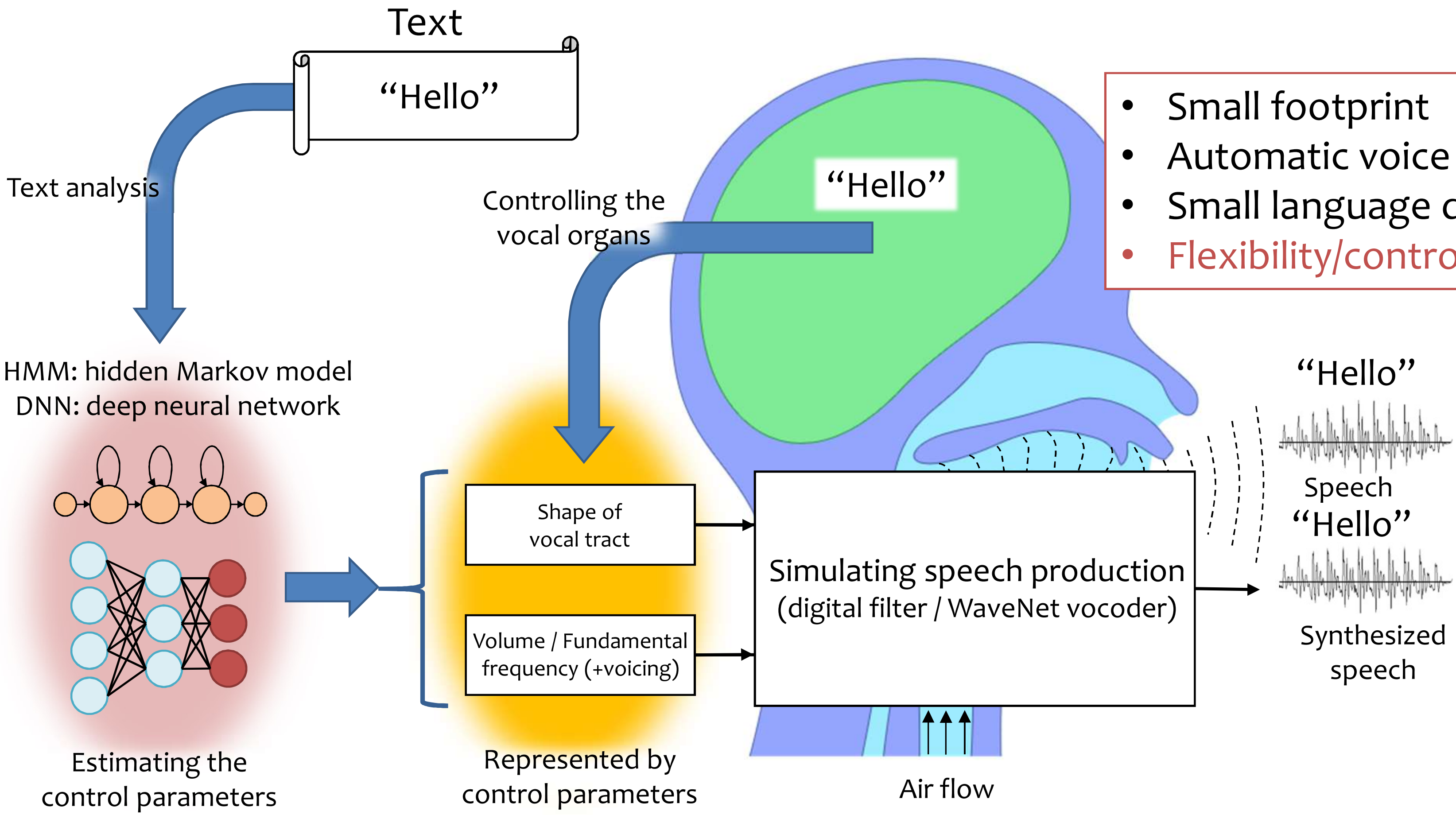
When we are unlucky:

Target cost

Concatenation cost

Minimize the total cost in runtime using dynamic programming

Statistical approach to speech synthesis



- Small footprint
- Automatic voice building
- Small language dependency
- Flexibility/controllability

Pros and cons

Solved by neural vocoding, e.g., WaveNet

Unit selection	Statistical parametric	
Waveform concatenation → Natural sounding	Vocoded → buzzy or muffled	☹
Discontinuity, hit or miss	Smooth	
Work better for larger databases	Can work for small databases	
Large footprint	Small footprint	
Fixed voice → fixed style, fixed emotional expression, etc.	Flexible → speaker adaptation, speaking style interpolation, etc.	☺

No reason to hesitate to move onto the statistical approach

Outline

- Statistical formulation of speech synthesis
- HMM-based speech synthesis
- Deep neural networks
- Evaluation / data & software tools
- Other related topics

Outline

- Statistical formulation of speech synthesis
- HMM-based speech synthesis
- Deep neural networks
- Evaluation / data & software tools
- Other related topics

The basic problem of speech synthesis

We have a speech database, i.e.,
a set of pairs 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

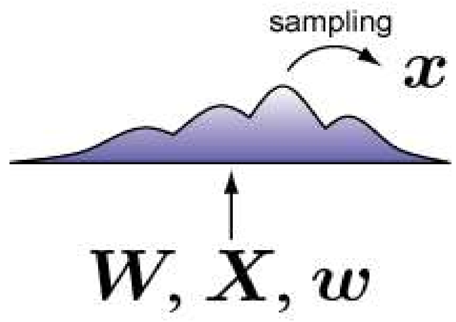
database

Given

- $\boldsymbol{w}$: text to be synthesized ($\boldsymbol{w} \notin \boldsymbol{W}$)
- $\boldsymbol{x}$: speech waveform

?

$$\boxed{\boldsymbol{x} \sim p(\boldsymbol{x} | \boldsymbol{w}, \boldsymbol{X}, \boldsymbol{W})}$$



Statistical formulation of speech synthesis (1/4)

- Estimating predictive distribution is hard.
  → Introduce generative representation (λ: model parameters)

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

- It is difficult to perform integral over auxiliary variables
  → Approximate integral by maximizing $p(\lambda|\boldsymbol{X},\boldsymbol{W})$

$$\boxed{\begin{array}{l}\hat{\lambda} = \arg\max_{\lambda} p(\hat{\lambda}|\boldsymbol{W},\boldsymbol{X}) \leftarrow \text{training} \\ \boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w},\hat{\lambda}) \leftarrow \text{generation}\end{array}}$$

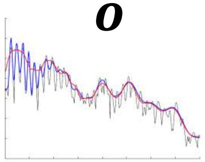
Statistical formulation of speech synthesis (2/4)

- Usually the generative model is decomposed into sub-modules, e.g.,

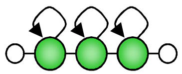
$$\boxed{p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)}$$

$\boldsymbol{x}$

$\boldsymbol{o}$



Feature extraction /waveform generation



Acoustic model

Text analysis

$\boldsymbol{o}$: parametric representation of speech waveform $\boldsymbol{x}$

$\boldsymbol{l}$: linguistic feature for $\boldsymbol{w}$

$\lambda = \{\lambda_A, \lambda_L\}$: generative model parameter

λ_A: acoustic model parameter

λ_L: text analysis module parameter

>>

Linguistic feature

Phoneme (or distinctive feature)

- {preceding, current, succeeding} phonemes

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, etc.

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
- Syntactic/dependency information, etc.

Phrase

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

⋮

(→)

+Frame-level Duration and positional information

+Speaking styles, emotional expressions, etc. when we have such tags/lables

Statistical formulation of speech synthesis (3/4)

- Decompose the generative model into sub-modules:

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\lambda}|\boldsymbol{W}, \boldsymbol{X}) \leftarrow \text{training}$$
$$\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w}, \hat{\lambda}) \leftarrow \text{generation}$$

$$\{\hat{\lambda}_A, \hat{\lambda}_L\} = \arg\max_{\lambda_A, \lambda_L} \int \sum_{L} p(\boldsymbol{X}|\boldsymbol{O})\, p(\boldsymbol{O}|\boldsymbol{L}, \lambda_A)\, P(\boldsymbol{L}|\boldsymbol{W}, \lambda_L)\, d\boldsymbol{O}\, p(\lambda_A) p(\lambda_L)$$

$$\uparrow \text{training}$$

$$\boldsymbol{x} \sim \int \sum_{l} p(\boldsymbol{x}|\boldsymbol{o})\, p(\boldsymbol{o}|\boldsymbol{l}, \hat{\lambda}_A)\, P(\boldsymbol{l}|\boldsymbol{w}, \hat{\lambda}_L)\, d\boldsymbol{o} \leftarrow \text{generation}$$

Statistical formulation of speech synthesis (4/4)

- It is difficult to perform integral and sum
  → Approximated by step-by-step maximization ←

$\hat{\lambda}_L$: pre-trained text analysis module parameter

$\hat{\boldsymbol{O}} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{X}|\boldsymbol{O})$ ← speech feature parameter extraction

$\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} P(\boldsymbol{L}|\boldsymbol{W}, \hat{\lambda}_L)$ or $p(\hat{\boldsymbol{O}}|\boldsymbol{L}, \hat{\lambda}_A)$ or $p(\hat{\boldsymbol{O}}|\boldsymbol{L}, \hat{\lambda}_A)P(\boldsymbol{L}|\boldsymbol{W}, \hat{\lambda}_L)$ ← labeling

$\hat{\lambda}_A = \arg\max_{\lambda_A} p(\hat{\boldsymbol{O}}|\hat{\boldsymbol{L}}, \lambda_A)p(\lambda_A)$ ← acoustic model training

↑ Training

↓ Synthesis

$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} P(\boldsymbol{l}|\boldsymbol{w}, \hat{\lambda}_L)$ ← text analysis

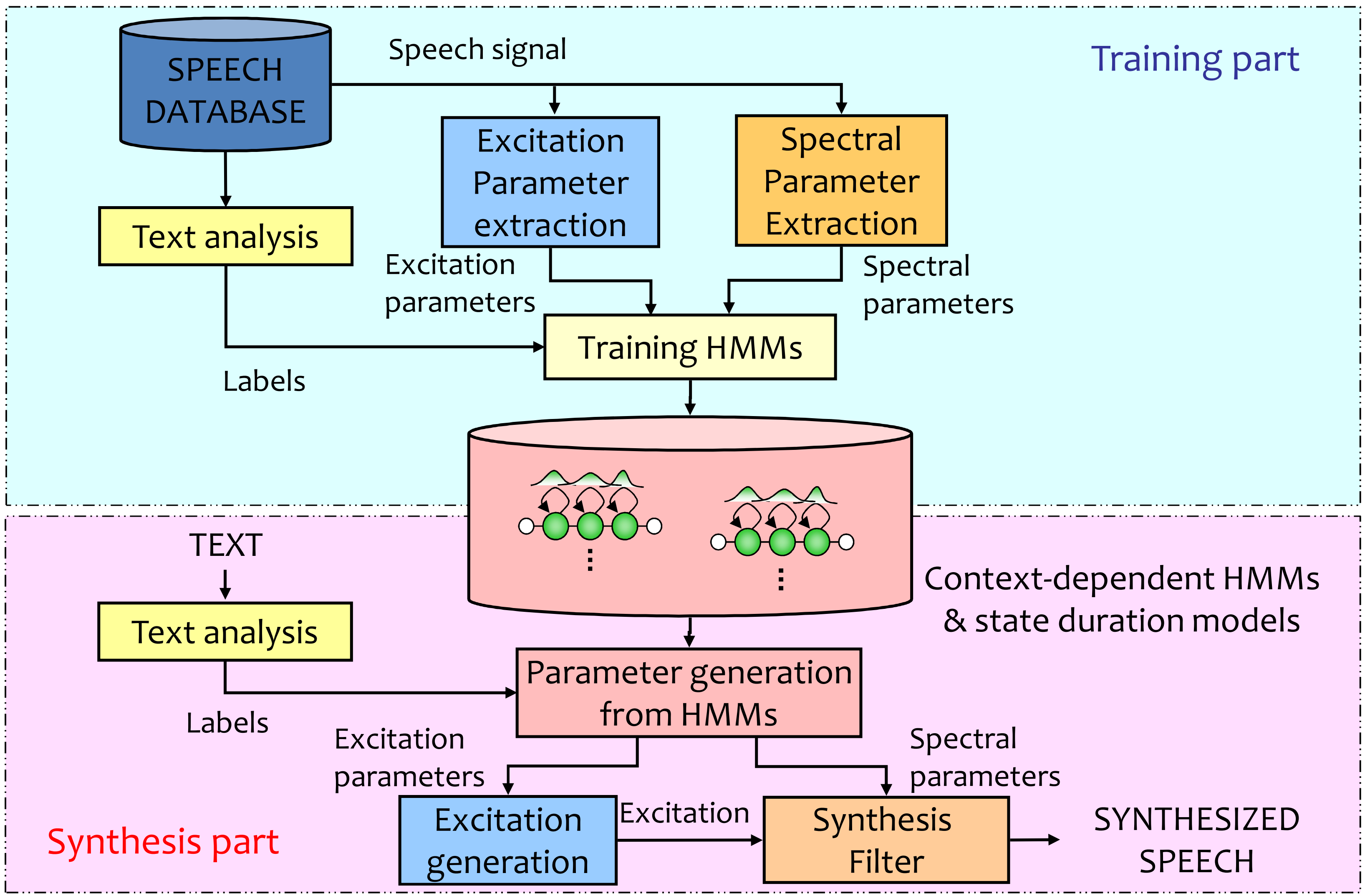
$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{c}} p(\boldsymbol{o}|\hat{\boldsymbol{l}}, \hat{\lambda}_A)$ ← speech parameter generation

$\boldsymbol{x} \sim p(\boldsymbol{x}|\hat{\boldsymbol{o}})$ ← waveform generation

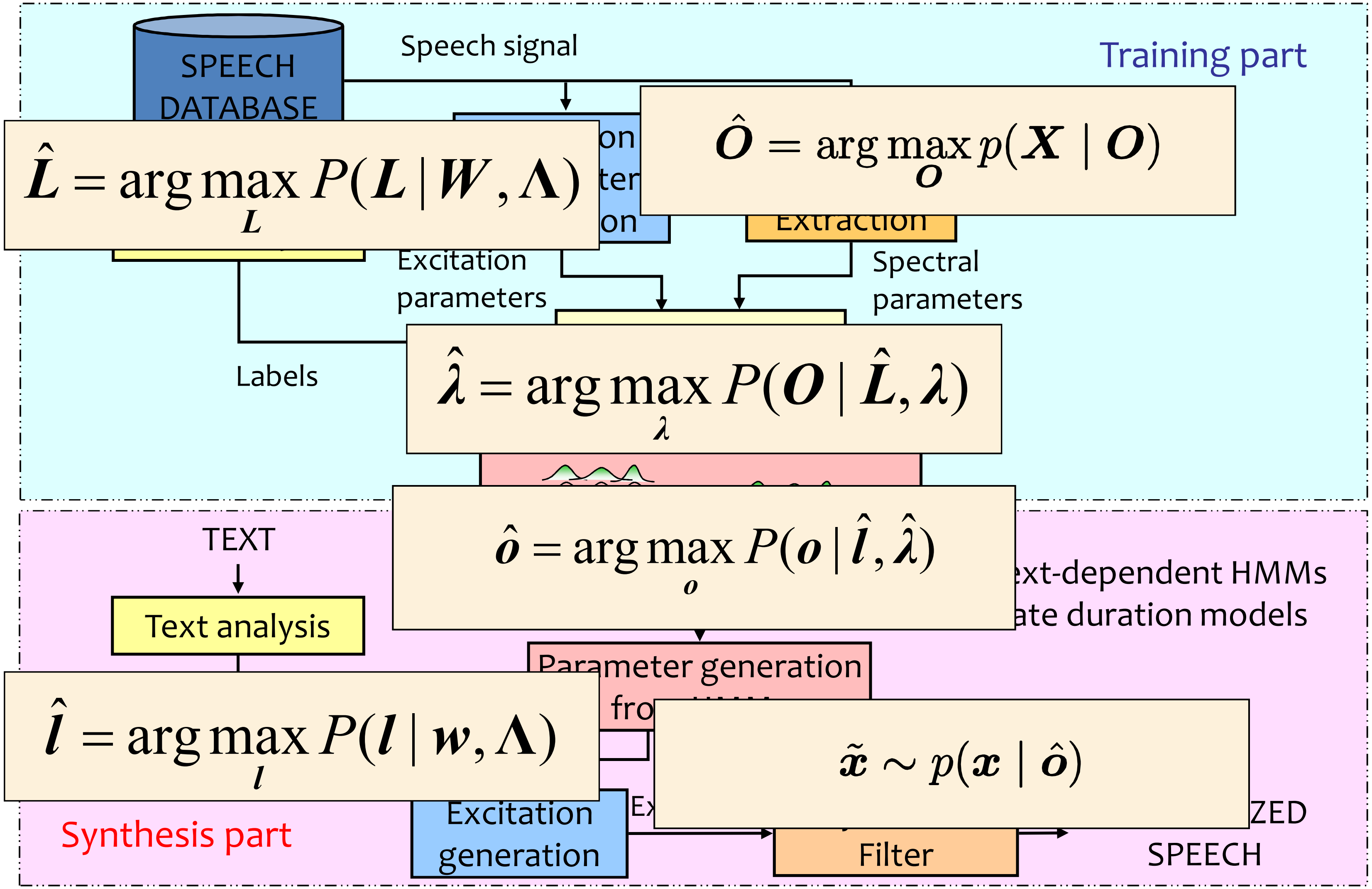
Outline

- Statistical formulation of speech synthesis
- HMM-based speech synthesis
- Deep neural networks
- Evaluation / data & software tools
- Other related topics

HMM-based speech synthesis system



HMM-based speech synthesis system



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

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

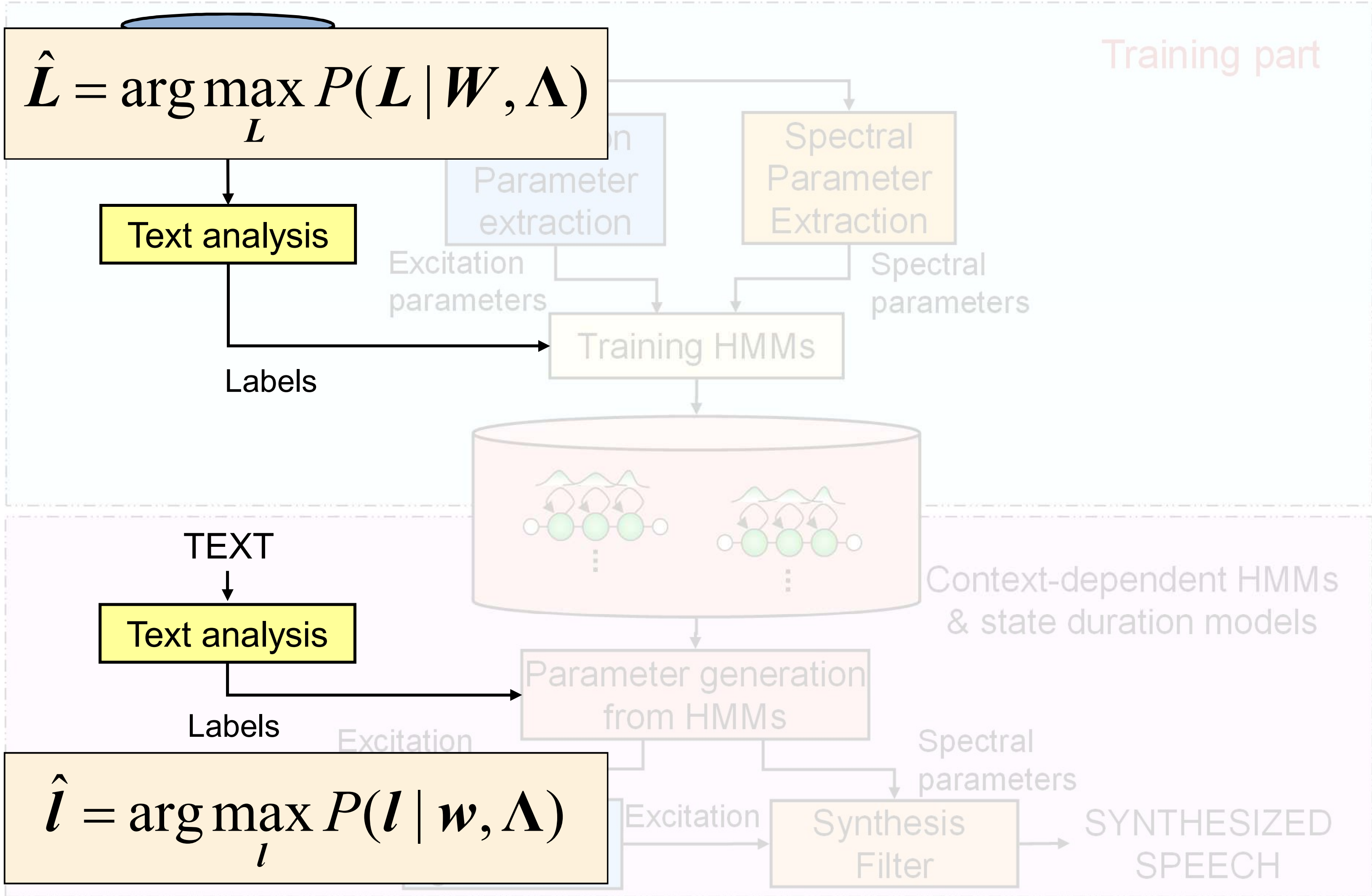
$$\hat{\lambda} = \arg\max_{\lambda} P(O \mid \hat{L}, \lambda)$$

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

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

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

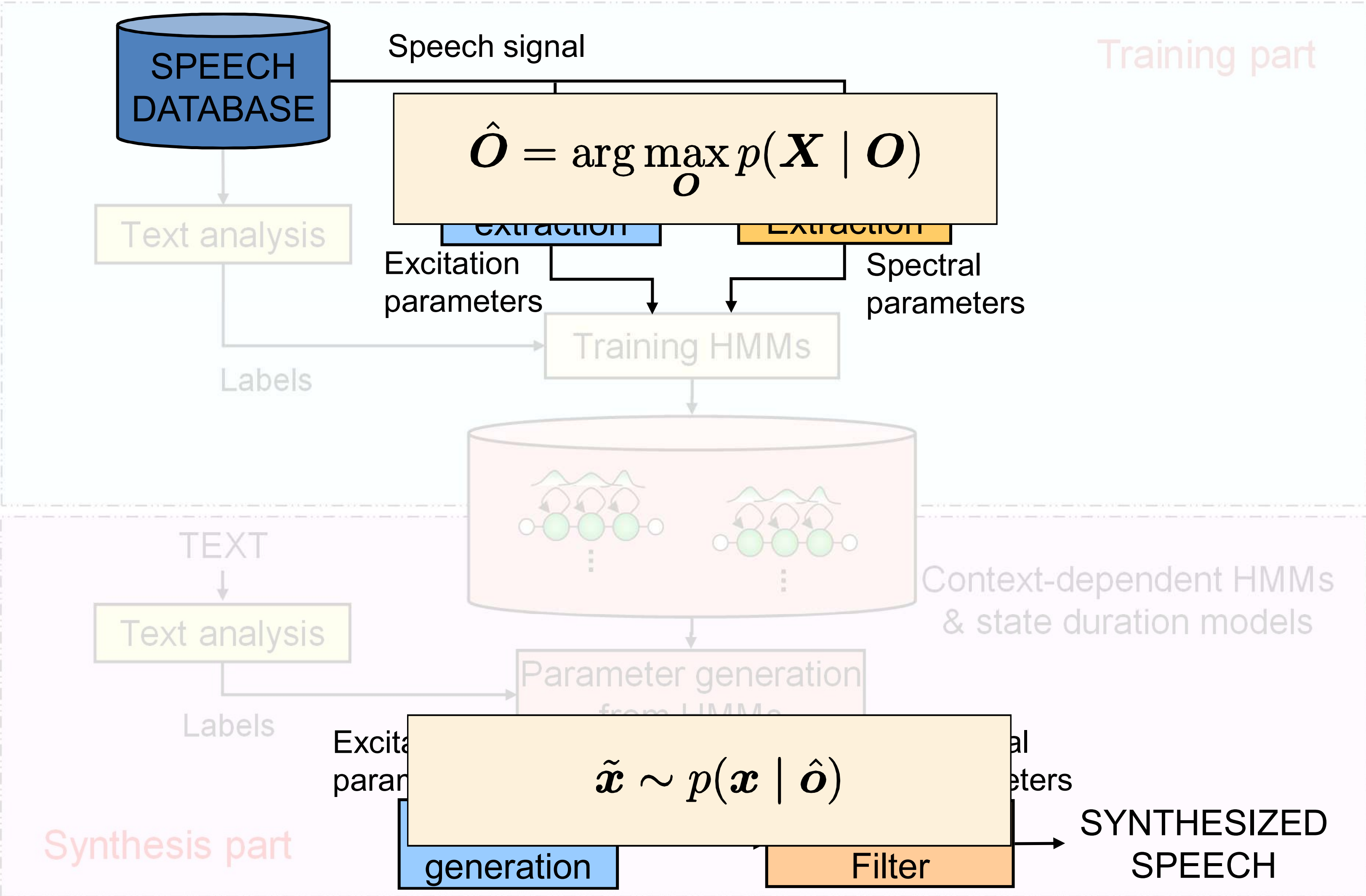
HMM-based speech synthesis system



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

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

HMM-based speech synthesis system



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

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

Mel-cepstrum-based spectral analysis

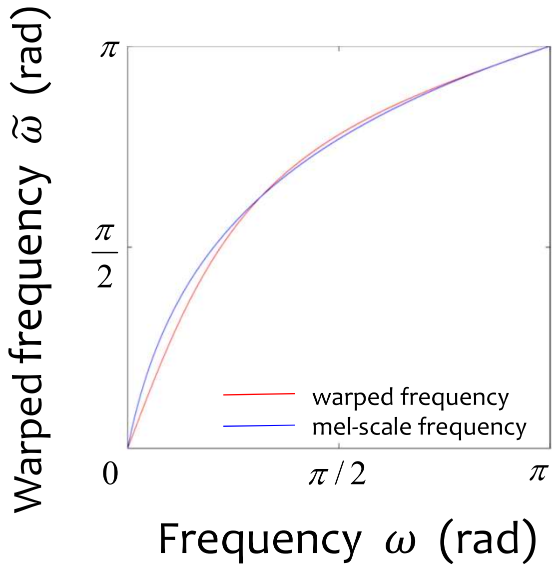
$$\boxed{H(e^{j\omega}) = \exp \sum_{m=0}^{M} c(m)\, e^{-j\tilde{\omega} m}, \qquad e^{-j\tilde{\omega}} = \frac{e^{-j\omega} - \alpha}{1 - \alpha e^{-j\omega}}}$$

$$\boldsymbol{c} = [c(0), c(1), \dots, c(M)]^{\mathrm{T}} \leftarrow \text{mel-cepstrum}$$

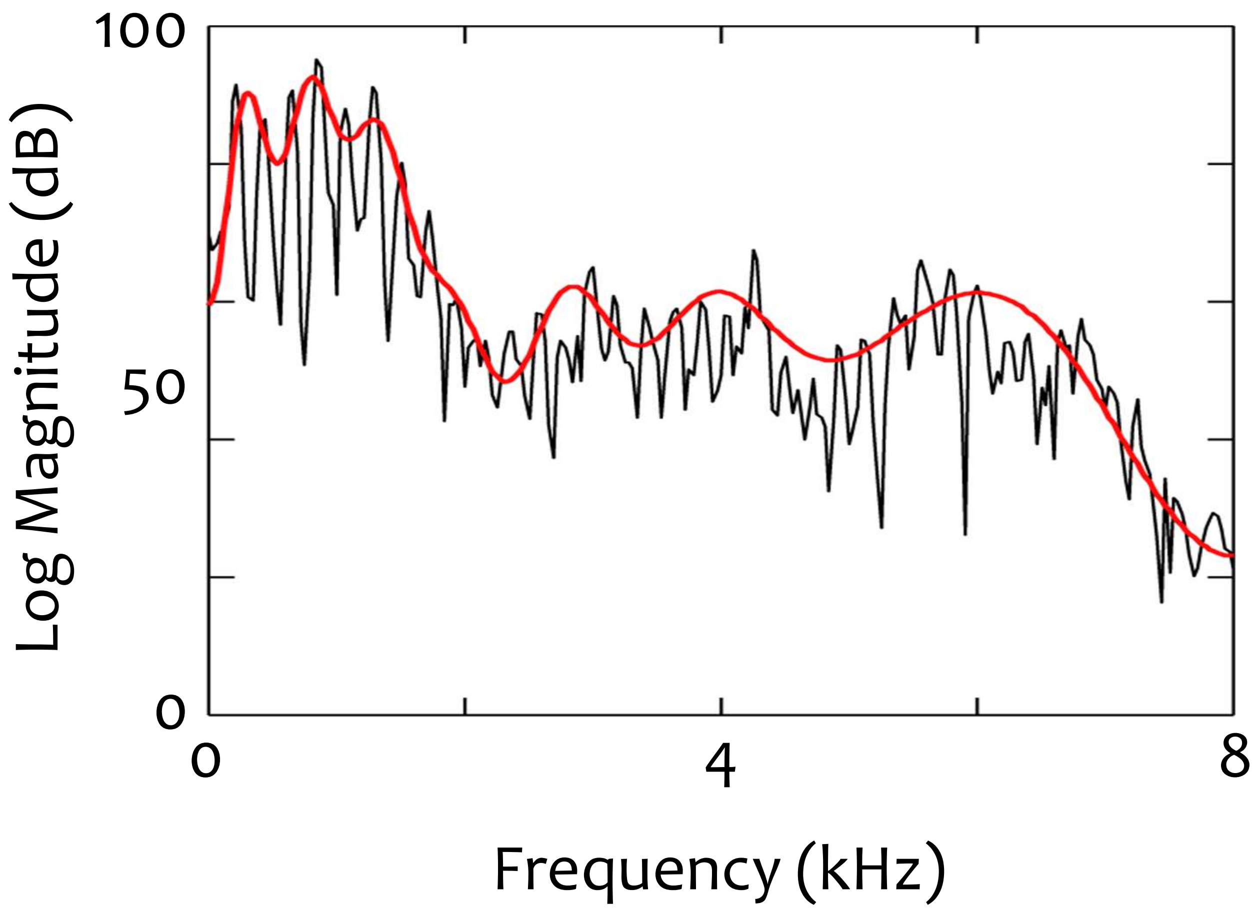
ML estimation of mel-cepstrum:

$$\boxed{\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} p(\boldsymbol{x}|\boldsymbol{c})} \quad \leftarrow$$

when $\boldsymbol{x}$ is Gaussian process,
$p(\boldsymbol{x}|\boldsymbol{c})$ is convex with respect to $\boldsymbol{c}$ [Fuka92]

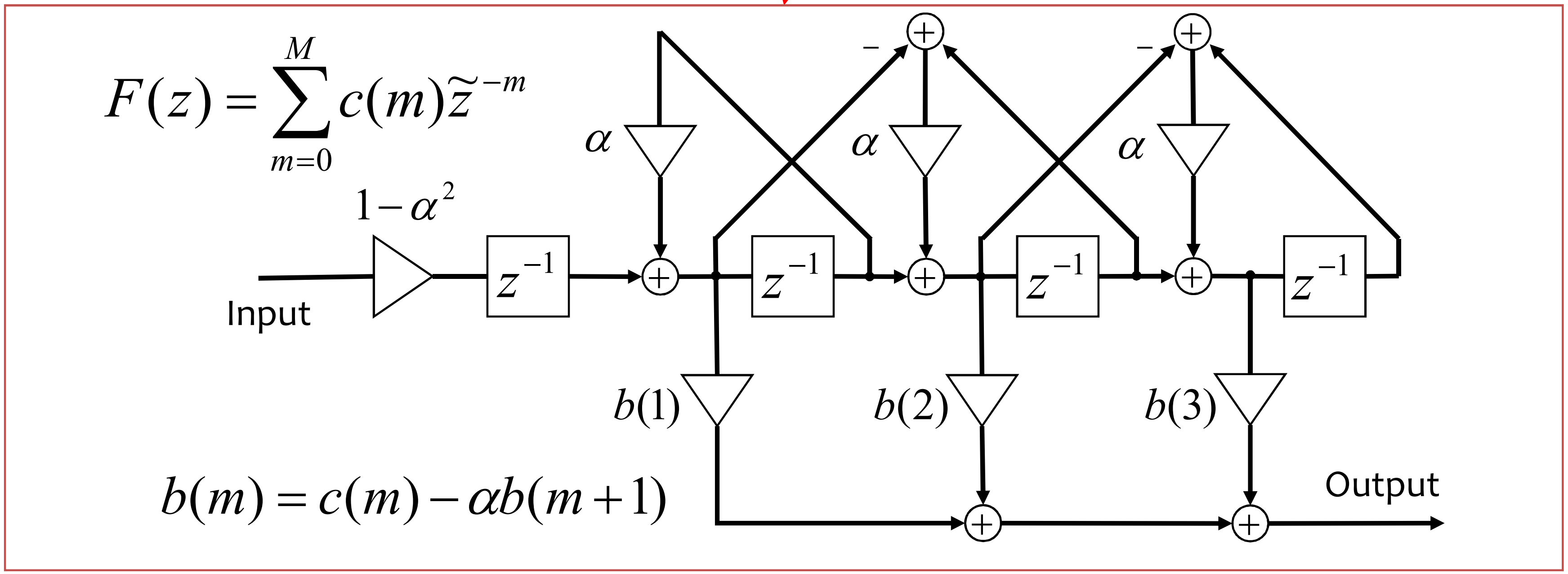


Spectral estimation example



MLSA filter (1/2) [Fukada '92]

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

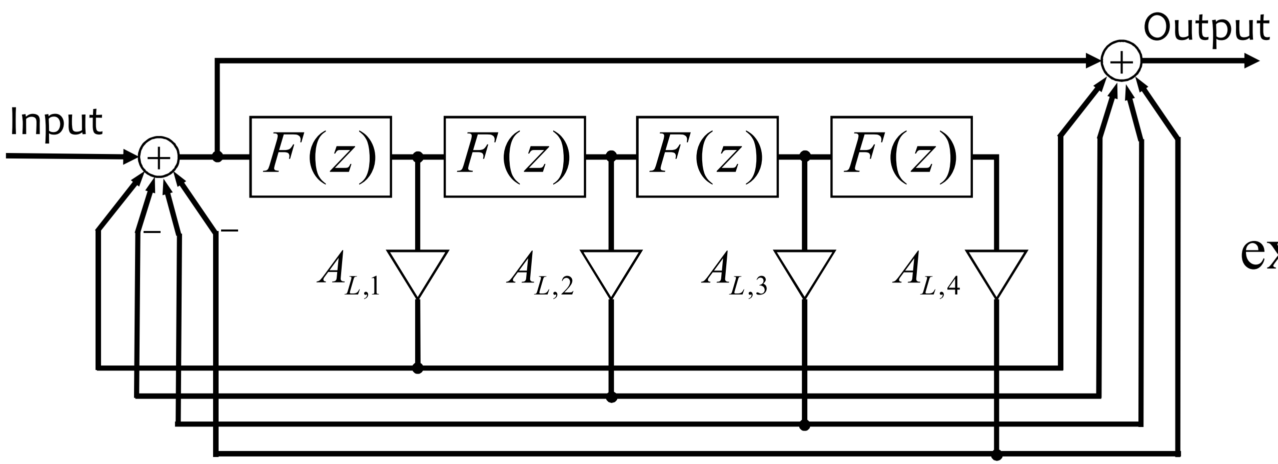


MLSA filter (2/2) [Fukada '92]

$$\exp x \cong \frac{1+\sum_{l=1}^{L} A_{L,l} x^l}{1+\sum_{l=1}^{L} A_{L,l} (-x)^l}$$

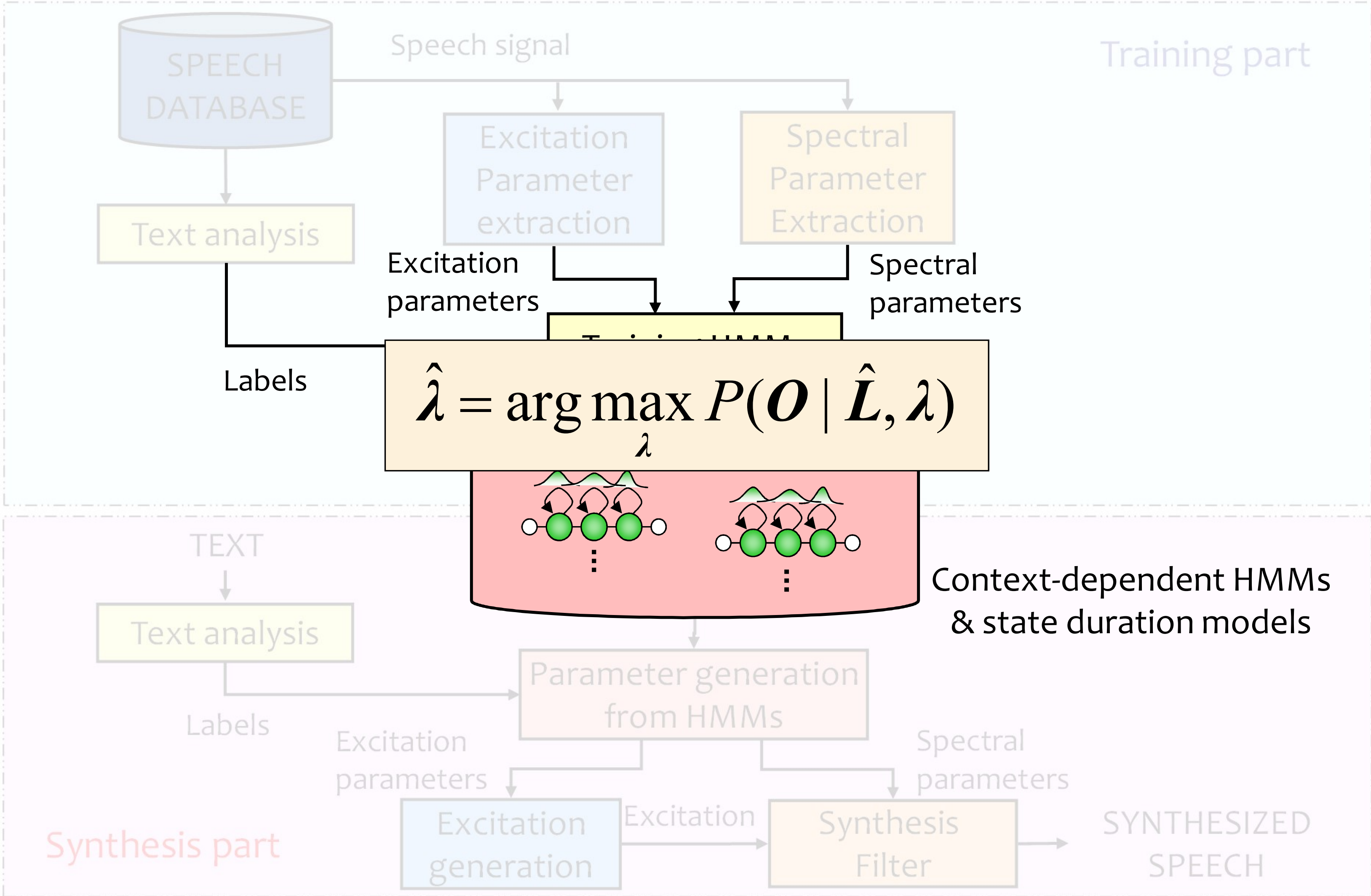
$$H(z) = \exp \sum_{m=0}^{M} c(m)\, \tilde{z}^{-m} = K \cdot \exp F(z)$$

- Approximation error < 0.24dB
- O(8M) operations/sample
- Stable filter

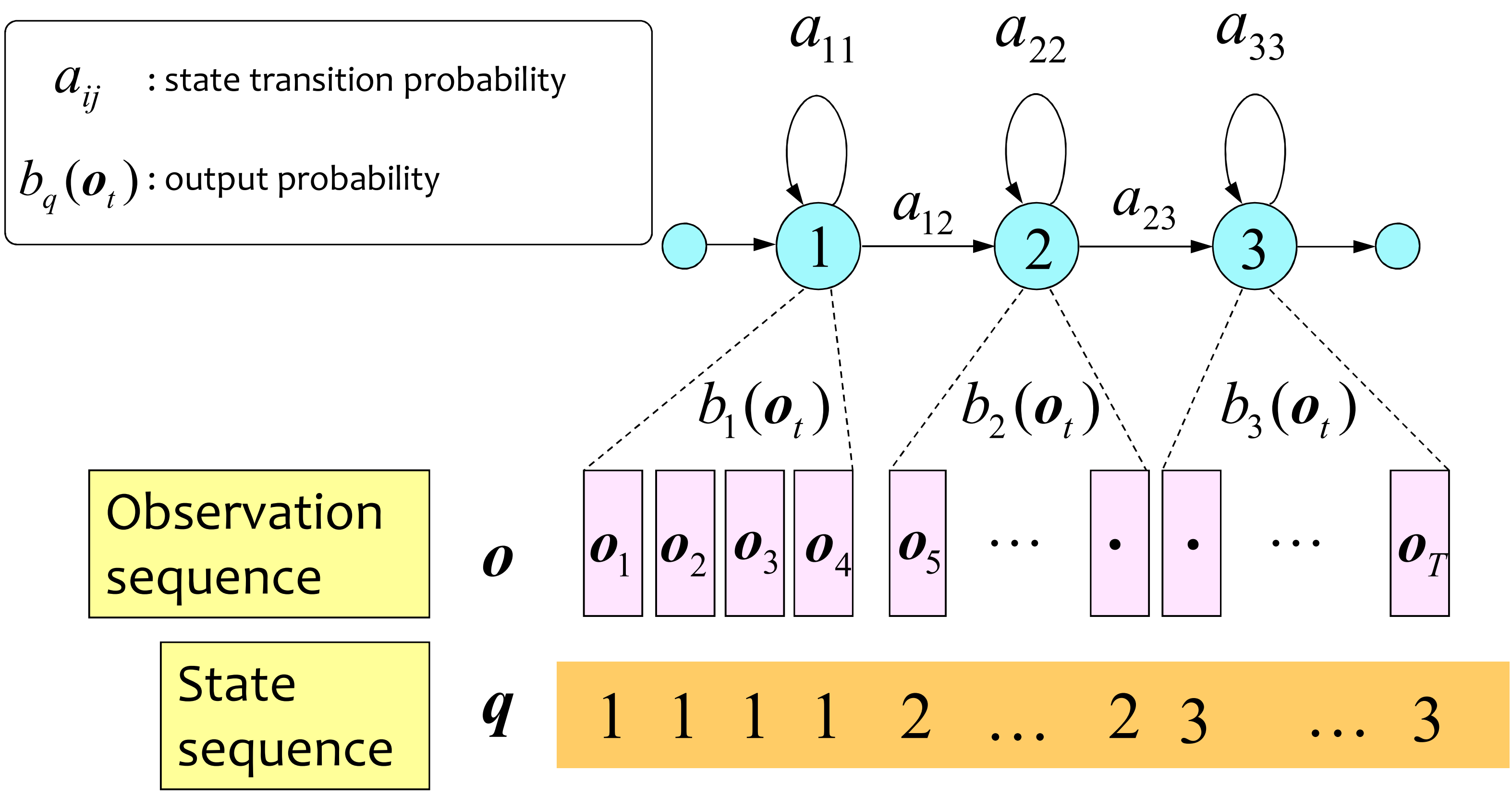


$$\exp F(z) \cong \frac{1+\sum_{l=1}^{L} A_{L,l} \{F(z)\}^l}{1+\sum_{l=1}^{L} A_{L,l} \{-F(z)\}^l}$$

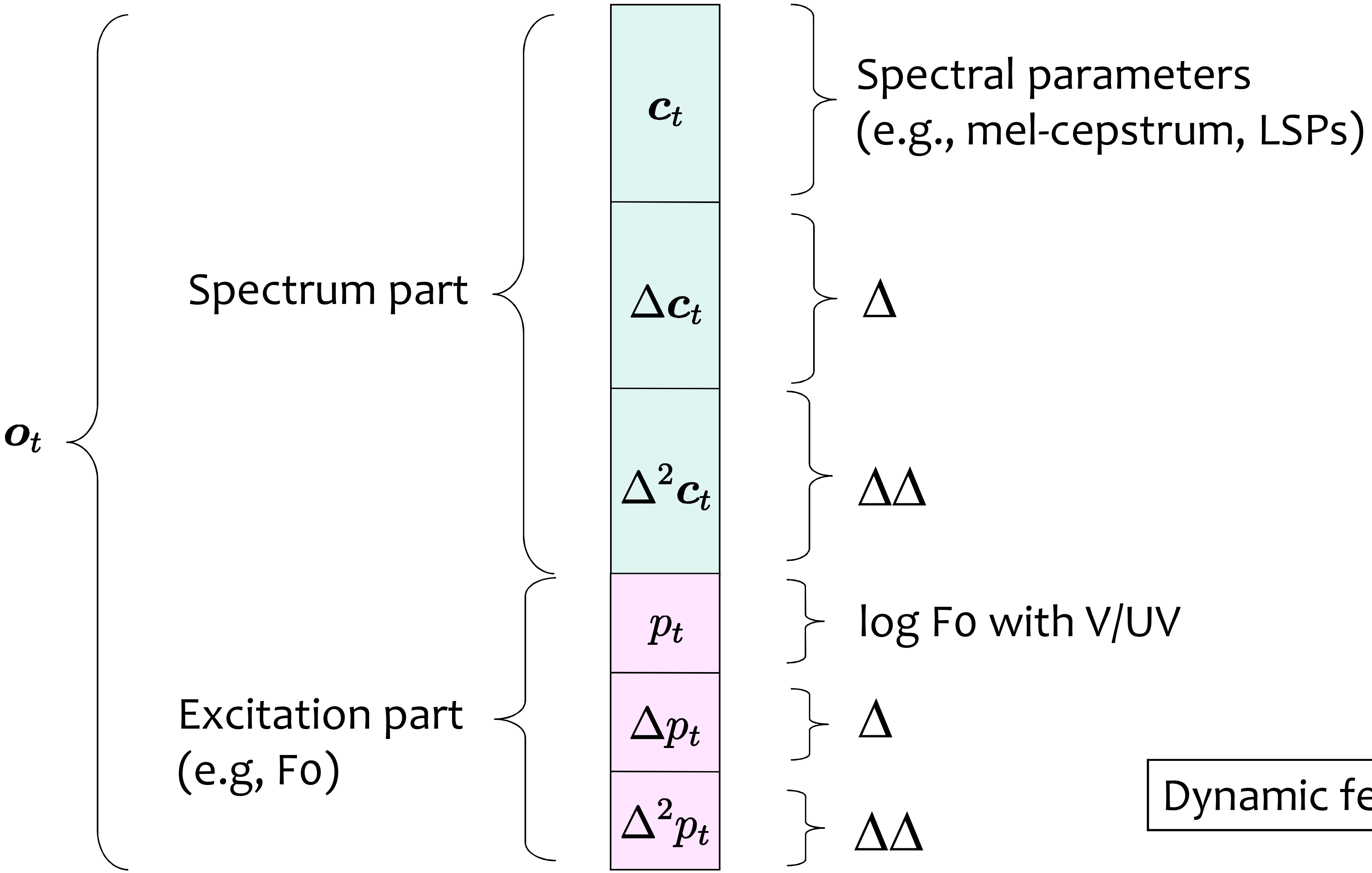
HMM-based speech synthesis system



Hidden Markov model (HMM)



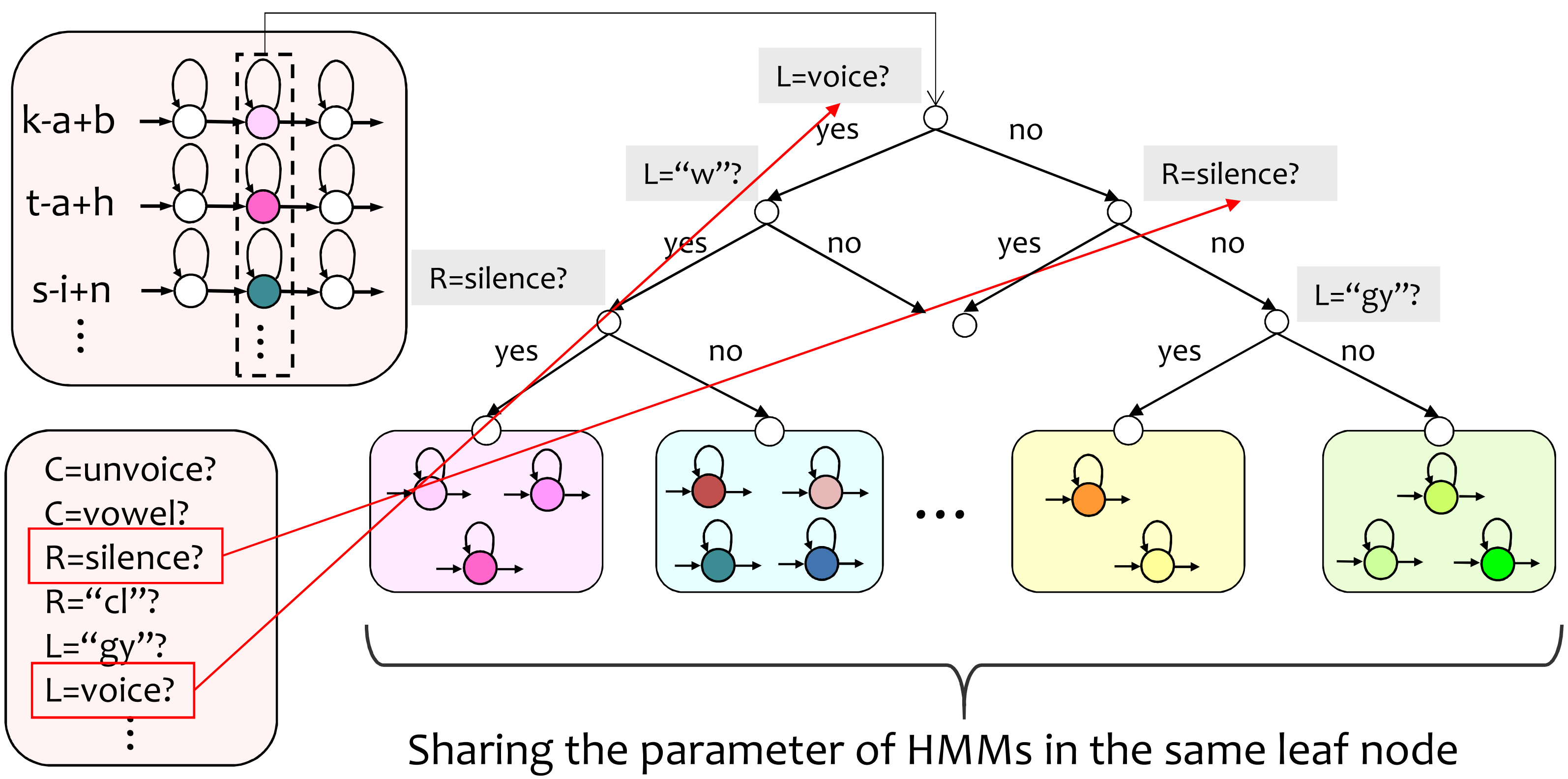
Structure of state output (observation) vector



Dynamic feature [Furui '86]

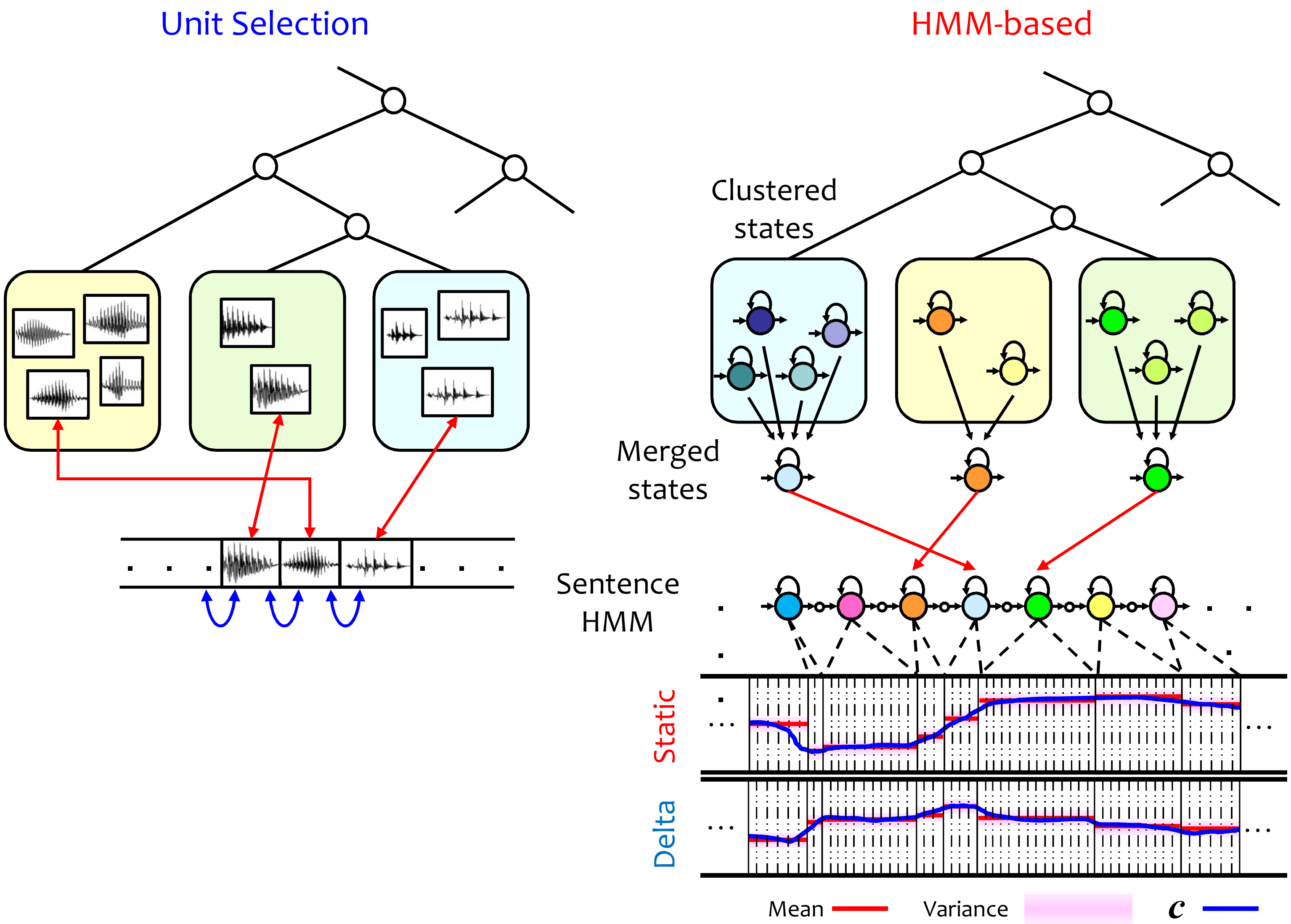
Decision tree-based state clustering [Odell; '95]

$p(\boldsymbol{o}|\boldsymbol{l}, \lambda_A)$: HMM, $\boldsymbol{l}$: linguistic feature

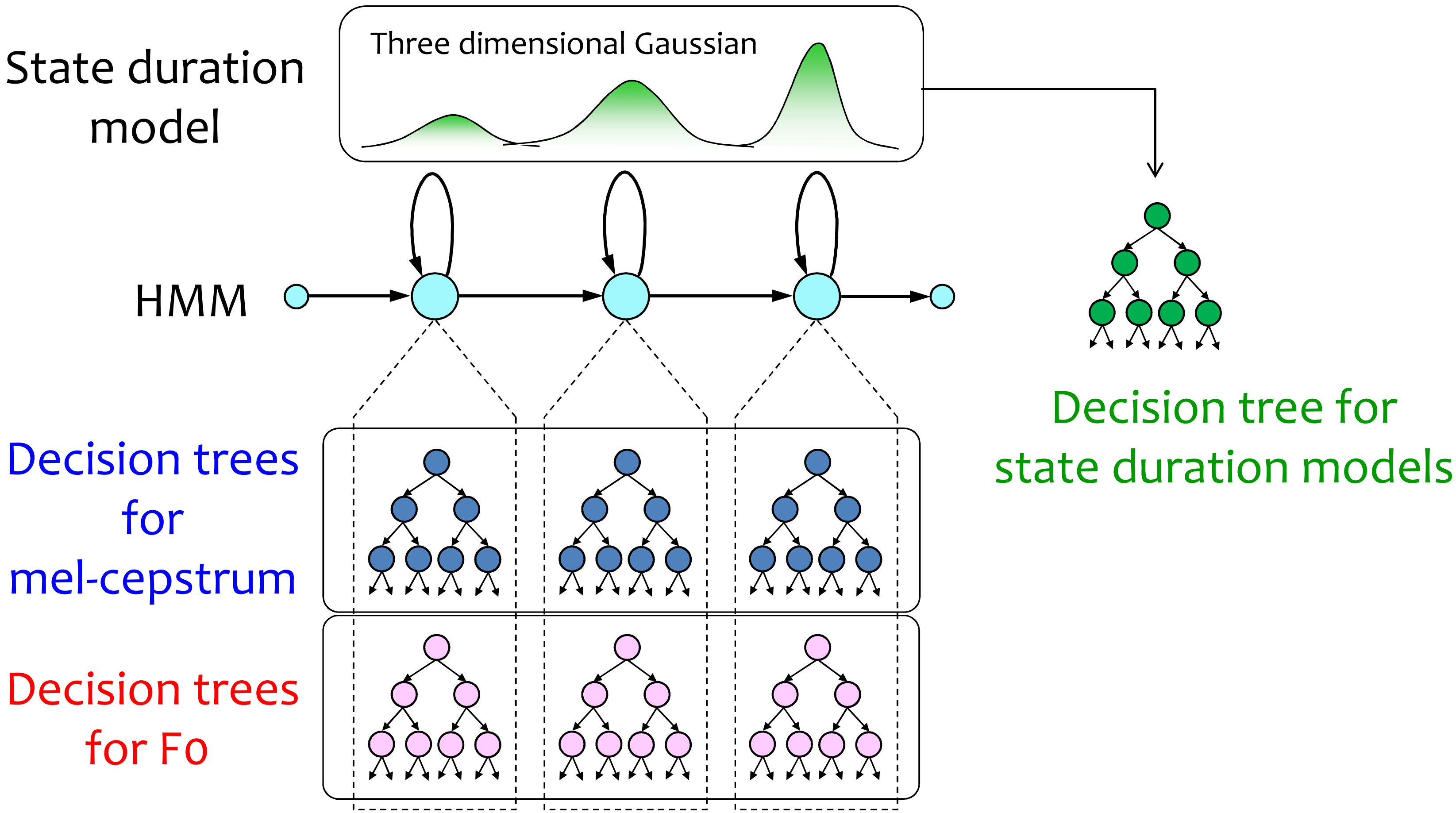


← >>

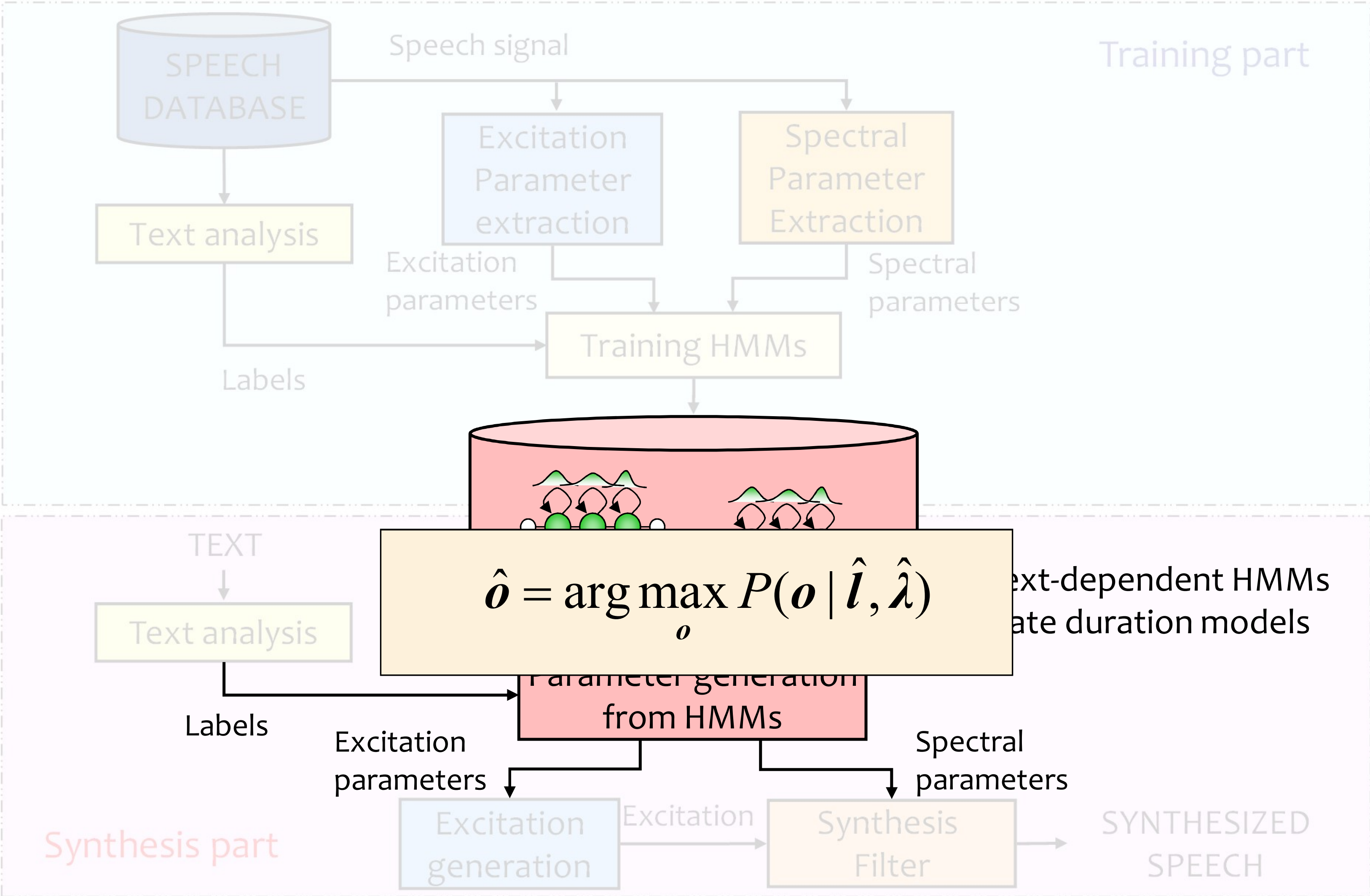
Relation between two approaches



Stream-dependent tree-based clustering



HMM-based speech synthesis system



Speech parameter generation algorithm

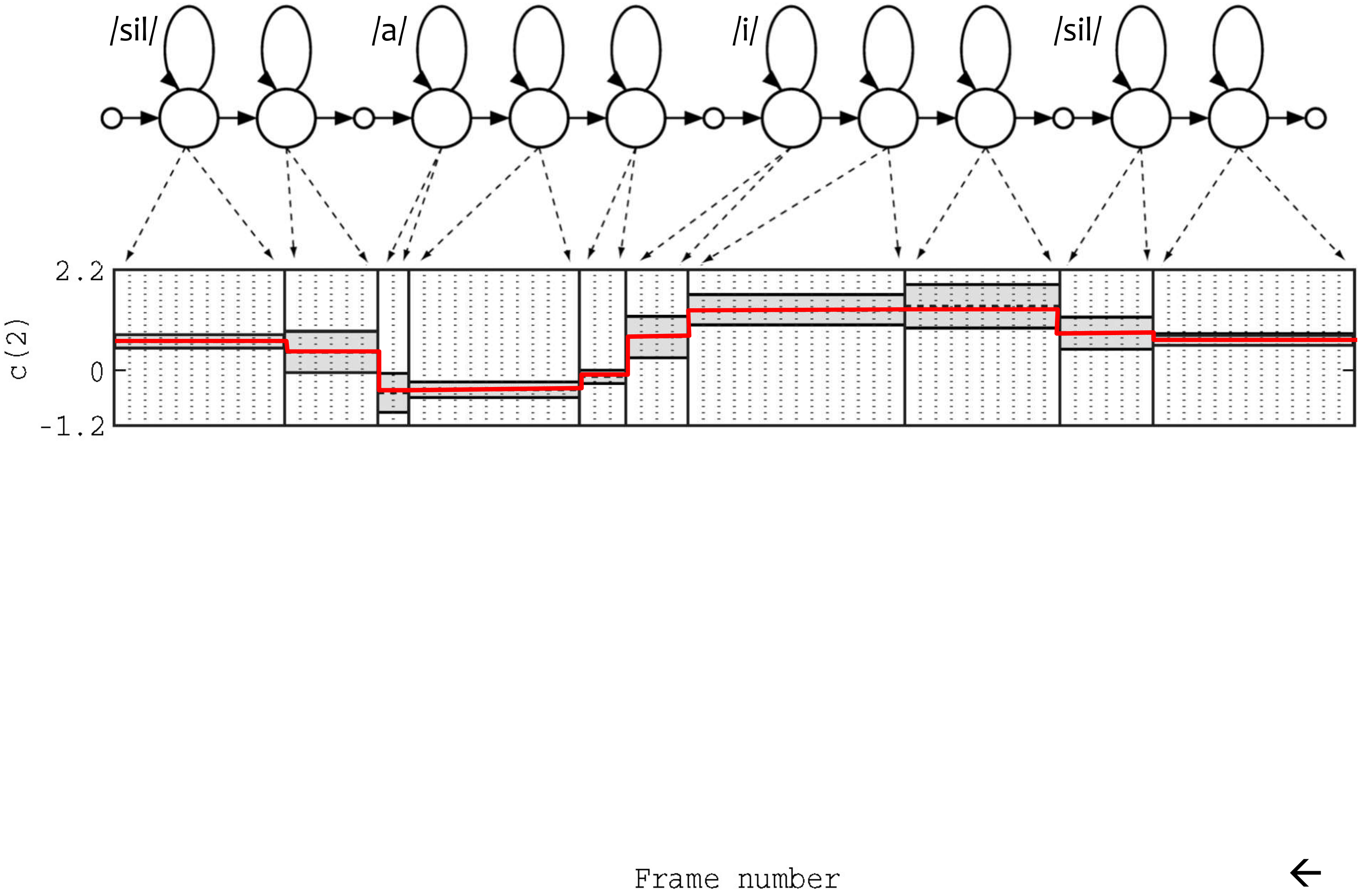
$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p\left(\boldsymbol{o} \middle| \hat{\boldsymbol{l}}, \hat{\lambda}_A\right) = \arg\max_{\boldsymbol{o}} \sum_{\boldsymbol{q}} P\left(\boldsymbol{o} \middle| \boldsymbol{q}, \hat{\lambda}\right) P\left(\boldsymbol{q} \middle| \hat{\boldsymbol{l}}, \hat{\lambda}\right)$$

$\boldsymbol{q}$: state sequence

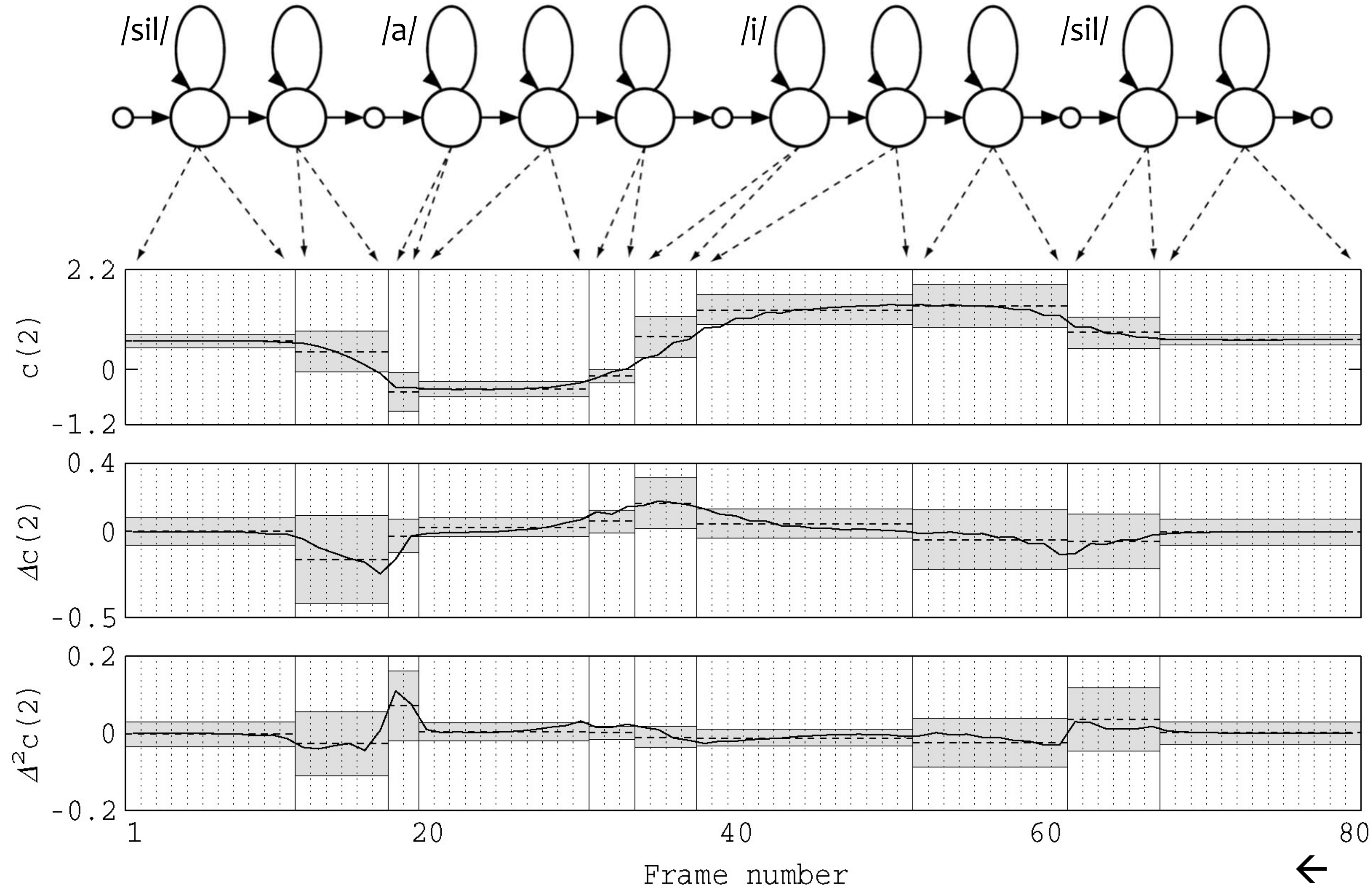
$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$ ← Determination of durations

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \hat{\lambda})$$ ← Determination of speech parameter

Generated speech parameter trajectory



Generated speech parameter trajectory



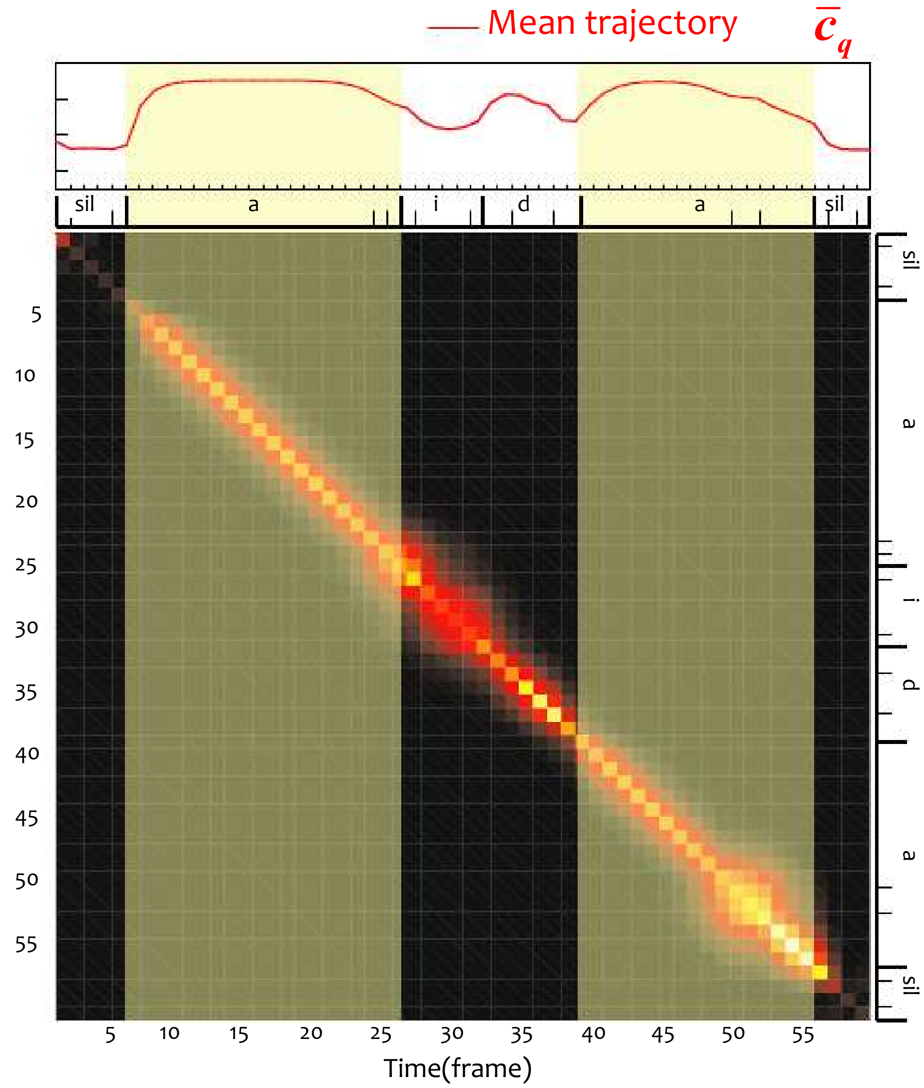
Trajectory HMM

with dynamic feature

w/o dynamic feature

$$\frac{1}{Z_{\boldsymbol{c}}} P(\boldsymbol{o}|\boldsymbol{q},\hat{\lambda}) = N(\boldsymbol{c}|\overline{\boldsymbol{c}}_{\boldsymbol{q}},\boldsymbol{P}_{\boldsymbol{q}})$$

$$Z_{\boldsymbol{c}} = \int P(\boldsymbol{o}|\boldsymbol{q},\hat{\lambda})\, d\boldsymbol{c}$$



Temporal covariance matrix $\boldsymbol{P}_{\boldsymbol{q}}$

Flexibility to control speech variations

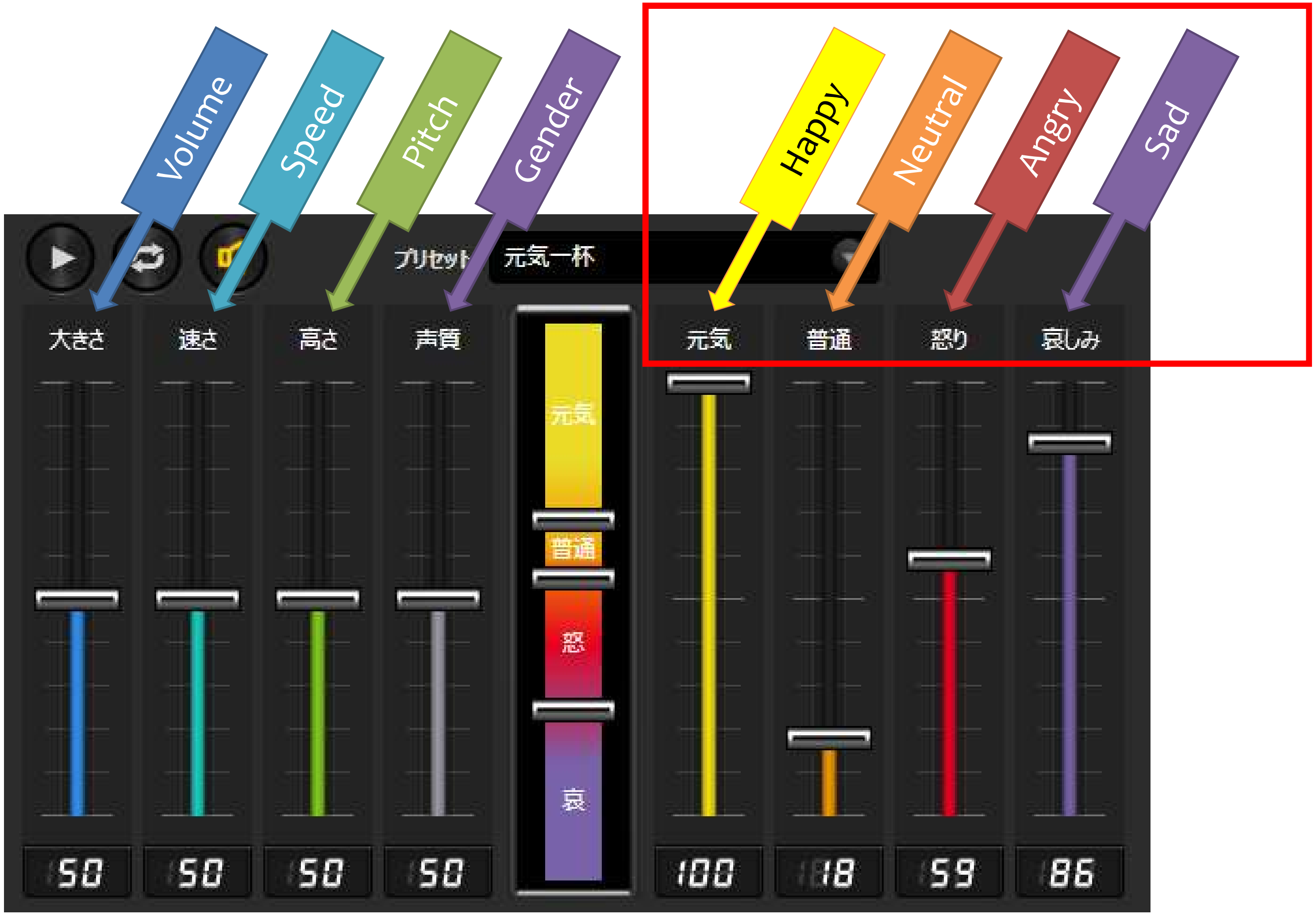
- **Speaker Adaptation** (mimicking voices)
  - [Tamura ’98], [Tamura ’01], [Yamagishi ’03], …

(→)

- **Speaker Interpolation** (mixing voices)
  - [Yoshimura ’97], …
- **Eigenvoice** (producing voices)
  - [Shichiri ’02], [Kazumi ’10], …
- **Multiple-regression** (controlling voices)
  - [Nose ’07], …

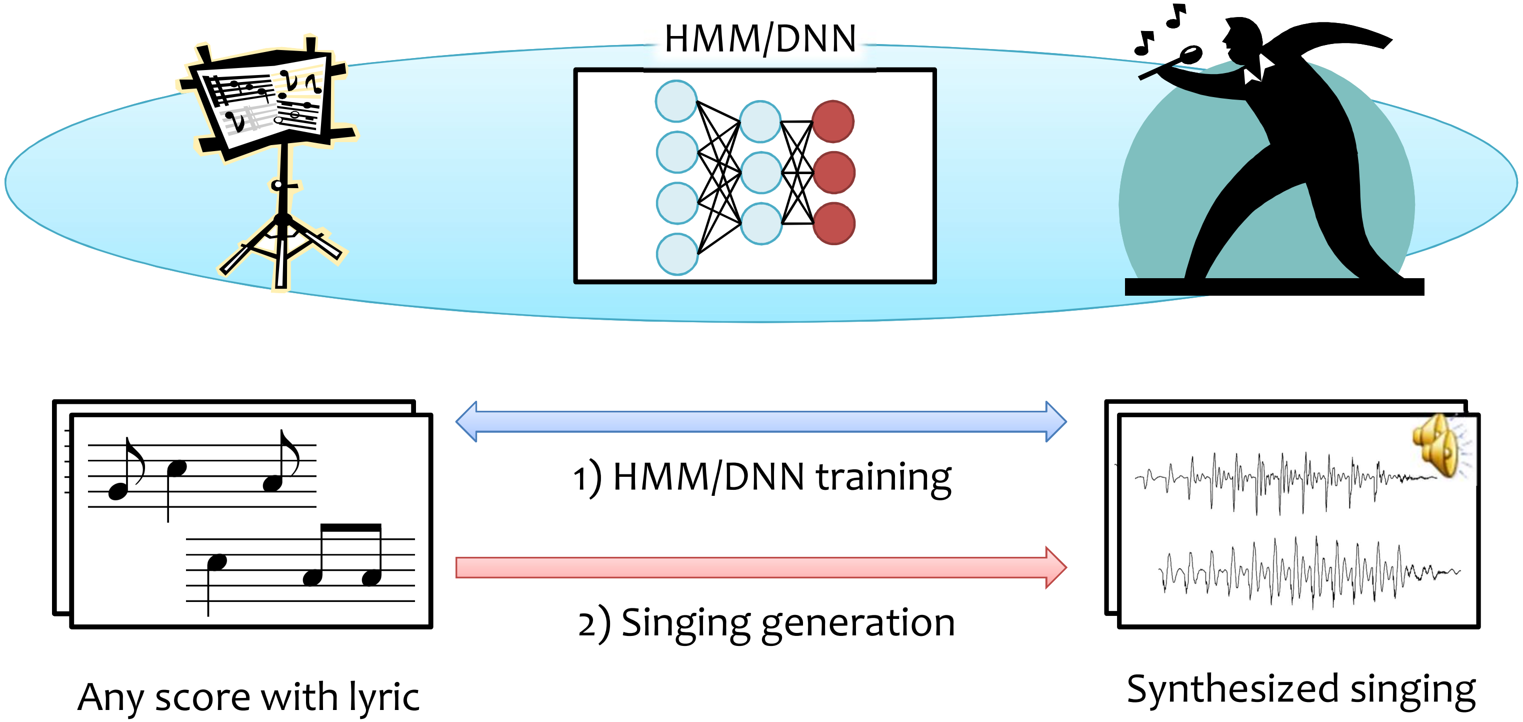
Only from publications by the HTS working group

Mixing emotional expressions



(←)

Singing synthesis



HMM+STRAIGHT

Outline

- Statistical formulation of speech synthesis
- HMM-based speech synthesis
- Deep neural networks
- Evaluation / data & software tools
- Other related topics

Hidden Markov model approach

STRAIGHT HMM

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\,p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\,P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

- GV-based parameter generation [Toda ’05]
- HSMM (hidden semi-Markov model) [Zen ’07]
- Trajectory HMM training [Zen ’07]
- MGE training [Wu ’08]
- Bayesian approach [Hashimoto ’09]
- Additive decision tree [Takaki ’10]
- Trainable excitation model [Maia ’07], etc.

Text analysis

Only from publications by the HTS working group

Recombining submodules

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\,p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\,P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

- Joint estimation of acoustic and excitation models [Maia '10]
- Log spectral distortion-version of MGE training [Wu '09]
- Factor analyzed trajectory HMM (STAVOCO) [Toda '08]
- Mel-cepstral analysis-integrated HMM [Nakamura '14]

Text analysis

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

Only from publications by the HTS working group

Deep neural network approaches (1/6)

FFNN, LSTM

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\, p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

Feature extraction /waveform generation

Acoustic model

Text analysis

- DNN-based speech synthesis [Zen '13]
- LSTM-based speech synthesis [Fan '14], etc.

DNN vs HMM

DNN

- Work for larger database?
- **Flat structure**
  - Easy to implement
  - Difficult to shouting troubles
- Often prior knowledge / model complexity is embedded in initialization and/or training process
- Suitable for parallel/distributed computation
- Optimization in continuous space

HMM (≅regression tree)

- Can work for small database?
- **Plausible structure**
  - Difficult to implement
  - Easy to shouting troubles
- Prior knowledge / model complexity can be given in an explicit form (e.g., model structure)
- Unsuitable for parallel/distributed computation
- Optimization in discrete space

Deep neural network approaches (2/6)

Source filter model FFNN, LSTM

$$\boxed{p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\, p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)}$$

→

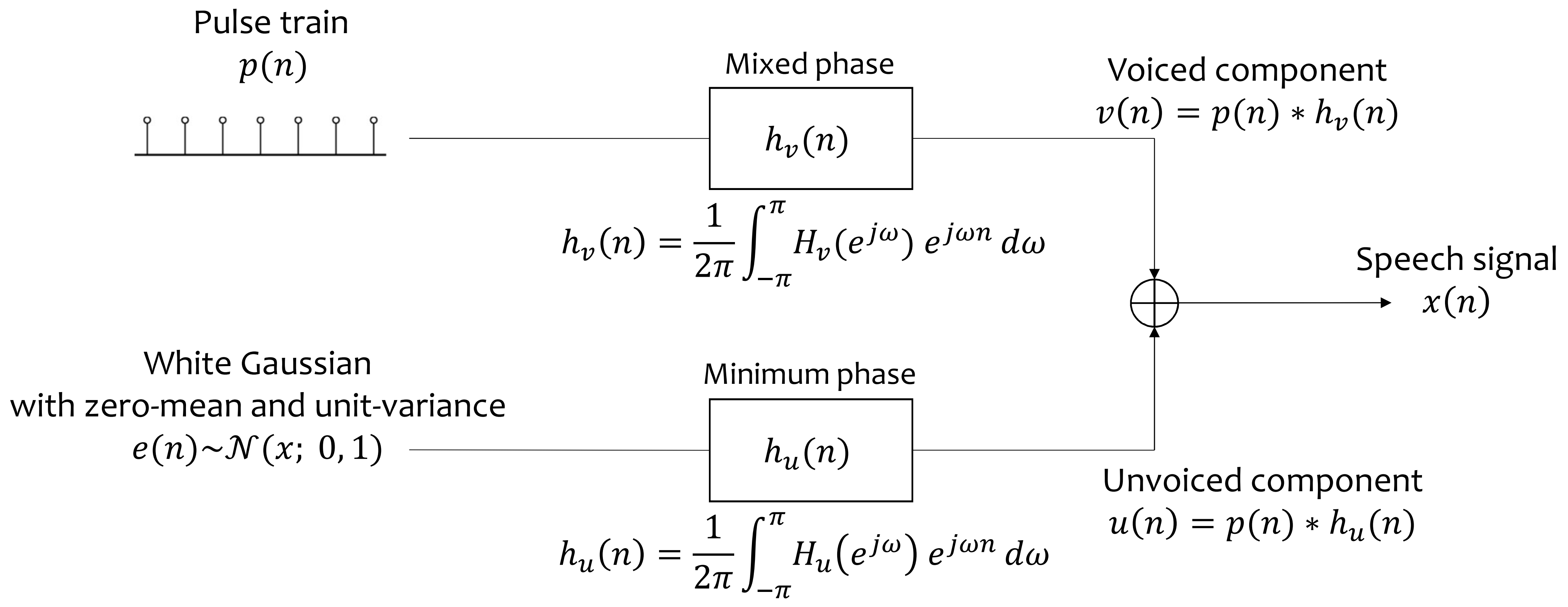
>>

1. Measuring likelihoods of speech waveform directly,
2. train a neural network
3. which models both voiced and unvoiced components.

t analysis

- Directly modeling speech waveforms by neural networks [Tokuda '15],
- Directly modeling voiced and unvoiced components by neural networks [Tokuda '16]

Speech signal model



Signal model for unvoiced+voiced sounds

Deep neural network approaches (3/6)

WaveNet, SampleRNN, WaveRNN, …
(autoregressive sturucture)

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\, p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

Feature extraction
/waveform generation

Acoustic model

Text analysis

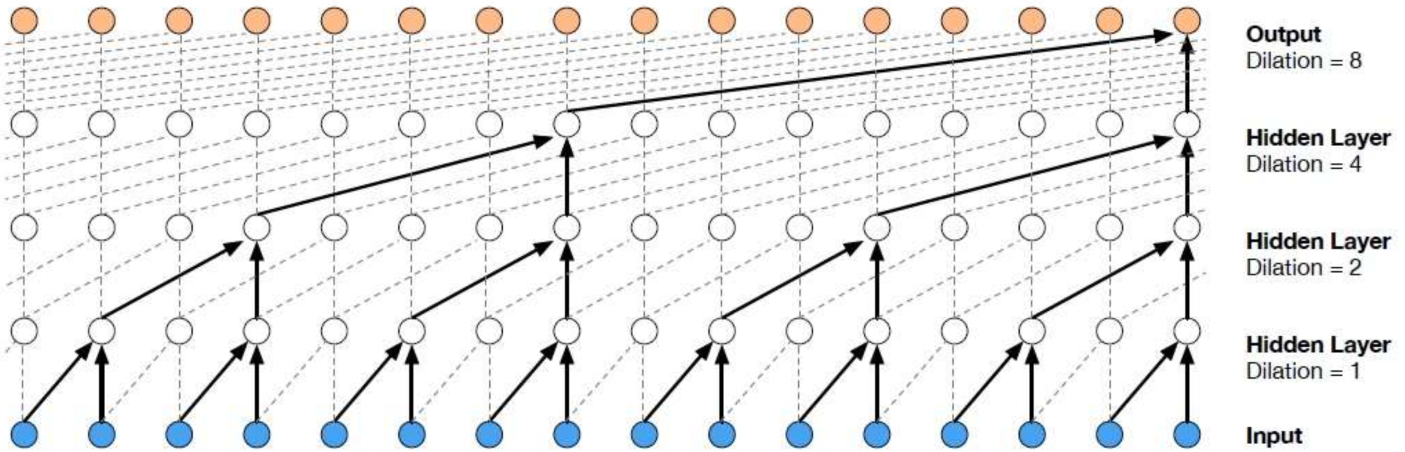
WaveNet

- Autoregressive generative model using convolutional NN
  - Directly modeling speech waveform

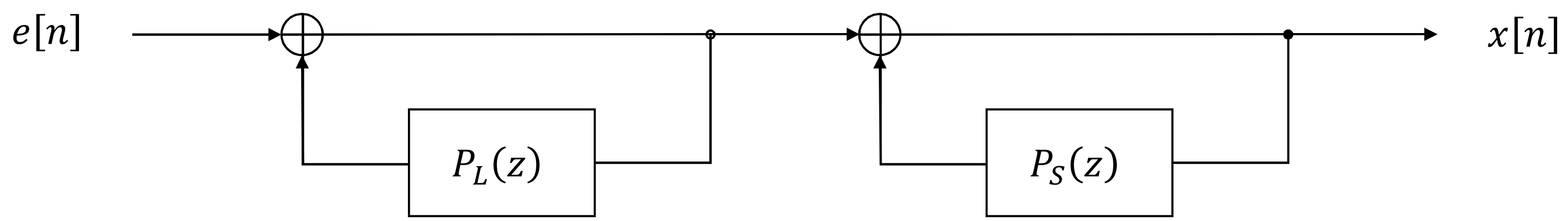
$\boldsymbol{x}$: waveform $\boldsymbol{h}$: acoustic and linguistic feature

$$p(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_{n=0}^{N-1} \underbrace{p(x[n] \mid x[0], \ldots, x[n-1], \boldsymbol{h})}_{\text{modeled by using CNN}}$$

  - Dilated causal convolution

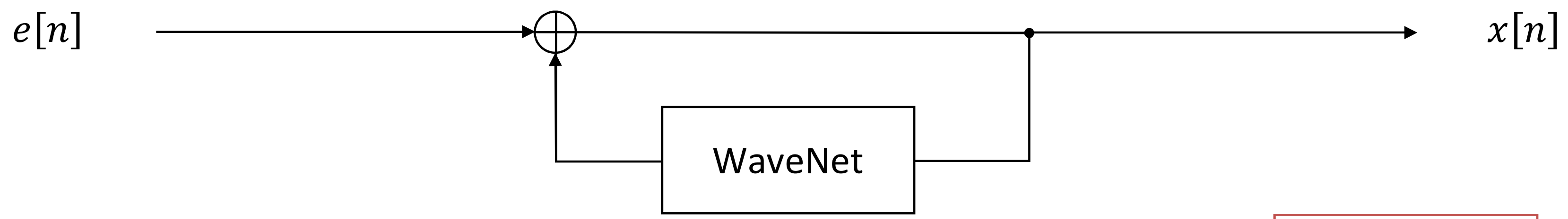


Speech signal generation model



Long-term predictor

Short-term predictor



Nonlinear Predictor

It works!

Famous words in speech technology (1980s)

"Every time I fire a linguist,
the performance of the speech recognizer goes up"
by Frederick Jelinek

"Every time I fire a speech technology researcher,
the performance of the speech synthesizer goes up"
by ????? ?????

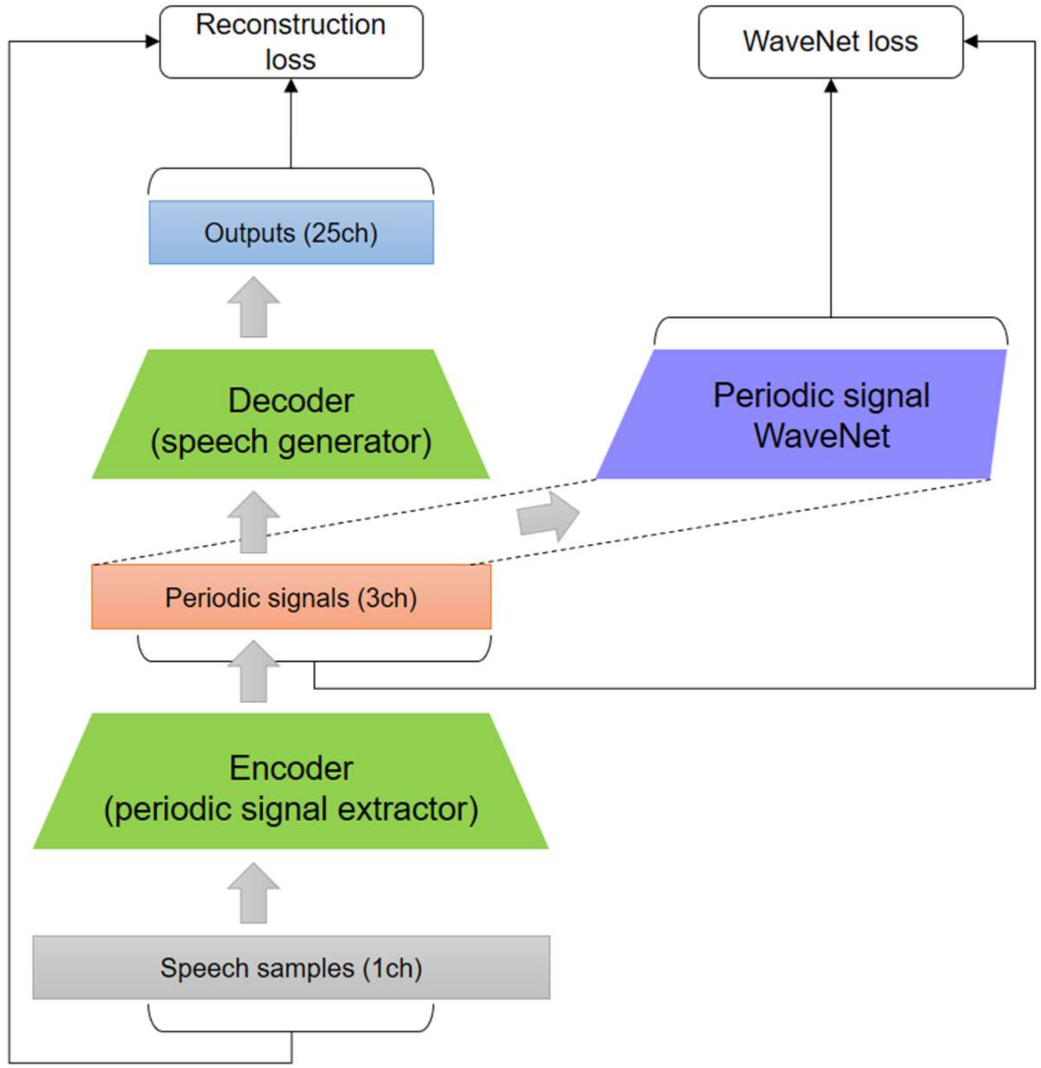
DNN variants for waveform modeling

- Autoregressive
  - WaveNet, SampleRNN, WaveRNN, …
- Normalizing flow
  - WaveGlow, Pallalel WaveNet, ClariNet, FloWaveNet, …
- Combining with source filter model
  - LPCNet, ExcitNet, GlotNet, LP-WaveNet, …
- Introducing signal processing technique
  - SubbandWaveNet, FFTNet, …

→ >>

DNN vocoder with periodic excitation [Oura '19]

- Autoencoder-type structure extracts 3 dimensional periodic signal
- Decoder generates periodic components and stochastic components
- WaveNet gives a constraints on the intermediate variable



Deep neural network approaches (4/6)

FFNN, LSTM

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\, p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

Feature extraction /waveform generation

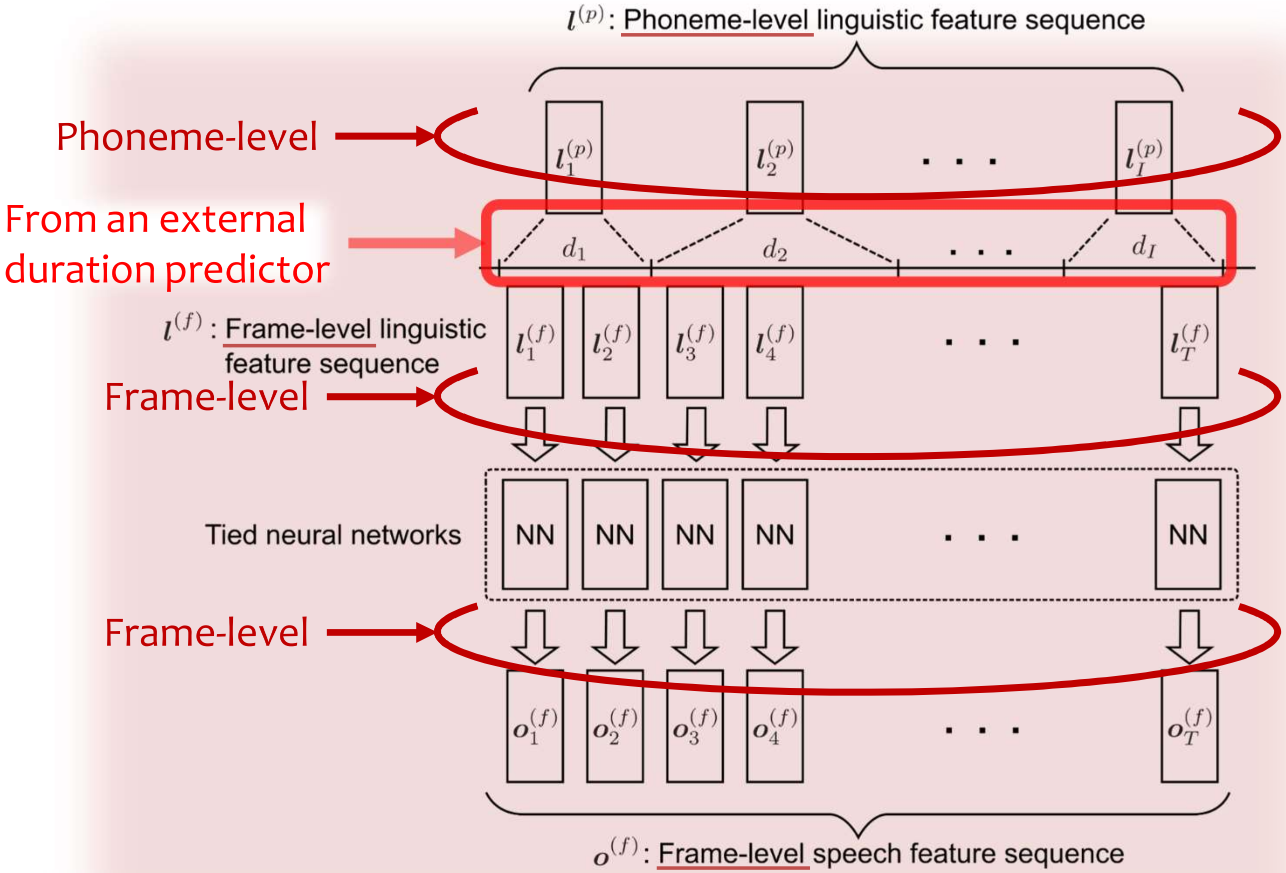
Acoustic model

Text analysis

- **HSMM:** duration model is included
- **FFNN, LSTM, WaveNet:** external duration predictor is required

Frame-by-frame conversion

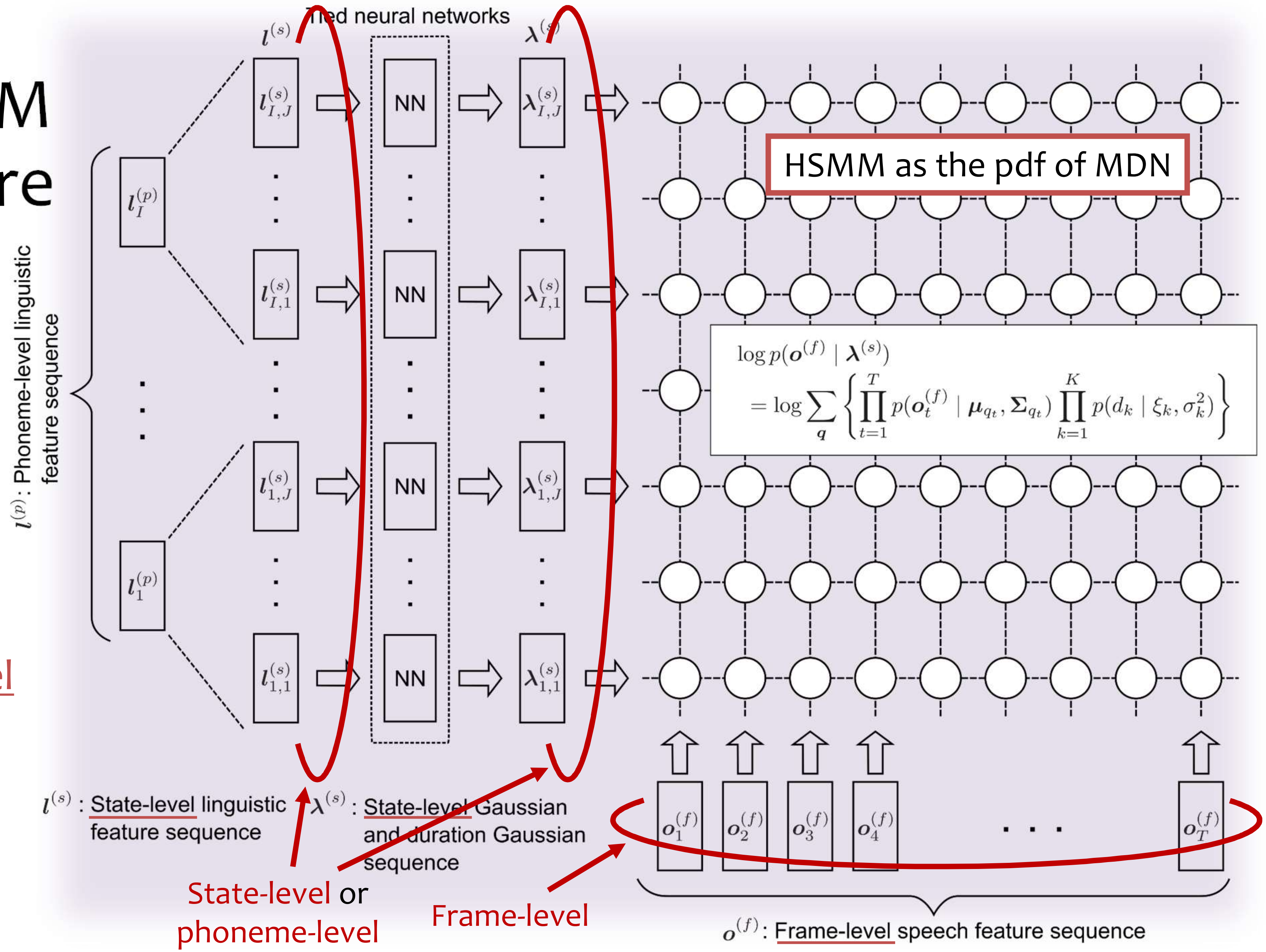
It needs an external duration predictor to determine phone durations



DNN-HSMM architecture

It runs at state-level
or phoneme-level

[Tokuda ’16]



Deep neural network approaches (5/6)

WaveNet vocoder
(autoregressive structure)

Tacotron, Char2Wav, DeepVoice, …
(attention mechanism)

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\, p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

Feature extraction
/waveform generation

Acoustic model

Text analysis

→

>>

Attention mechanism

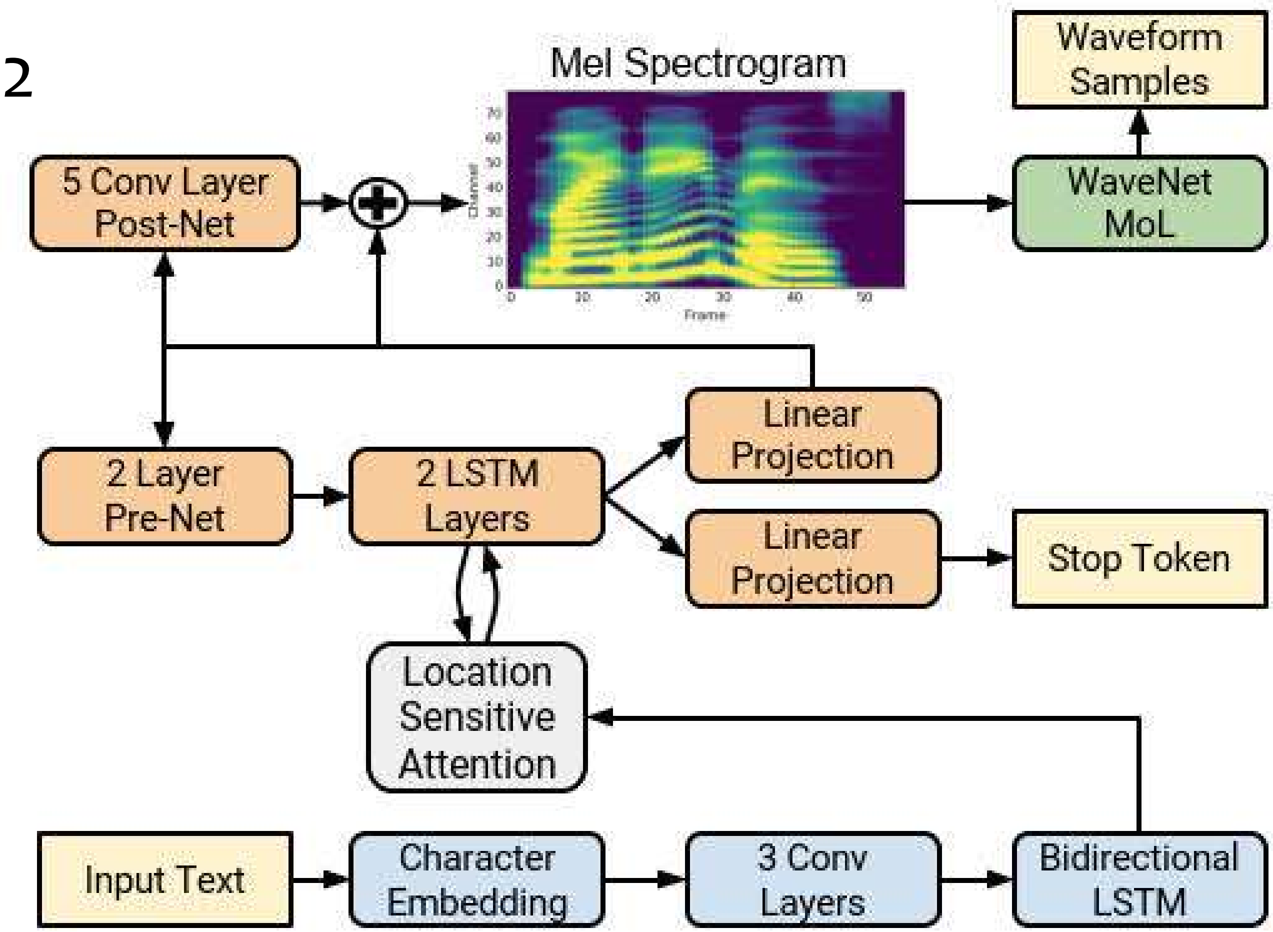
[from arXiv:1712.05884]

Tacotron 2

Decoder
Acoustic feature generation

Attention
mapping b/w
text and acoustic feature

Encoder
≅ text analysis



Deep neural network approaches (6/6)

WaveNet vocoder
(autoregressive structure)

Attention

$$\boxed{p(\boldsymbol{x}|\boldsymbol{w},\lambda) = p(\boldsymbol{x}|\boldsymbol{c})\, p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)\, P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)}$$

Feature extraction
/waveform generation

Acoustic model

Text analysis

←

Controlling intermediate variables in the hierarchical structure

- Language
  - Japanese, English, Chinese, …
- Dialect
- Pronunciation
- Pause
- Allophone
- Prosody
  - Accent, stress, tone, …
- Speaking style, emotional expression
- Emphasis
- Nonverbal, paralinguistic information
- Voice characteristics
  - Male, female, child, adult, eldarly
- Speech parameter
  - Fundamental frequency, volume, duration, aperiodic component, …

High level

Text analysis

Acoustic model

Low level

Vocoding

Singing synthesis with CNN+WaveNet

$$p(\boldsymbol{x}|\boldsymbol{w},\lambda) = \overbrace{p(\boldsymbol{x}|\boldsymbol{c})}^{\text{WaveNet vocoder}}\ \overbrace{p(\boldsymbol{c}|\boldsymbol{l},\lambda_A)}^{\text{CNN}}\ P(\boldsymbol{l}|\boldsymbol{w},\lambda_L)$$

Feature extraction /waveform generation

Acoustic model

Text analysis

Other DNN techniques and architectures

- GAN
- VAE/VQ-VAE
- Transformer (self attention)

Outline

- Statistical formulation of speech synthesis
- HMM-based speech synthesis
- Deep neural networks
- Evaluation / data & software tools
- Other related topics

Blizzard Challenge

- Performance of TTS system depends on the database
- Difficult to compare techniques themselves

“Blizzard Challenge”

Evaluating corpus-based speech synthesis

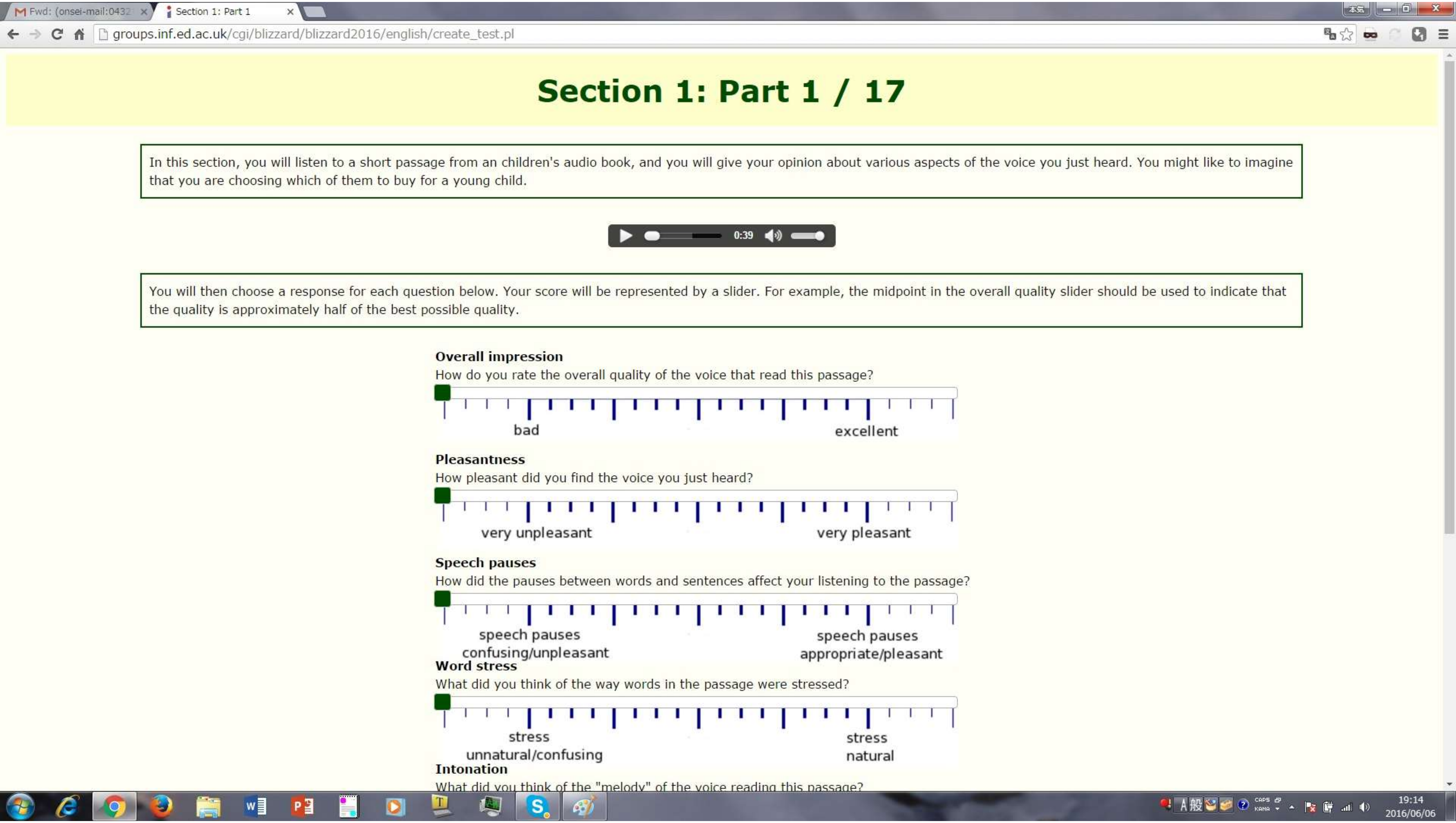
on **common datasets** [Black ’05]

Since 2005

Evaluation methodology

- Naturalness
  - Mean Opinion Score
- Speaker similarity
  - Degradation Mean Opinion Score
- Intelligibility (dictation of SUS, PCS, etc.)
  - Word accuracy

Not enough for spontaneous speech, audio book task, etc.


Fwd: (onsei-mail:0432
Section 1: Part 1
groups.inf.ed.ac.uk/cgi/blizzard/blizzard2016/english/create_test.pl
Section 1: Part 1 / 17
In this section, you will listen to a short passage from an children's audio book, and you will give your opinion about various aspects of the voice you just heard. You might like to imagine that you are choosing which of them to buy for a young child.
0:39
You will then choose a response for each question below. Your score will be represented by a slider. For example, the midpoint in the overall quality slider should be used to indicate that the quality is approximately half of the best possible quality.
Overall impression
How do you rate the overall quality of the voice that read this passage?
bad
excellent
Pleasantness
How pleasant did you find the voice you just heard?
very unpleasant
very pleasant
Speech pauses
How did the pauses between words and sentences affect your listening to the passage?
speech pauses confusing/unpleasant
speech pauses appropriate/pleasant
Word stress
What did you think of the way words in the passage were stressed?
stress unnatural/confusing
stress natural
Intonation
What did you think of the "melody" of the voice reading this passage?
19:14
2016/06/06

Common datasets for speech synthesis

- ELRA http://www.elra.info/
- ELDA http://www.elda.org/
- LDC https://www.ldc.upenn.edu/
- OpenSLR http://www.openslr.org/

- ARCTIC
- VCTK
- LibriTTS, …

Not so many because it needs studio-quality recordings

Software tools

- ISCA SynSig https://www.synsig.org/index.php/Software
- ISCA SCOOT https://www.isca-speech.org/iscaweb/index.php/scoot

Software tools

MMDAgent
- Toolkit for building voice interaction systems -

Toolkit for building voice interaction systems

hts_engine
Speech synthesis engine

HTS
Training toolkit

Open JTalk
Japanese TTS system

SPTK
Speech signal processing toolkit

Sinsy
Singing synthesis system

Takashi Masuko, Noboru Miyazaki, Kazuhito Koishida, Takayoshi Yoshimura,
Heiga Zen, Junichi Yamagishi, Keiichiro Oura, Akinobu Lee and others contributed

Outline

- Statistical formulation of speech synthesis
- HMM-based speech synthesis
- Deep neural networks
- Evaluation / data & software tools
- Other related topics
    - Text normalization
    - Voce conversion
    - Speech coding
    - Anti-spoofing
    - Physical simulation

Text normalization

- Text normalization is excluded from the end-to-end systems
- Still rule-based approach is the mainstream

- It would be included in the end-to-end process in the near future

Voice conversion

- Close relationship to speech synthesis
- DNN-approach has emerged also in voice conversion research
- Realtime application is essential
- Realtime (or low-latency) prosody conversion is a challenging problem

Speech coding

- WaveNet and other waveform modeling approaches seems to bring a revolution to speech coding.

- WaveNet based low rate speech coding [Kleijn '18]
- A Real-Time Wideband Neural Vocoder at 1.6 kb/s Using LPCNet [Valin '19]
- Low Bit-rate Speech Coding with VQ-VAE and WaveNet [Garbacea '19]
- High-quality speech coding with sample RNN [Klejsa '19]
- WaveNet-based zero-delay lossless speech coding [Yoshimura '18]
- Wavenet-based delay-free ADPCM Speech Coding [Yoshimura '19]

Imposture using speech synthesis

- Fear for spoofing with speech synthesis
  - On the security of HMM-based speaker verification systems against imposture using synthetic speech [Masuko '99]
- Detecting synthesized speech
  - A robust speaker verification system against imposture using an HMM-based speech synthesis system [Satoh '01]
- ASVspoof 2015
  - The First Automatic Speaker Verification Spoofing and Countermeasures Challenge

Physical simulation vs Deep neural network

- In the future, techniques for measuring dynamics of vocal tract will be significantly progressed.
- Also, techniques for simulating speech production system will be progressed.

↓

will it be possible to generate natural-sounding speech
based on the physical simulation approach?

- Advantage: realistic constraints, lower dimensional representation

→ latent representation in DNN-based system?

Summary

Statistical approach to speech synthesis

- Now it has reached at the level that we cannot tell the difference between human and machine
- Still we have a lot of problems to be solved →
- More flexibility and controllability for realizing diversity of speech

Let us enjoy speech synthesis research!

Thank you!

>>

Speech synthesis in the future

- Spoken dialog system
- S
- Support for people with disabilities
- CALL
- Content creation

Cross-lingual/multilingual

Flexibility/diversity

rce language

Editor design

Common data

Let us enjoy speech synthesis research for realizing diversity of speech!

Thank you!

Special thanks

- **Supervisors**: Satoshi Imai, Tadashi Kitamura, Takao Kobayashi
- **Colleagues & students**: Takashi Masuko, Noboru Miyazaki, Takayoshi Yoshimura, Shinji Sako, Masatsune Tamura, Junichi Yamagishi, Tomoki Toda, Heiga Zen, Kazuhito Koishida, Tetsuya Yamada, Nobuaki Mizutani, Ryuta Terashima, Akinobu Lee, Keiichiro Oura, Keijiro Saino, Kenichi Nakamura, Yi-Jian Wu, Ling-Hui Chen, Shifeng Pan, Yoshihiko Nankaku , Ranniery Maia, Sayaka Shiota, Chiyomi Miyajima, Kei Hashimoto, Shinji Takaki, Kazuhiro Nakamura, Kei Sawada, Takenori Yoshimura, Daisuke Yamamoto, …
- **Collaborators and advisors**: Junichi Takami, Naoto Iwahashi, Mike Schuster, Satoshi Nakamura, Frank Soong, Mchiael Picheny, Simon King, Steve Young, Mari Ostendorf, Alan Black, Alex Acero, Bill Byrne, Phil Woodland, Thomas Hain, Phil Garner, Masataka Goto, Shigeru Katagiri, Hideki Kenmochi, Kazuya Takeda, Tatsuya Kawahara, Sadaoki Furui, Seiichi Nakagawa, Keikichi Hirose, Tetsunori Kobayashi, Miko Kurimo, Shigeki Sagayama, Kiyohiro Shikano, Hisashi Kawai, Nobuyuki Nishizawa, Minoru Tsuzaki, Yoichi Yamashita, Nobuaki Minematsu, Mat Shannon, Mark Gales, Kai Yu, John Dines, …

in random order. I am sorry but I may have missed many…

References

- K. Tokuda, T. Kobayashi, S. Shiomoto, S. Imai, “Adaptive filtering based on cepstral representation —adaptive cepstral analysis of speech,” ICASSP 1990a.
- K. Tokuda, T. Kobayashi, S. Imai, “Generalized cepstral analysis of speech — unified approach to LPC and cepstral method,” ISCSLP 1990b.
- T. Fukada, K. Tokuda, T. Kobayashi, S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” ICSSP 1992.
- K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, “Mel-generalized cepstral analysis —a unified approach to speech spectral estimation,” ICSLP 1994.
- K. Tokuda, T. Kobayashi, S. Imai, “Speech parameter generation from HMM using dynamic features,” ICASSP 1995a.
- K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, S. Imai, “An algorithm for speech parameter geeration from continuous mixture HMMs with dynamic features,” EUROSPEECH 1995b.

Only from HTS group

- T. Masuko, K. Tokuda, T. Kobayashi, S. Imai, “HMM-based speech synthesis with various voice characteristics,” ASA/ASJ Joint Meeting 1996.
- T. Masuko, K. Tokuda, T. Kobayashi, S Imai, “Voice characteristics conversion for HMM-based speech synthesis system,” ICASSP 1997.
- T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” EUROSPEECH 1997.
- T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, K. Tokuda, “Text-to-visual speech synthesis based on parameter generation from HMM,” ICASSP 1998.
- M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, ”Speaker adaptation for HMM-based speech synthesis system using MLLR,” SSW 1998.
- T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” EUROSPEECH 1999.
- K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” ICASSP 2000.

SSW: ISCA Speech Synthesis Workshop

- T. Masuko, T. Hitotsumatsu, K. Tokuda, T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," EUROSPEECH 1999.
- T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," EUROSPEECH 1999.
- T. Masuko, K. Tokuda, T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," ICSLP/INTERSPEECH 2000.
- S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "HMM-based text-to-audio-visual speech synthesis," ICSLP/INTERSPEECH 2000.
- M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," EUROSPEECH 2001.
- T. Satoh, T. Masuko, T. Kobayashi, K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," EUROSPEECH 2001.
- K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Eigenvoices for HMM-based speech synthesis," ICSLP 2002.

- J. Yamagishi, T. Masuko, K. Tokuda, T. Kobayashi, “A training method for average voice model based on shared decision tree context clustering and speaker adaptive training,” ICASSP 2003.
- K. Tokuda, H. Zen, T. Kitamura, “Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features,” EUROSPEECH 2003.
- H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Hidden semi-Markov model based speech synthesis,” ICSLP 2004.
- T. Toda, A. Black, K. Tokuda, “Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter,” ICASSP 2005.
- T. Toda, K. Tokuda, “Speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” INTERSPEECH 2005.
- A. Black, K. Tokuda, “The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets,” INTERSPEECH 2005.
- K. Saino, H. Zen, Y. Nankaku, A. Lee, K. Tokuda, “HMM-based singing voice synthesis system,” Interspeech 2006.

- R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, “An excitation model for HMM-based speech synthesis based on residual modeling,” SSW 2007.
- K. Oura, Y. Nankaku, T. Toda, K. Tokuda, R. Maia, S. Sakai, S. Nakamura, “Simultaneous Acoustic, Prosodic, and Phrasing Model Training for TTS Conversion Systems,” ISCSLP 2008.
- K. Hashimoto, H. Zen, Y. Nankaku, K. Tokuda, “A Bayesian approach to HMM-based speech synthesis,” ICASSP 2009.
- K. Kazumi, Y. Nankaku, K. Tokuda, “Factor analyzed voice models for HMM-based speech synthesis,” ICASSP 2010.
- K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, K. Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System – Sinsy,” SSW 2010.
- K. Nakamura, K. Hashimoto, Y. Nankaku, K. Tokuda, “Integration of Acoustic Modeling and Mel-cepstral analysis for HMM-based Speech Synthesis,” ICASSP 2013.
- K. Tokuda, H. Zen, ``Directly modeling voiced and unvoiced components in speech waveforms by neural networks,'' ICASSP 2016.

- K. Tokuda, K. Hashimoto, K. Oura, Y. Nankaku, “Temporal modeling in neural network based statistical parametric speech synthesis,” SSW 2016.
- T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, “Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1173-1180, July, 2018.
- J. Niwa, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, “Statistical voice conversion based on WaveNet,” ICASSP 2018.
- T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, “WaveNet-based zero-delay lossless speech coding,” SLT 2018.
- T Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, ``Speaker-dependent WaveNet-based delay-free ADPCM speech coding,'' ICASSP 2019.
- S. Takaki, Y. Nankaku, K. Tokuda, “Spectral modeling with contextual additive structure for HMM-based speech synthesis,” SSW 2010.