統計的音声合成の展開と展望

徳田恵一（名古屋工業大学）

2019年12月6日　音声言語シンポジウム

音声合成のアプローチ（超簡略版）

- ルールベース, フォルマント音声合成 (~'90s)
  人手によるルールに基づいて各音素の素片を構築

- コーパスベース, 波形接続型音声合成 ('90s~)
  音声データベースから音声素片（音響特徴あるいは波形）を接続して合成
  - 単一インベントリ：ダイフォン音声合成
  - 複数インベントリ：単位選択型音声合成　→

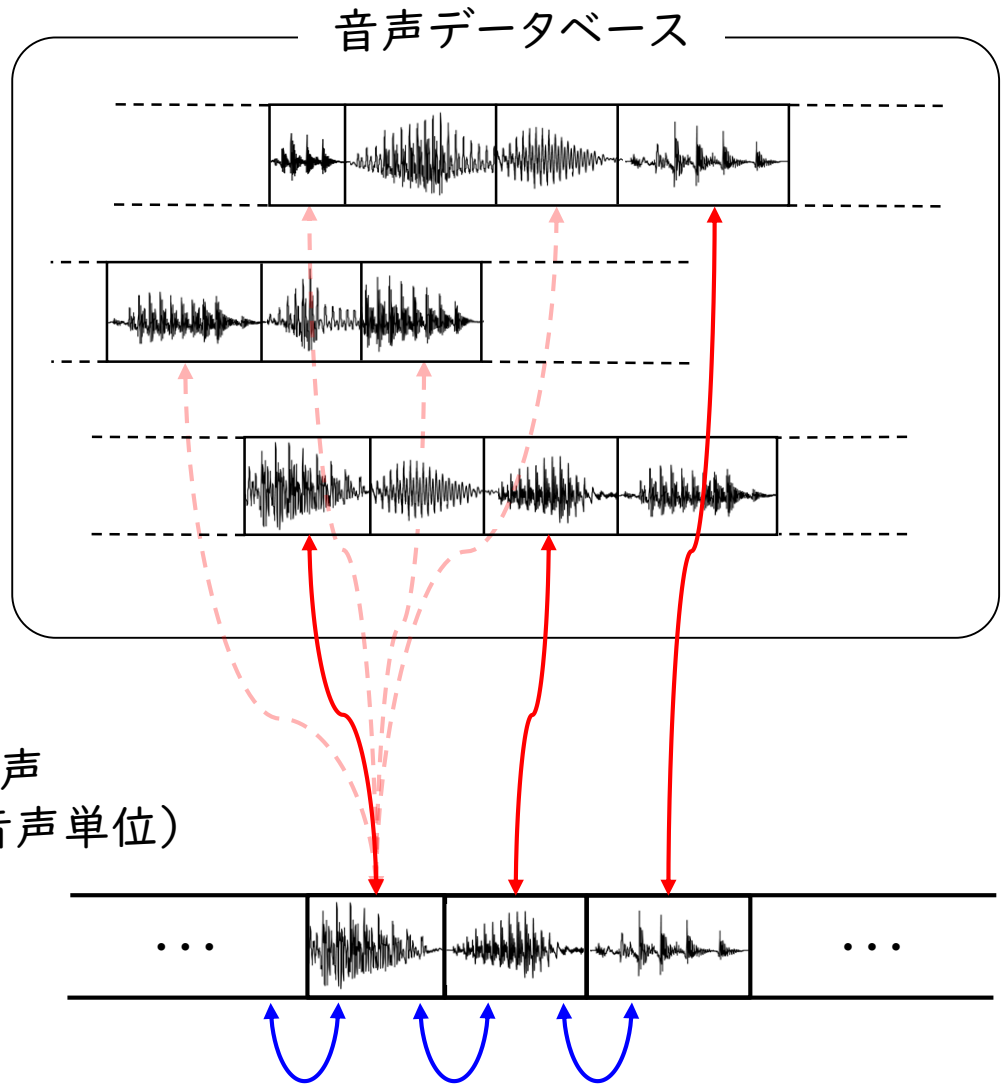
- コーパスベース, 統計的音声合成 (late '90s~)
  ソースフィルタモデル + 統計的音響モデル
  - HMM (hidden Markov model) (1995~)　→
  - DNN (deep neural networks) (2013~)
  - WaveNet (2016~)

主としてこちらを
やってきました

>>

統計的音声合成へのパラダイムシフト

単位選択型音声合成



運が良いとき:

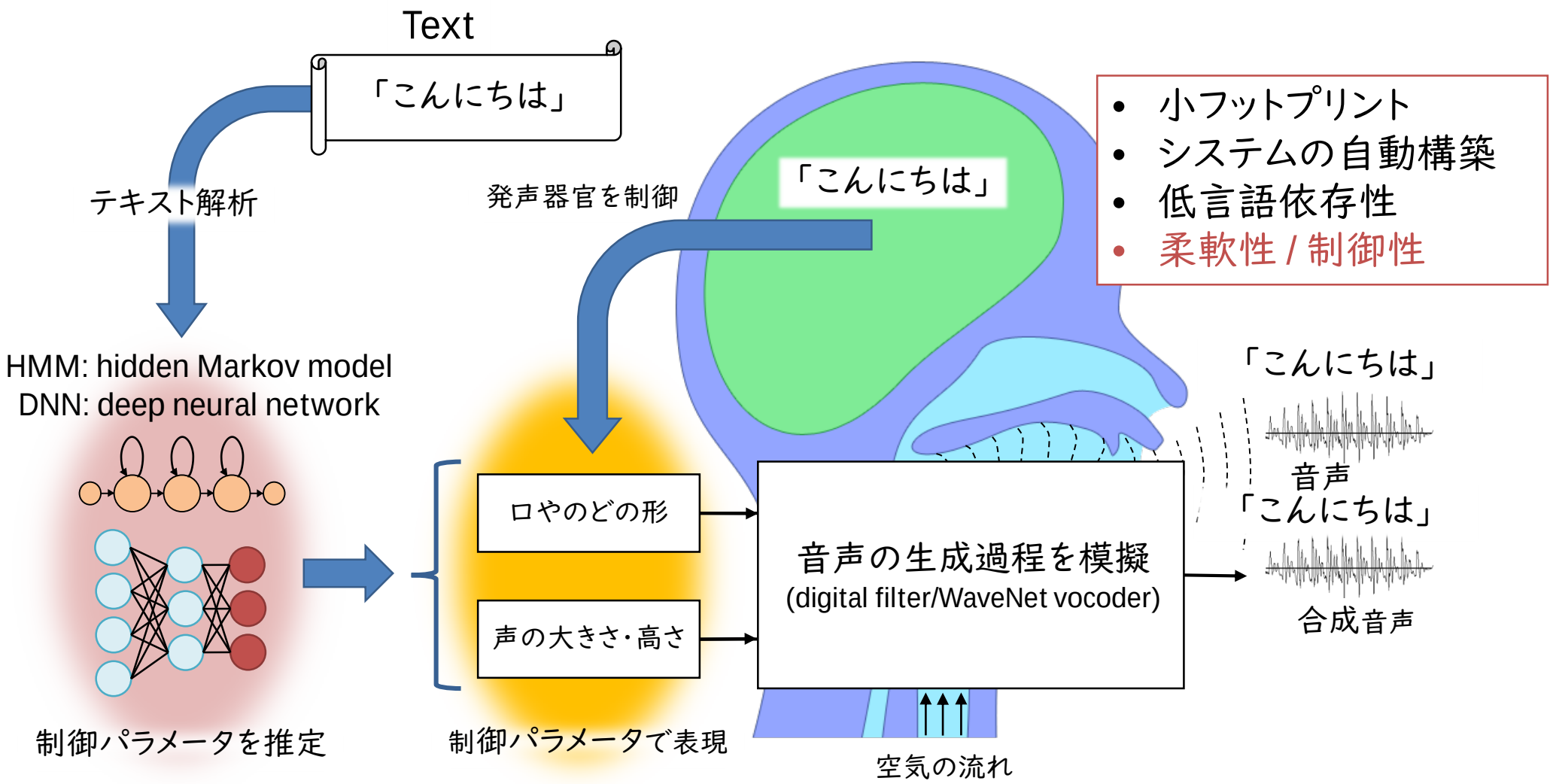
運が悪いとき:

ターゲットコスト

接続コスト

ランタイムに動的計画法により
総コストを最小化

統計的音声合成



- 小フットプリント
- システムの自動構築
- 低言語依存性
- 柔軟性 / 制御性

長所・短所

WaveNet等のニューラルボコーダにより解決

単位選択型音声合成	統計的音声合成	
波形接続 → 高い自然性	ボコーダ → バジー＆こもった音	☹
接続歪（当たり外れあり）	滑らか	
必要データ量比較的大？	必要データ量比較的小？	
フットプリント大	フットプリント小	
柔軟性小 → 発話スタイル、感情表現等が固定	柔軟性大 → 話者適応、発話スタイル補間など	☺

統計的音声合成へのパラダイムシフト

あらまし

- 音声合成の統計的定式化
- HMM音声合成
- DNN音声合成
- 評価 / データ&ソフトウェアツール
- その他の関連トピックス

あらまし

- 音声合成の統計的定式化
- HMM音声合成
- DNN音声合成
- 評価 / データ&ソフトウェアツール
- その他の関連トピックス

音声合成の基本問題

テキストとそれに対応する音声波形の組の集合があるとき，
任意に与えられたテキストに対応する音声波形を求めよ。

- $\boldsymbol{W}$: テキスト
- $\boldsymbol{X}$: 音声波形

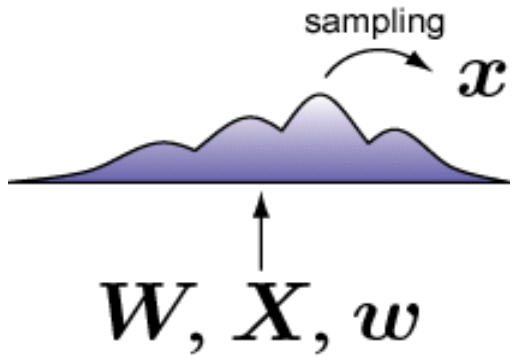
データベース

既知

- $\boldsymbol{w}$: 任意のテキスト $(\boldsymbol{w} \notin \boldsymbol{W})$
- $\boldsymbol{x}$: 合成音声波形

? ←

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



音声合成の統計的定式化 (1/4)

- 予測分布の推定は簡単ではない
  → 生成モデルに基づいたパラメトリック表現を導入

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

λ: モデルパラメータ

- 積分は容易ではない
  → $p(\lambda|\boldsymbol{X},\boldsymbol{W})$ の最大化で積分を近似

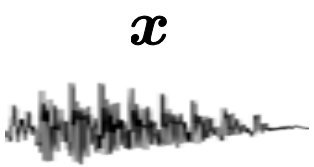
$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\lambda}|\boldsymbol{W},\boldsymbol{X}) \leftarrow \text{学習}$$

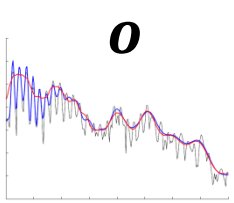
$$\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w},\hat{\lambda}) \leftarrow \text{生成}$$

音声合成の統計的定式化 (2/4)

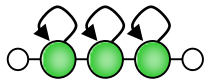
- 通常，生成モデルは部分モデルに分解される

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





特徴抽出・波形生成
（ボコーダ）



音響モデル

テキスト解析

$\boldsymbol{o}$: 音声波形 $\boldsymbol{x}$ のパラメトリック表現（音響特徴）

$\boldsymbol{l}$: テキスト $\boldsymbol{w}$ の 言語特徴（ラベル）

$\lambda = \{\lambda_A, \lambda_L\}$: 生成モデルのパラメータ

λ_A: 音響モデルのパラメータ

λ_L: テキスト解析部のパラメータ

←

>>

言語的特徴（ラベル/コンテキスト）

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.

⋮

(→)

+フレームレベルの継続長・位置情報

+発話スタイル, 感情表現等
（このようなタグ・ラベルがある場合）

音声合成の統計的定式化 (3/4)

- 生成モデルを部分モデルに分解すると

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\lambda}|\boldsymbol{W}, \boldsymbol{X}) \leftarrow \text{学習}$$

$$\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{w}, \hat{\lambda}) \leftarrow \text{生成}$$

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

↑ 学習

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

音声合成の統計的定式化 (4/4)

- 積分と総和の全体最大化は困難
  → 積分と総和の逐次最大化で近似 ← ←

$\hat{\lambda}_L$:事前学習されたテキスト解析モジュールのパラメータ

$\hat{\boldsymbol{O}} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{X}|\boldsymbol{O})$ ← 音響特徴の抽出

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

$\hat{\lambda}_A = \arg\max_{\lambda_A} p(\hat{\boldsymbol{O}}|\hat{\boldsymbol{L}}, \lambda_A)p(\lambda_A)$ ← 音響モデル学習

⬆ 学習

$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} P(\boldsymbol{l}|\boldsymbol{w}, \hat{\lambda}_L)$ ← テキスト解析

⬇ 合成

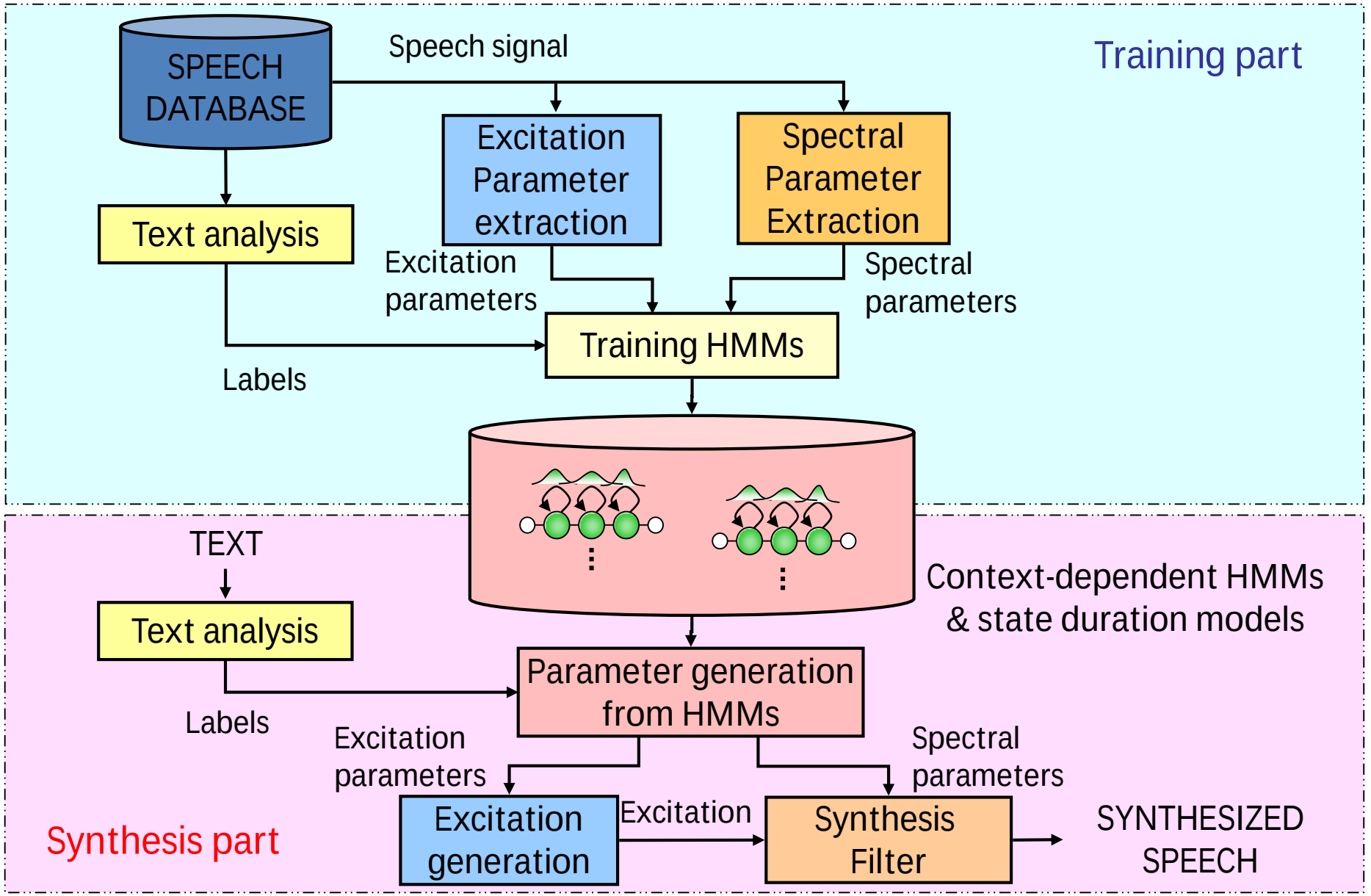
$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{c}} p(\boldsymbol{o}|\hat{\boldsymbol{l}}, \hat{\lambda}_A)$ ← 音響特徴生成

$\boldsymbol{x} \sim p(\boldsymbol{x}|\hat{\boldsymbol{o}})$ ← 音声波形生成

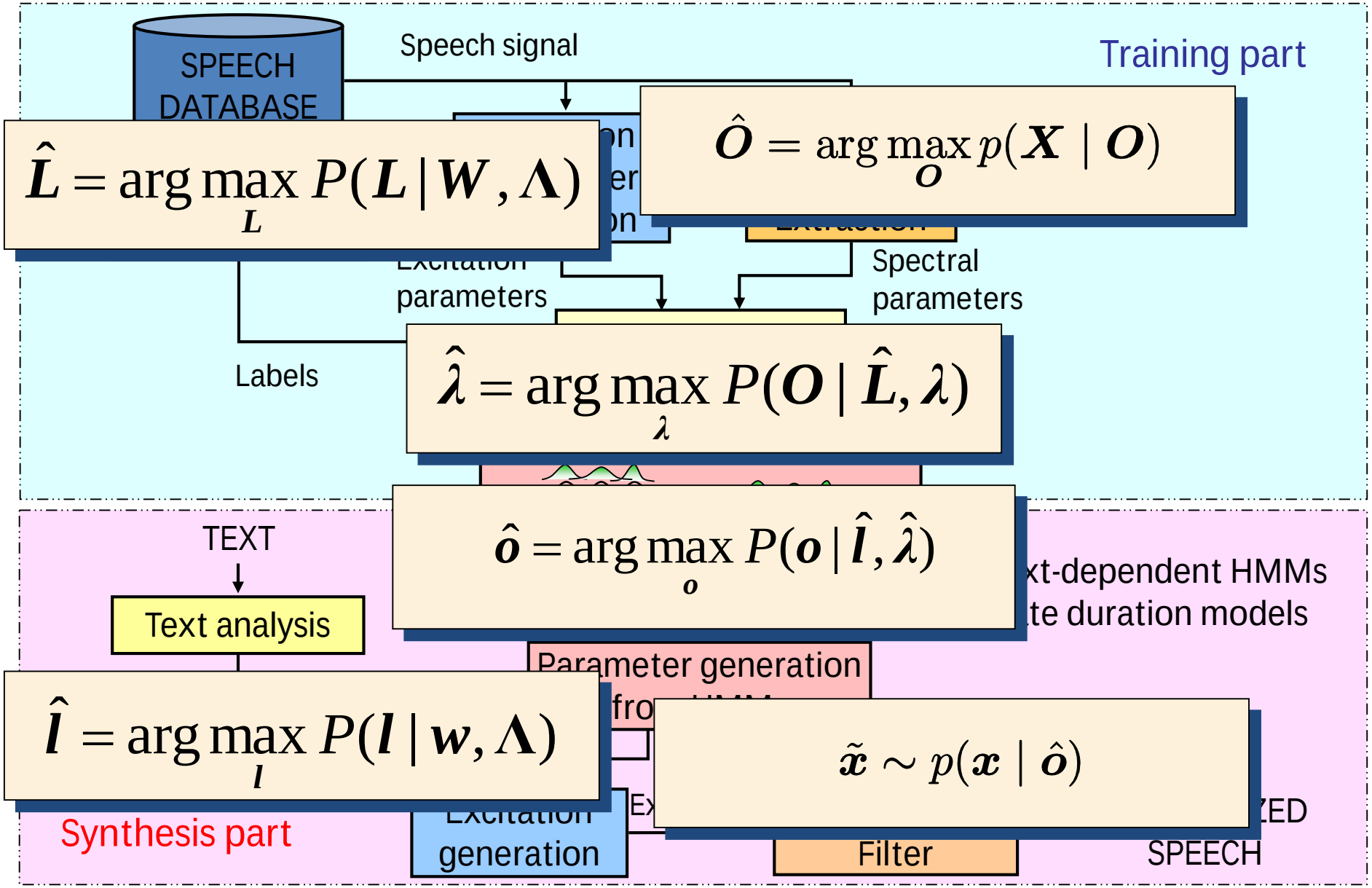
あらまし

- 音声合成の統計的定式化
- HMM音声合成
- DNN音声合成
- 評価 / データ&ソフトウェアツール
- その他の関連トピックス

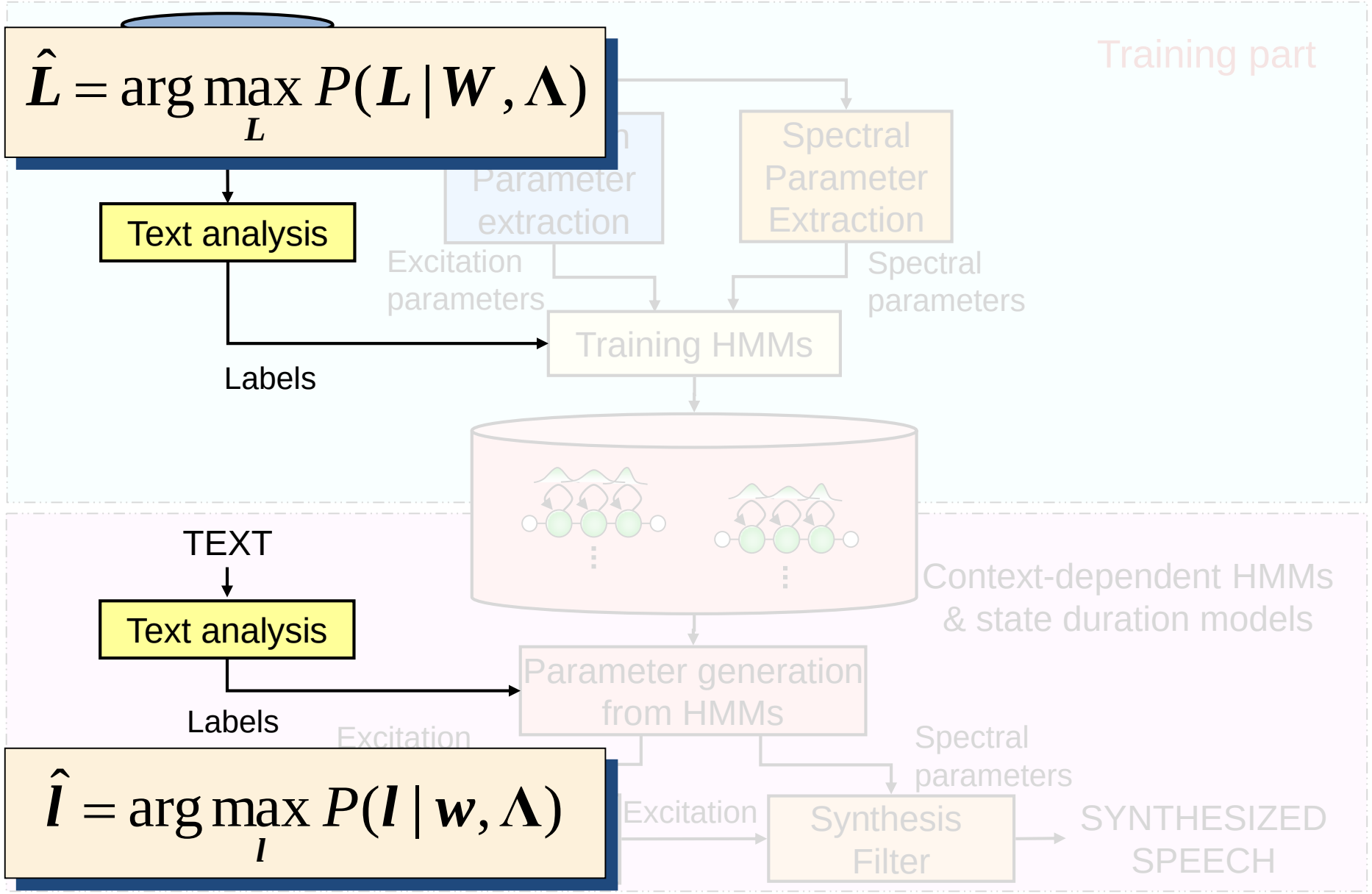
HMM音声合成の枠組み



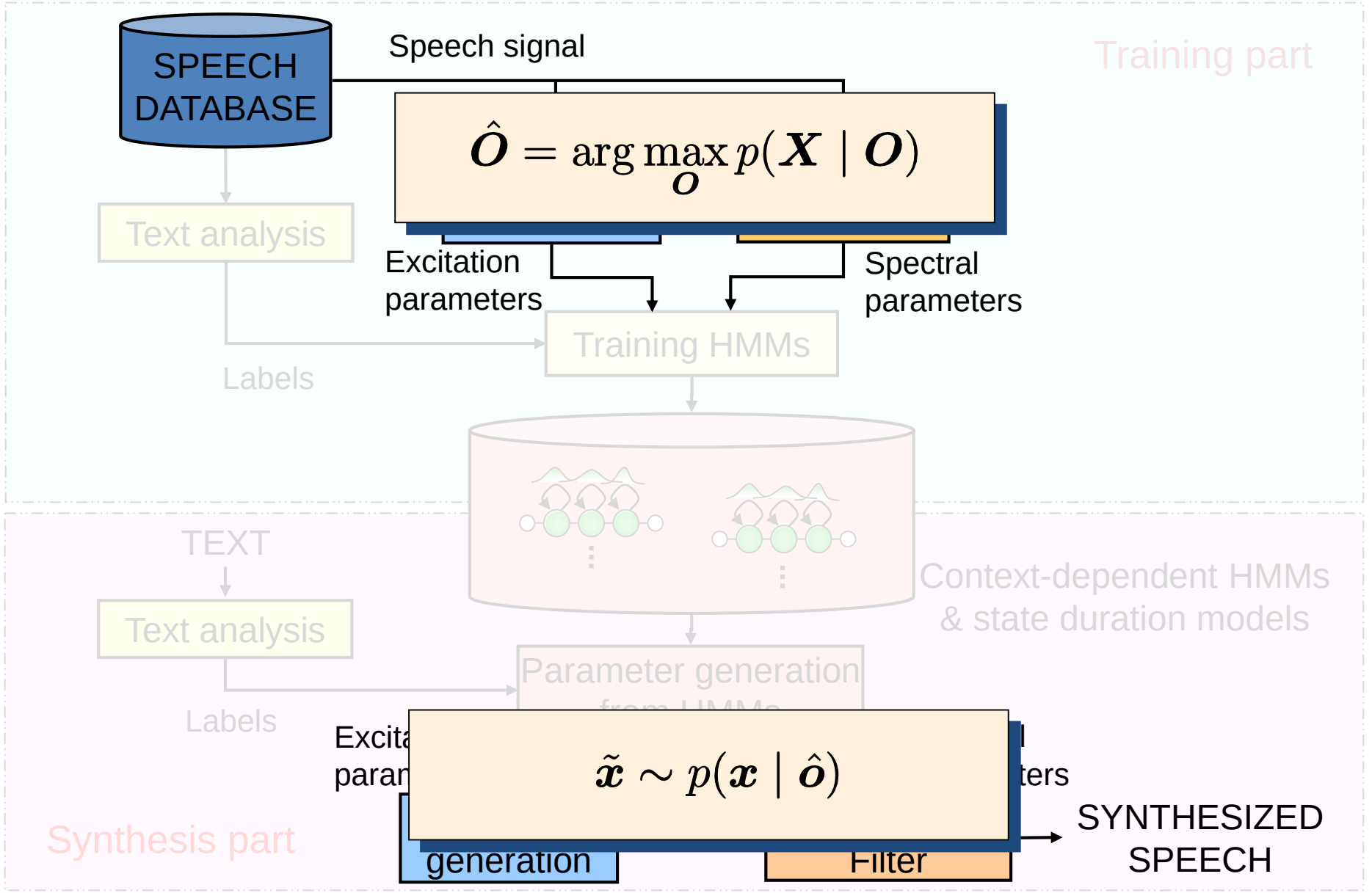
HMM音声合成の枠組み



HMM音声合成の枠組み



HMM音声合成の枠組み



メルケプストラムに基づくスペクトル分析

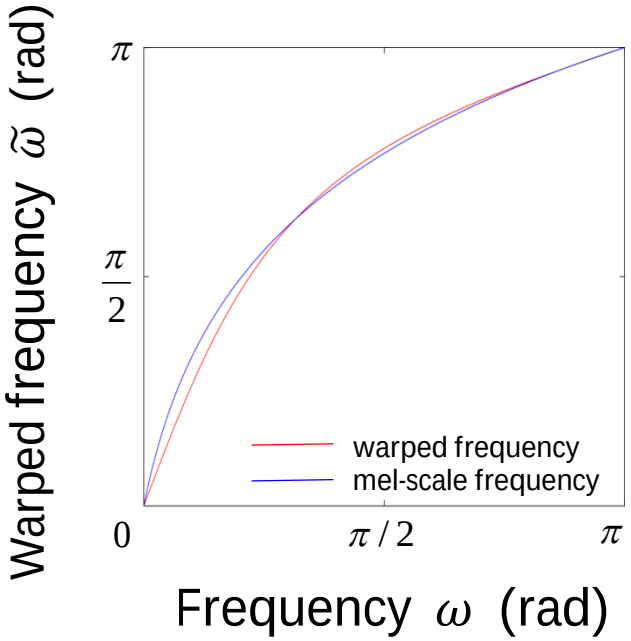
$$H\left(e^{j\omega}\right) = \exp \sum_{m=0}^{M} c(m)\, e^{-j\tilde{\omega}m}, \qquad e^{-j\tilde{\omega}} = \frac{e^{-j\omega} - \alpha}{1 - \alpha e^{-j\omega}}$$

$\boldsymbol{c} = [c(0), c(1), \ldots, c(M)]^{\mathrm{T}}$ ← メルケプストラム

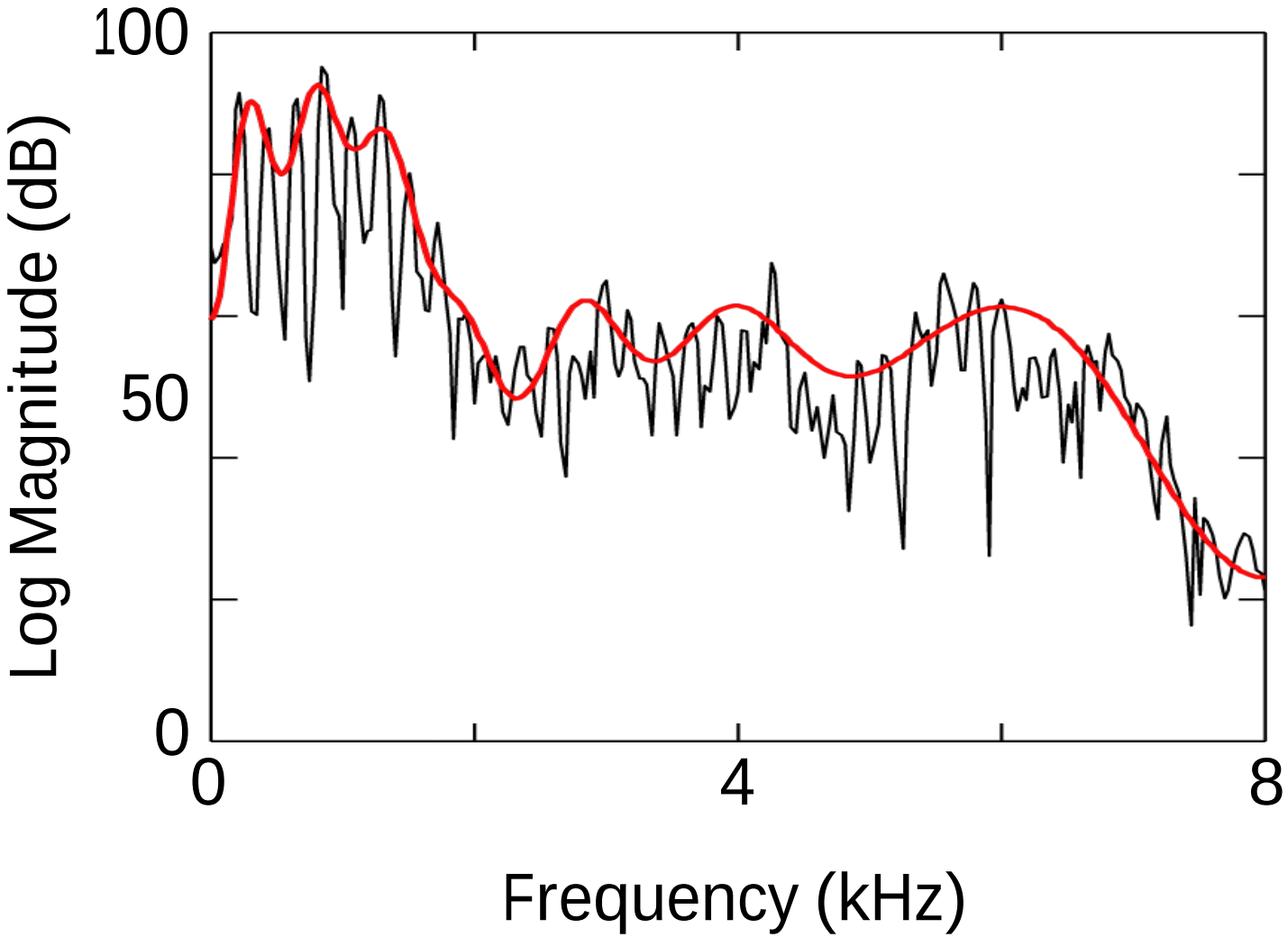
メルケプストラムの最尤推定:

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} p(\boldsymbol{x}|\boldsymbol{c})$$ ←

音声波形 $\boldsymbol{x}$ がガウス過程のとき
$p(\boldsymbol{x}|\boldsymbol{c})$ は $\boldsymbol{c}$ に関して凸 [Fuka92]

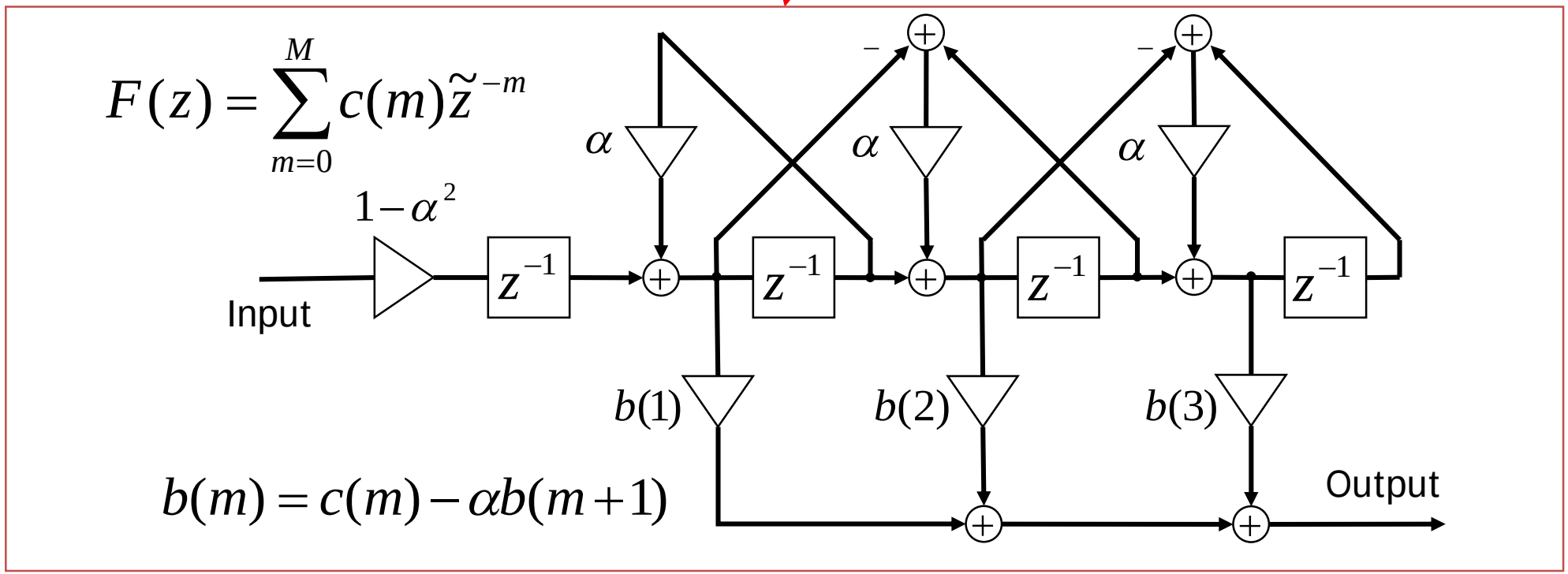


スペクトル推定例



MLSA フィルター (1/2) [Fukada '92]

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

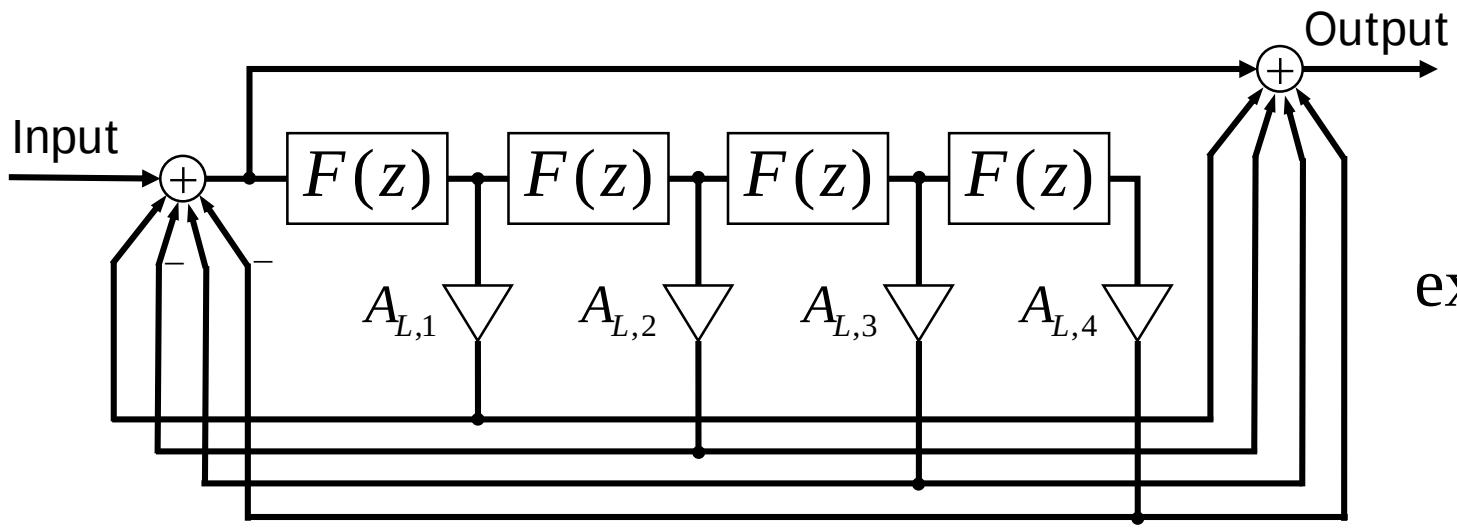


MLSA フィルター (2/2) [Fukada '92]

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

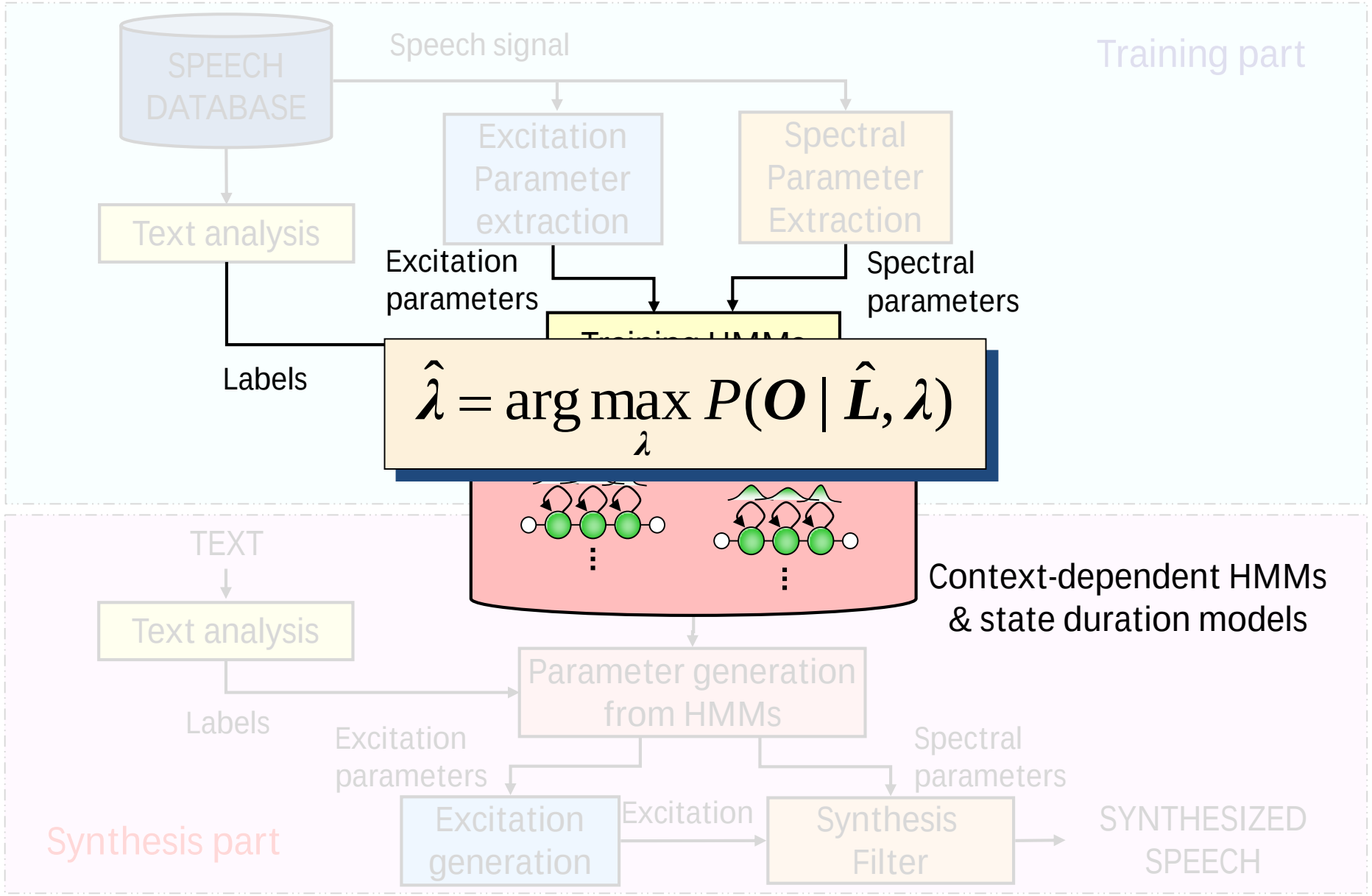
$$\exp x \cong \frac{1 + \sum_{l=1}^{L} A_{L,l} x^{l}}{1 + \sum_{l=1}^{L} A_{L,l} (-x)^{l}}$$

- 近似誤差 < 0.24dB
- O(8M) 積和/サンプル
- 安定性の保証



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

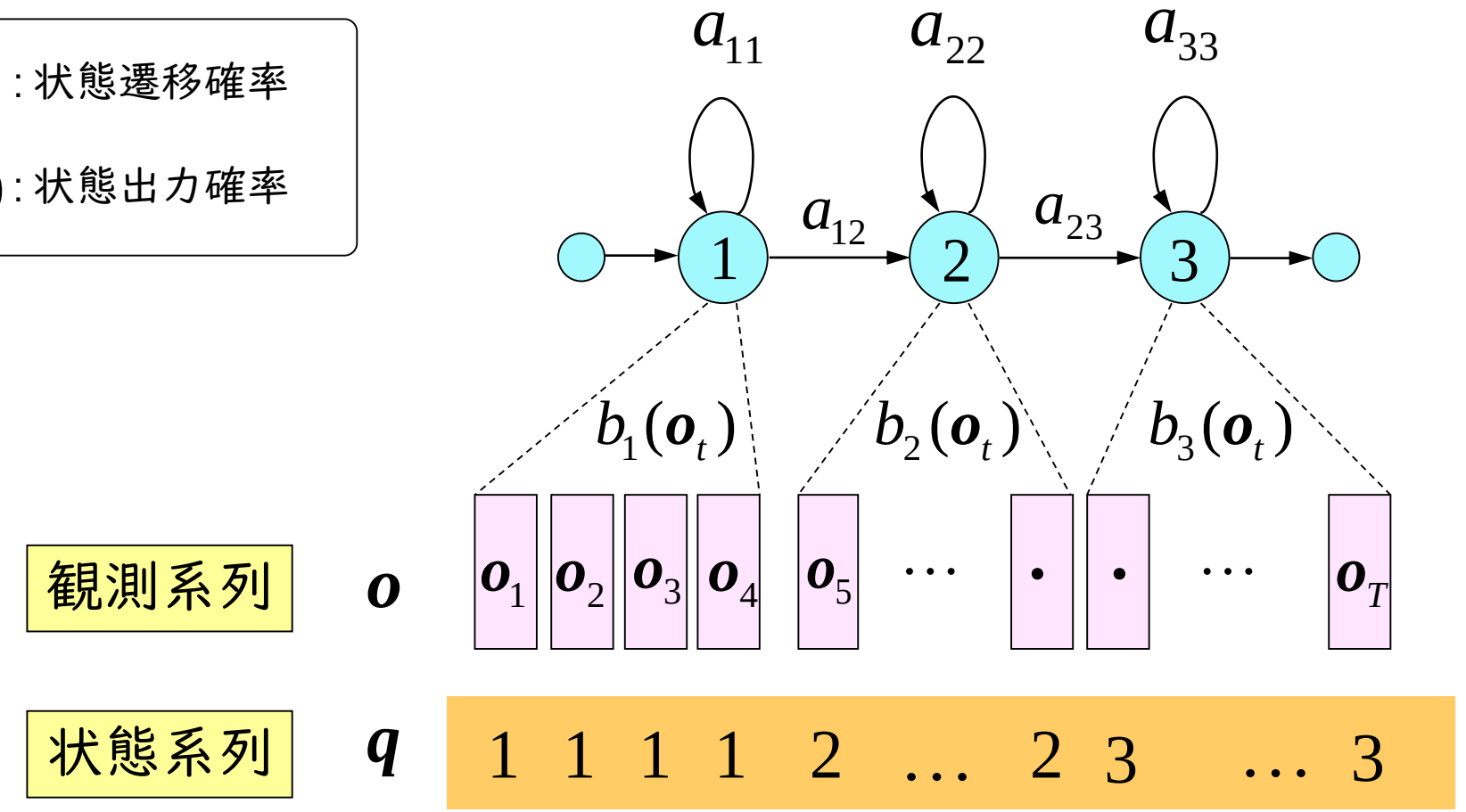
HMM音声合成の枠組み



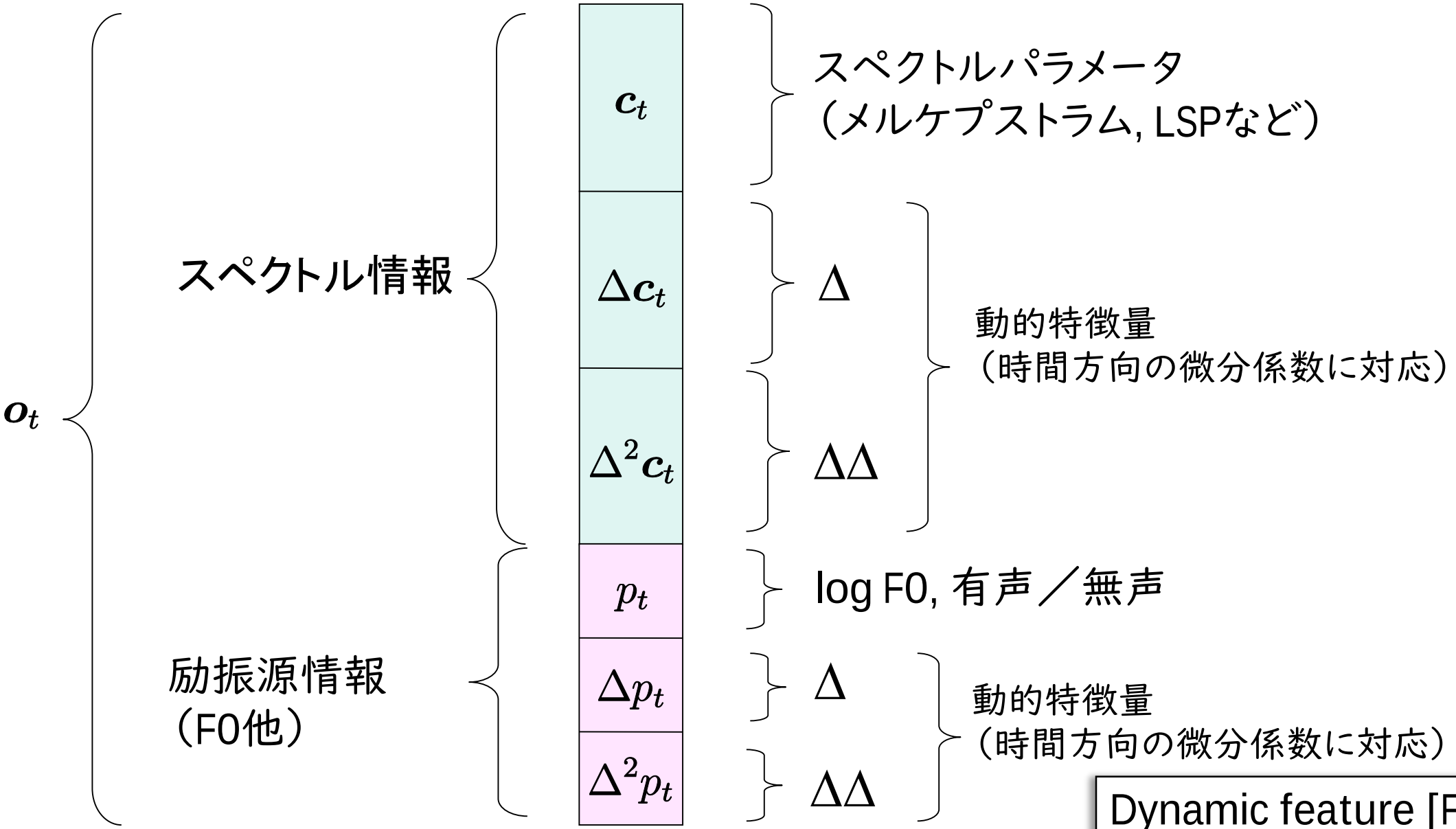
隠れマルコフモデル (HMM)

a_{ij} : 状態遷移確率

$b_q(\boldsymbol{o}_t)$: 状態出力確率



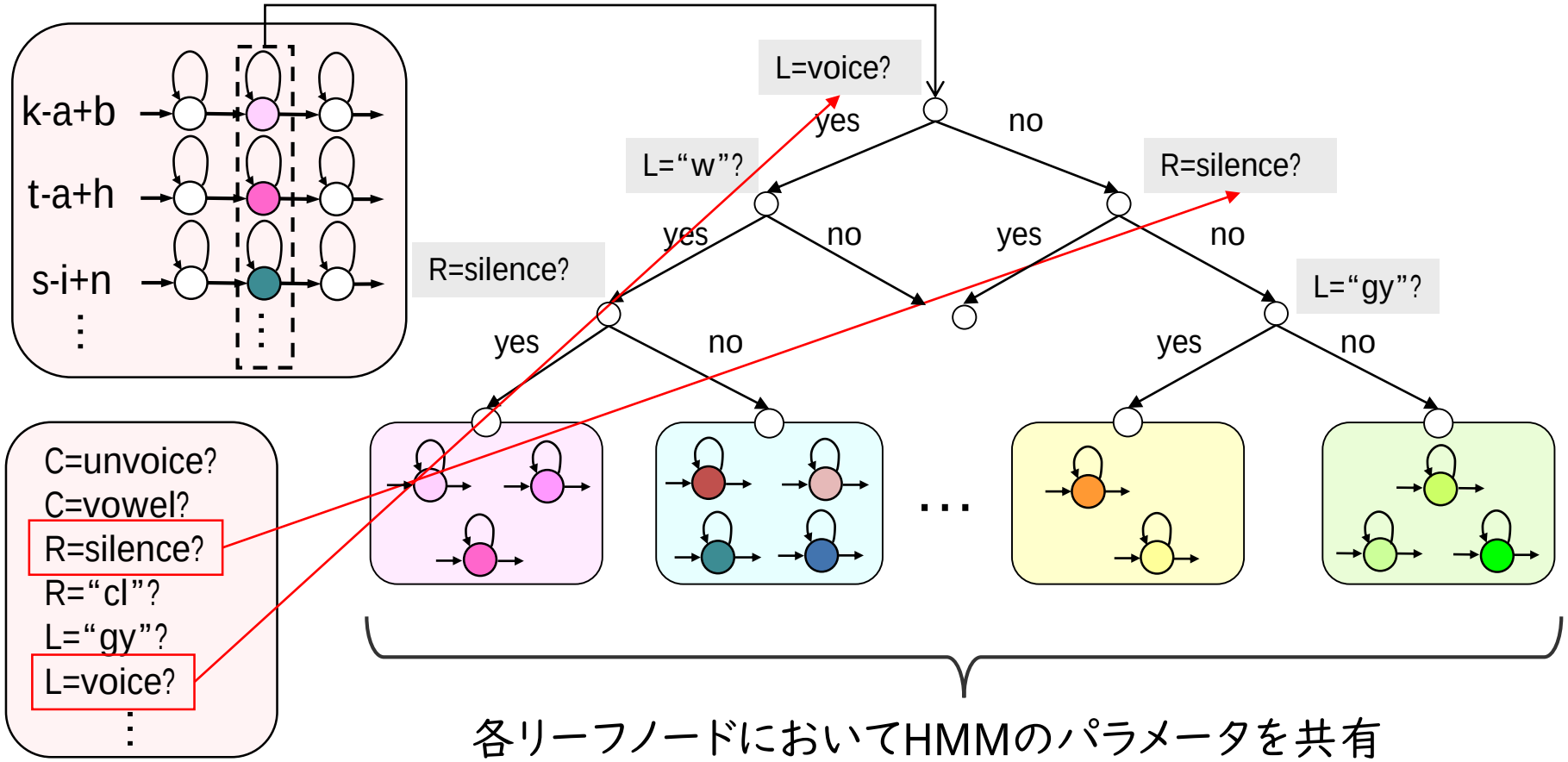
観測ベクトル（音響特徴量ベクトル）の構造



Dynamic feature [Furui ’86]

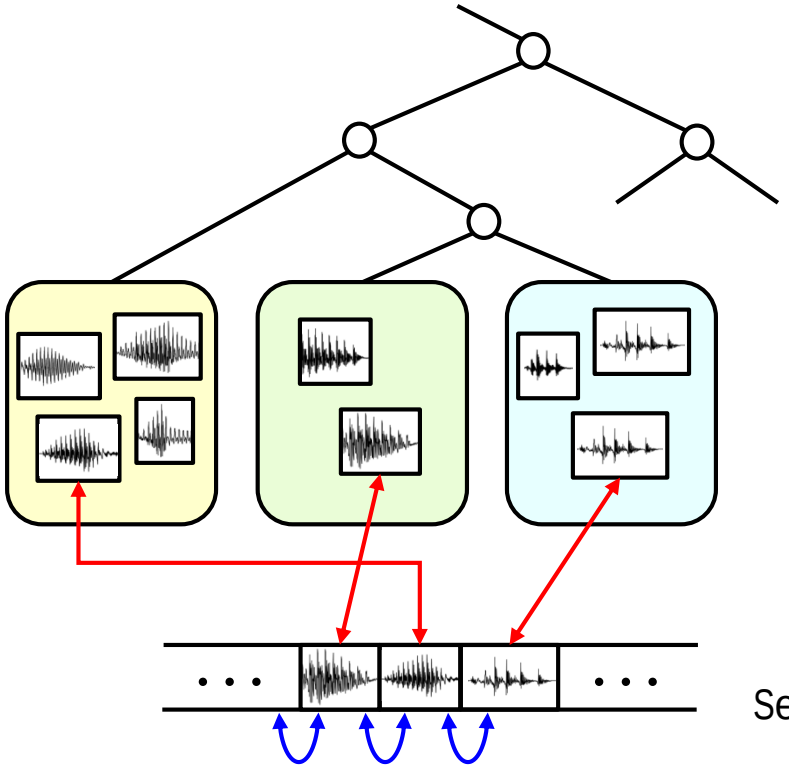
決定木に基づくコンテキストクラスタリング [Odell; '95]

$p(\boldsymbol{o}|\boldsymbol{l}, \lambda_A)$: HMM, $\boldsymbol{l}$: 言語特徴

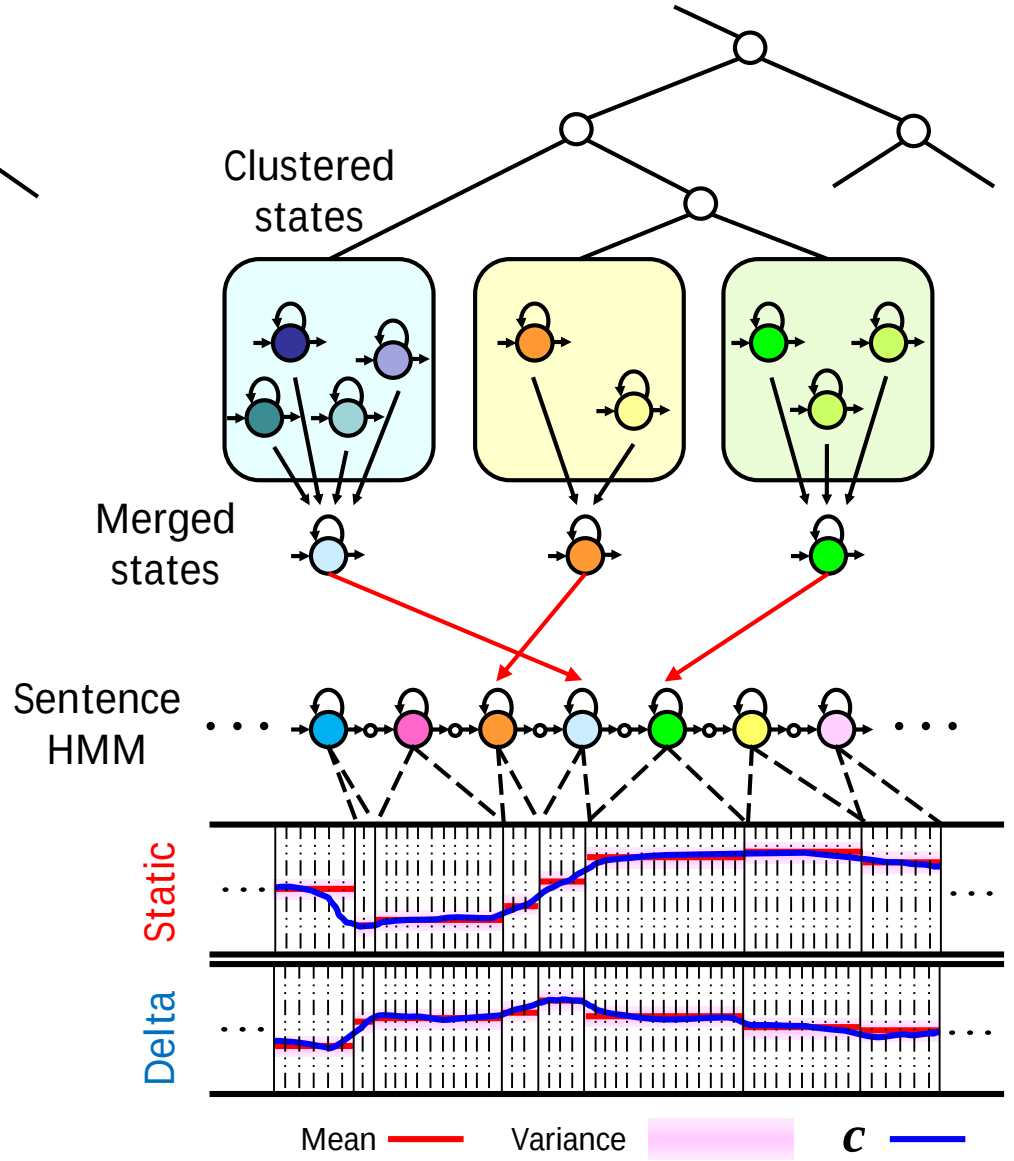


それぞれのアプローチを対照 (1/2)

単位選択型音声合成

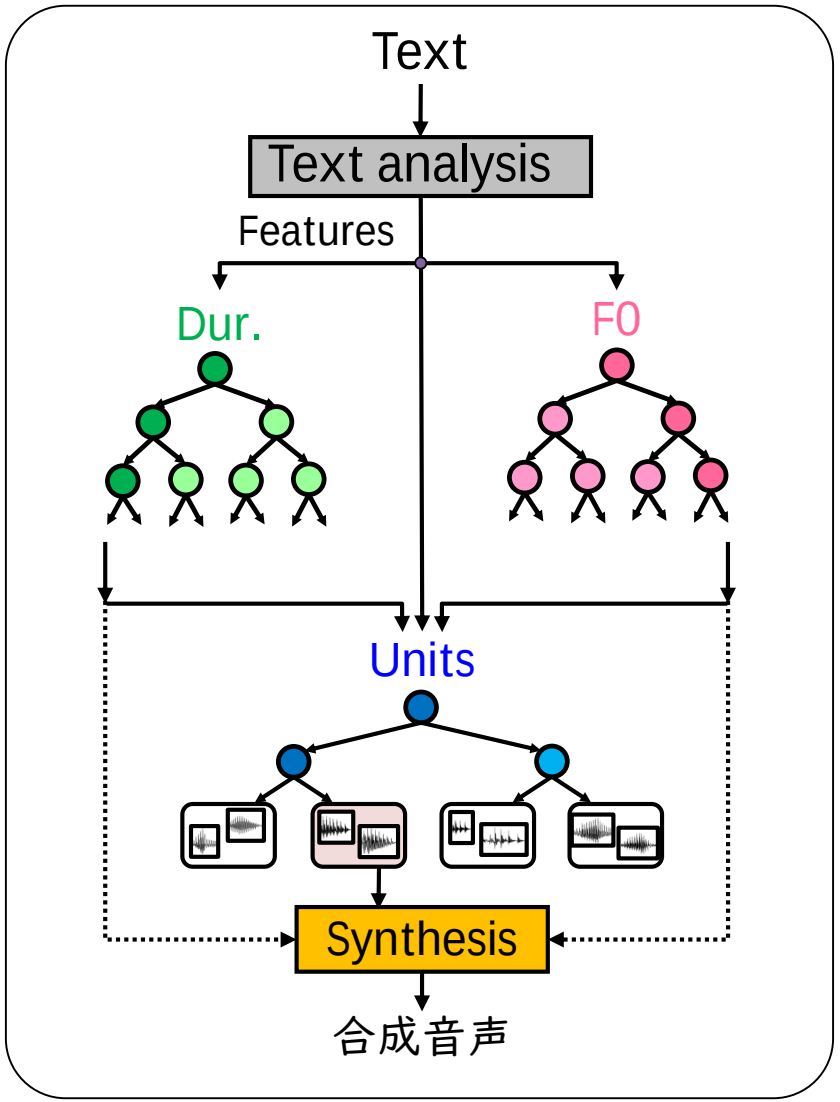


HMM音声合成

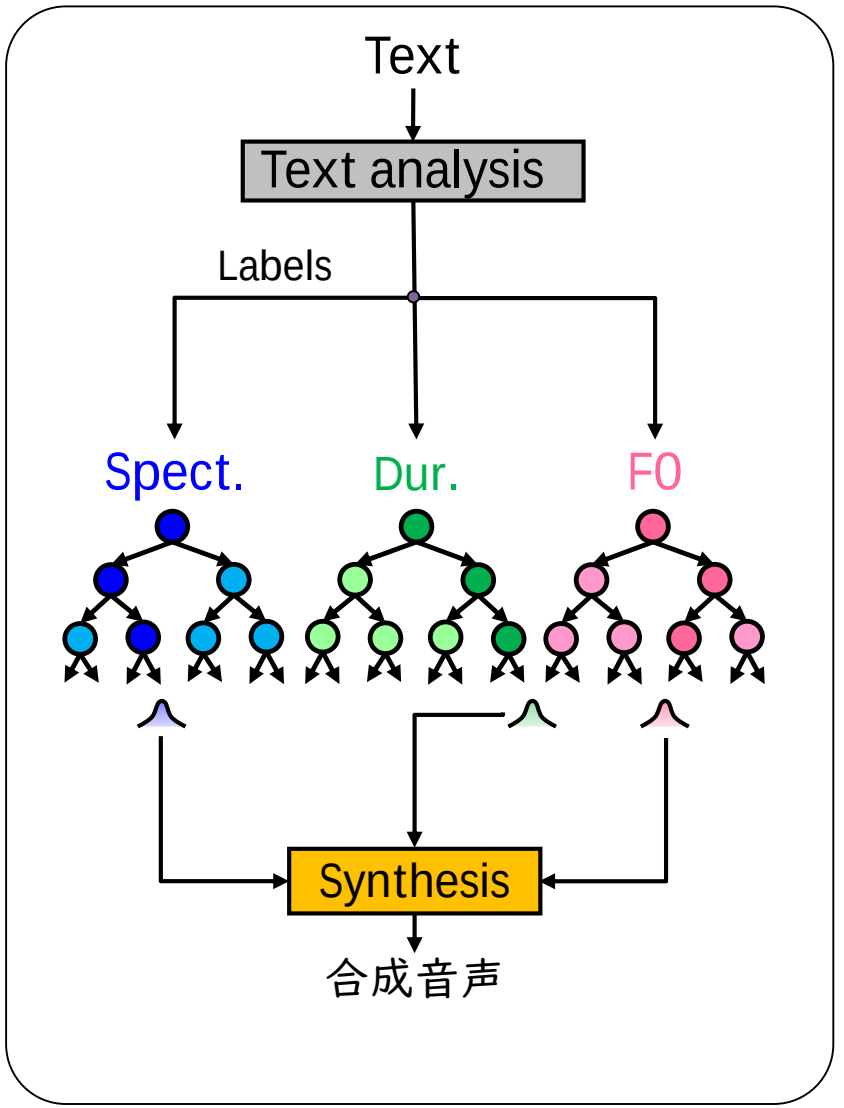


それぞれのアプローチを対照 (2/2)

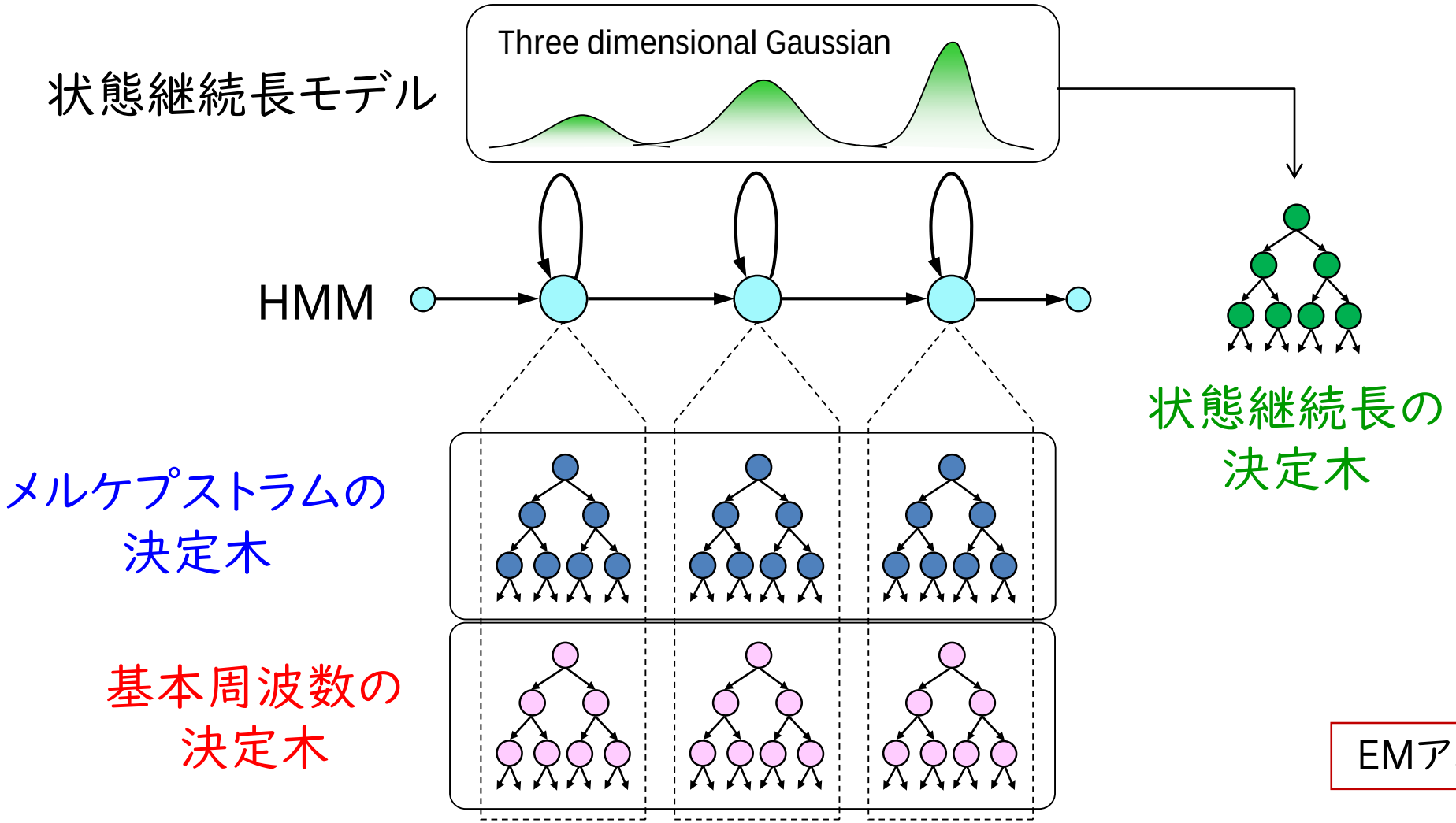
単位選択型音声合成



HMM音声合成

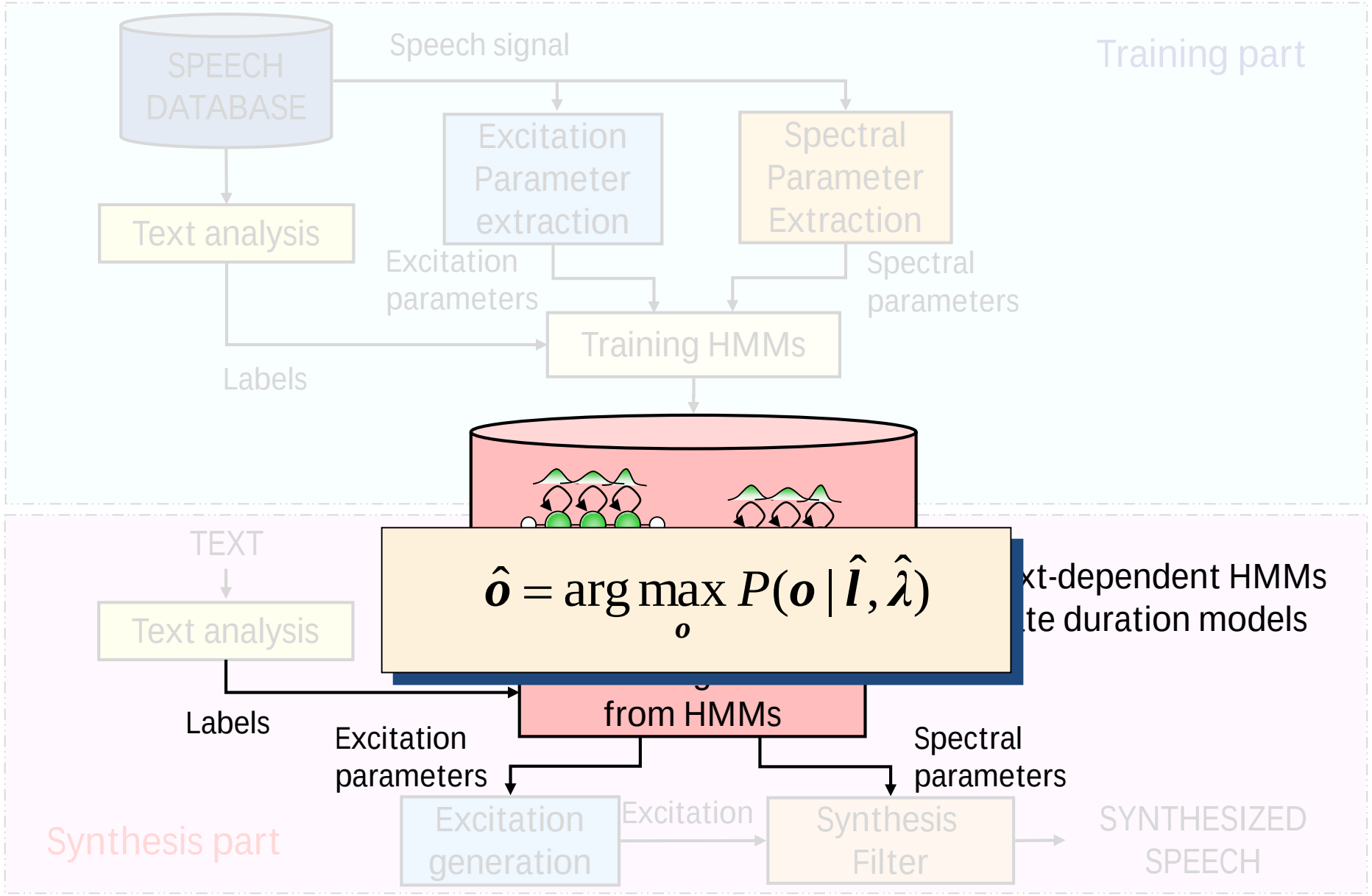


ストリーム依存決定木クラスタリング



EMアルゴリズムにより学習

HMM音声合成の枠組み



音声パラメータ（音響特徴）生成アルゴリズム

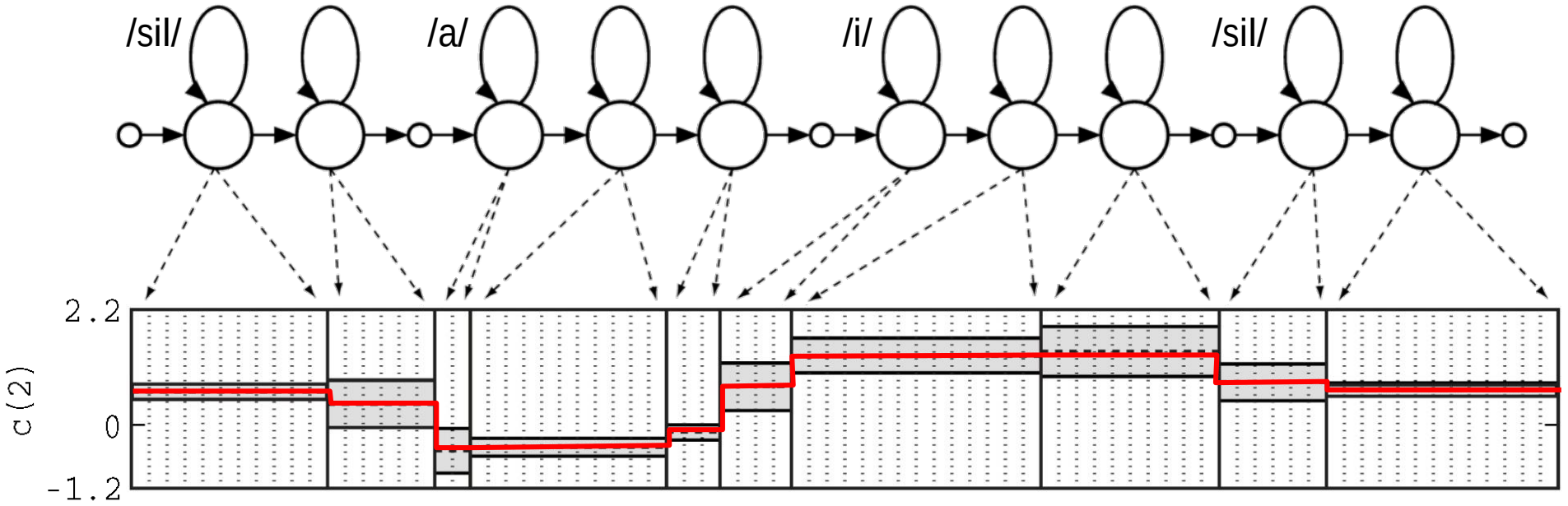
$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p\left(\boldsymbol{o}\middle|\hat{\boldsymbol{l}}, \hat{\lambda}_A\right) = \arg\max_{\boldsymbol{o}} \sum_{\boldsymbol{q}} P\left(\boldsymbol{o}\middle|\boldsymbol{q}, \hat{\lambda}\right) P\left(\boldsymbol{q}\middle|\hat{\boldsymbol{l}}, \hat{\lambda}\right)$$

$\boldsymbol{q}$: 状態系列

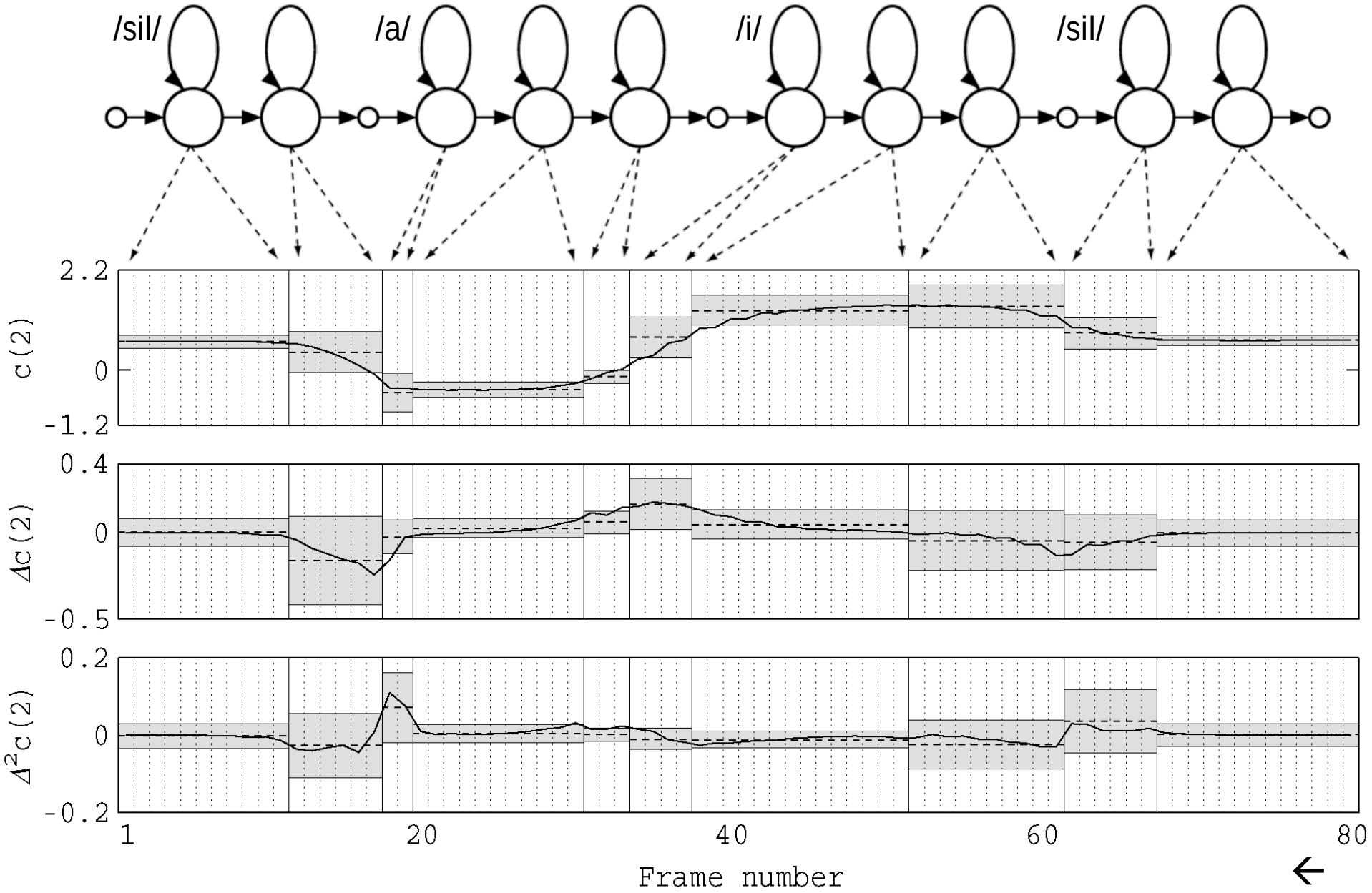
$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid \hat{\boldsymbol{l}}, \hat{\boldsymbol{\lambda}})$$ ← 状態継続長の決定

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \hat{\boldsymbol{\lambda}})$$ ← 音声パラメータの決定

生成された音声パラメータ列の軌跡



生成された音声パラメータ列の軌跡



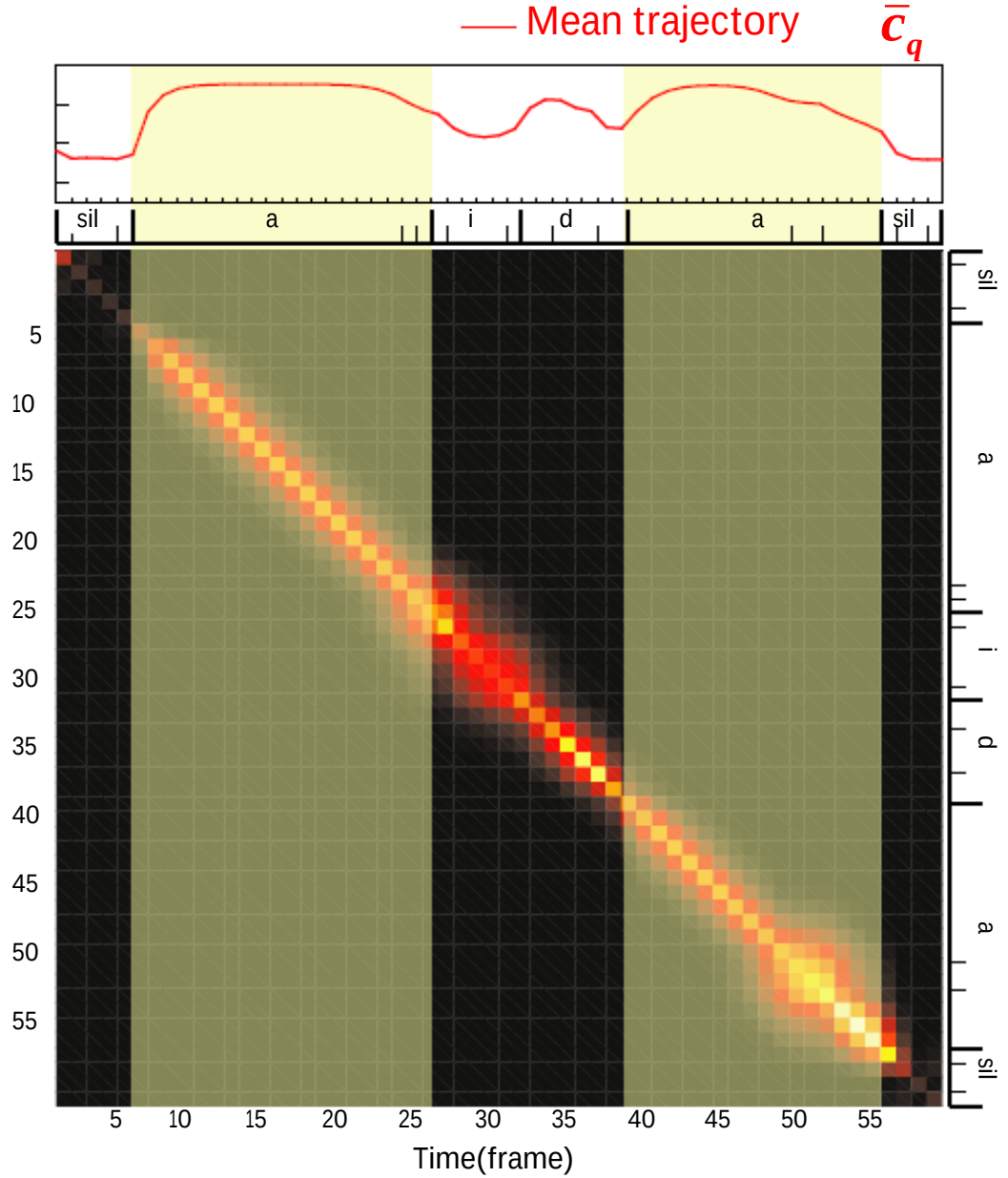
トラジェクトリ HMM

動的特徴量を含む

動的特徴量を含まない

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

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



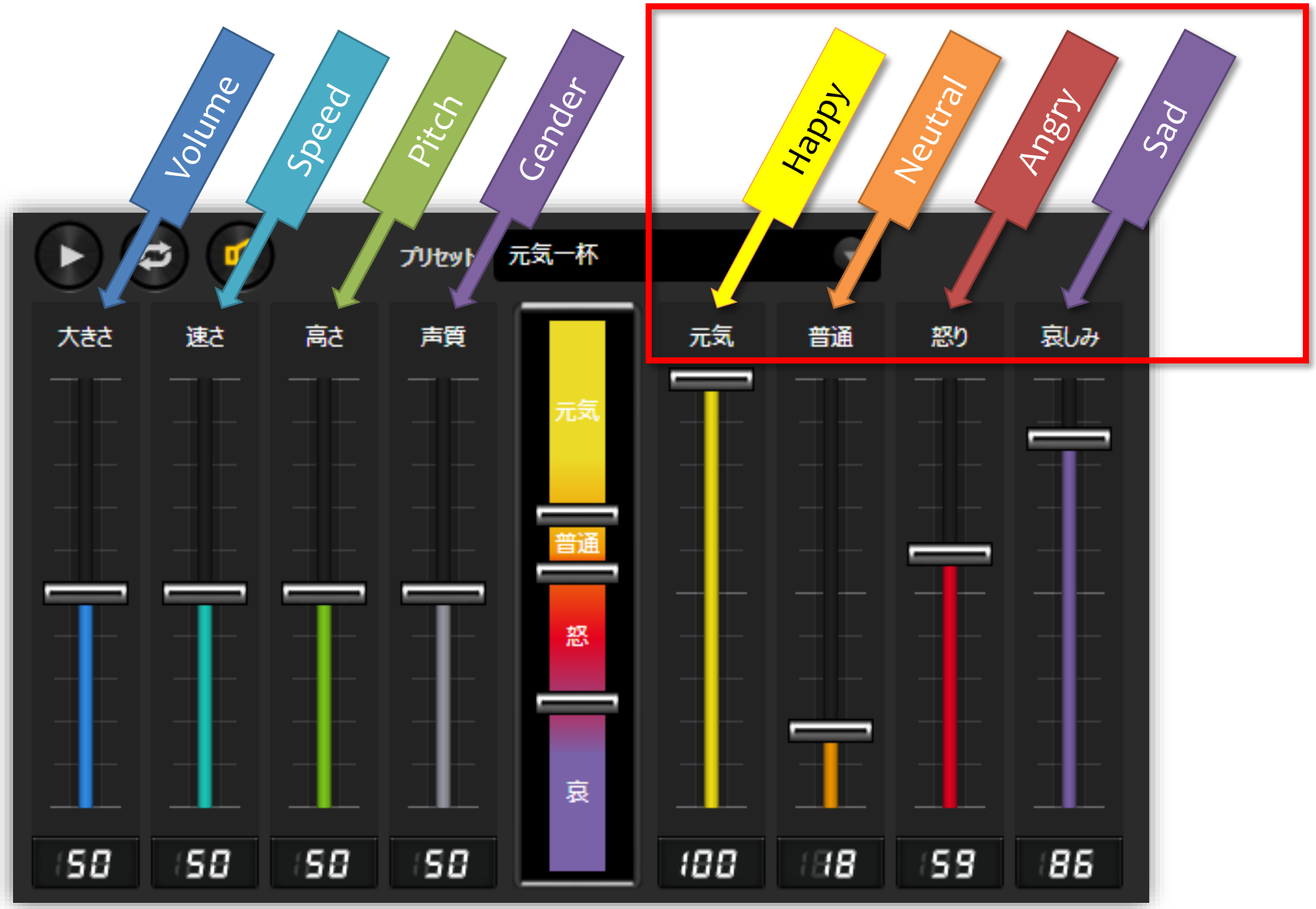
音声変化を制御する柔軟性

- Speaker Adaptation（声を真似る）
  - [Tamura ’98], [Tamura ’01], [Yamagishi ’03], …
- Speaker Interpolation（声を混ぜる）
  - [Yoshimura ’97], …
- Eigenvoice（声をつくる）
  - [Shichiri ’02], [Kazumi ’10], …
- Multiple-regression（声を制御する）
  - [Nose ’07], …

(→)

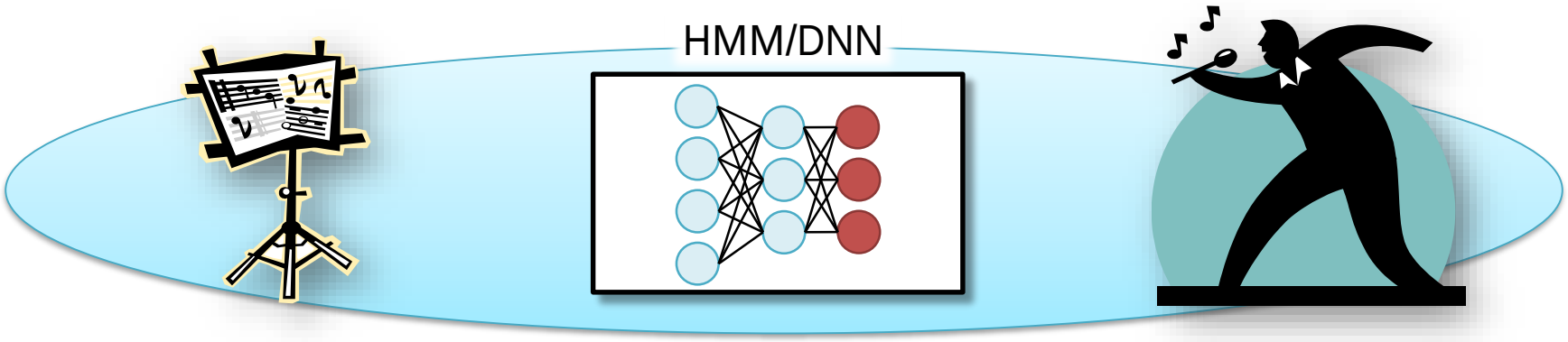
Only from publications by the HTS working group

感情表現を混ぜる


Volume
Speed
Pitch
Gender
Happy
Neutral
Angry
Sad
プリセット
元気一杯
大きさ
速さ
高さ
声質
元気
普通
怒り
哀しみ
元気
普通
怒
哀
50
50
50
50
100
18
59
86

(←)

歌声合成への応用



HMM+STRAIGHT

あらまし

- 音声合成の統計的定式化
- HMM音声合成
- DNN音声合成
- 評価 / データ&ソフトウェアツール
- その他の関連トピックス

隠れマルコフモデル (HMM) によるアプローチ

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.

テキスト解析

λ_L: テキスト解析部のパラメータ

Only from publications by the HTS working group

二つのモジュールの再結合

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

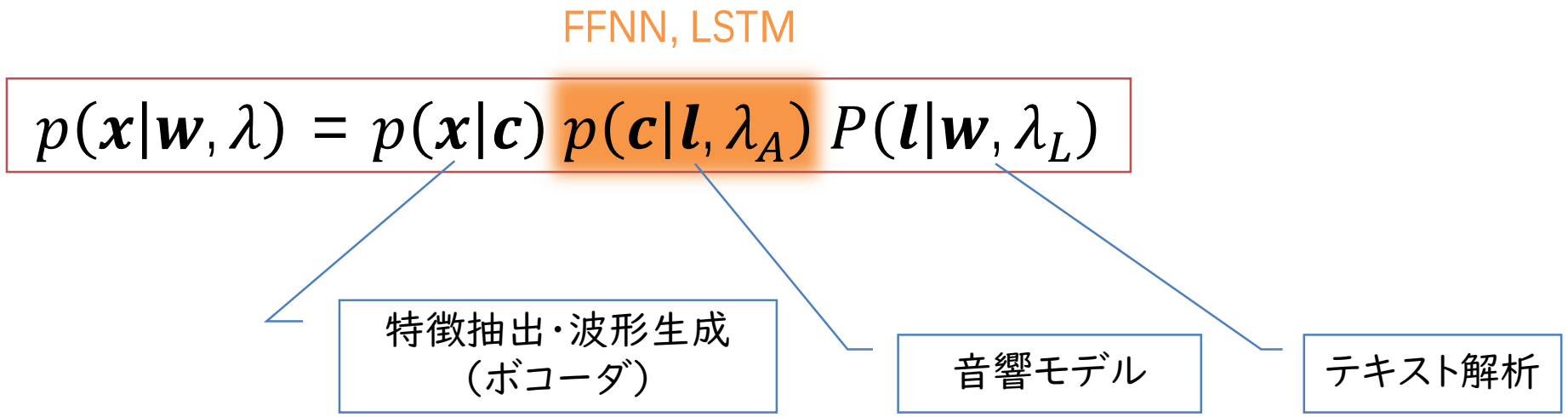
- Joint estimation of acoustic and excitation models [Maia ’10]
- Log spectral distortion-version of MGE training [Wu ’09]
- Factor analyzed trajectory HMM (STAVOCO) [Toda ’08]
- Mel-cepstral analysis-integrated HMM [Nakamura ’14]

テキスト解析

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

Only from publications by the HTS working group

ディープニューラルネットワーク (DNN) によるアプローチ (1/6)



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

DNN vs. HMM

DNN

- データ量比較的大？
- **フラットな構造**
  - トラブル解決が困難
  - 実装が容易
- 事前情報や知識が初期化や学習手順に埋め込まれる
- 並列・分散処理しやすい
- 連続空間における最適化

HMM (≅regression tree)

- データ量比較的小？
- **意味付けのある構造**
  - トラブル解決が容易
  - 実装が困難
- 事前情報や知識を明確な形で与えやすい
- 並列分散処理しにくい
- 離散空間における最適化

いずれもダイナミクスのモデル化構造は必要そう

ディープニューラルネットワーク (DNN) によるアプローチ (2/6)

Source filter model

FFNN, LSTM

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

→

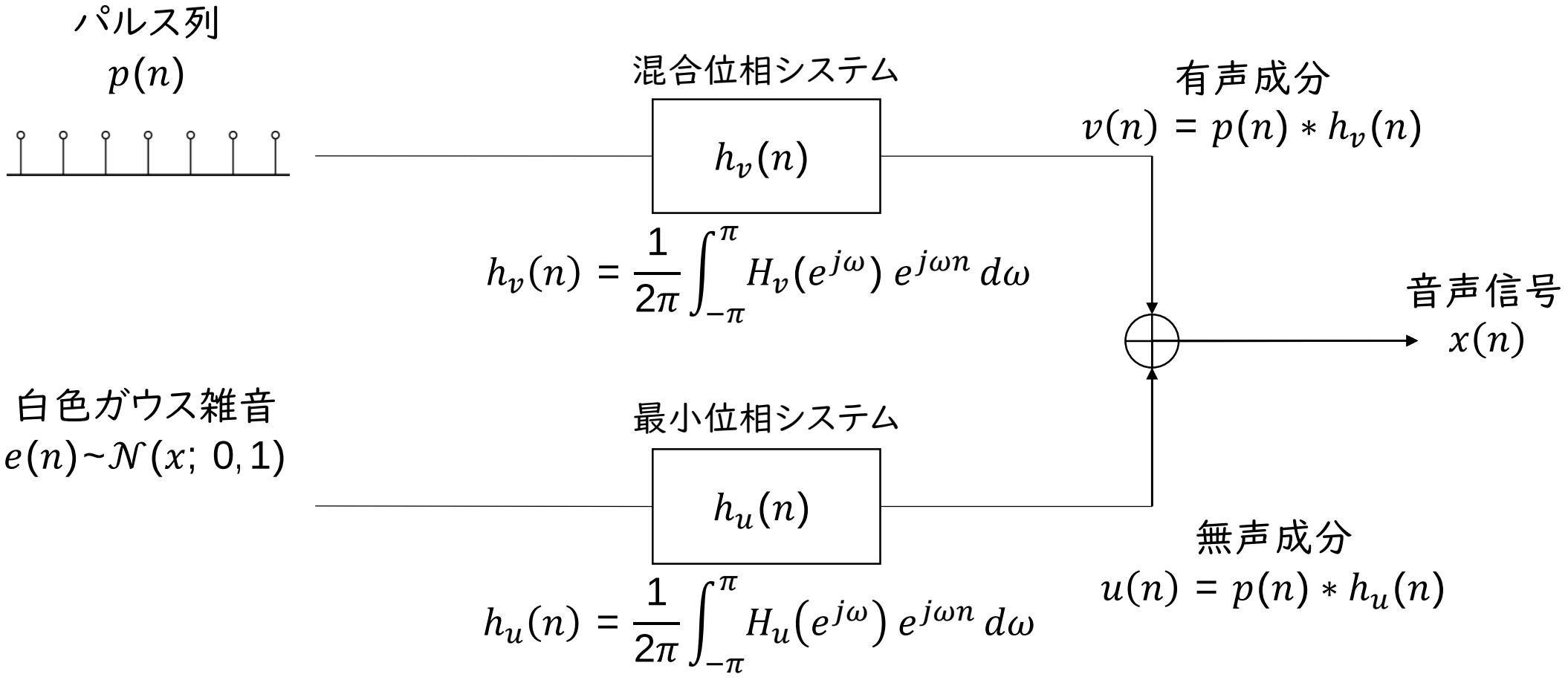
>>

1. 音声波形の尤度を直接測る形で
2. 有声成分・無声成分を同時にモデル化する
3. ニューラルネットワークを学習する

テキスト解析

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

音声信号モデル



有声成分+無声成分の信号モデル

ディープニューラルネットワーク (DNN) によるアプローチ (3/6)

WaveNet, SampleRNN, WaveRNN, …
(autoregressive sturucture)

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

特徴抽出・波形生成
(ボコーダ)

音響モデル

テキスト解析

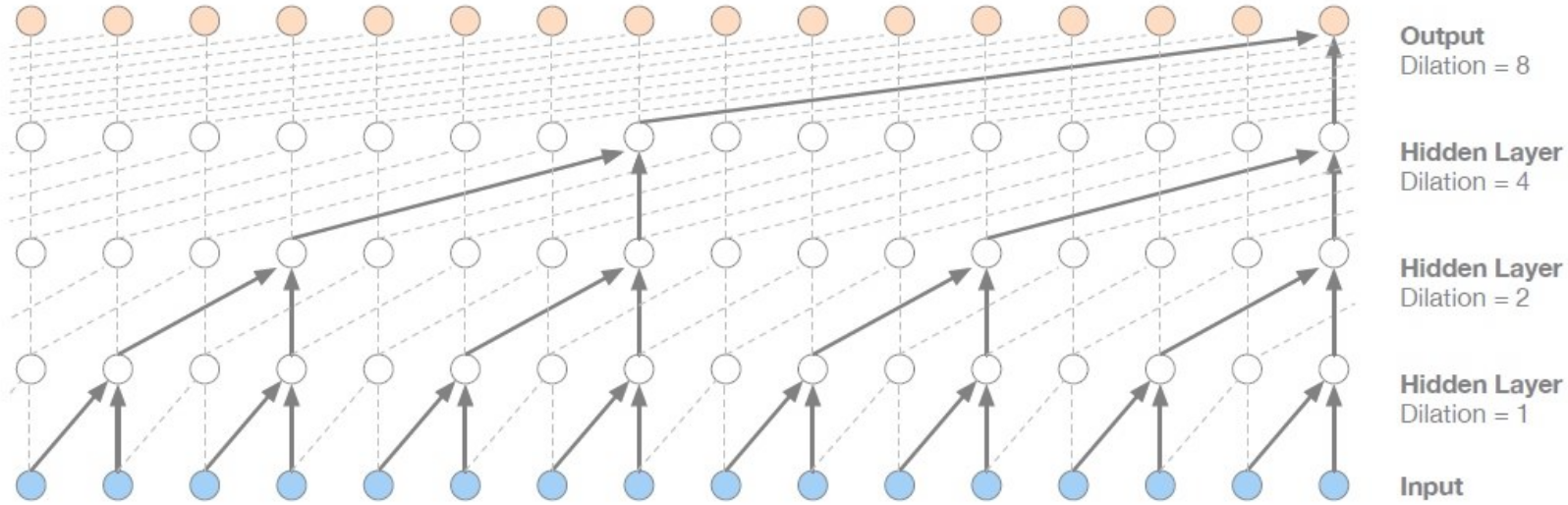
WaveNet

- たたみ込みニューラルネットワーク (CNN) に基づいた自己回帰型生成モデル
  - 音声波形の直接モデル化

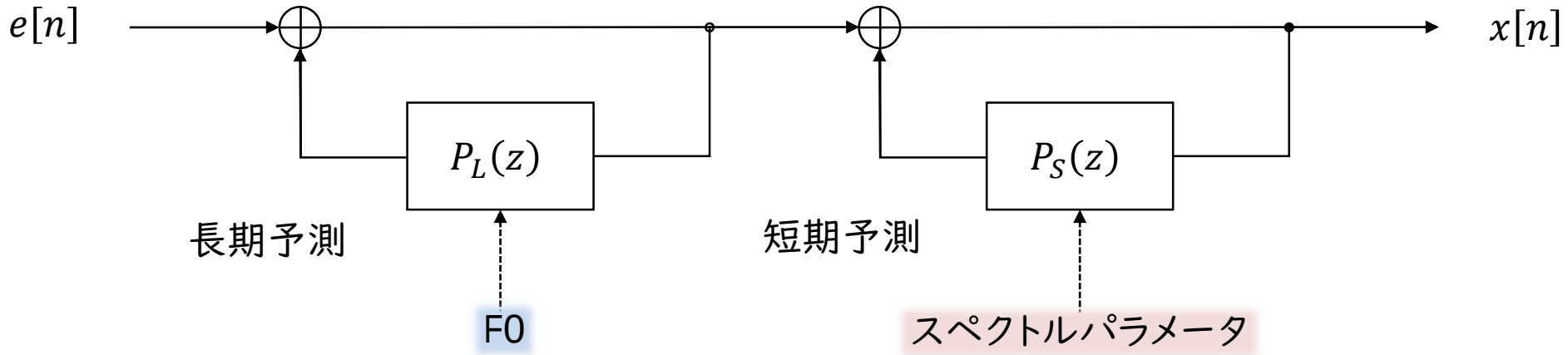
$\boldsymbol{x}$: 音声波形　　$\boldsymbol{h}$: 音響特徴 or 言語特徴

$$p(\boldsymbol{x} \mid \boldsymbol{h}) = \prod_{n=0}^{N-1} \underbrace{p(x[n] \mid x[0], \ldots, x[n-1], \boldsymbol{h})}_{\text{CNNによりモデル化}}$$

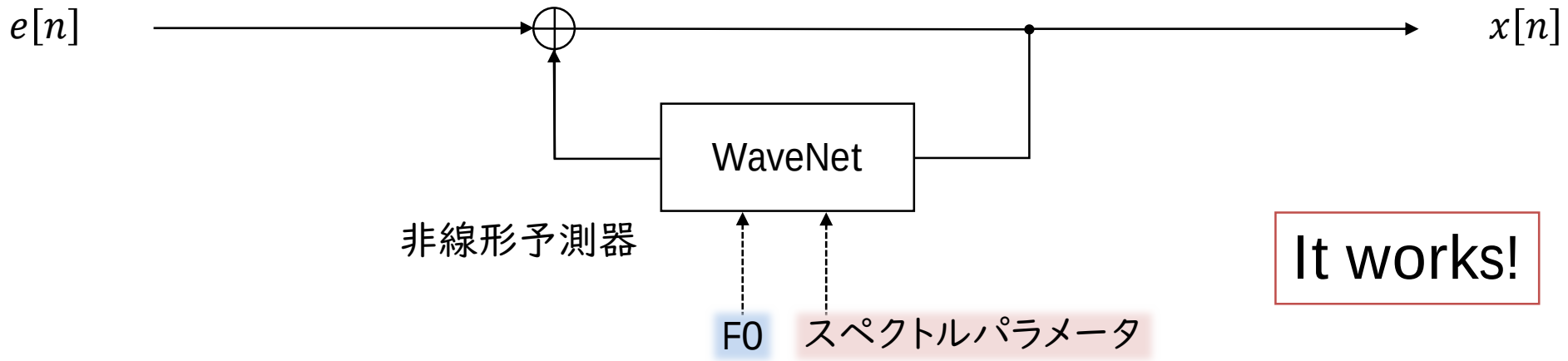
  - Dilated causal convolution



音声信号の生成モデル



←



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

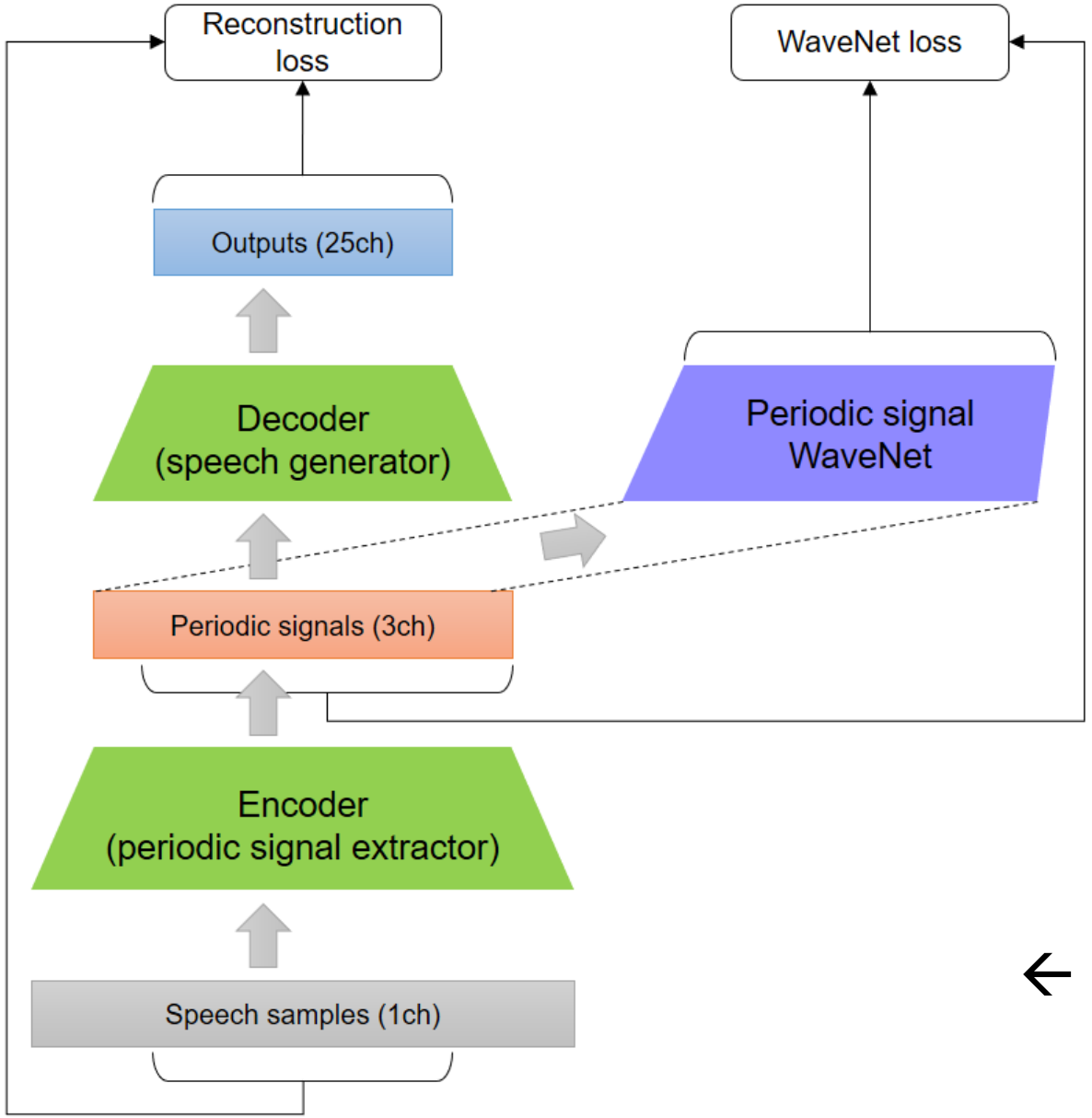
"Every time I fire a speech synthesis researcher,
the performance of the speech synthesizer goes up"
by ????? ?????

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



ディープニューラルネットワーク (DNN) によるアプローチ (4/6)

FFNN, LSTM

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

特徴抽出・波形生成
(ボコーダ)

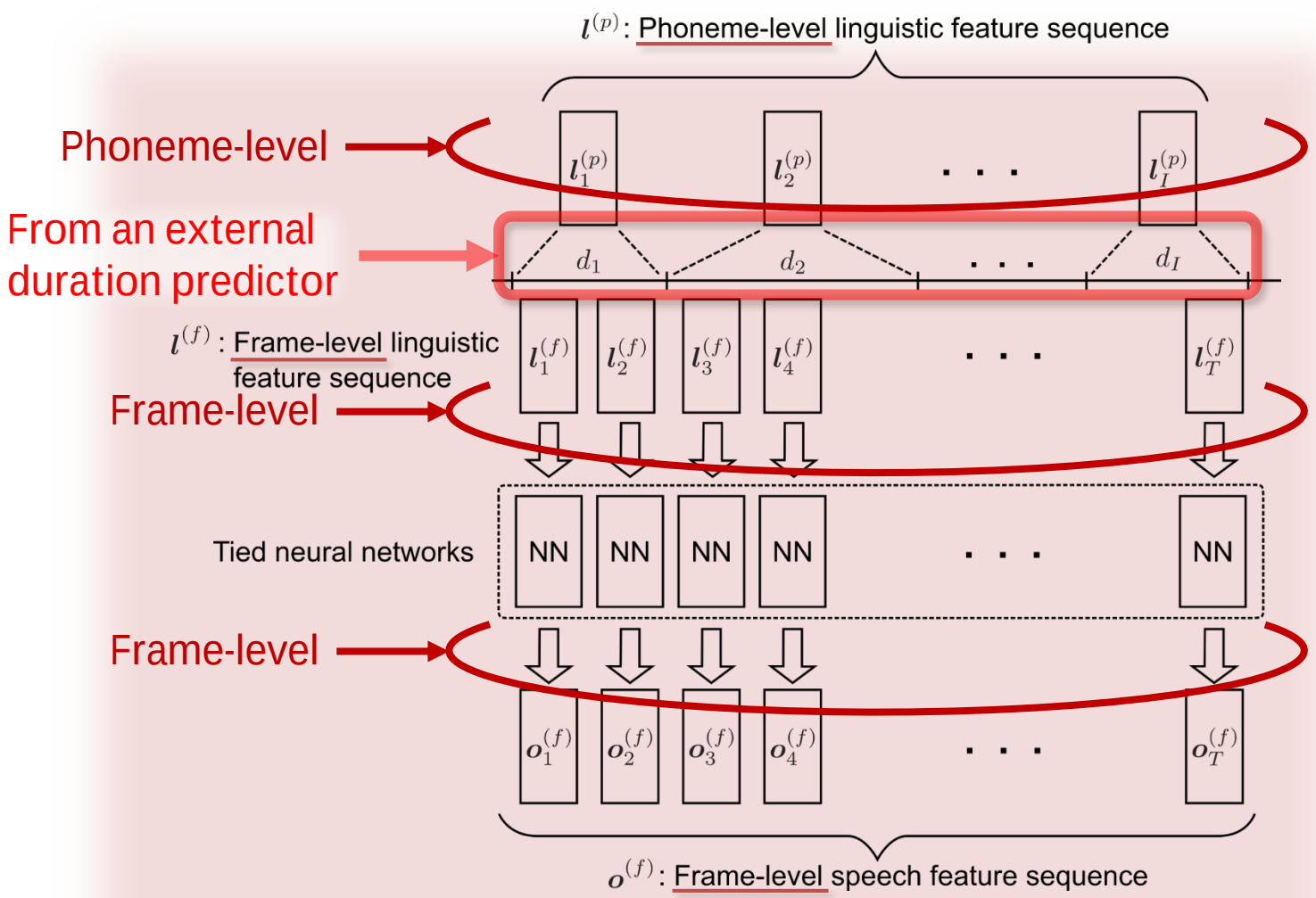
音響モデル

テキスト解析

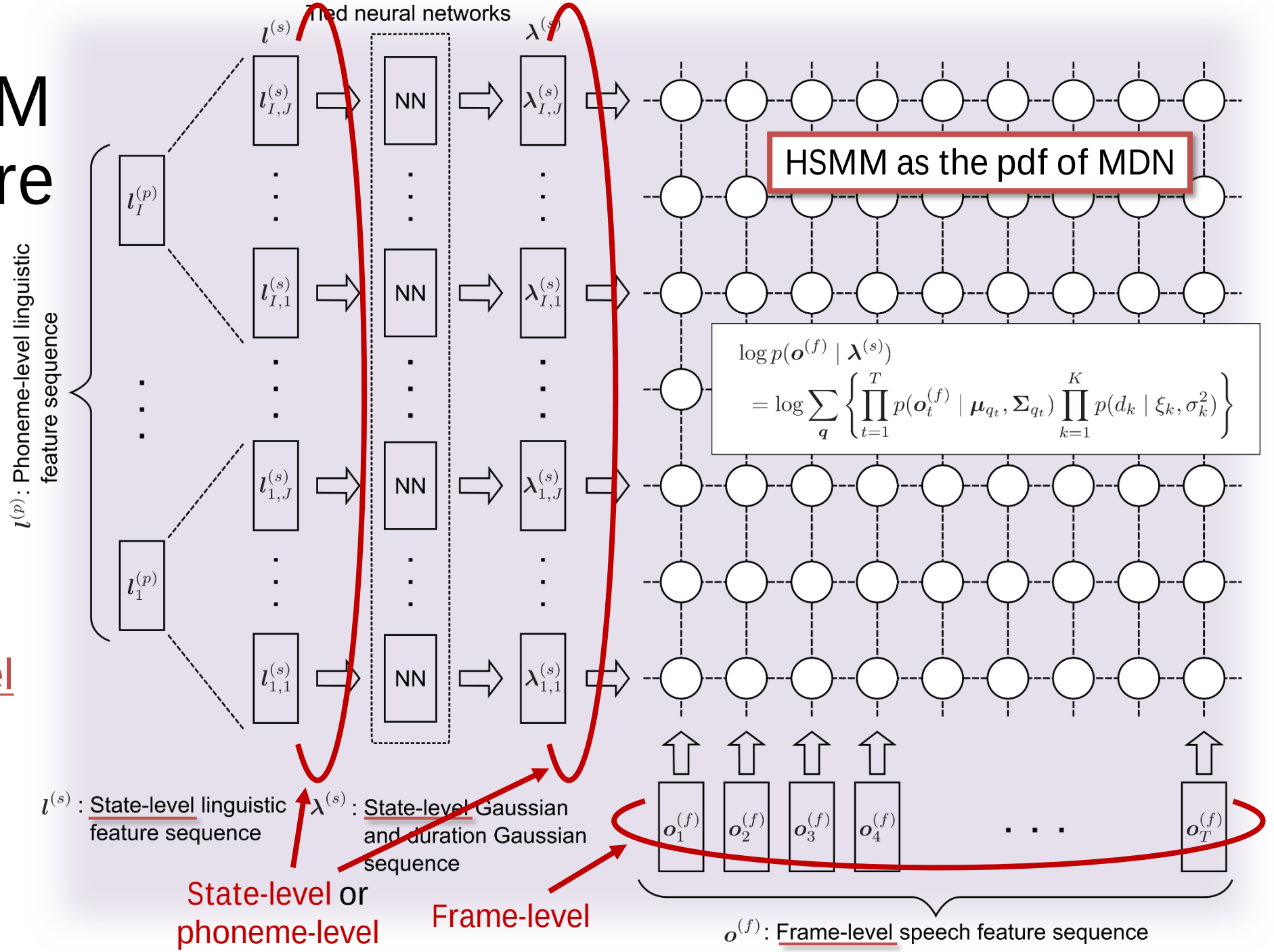
- HSMM: 継続長モデルを含む
- FFNN, LSTM, WaveNet: 外部に継続長予測器が必要

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]

ディープニューラルネットワーク (DNN) によるアプローチ (5/6)

WaveNet vocoder
(autoregressive structure)

Tacotron, Char2Wav, DeepVoice, …
(attention mechanism)

→

>>

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

特徴抽出・波形生成
(ボコーダ)

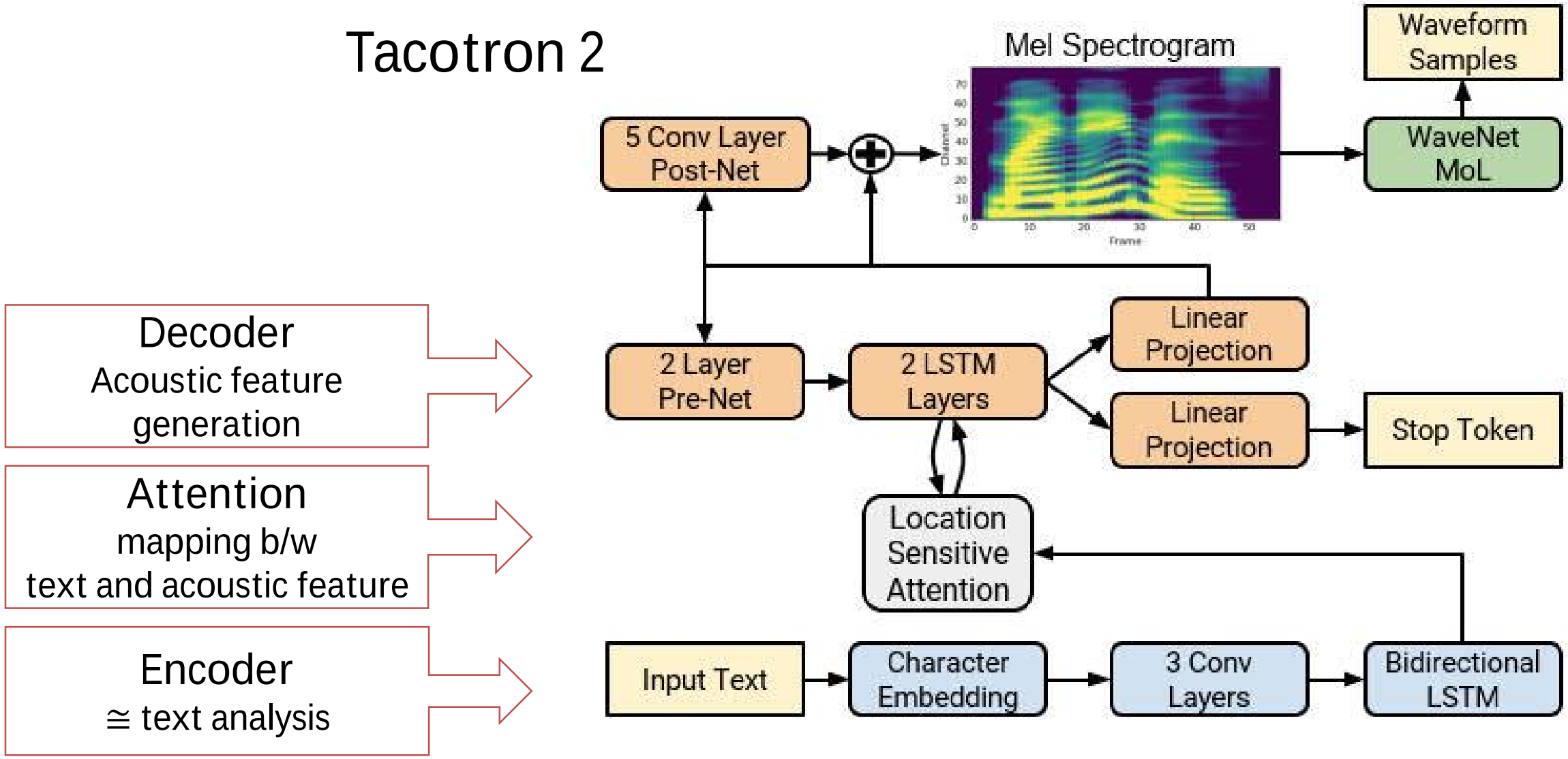
音響モデル

テキスト解析

Attention mechanism

[from arXiv:1712.05884]

Tacotron 2



ディープニューラルネットワーク (DNN) によるアプローチ (6/6)

WaveNet vocoder
(autoregressive structure)

Attention

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

←

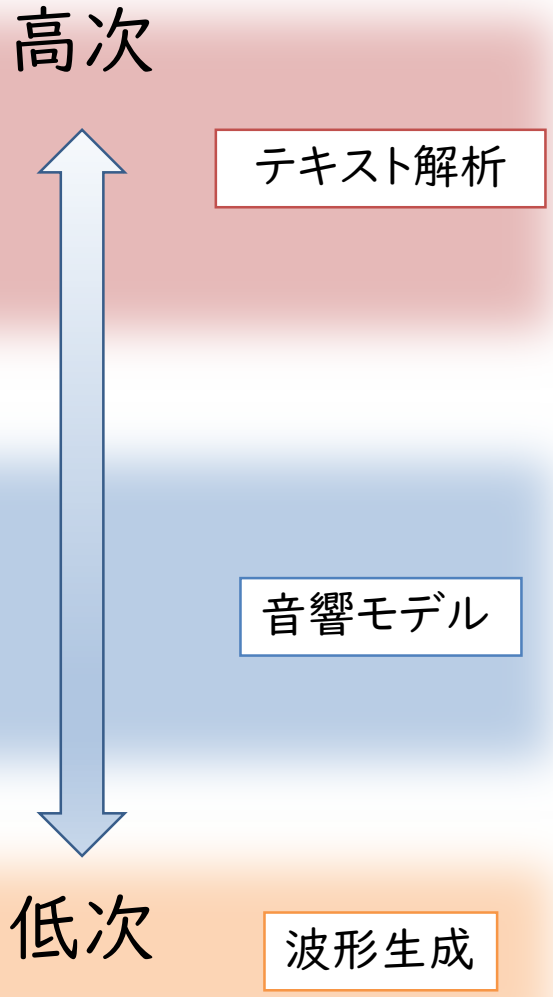
特徴抽出・波形生成
(ボコーダ)

音響モデル

テキスト解析

様々な階層での制御・編集機能

- 言語
  - 日本語, 英語, 中国語, 日本語英語, …
- 読み
- ポーズ
- 発音変形
  - 音声 → /o N s e:/, します → /sh I ma s/
- 韻律変形
  - アクセント結合・変形
- 発話スタイル・感情表現等
- 単語の強調
- ノンバーバル情報・パラ言語情報
- 声質
  - 男性, 女性, 大人, 子供
- 音声パラメータ
  - 基本周波数パターン, 音量変化パターン, 継続長, …



FFNN+CNN+WaveNet による歌声合成

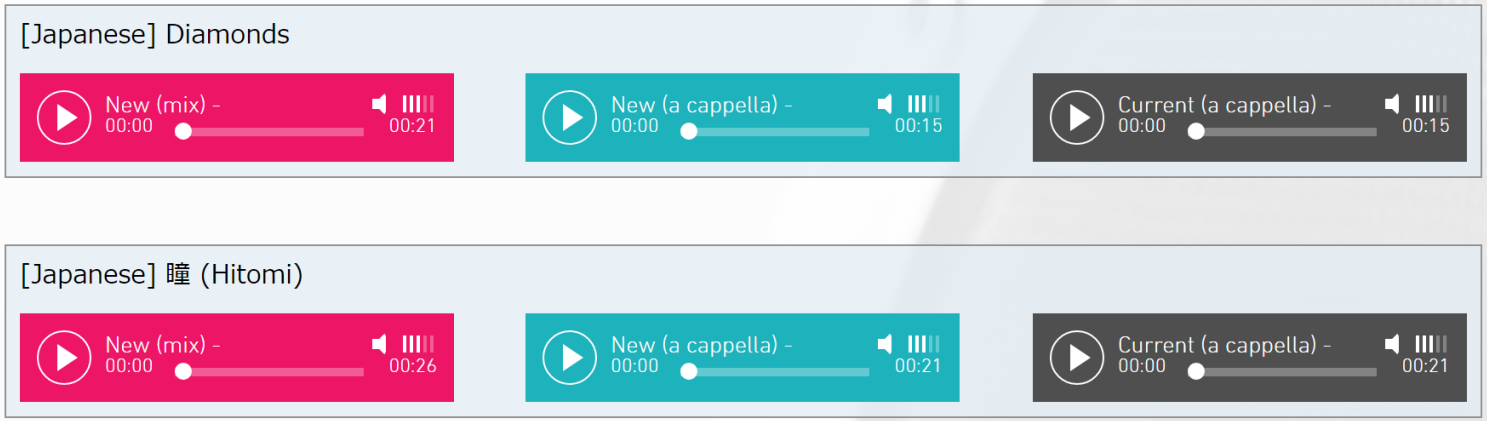
WaveNet vocoder　FFNN+CNN

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

特徴抽出・波形生成
（ボコーダ）

音響モデル

テキスト解析



その他のDNN技法とアーキテクチャ

- GAN
- VAE / VQ-VAE
- Transformer / BERT

あらまし

- 音声合成の統計的定式化
- HMM音声合成
- DNN音声合成
- 評価 / データ&ソフトウェアツール
- その他の関連トピックス

Blizzard Challenge

- TTSシステムの性能はデータベースに強く依存
- 技術そのものを評価することの難しさ

"Blizzard Challenge"
Evaluating corpus-based speech synthesis on common datasets [Black '05]

Since 2005

評価方法

- 自然性
  - Mean Opinion Score (MOS)
- 話者類似性
  - Degradation Mean Opinion Score (DMOS)
- 明瞭性 (dictation of SUS, PCS, etc.)
  - 単語正解精度

自由発話音声、オーディオブックタスクなどでは、十分ではない

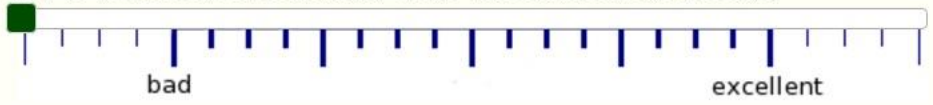
Section 1: Part 1 / 17

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.

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.

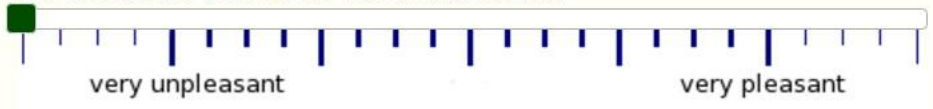
Overall impression

How do you rate the overall quality of the voice that read this passage?



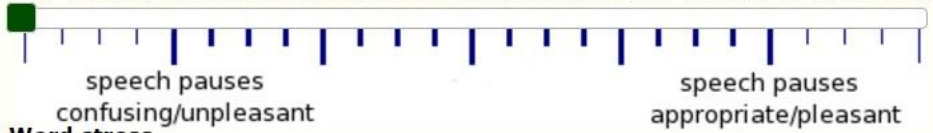
Pleasantness

How pleasant did you find the voice you just heard?



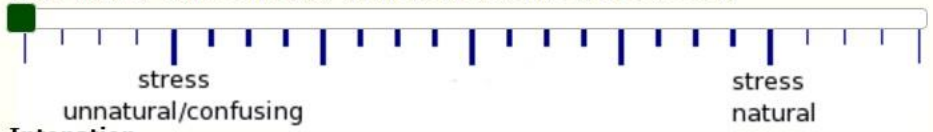
Speech pauses

How did the pauses between words and sentences affect your listening to the passage?



Word stress

What did you think of the way words in the passage were stressed?



Intonation

What did you think of the "melody" of the voice reading this passage?

音声合成のための共通のデータセット

- ELRA http://www.elra.info/
- ELDA http://www.elda.org/
- LDC https://www.ldc.upenn.edu/
- OpenSLR http://www.openslr.org/
- SRC http://research.nii.ac.jp/src/

- ARCTIC
- VCTK
- LibriTTS, …

音声認識に比べて数や量が少ないのは
スタジオ品質の収録が必要なため？

ソフトウェアツール (1/2)

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

ソフトウェア・ツール (2/2)

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

あらまし

- 音声合成の統計的定式化
- HMM音声合成
- DNN音声合成
- 評価 / データ&ソフトウェアツール
- その他の関連トピックス
  - テキスト正規化
  - 声質変換
  - 音声符号化
  - Anti-spoofing
  - 物理的シミュレーション

テキスト正規化

- End-to-endシステムにテキスト正規化は含まれない
- ルールベースアプローチが主流

- 近い将来end-to-endプロセスに組み込むことが可能となる？

声質変換

- 音声合成と緊密な関係
- 声質変換研究においてもDNNによるアプローチが台頭している
- リアルタイム応用が本質
- リアルタイム（あるいは低遅延）の韻律変換はチャレンジングな課題

音声符号化

- WaveNetあるいはその他の波形生成アプローチは音声符号化に革命をもたらす可能性

- 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]

合成音声による詐称

- 音声合成を用いたなりすましへの懸念
  - On the security of HMM-based speaker verification systems against imposture using synthetic speech [Masuko ’99]
- 合成音声の検出
  - 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

物理的シミュレーションとDNN

- 将来, 声道の動的計測技術が進展
- 音声生成過程のシミュレーション技術の向上

物理的シミュレーションに基づいて
自然な音声を生成することは可能になる？

- 利点：現実的なモデル制約の低次元表現
  → DNNベースシステムの潜在表現として有効？

まとめ

統計的アプローチによる音声合成

- 人間と機械の区別がつかなくなるレベルに達しつつあるが、
- まだまだ多くの課題がある　→
- 音声の多様性を実現するために更なる柔軟性と制御性が必要

まだまだ楽しめる音声合成研究！

ありがとうございました！

>>

音声合成の未来

- 音声対話システム
- 音
- 障
- 言語学習
- コンテンツ制作

マルチリンガル/クロスリンガル

柔軟性/多様性

エディターデザイン

共通のデータ/ツール

更なる音声の多様性の実現を目指して
まだまだ楽しめる音声合成研究!

ありがとうございました!

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.

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