東京大学 信号処理論特論第7回 (2018/06/05)

音声合成・変換 その1

猿渡 洋・高道 慎之介

講義予定

04/10: 第1回 統計的音声音響信号処理概論
05/01: 第2回 非負値行列因子分解
05/08: 第3回 ブラインド音源分離その１
05/15: 第4回 ブラインド音源分離その２
05/22: 第5回 エンハンスメント・高次統計量解析とその応用
05/29: 第6回 【レポート課題１】
06/05: 第7回 音声合成・変換その１
06/12: 第8回 音声合成・変換その２
06/19: 第9回 音場再現の基礎
06/26: 第10回 学外講師・未定
07/03: 第11回 【レポート課題２】

講義資料と成績評価

講義資料

- http://www.sp.ipc.i.u-tokyo.ac.jp/~saruwatari/
- (システム情報第一研究室からたどれるようになってます)

成績評価

- 出席点
- レポート点 (2回の提出が必須)

はじめに

本講義の目的

音声合成・変換とは何？その基盤技術は？
（応用やホットな話題に関しては合成変換2で扱います）

音声合成：音声を人工的に作り出す技術

狭義の音声合成

- テキスト音声合成 (Text-To-Speech: TTS)
  - 音声認識 (speech-to-text) の逆

広義の音声合成 (xxx-to-speech)

- テキスト音声合成
- 音声変換 (Voice Conversion: VC)
- ボイスチェンジャ
- 概念音声合成 (Concept-To-Speech: CTS)
  - 概念 → 言語生成 → 音声合成
- 調音・音響間マッピング
  - 調音機構特性と音声の変換
- マルチモーダル音声合成
  - 動画像などを含む音声合成

テキスト音声合成・変換

テキスト音声合成 (Text-To-Speech: TTS)

- テキスト等から音声を合成
- ヒト以外のモノのコミュニケーションのため

音声変換 (Voice Conversion: VC)

- 音声を異なる音声に変換
- ヒトの発声制約をこえたコミュニケーションのため

音声合成の役目：
モノの違いを超えたコミュニケーション

音声変換
(声をかえる)

テキスト音声合成
(声をつくる)

製品例

VOCALOID

音声合成の声優事務所

マツコロイド & totto

R.C. voice

クリムゾン

Google Home

コエステーション

https://www.vocaloid.com/products
https://www.ai-j.jp/archives/7889
http://voicetext.jp/voiceactor/
https://store.google.com/jp/product/google_home
https://crimsontech.jp/works/rcvoice/
https://coestation.jp/

音声の持つ情報

狭義の音声認識
(speech-to-text)

言語情報

テキスト化できる情報

感情認識など
(emotion recognition)

パラ言語情報

話し手が意図的に付与する,
テキスト化できない情報 (例：感情)

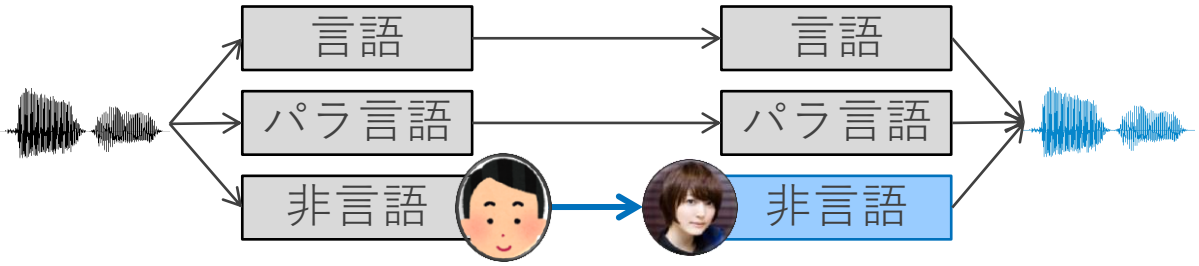
話者認識など
(speaker recognition)

非言語情報

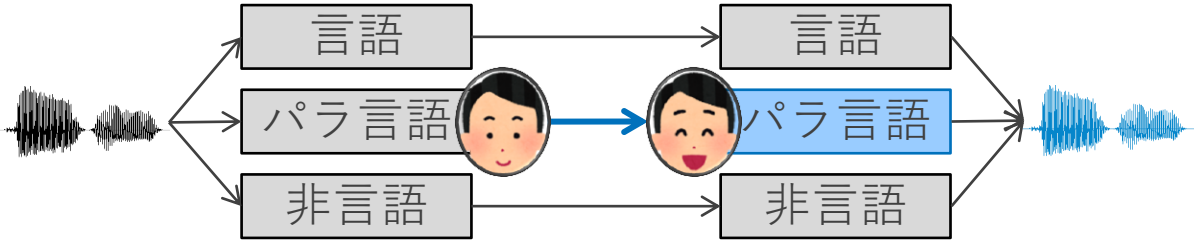
話し手の意図とは無関係に付与される,
テキスト化できない情報（例：話者性）

音声変換は何の情報を保持・変換する？

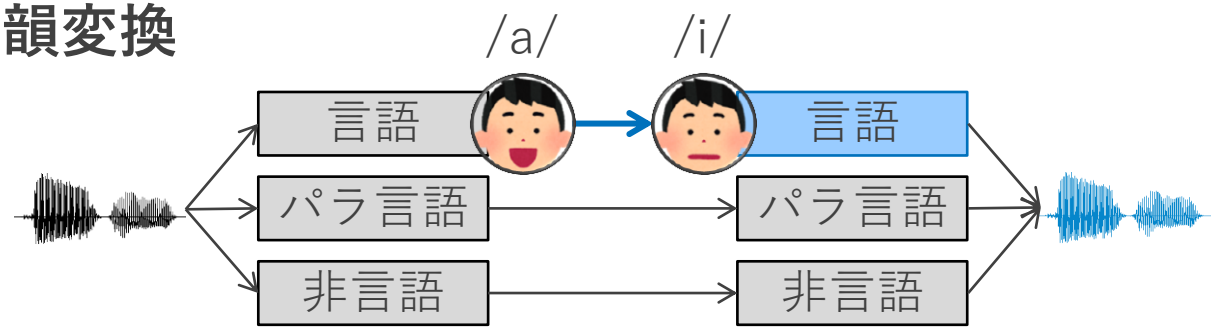
例1：話者変換（名探偵コナンの蝶ネクタイ型変声器）



例2：感情変換

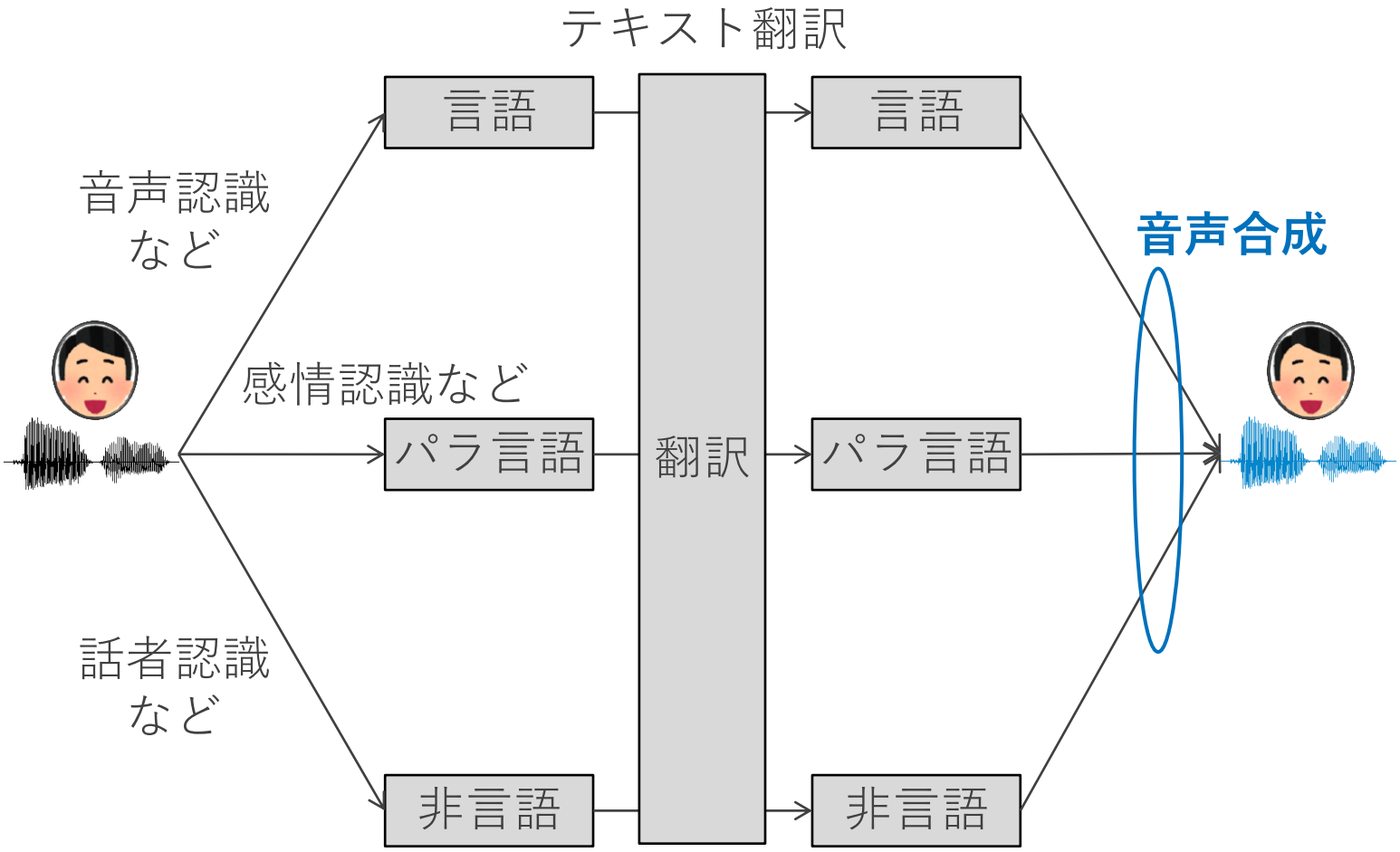


例3：音韻変換



音声合成は何の情報を保持・変換する？

例：究極の音声翻訳（ドラえもんのホンヤクこんにゃく）



コンテキスト・音声特徴量

コンテキスト・音声特徴量

音声合成では入出力情報から特徴量を抽出

コンテキスト：音声を制御する特徴量

- 言語特徴量
- パラ言語特徴量
- 非言語特徴量

音声特徴量：音声を効率的に表す特徴量

- 声道の特徴量
- 声帯の特徴量

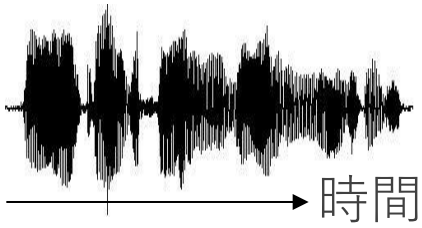
音声の生成過程：ソース・フィルタモデル

音色の付与

口や舌を動かして，
音色をつける！

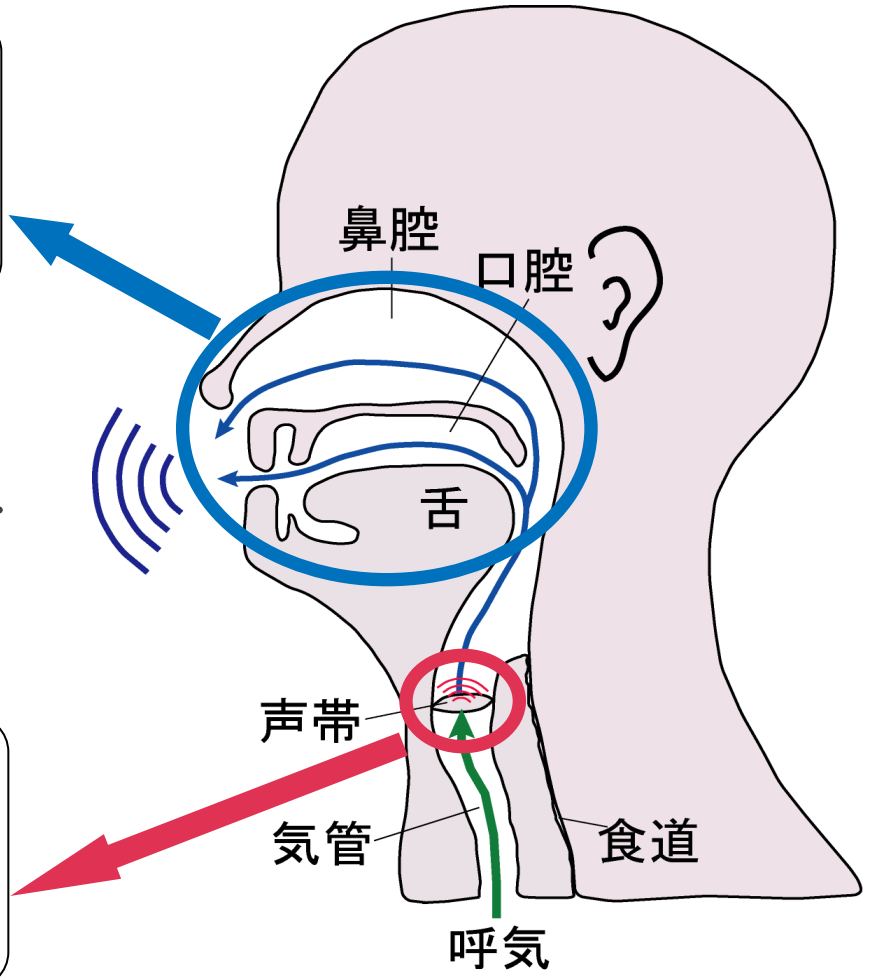
畳み込むと…

声になる！



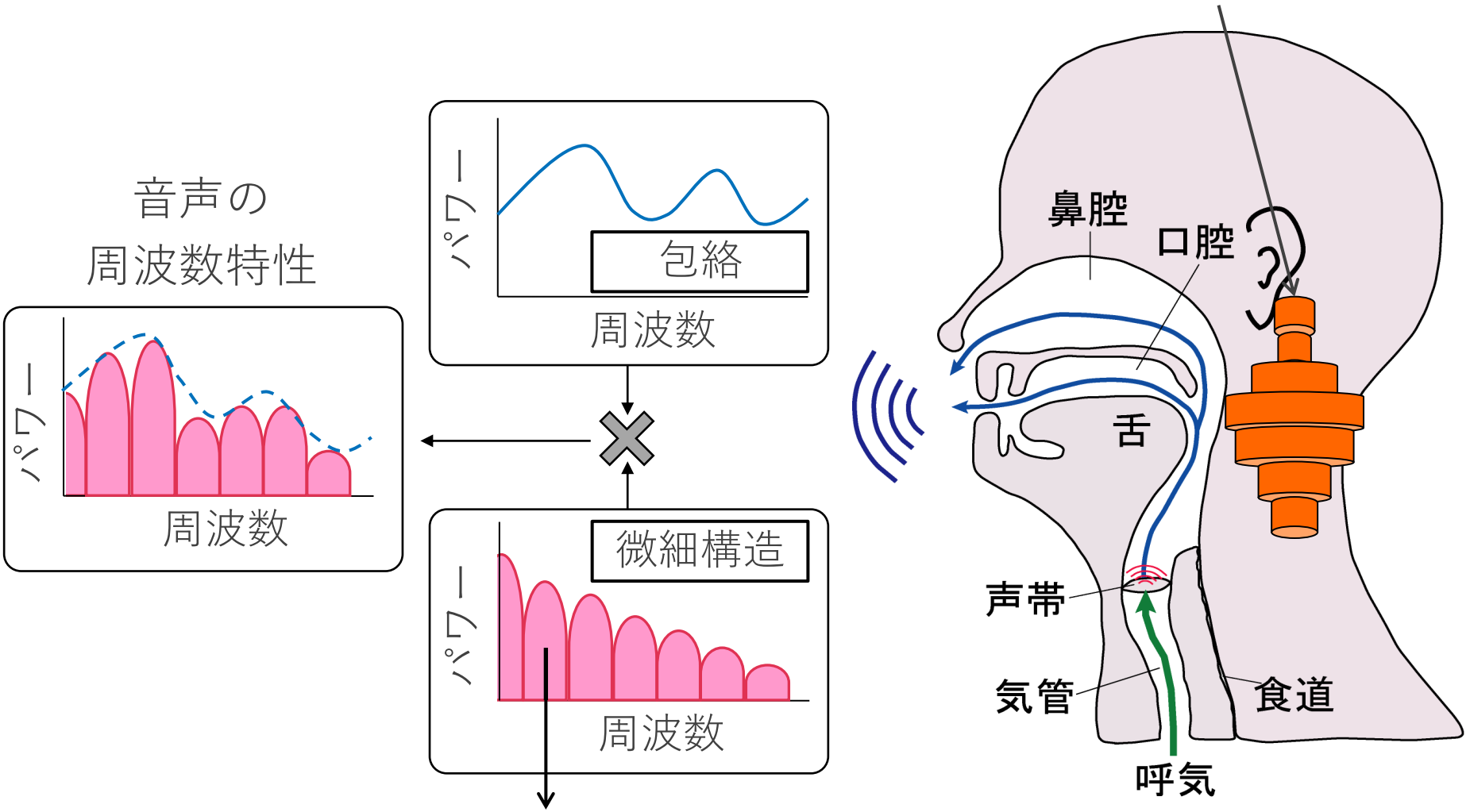
音高の生成

声帯を開閉させて，
空気を振動させる！



音声のスペクトル構造
（音声のスペクトル構造の2要素）

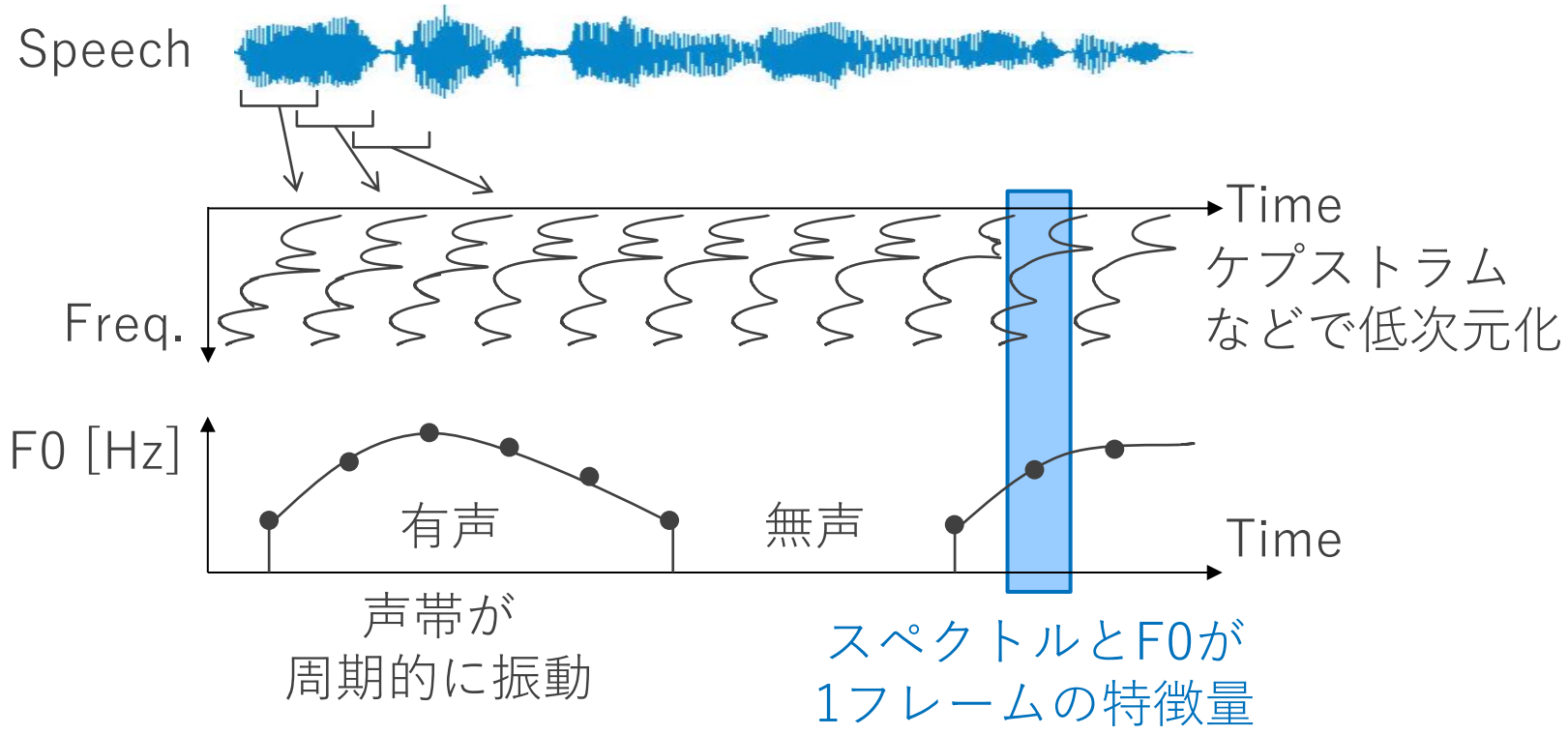
音響管連接でモデル化可能



フレーム分析と音声特徴量

音声の準定常性を仮定してフレーム分析

– 20~30ms程度であれば，音声は定常信号

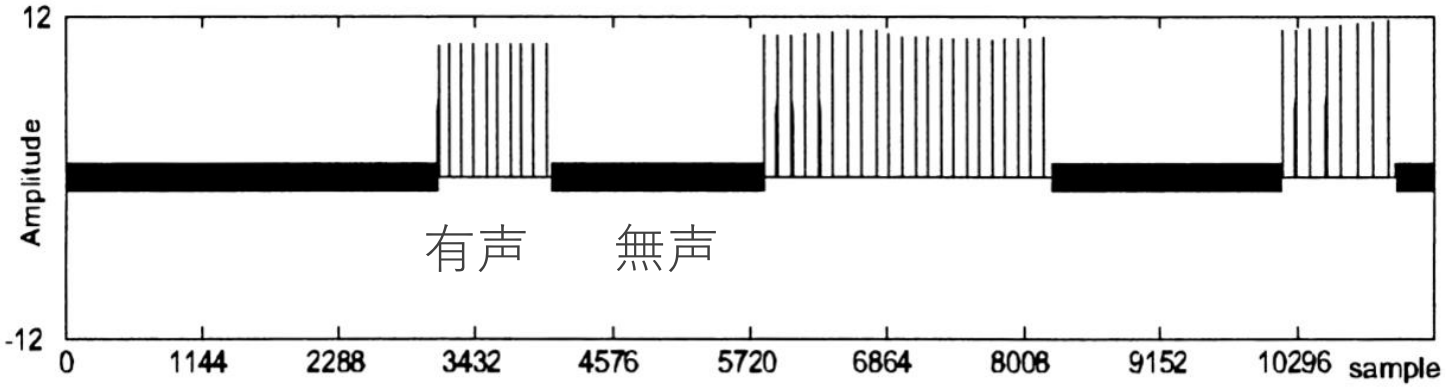


音声波形生成

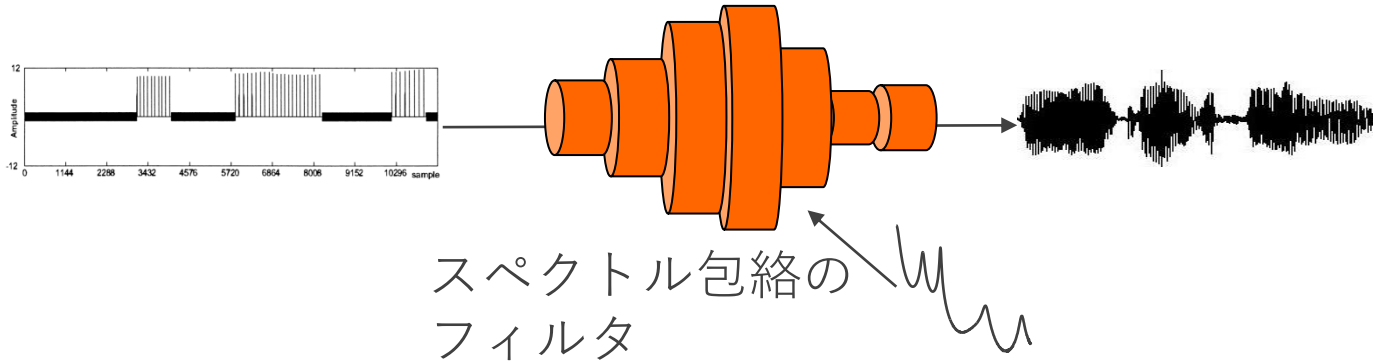
F0に基づいて音源信号を駆動

– 有声音はF0の逆数の周期のインパルス列，無声音は白色信号

[吉村 他, 2004.]



この駆動信号を，スペクトル包絡によりフィルタリング



音声に関与する言語特徴量

言語寄りの特徴量

- 言語 (mixed languageも含む)
- 形態素、Part-Of-Speech (POS)
- 係り受け

音声寄りの特徴量

- 発音・音節
  - 音韻交替：二本 (に**ほ**ん) → 三本 (さん**ぼ**ん)
- アクセント・ストレス
  - アクセント結合：にひゃく + メートル → にひゃくメートル
- リズム・等時性

①発音・音節

発音

- 発声の最小単位である音素の違い
- /a/, /i/, /u/, /e/, /o/ …

音節（シラブル）

- 音節 … 言語依存の発声単位（日本語ならほぼひらがな一つに対応）
  - 開音節 … 母音で終わる音節。日本語の”か(k a)”など。
  - 閉音節 … 子音で終わる音節。例： 英語の”it (i t)”など。
- 子音連結 … 同一音節中で連続する子音
  - 日本語 … ほとんどCV（C: 子音、V: 母音）
  - 英語 … CCCV、CCV、VCC、VCCCなどが頻出
    - straight = stra + ight

②アクセント・ストレス

音声のアクセント・ストレス

– 言語に依存してスペクトルとF0に現れる

例1: 日本語（アクセント）

高いF0

わ た し は と しょ か ん へ い き ま し た。

低いF0

例2: 中国語（アクセント: 四声）

我 去 图 书 馆

∨ ＼ ／ ー ∨ F0の変化

例3: 英語（ストレス）

ストレス

I went to the library to study for the exam.

③リズム・等時性

音声の等時性

– 言語に依存した音声的単位が、時間的に等間隔に現れる

例1: 日本語（モーラ等時性）

わ た し は と しょ か ん へ い き ま し た。

例2: 中国語（シラブル等時性）

各点は一定時間周期で現れる

我 去 图 书 馆

例3: 英語（ストレス等時性）

I went to the library to study for the exam.

アクセントは誰が決めてる？：NHKアクセント辞典

2016年に改定！

– 18年ぶり6回目。初版は1943年

前回から何が変わった？

[太田 他, 2016.]

ついに「ク＼マ」が出た！

– ”クマが出た” のアクセントは？

図 2 「熊」の NHK アナウンサー調査の結果（語別）

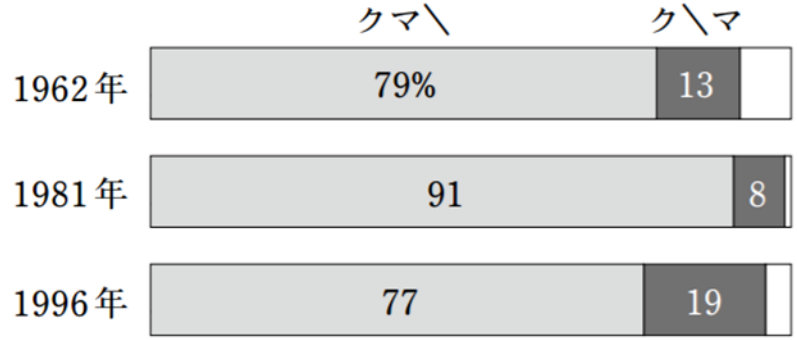
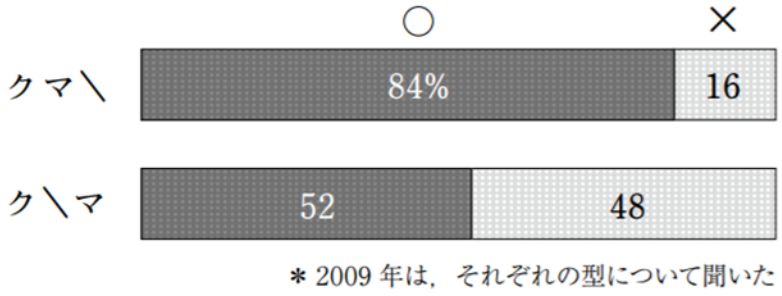


図 3 「熊」の NHK アナウンサー調査の結果（2009 年／型別）



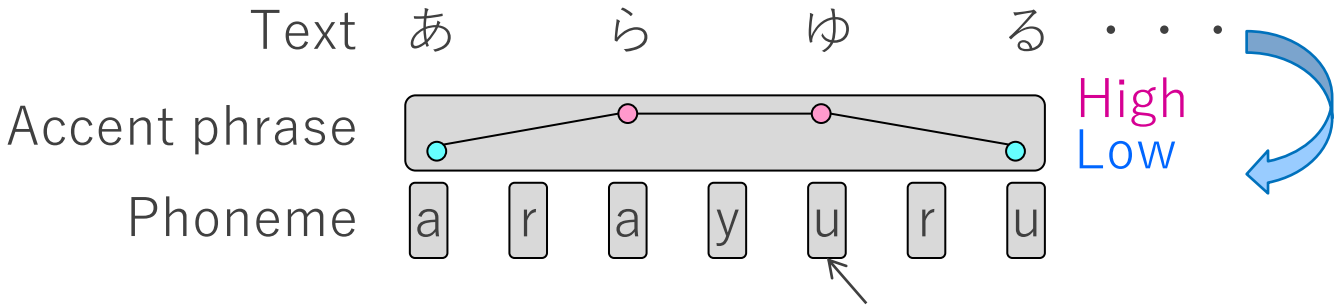
＊ 2009 年は，それぞれの型について聞いた

– 外来語は平板化
– 複合語（歩み＋寄るなど）は平板から起伏化
– などなど

ここまでまとめ

言語特徴量

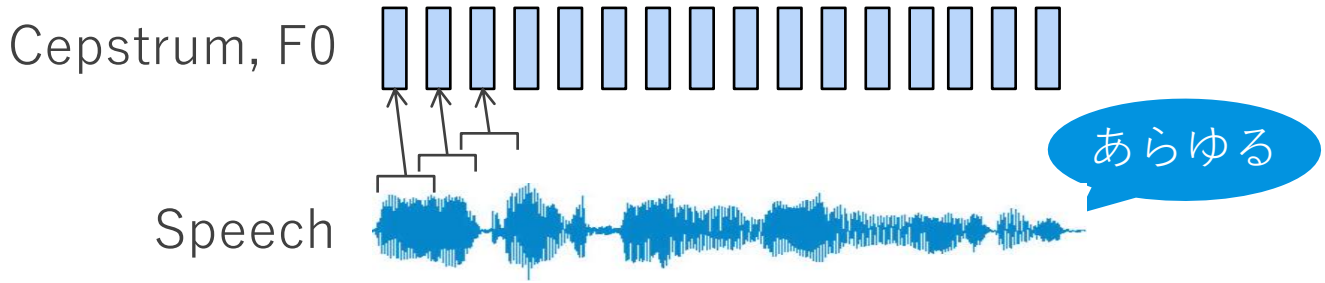
– テキストから，音素・音節・アクセントなどの特徴量を抽出



前の音素は/y/，後の音素は/r/，高いアクセント，
形容詞である単語の中の3モーラ目である/u/

音声特徴量

– 音声から，声道・声帯の特徴量を抽出



音声合成

音声合成の長い歴史

1939: Voder（ベル研究所）
- その前身はvocoder（voice + coder）

1961: 音声合成による ‘Daisy Bell’（ベル研究所）

~1990: フォルマント音声合成
- 専門家による音声規則設計

1990~: 素片選択型音声合成
- ダイフォン音声合成，単位選択型音声合成

1995~: 統計的パラメトリック音声合成
- HMM・DNN音声合成
- GMM・DNN音声変換

事前収録音声コーパスを用いて合成を行う
コーパスベース合成方式

コーパスベース音声合成の種類

素片選択型合成 (unit selection synthesis)

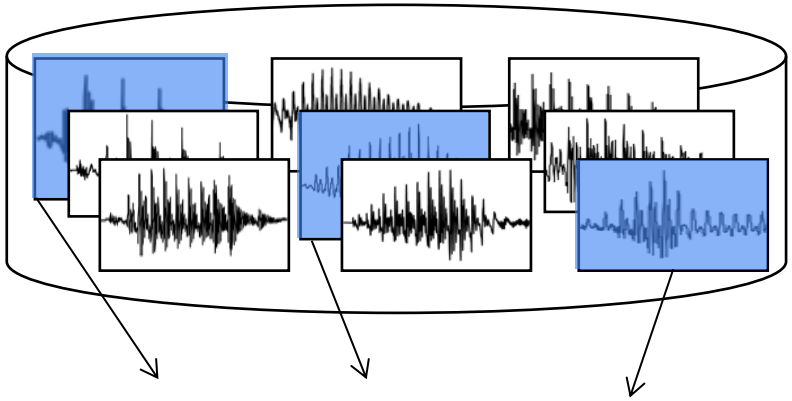
- 音声波形・パラメータを保存し、その接続・加工で音声合成
- 長所：非常に肉声感の高い合成音
- 短所：声質を制御しにくい、フットプリントが大きい

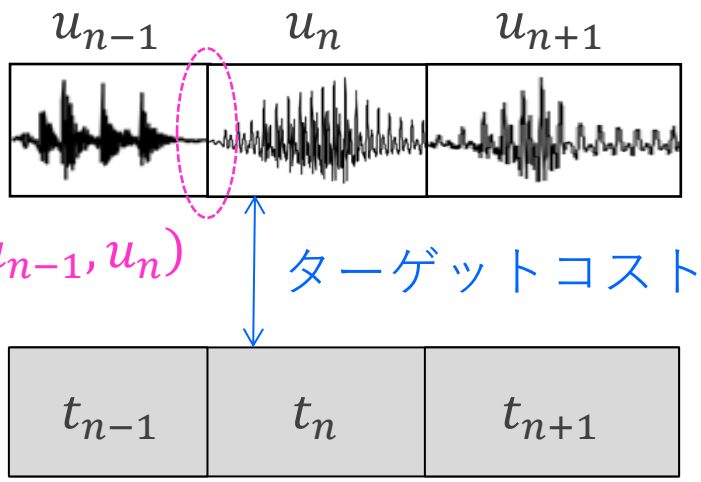
統計的音声合成 (statistical speech synthesis)

- 音声波形・パラメータを統計モデルでモデル化
- 長所: 声質を制御しやすい、フットプリントが小さい，機械学習の知見を大いに使える
- 短所: 低い音質 (最近は非常に改善されてきた)

サンプルベース方式 (波形接続型)

音声データベースにある
音声セグメント





接続コストとターゲットコストの和を最小化するように
音声セグメントを選択

コスト関数

最小化されるコスト関数

– これを最小化するように セグメント系列 $u_1, \cdots u_n, \cdots, u_N$ を決定

– 動的計画法などを利用

$$C^{(\mathrm{us})} = \sum_{n=1}^{N} \omega_{\mathrm{t}}^{(n)} C_{\mathrm{t}}^{(\mathrm{us})}(t_n, u_n) + \sum_{n=2}^{N} \omega_{\mathrm{c}}^{(n)} C_{\mathrm{c}}^{(\mathrm{us})}(t_n, u_n)$$

ターゲットコストの重み　接続コストの重み

通常，ヒューリスティックに決定

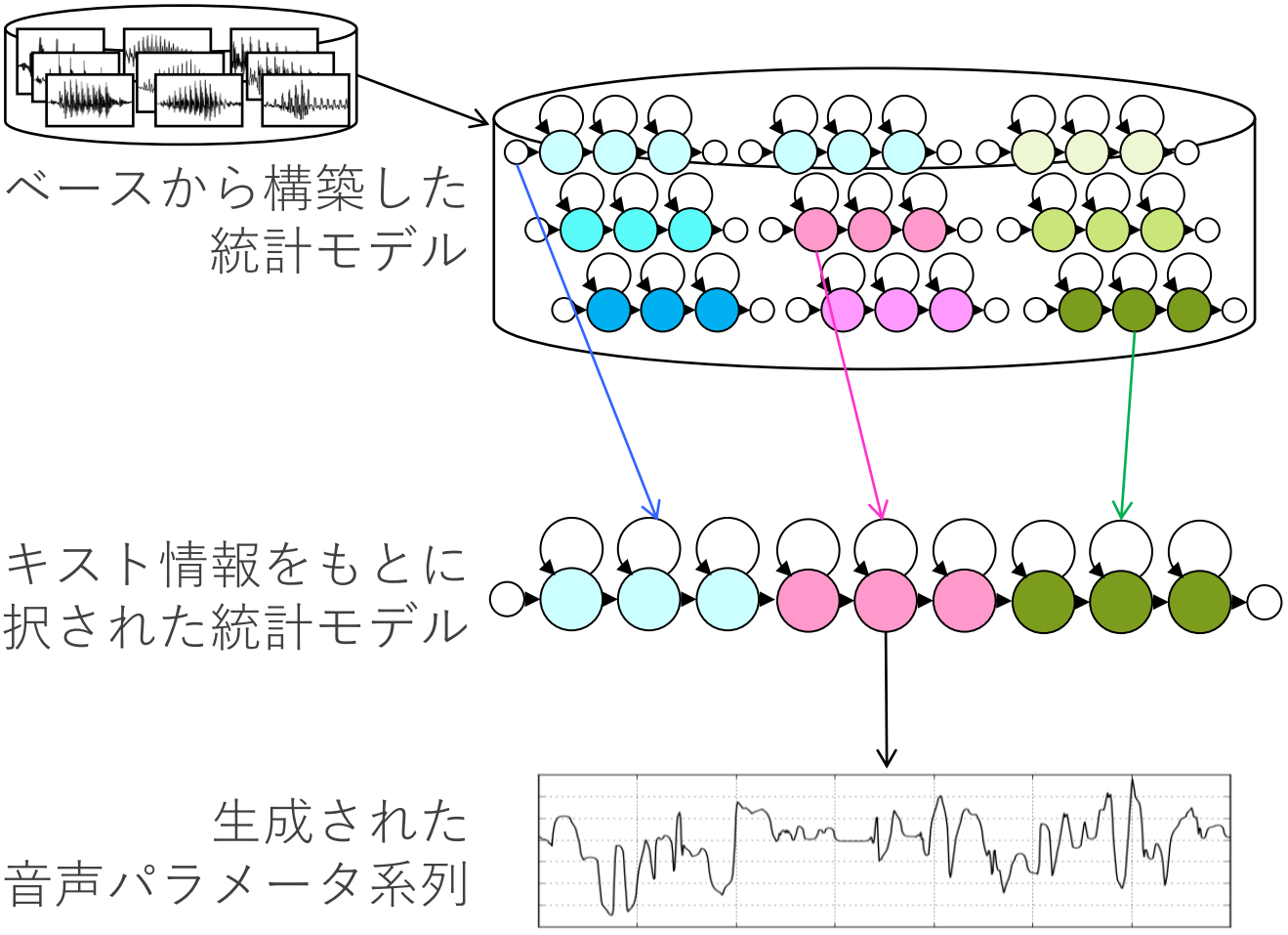
コスト関数の例（テキストからの予測特徴量をF0系列とする）

– ターゲットコスト：予測特徴量とセグメントの特徴量の二乗誤差

– 接続コスト：セグメントの接続フレーム前後の変動量

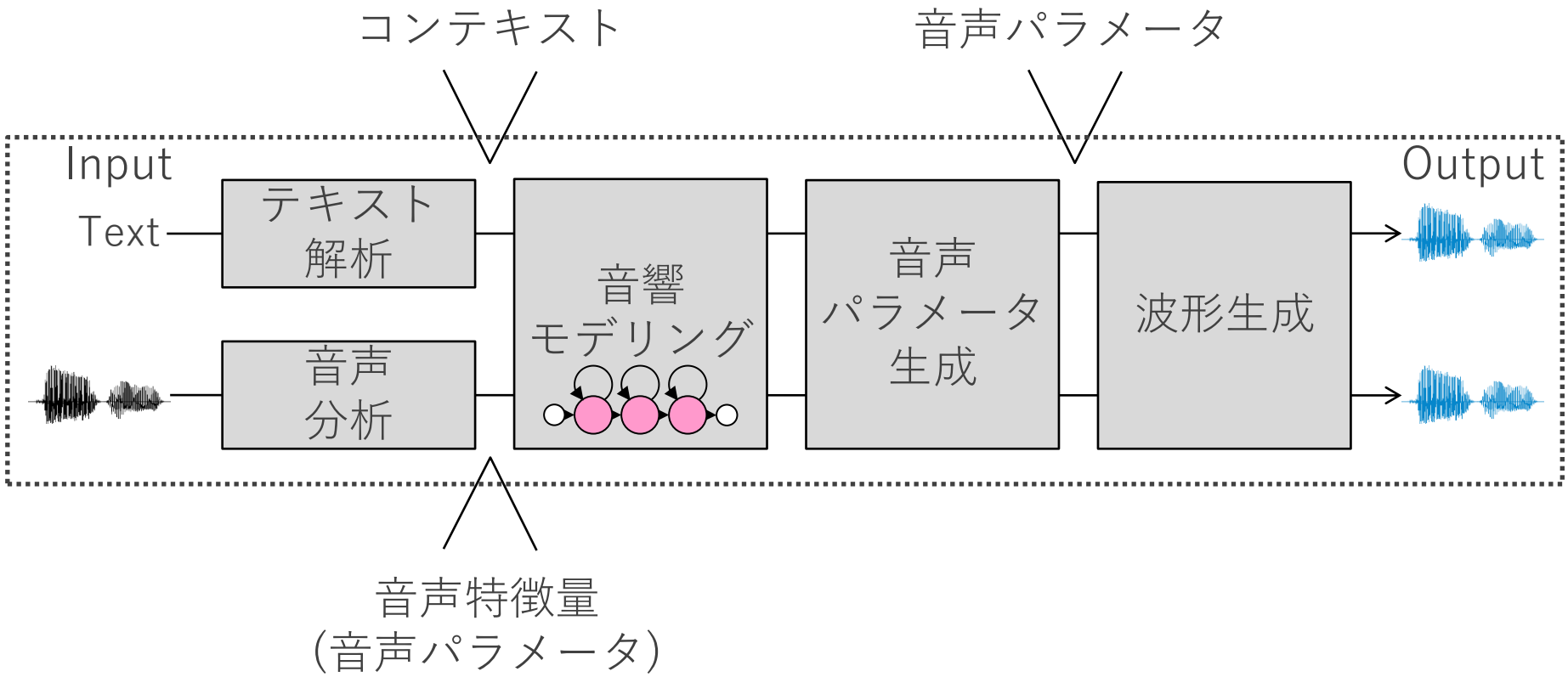
• 各コストがサブコストの重み付き和の場合もある

統計ベース方式



音声波形の代わりに統計モデルを保存して
統計モデルから音声パラメータを生成

統計ベース方式の手順


コンテキスト
音声パラメータ
Input
Text
テキスト
解析
音声
分析
音響
モデリング
音声
パラメータ
生成
波形生成
Output
音声特徴量
(音声パラメータ)

統計的音声合成の方式

テキスト音声合成

- Hidden Markov Model (HMM)
- Gaussian Process Regression (GPR)
- Classification And Regression Tree (CART)
- Hybrid (unit selection & statistical models)
- Deep Neural Network (DNN)
  - FFNN/LSTM, GAN, MMD, WaveNet, Seq2Seq, MemoryNet, …

音声変換 (テキストを介さず，音声を音声に直接変換する手法)

- Gaussian Mixture Model (GMM)
- Nonnegative Matrix Factorization (NMF)
- Hybrid
- DNN

* テキスト依存音声変換 (音声認識＋テキスト音声合成) もあるが，本講義では扱わない

HMM音声合成

歴史

- 1990年代初頭にHMM音声認識が隆盛
- 「音声認識が上手くいくなら音声合成もイケるだろう」
  - 後述するDNN音声合成も同様
- 1995年頃，名工大 徳田先生らによって提案 [Tokuda et al., 1995.]

> The hidden Markov models (HMMs) are widely-used statistical models to characterize the sequence of speech spectra and have successfully been applied to speech recognition systems. From these facts, we surmise that the HMMs are also useful for speech synthesis by rule, and we have pro-

貢献

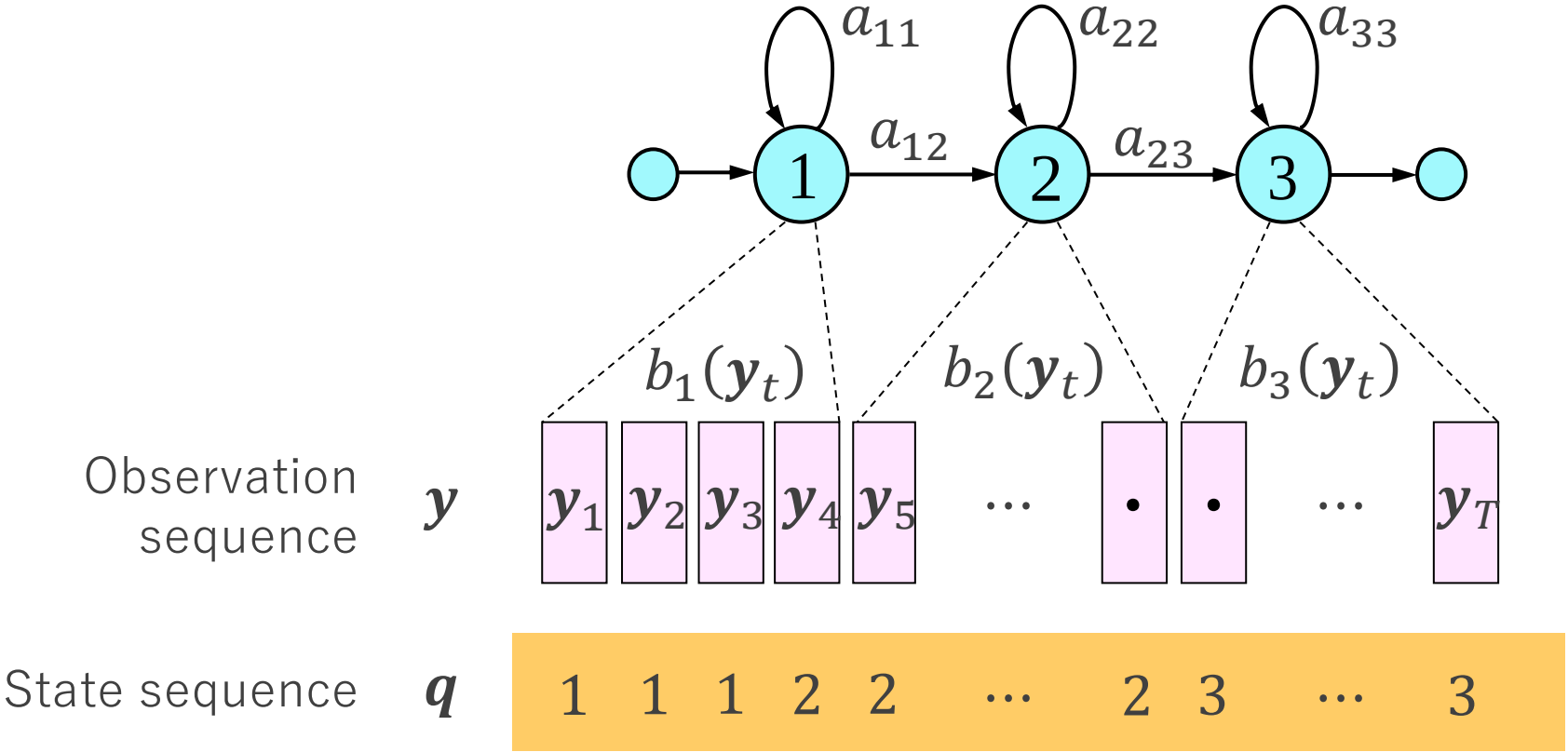
- 現在に至るまでの，統計的音声合成の基盤を確立
- ヒューリスティックだった音声合成に機械学習を導入し，音声合成エンジンの (半)自動構築を可能に

隠れマルコフモデル (HMM) とは

HMM：状態系列 q の隠れたマルコフ連鎖

– モデルパラメータ $\boldsymbol{\lambda}$ は遷移確率 a_{pq} と出力確率 $b_q(\cdot)$

• $b_q(\cdot) = N(\cdot, \mu_q, \Sigma_q)$ (正規分布)とする

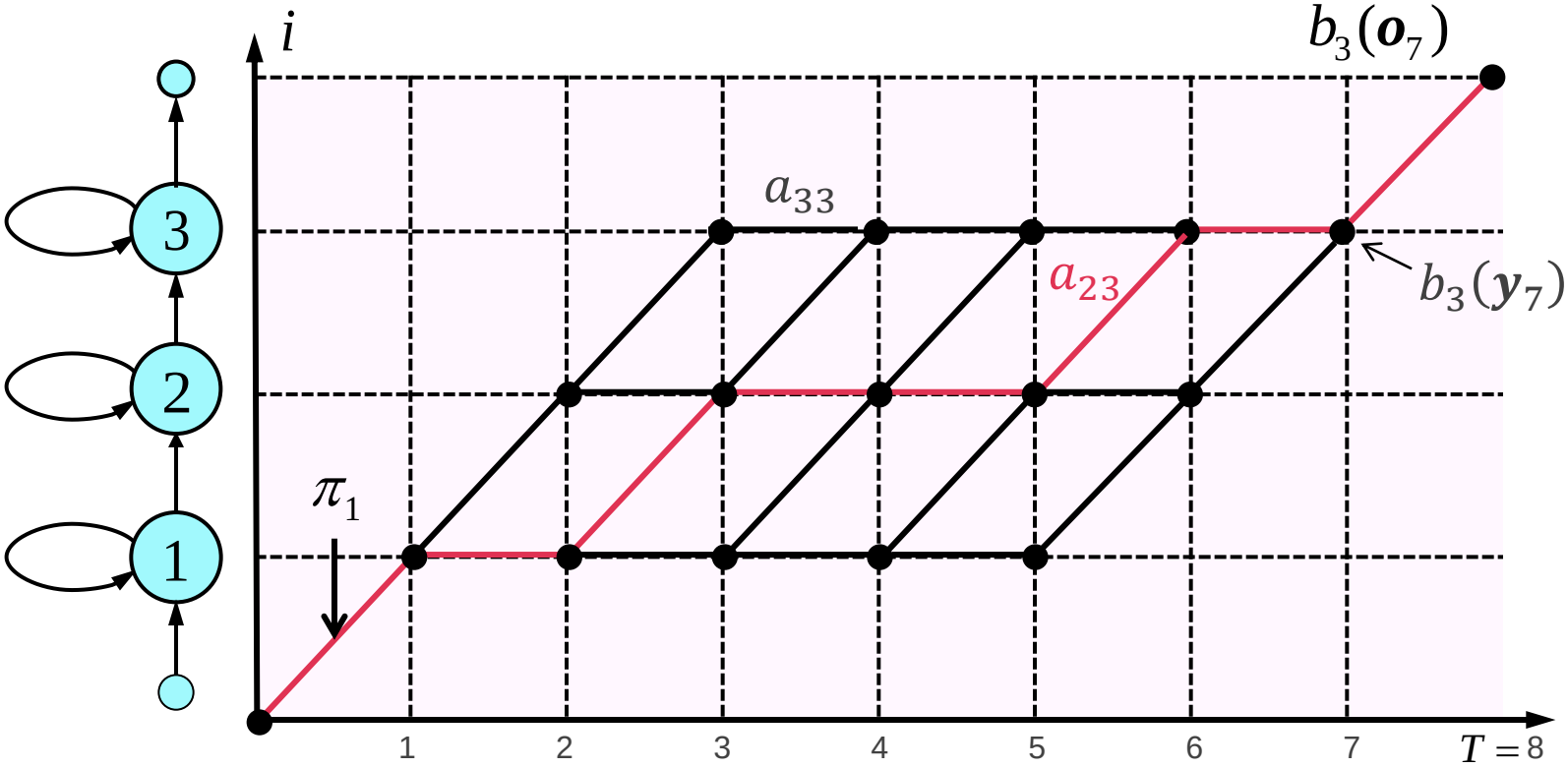


HMMの学習

最尤基準に基づくモデルパラメータの学習

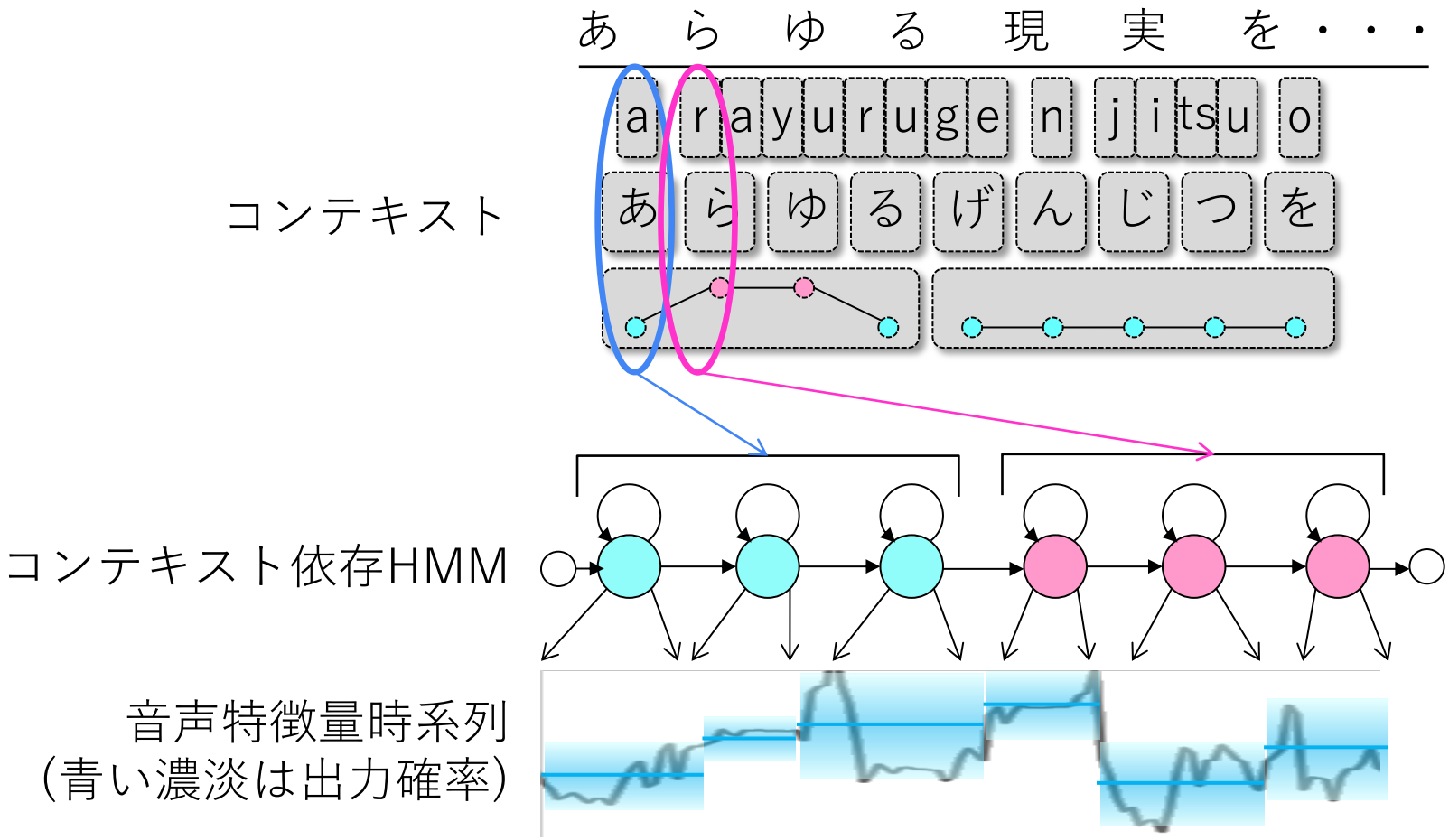
– 状態系列 $\boldsymbol{q}$ を隠れ変数とした EM アルゴリズム

$$\hat{\boldsymbol{\lambda}} = \operatorname{argmax} \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{y}|\boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q}|\boldsymbol{\lambda})$$



コンテキスト依存HMMの学習

各コンテキスト毎にHMMを学習．各HMM 状態でセグメ ントの最初・真ん中・最後あたりをモデル化

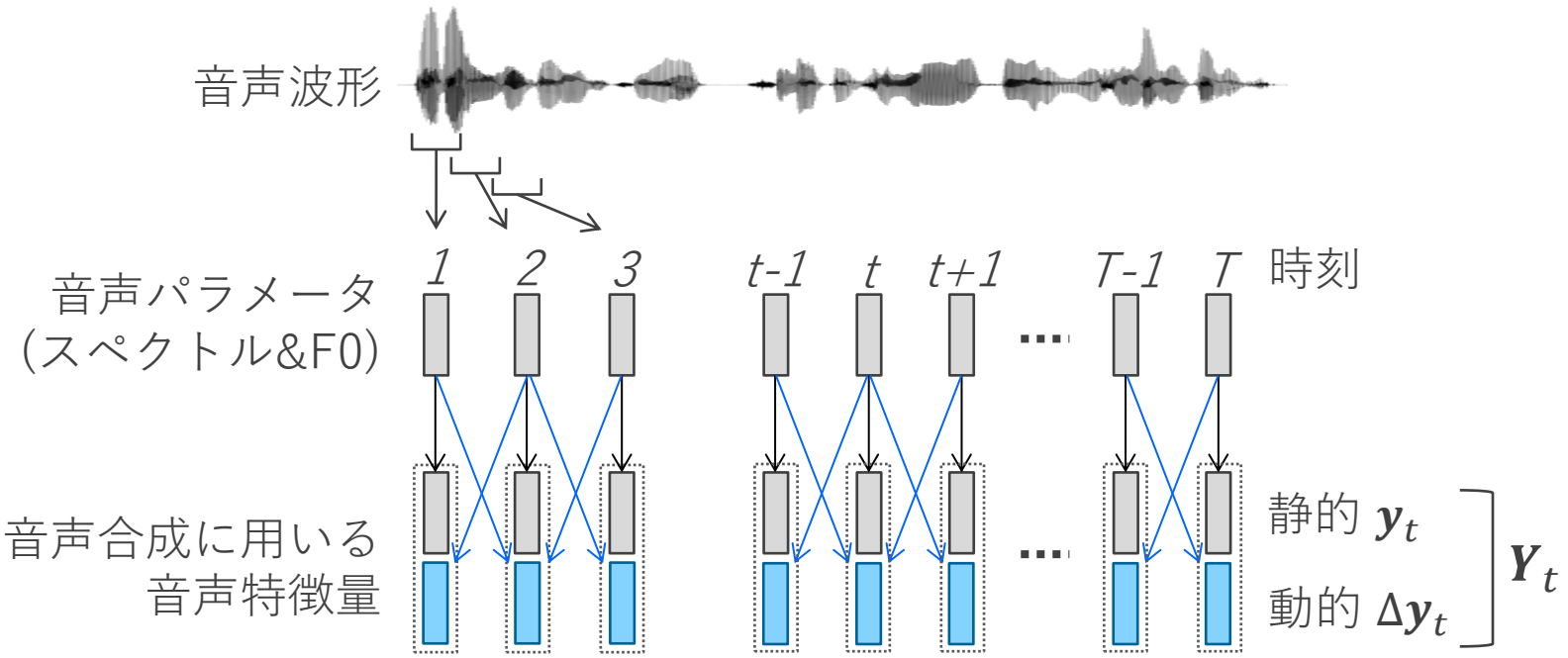


動的特徴量の導入

動的特徴量：特徴量の時間変化

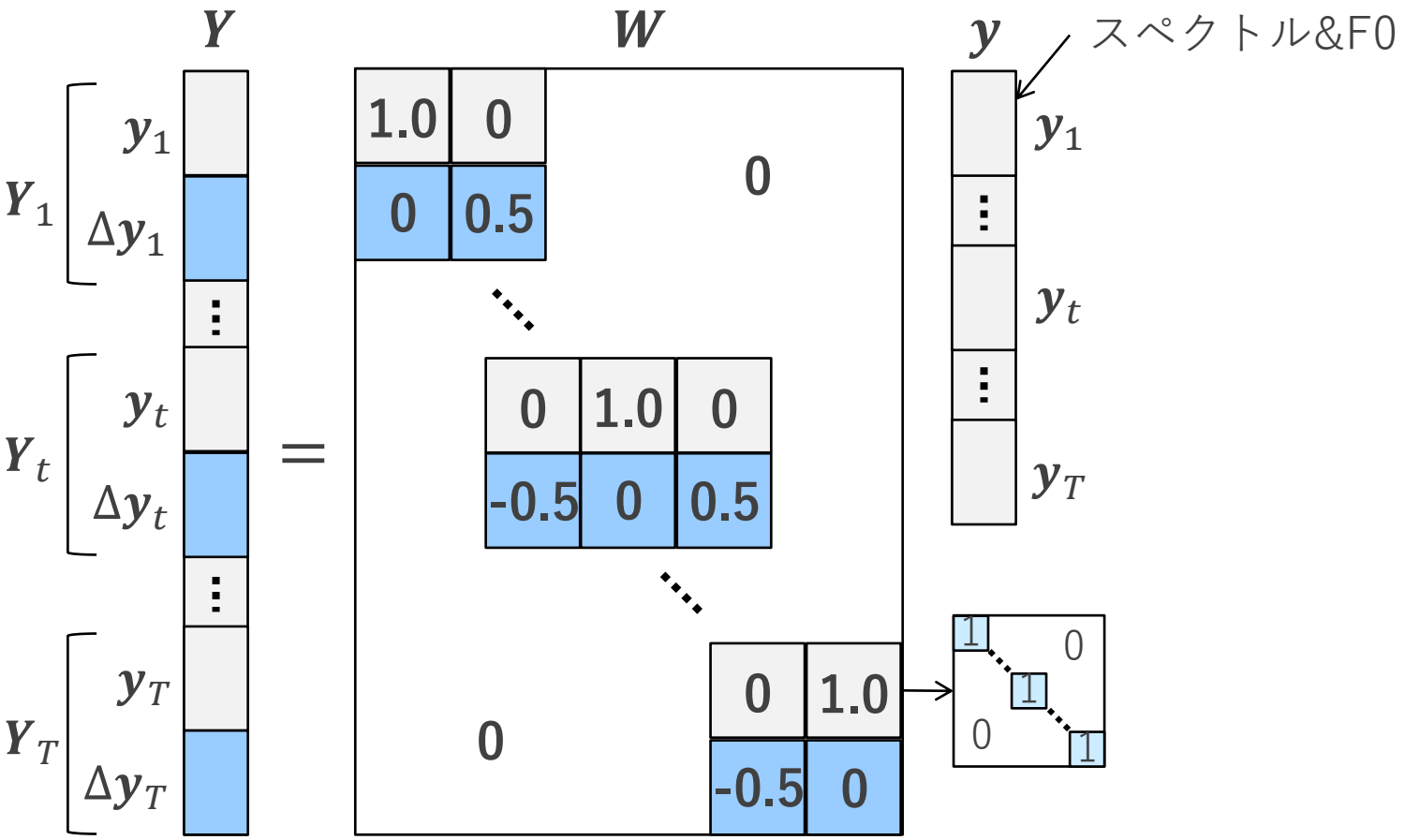
– 差分量を導入し，静的・動的特徴量系列からHMMを学習

– (理由は後述)



$$\Delta\boldsymbol{y}_t = 0.5(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t-1})$$

動的特徴量計算の行列表現



F0系列のモデリング：MSD-HMM

[Tokuda et al., 2002.]

F0系列は，時刻毎に次元数の変化する特徴量系列

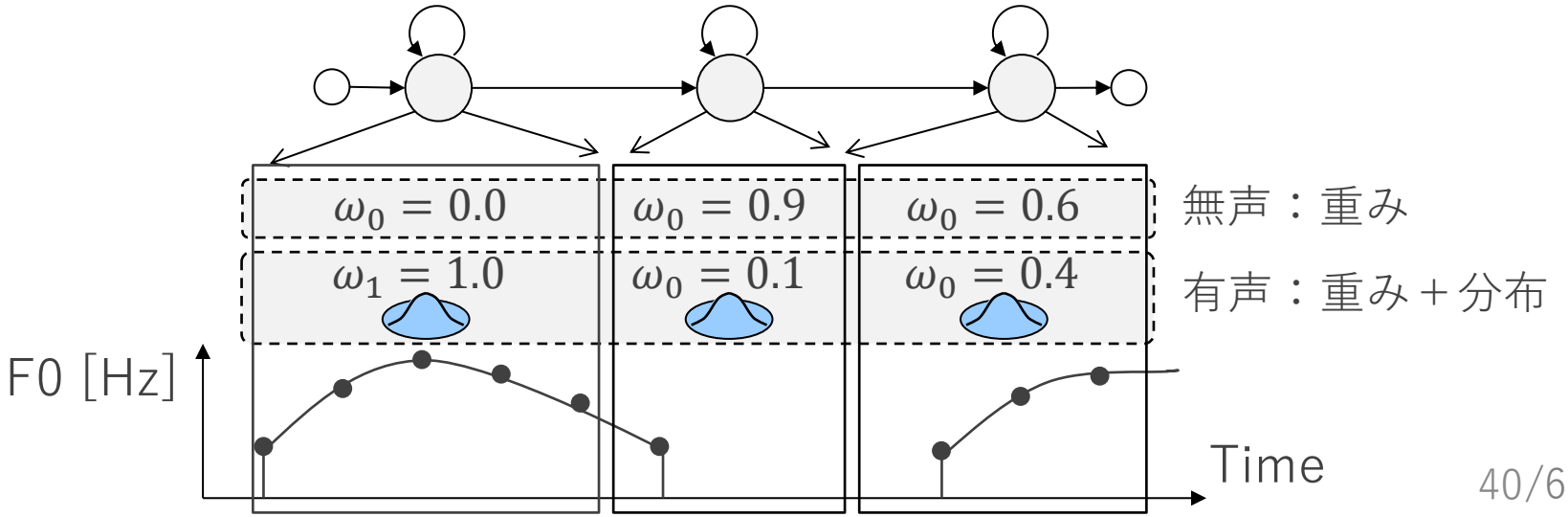
- 単一の出力分布（正規分布）ではモデル化できない
- 有声音は1次元，無声音は0次元とみなす

MSD-HMM（Multi-Space probability Distribution HMM）

- 複数次元の特徴量に対応する確率分布を重み付きで持つ

$$P(\boldsymbol{y}_t) = \sum_{\text{all } d} \omega_d P_d(\boldsymbol{y}_t)$$

$P_d(\boldsymbol{y}_t)$: d次元特徴量に対する確率（密度）



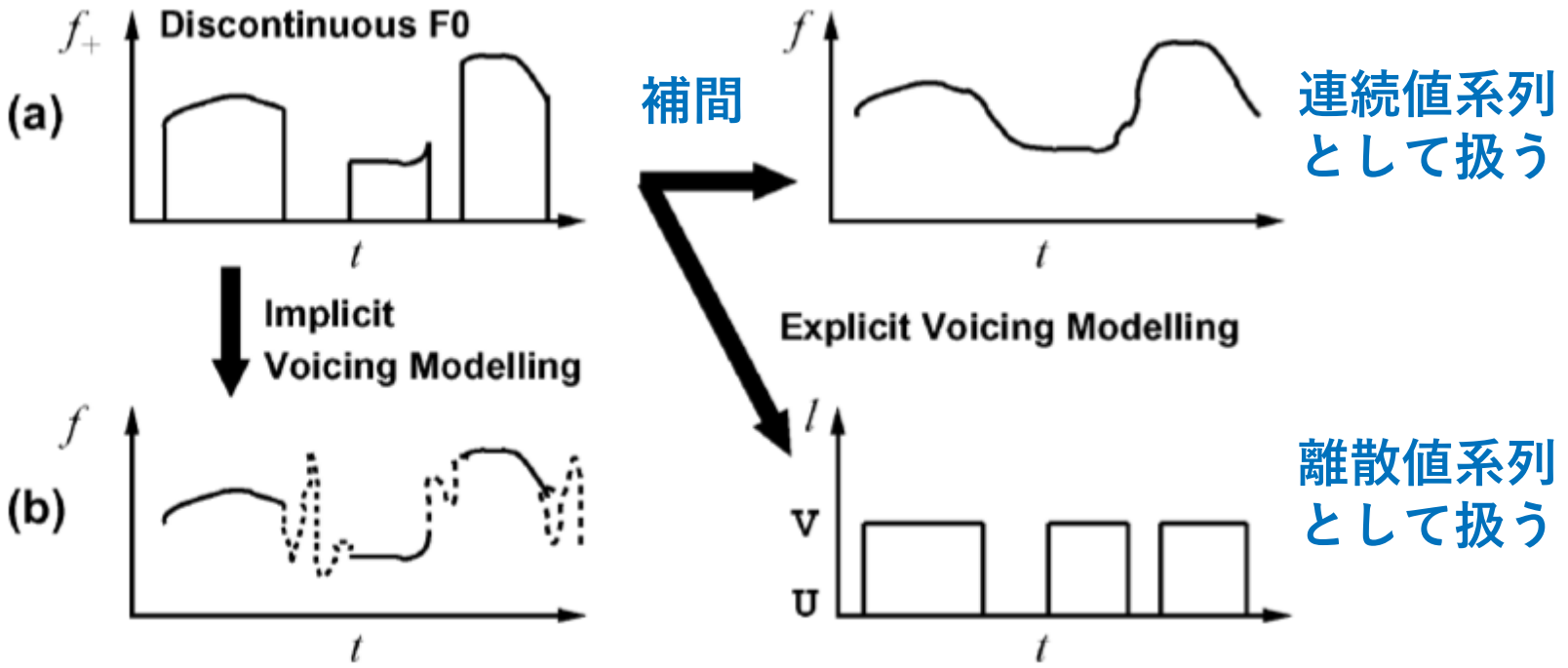
F0系列のモデリング：連続F0モデル

[Yu et al., 2011.]

MSD-HMMによるモデリングの欠点

– 確率と確率密度のスケールの違い．動的特徴量との整合性の乏しさ

連続F0モデル：連続F0系列と有声／無声ラベルに分割



コンテキストクラスタリング

[Shinoda et al., 2000.]

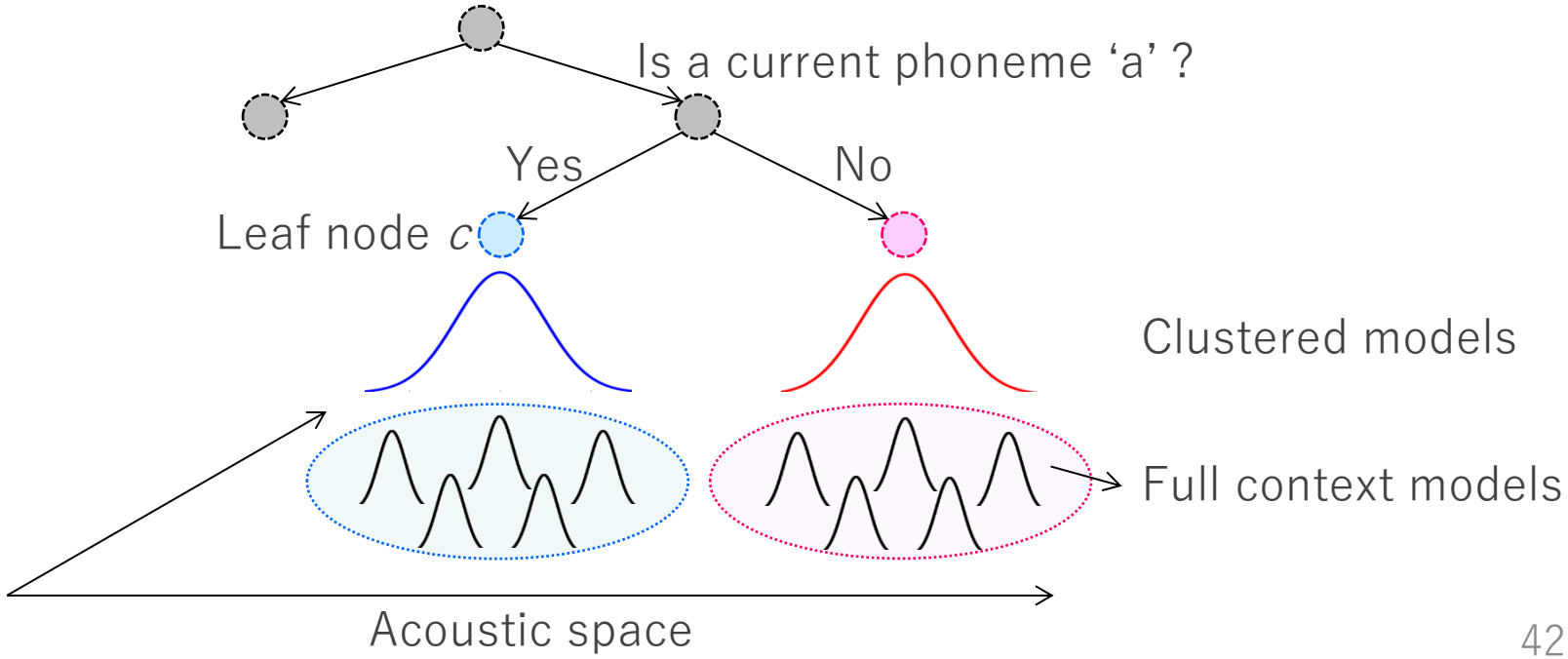
コンテキストのスパース性の問題

– 素性の多さから同じコンテキストは学習データに二度と登場しない

コンテキストクラスタリング

– HMMの出力分布をMDL基準＋二分木でクラスタリング

– 分割要素はコンテキストに対する質問

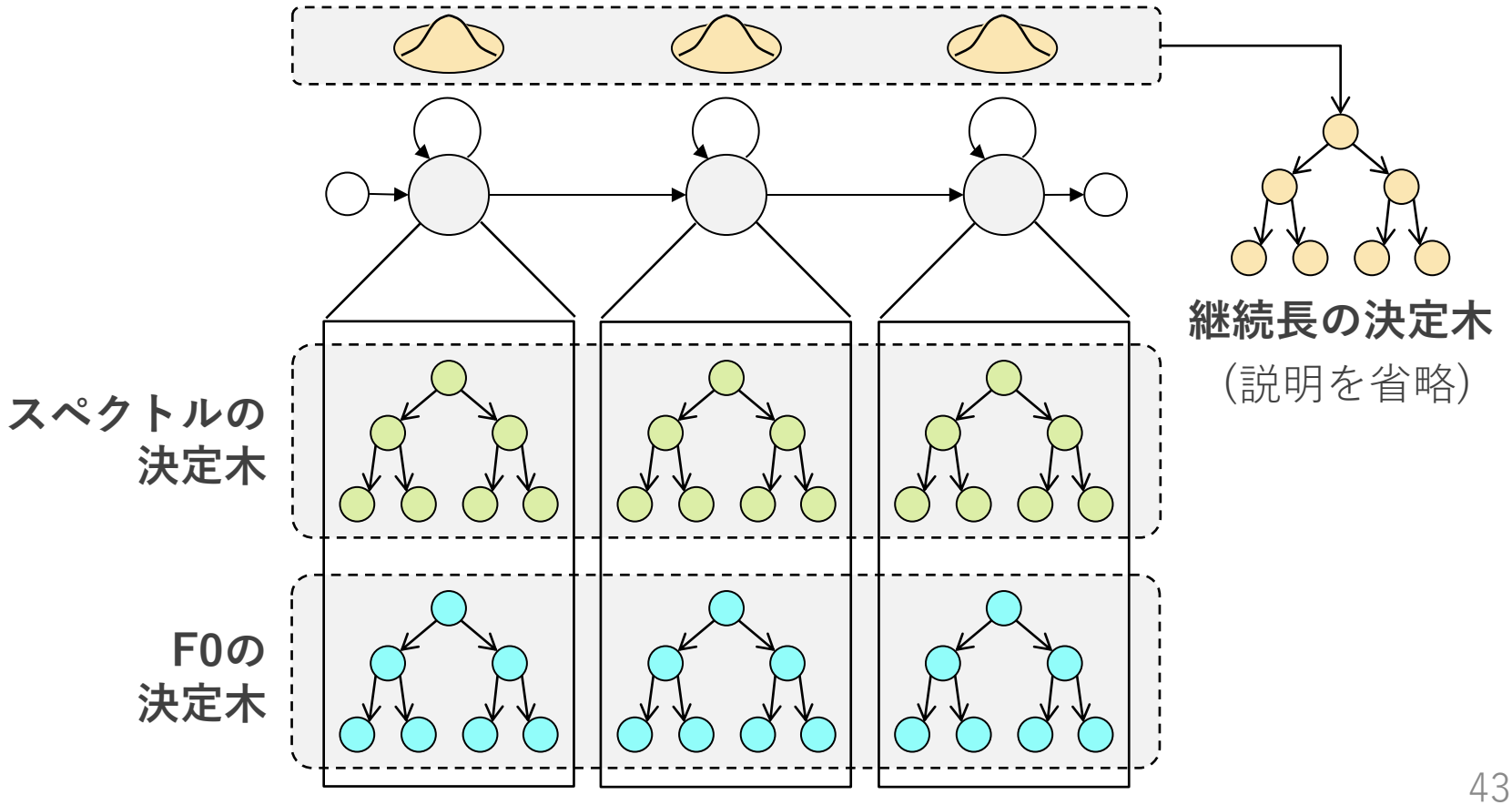


最終的に学習される音響モデル

[Tokuda et al., 2013.]

最終的に得られるモデル

- 特徴量毎・HMM状態毎に二分木クラスタリングを行う.
- 各リーフに単一の出力分布を有する.



音声合成：音声パラメータの確率分布

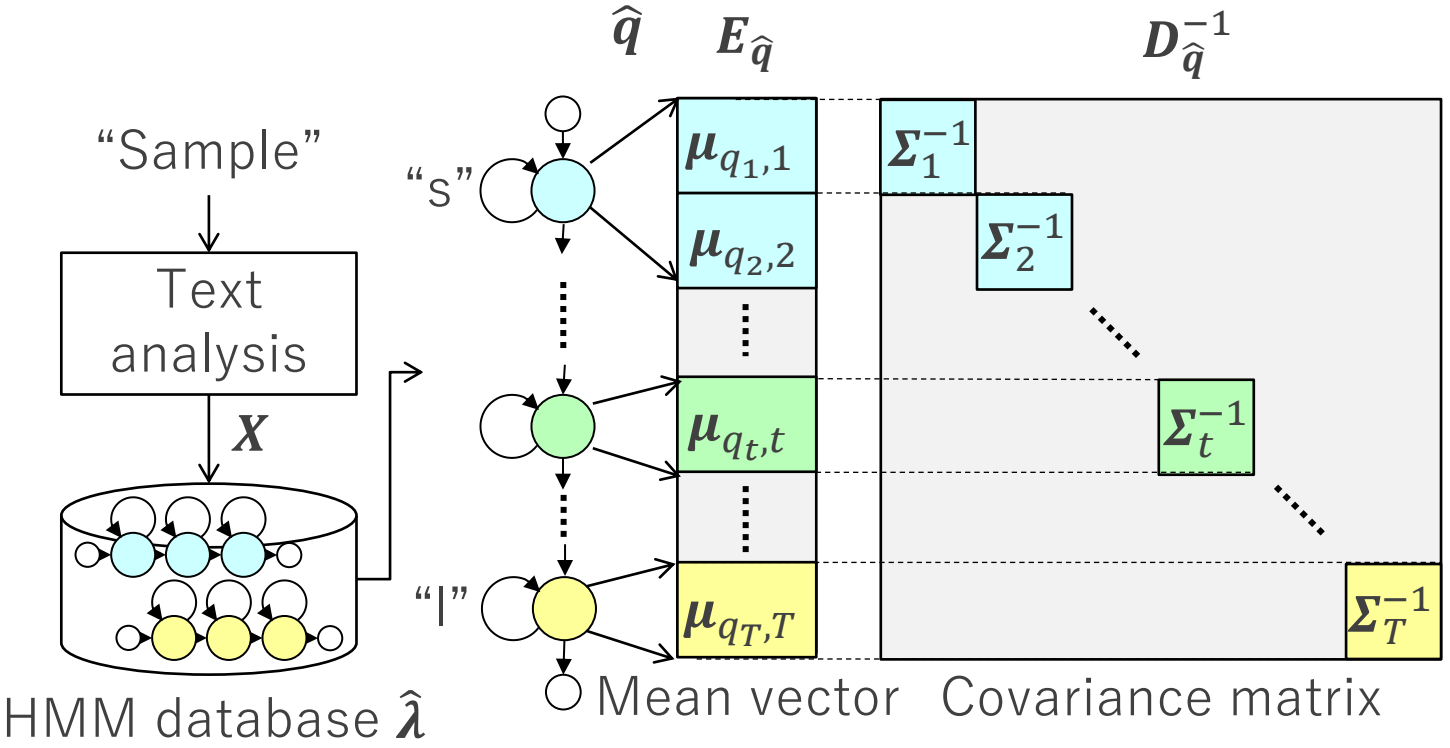
[Tokuda et al., 2000.]

入力テキストと学習済みHMM λ から音声パラメータ $\hat{y}$ を生成

– 決定木をたどり，対応する出力分布を決定．継続長 (時間長) を Viterbi系列 $\hat{q}$ で近似すると，Yの生成確率は正規分布で得られる

$$P(Y|\hat{q},\hat{\lambda}) = N(Y; E_{\hat{q}}, D_{\hat{q}})$$

Yは静的・動的特徴量系列



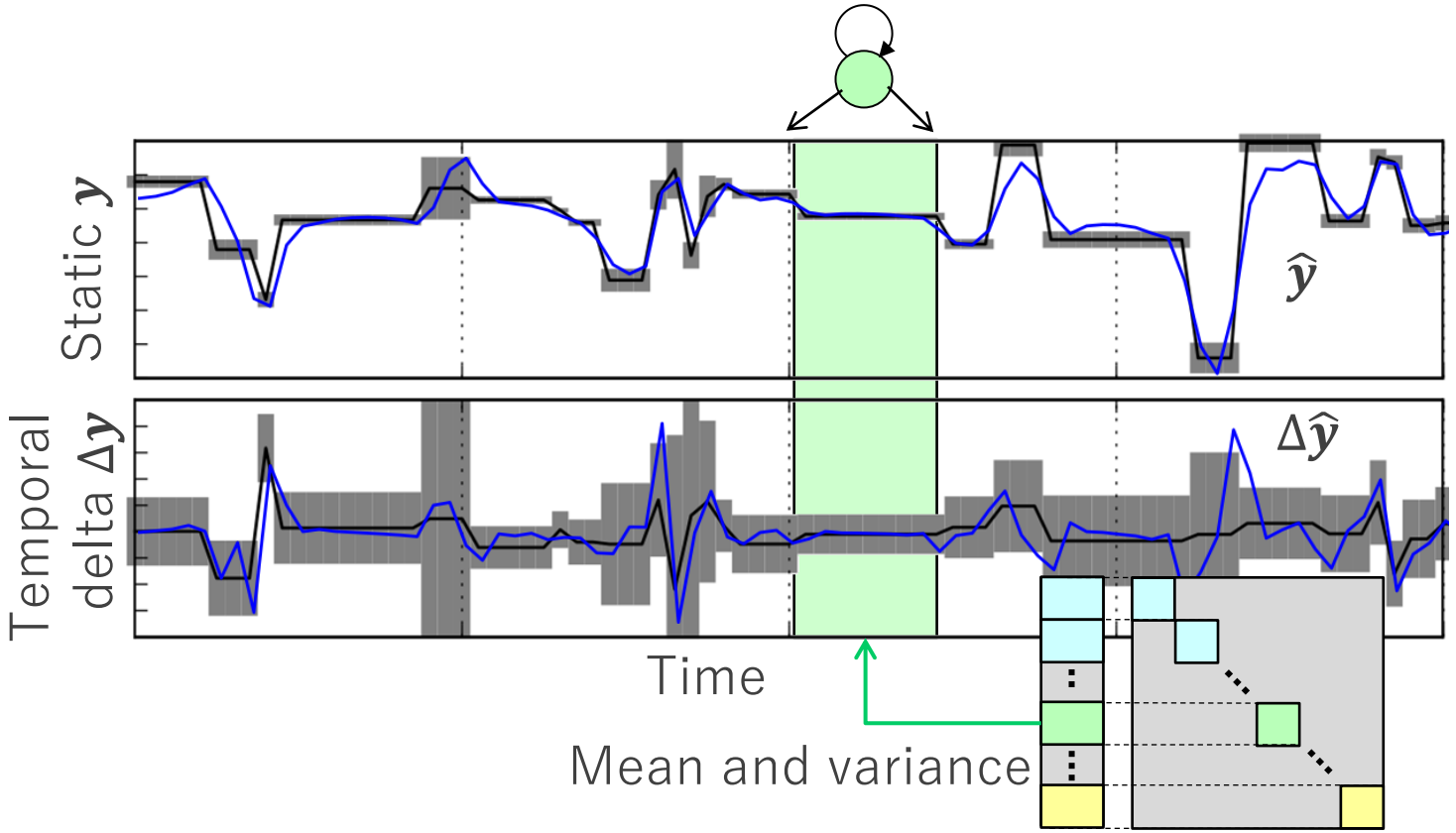
動的特徴量を考慮した最尤パラメータ生成

[Tokuda et al., 2000.]

音声パラメータ $\hat{\boldsymbol{y}}$ は動的特徴量の制約下の最尤推定で得られる

– $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ (少し前のページを参照)

$$\hat{\boldsymbol{y}} = \operatorname{argmax} N\left(\boldsymbol{Y}; \boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right) = \operatorname{argmax} N\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right) = \left(\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}$$



何故，動的特徴量を用いるか？

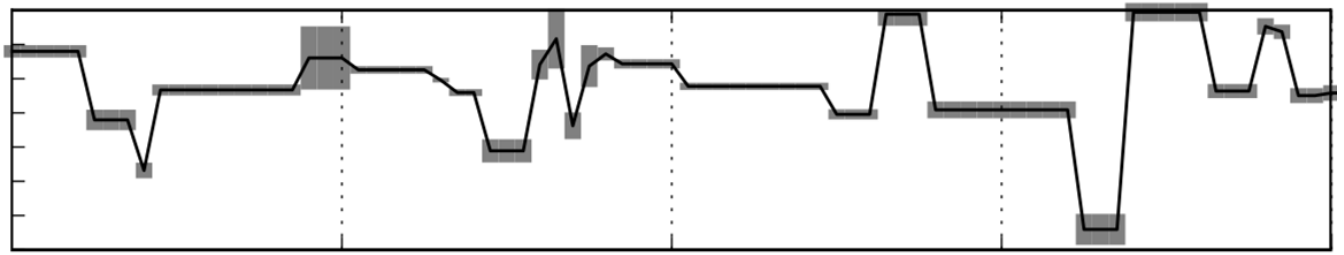
[Tokuda et al., 1995.]

HMMは時間を量子化する

- Tフレームの系列を (例えば) 3状態のHMMで表現.
- 状態内は定常と仮定

動的特徴量を用いずに最尤推定すると…？

- 平均のみが出力され，階段状の音声パラメータ系列に → 不連続



HMMからサンプリングすれば…？

- HMMからのサンプリングでは，音質が顕著に劣化する
- (時間量子化，正規分布の過程などが原因)

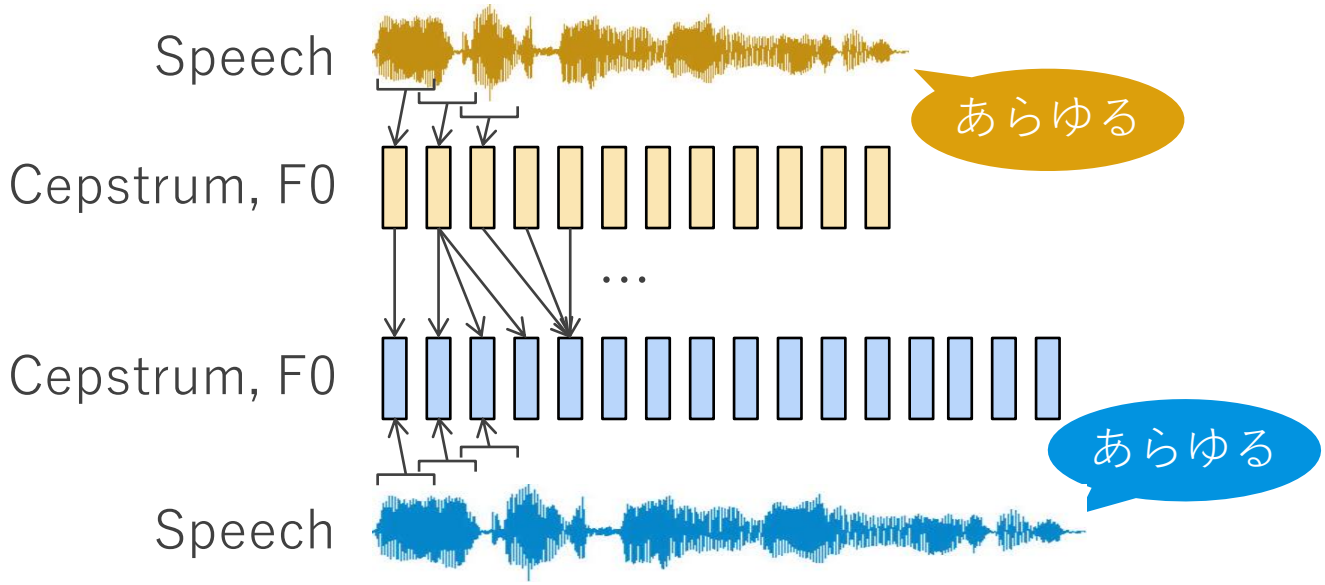
GMM音声変換

歴史

– 1998 年，クレタ大 Dr. Stylianou らによって提案
– HMM音声合成の技術を応用し，名大 戸田教授らにより発展
– 同一文を発話した音声対から自動学習

事前準備

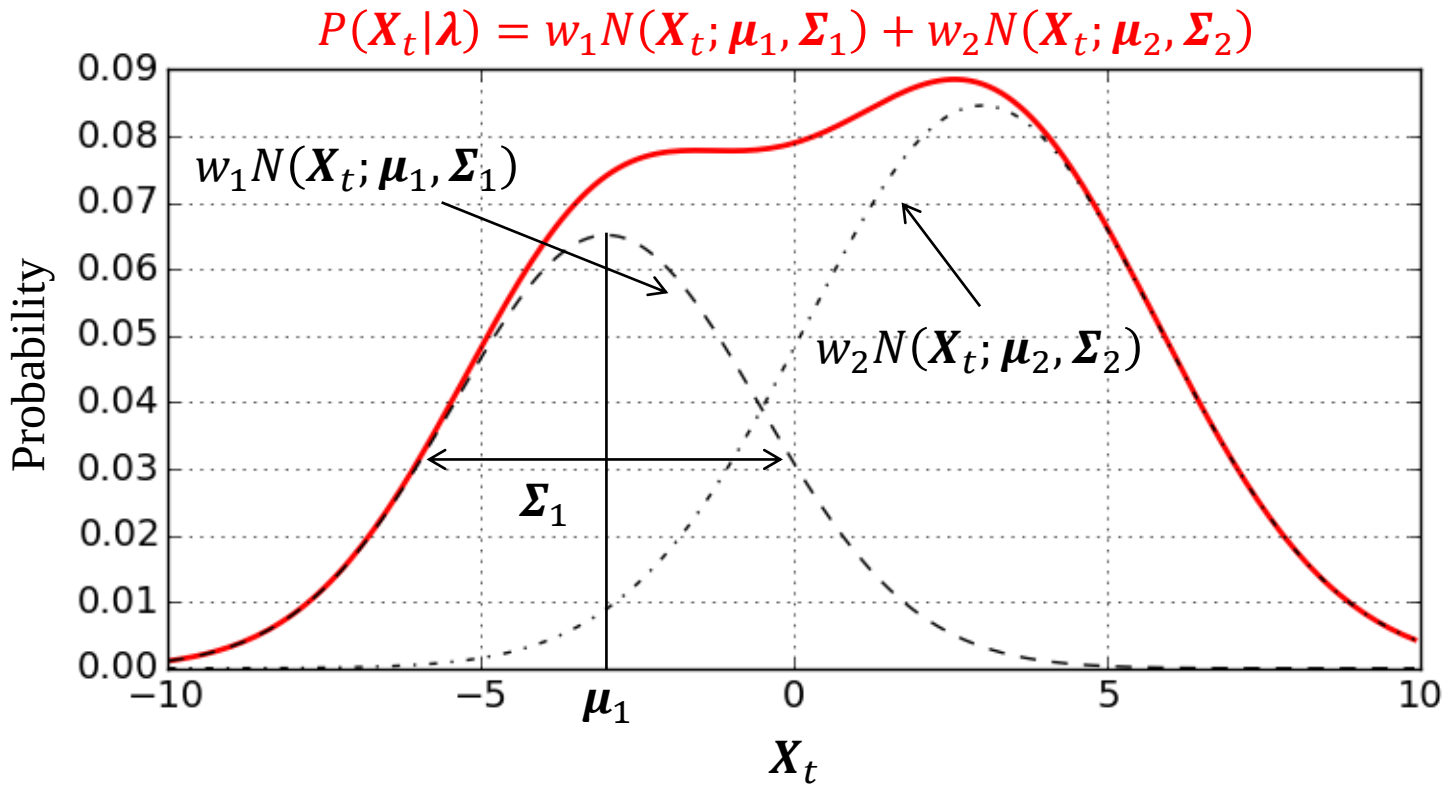
– 入出力話者の話速の違いは DTW (動的時間伸縮) で補正



GMM (Gaussian Mixture Model)とは

正規分布の混合モデル (下図は 2 混合).

モデルパラメータ $\boldsymbol{\lambda}$ (重み w_q, 平均ベクトル μ_q, 共分散行列 Σ_q) は EM アルゴリズムで推定可能



GMMによる同時確率のモデル化

[Stylianou et al., 1998.]

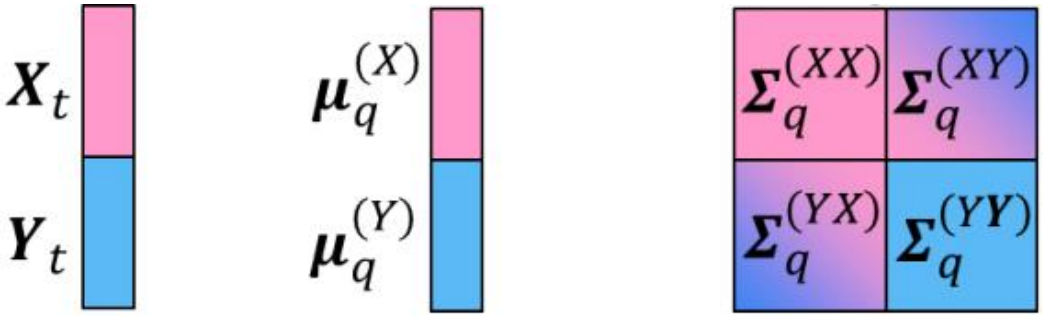
入出力話者から音声パラメータ(スペクトル, F0)を抽出

- 入力 $\boldsymbol{X}_t$, 出力 $\boldsymbol{Y}_t$ (t はフレームインデックス)
- それぞれ，静的・動的特徴量から成る

同時確率をGMMでモデル化

- 学習は，通常のGMMと同様に学習可能

$$P\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{bmatrix}|\boldsymbol{\lambda}\right)=\sum_{q=1}^{\boldsymbol{1}}\omega_q N\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_q^{(X)}\\ \boldsymbol{\mu}_q^{(Y)}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_q^{(XX)} & \boldsymbol{\Sigma}_q^{(XY)}\\ \boldsymbol{\Sigma}_q^{(YX)} & \boldsymbol{\Sigma}_q^{(YY)}\end{bmatrix}\right)$$



音声変換：出力分布を計算

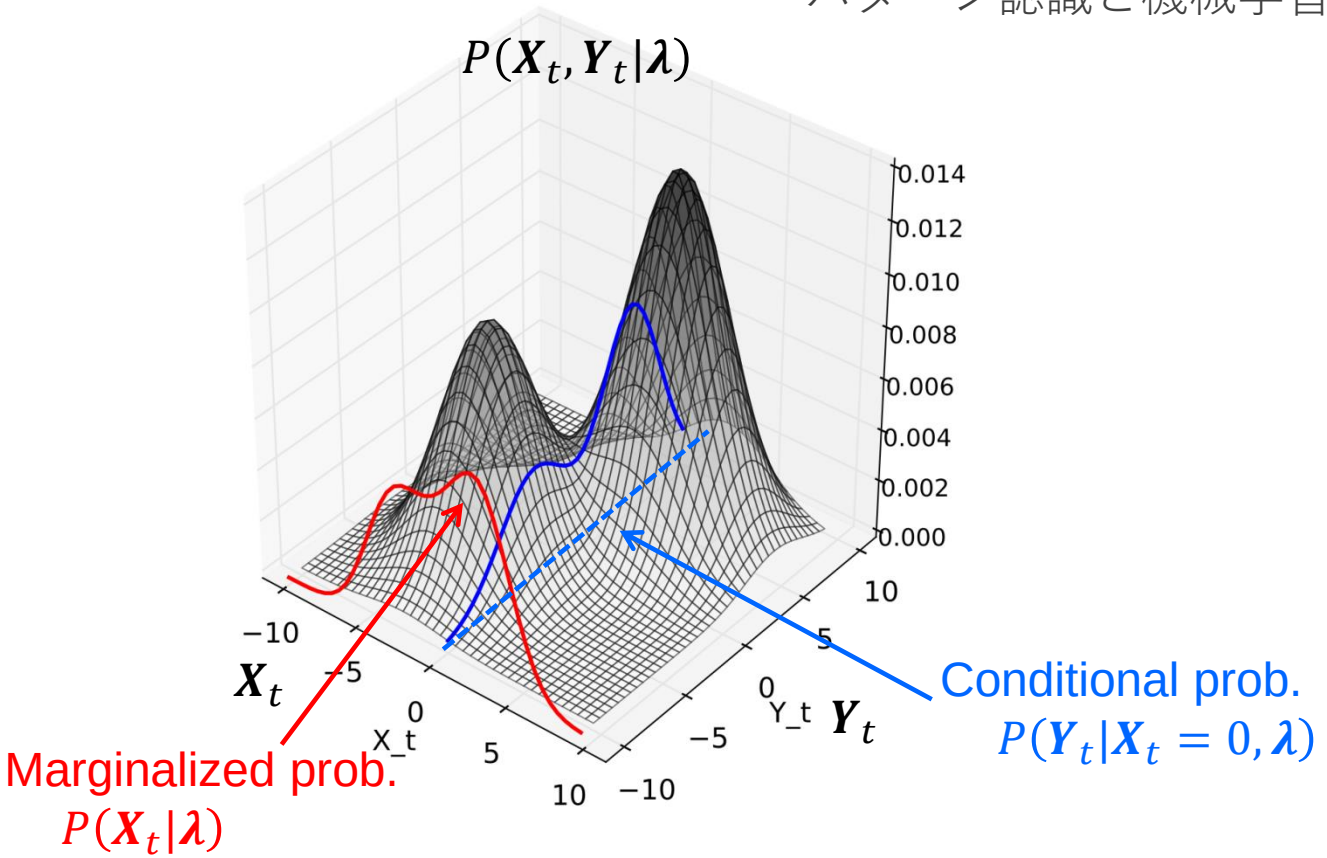
[Toda et al., 2007.]

入力特徴量 $[\boldsymbol{X}_1, \cdots, \boldsymbol{X}_t \cdots, \boldsymbol{X}_T]$ に対する音声パラメータ $\hat{\boldsymbol{y}}$ を生成

– まず，GMMを単一混合要素 $\hat{\boldsymbol{q}} = [\hat{q}_1, \cdots, \hat{q}_t, \cdots, \hat{q}_T]$ で近似

• $\hat{q}_t = \mathrm{argmax}\, P(q|\boldsymbol{X}_t, \hat{\boldsymbol{\lambda}})$ …周辺分布 $P(\boldsymbol{X}_t|\hat{\boldsymbol{\lambda}})$ から解析的に導出

“パターン認識と機械学習”を参照

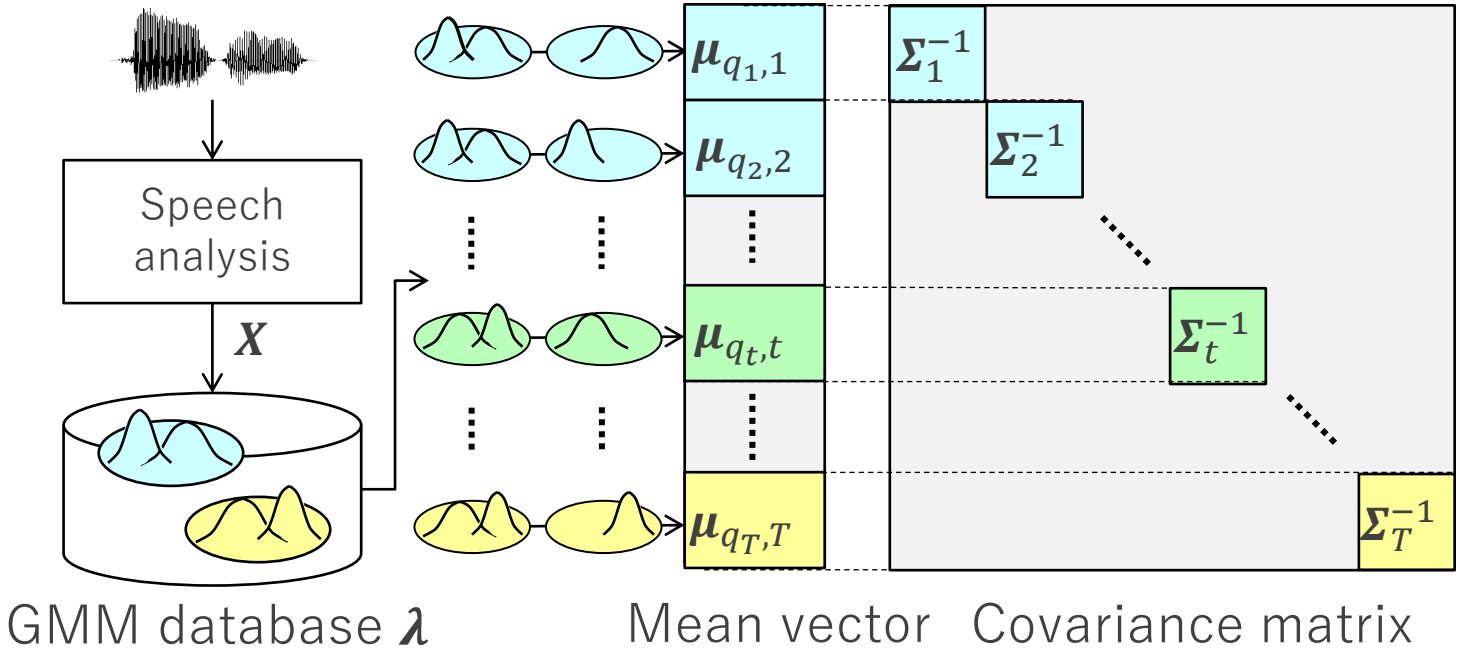


最尤パラメータ生成

[Toda et al., 2007.]

単一混合近似により，HMMと同じように最尤生成可能

- 平均 $\boldsymbol{\mu}_{q_t,t} = \boldsymbol{A}_{q_t}\boldsymbol{X}_t + \boldsymbol{b}_{q_t}$ (線形変換)
- 共分散 $\boldsymbol{\Sigma}_{q_t} = \boldsymbol{\Sigma}_q^{(YY)} - \boldsymbol{A}_{q_t}^{\top}\boldsymbol{\Sigma}_q^{(XX)}\boldsymbol{A}_{q_t}$
- $\boldsymbol{A}_{q_t} = \boldsymbol{\Sigma}_q^{(YX)}\boldsymbol{\Sigma}_q^{(XX)^{-1}}, \boldsymbol{b}_{q_t} = \boldsymbol{\mu}_{q_t}^{(Y)} - \boldsymbol{A}_{q_t}\boldsymbol{\mu}_{q_t}^{(X)}$



HMM/GMM から DNN へ

DNN隆盛へ

– 音声認識での成功、学習アルゴリズム等の改良により、音声合成・変換にも DNN の波が到来 [Zen et al., 2013]

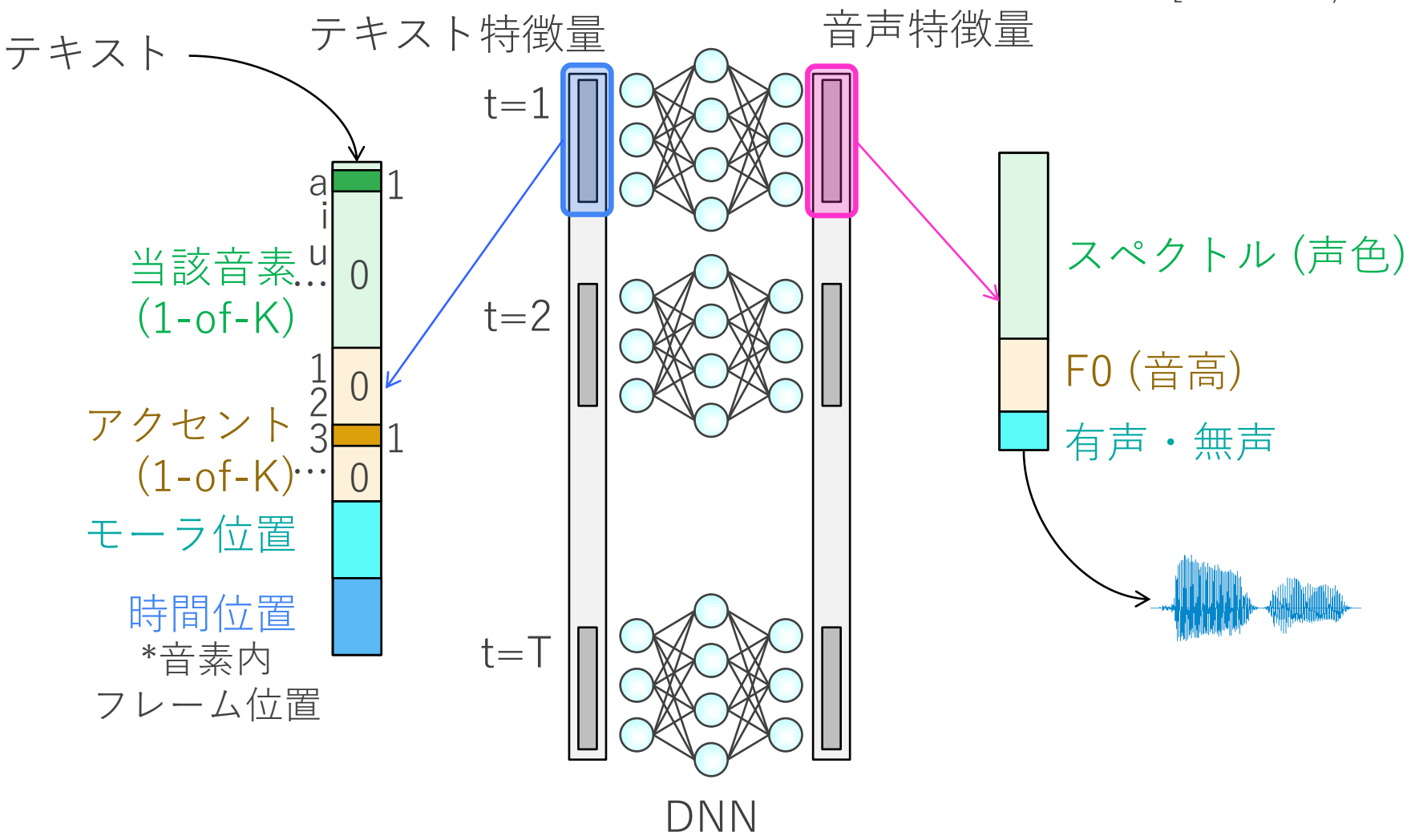
> ing data. Deep neural networks have achieved large improvements over conventional approaches in various machine learning areas including speech recognition [22] and acoustic-articulatory inversion

– HMM 音声合成・ GMM 音声変換の知見と技術をそのまま利用可能

– 他分野のDNN技術を積極的に流用可能

Text-to-speechでの利用

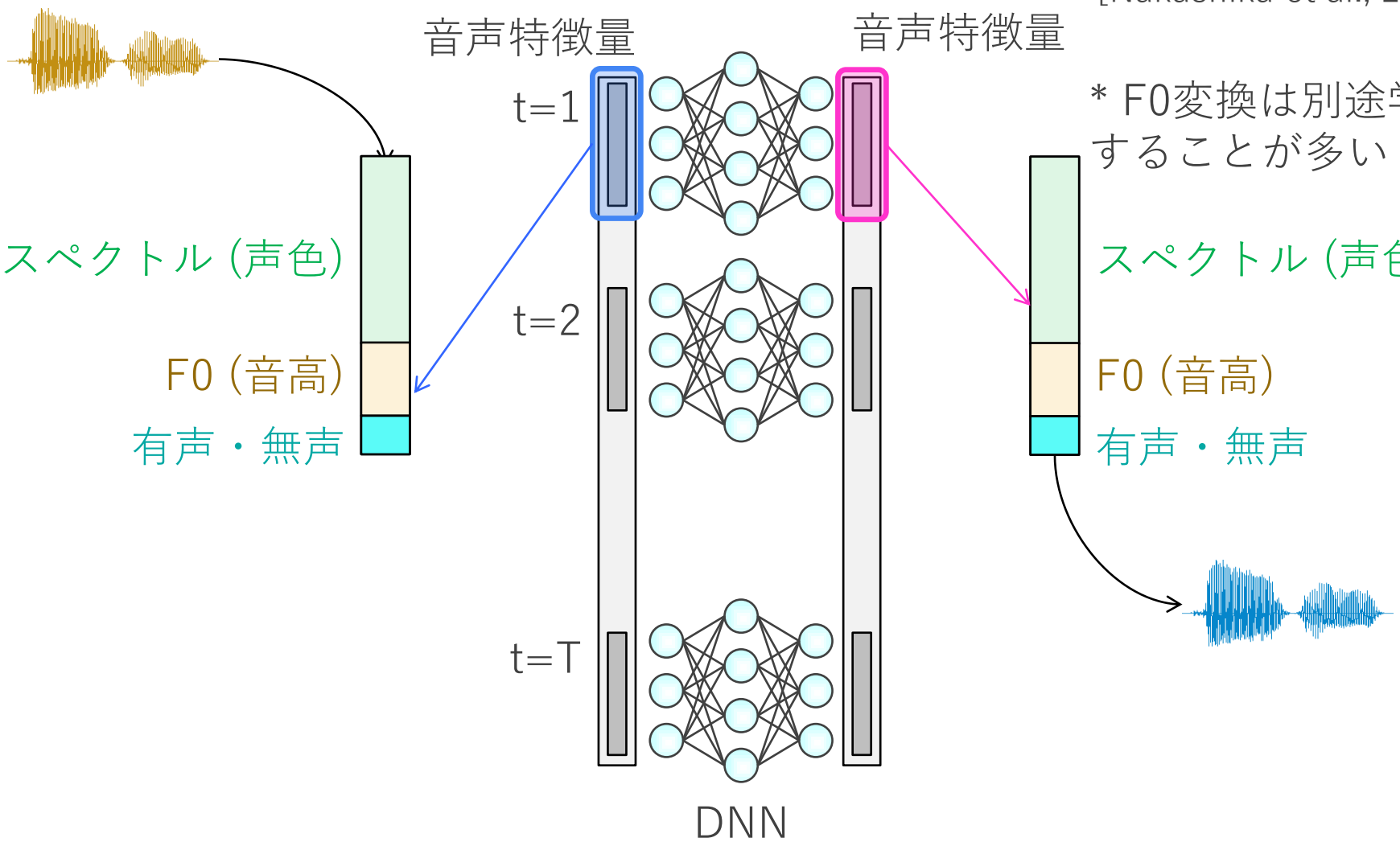
[Zen et al., 2013.]



DNNは自然音声特徴量との二乗誤差を最小化するように学習

Voice conversionでの利用

[Nakashika et al., 2013.]



* F0変換は別途学習することが多い

HMM/GMM と比べて 何が良くなった？

[Zen et al., 2013.][Merritt et al., 2016.]

HMM音声合成と比較して

- 時間量子化の緩和：HMM状態 → フレーム
- 予測の精微化：クラスタリング→ 回帰
- 大規模データが利用可能に

GMM音声変換と比較して

- 区分線形変換 (各混合要素は線形変換) → 非線形変換

もう少し詳しい話は「音声合成・変換 その2」で.

GPR音声合成・変換

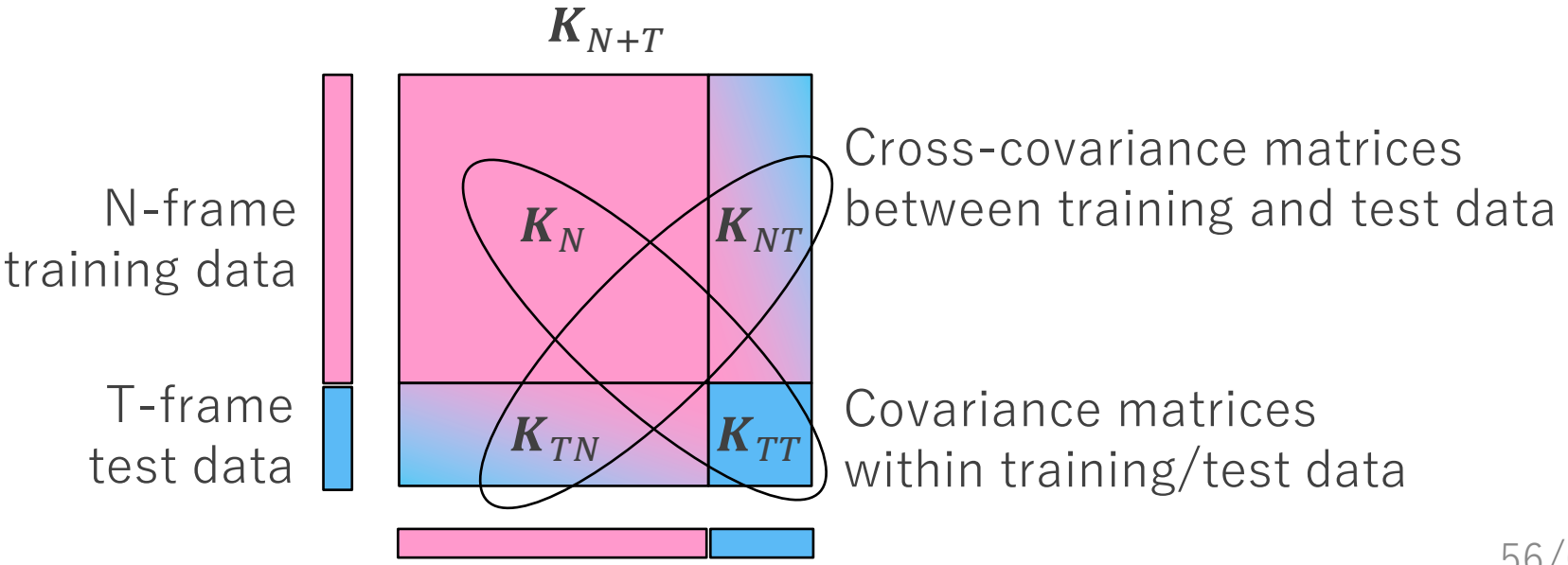
[Koriyama et al., 2014.][Pilkington et al., 2011.]

HMM/GMMの低い表現能力を緩和するために提案

- HMMの時間量子化など，GMMの(区分)線形変換に対処
- データ量に応じた柔軟性

学習データ・テストデータの同時分布を計算

- $P(\boldsymbol{Y}, \boldsymbol{Y}' | \boldsymbol{X}, \boldsymbol{X}') = N(\boldsymbol{Y}, \boldsymbol{Y}'; \boldsymbol{0}, \boldsymbol{K}_{N+T} + \sigma \boldsymbol{I}_{N+T})$
- 生成時には，これから$P(\boldsymbol{Y} | \boldsymbol{Y}', \boldsymbol{X}, \boldsymbol{X}')$を計算



カーネルの設計

[Koriyama et al., 2014.]

コンテキスト間のカーネル（距離）をどう設計する？

– 音素の属性をバイナリ表現

	a	i	u	e	o	k	t	n	s	m
vocalic	+	+	+	+	+	−	−	−	−	−
high	−	+	+	−	−	+	−	−	−	−
low	+	−	−	−	−	−	−	−	−	−
anterior	−	−	−	−	−	−	+	+	+	+
back	+	−	+	−	+	+	−	−	−	−
coronal	−	−	−	−	−	−	+	+	+	−
plosive	−	−	−	−	−	+	+	−	−	−
affricative	−	−	−	−	−	−	−	−	−	−
continuant	+	+	+	+	+	−	−	−	+	−
voiced	+	+	+	+	+	−	−	+	−	+
nasal	−	−	−	−	−	−	−	+	−	+
semi-vowel	−	−	−	−	−	−	−	−	−	−
silent	−	−	−	−	−	−	−	−	−	−

GPR/NMF における事前クラスタリング

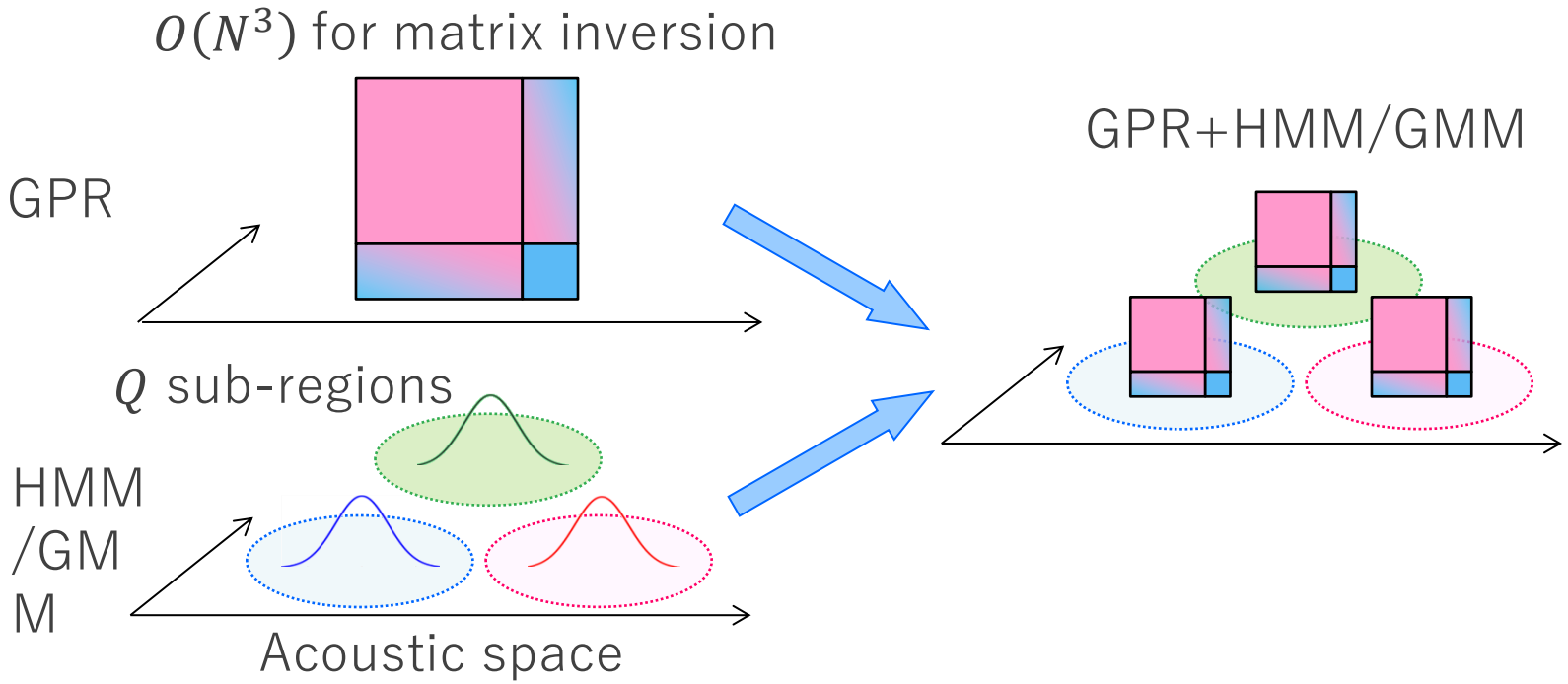
[Koriyama et al., 2014.][Pilkington et al., 2011.]

GPR/NMFにおけるスケーラビリティ

– 学習データ量に応じて計算量が爆発

HMM/GMMによる事前クラスタリング

– 音響空間をクラスタリングして，その部分空間ごとにGPR/NMF



ハイブリッド型

ハイブリッド型

– 素片選択と統計モデル (機械学習)の両方を使う

素片選択から見た利点

– 素片選択のコスト関数の設計を自動化
– 機械学習技術を導入可能

統計ベースから見た利点

– 統計モデリングによる平滑化を緩和して高品質化

HMM/DNN-based unit selection

[Ling et al., 2007.]

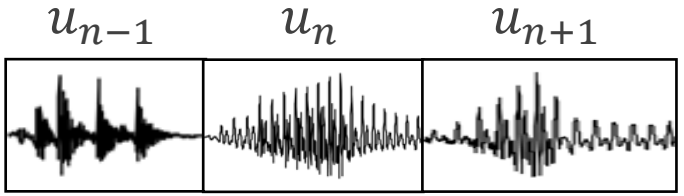
学習時

– 素片選択選択データベースと別にHMM/DNNを学習

合成時

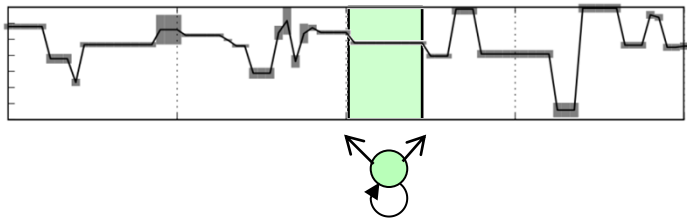
– HMM/DNN尤度を最大化するように素片を選択

選択された音声セグメント系列



コスト＝負の尤度

学習済みHMMの出力分布系列

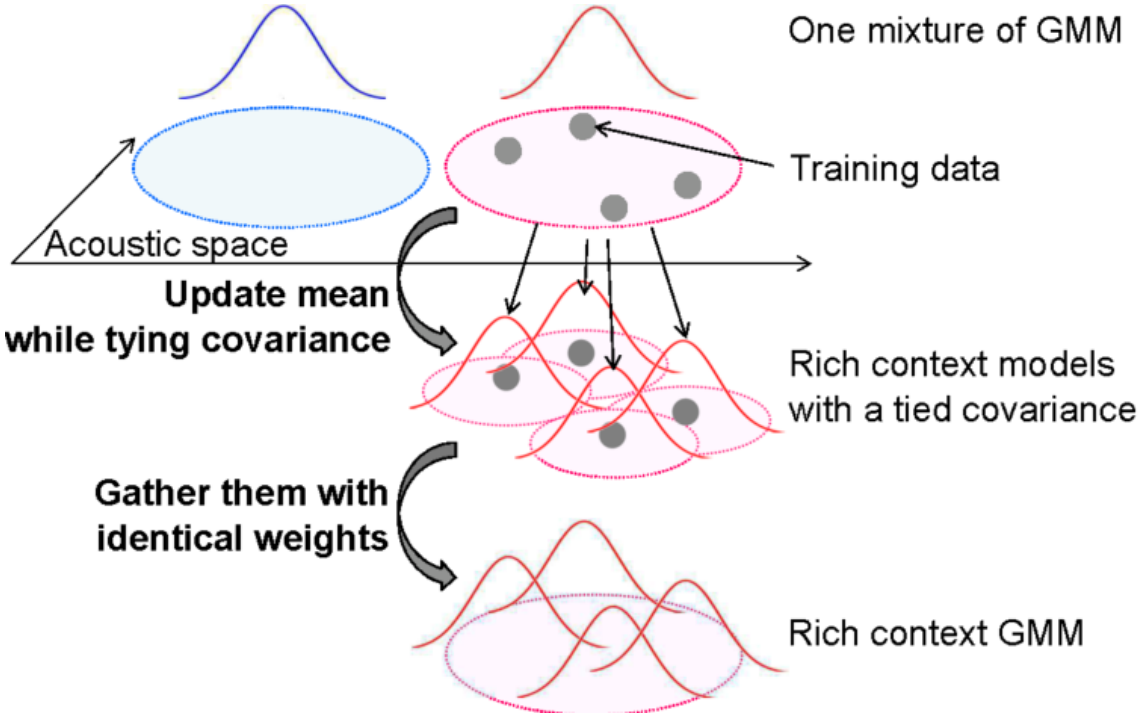


Tied-covariance HMM/GMM

[Takamichi et al., 2014, 2016.]

学習時

– 学習データの各サンプルに対し，部分空間をカバーする共分散行列
→ 未知データに対する頑健性を情報



生成時

– 通常のHMM音声合成・GMM音声変換と同様

まとめ

まとめ

音声合成の基礎

- コンテキスト・音声特徴量
- 素片選択型合成法
- 統計的音声合成法
  - HMM, GMM, DNNなど

次回

- 近年のホットな話題
- 音声合成の応用

参考文献

- http://www.sp.ipc.i.u-tokyo.ac.jp/~saruwatari/SP-Grad2016_05.pdf を参照