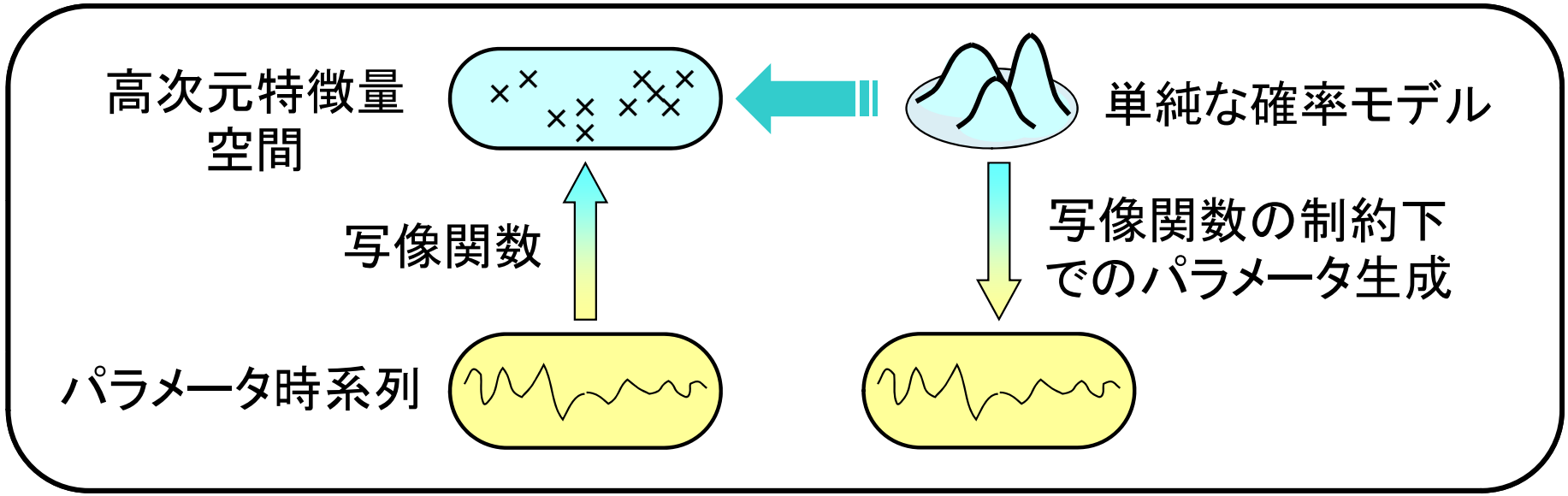
音声音響信号処理
～統計的手法による音声変換～



戸田 智基

奈良先端科学技術大学院大学

情報科学研究科

2014年1月20日（月）

内容

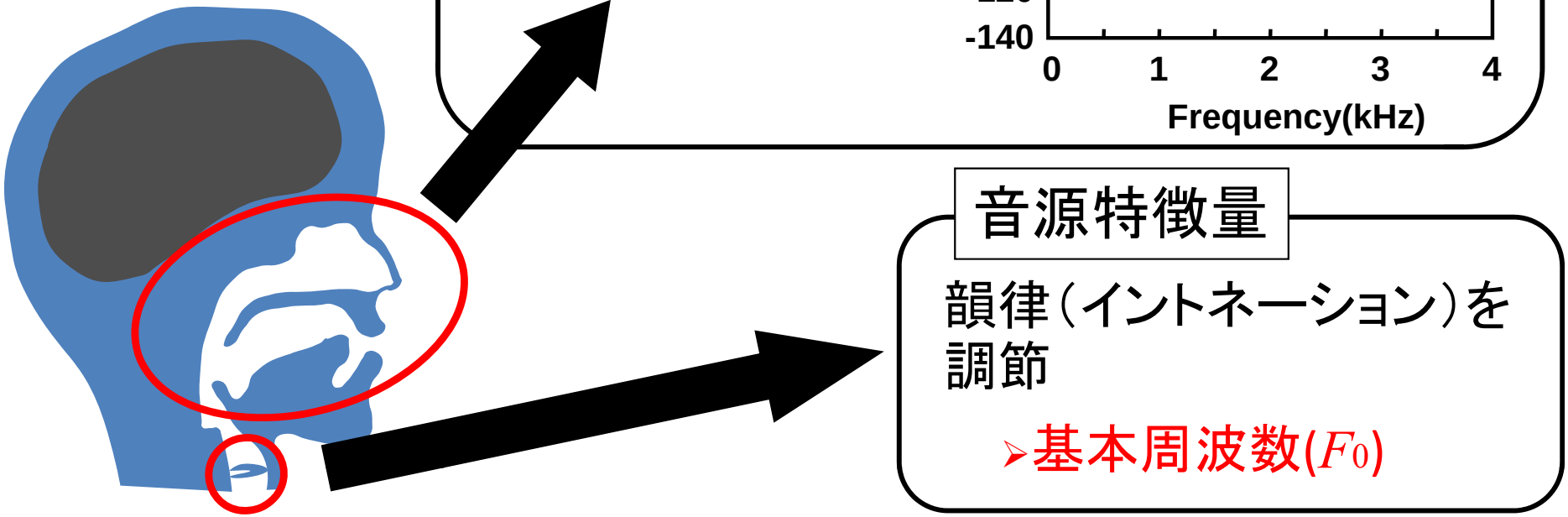
1. 音声変換のしくみ
2. 統計的手法による声質変換
3. 応用例

内容

1. 音声変換のしくみ
2. 統計的手法による声質変換
3. 応用例

音声の特徴量

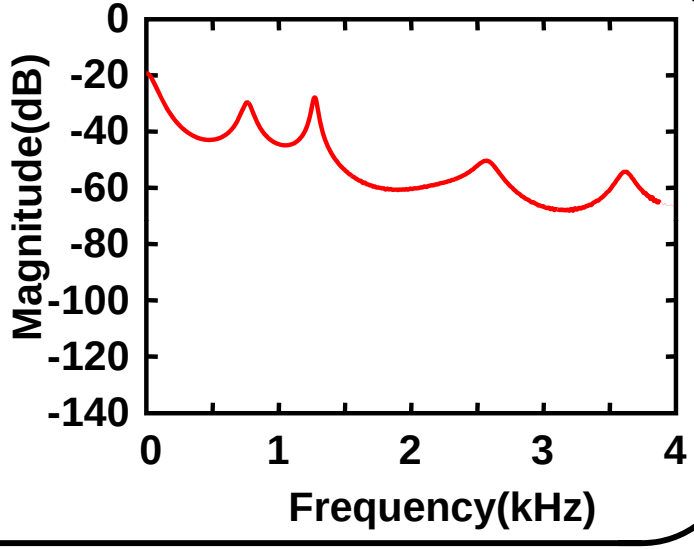
- 音声の生成過程に基づいており，音声波形から抽出可能



声道特徴量

音韻性や声質を調節

➢スペクトル包絡



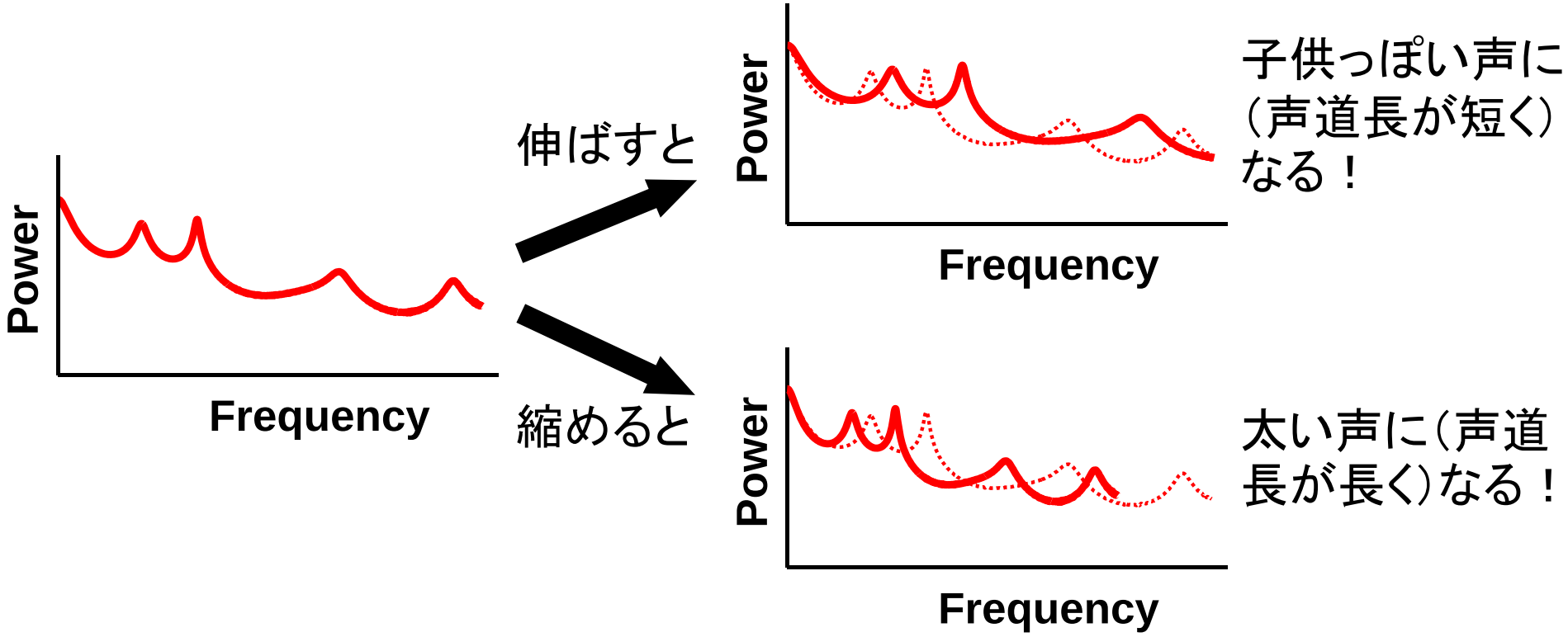
音源特徴量

韻律（イントネーション）を調節

➢基本周波数(F_0)

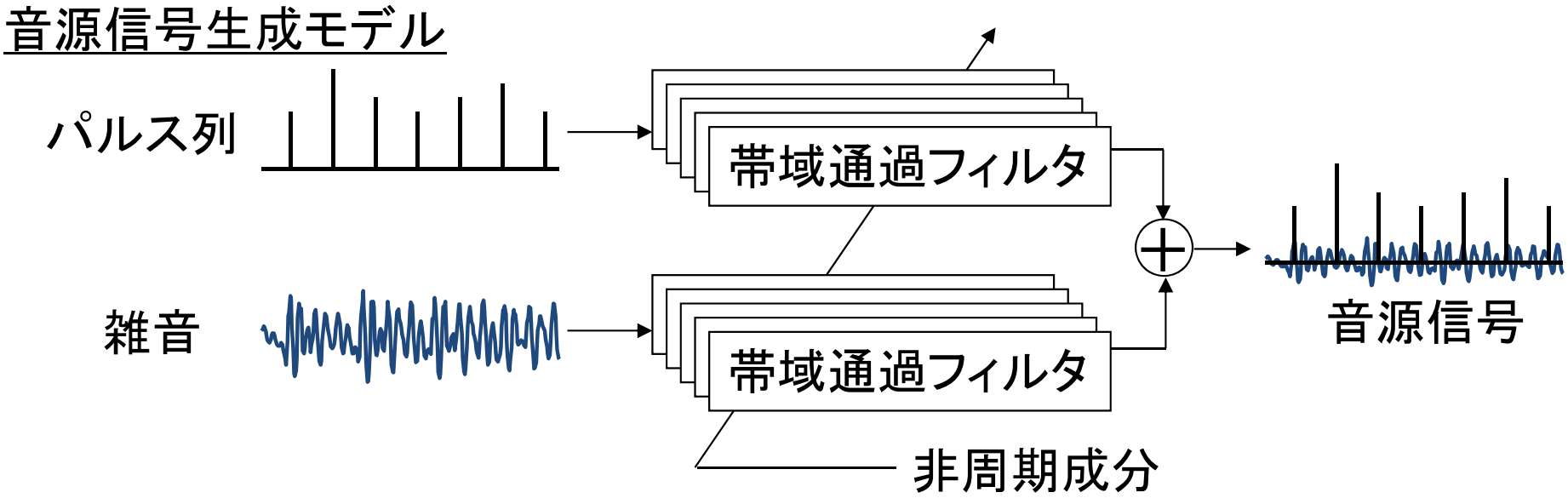
声道特徴量の変換

- 特徴量パラメータ：スペクトル包絡
- 音韻性や声質などを表現
- 周波数軸方向に一律に伸縮させることで，音韻性を保ったまま声質を変換することが可能



音源特徴量の変換

- 特徴量パラメータ：基本周波数（F_0）や非周期成分
- 声の高さや声のかすれなどを表現
- 基本周波数を高くすれば高い声に，低くすれば低い声へと変換
- 非周期成分を大きくすればかすれた声へと変換



リアルタイム音声変換デモ

- 簡単な規則による声道／音源特徴量の変換

リアルタイム音声変換ソフト
Herium: High Entertaining Real-time
Input Utterance Modifier

作成者： 名城大学　坂野秀樹先生

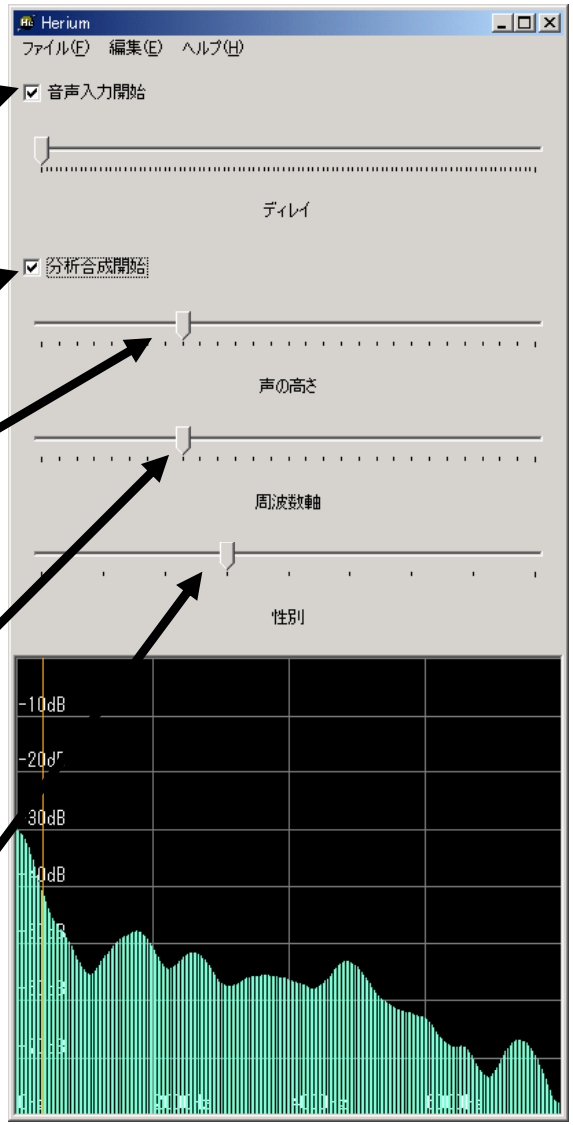
入力開始

分析合成開始

声の高さ成分の調節

音色成分の調節
➢スペクトルの変換
（周波数軸の伸縮）

性別の調節
➢両成分の変換



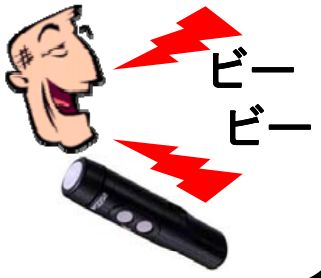
http://www.sp.m.is.nagoya-u.ac.jp/people/banno/spLibs/herium/index-j.html

さらに高度な変換処理を実現するには？

- 例えば，電気式人工喉頭を用いて発声された機械的な音声（電気音声）を通常音声へとリアルタイムで変換したい・・・

話し手側

機械的な声質を持つ音声の生成



音声変換

聞き手側

より自然な声質を持つ音声の提示

失った声を取り戻せるかもしれない!

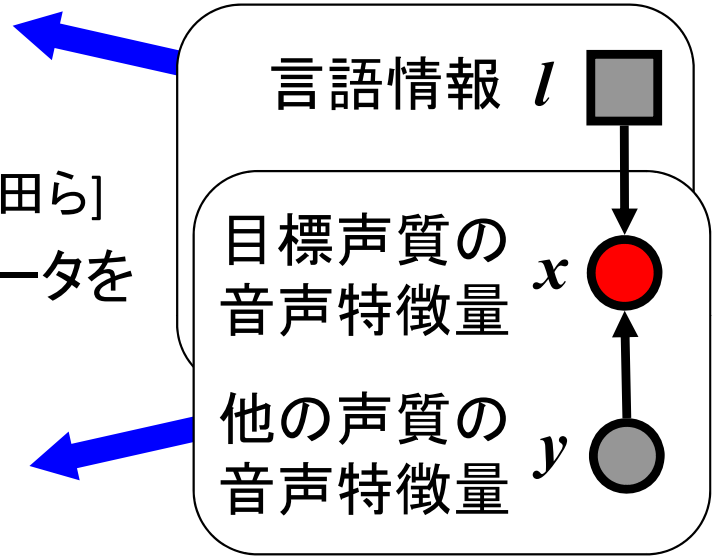
統計的手法により高度な変換処理を実現可能！

内容

1. 音声変換のしくみ
2. 統計的手法による声質変換
   2.1. 基本的な枠組み
   2.2. フレームベース変換法
   2.3. 系列ベース変換法
   2.4. リアルタイム変換法
3. 応用例

統計的手法に基づく声質モデリング

- 確率モデルによる音声特徴量のモデル化
  - テキスト音声合成(TTS)
    - 確率密度関数(*p.d.f.*)$P(\boldsymbol{x}\,|\,\boldsymbol{l})$ のモデル化
    - 隠れマルコフモデル(HMM)による手法 [徳田ら]
    - 言語情報が付与された目標声質の音声データを用いて学習
  - 声質変換
    - $P(\boldsymbol{x}\,|\,\boldsymbol{y})$ のモデル化
    - 混合正規分布モデル(GMM)による手法 [Stylianou *et al.*, 1998]
    - 言語情報は同一で，所望の声質成分のみが異なる音声データ(パラレルデータ)を用いて学習

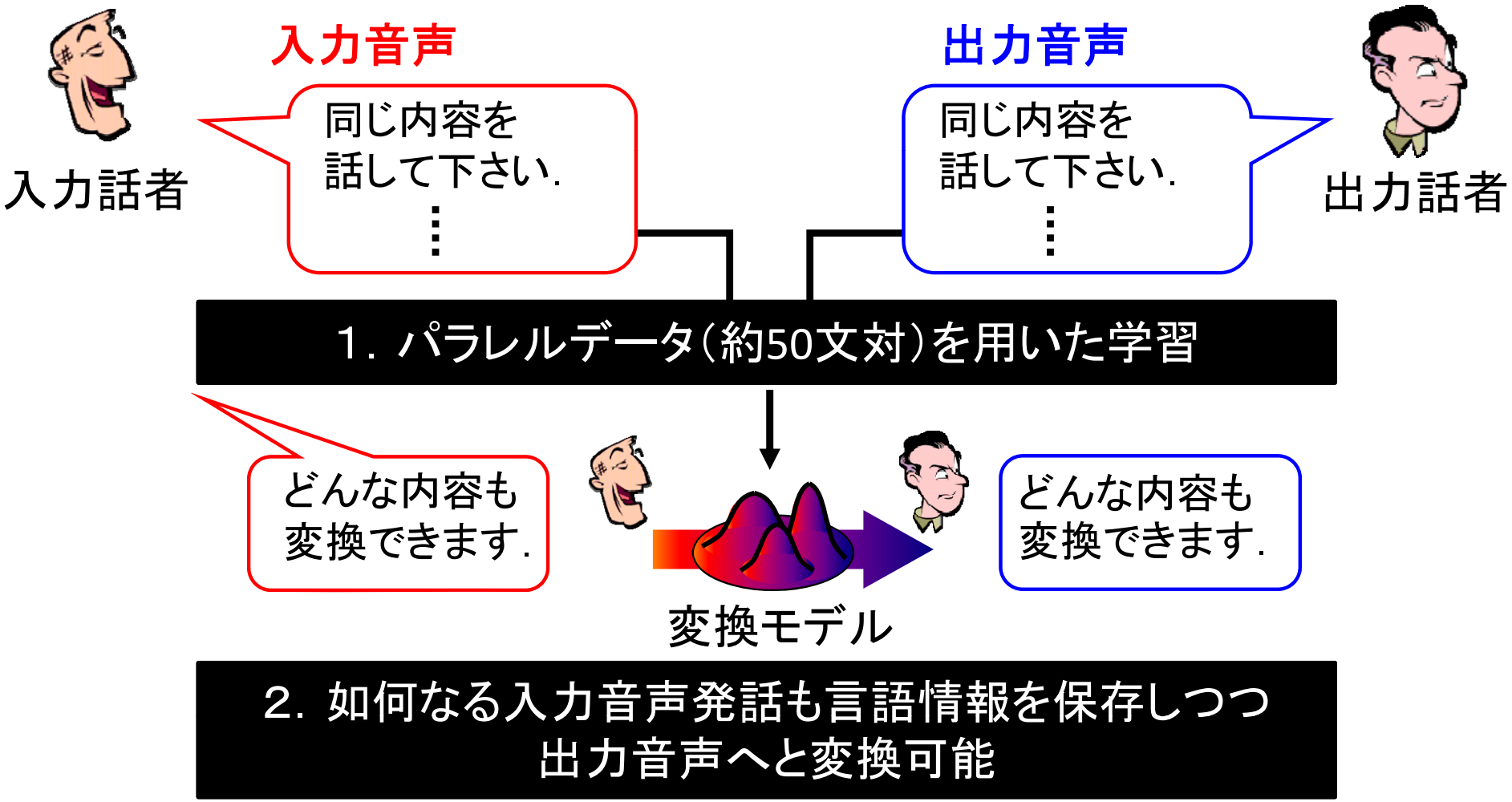


- モデル化される声質成分は**学習データ**により決定

統計的手法に基づく声質変換の枠組み

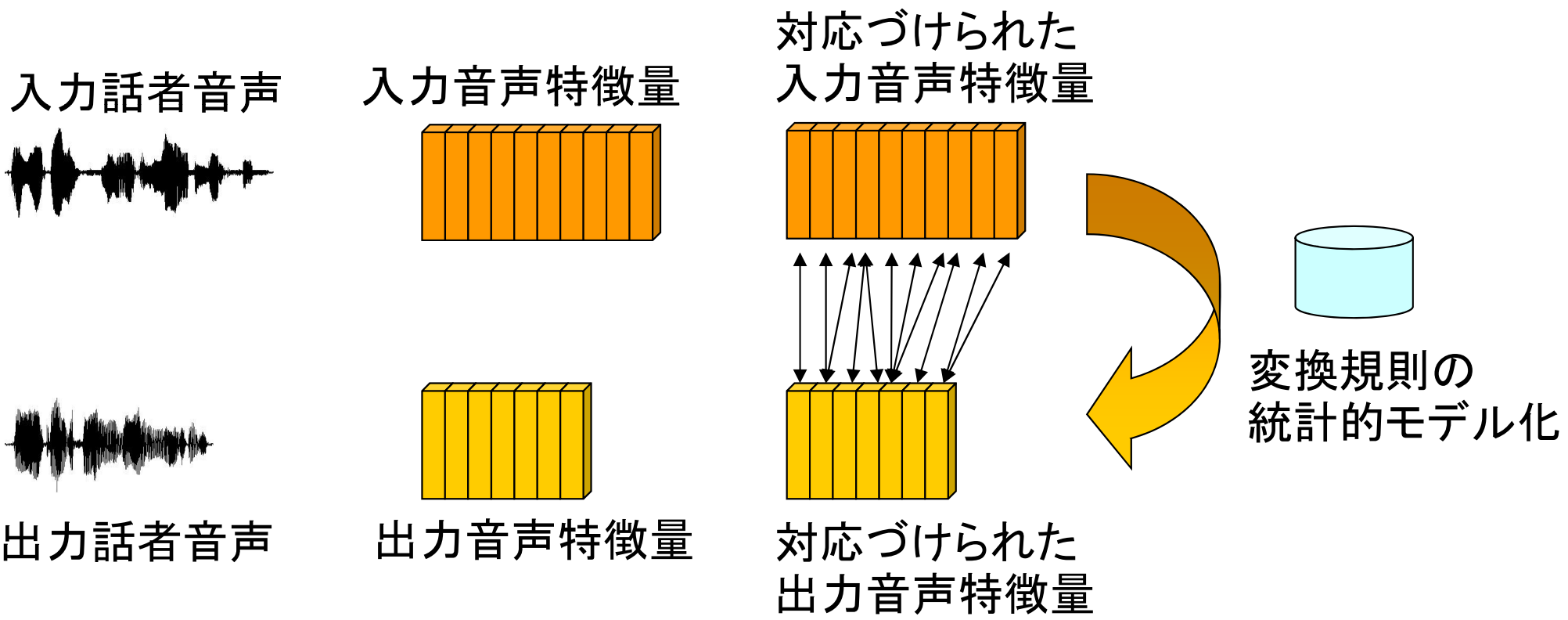
[Abe *et al.*, 1990]

- 異なる話者や発話様式間における変換を実現
- 入・出力話者による同一内容発声の音声データ(パラレルデータ)を用いて入・出力音声特徴量間の対応関係を統計的にモデル化



学習処理

音声波形のペア → 音声特徴量のペア → 時間の対応付けがなされた音声特徴量ペア → 変換規則



内容

1. 音声変換のしくみ
2. 統計的手法による声質変換
   2.1. 基本的な枠組み
   2.2. フレームベース変換法
   2.3. 系列ベース変換法
   2.4. リアルタイム変換法
3. 応用例

フレームベースの変換関数

- 各フレーム（各時間）において独立に変換処理を実施
  - 入力特徴量ベクトル　　：$\boldsymbol{x}_t$
  - 出力特徴量ベクトル　　：$\boldsymbol{y}_t$
  - モデルパラメータセット　：$\boldsymbol{\lambda}$
  - 変換特徴量ベクトル　　：$\hat{\boldsymbol{y}}_t$

$$\hat{\boldsymbol{y}}_t = F_{\lambda}(\boldsymbol{x}_t)$$

例えば

$$= E_{\lambda}[\boldsymbol{y}_t \mid \boldsymbol{x}_t] = \int \boldsymbol{y}_t P(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}) d\boldsymbol{y}_t$$

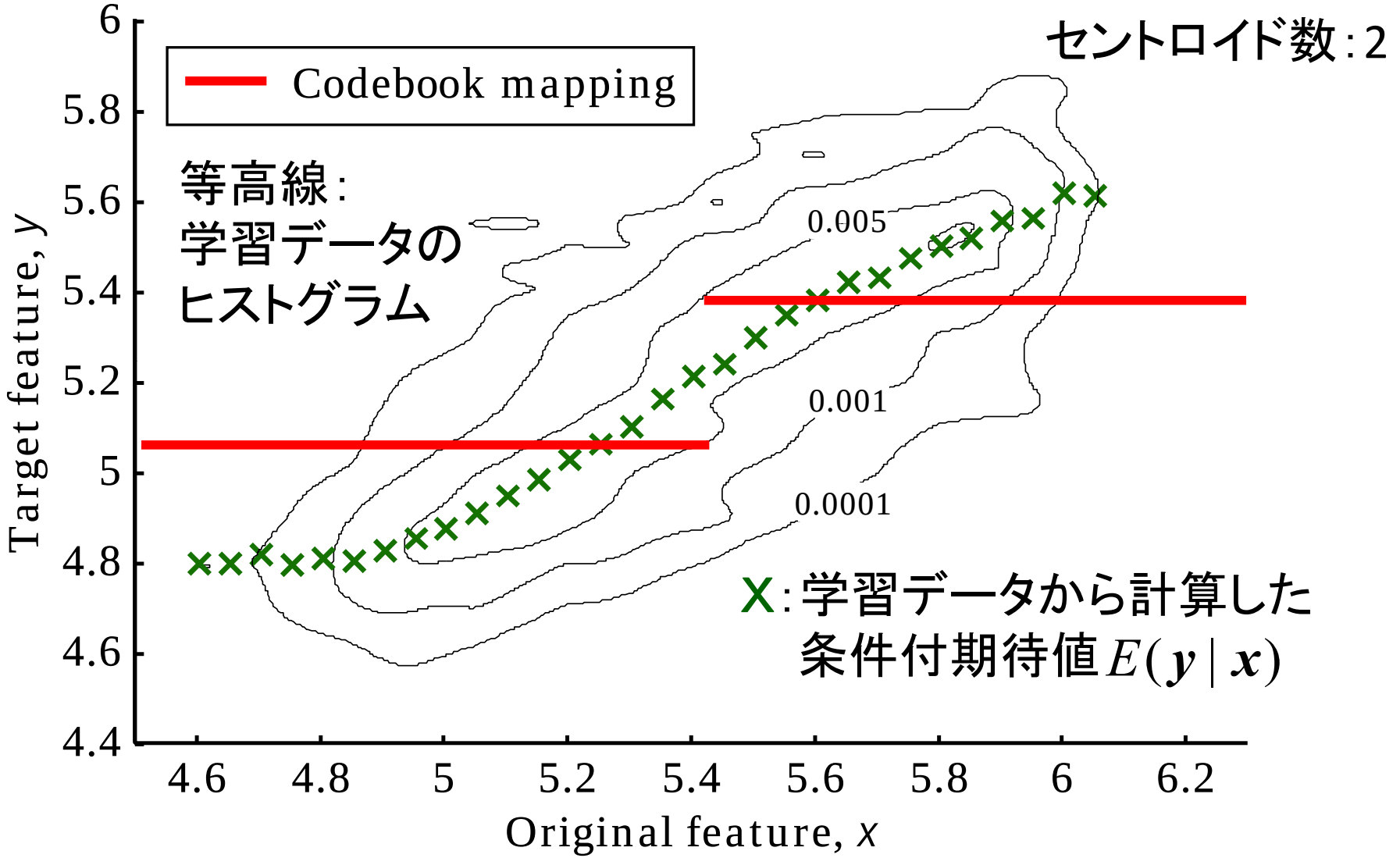
とか

$$= \arg\max_{\boldsymbol{y}_t} P(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda})$$

代表的な変換法

1. コードブックマッピング法 [Abe *et al.*, 1990]	$\hat{\boldsymbol{y}}_t = \boldsymbol{\mu}_m^{(y)}$ ハードクラスタリングと離散的なマッピング
2. ファジーベクトル量子化に基づくコードブックマッピング法 [中村 他, 1989]	$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} \gamma_{m,t}^{(x)} \boldsymbol{\mu}_m^{(y)}$ ソフトクラスタリングと離散的なマッピング
3. ファジーベクトル量子化と差分ベクトルに基づくコードブックマッピング法 [Matsumoto *et al.*, 1993]	$\hat{\boldsymbol{y}}_t = \boldsymbol{x}_t + \sum_{m=1}^{M} \gamma_{m,t}^{(x)} \left(\boldsymbol{\mu}_m^{(y)} - \boldsymbol{\mu}_m^{(x)} \right)$ ソフトクラスタリングと連続的なマッピング
4. 線形回帰法 [Valbret *et al.*, 1992]	$\hat{\boldsymbol{y}}_t = \boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m$ ハードクラスタリングと連続的かつ高精度な変換
5. 混合正規分布モデルに基づく変換法 [Stylianou *et al.*, 1998]	$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} \gamma_{m,t}^{(x)} \left(\boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m \right)$ ソフトクラスタリングと連続的かつ高精度な変換

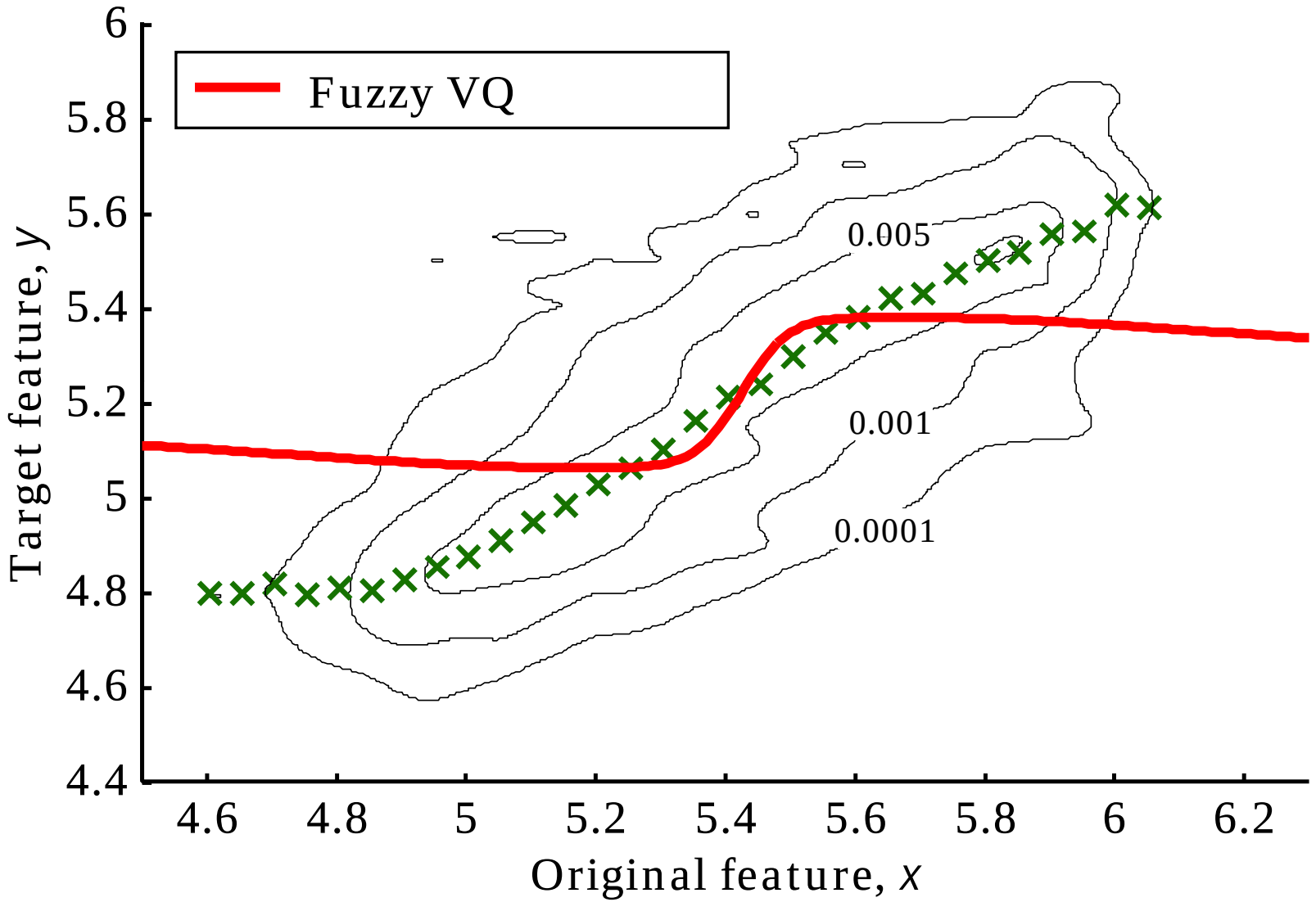
コードブックマッピング法の変換関数



$$\hat{y}_t = \mu_m^{(y)}$$

ハードクラスタリングと離散的なマッピング

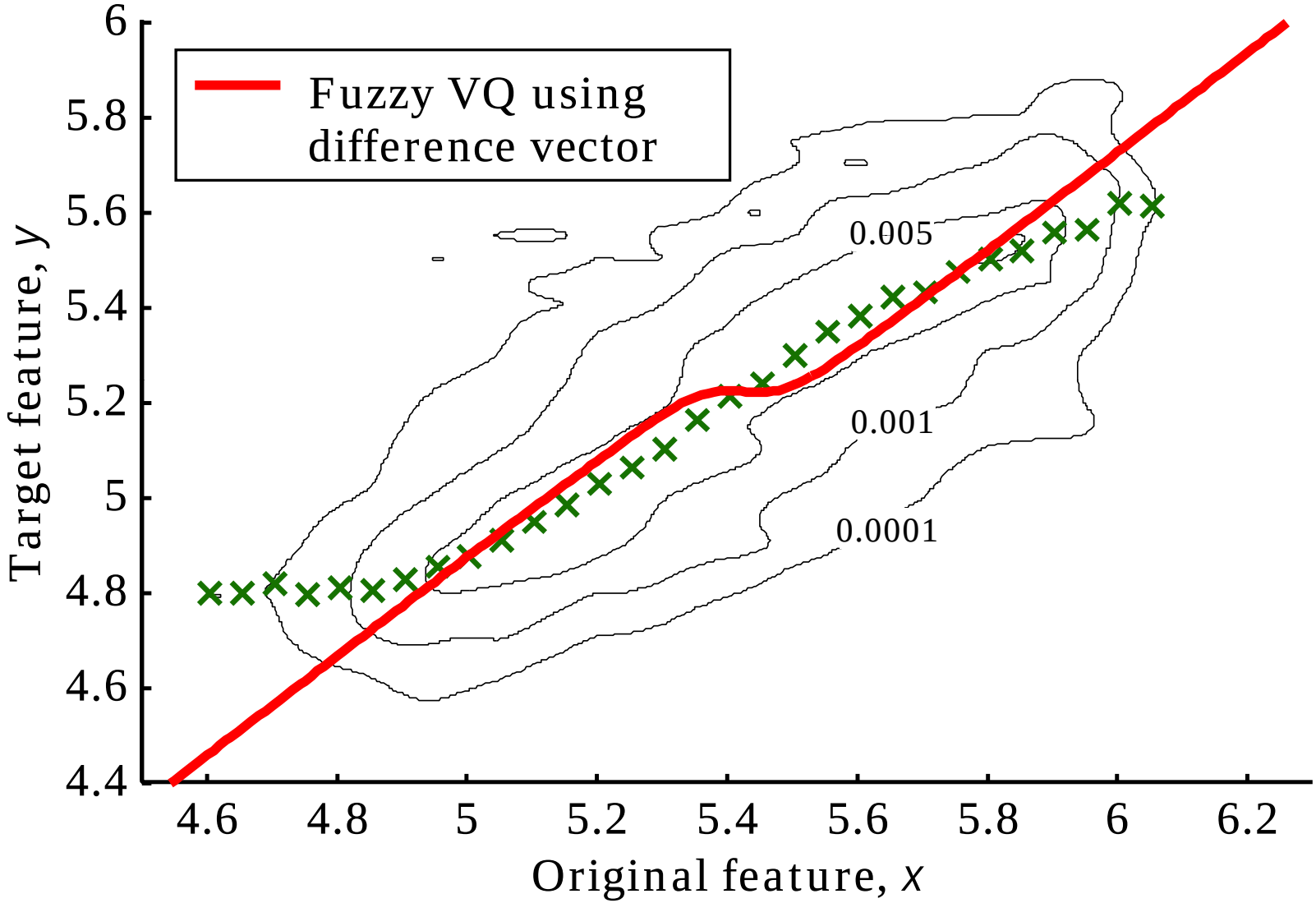
ファジーVQ使用時の変換関数



$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} \gamma_{m,t}^{(x)} \boldsymbol{\mu}_m^{(y)}$$

ソフトクラスタリングと離散的なマッピング

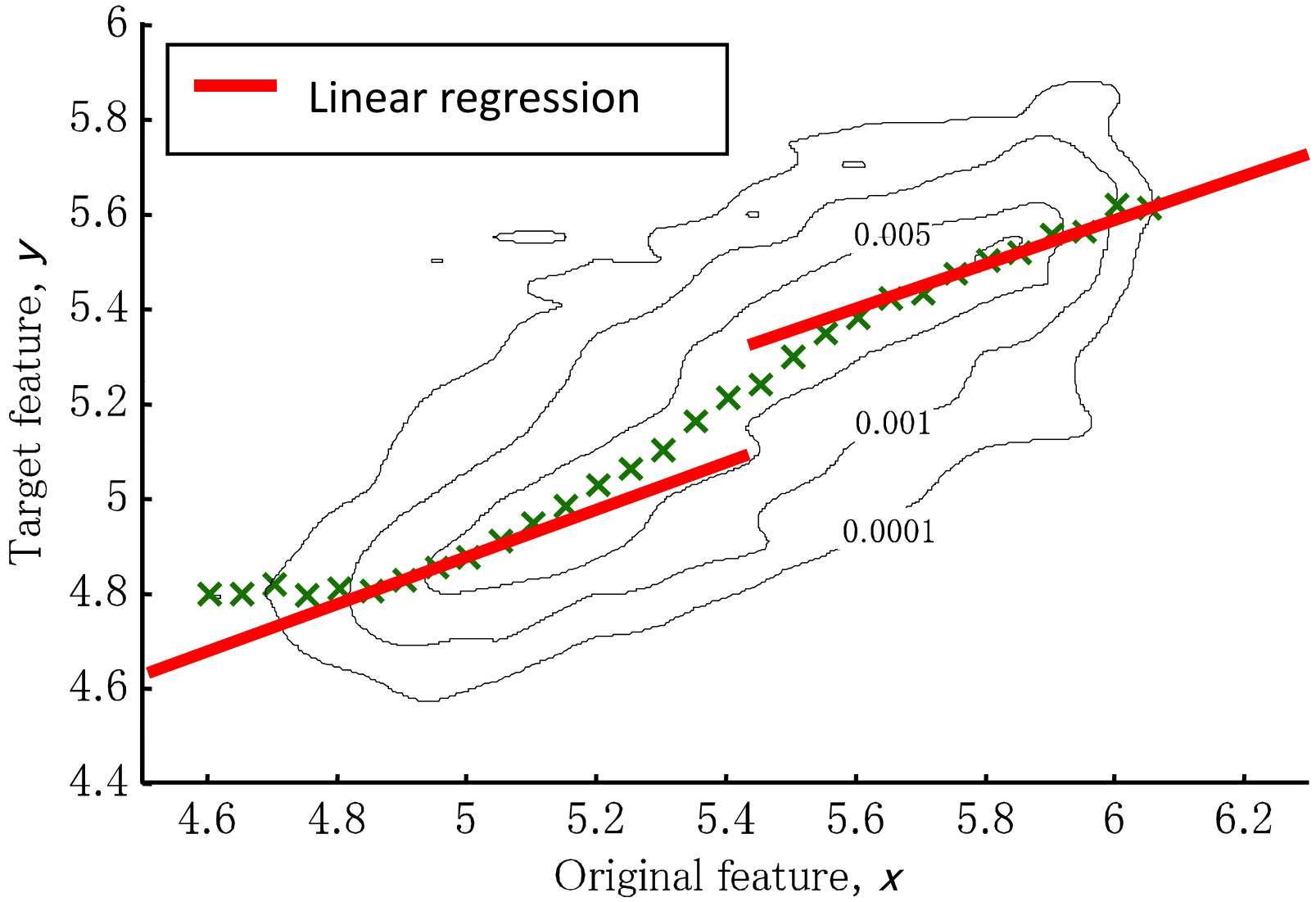
差分ベクトル使用時の変換関数



$$\hat{\boldsymbol{y}}_t = \boldsymbol{x}_t + \sum_{m=1}^{M} \gamma_{m,t}^{(x)} \left(\boldsymbol{\mu}_m^{(y)} - \boldsymbol{\mu}_m^{(x)} \right)$$

ソフトクラスタリングと連続的なマッピング

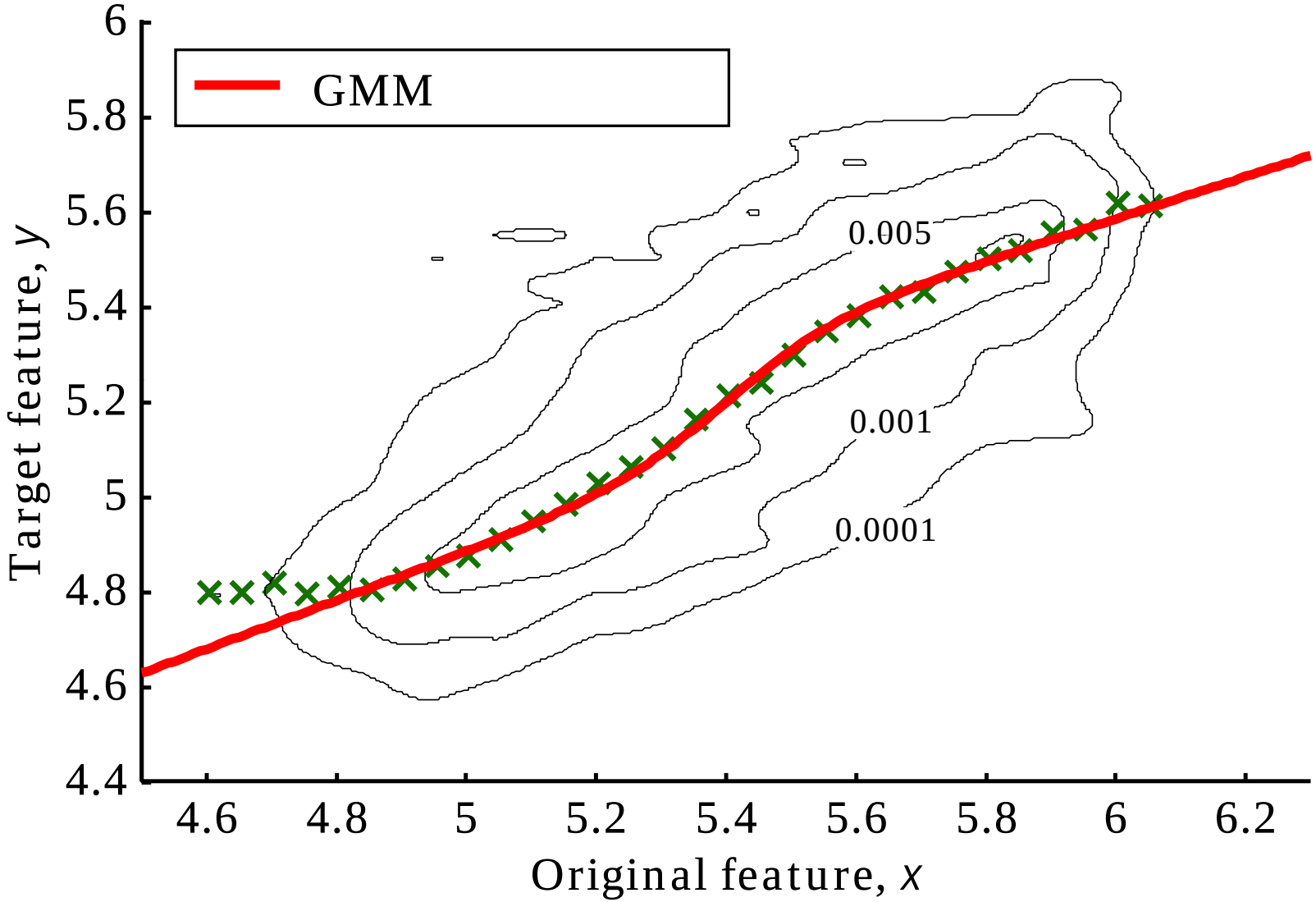
線形回帰に基づく変換関数



$$\hat{\boldsymbol{y}}_t = \boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m$$

ハードクラスタリングと連続的かつ高精度な変換

混合正規分布モデルに基づく変換関数



$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} \gamma_{m,t}^{(x)} \left(\boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m \right)$$ ソフトクラスタリングと連続的かつ高精度な変換

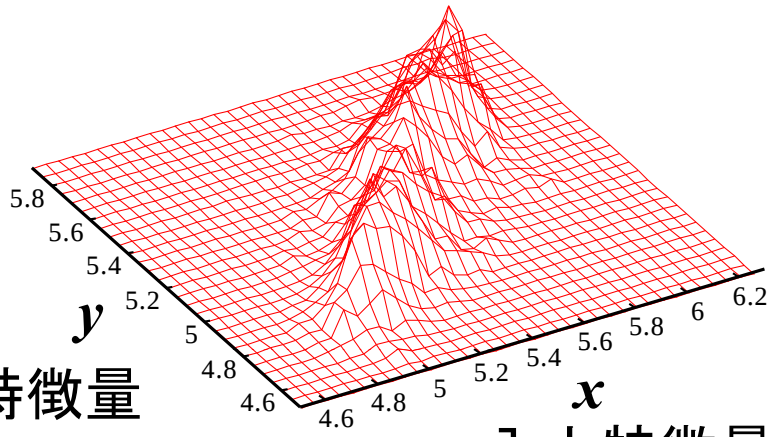
学習処理：結合$p.d.f.$のモデル化

[Kain & Macon, 1998]

- パラレルデータを用いて入力特徴量と出力特徴量の結合$p.d.f.$をGMMによりモデル化

学習データのヒストグラム

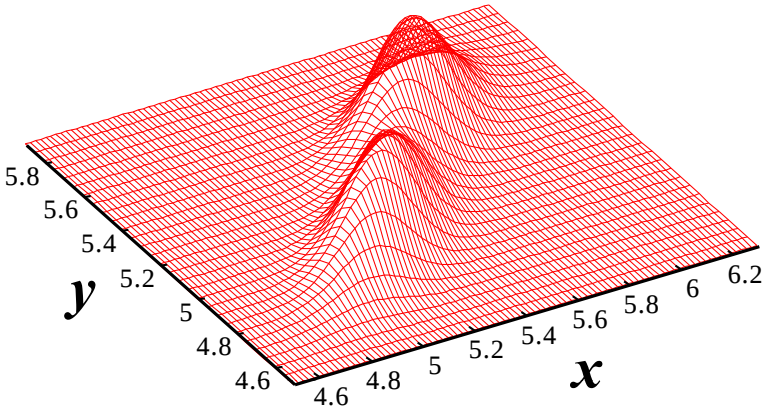
結合特徴量ベクトル $\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y}_t \end{bmatrix}$



GMM:

結合$p.d.f.$ $P(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\lambda}^{(x,y)}) = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{\lambda}^{(x,y)}\right) P\left(\boldsymbol{x}_t, \boldsymbol{y}_t \mid m, \boldsymbol{\lambda}^{(x,y)}\right)$

$$= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y}_t \end{bmatrix} \mid \boldsymbol{\mu}_m^{(x,y)}, \boldsymbol{\Sigma}_m^{(x,y)}\right)$$



平均 $\boldsymbol{\mu}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}$ 分散 $\boldsymbol{\Sigma}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}$

条件付き*p.d.f.*

- 結合*p.d.f.*: $P\left(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\lambda}^{(x,y)}\right) = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{\lambda}^{(x,y)}\right) P\left(\boldsymbol{x}_t \mid m, \boldsymbol{\lambda}^{(x,y)}\right) P\left(\boldsymbol{y}_t \mid \boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(x,y)}\right)$

  周辺化すると $P\left(\boldsymbol{x}_t \mid \boldsymbol{\lambda}^{(x,y)}\right) = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{\lambda}^{(x,y)}\right) P\left(\boldsymbol{x}_t \mid m, \boldsymbol{\lambda}^{(x,y)}\right)$

- 入力 $\boldsymbol{x}_t$ が与えられた際の条件付き*p.d.f.*:

$$P\left(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(x,y)}\right) = \frac{P\left(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{\lambda}^{(x,y)}\right)}{P\left(\boldsymbol{x}_t \mid \boldsymbol{\lambda}^{(x,y)}\right)}$$

$$= \sum_{m=1}^{M} P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(x,y)}\right) P\left(\boldsymbol{y}_t \mid \boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(x,y)}\right)$$

GMMでモデル化される.

事後確率

$$\frac{P\left(m \mid \boldsymbol{\lambda}^{(x,y)}\right) P\left(\boldsymbol{x}_t \mid m, \boldsymbol{\lambda}^{(x,y)}\right)}{\sum_{n=1}^{M} P\left(n \mid \boldsymbol{\lambda}^{(x,y)}\right) P\left(\boldsymbol{x}_t \mid n, \boldsymbol{\lambda}^{(x,y)}\right)}$$

正規分布

平均: $\boldsymbol{\mu}_{m,t}^{(y|x)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right)$

$= \boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m$

共分散: $\boldsymbol{\Sigma}_m^{(y|x)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)}$

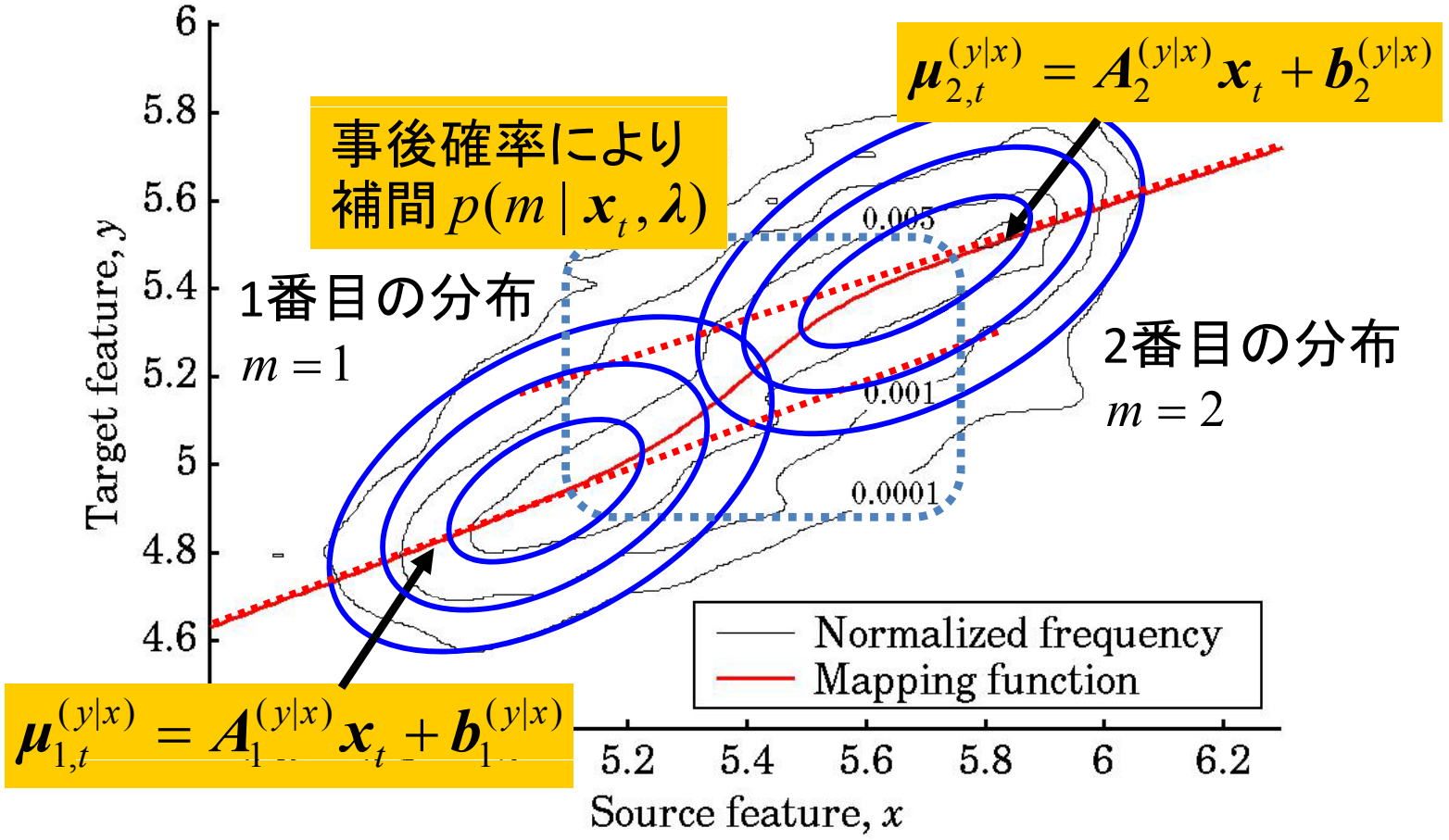
最小平均二乗誤差推定

[Stylianou *et al.*, 1998]

推定値: $\hat{\boldsymbol{y}}_t = \int \boldsymbol{y}_t p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{y}_t = \sum_{m=1}^{M} p(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}) \boldsymbol{\mu}_{m,t}^{(y|x)}$

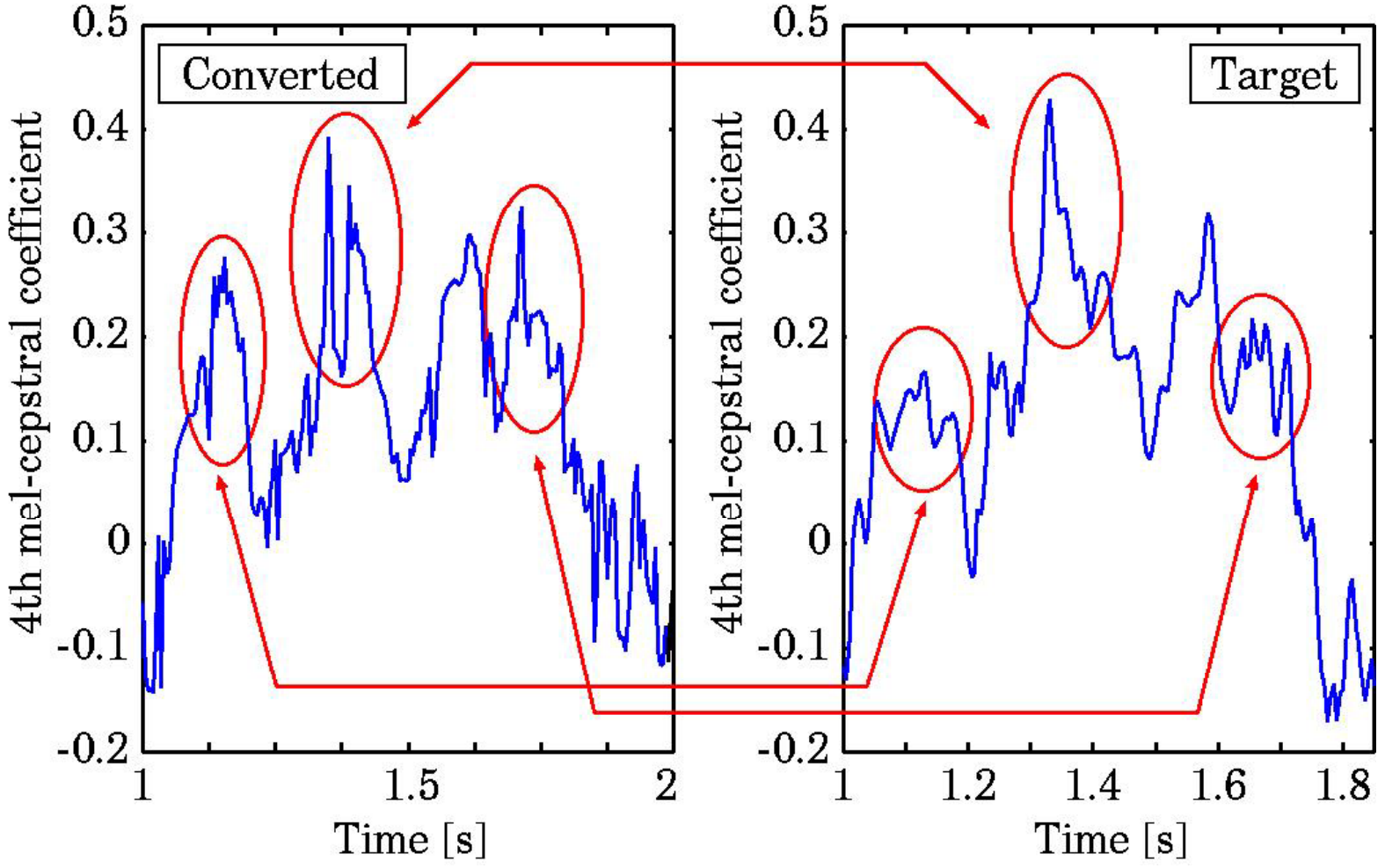
ソフトクラスタリング

線形回帰 $= \boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m$



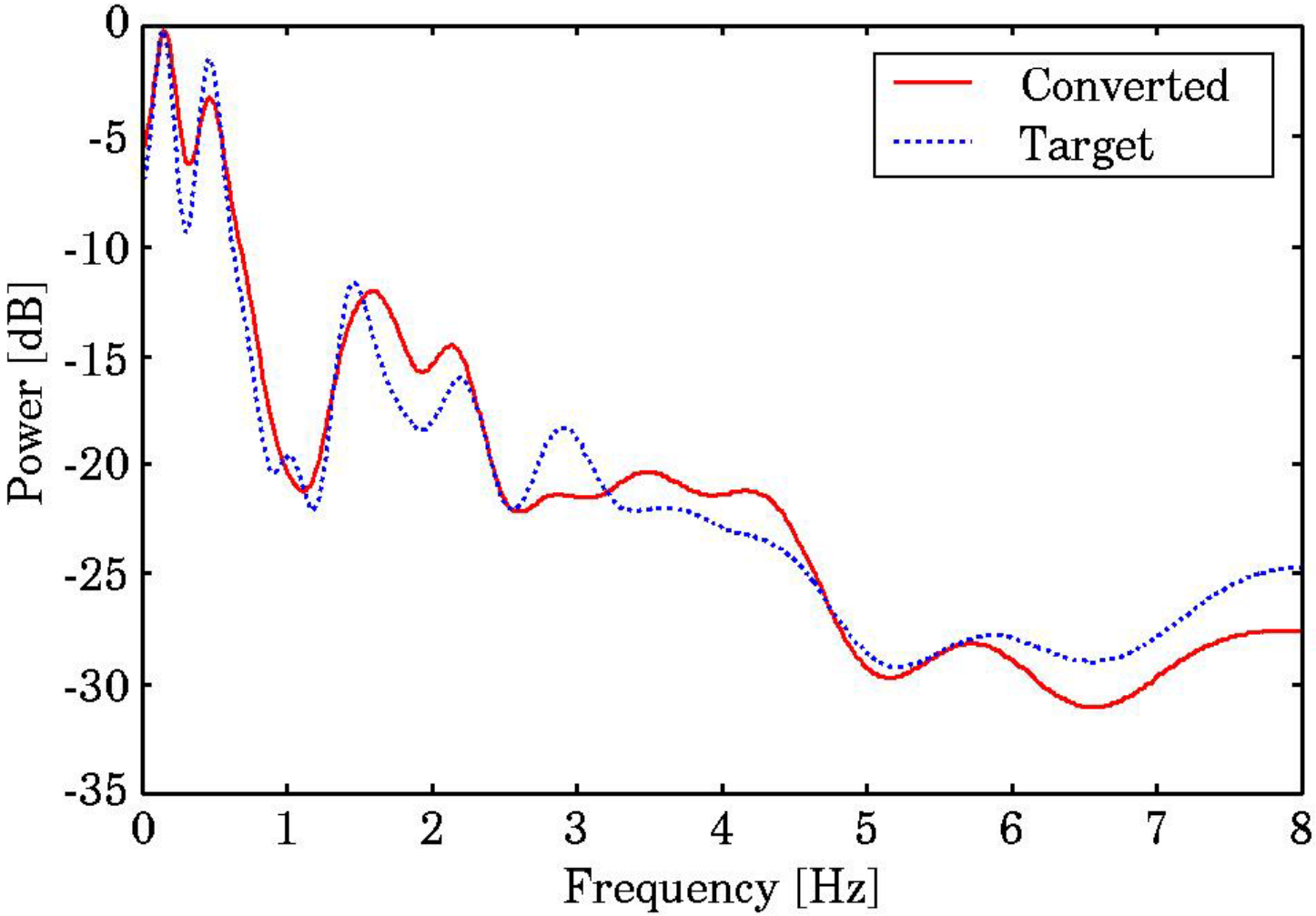
問題点1：時間的依存関係の無視

- フレーム毎に独立して変換処理が行われるため，しばしば不適切なパラメータ遷移が発生



問題点2：過剰な平滑化

- 汎化処理により詳細なスペクトル構造が消失し，変換音声の肉声感が大幅に劣化



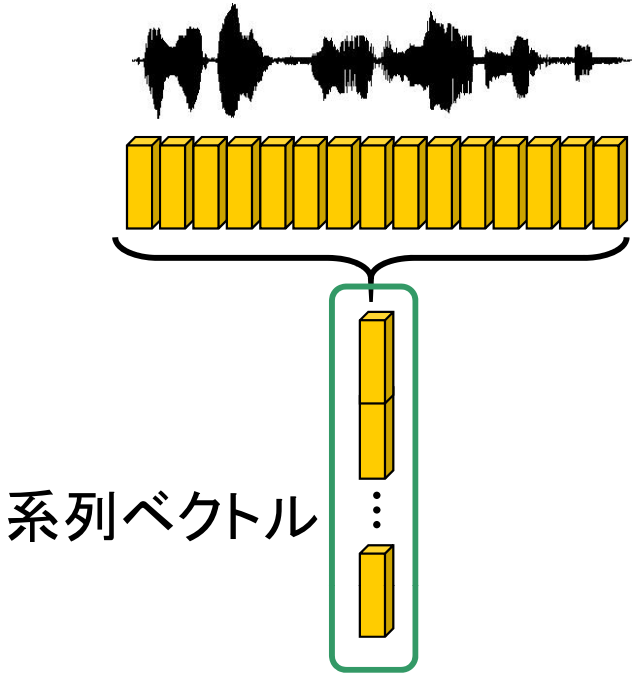
内容

1. 音声変換のしくみ
2. 統計的手法による声質変換
   2.1. 基本的な枠組み
   2.2. フレームベース変換法
   2.3. 系列ベース変換法
   2.4. リアルタイム変換法
3. 応用例

系列ベースの変換関数

- 時系列単位で変換処理を実施
  - 入力特徴量系列ベクトル： $\boldsymbol{x} = [\boldsymbol{x}_1^{\mathrm{T}} \quad \boldsymbol{x}_2^{\mathrm{T}} \quad \cdots \quad \boldsymbol{x}_T^{\mathrm{T}}]^{\mathrm{T}}$
  - 出力特徴量系列ベクトル： $\boldsymbol{y} = [\boldsymbol{y}_1^{\mathrm{T}} \quad \boldsymbol{y}_2^{\mathrm{T}} \quad \cdots \quad \boldsymbol{y}_T^{\mathrm{T}}]^{\mathrm{T}}$
  - モデルパラメータセット　： $\boldsymbol{\lambda}$
  - 変換特徴量系列ベクトル： $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^{\mathrm{T}} \quad \hat{\boldsymbol{y}}_2^{\mathrm{T}} \quad \cdots \quad \hat{\boldsymbol{y}}_T^{\mathrm{T}}]^{\mathrm{T}}$

$$\hat{\boldsymbol{y}} = F_{\lambda}(\boldsymbol{x})$$
$$= \arg\max P(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\lambda})$$



最尤系列変換法

[Toda *et al.*, 2007]

- 系列に特化した特徴量を考慮した変換処理

1. 動的特徴量を考慮した系列ベースの最尤推定
   - フレーム間相関を考慮した変換を実現
   - 問題点1（時間的依存関係の無視）を解決
2. 系列内変動の明示的なモデル化の導入
   - 2次モーメントを考慮した変換を実現
   - 問題点2（過剰な平滑化）の影響を大幅に緩和

1. 動的特徴量の導入

[Toda *et al.*, 2007]

- 学習時には，静的特徴量のみでなく動的特徴量も含んだ結合*p.d.f.*をGMMでモデル化

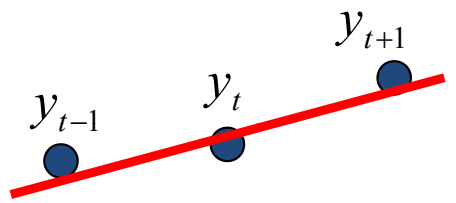
  動的特徴量（微分係数）の計算

  例えば

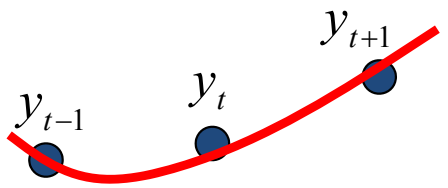
$$\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$$

  や

$$\begin{cases} \Delta \boldsymbol{y}_t = \dfrac{1}{2}\left(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t-1}\right) \\ \Delta\Delta \boldsymbol{y}_t = \boldsymbol{y}_{t-1} - 2\boldsymbol{y}_t + \boldsymbol{y}_{t+1} \end{cases}$$



回帰直線（$y=at+b$）の1次微分係数a



回帰曲線（$y=at^2+bt+c$）の2次微分係数

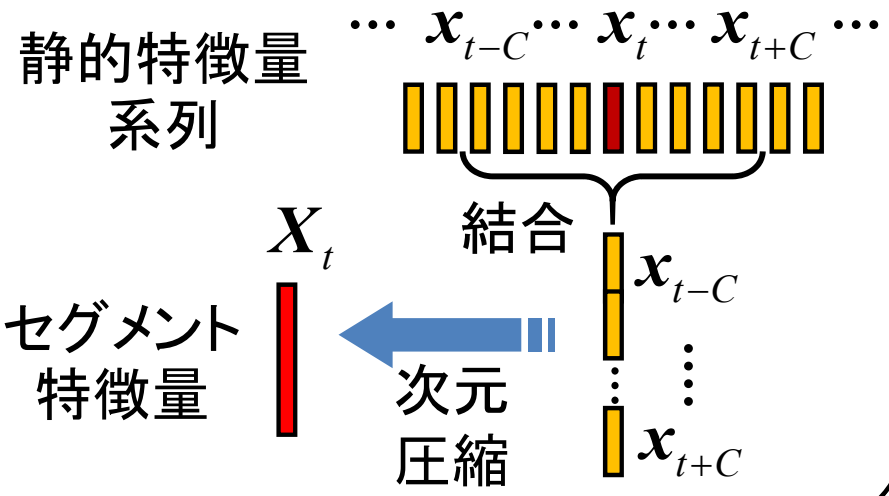
- 変換時には，静的・動的特徴量間の明示的な関係を考慮して，条件付き*p.d.f.*の尤度最大化に基づき入力特徴量ベクトルを変換

学習処理：結合確率密度のモデル化

- 時間変化を効果的に捉える特徴量空間において入力・出力結合*p.d.f.*をモデル化

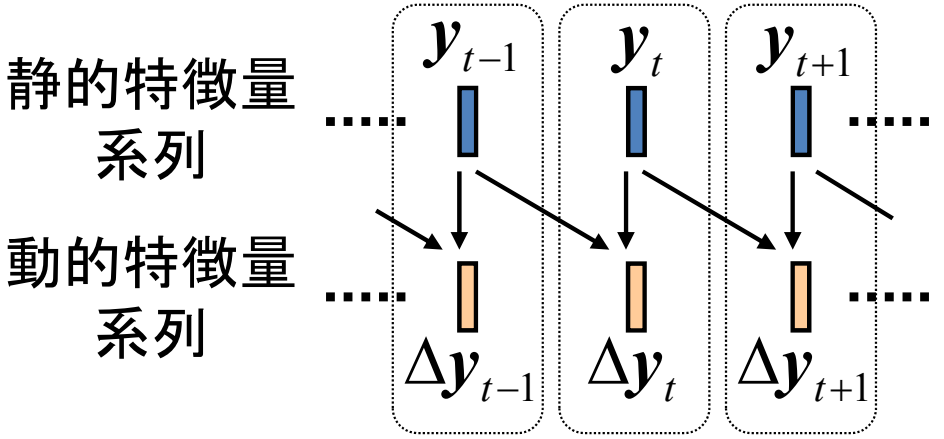
入力：セグメント特徴量

$$\boldsymbol{X}_t = \boldsymbol{A}\left[\boldsymbol{x}_{t-C}^{\mathsf{T}}, \cdots, \boldsymbol{x}_{t+C}^{\mathsf{T}}\right]^{\mathsf{T}} + \boldsymbol{b}$$



出力：静的・動的特徴量

$$\begin{bmatrix} \boldsymbol{y}_t \\ \Delta\boldsymbol{y}_t \end{bmatrix} \quad \text{ここで} \quad \Delta\boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$$

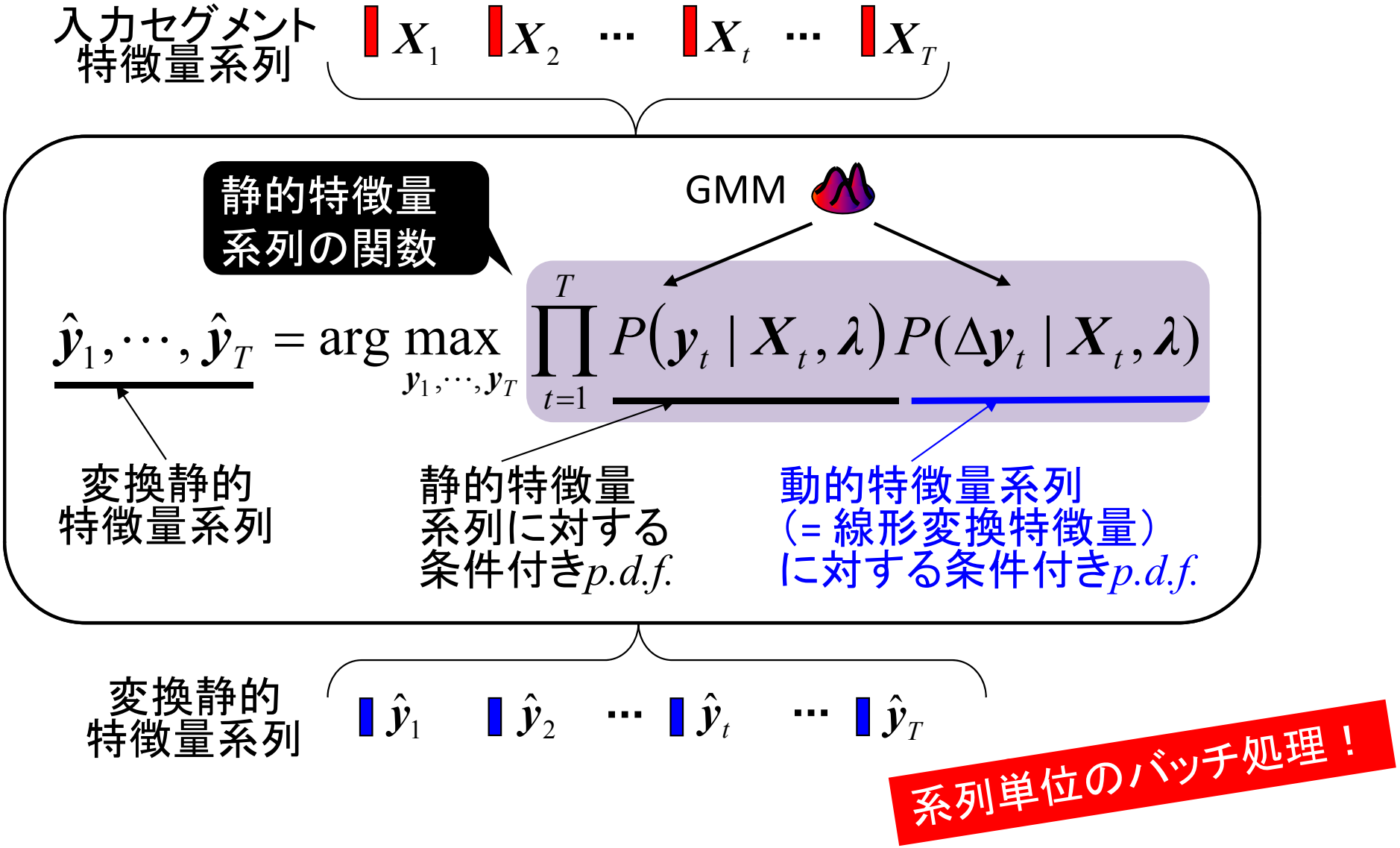


混合正規分布
モデル（GMM）

$$\prod_t p\left(\boldsymbol{X}_t, \boldsymbol{y}_t, \Delta\boldsymbol{y}_t \mid \boldsymbol{\lambda}\right)$$

変換処理：動的特徴量を考慮した変換

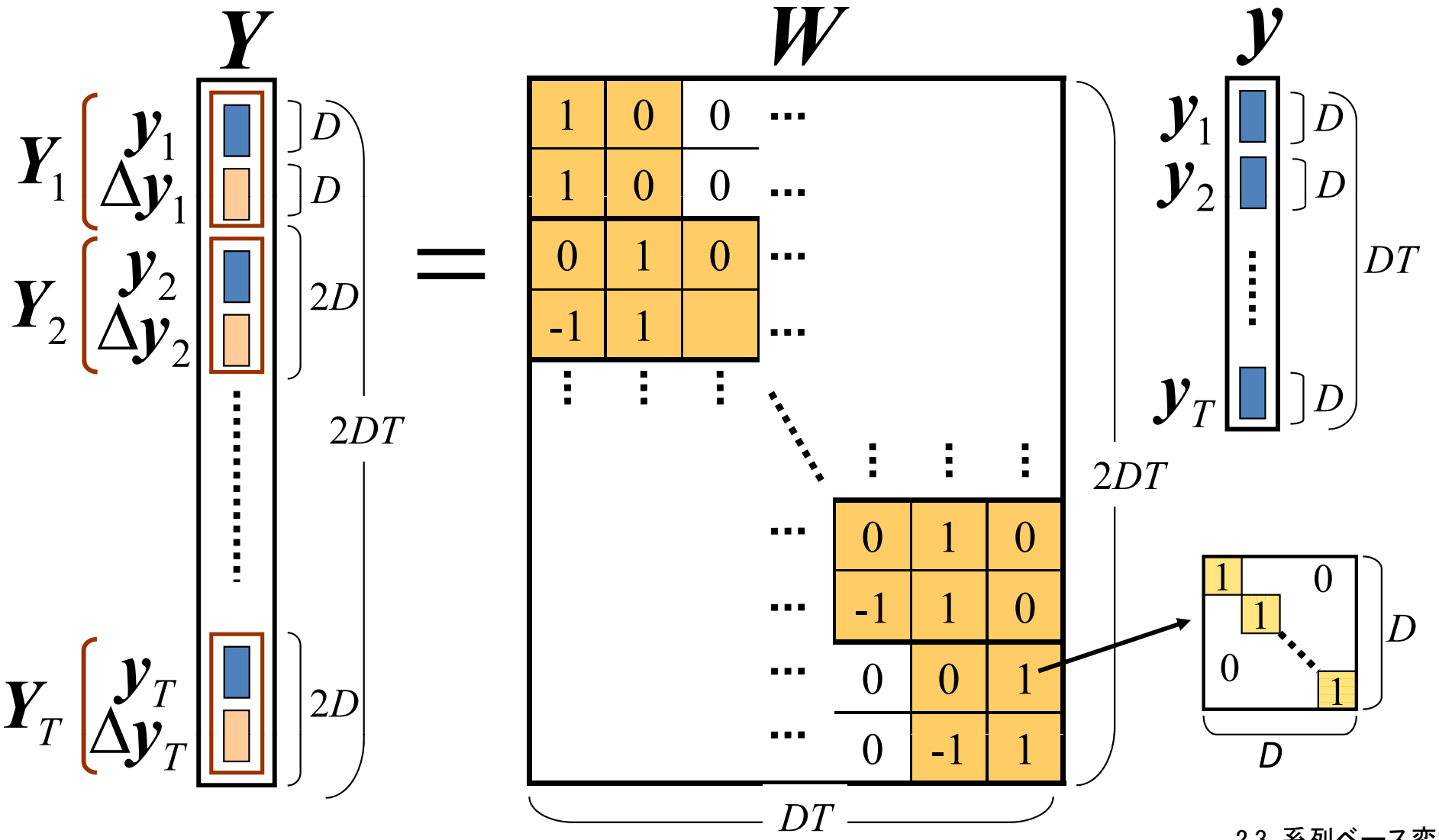
- 時系列単位での変換処理（全時間フレームを同時に変換）



静的・動的特徴量間の明示的な関係

[Tokuda *et al.*, 1997]

- 静的・動的特徴量系列 $\boldsymbol{Y} = [\boldsymbol{Y}_1^{\mathrm{T}} \ \boldsymbol{Y}_2^{\mathrm{T}} \ \cdots \ \boldsymbol{Y}_T^{\mathrm{T}}]^{\mathrm{T}}$ は静的特徴量系列 $\boldsymbol{y} = [\boldsymbol{y}_1^{\mathrm{T}} \ \boldsymbol{y}_2^{\mathrm{T}} \ \cdots \ \boldsymbol{y}_T^{\mathrm{T}}]^{\mathrm{T}}$ の線形変換 $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ により表される．



系列単位の尤度関数

尤度関数

$$P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) = \prod_{t=1}^{T} \sum_{m=1}^{M} P\left(m \mid \boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\right) P\left(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(X,Y)}\right)$$

$$= \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(X,Y)}\right)$$

事後確率

$$P\left(\boldsymbol{m} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) = \prod_{t=1}^{T} P\left(m_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\right)$$

分布系列: $\boldsymbol{m} = \{m_1, m_2, \cdots m_T\}$

正規分布

$$P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(X,Y)}\right) = \prod_{t=1}^{T} P\left(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, m_t, \boldsymbol{\lambda}^{(X,Y)}\right)$$

$$= P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(X,Y)}\right)$$

$$= N\left(\boldsymbol{Y}; \boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)}, \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)}\right)$$

観測ベクトル

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_T \end{bmatrix}$$

平均ベクトル

$$\boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)} = \begin{bmatrix} \boldsymbol{\mu}_{m_1}^{(Y|X)} \\ \boldsymbol{\mu}_{m_2}^{(Y|X)} \\ \vdots \\ \boldsymbol{\mu}_{m_T}^{(Y|X)} \end{bmatrix}$$

共分散行列

$$\boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)} = \begin{bmatrix} \boldsymbol{\Sigma}_{m_1}^{(Y|X)} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_{m_2}^{(Y|X)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Sigma}_{m_T}^{(Y|X)} \end{bmatrix}$$

最尤特徴量系列の導出

- 単一分布系列 $\boldsymbol{m} = \{m_1, m_2, \cdots m_T\}$ による近似を用いる場合：

尤度関数

$$P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(X,Y)}\right)$$
$$\approx P\left(\boldsymbol{m} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(X,Y)}\right)$$

準最適な分布系列の決定（ハードクラスタリング）

$$\hat{\boldsymbol{m}} = \arg\max P\left(\boldsymbol{m} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right)$$

出力静的特徴量系列の推定

$$\hat{\boldsymbol{y}} = \arg\max P\left(\hat{\boldsymbol{m}} \mid \boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) P\left(\boldsymbol{Y} \mid \boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(X,Y)}\right)$$
$$= \arg\max -\frac{1}{2}\left(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Y|X)}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Y|X)^{-1}} \left(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Y|X)}\right)$$
$$= \left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Y|X)^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^{\mathrm{T}} \boldsymbol{\Sigma}_{\hat{\boldsymbol{m}}}^{(Y|X)^{-1}} \boldsymbol{\mu}_{\hat{\boldsymbol{m}}}^{(Y|X)}$$

※EMアルゴリズムを用いてソフトクラスタリングを行うことも可能

トラジェクトリモデルとしての解釈

[Zen *et al.*, 2007]

静的特徴量系列ベクトル $\boldsymbol{y}$ を確率変数としたモデルの尤度最大化

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) = P(\boldsymbol{W}\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda})$$

$$= \frac{1}{\sqrt{(2\pi)^{2DT}\left|\boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)}\right|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)-1}\left(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)}\right)\right)$$

$$= Z_{\boldsymbol{m}} \frac{1}{\sqrt{(2\pi)^{2DT}\left|\boldsymbol{P}_{\boldsymbol{m}}\right|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{m}}\right)^{\mathrm{T}} \boldsymbol{P}_{\boldsymbol{m}}^{-1}\left(\boldsymbol{y} - \bar{\boldsymbol{y}}_{\boldsymbol{m}}\right)\right)$$

$$= Z_{\boldsymbol{m}} P(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda})$$

正規分布

平均：$\boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)}$

共分散：$\boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)}$

正規化項

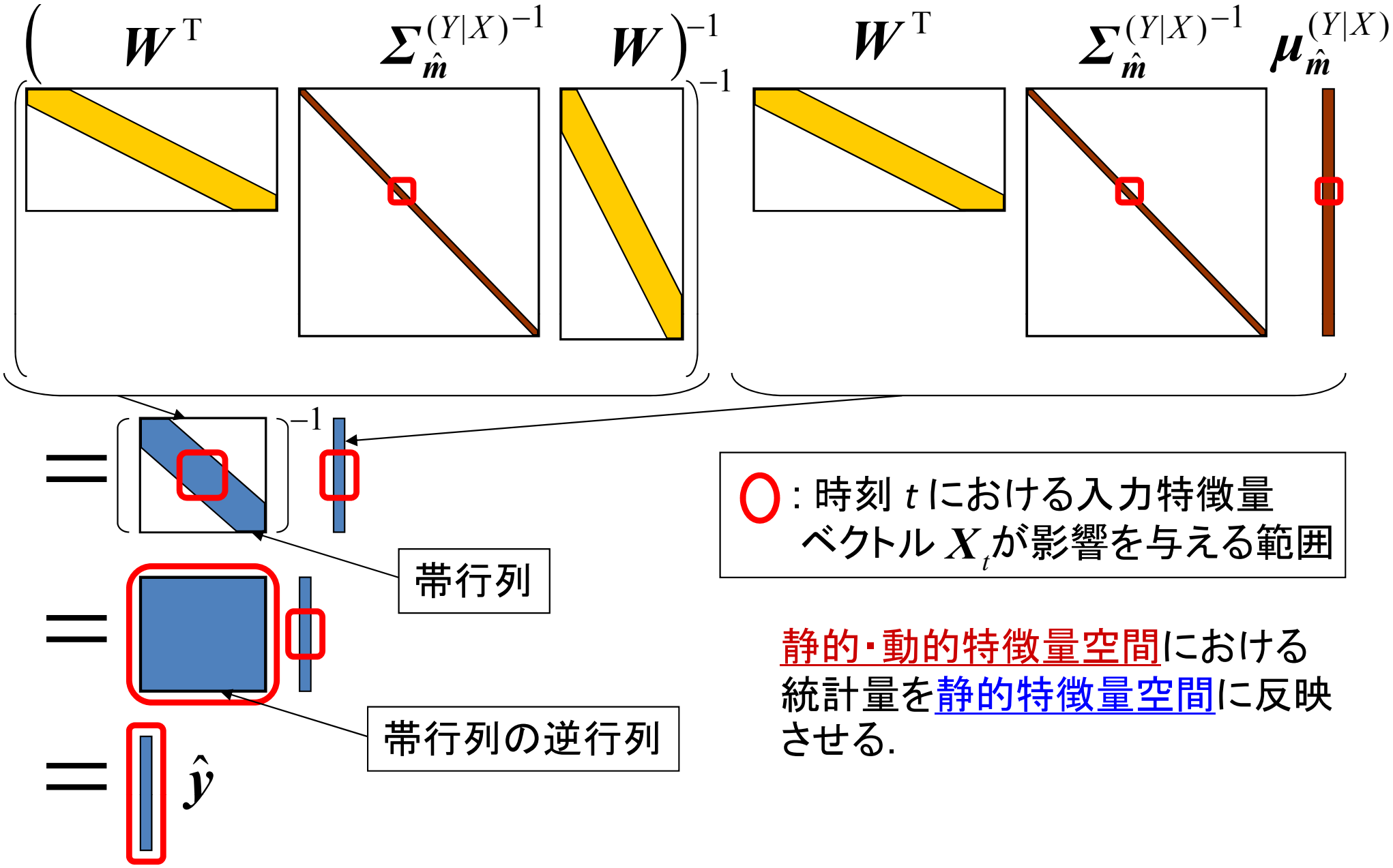
$$\frac{\sqrt{(2\pi)^{DT}\left|\boldsymbol{P}_{\boldsymbol{m}}\right|}}{\sqrt{(2\pi)^{2DT}\left|\boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)}\right|}} \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)-1} \boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)} - \bar{\boldsymbol{y}}_{\boldsymbol{m}}^{\mathrm{T}} \boldsymbol{P}_{\boldsymbol{m}}^{-1} \bar{\boldsymbol{y}}_{\boldsymbol{m}}\right)\right)$$

正規分布

平均：$\bar{\boldsymbol{y}}_{\boldsymbol{m}} = \left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)-1} \boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)}$

共分散：$\boldsymbol{P}_{\boldsymbol{m}} = \left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{m}}^{(Y|X)-1} \boldsymbol{W}\right)^{-1}$

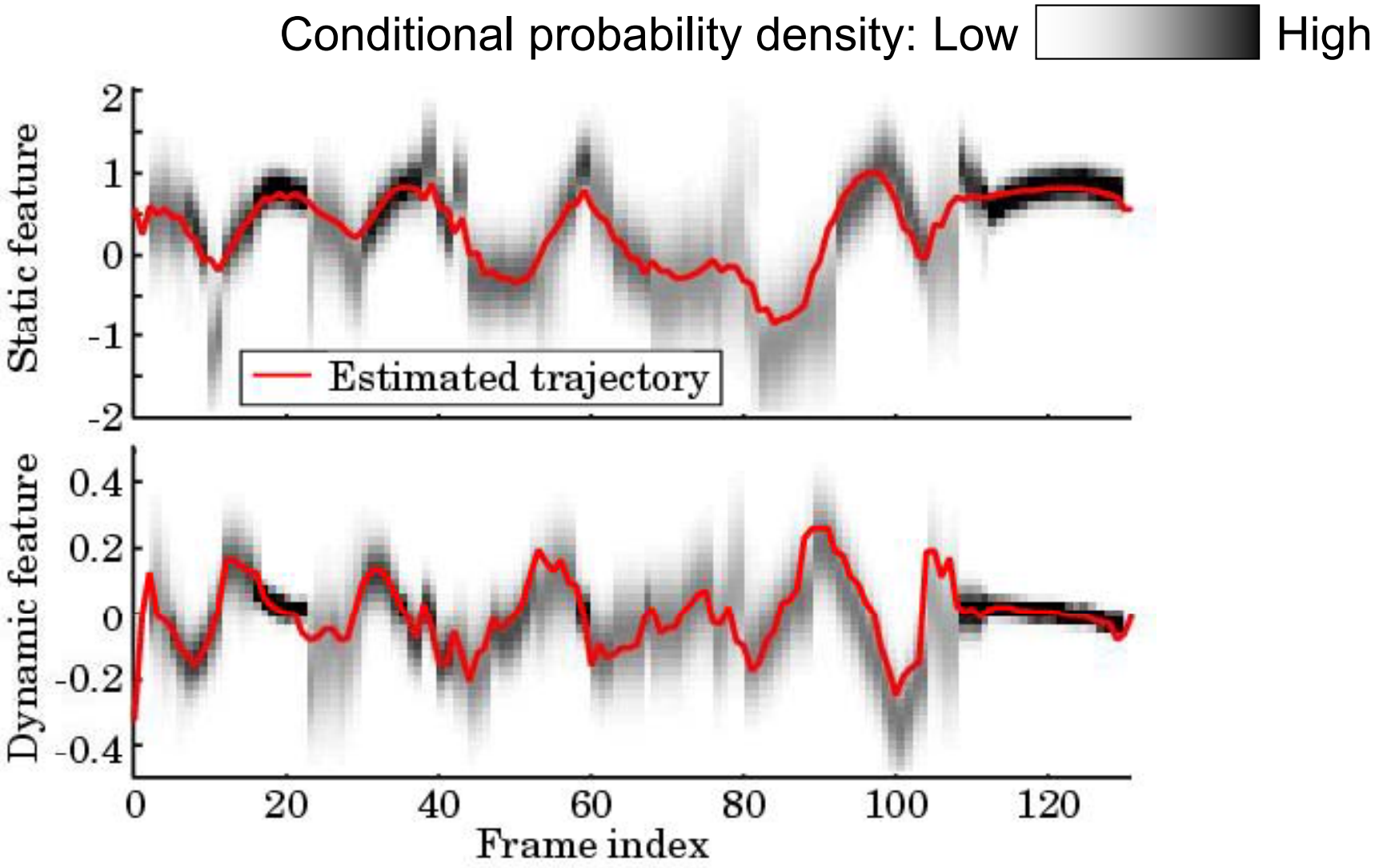
最尤特徴量系列の演算



○：時刻 t における入力特徴量ベクトル $\boldsymbol{X}_t$ が影響を与える範囲

静的・動的特徴量空間における統計量を静的特徴量空間に反映させる.

最尤特徴量系列の一例


Conditional probability density: Low
High
Static feature
Dynamic feature
Estimated trajectory
Frame index
2
1
0
-1
-2
0.4
0.2
0
-0.2
-0.4
0
20
40
60
80
100
120

最尤系列変換法

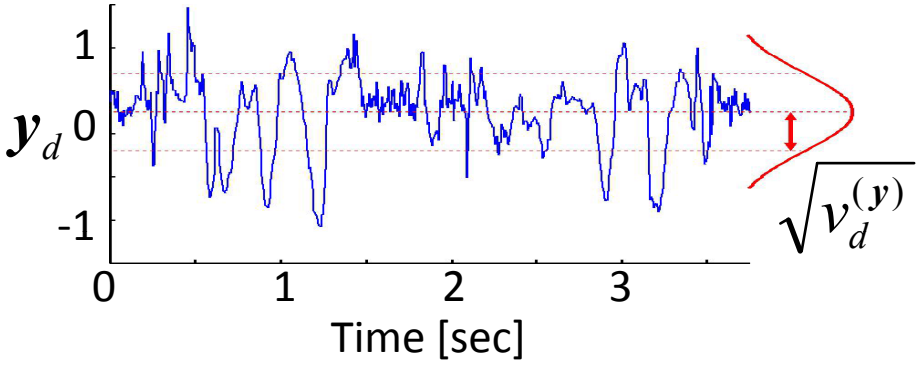
[Toda *et al.*, 2007]

- 系列に特化した特徴量を考慮した変換処理
  1. 動的特徴量を考慮した系列ベースの最尤推定
     - フレーム間相関を考慮した変換を実現
     - 問題点1（時間的依存関係の無視）を解決
  2. 系列内変動の明示的なモデル化の導入
     - 2次モーメントを考慮した変換を実現
     - 問題点2（過剰な平滑化）の影響を大幅に緩和

2. 系列内変動モデリングの導入

[Toda *et al.*, 2007]

- 学習時には，系列内（例えば一発話）の全フレームにわたって計算される分散（Global Variance: GV）の*p.d.f.*をモデル化



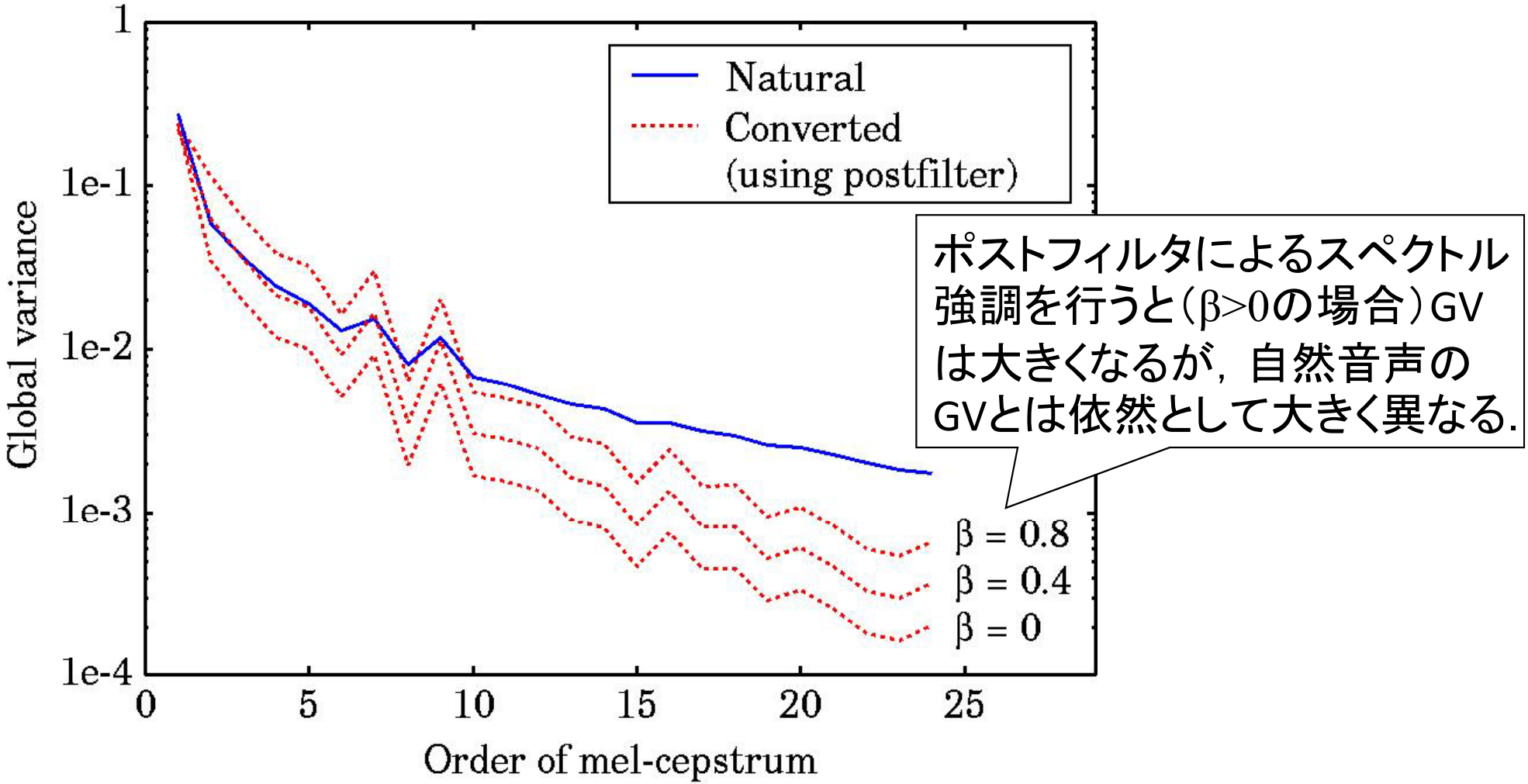
フレーム1～Tにおいて
計算されるd次元目のGV:

$$v_d^{(y)} = \frac{1}{T}\sum_{t=1}^{T}\left(y_{t,d} - \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau,d}\right)^2$$

- 変換時には，静的特徴量とGV間の明示的な関係を考慮して，条件付き*p.d.f.*及びGVの*p.d.f.*の尤度最大化に基づき入力特徴量ベクトルを変換

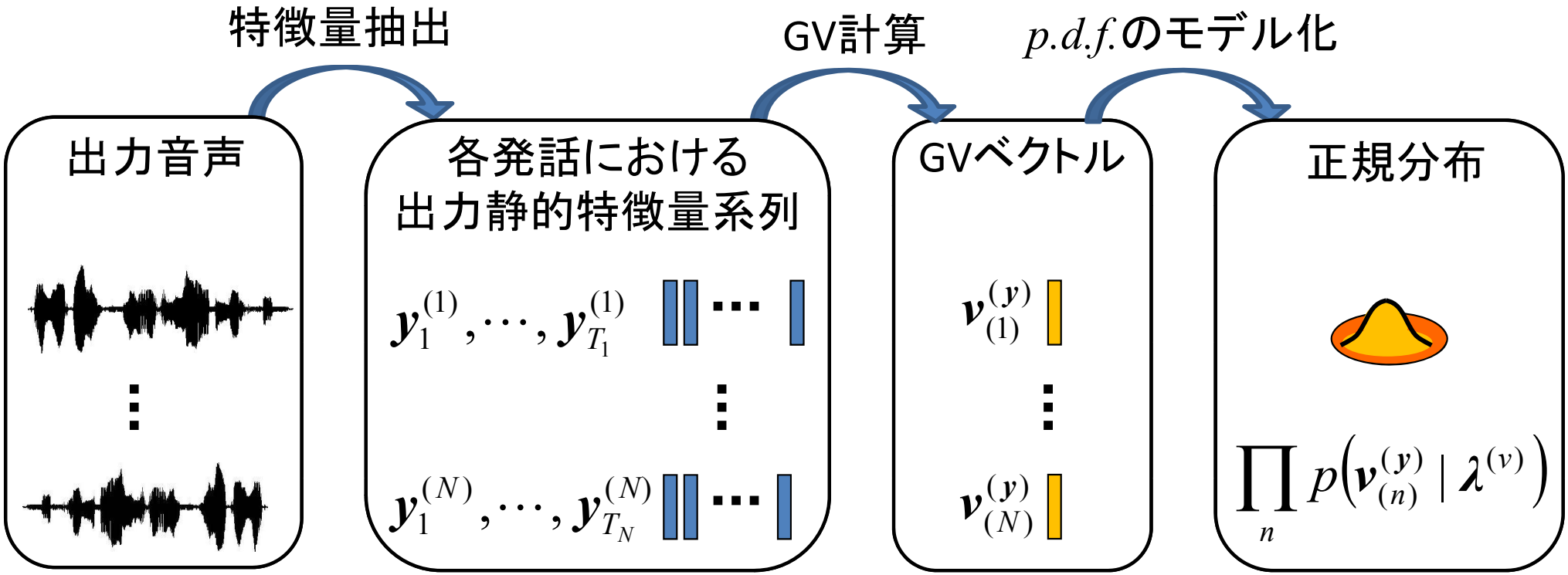
GVの統計的特徴

- 統計的変換処理で得られる最尤変換特徴量系列のGVは減少



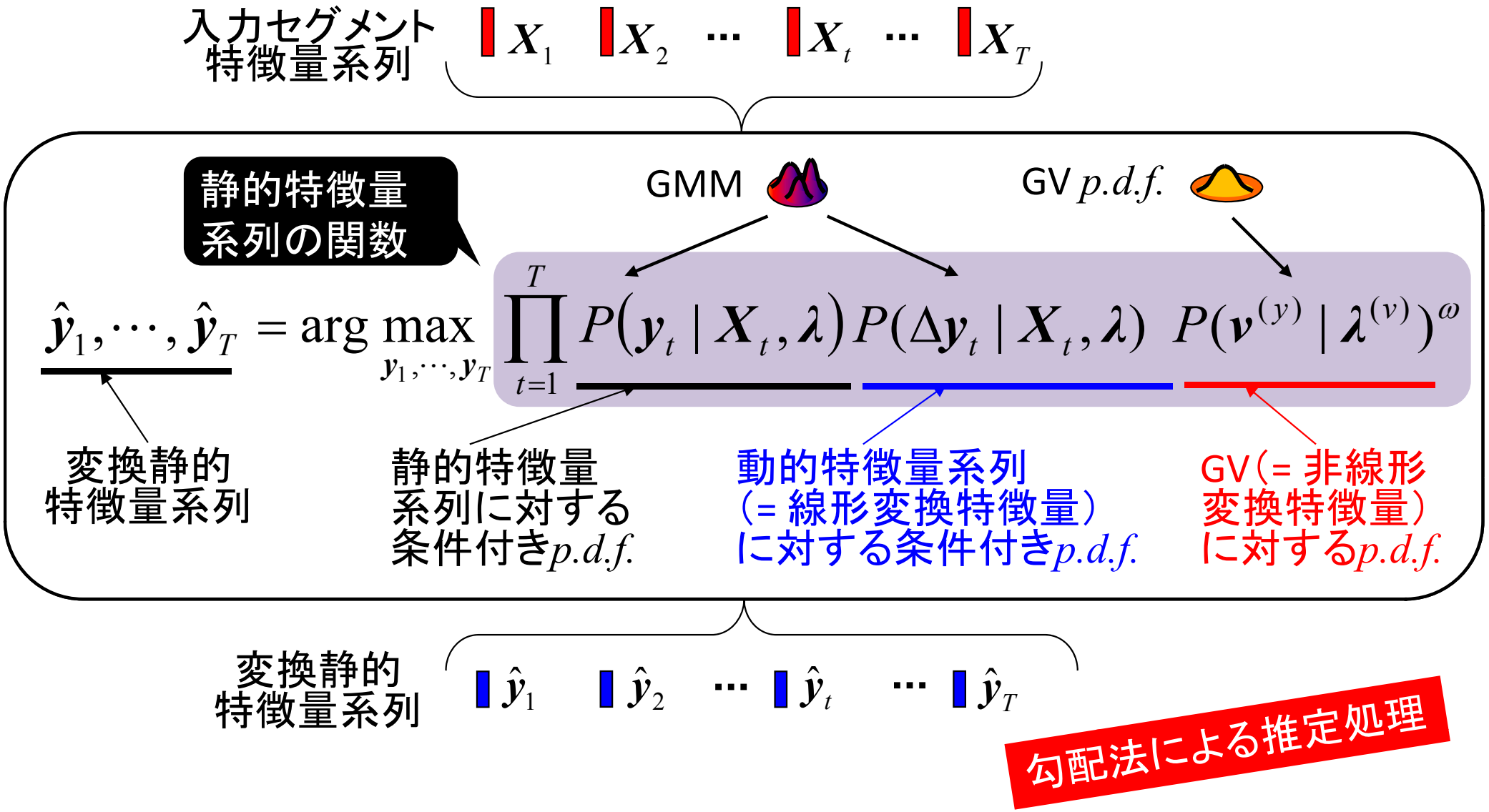
学習処理：GVのモデル化

- 出力静的特徴量系列のGVの*p.d.f.*をモデル化



変換処理：GVを考慮した変換

- 時系列単位での変換処理（全時間フレームを同時に変換）

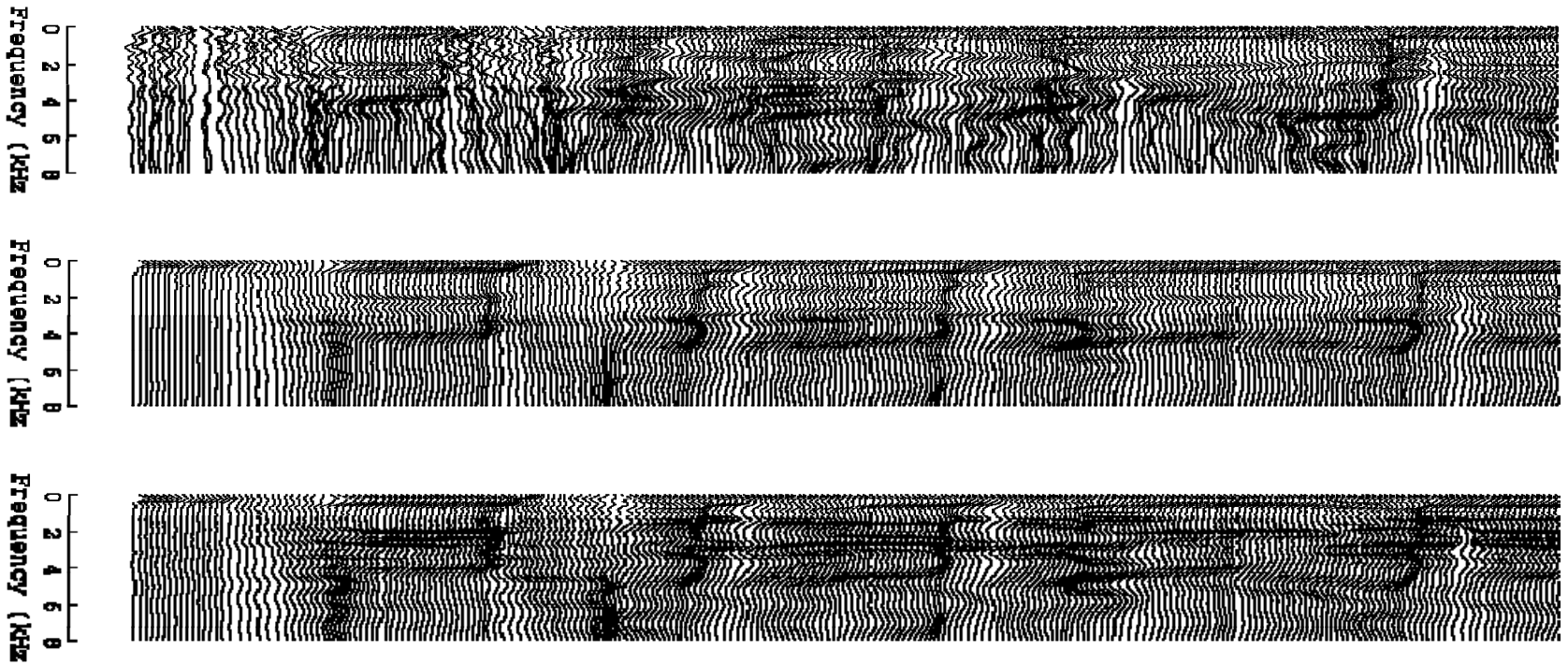


GVを考慮する効果

(a)

(b)

(c)



(a) 目標スペクトル
(b) 変換スペクトル（GV未使用）
(c) 変換スペクトル（GV使用）

話者変換音声サンプル

[Toda *et al.*, 2007]

- 話者：4名（男：bdl, 男：rms, 女：clb, 女：slt）
- 学習：50文対（完全自動学習）

目標話者

元話者	bdl	rms	clb	slt
bdl				
rms				
clb				
slt				

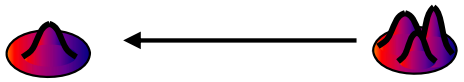
内容

1. 音声変換のしくみ
2. 統計的手法による声質変換
   2.1. 基本的な枠組み
   2.2. フレームベース変換法
   2.3. 系列ベース変換法
   2.4. リアルタイム変換法
3. 応用例

フレーム単位の変換処理への近似

[Toda *et al.*, 2012]

単一分布系列による近似：$\hat{m}_t = \arg\max_m P(m \mid \boldsymbol{X}_t, \boldsymbol{\lambda})$

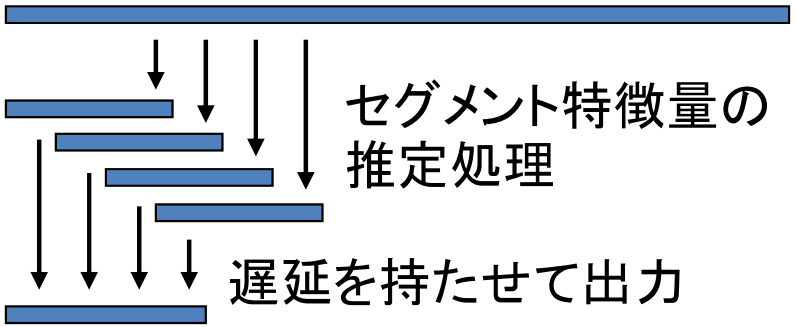


GMM 

GV *p.d.f.* 

$$\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_T = \arg\max_{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T} \prod_{t=1}^{T} P(\boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda}) P(\Delta \boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda}) \quad P(\boldsymbol{v}^{(y)} \mid \boldsymbol{\lambda}^{(v)})^{\omega}$$

短遅延変換（カルマンフィルタ）による近似



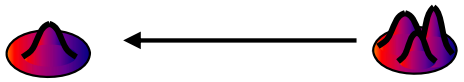
時不変ポストフィルタによる近似

$$\hat{\boldsymbol{y}}_t^{(\mathrm{GV})} = \boldsymbol{A}\, \hat{\boldsymbol{y}}_t^{(\mathrm{GV})} + \boldsymbol{b}$$

フレーム単位の変換処理への近似

[Toda *et al.*, 2012]

単一分布系列による近似：$\hat{m}_t = \arg\max_m P(m \mid \boldsymbol{X}_t, \boldsymbol{\lambda})$

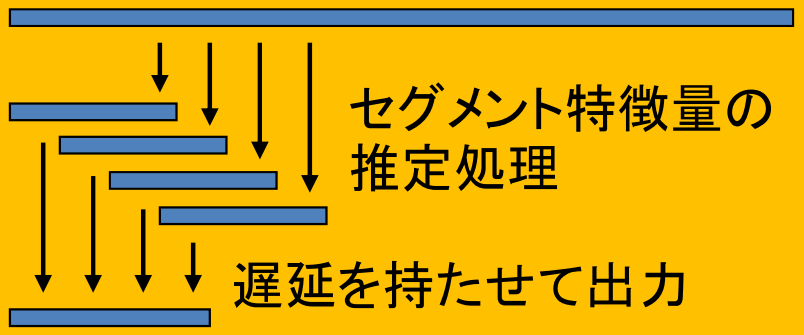


GMM 

GV *p.d.f.* 

$$\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_T = \arg\max_{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T} \prod_{t=1}^{T} P(\boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda}) P(\Delta \boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda}) \quad P(\boldsymbol{v}^{(y)} \mid \boldsymbol{\lambda}^{(v)})^{\omega}$$

短遅延変換（カルマンフィルタ）による近似



時不変ポストフィルタによる近似

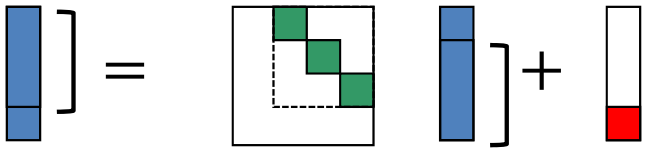
$$\hat{\boldsymbol{y}}_t^{(\mathrm{GV})} = \boldsymbol{A}\, \hat{\boldsymbol{y}}_t^{(\mathrm{GV})} + \boldsymbol{b}$$

線形動的システムによる表現

- セグメント特徴量（当該～前Lフレームの結合）を状態空間でモデル化

状態方程式：セグメント特徴量空間

$$\boldsymbol{y}_d^{(t)} = \begin{bmatrix} y_{t-L,d} \\ \vdots \\ y_{t,d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0}_{L\times 1} & \boldsymbol{I}_{L\times L} \\ 0 & \boldsymbol{0}_{1\times L} \end{bmatrix} \boldsymbol{y}_d^{(t-1)} + \begin{bmatrix} \boldsymbol{0}_{L\times 1} \\ \mu_{\hat{m}_t,t,d}^{(y|X)} + n_{\hat{m}_t,d} \end{bmatrix}$$

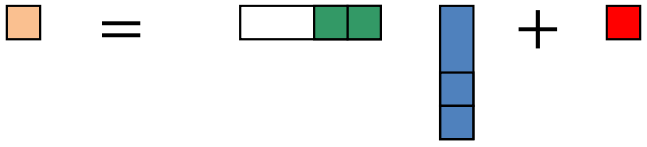


雑音：$\mathcal{N}\left(n_{\hat{m}_t,d};0,\Sigma_{\hat{m}_t,d}^{(y|X)}\right)$

静的特徴量の$p.d.f.$パラメータ

観測方程式：動的特徴量空間

$$\mu_{\hat{m}_t,t,d}^{(\Delta y|X)} = \begin{bmatrix} \boldsymbol{0}_{1\times(L-1)} & -1 & 1 \end{bmatrix} \boldsymbol{y}_d^{(t)} + e_{\hat{m}_t,d}$$



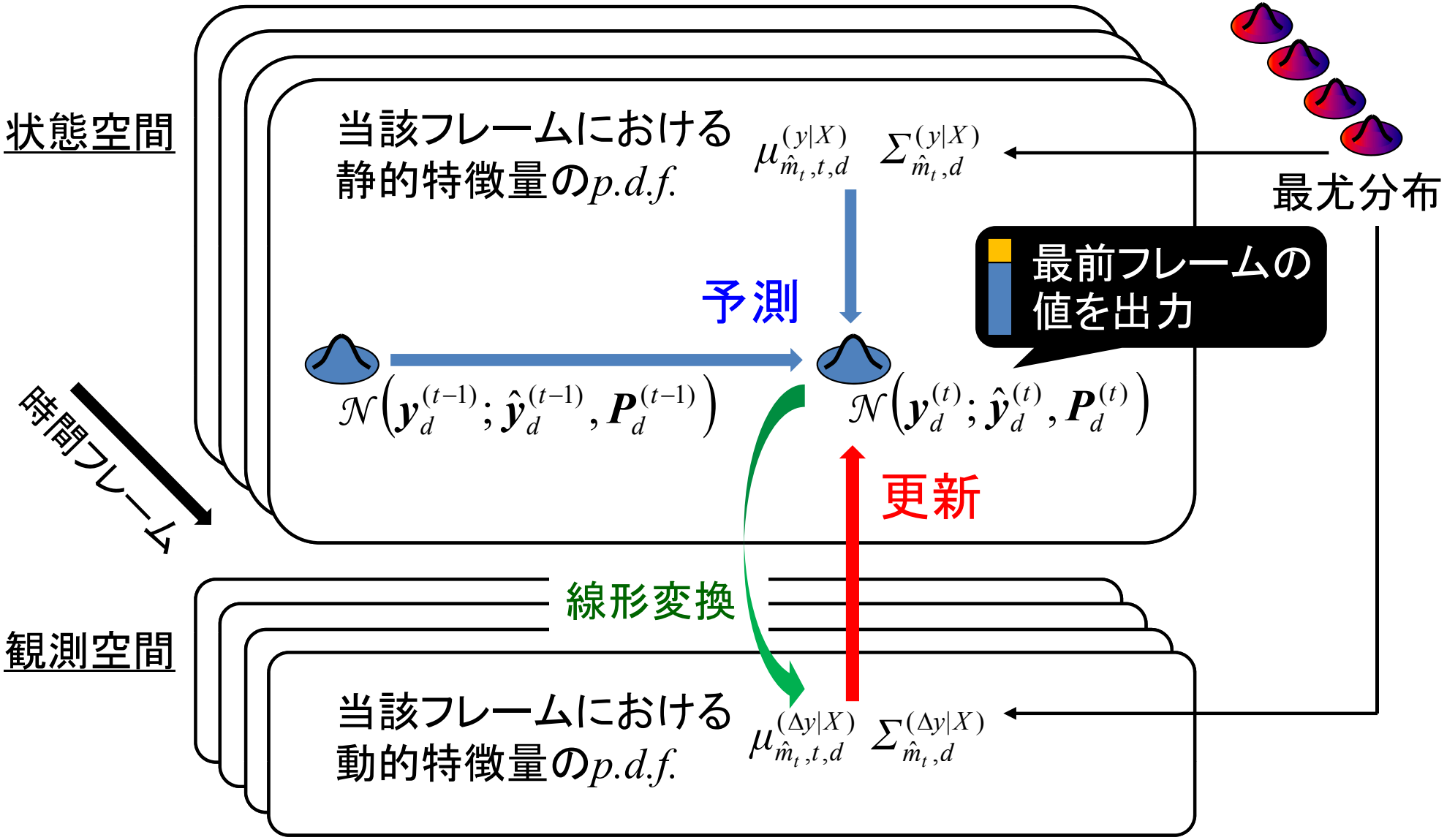
雑音：$\mathcal{N}\left(e_{\hat{m}_t,d};0,\Sigma_{\hat{m}_t,d}^{(\Delta y|X)}\right)$

動的特徴量の$p.d.f.$パラメータ

短遅延変換処理

[Muramatsu *et al.*, 2008]

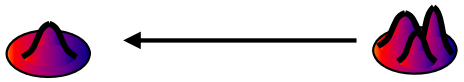
- カルマンフィルタにより状態空間分布を順次更新し平均ベクトルを出力



フレーム単位の変換処理への近似

[Toda *et al.*, 2012]

単一分布系列による近似：$\hat{m}_t = \arg\max_m P(m \mid \boldsymbol{X}_t, \boldsymbol{\lambda})$



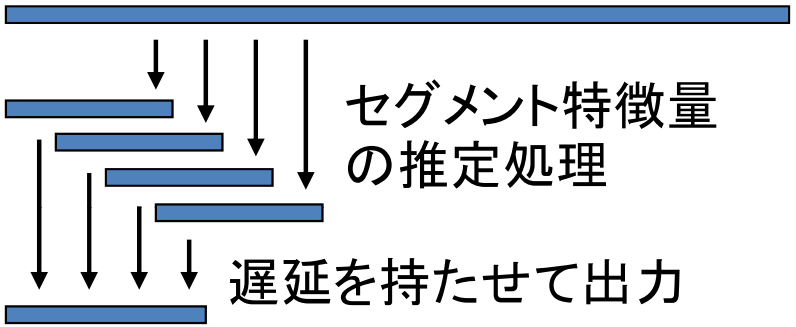
GMM

GV *p.d.f.*

$$\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_T = \arg\max_{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T} \prod_{t=1}^{T} P(\boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda}) P(\Delta \boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{\lambda})$$

$$P(\boldsymbol{v}^{(y)} \mid \boldsymbol{\lambda}^{(v)})^{\omega}$$

短遅延変換（カルマンフィルタ）による近似



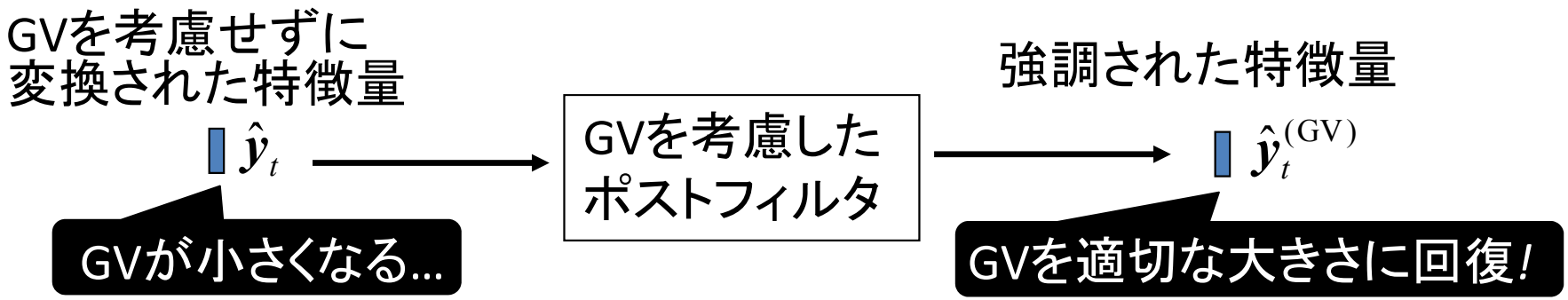
時不変ポストフィルタによる近似

$$\hat{\boldsymbol{y}}_t^{(\mathrm{GV})} = \boldsymbol{A}\, \hat{\boldsymbol{y}}_t^{(\mathrm{GV})} + \boldsymbol{b}$$

GVポストフィルタ

[Toda *et al.*, 2012]

- 時不変の線形変換による分散のスケーリング



出力特徴量のGV平均

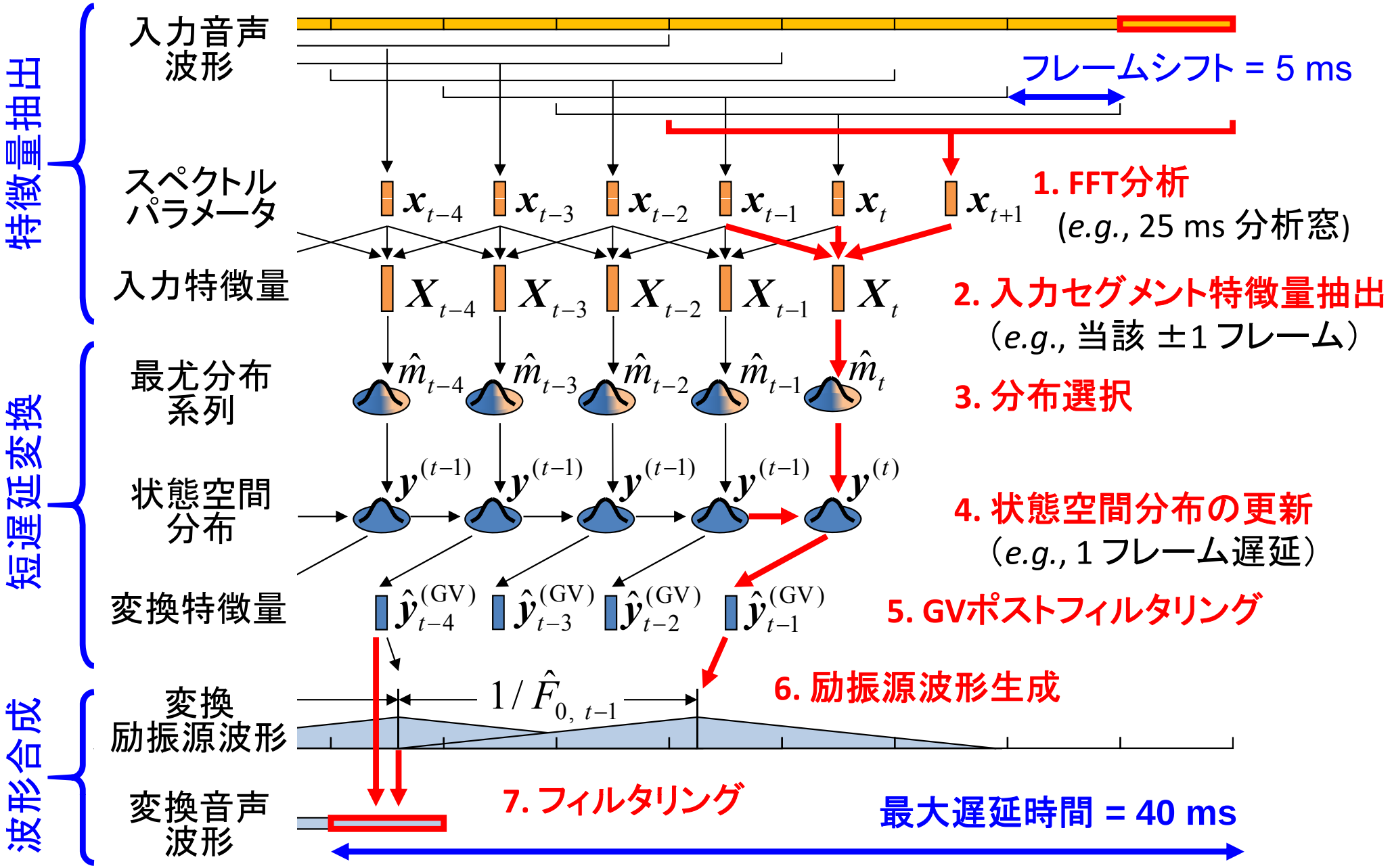
$$\hat{y}_{t,d}^{(\mathrm{GV})} = \sqrt{\frac{\mu_d^{(v)}}{\hat{\mu}_d^{(v)}}}\left(\hat{y}_{t,d} - \langle \hat{y}_d \rangle \right) + \langle \hat{y}_d \rangle$$

強調後　　強調前

学習データに対する
変換特徴量のGV平均

学習データに対する
変換特徴量の平均

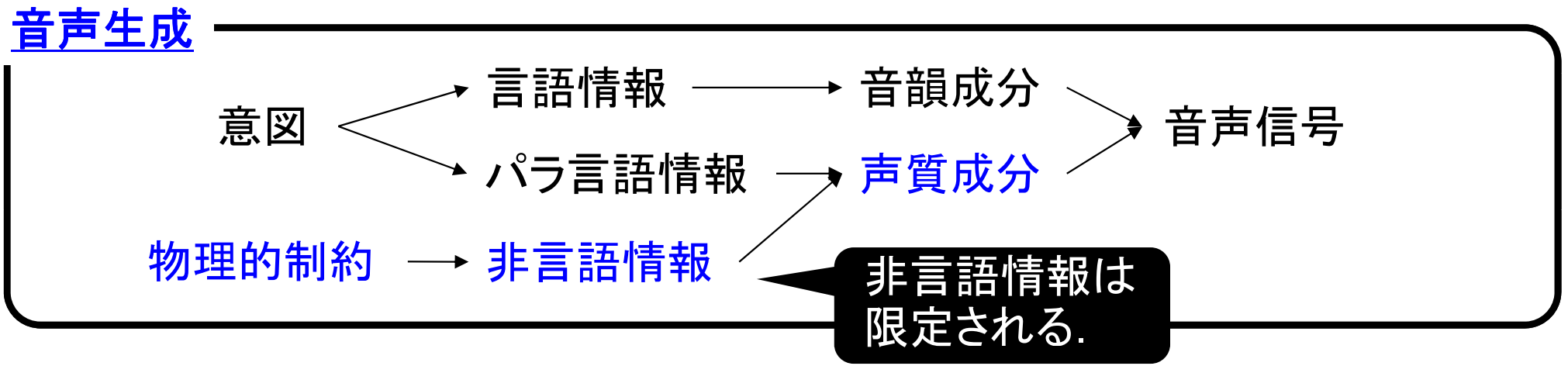
リアルタイム変換処理の流れ



内容

1. 音声変換のしくみ
2. 統計的手法による声質変換
3. 応用例
   3.1. 体内伝導音声通話
   3.2. 発声障害者補助
   3.3. ボイスチェンジャー／ボーカルエフェクター
   3.4. 帯域拡張
   3.5. 調音運動制御による音声変換

音声生成過程における物理的制約



しかし，強い物理的制約は時として障壁を生み出す！

- ✓音声生成機能の一部が失われたら深刻な発声障害を患う･･･
- ✓周囲の状況によっては声を発することが躊躇される･･･
- ✓演劇や歌唱における表現が限定される･･･

声質変換により何が実現できるのか？

機能拡張された音声生成

意図 → 言語情報 → 音韻成分
意図 → パラ言語情報 → 声質成分
物理的制約 → 非言語情報 → 声質成分
音韻成分 + 声質成分 → 音声信号

意図 → 所望の非言語情報 → **声質変換** → 変換音声信号

物理的制約を超えた非言語情報制御が可能となる！

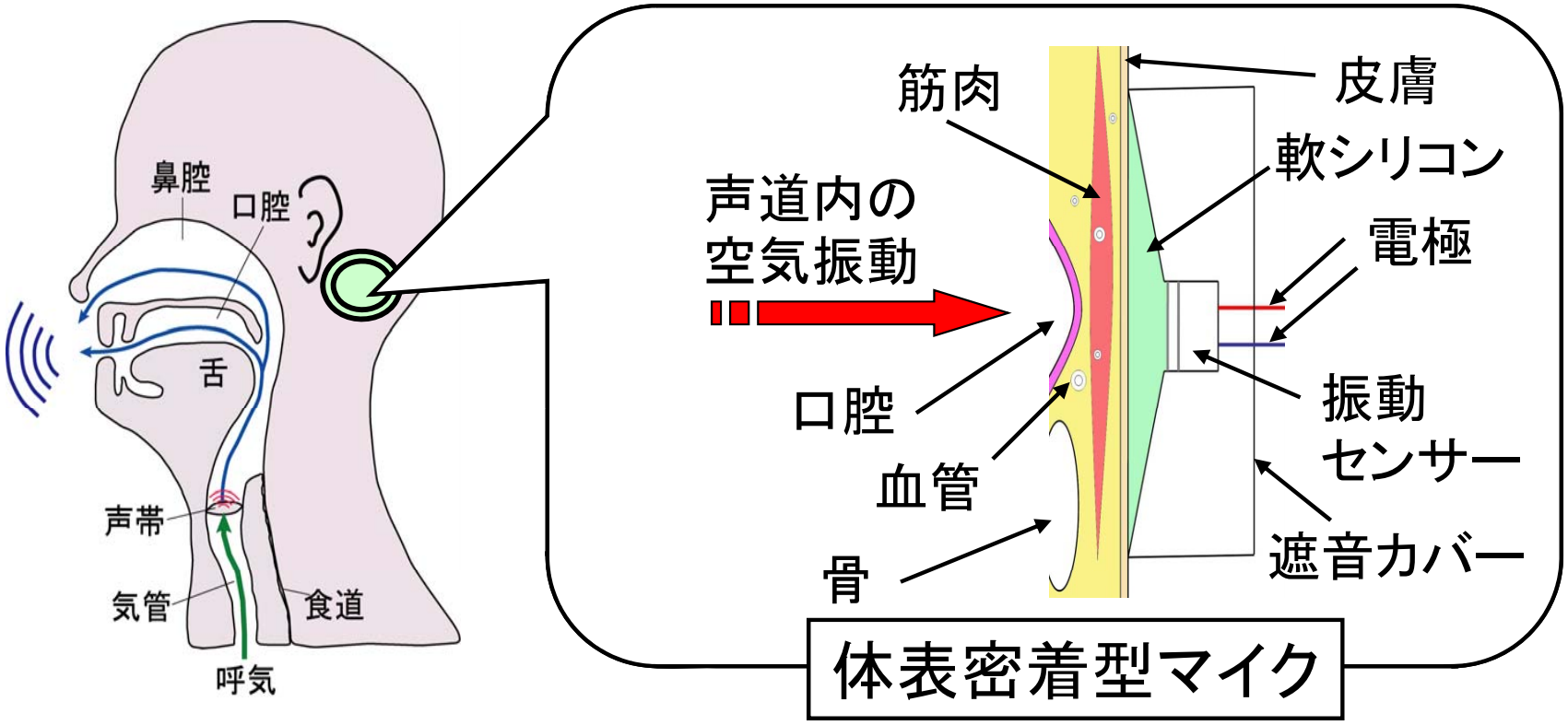
リアルタイム処理が重要！

＋声質変換 ➡ 不可能を可能とするバリアフリー音声コミュニケーションの実現！

応用1. 体内伝導音声通話

[中島 他, 2004]

- 体表密着型マイクを用いて軟組織を伝わる音声（体内伝導音声）を体表から収録し，音声通話に利用
  - 普通の音声から周囲に聞こえないぐらい小さな音声まで収録可能



サイレント音声通話

[Toda *et al.*, 2012]

- 周囲に聞こえないぐらい小さな体内伝導音声（非可聴つぶやき）をより自然な音声へとリアルタイムで変換

話し手側

周囲に聞かれたくない内容を
非可聴つぶやきにより発声

非可聴
つぶやき
マイク

変換

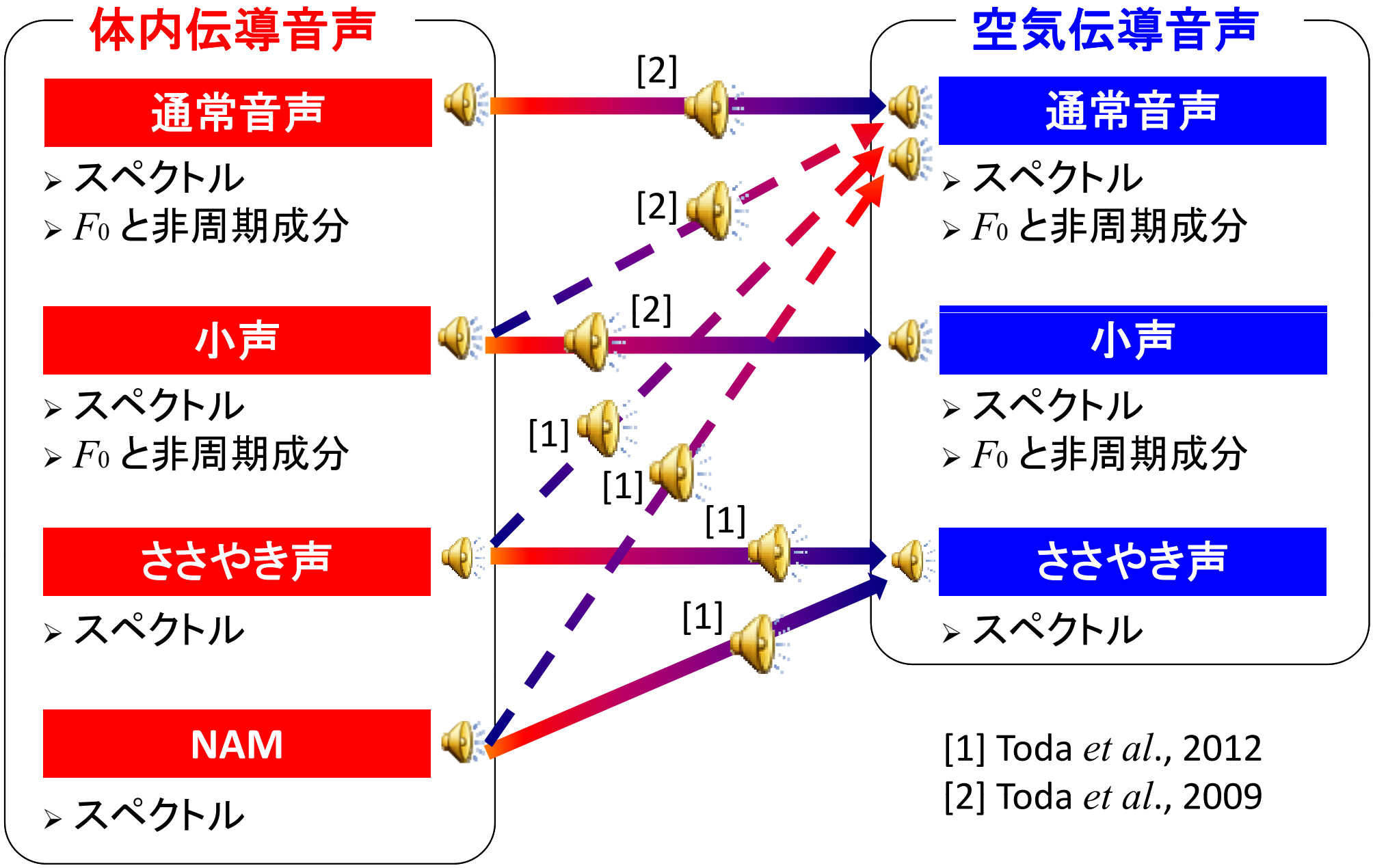
聞き手側

より自然で明瞭な音声を
聞き手にのみ提示

空気伝導音声への変換
- ✓ 通常音声（F_0予測が必要）
- ✓ ささやき声（F_0予測は不要）

テレパシーのような音声コミュニケーションを可能とする技術

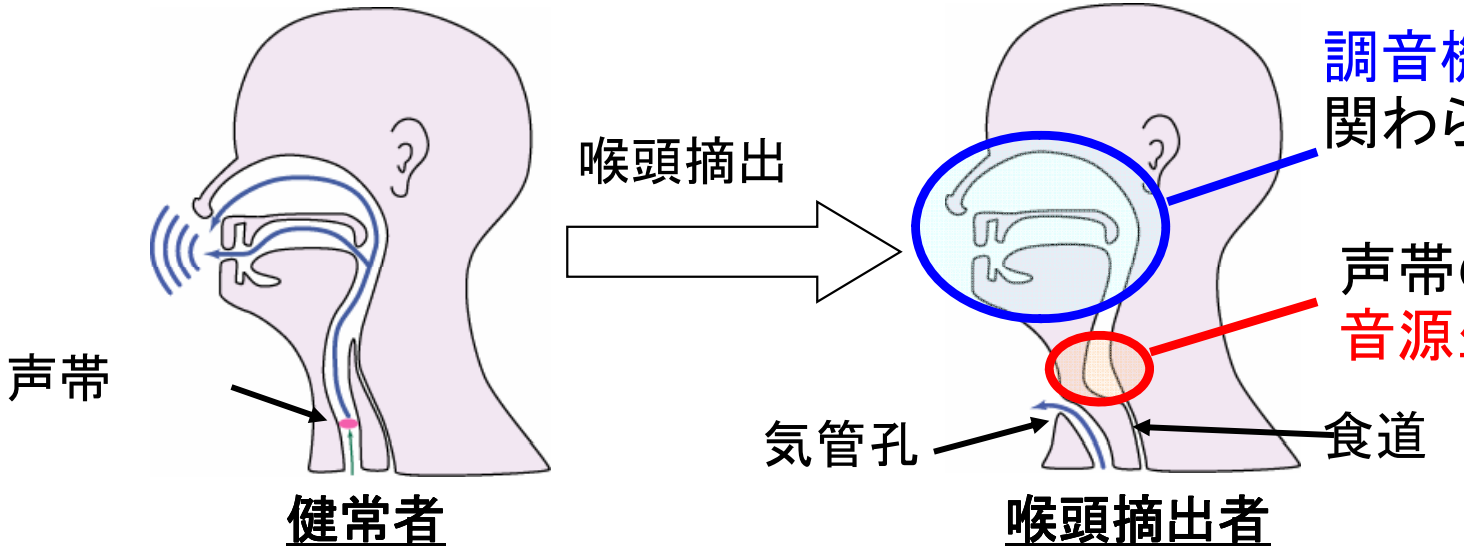
各種体内伝導音声に対する変換音声



[1] Toda *et al.*, 2012
[2] Toda *et al.*, 2009

応用2. 発声障害者補助

- 発声障害の一例：喉頭摘出による発声障害
  - 喉頭摘出者
    - 喉頭癌等の理由により喉頭を摘出した人
    - 全国に約2万人
    - 気管と食道が完全に分離

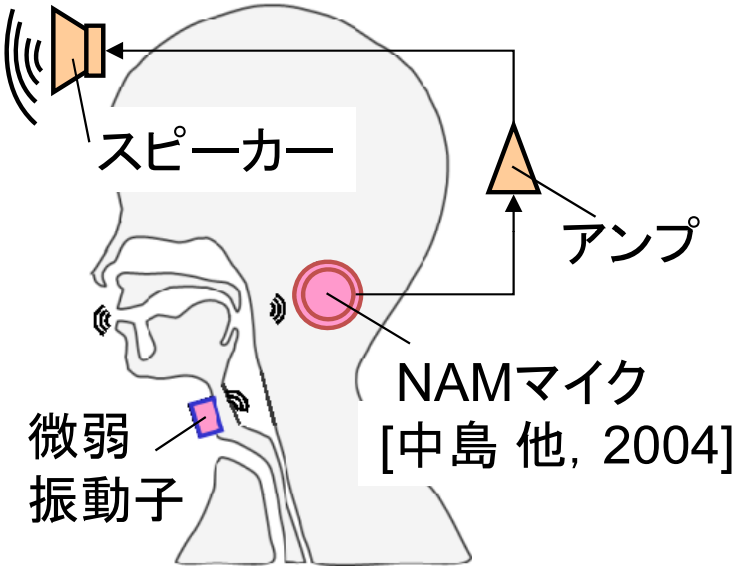


調音機能は正常であるにも関わらず発声不可能！

声帯の除去による音源生成機能の消失

発声能力が失われることによる生活の質（Quality of Life: QOL）の低下は極めて深刻な問題！

喉頭摘出者のための代替発声法

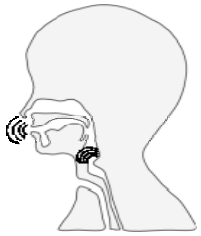
	食道発声	電気式人工喉頭を用いた発声 [橋場 他, 2001]	微弱振動子及びNAMマイクを用いた発声 [中村 他, 2007]
			
補助器具	不要	必要	必要
習得	困難	容易	比較的容易
伝達される音信号	無喉頭音声	無喉頭音声及び人工喉頭の音源信号	スピーカーから提示される無喉頭音声
無喉頭音声	食道音声	電気音声	肉伝導微弱電気音声

無喉頭音声から通常音声への変換

目標話者：

食道音声の変換

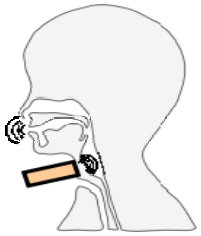
[Doi *et al*., 2009－2010]



電気音声の変換

[Nakamura *et al*., 2009－2010]

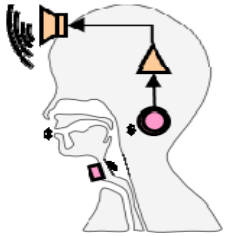
[Doi *et al*., 2010]

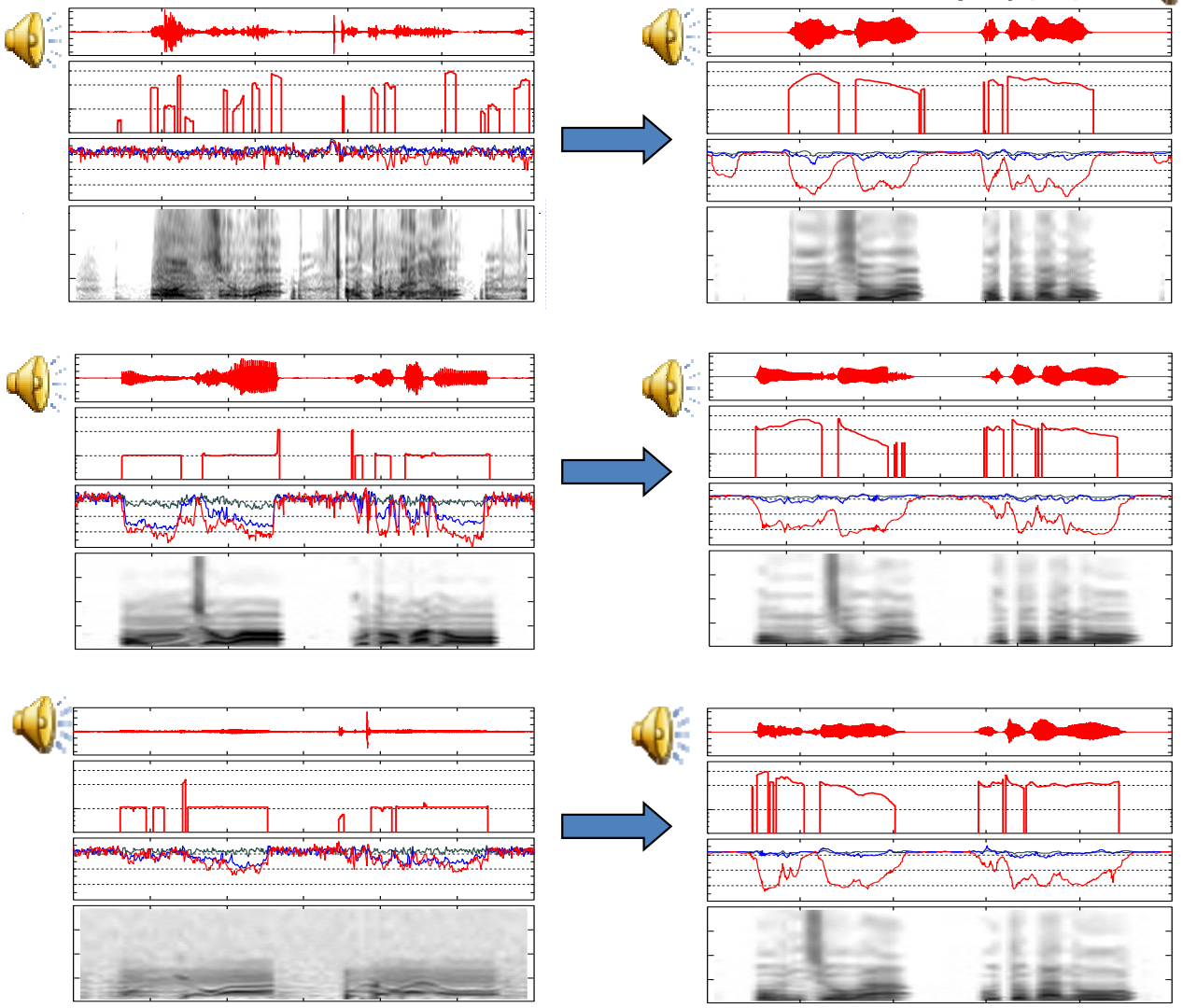


肉伝導微弱電気音声の変換

[Nakamura *et al*., 2007－2010]

[Doi *et al*., 2010]



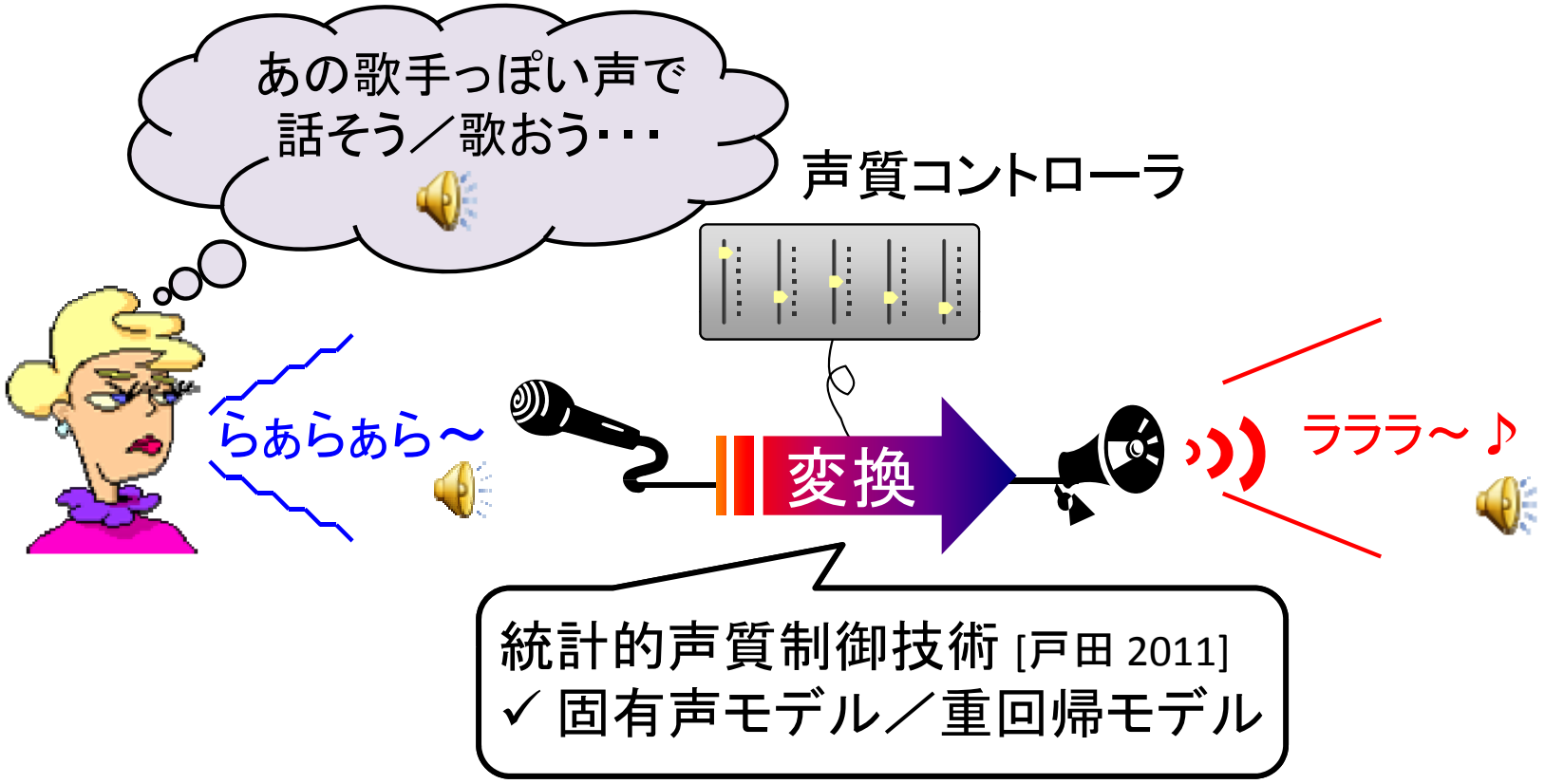


身体的制約を超えた発声（失われた声の回復）を可能とする技術

応用3. ボイス／ボーカルエフェクタ

[Doi *et al.*, 2012]

- ユーザの音声／歌声を所望の声質へとリアルタイムで変換

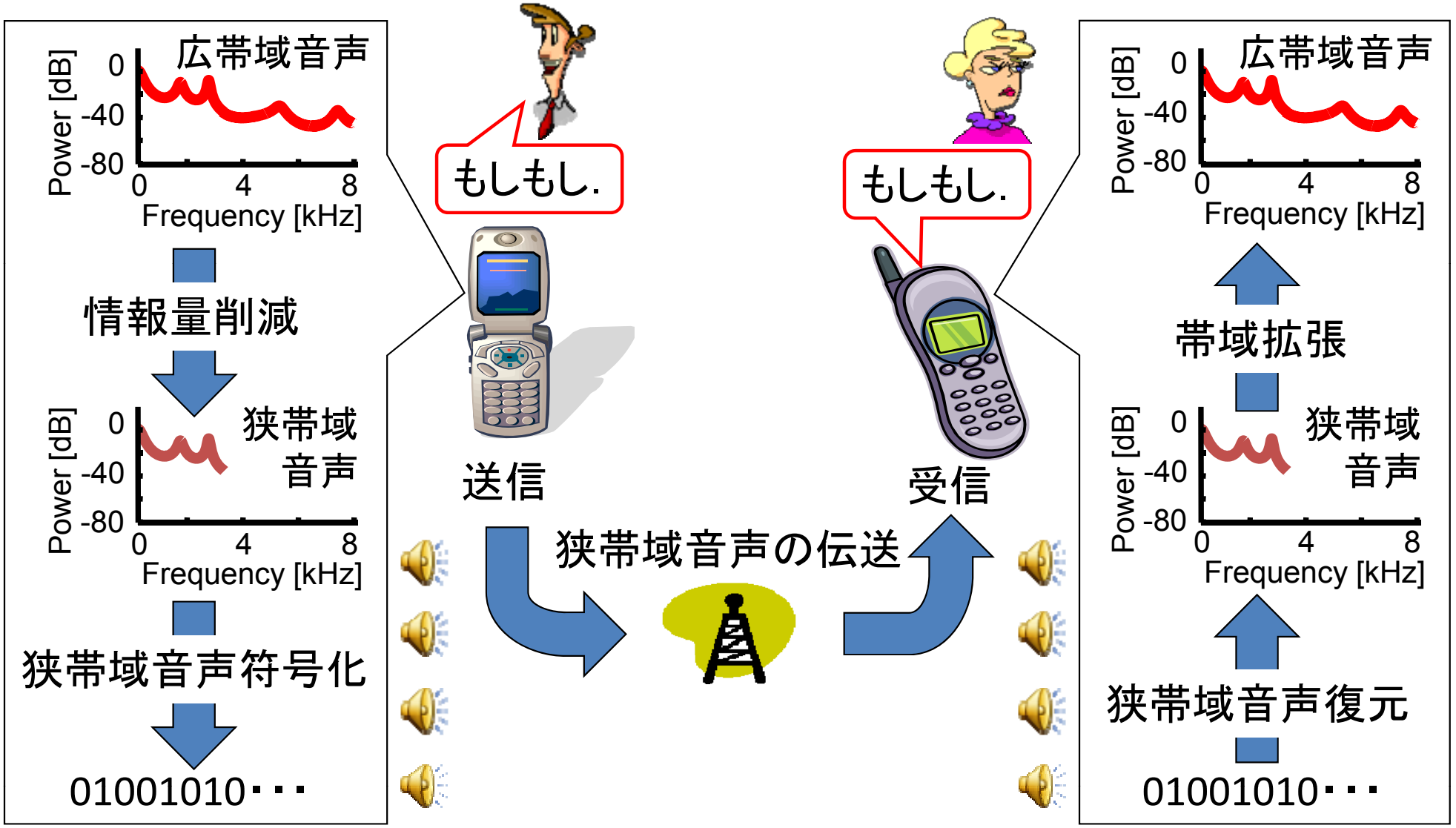


身体的制約を超えた発声／歌唱表現を可能とする技術

応用4. 携帯電話音声の帯域拡張

[Jax and Vary, 2003]

- 従来の情報伝送量で高品質な広帯域音声を受聴可能



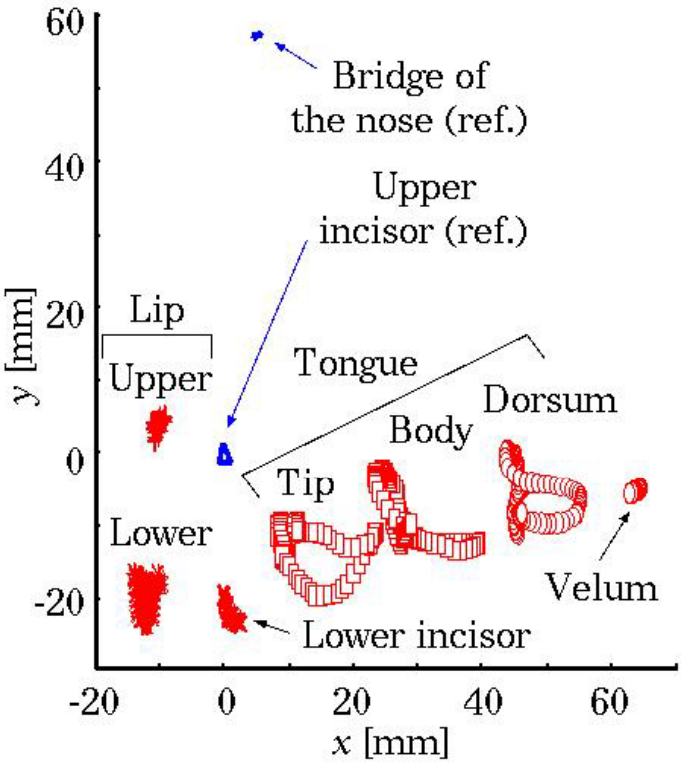
応用5. 調音運動制御による音声変換

[Toda *et al.*, 2008]

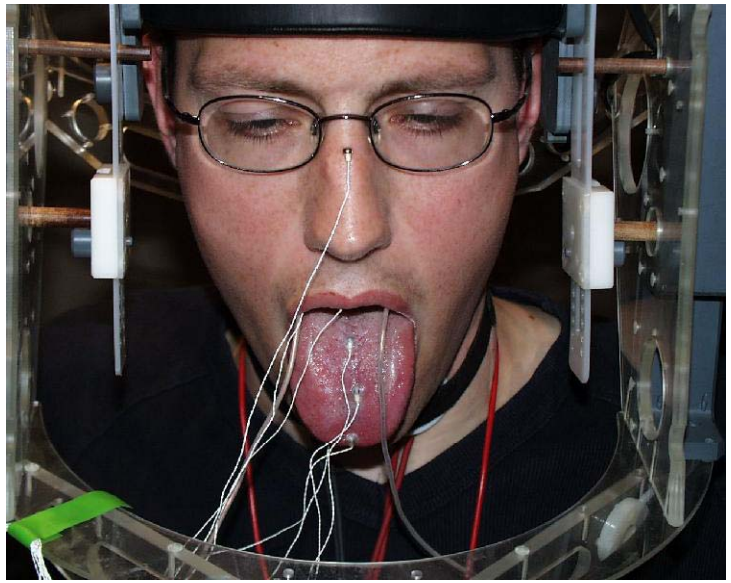
- 調音運動による音声変換を実現

  手順1. 調音運動と音声信号を同時に収録

  手順2. 調音運動と音声特徴量の対応関係をモデル化



Electromagnetic articulograph (EMA)データ

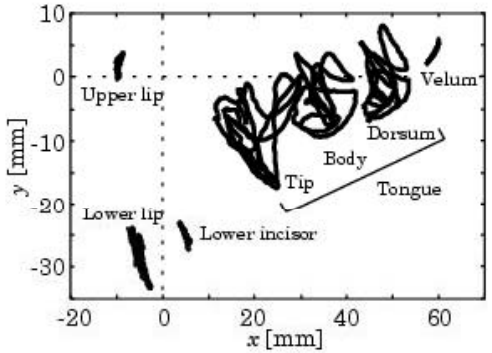


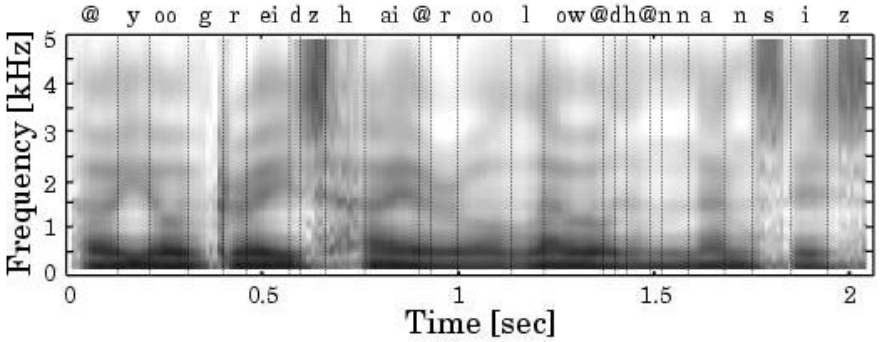
MOCHAデータベース
Edinburgh大から公開

http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

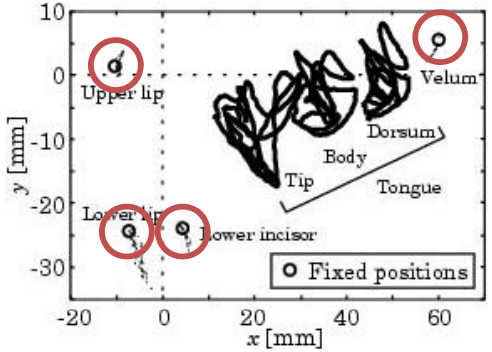
調音運動制御による音声変換の一例

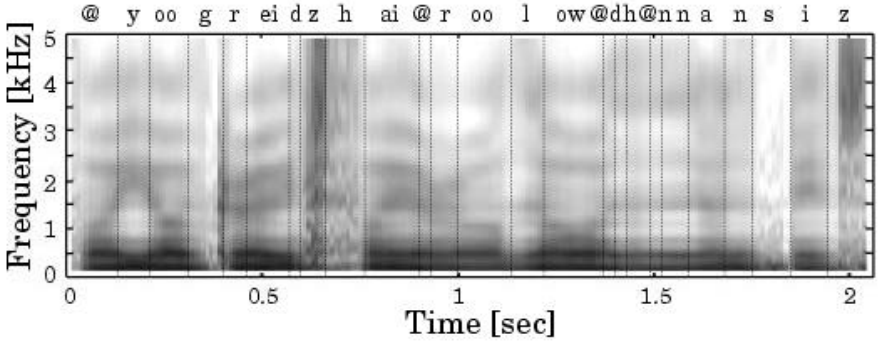
1. 自然な調音運動からの合成音



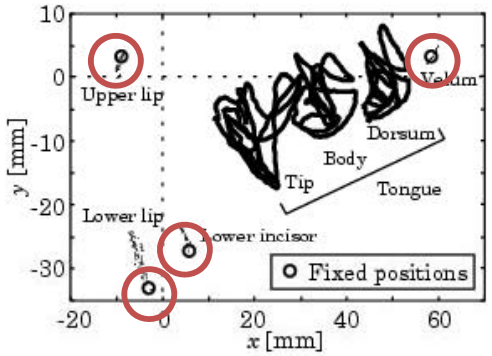


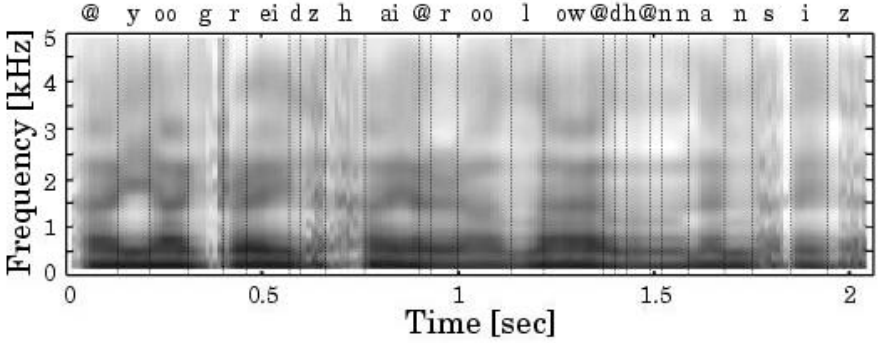
2. 口をすぼめたままの合成音





3. 口を開けたままの合成音





まとめ

統計的手法による音声変換

1. 音声変換のしくみ
2. 統計的手法による声質変換
3. 応用例

- 声質変換の進展

  統計処理　＋　リアルタイム処理　＝　音声生成機能拡張

- 声質変換の**有効性**と**危険性**
  - 「包丁のようなものである」

    by 産総研 後藤

参考文献

- M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, “Voice conversion through vector quantization,” *J. Acoust. Soc. Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.
- 中村哲，鹿野清宏，“ファジィベクトル量子化を用いたスペクトログラムの正規化,” *日本音響学会誌*, vol. 45, no. 2, pp. 107－114, 1989.
- H. Matsumoto and Y. Yamashita, “Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function,” *J. Acoust. Soc. Jpn. (E)*, vol. 14, no. 5, pp. 353－361, 1993.
- H. Valbret, E. Moulines, and J. P. Tubach, “Voice transformation using PSOLA technique,” *Speech Communication*, vol. 11, no. 2－3, pp. 175－187, 1992.
- Y. Stylianou, O. Cappe, E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.
- A. Kain, M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *Proc. ICASSP*, Seattle, USA, pp. 285–288, May 1998.
- T. Toda, A.W. Black, K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Trans. Audio, Speech and Language Process*., vol. 15, no. 8, pp. 2222–2235, 2007.
- H. Zen, K. Tokuda, Alan W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol.51, no.11, pp.1039–1154, 2009.
- K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.
- H. Zen, K. Tokuda, and T. Kitamura, “Reformulating the HMM as a trajetory model by imposing explicit relationships between static and dynamic feature vector sequences,” *Computer Speech and Language*, vol. 21, no. 1, pp. 153–173, 2007.
- 徳田恵一, 益子貴史, 小林隆夫, 今井 聖, “動的特徴を用いた HMMからの音声パラメータ生成アルゴリズム,” 日本音響学会誌, vol.53, no.3, pp.192–200, Mar. 1997.
- T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano, “Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.
- T. Toda, T. Muramatsu, H. Banno, “Implementation of computationally efficient real-time voice conversion,” *Proc. INTERSPEECH*, Portland, USA, Sep. 2012.

参考文献

- M. Abe, K. Shikano, and H. Kuwabara, “Statistical analysis of bilingual speaker’s speech for cross-language voice conversion,” *J. Acoust. Soc. Am.*, vol. 90, no. 1, pp. 76–82, 1991.
- M. Mashimo, T. Toda, H. Kawanami, K. Shikano, N. Campbell, “Cross-language voice conversion evaluation using bilingual databases,” *IPSJ Journal*, vol. 43, no. 7, pp. 2177–2185, 2002.
- 中島淑貴，柏岡秀紀，ニック キャンベル，鹿野清宏，“非可聴つぶやき認識,” *電子情報通信学会論文誌*, vol. J87-D-II, no. 9, pp. 1757–1764, 2004.
- T. Toda, M. Nakagiri, K. Shikano, “Statistical voice conversion technique for body-conducted unvoiced speech enhancement,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 20, no. 9, pp. 2505–2517, 2012.
- T. Toda, K. Nakamura, H. Sekimoto, K. Shikano, “Voice conversion for various types of body transmitted speech,” *Proc. ICASSP*, pp. 3601–3604, Taipei, Taiwan, Apr. 2009.
- 中村圭吾，戸田智基，猿渡洋，鹿野清宏，“肉伝導人工音声の変換に基づく喉頭全摘出者のための音声コミュニケーション支援システム,” *電子情報通信学会論文誌*, vol. J90-D, no. 3, pp. 780–787, 2007.
- H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, “Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models,” *IEICE Trans. on Inf. and Syst.*, vol. E93-D, no. 9, pp. 2472—2482, 2010.
- K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” *Speech Communication*, vol. 54, no. 1, pp. 134—146, 2012.
- H. Doi, T. Toda, K. Nakamura, H. Saruwatari, K. Shikano, “Alaryngeal speech enhancement based on one-to-many eigenvoice conversion ,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 22, no. 1, pp. 172—183, 2014.
- H. Doi, T. Toda, T. Nakano, M. Goto, S. Nakamura, “Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system,” *Proc. APSIPA ASC*, Hallywood, USA, Nov. 2012.
- P. Jax, P. Vary, “On artificial bandwidth extension of telephone speech,” *Signal Processing*, vol. 83, pp. 1707–1719, 2003.
- T. Toda, A.W. Black, K. Tokuda, “Mapping between articulatory movements and acoustic spectrum with a Gaussian mixture model, ” *Speech Communication*, vol. 50, no. 3, pp. 215–227, Mar. 2008.