スペシャルセッション
「音楽情報処理と機械学習」

歌声情報処理：声質変換モデル

戸田 智基

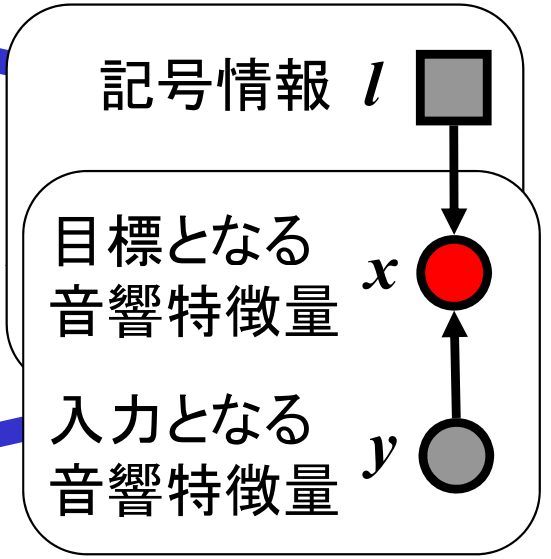
奈良先端科学技術大学院大学 情報科学研究科

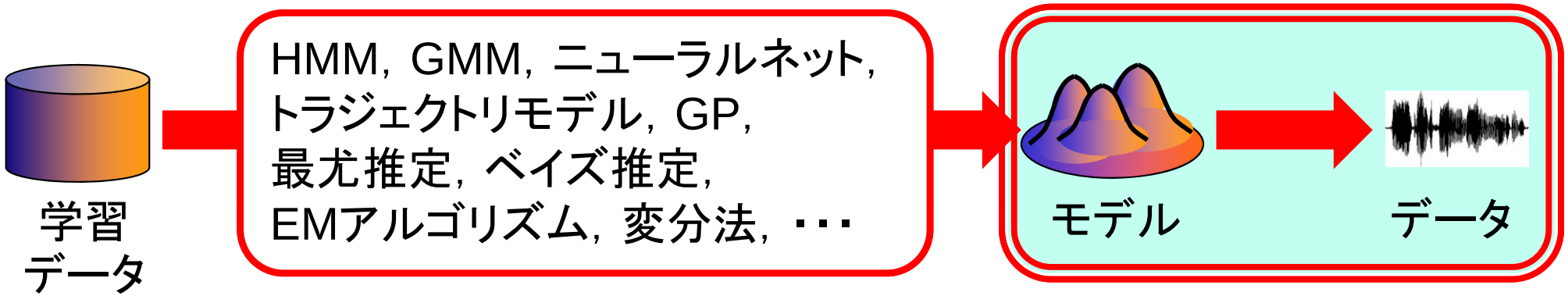
2012年8月9日

機械学習を用いた研究の進展

歌声／音声変換・合成の定式化

- 確率モデルによる音響特徴量のモデル化
  - **歌声合成，テキスト音声合成**
    - 確率密度関数（$p.d.f.$）$P(\boldsymbol{x}\,|\,\boldsymbol{l})$ のモデル化
    - 記号情報が付与された歌声／音声データを用いて**自動的に学習**
  - **歌声声質変換，声質変換**
    - $P(\boldsymbol{x}\,|\,\boldsymbol{y})$ のモデル化
    - 記号情報は同一で所望の声質成分のみが異なる歌声／音声データを用いて**自動的に学習**





確率分布からの時系列データ生成法

[徳田 他, 1995]

静的特徴量と**動的特徴量**（各時刻における時間方向の変化量）の関係を導入することで，適切に遷移する時系列データを生成

$\boldsymbol{c}$：静的特徴量系列

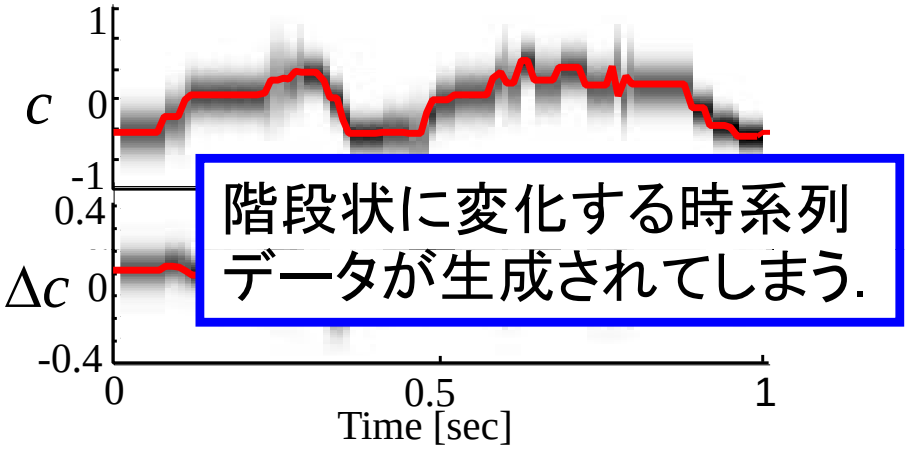
$\boldsymbol{o}$：静的・動的特徴量系列

$\boldsymbol{q}$：$\boldsymbol{o}$ をモデル化する分布系列

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

$=$平均ベクトル系列

静的・動的特徴量系列の*p.d.f.*



階段状に変化する時系列データが生成されてしまう．

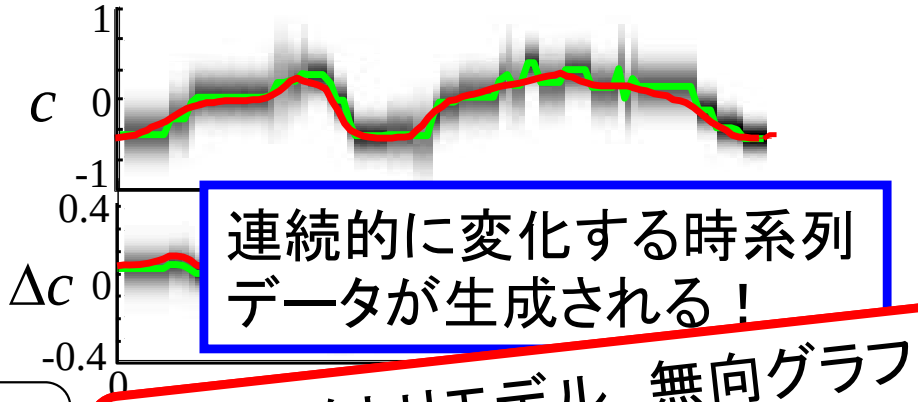
動的特徴量を考慮した時系列データ生成法

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

$$= \arg\max_{\boldsymbol{c}} P(\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\lambda}) P(\Delta\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\lambda})$$

静的特徴量系列の*p.d.f.*　　動的特徴量系列の*p.d.f.*

静的特徴量系列を線形変換 $\Delta\boldsymbol{c} = f_{\Delta}(\boldsymbol{c})$ した変数の*p.d.f.*



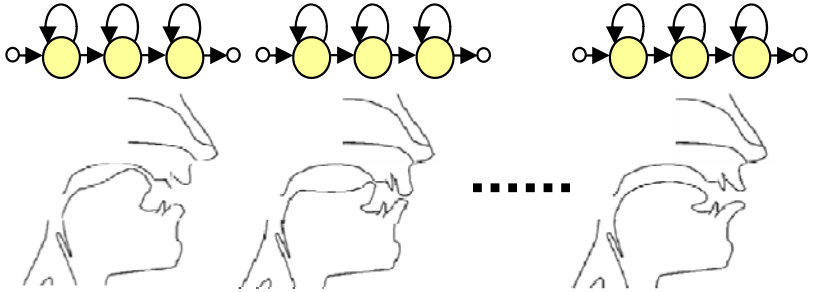
連続的に変化する時系列データが生成される！

トラジェクトリモデル，無向グラフ，線形動的システム，・・・

変換・合成：もっと音を良くしたい！

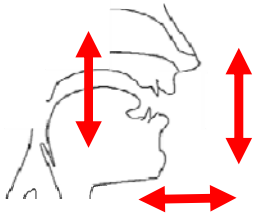
2004年当初・・・，GMMに基づく声質変換およびHMM音声合成による合成音声はこもり感が強く，自然性が低かった・・・

人がHMM尤度最大化基準で発声したらどうなるかな？



口・舌を動きが小さくなりそう・・・
はっきりと話さなくなりそう・・・

各音韻を発声する際の**平均的**な
口・舌の形を使って発声する.

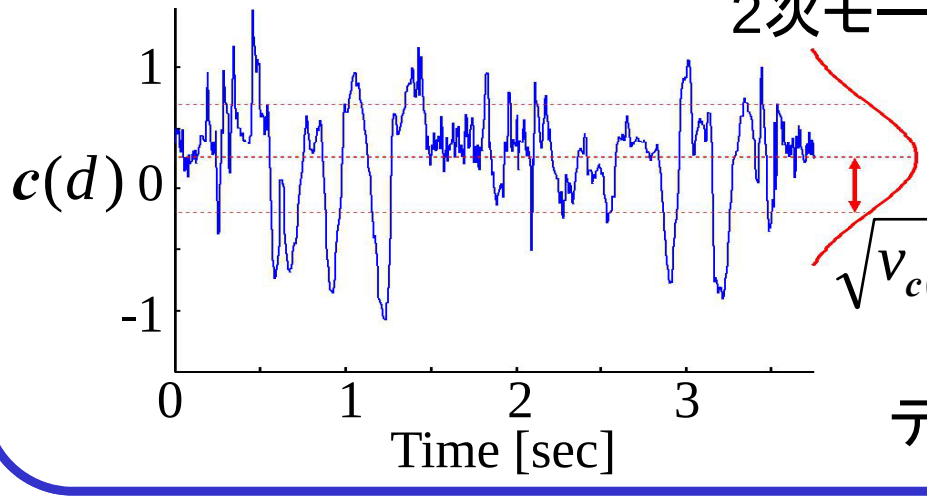


発話全体を通して口・舌をどの程度動かすかという**変動量**という基準も使っている気がする！

系列内変動（GV）を考慮した生成法

[Toda *et al.*, 2004]

GV (global variance)



2次モーメント（= 静的特徴量系列の**非線形変換**）

$$\boldsymbol{v}_c = \left[v_{c(1)}, \cdots, v_{c(D)}\right]^{\mathrm{T}} = f_v(\boldsymbol{c})$$

$$v_{c(d)} = \sum_{t=1}^{T}\left(c_t(d) - \sum_{\tau=1}^{T} c_\tau(d)\right)^2$$

データからGVの*p.d.f.*も学習

GVを考慮した時系列データ生成法

$$\bar{\boldsymbol{c}}_q = \arg\max_{\boldsymbol{c}} P(\boldsymbol{c}\,|\,\boldsymbol{q},\boldsymbol{\lambda})P(\Delta\boldsymbol{c}\,|\,\boldsymbol{q},\boldsymbol{\lambda})P(\boldsymbol{v}_c\,|\,\boldsymbol{\lambda}_v)$$

静的特徴量系列 $\boldsymbol{c}$ の関数

静的特徴量系列の*p.d.f.*　動的特徴量系列の*p.d.f.*　系列内変動の*p.d.f.*

※線形変換 $\Delta\boldsymbol{c} = f_\Delta(\boldsymbol{c})$　※非線形変換 $\boldsymbol{v}_c = f_v(\boldsymbol{c})$

HMM音声合成によるサンプル音声

GVなし:　GVあり:

PoE，正則化，特徴空間，非線形写像，・・・

変換：学習処理を無くしたい！

2005年当初・・・，声質変換の障害者補助応用を考え始める・・・

目標とする声質の音声が既に存在しない・・・どうしよう？

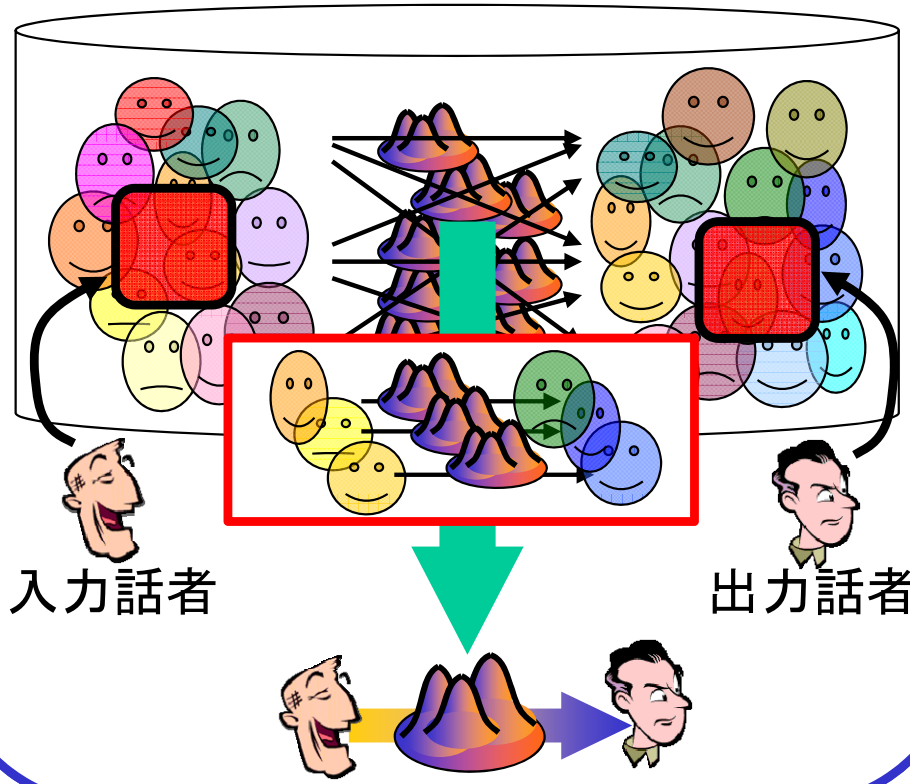
他人の変換モデルで代用できるのでは・・・

多人数集めれば似た声の人はいるのでは・・・

声質コントローラみたいな仕組みを作れないか・・・

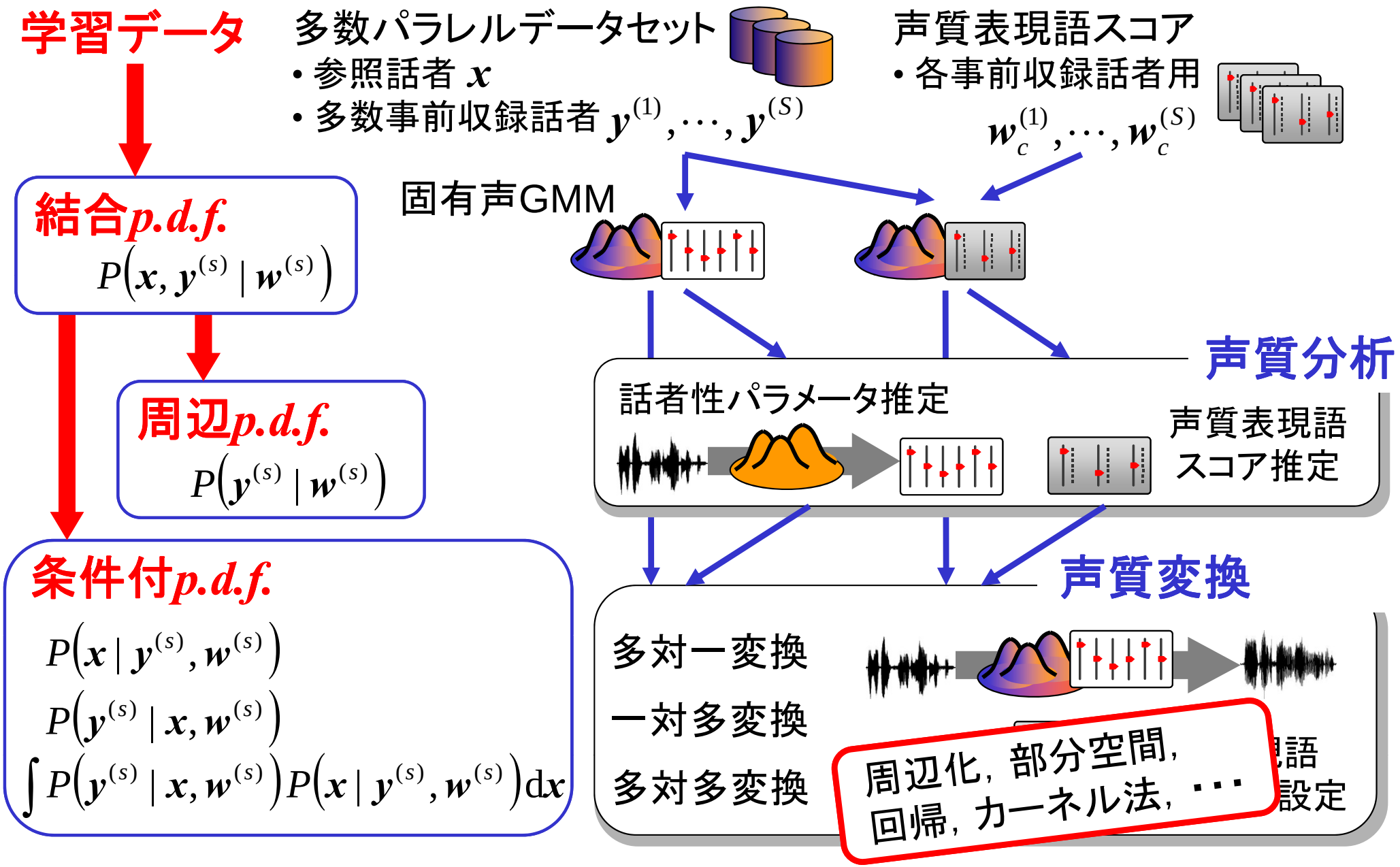
音声合成／認識分野における数多くの研究成果を活用できないか・・・

他人の声を上手く**混ぜる**ことで
別の人の声を創り出そう！



固有声変換

[Toda *et al.*, 2006]



分析：統計処理で誤差を減らしたい！

2006年夏・・・，声質変換の分析エラーが気になり，分析合成も統計的にできるのではないかと思い始める・・・

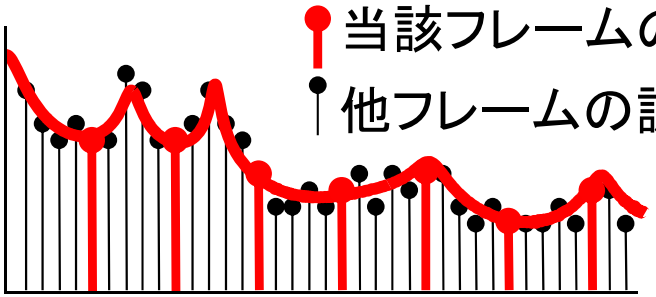
スペクトル分析は見えない箇所を見る問題・・・人は障害物の向こうにある物をどうやってイメージしているんだろう？

別の箇所が見えている
他のサンプルも使えば
推測できるのでは？

他フレームから得られる情報も使おう！

分析フレーム

t-a+i

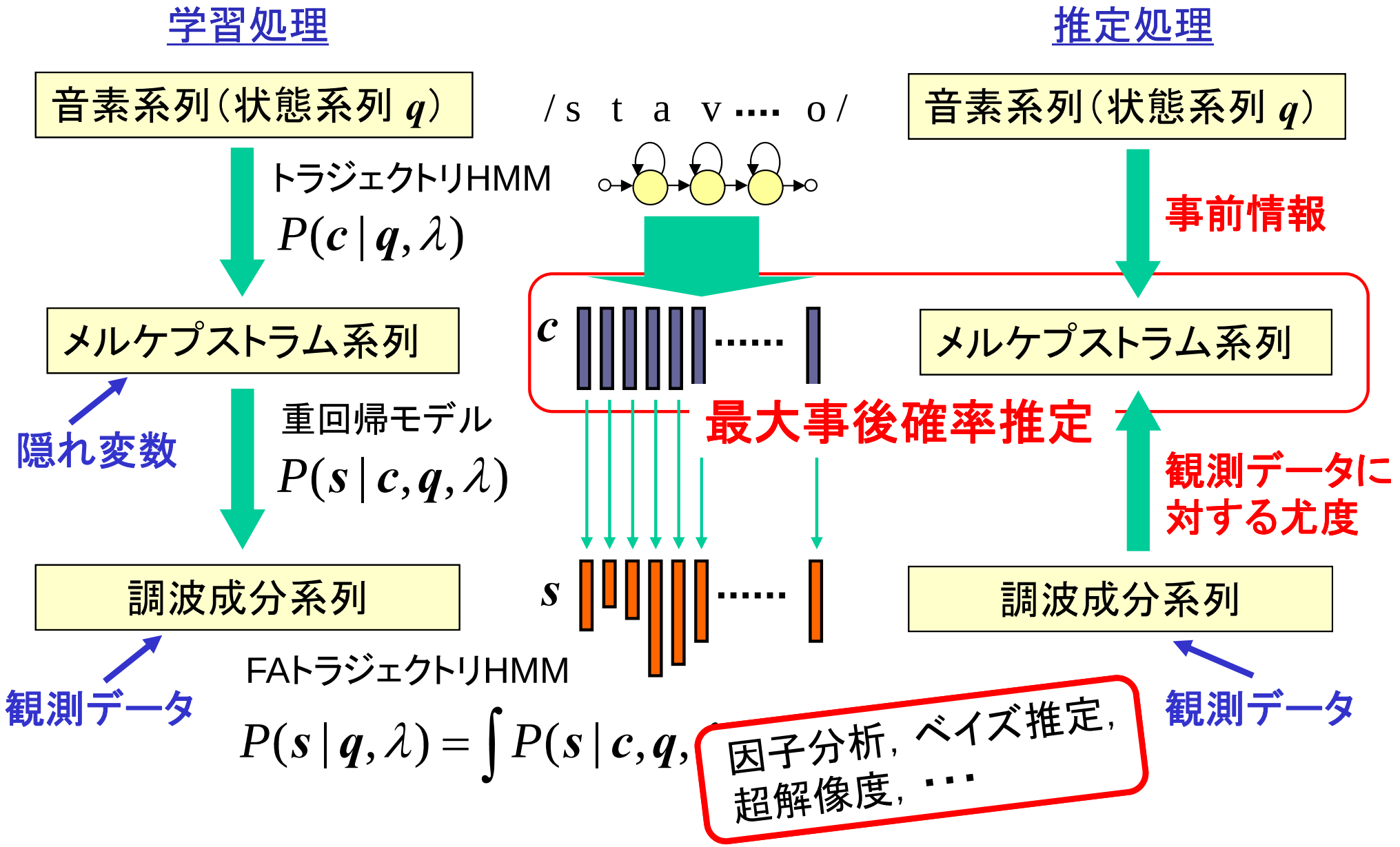


類似した音素環境を
持つ他のフレーム

k-a+i　t-a+i　p-a+i　……

STAVOCO(*STAtistical VOCOder*)のスペクトル分析

[Toda *et al.*, 2007]



機械学習の効果と限界

なぜ機械学習を使うのか？

- **こんなにも良いことがある！**
  - 解くべき問題を**数理的に記述**することで，**高い汎用性と実用性**が得られる！
  - 容易に**技術を共有**することができ，色々な問題へと適用できる！
  - **データドリブン**な枠組みであるため，システムを**自動的に構築**できる！
  - **統計処理**を用いることで，物理的に**不可能なことも可能とする**技術を生み出すことができる！

人の能力　＋　機械学習　＝　世界が変わる！

機械学習を適用する際に重要なことは？

- **適用する問題に対して特化させよう！**
  - 応用先の処理（例えば合成）の特徴，モデル化対象となる信号の特徴，物理的な生成過程などを考慮して，モデルを構築する.
  - （機械学習のみに限らず）従来研究の知見を活用する.

- **解決すべき課題は多い！**
  - 主観的・知覚的な評価との対応を上手くとるには？
  - 如何に対象となるデータを大量に手に入れるか？
    - システム使用時に得られるデータを大量に集める必要あり・・・

今後の展開および実用化に向けて

今後の展開および実用化に向けて

- **歌声・音楽情報処理における非言語／パラ言語情報のモデル化技術をさらに発展させていきたい！**
  - **イメージを具現化できる歌声／音声／楽器音合成・変換技術**
    - 全ての音が合成音や変換音で作られた芸術作品（映画など）も可能に？
  - 人の主観値に基づくデータを自動的に獲得する仕組みの構築
    - 機械学習が苦手なところは人の力を借りる枠組みへ
  - 創造力をかきたてる技術を構築し，これまでには存在しなかった歌声／音声／楽器音データの作成を支援
    - **「人によるデータの創出」**と**「機械学習」**のポジティブスパイラルへ

人と機械学習が協力しあう仕組みが重要！
2次創作文化を大いに活用！

スペシャルセッション
「音楽情報処理と機械学習」

続き：機械学習の効果と限界

歌声情報処理：声質変換モデル

戸田 智基

奈良先端科学技術大学院大学 情報科学研究科

2012年8月10日

なぜ機械学習を使うのか？

- **こんなにも良いことがある！**
  - 解くべき問題を**数理的に記述**することで，**高い汎用性と実用性**が得られる！
  - **技術の共有**が容易であり，色々な問題へと適用できる！
  - **データドリブン**な枠組みであるため，システムを**自動的に構築**できる！
    - まだ発見されていない知見も見えてくるかも・・・
  - **統計処理**を用いることで，物理的に**不可能なことも可能とする**技術を生み出すことができる！
    - 声を真似る／混ぜる／創る／操るという処理を実現できる！

人の能力　＋　機械学習　＝　世界が変わる！

機械学習を適用する際に重要なことは？

- **適用する問題に対して特化させよう！**
  - 応用先の処理（例えば合成）の特徴，モデル化対象となる信号の特徴，物理的な生成過程などを考慮して，モデルを構築する.
    - 何を誤差（ゆらぎ）とみなしてモデル化するのかを考える.
  - （機械学習のみに限らず）従来研究の知見を活用する.
- **解決すべき課題は多い！**
  - **感性**：主観的・知覚的な評価との対応を上手くとるには？
  - **データ**：如何に対象となるデータを大量に手に入れるか？
    - 未知のデータを創出するのは苦手（想像力に乏しい）・・・

✓人と機械学習が協力しあう仕組みが重要！
「人によるデータの創出」と「機械学習」のポジティブスパイラルへ
✓二次創作文化を大いに活用！