信号処理論特論
第8回 (11/22)

情報理工学系研究科　猿渡　洋
首都大学東京・システムデザイン学部
塩田さやか
hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp
sayaka@tmu.ac.jp

講義予定

- 9/27: 第1回　統計的音声音響信号処理概論
- 10/04: 第2回　非負値行列因子分解
- 10/11: 第3回　独立因子分析（ICA, IVA, ILRMA）
- 10/18: 第4回　独立因子分析（続き）
- 10/25: 第5回　音場再現・スパース最適化
- 11/01: 第6回　音声合成・変換１
- 11/15: 第7回　【レポート課題１】
- 11/22: 第8回　話者認識
- 11/29: 休講
- 12/06: 第9回　エンハンスメント・高次統計量解析
- 12/13: 休講
- 12/20: 第10回　音声合成・変換２
- 01/10: 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）

話者認識

本日の主な内容

話者認識

導入：バイオメトリクスの観点から見た話者認識
具体的な想定内容
研究の推移
歴史的な話から最近の傾向まで

どういう前提を持ってモデル化が定義されているのか

話者認識の利用

想定されるシーン

PCが誰かを認識した上で
その人に適したサービスを提供

入力された音声がAさん本人かを確認し
スマートフォンのロックを解除

つまり喋ってる人が誰かを認識する技術

バイオメトリクスという観点からの話者照合

バイオメトリクス（生体認証）

- 本人認証に生体情報を用いた認証技術
  - 誰もが持っている特徴
  - 全員が異なる特徴

バイオメトリクスの実用化例

身体的特徴	センサ	受容性	導入コスト	問題点
指紋	静電容量形 感圧式、光学式	中 （高？）	安	乾燥指、水濡れの影響 特徴量の互換性
顔	CCDカメラ	高	中	化粧、メガネ、照明 加齢、双生児
虹彩	CCDカメラ	中	高	まつ毛の影響、装置
静脈	赤外線CCDカメラ	中	中	装置
音声	マイクロホン	高	安	体調、双生児、経時変化

行動的特徴 ——→ 他にもDNAや署名、掌形など

行動的特徴

行動的特徴の考え方

- 音声：呼気で声道を振動させた音という生成物
- 署名：筆記用具を手に持ち複数の運動形を動かして得られる生成物

特徴

- 採取されることに心理的負担が低い
  - 非接触だとより低い
- 採取時の環境や生体の経時変化の影響あり
  - 永続性および精度に影響

精度向上が目下の目標

話者認識の分類と応用

話者認識（Speaker recognition）

- 話者照合（Speaker verification）
  - 本人か否かを判定
- 話者識別（Speaker identification）
  - 誰の音声かを識別

派生する研究分野

- 音声インデキイング
- ダイアライゼーション

大局的に見れば判定をどうするかの違い

発話内容からの分類

テキスト依存型（text-dependent）

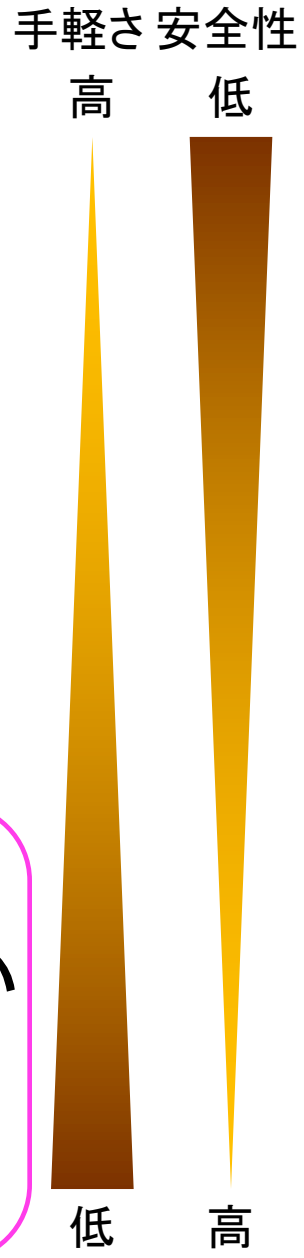
- 登録と照合で同じ発話内容を使用
- 欧米での主流（以前は）

テキスト指定型（text-prompted）

- 発話内容をシステムから提示（指定）
- 依存型と独立型の中間

テキスト独立型（text-independent）

- 登録と照合で発話内容は指定や限定をしない
- 話者認識の世界的なコンペィションでのタスク
- 応用範囲が広いことから主流の研究に



本日の主な内容

話者認識

導入：バイオメトリクスの観点から見た話者認識

具体的な想定内容

⟹ 研究の推移

歴史的な話から最近の傾向まで

どういう前提を持ってモデル化が定義されているのか

話者認識の歴史

1962年

- Bell研のKerstaがスペクトログラムからの話者認識の可能性を発表
- 当初は声紋を研究者が見て判断

1960～1970年代

- DTW(dynamic time warping)に基づくテキスト依存型
  - 静かな環境パスワードを発生する小規模なデータベース
- ベクトル量子化（VQ）に基づくテキスト非依存型
  - パスワードを設定しない手法に注目

1980年代

- 確率モデル（生成モデル）の導入

何をモデル化すればいいのか

音声信号

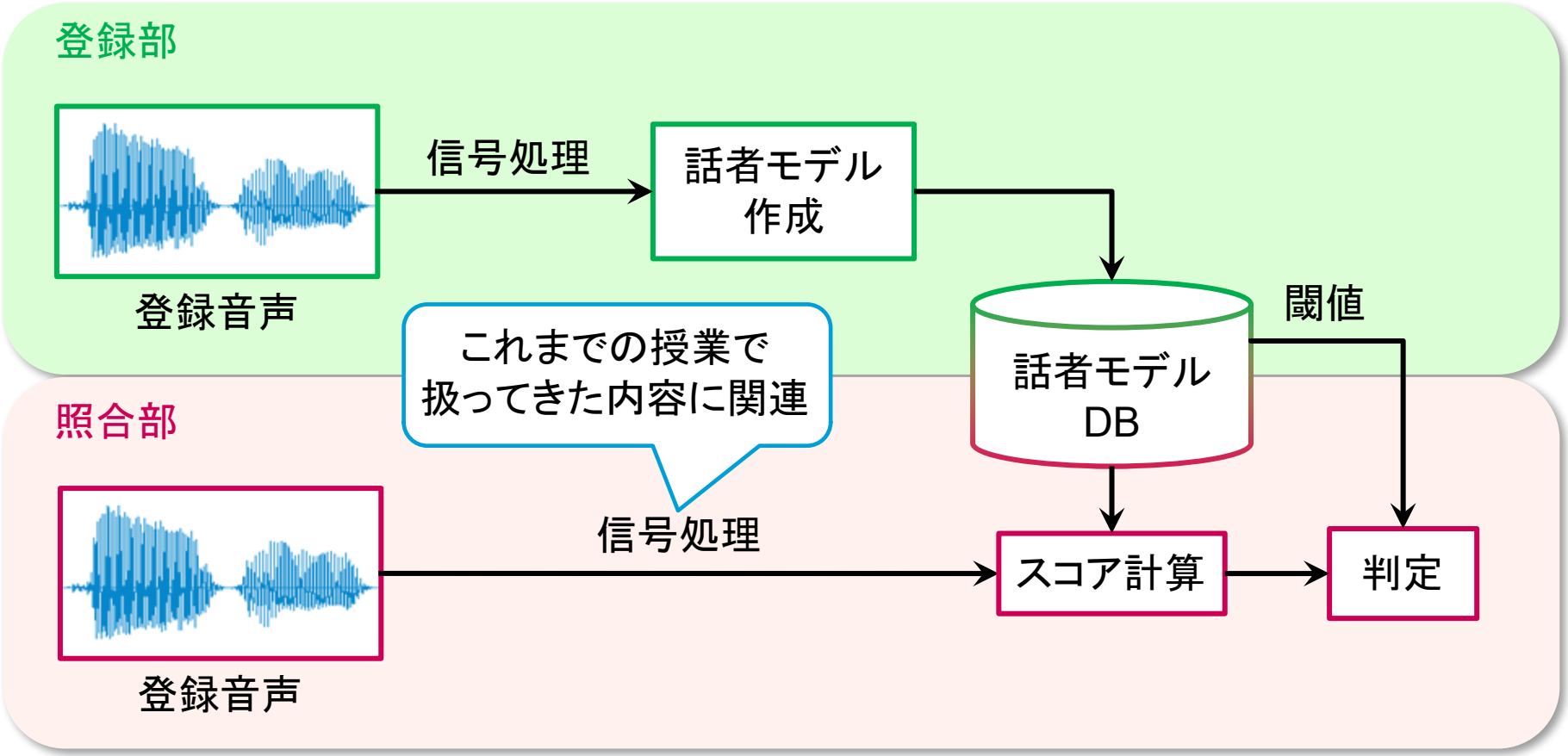
- 言葉の意味内容に関連した音韻性情報
- 誰が話しているかの個人性情報

音声の個人性情報の分類

- 先天的特徴：発声器官の個人差によるもの
  - 音声の共振周波数であるフォルマント周波数の高低
  - 共振の強さを表すフォルマんと帯域幅の大小
  - 声帯の振動周波数を与える平均ピッチ（基本周波数）
  - 音声スペクトル概形の傾斜
- 後天的特徴：発声習慣の個人差によるもの
  - アクセントやなまり
  - ピッチやフォルマント周波数の時間、パターン単語の長さ

なりすましにくい要素

話者照合のフロー図


登録部
信号処理
話者モデル
作成
登録音声
閾値
これまでの授業で
扱ってきた内容に関連
話者モデル
DB
照合部
信号処理
スコア計算
判定
登録音声

発声機構(復習)

音色の付与

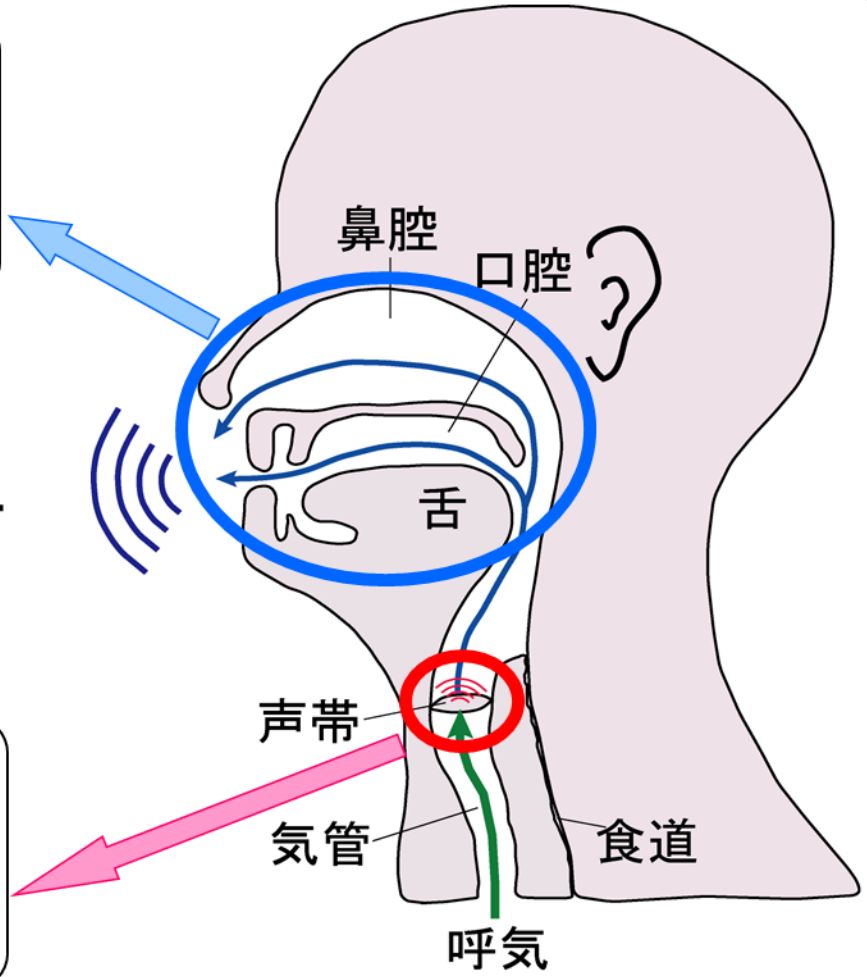
口や舌を動かして，
音色をつける！

声になる！

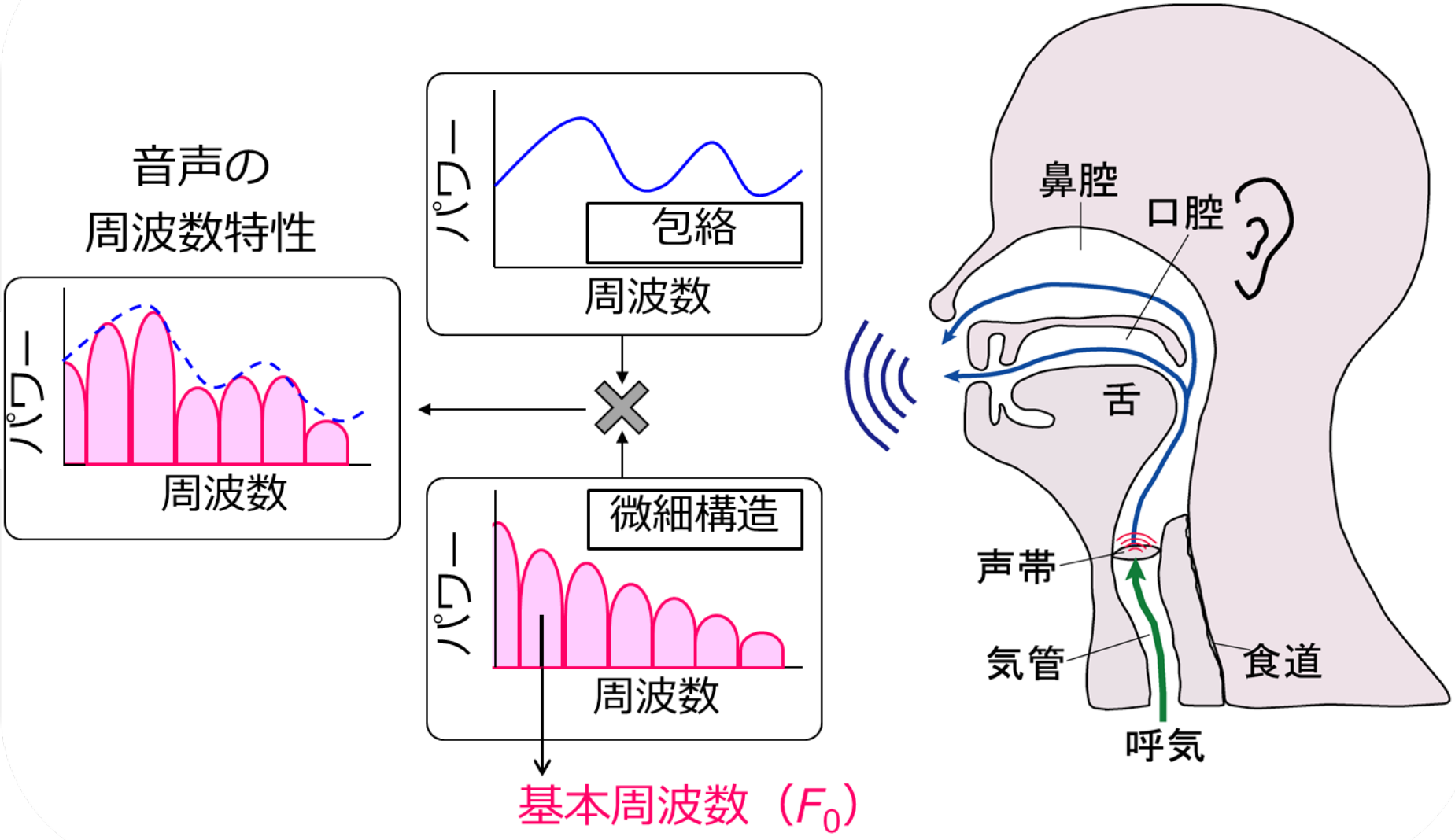
畳み込むと…

音高の生成

声帯を開閉させて，
空気を振動させる！



発声機構(復習)


音声の
周波数特性
パワー
周波数
パワー
包絡
周波数
パワー
微細構造
周波数
基本周波数（F_0）
鼻腔
口腔
舌
声帯
気管
食道
呼気

音声スペクトルの構成

構成要素

- 周波数とともに緩やかに変化する成分[スペクトル包絡]
  ⇒ 発声器官の共振・反共振特性を表す
  （つまり人間の喉・口の形をあらわす特徴量）
- 細かく周期的（有声音；母音などの場合）または非周期的（無声音の場合）に変化する成分
  [スペクトル微細構造]
  ⇒ 音源の周期性
  （つまり声帯の基本周期・声の高低を表す特徴量）

スペクトル分析

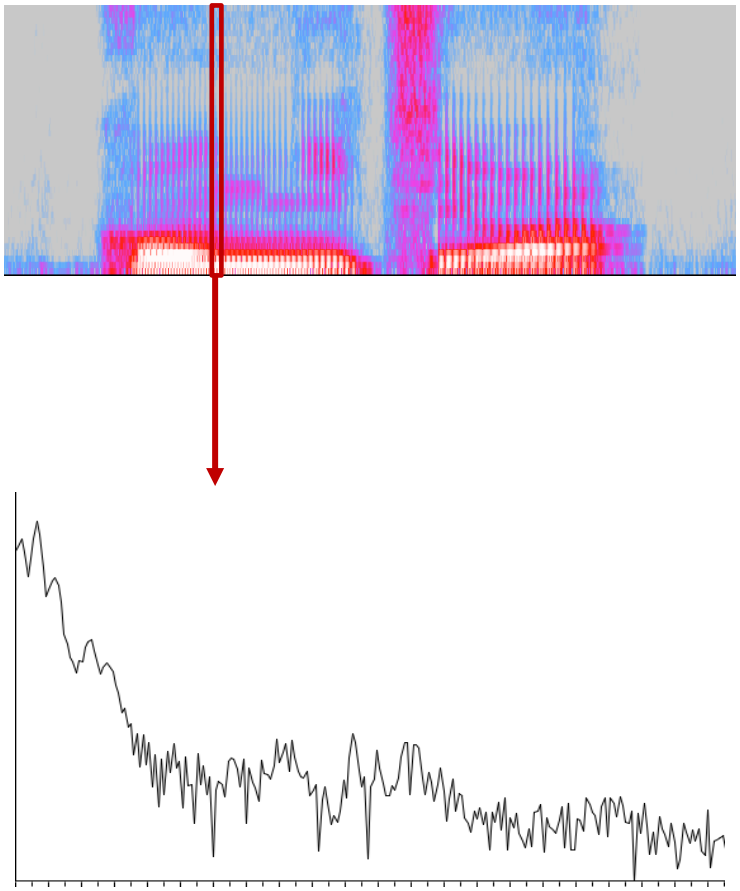
- 音源信号スペクトラム $G(e^{j\omega})$
- 調音フィルタの伝達特性 $H(e^{j\omega})$
- 音声振幅スペクトル

$$|S(e^{j\omega})| = |G(e^{j\omega})| * |H(e^{j\omega})|$$

- 対数化による分離を想定

$$\log|S(e^{j\omega})| = \log|G(e^{j\omega})| + \log|H(e^{j\omega})|$$

  - 現実問題は簡単ではない



- 必要な情報を必要な形で抽出
  - 一例としてMFCCの抽出を紹介

何をモデル化すればいいのか（再掲）

音声信号

- 言葉の意味内容に関連した音韻性情報
- 誰が話しているかの個人性情報

音声の個人性情報の分類

- 先天的特徴：発声器官の個人差によるもの
  - 音声の共振周波数であるフォルマント周波数の高低
  - 共振の強さを表すフォルマんと帯域幅の大小
  - 声帯の振動周波数を与える平均ピッチ（基本周波数）
  - 音声スペクトル概形の傾斜
- 後天的特徴：発声習慣の個人差によるもの
  - アクセントやなまり
  - ピッチやフォルマント周波数の時間、パターン単語の長さ

なりすましにくい要素

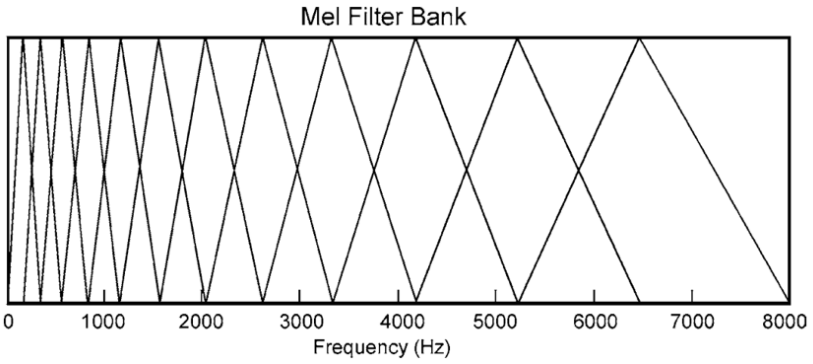
メル周波数ケプストラム係数(MFCC)(1/2)

MFCC

- 音声認識や音声合成で一番標準的に使われる特徴量
- ケプストラムパラメータを計算する方法の一つ
- 声道特性を表す特徴量

メル周波数

- 人間の音の高さの知覚特性から得られた尺度
- 低周波ほど間隔が狭く、高周波ほど広い

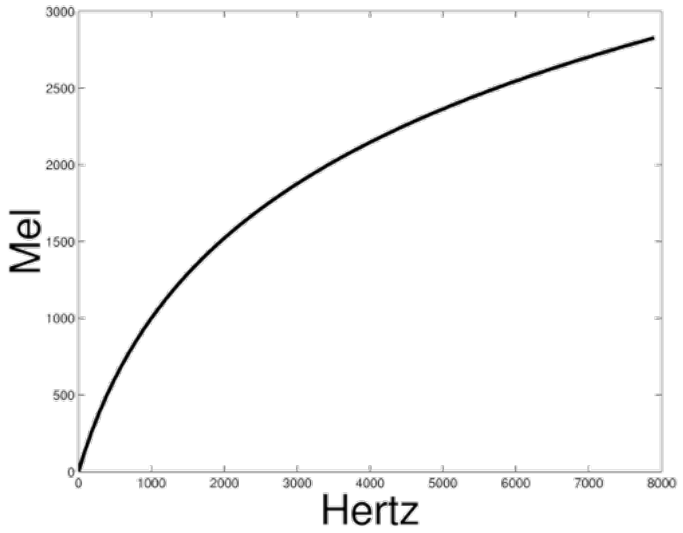


メル尺度の計算

メル尺度と周波数の関係式

$$Mel(f) = (\frac{1000}{\log 2}) \log(1 + \frac{f}{1000})$$

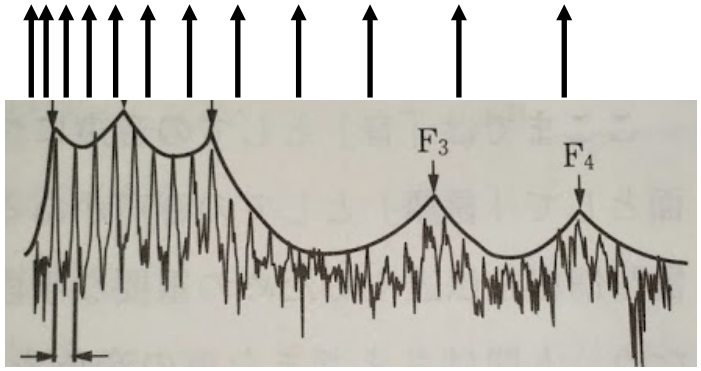
- 周波数とは対数の関係
- 1000Hz,音圧レベル40 dB の純音を基準の音1000mel
- 基準音より2倍の高さあるいは1/2の高さに知覚される音をマグニチュード測定法などで測定し，それぞれ2000mel，500melと規定



メル周波数ケプストラム係数（MFCC）(2/2)

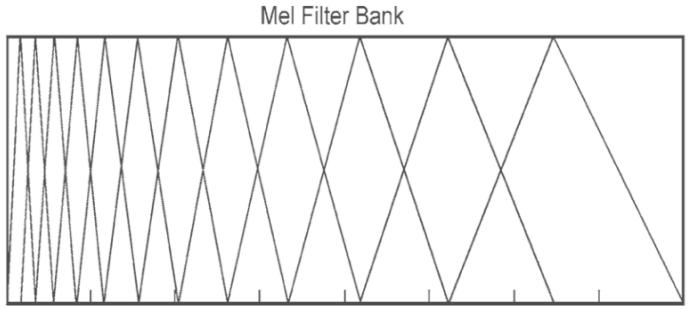
1. フレーム処理→窓関数をかけてFFTから振幅スペクトルを求めるところまではだいたい同じ
2. 振幅スペクトルにメルフィルタバンクをかけて伸縮
3. 伸縮した数字列を信号とみなしDCT
   低次元を細かく、高次元を大雑把にとる
4. 得られたケプストラムの低次成分がMFCC

フィルタバンクの数の値が出てくる



振幅スペクトル

メル周波数軸上で等間隔になる

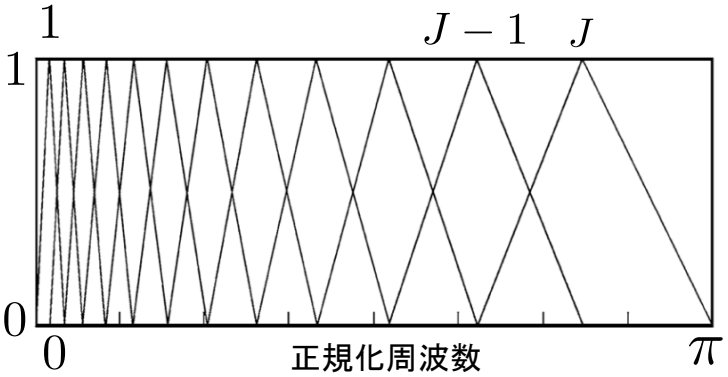


MFCCを式から見ると(1/2)

サンプル数:N

音声サンプルのDFT係数:

$$S(\omega_n), 0 \leq n < N$$



j 番目のフィルタでのスペクトル推定量:

$$q(j) = \log[\sum_{n=l_j}^{u_j} \|S(\omega_n)\|^2 w_j(n)]$$

$w_j(n)$ は j 番目のフィルタの ω_n 成分に対する重み付け係数

$$w_j(n) = \begin{cases} \frac{f_s - f_{c_{j-1}}}{f_{c_j} - f_{c_{j-1}}} & (l_j \leq n \leq c_j) \\ \frac{f_{c_{j+1}} - f_s}{f_{c_{j+1}} - f_{c_j}} & (c_j \leq n \leq u_j) \\ 0 & (n < l_j \ or \ u_j < n) \end{cases}$$

フィルタの通過周波数帯域の下限 l_j 、中心 c_j 、上限 u_j

パスバンドの下限と上限

$f_{c_{j-1}}$　　$f_{c_{j+1}}$

MFCCを式から見ると(2/2)

フィルタバンクにより得られた$q(j)$をDCTすることでMFCCが得られる

$$C(i) = \sum_{j=0}^{J} q(j) \cos[n(j - \frac{1}{2})\frac{\pi}{J}], i = 1, 2, \ldots, M$$

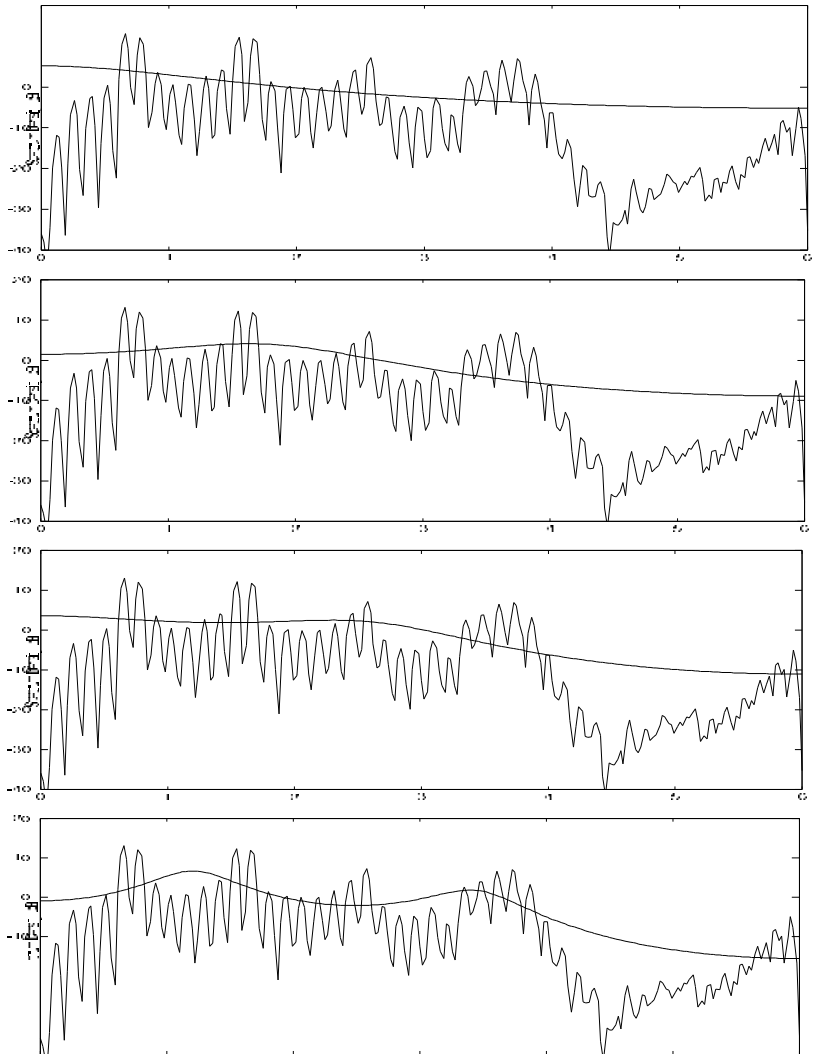
- MがMFCCの次数に

利点

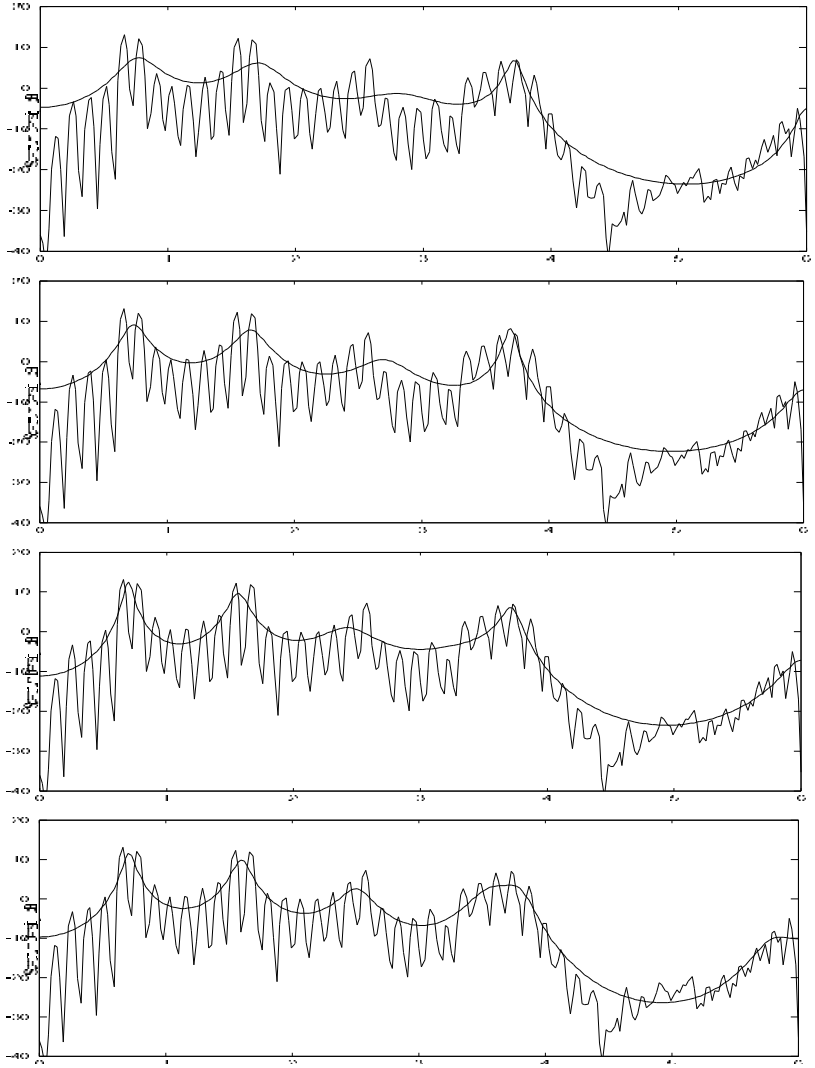
- メル尺度を用いることで認識に必要な情報を少ない次元数で表現可能
- 聴覚特性を取り入れることで雑音に対する頑健性も向上

最尤推定による音声スペクトル推定の例

1～4次

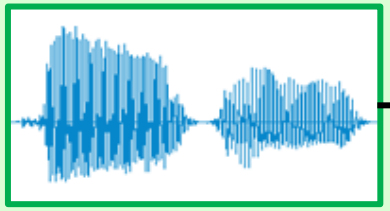


9, 10, 12, 14次



音声データの表現

登録部



登録音声

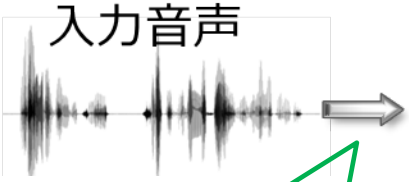
信号処理

話者モデル
作成

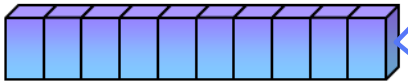
特徴ベクトル系列

$$X = \{x_1, x_2, ..., x_T\}$$

入力音声

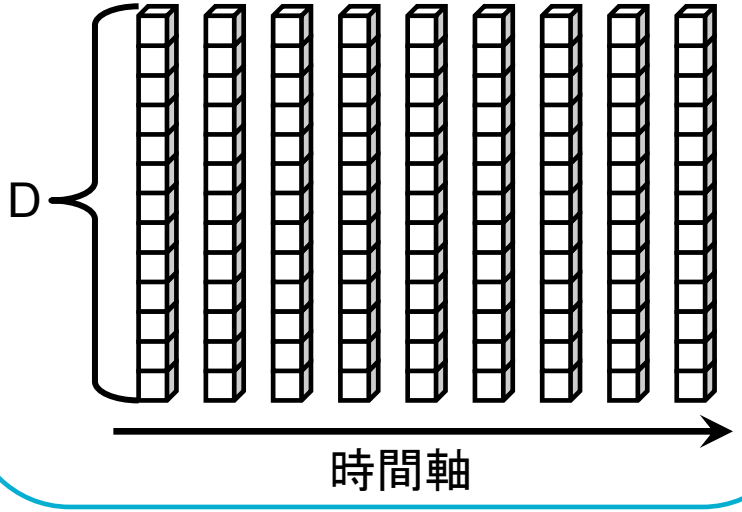


入力特徴量

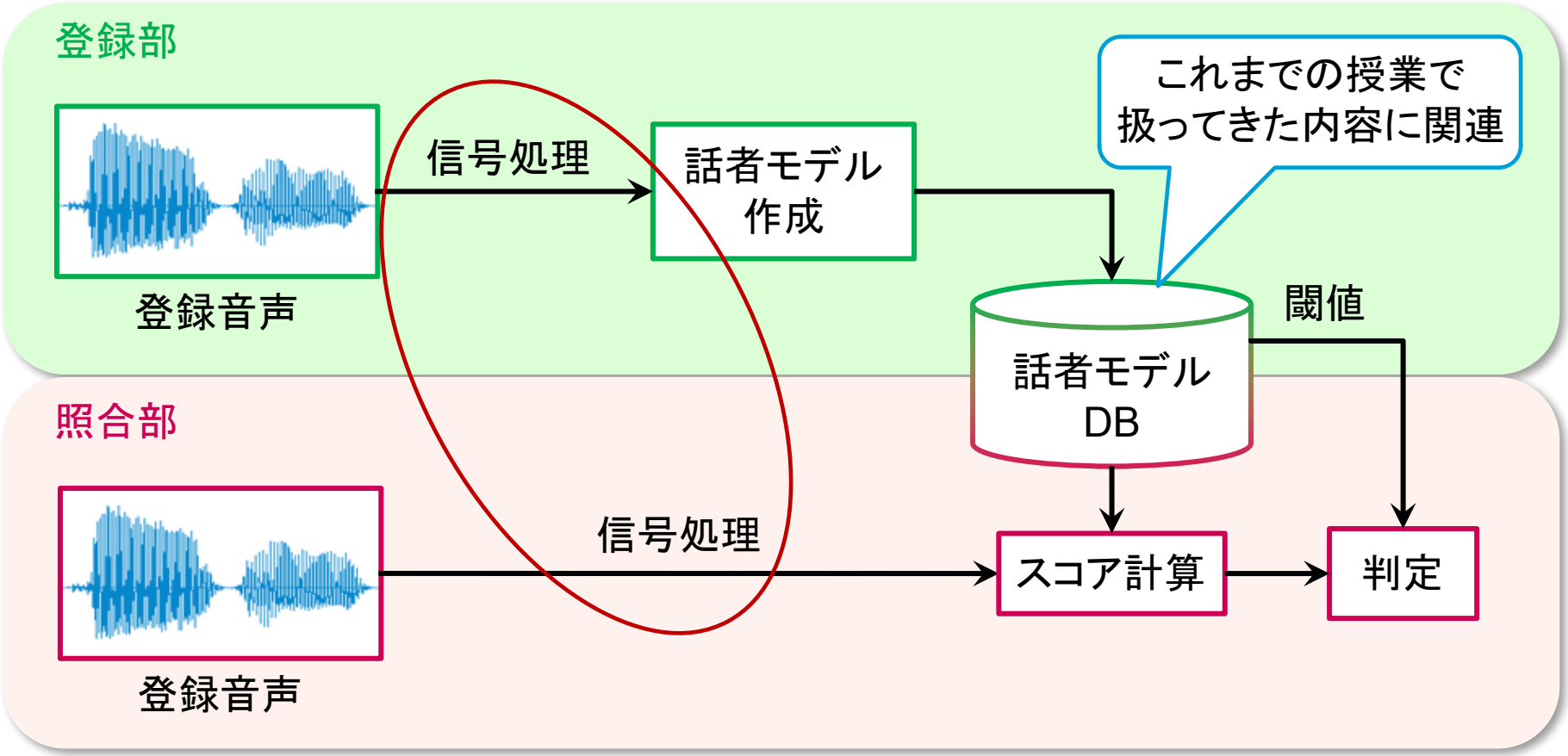


フレーム処理
窓関数の掛け算
特徴量抽出

例えばD次元のMFCCなら…



話者照合のフロー図（再掲）



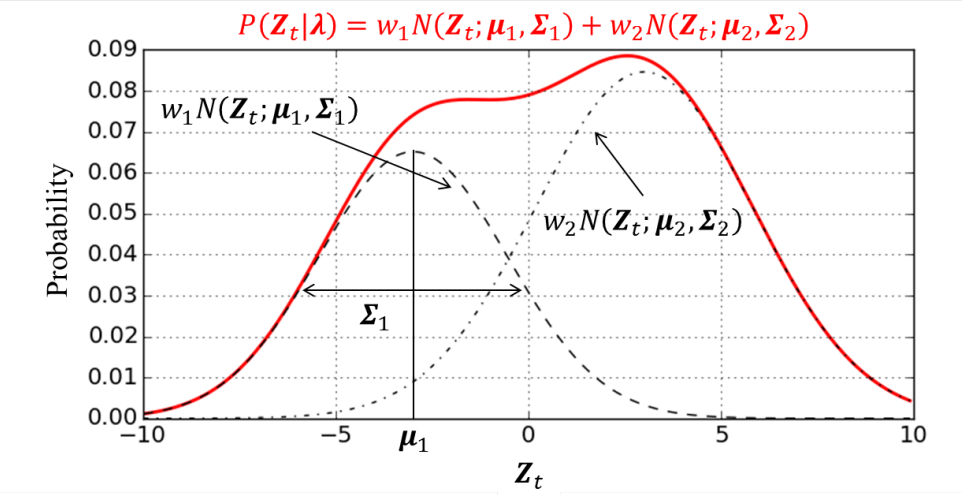
使用されるモデルの推移を紹介

GMMに基づくアプローチ

Reynoldsらにより提案（1990年代半ば）

- 特徴ベクトル系列の生成過程がGMMに従う
- 仮定：特徴ベクトルの分布が時刻や発話内容によらず話者のみに依存（思い切った仮定）
- 話者性をよく表現できるとしてテキスト独立型の標準に
  - 1つのGMMが1人の話者を表現
  - 当初は32混合程度のGMMを使用

$$P(Z_t|\lambda) = \sum_{k=1}^{K} w_k N(Z_t; \mu_k, \Sigma_k)$$



GMM-UBM（1/3）

UBM(Universal Background Model)

- 不特定話者の平均的な音声モデル

特定話者モデルの学習

- UBMを基点に最大事後確率(Maximum A posteriori Probability; MAP)推定で学習
- MAP適応
  - 入力データ X が得られたとき、パラメータ λ を確率変数として扱い、その事後確率を最大化するように推定

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} p(\lambda|X)$$
$$= \arg\max_{\lambda} p(X|\lambda)p(\lambda)$$

GMM-UBM（2/3）

MAP適応によるGMMの平均ベクトル更新

- 観測データ　が　番目のガウス分布から生成されたとする確率

$$p(k|x_t) = \frac{w_k N(x_t; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j N(x_t; \mu_j, \Sigma_j)}$$

- 観測データ　について　番目のガウス分布かあら生成されるデータの確率的サンプル数と確率的平均ベクトル（十分統計量）

$$n_k = \sum_{t=1}^{T} p(k|x_t)$$

$$e_k(x) = \frac{1}{n_k} \sum_{t=1}^{T} p(k|x_t) x_t$$

$$e_k(x^2) = \frac{1}{n_k} \sum_{t=1}^{T} p(k|x_t) x_t^2$$

- 平均、分散、重みの更新式

$$\hat{w}_k = [\frac{\alpha_k n_k}{T} + (1 - \alpha_i) w_k]\gamma$$

$$\hat{\mu}_k = \beta_k e_k + (1 - \beta_k)\mu_k$$

$$\hat{\Sigma}_k^2 = \eta_k e_k(x^2) + (1 - \eta_k)(\Sigma_k^2 + \mu_k^2) - \hat{\mu}_k^2$$

$\alpha_i, \beta_k, \eta_k$ は更新量のバランスパラメータ

GMM-UBM（3/3）

照合時

- 入力データXに対して対数尤度比を使用

$$\log p(X|\lambda_{UBM}) - \log p(X|\lambda_s)$$

- 設定した閾値により受理か棄却かを決定

特徴

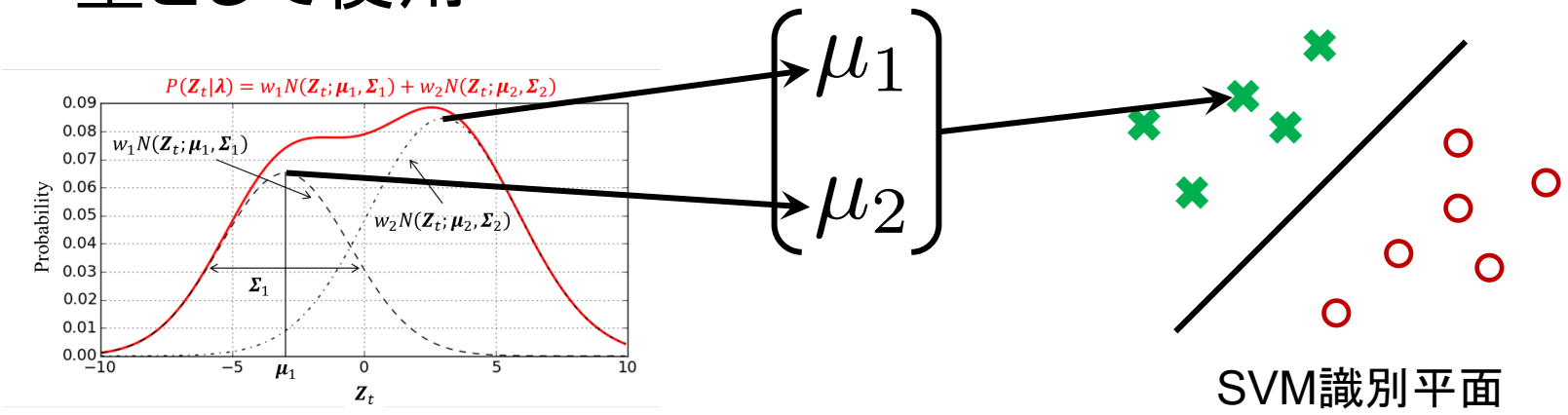
- UBMを基点とすることで話者性の違いが明確に
- UBMの安定した分散を使用可能
- 過学習の問題を回避

現在でもベースラインとして使用

サポートベクトルマシンに基づくアプローチ

Campbellらにより提案（2000年代前半）

- 識別指向な学習機械器SVMを話者照合に適用
- GMMを構成する各ガウス分布の平均ベクトルを全て連結した高次元ベクトル（GMMスーパーベクトル）を特徴量として使用



- 識別平面の決め方をSVMに
- 2クラス分類を得意とするSVMとは親和性が高
- ベクトルの長さはD次元x混合数（512）＝高次元

因子分析によるモデル化

接合因子分析（JFA）（2000年代後半）

- 回線（チャネル）の違いによる音声の変動
- 同一話者における時期差などによる変動

} を緩和

GMMスーパーベクトルの表現を新たに定義

話者部分空間を定義する固有声行列　　チャネル部分を定義する固有チャネル行列

$$M_u = m + Vy + Ux_u + \epsilon_u$$

残差成分

UBMから得られる話者およびチャネルに非依存なGMMスーパーベクトル

- 話者ベクトルの次元数を大幅に削減
- 話者照合では話者性を表すx_uにて識別

x_u チャネル固有因子
y 話者固有因子

チャネル因子にも話者情報が含まれている

i-vectorに基づく話者照合

JFAを簡略かした手法

- JFAの話者とチャネルの因子を区別せずに因子分析を行い、後段で線形判別分析（Linear Discriminant Analysis; LDA）などで話者性を抽出
  - 話者とチャネルの同時モデル化を諦めた

GMMスーパーベクトルの定義

$$M = m + Tw$$ i-vector

- 識別は登録と照合のi-vectorのベクトル同士のcos類似度で簡単に比較可能かつ高精度

$$\cos(w_{target}, w_{test}) = \frac{w_{target} \cdot w_{test}}{\|w_{target}\| \|w_{test}\|}$$

Probabilistic Linear Discriminant Analysis (PLDA)

i-vector空間において話者間変動とチャネル変動をモデル化

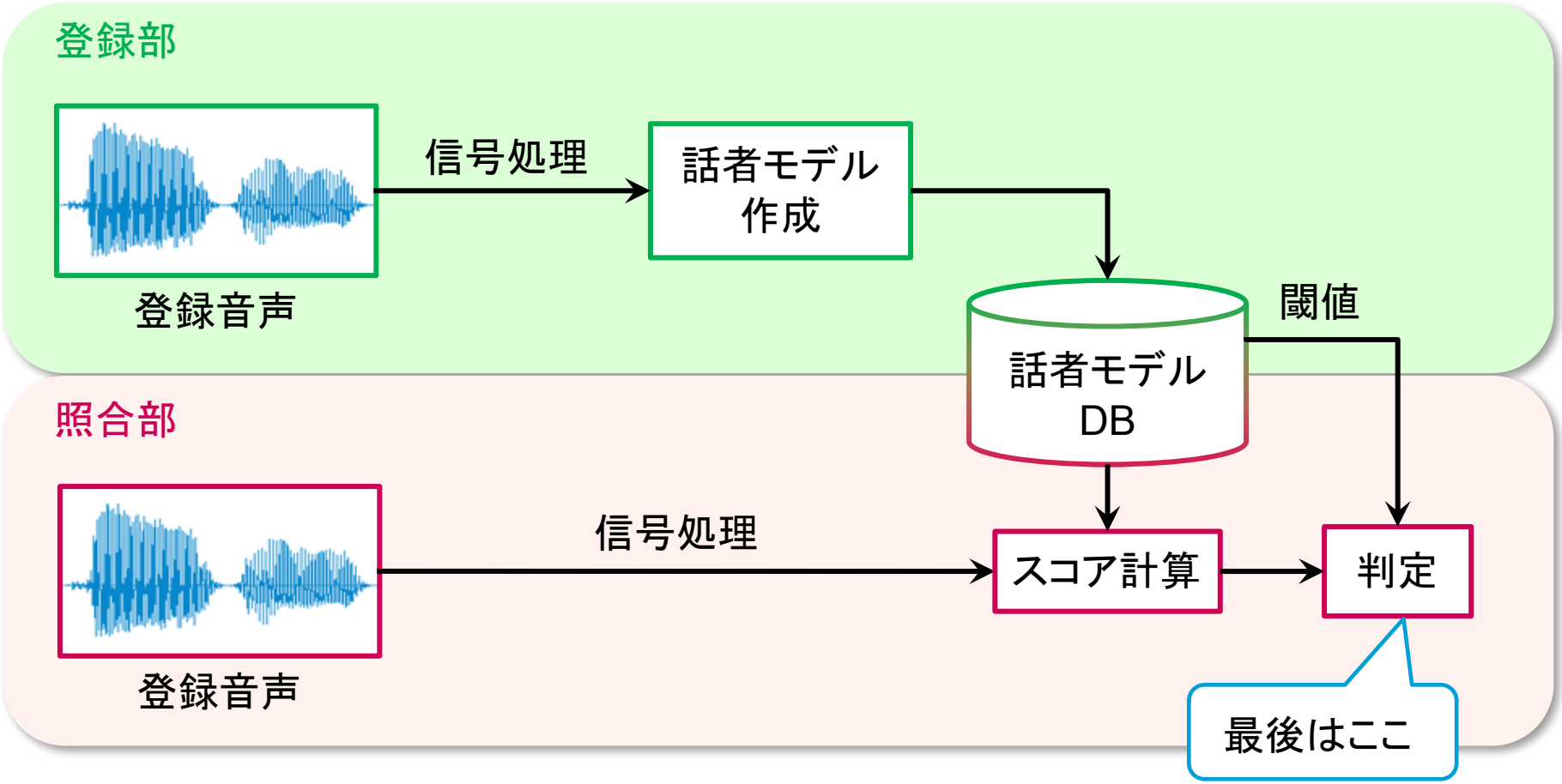
- i-vectorを抽出したあとにi-vectorを確率的生成モデルの観測とみなしてモデル化

$$w_u = \bar{w} + \Phi\beta + \Gamma\alpha_u + \epsilon_u$$

$\bar{w}$ ：i-vector空間のオフセット
Φ ：話者部分空間を張る基底行列
Γ ：チャネル部分空間を張る基底行列
β ：話者因子
α_u ：チャネル因子
ϵ_u ：残差成分

- i-vectorの抽出仮定を無視
- モデル化のために必要な仮定に疑問
- 様々な研究機関が工夫中

話者照合のフロー図（再掲）



精度計測について

性能評価方法

- どのようなシステムが適切か
  - もちろん本人を受理し、他人を拒否してくれるシステム
    - 本人を拒否する確率: False rejection rate (FRR) [Type I error]
    - 他人を受け入れる確率: False acceptance rate (FAR) [Type II error]

$$\mathrm{FRR} = \frac{\text{本人なのに本人でないと拒否された回数}}{\text{本人による認証回数}}$$

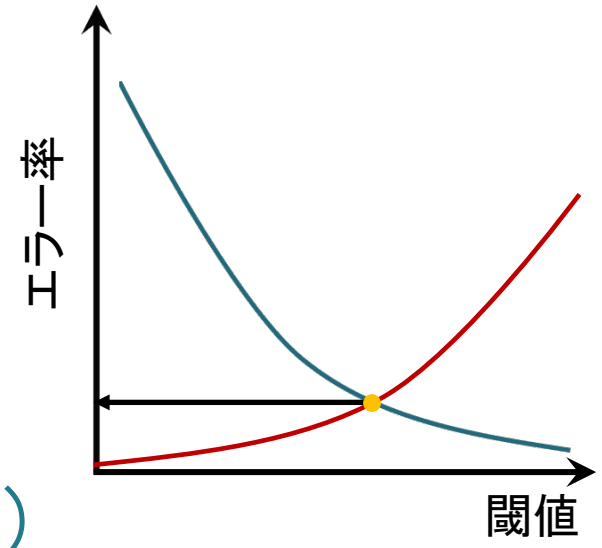
$$\mathrm{FAR} = \frac{\text{他人にもかかわらず本人であると受け入れられた回数}}{\text{他人による認証回数}}$$

  - 閾値の設定の仕方によって大きく変化

EERとROC

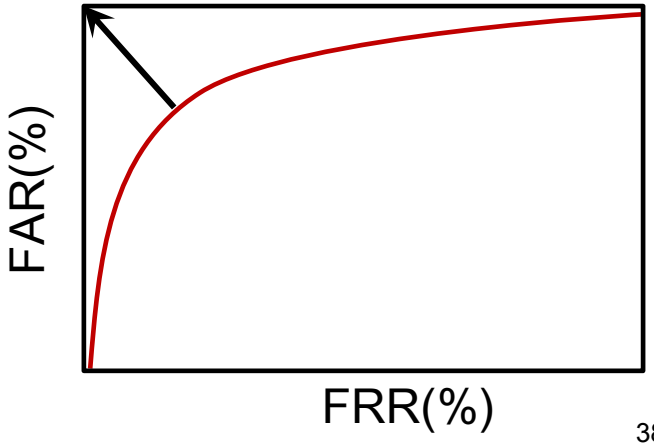
等価エラー率（Equal Error Rate）

- FARとFRRが等しくなる点
- FARとFRRはトレードオフの関係
- どちらも低くなることが理想



Receiver Operating Curve（ROC）

- 登録データとテストデータの類似度と閾値で判定
- 閾値をパラメータとしてFARとFRRをプロット
- 要求するFARとFRR,その時の閾値（信頼区間）が容易に把握可能

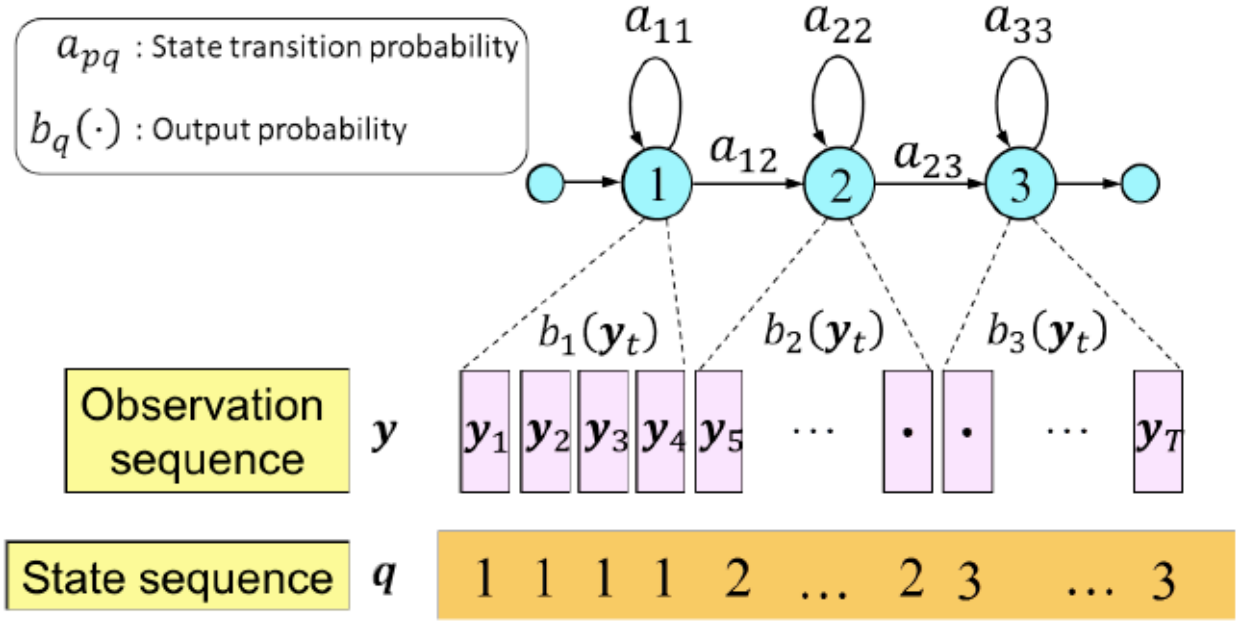


テキスト依存型の話者照合

認証の前提条件とモデル

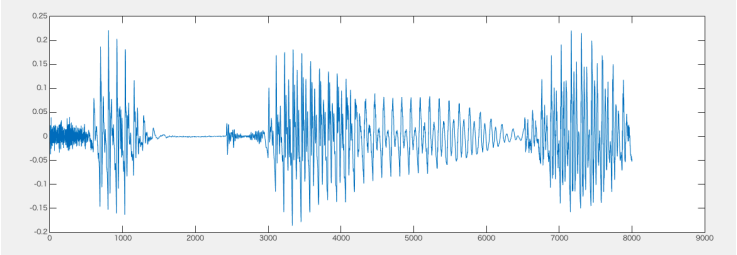
テキスト依存型の場合

- 発声内容が決まっている
- スペクトル情報と時系列情報両方を利用可能
- 時系列情報のモデル化＝HMM
- 時間の不可逆生からleft-to-right

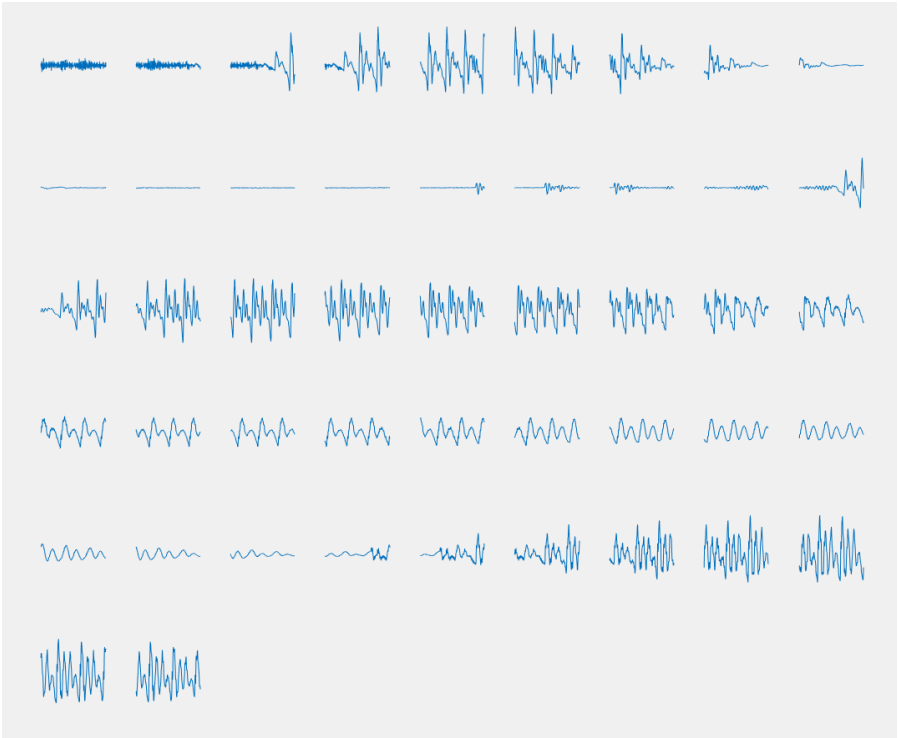


音声フレームに窓関数をかけた結果

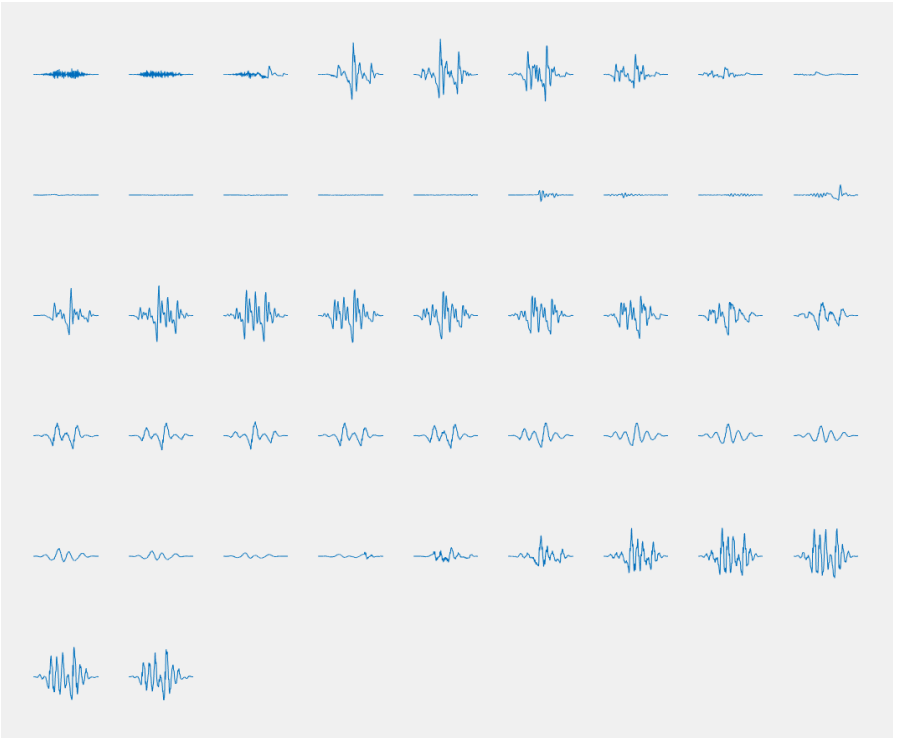
元音声



窓かけ前



窓かけ後



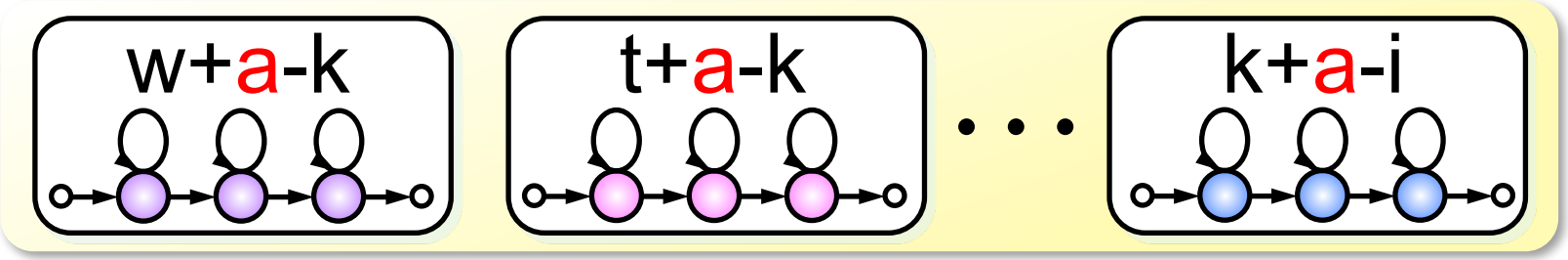
音素

音素とは

- 発声の最小単位
  - その音が変わった時に言葉の持つ意味が変わる音
  - 例: 若い（ｗａｋａｉ）と高い（ｔａｋａｉ）

認識・合成における音素の扱い

- 1つのaという音についても細かく見ることが可能



- 合成ではもっと細かいところまで分けて考える

テキスト依存型の照合

学習

- 登録話者sのHMMモデル

$$\lambda^s = (A, B, \pi)$$

A :状態遷移確率
π :初期状態確率
B : x が出力される出力確率

- left-to-right型では $\pi_0 = 1$
- 出力確率はGMMとして表現

$$b_j(x) = \sum_{k=1}^{K_j} w_{jk} \mathcal{N}(x; \mu_{jk}, \Sigma_{jk})$$

- K_j は状態 j の混合数、w_{jk} は状態 j におけるインデックス k のときの重み、$\mathcal{N}(x; \mu_{jk}, \Sigma_{jk})$ は平均 μ_{jk}、共分散行列 Σ_{jk} をもつガウス分布

おすすめ文献

日本音響学会誌69巻7号（2013） 小特集 「話者認識に関する研究の動向」にあたって

日本音響学会誌70巻6号（2014） 解説 「i-vectorを用いた話者認識」、小川哲司、塩田さやか

バイオメトリクス教科書 -原理からプログラミングまで- 映像情報メディア学会編

音響学入門 日本音響学会編

Speaker Classification I & II, Springer (Christian Muller (Ed.))