話者認識における機械学習

- 話者認識を含めパターン認識の課題全般にわたり，機械学習の有効性が確認されている
  - 機械学習: データから帰納的に知見を抽出する方法・仕組み
    - 大まかに4ステップ
      - データから特徴量を抽出する
      - **特徴量をモデル化する**
      - モデルのパラメータを推定するための評価基準
      - パラメータの最適化
- モデル化に焦点を当てる
  - 生成モデル
  - 識別モデル
  - 因子分析モデル

生成モデル

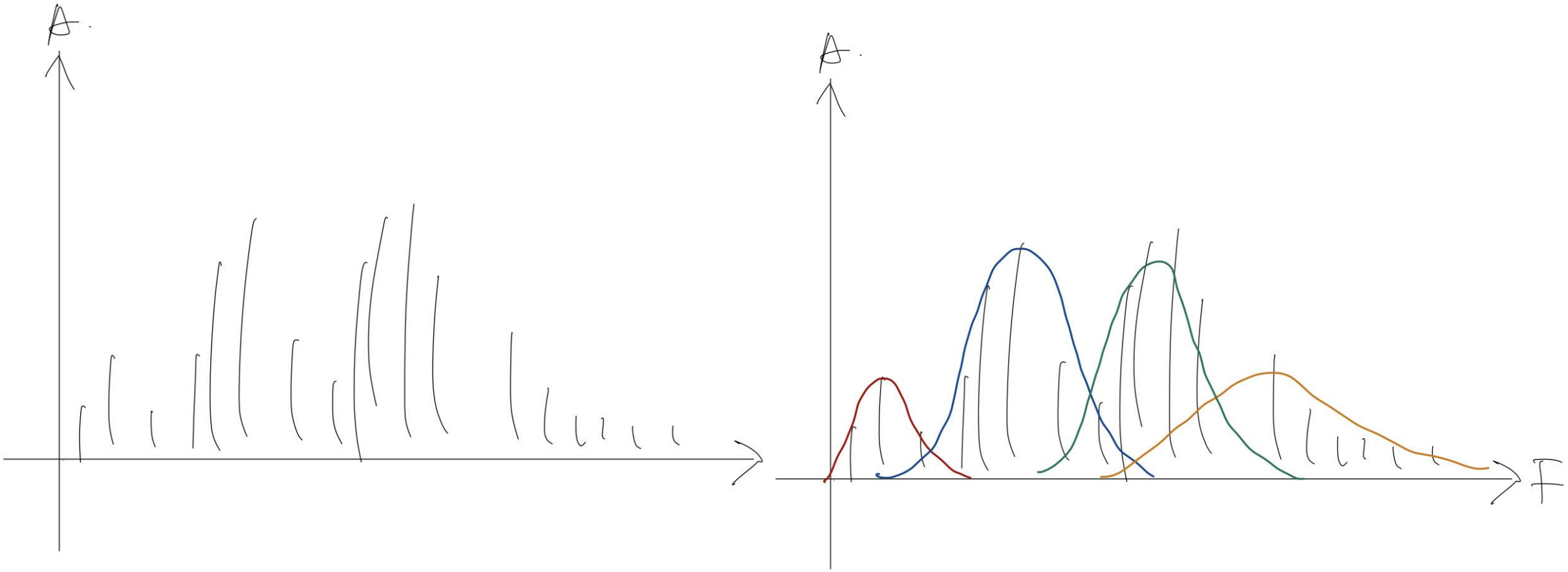
- 各話者のとき超量を生成する分布のモデル
  - 特徴量ベクトルがある話者を表す確率分布から生成されるという仮定
    - 仮に特徴量を振幅スペクトルとしたとき，どの周波数成分が多いがち・少ないがち…等は話者に依存する
  - **GMM-UBM**
  - **GMMスーパーベクトル**
- データから得られる特徴量ベクトル達を用いて，分布のパラメータを決定
- 尤度(対数尤度)で評価する

GMM-UBM

- Gaussian mixture model – Universal background model
  - GMM: 混合ガウスモデル
    - データが複数の正規分布から生成されたと考える手法
- 一般的な音声らしいモデル(UBM)を当該話者の分布に適合させることで，モデルを構築
  - 不特定話者データを用いて一般的な分布のモデルを構築しておく
  - 当該話者モデルの構築の際は，当該話者のデータで分布のパラメータを調整
- 詐称者モデルと対象話者モデルの尤度比で照合
  - UBMを詐称者モデルとする
- モデル調整に用いる(登録)データと評価するデータの音素に偏りがあると精度が下がる
  - 登録データと評価データを大きくしないといけない
    - 音素や音節ごとに話者UBMを構築する
    - 音素分類木を用いる

GMM-UBM*

- データが複数の正規分布から生成されたと考える手法



GMMスーパーベクトル

- GMMの平均ベクトルを結合して得られるベクトル
  - 二次元の分布で，構成分布の平均ベクトルが(1,2),(3,4)のとき，GMMスーパーベクトルは((1,2)',(3,4)')
- 近年はこれを特徴量として用いるのが主流
- 特徴量次元より高次元
- 時系列データをベクトル空間上の1点で表せる
  - 一発話のデータでUBMを適応させて得たGMMの平均ベクトルから，その発話データのGMMスーパーベクトルが得られる
- SVM等の識別モデルの入力として用いられる
- 因子分析アプローチに繋がる

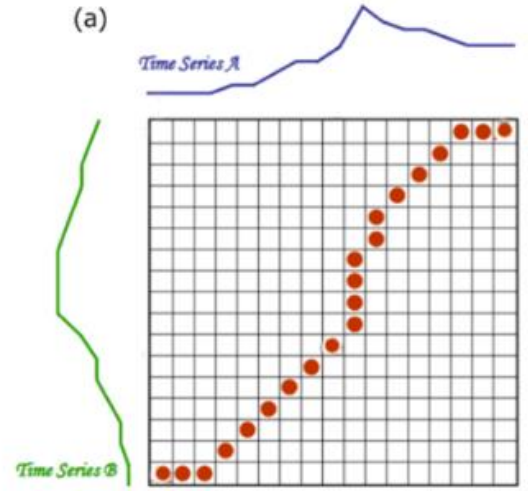
識別モデル

- カーネルマシンを用いたアプローチ
  - 特徴量空間を高次元空間に写像する
    - 写像の仕方はカーネル関数による
- どのようなカーネル関数や特徴量を用いるかが，主な検討事項
  - カーネル関数
    - 系列(sequence)カーネル
    - **GMMスーパーベクトルカーネル**
    - 共分散カーネル
    - **global alignment(GA)カーネル**
  - 特徴量
    - 分析フレーム列
    - GMMスーパーベクトル
    - 対数パワースペクトル
- 適切にカーネル関数やそのパラメータを選択しないと性能がでない

GMMスーパーベクトルカーネル

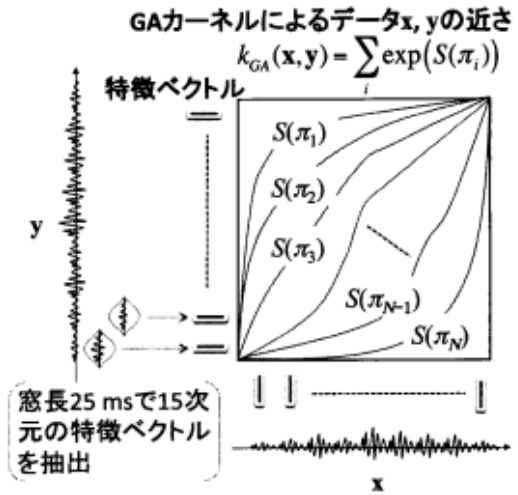
- GMMスーパーベクトル特徴量にあわせて考案された
- 特徴量がD次元，GMMのコンポーネント数がNのとき，GMMスーパーベクトルはD*N次元
  - 高次元空間に写像している

GAカーネル



- Global Alignmentカーネル
- ベクトル時系列を引数に取る
  - 2つの音声それぞれに窓関数を適用し，特徴ベクトル時系列を得る
- 2つの時系列間のいろいろな非線形伸縮軸上での距離を求め，その距離を用いてカーネルの値を計算

$$1 = \pi_1(1) \leq \ldots \leq \pi_1(|\pi|) = n,$$
$$1 = \pi_2(1) \leq \ldots \leq \pi_2(|\pi|) = m,$$
$$S(\pi) = \sum_{i=1}^{|\pi|} \varphi\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right)$$



図–1　GA カーネルの計算

$$k_{\mathrm{GA}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\pi \in \mathcal{A}(\boldsymbol{x}, \boldsymbol{y})} \exp(S(\pi)) = \sum_{\pi \in \mathcal{A}(\boldsymbol{x}, \boldsymbol{y})} \exp\left(\sum_{i=1}^{|\pi|} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})\right) = \sum_{\pi \in \mathcal{A}(\boldsymbol{x}, \boldsymbol{y})} \prod_{i=1}^{|\pi|} k\left(\varphi(x_{\pi_1(i)}, y_{\pi_2(i)})\right) \quad (6)$$

因子分析モデル

- スーパーベクトル空間上で表された発話データについて，話者以外の要因を除去しようとするアプローチ
  - 録音環境等の影響を除去したい
  - 一大トレンド
  - **JFA**
  - **i-vctor**
  - **PLDA**

JFA

- Joint Factor Analysis
- 話者の声とチャネルの影響を明示的にモデル化する
- 話者内の音響特徴の変動(録音環境の違い，セッションによる変化)に強い
- GMMスーパーベクトルが，話者を表現するベクトル，チャネルを表現するベクトルに分解できることを仮定している

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{D}\boldsymbol{z}$$

- 大規模なラベル付き学習データでV,U,Dを推定
- 与えられた発話データでy,x,zを推定
- チャネル変動成分補正後の話者モデルに対して，入力発話の特徴ベクトルの尤度を計算し，評価する

i-vector

- 因子分析では発話データを話者・チャネル依存の全変動空間に写像し，チャネル変動については別に除去するアプローチ
  - JFAでは話者以外の要因をモデル化して除去することの限界がある
    - DehakがJFAによって得られたチャネル因子にも話者情報が含まれていることを示した

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}$$

  - w: 発話に対するi-vector
    - GMMスーパーベクトル空間における一般的な話者(UBM)からの差として話者を表現したもの
    - 話者とチャネル依存の部分空間上のベクトル
- チャネル変動補正の正確さが全体の性能の鍵
  - PLDAによるアプローチもチャネル変動補正の1方式といえる
- 入力データに対して得たi-vectorと，話者モデルとして登録したi-vectorのコサイン類似度によって照合する

PLDA

- Probabilistic Liner Discriminant Analysis
  - i-vector空間において直接的に話者変動やチャネル変動をモデル化する試み

$$\boldsymbol{w} = \bar{\boldsymbol{w}} + \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

  - 仮定には疑問が残る部分もある
    - 話者成分とチャネル成分は統計的に独立である
    - 話者因子とチャネル因子はガウス分布に従う
      - i-vectorの非ガウス性を扱うため，Student-t分布を用いたPLDAが報告されている
        - G-PLDAより高性能
- 評価については，「2つのi-vectorが同一の話者モデルから生成されたか否か」について，対数尤度比を評価できる．
- 登録データ，評価データの発話継続長が短い場合，及び発話継続長にミスマッチがある場合は，PLDAをはじめとするアプローチは弱い

窓長変化による
スペクトル変化

- 窓長+: 時間分解能-, 周波数分解能+