Automatic speaker verification systems and spoof detection techniques: review and analysis

3.1.1~

Automatic speaker verification, 話者照合

- フロントエンド，バックエンドモデル
  - Frontend
    - 入力される信号から，特徴量を抽出するフェーズ
      - AD変換
      - 特徴量抽出
  - Backend
    - 特徴量に基づき，照合スコアを算出するフェーズ

- End to endモデル
  - 登録音声と照合音声のペアから直接スコアを計算する

フロントエンド(特徴量抽出)のアプローチ

- メル周波数ケプストラム係数（MFCCs）
- 逆メル周波数ケプストラム係数（IMFCCs）
- 線形周波数ケプストラム係数（LFCC）
- 定常Q値ケプストラム係数（CQCC）
- 線形予測ケプストラム係数（LPCC）
- Perceptual linear prediction (PLP)
- Power normalized cepstrum coefficients (PNCC)
- All-pole group delay function (APGDF)
- Sub-band centroid frequency coefficients (SCFC)
- 深層学習に基づく特徴量抽出

- メル周波数ケプストラム係数（MFCCs）
  - なりすまし検出に対する一般的な特徴量
  - 発話をDFT or FFTし，ガウシアンフィルタを適用，メルスケールへ変換
  - 対数を取り，離散コサイン変換する
  - 定数項や頭12~14項の1,2次導関数が，話者照合において有用

- 逆メル周波数ケプストラム係数（IMFCCs）
  - MFCCは低周波の領域を重視していたが，IMFCCは高周波の領域を重視する
  - MFCCとは相補的な情報が得られる

- 線形周波数ケプストラム係数（LFCC）
  - 全ての周波数領域を等しくモデル化する
  - MFCCより，ASVにおいてはよく働く
  - 音声認識だけでなく，話者識別にも使える
- 定常Q値ケプストラム係数（CQCC）
  - フーリエ変換ベースの手法では，周波数のビンは規則的な間隔で配置されるので，Q値が変動してしまう
  - CQCCでは，周波数のビンは幾何学的な間隔で配置される
  - 信号に定常Q変換を適用し，Q値が一定となることが保証される
    - Q値：品質係数，値が大きいほど振動が安定

- 線形予測ケプストラム係数（LPCC）
  - Linear predictive Coding (LPC)を適用し，話者の特徴を得る
  - LPC係数が得られるため，自己再帰関数によるLPCCに変換して得られる
- Perceptual linear prediction (PLP)
  - 知覚的な動機に基づく線形予測符号化に基づく特徴で，人間の聴覚システムの特徴をモデル化.
- Power normalized cepstrum coefficients (PNCC)
  - パワー正規化ケプストラム係数
  - MFCC等に比べて，ノイズ（0~15dB SNR）が乗った発話に対して有用
  - 計算量が大きいが，他の特徴量と組み合わせると性能が向上する
- All-pole group delay function (APGDF)
  - 人間が認識できない位相情報も扱う
- Sub-band centroid frequency coefficients (SCFC)
  - フォルマントベースの特徴量
  - ケプストラム特徴量で補足できない，サブバンドの補完的な情報を得られる
- 深層学習に基づく特徴量抽出
  - 2011くらいから出てきた
  - 物体認識や画像認識等，コンピュータビジョンの文脈では，CNNやRNNが使われる
  - CNNについては，音声信号が適切に表現されていれば，音声に適合させることもできる
    - 中間層が発話の特徴ベクトルとして抽出できる
      - d-vector, j-vector, x-vector

バックエンドの設計

- 話者照合のバックエンドは，音声の特徴量とその話者情報を入力とするような分類モデルといえる
- 学習により，実発話かなりすましか等の識別パターンを見つけ出し，異なるクラスの特性を学習する
- 照合を行う時は，システム内にある申請者のデータとマッチングを試み，受理するか否かを決める

- 旧来の機械学習手法によるアプローチ
  - なりすまし検出に有用
- 深層学習によるアプローチ
  - 複雑な分布構造を持つ大規模なデータセットを処理できる

バックエンド/ 機械学習手法

- GMMに基づくモデル
  - 混合ガウスモデル(GMM)に基づく手法，デファクト
  - 話者依存の形状は，害す成分で表せることに影響されている
  - 本物か合成かの違いが効率的にモデル化できる
- SVMに基づくモデル
  - 話者検証（話者照合？）やなりすまし検出では，いい仕事をする
  - 1クラスのSVMは，実発話の外れ値検出にも使える
  - Radial Bias Function (RBS)カーネルに基づくSVM
    - 発話に混ざった未知のなりすましをうまく検出できる
- 隠れマルコフモデルに基づくモデル
  - ASVではよく知られた技術
  - TD-ASVシステムをデザインする際によく使われる
- K-means
  - 分類タスクともとらえられるので，分類アルゴリズムも有用

バックエンド/ 深層学習手法1

- DNNに基づくモデル
  - 学習後に，隠れ層から特徴量を抽出できる
  - 各フレームの隣接する文脈も一緒に供給されるので，データの前処理が大切
  - 特徴ベクトルは共通して同じ次元となるため，実用的
- CNNに基づくモデル
  - 畳み込みニューラルネット
  - 本物かなりすましかの区別に適している
  - 畳み込み層のカーネルを用いるため，前処理がほとんど必要ない
  - 隠れ層がパラメータの調整により特徴を学習し，接続層が分類を行う
- RNNに基づくモデル
  - 音声信号の時間履歴を保持できる
  - 全ての入力ベクトルx_tに対して，時刻tでの出力ベクトルを計算，ラベルを付ける
  - タイムステップごとにラベルが付くので，これらを単一のベクトルに集約する必要がある

バックエンド/ 深層学習手法2

- LSTM
  - Long short-term memory network
  - 情報を長期に保持できるような，RNNの特別版
  - LSTMは4つのニューロンそれぞれが特殊な方法で接続されている
  - 過学習が起こりにくくなる
- Wave-U-Net
  - 音声信号は長い範囲で時間的な相関が高いため，高品質の分離が必要
  - リサンプリングを繰り返すことで，対照的な時間領域で特徴マップを計算し，結合する
- 複数モデルのアンサンブル
  - アンサンブル: 複数の機械学習モデルを組み合わせること
  - 複数のモデルを組み合わせることで性能が向上することがある
  - 1つのモデルではバイアスやバリアンスが高くて性能が出なくても，複数組み合わせると，補完しあっていい感じになる