

Automatic speaker verification

Front end

人力による音声の特徴量を
抽出するアーキ

- AD変換
- 特徴量抽出

Backend.

特徴に基づき照合スコアを
算出する。



end to end.

登録・照合音声のペアを
照合スコアと直接計算する。

Chapter 3. | Frontend ⇒ 12-4 .

3.1.1

Xル周波数ケーブル係数 (MFCC_s)

- オーディオ信号における最も一般的な特徴量
- Xルスケールに基づいた音高の周波数表現の比較

Xルスケール

(人間の耳聴覚の性質に基づいた尺度 (低音が強調される))

係数の抽出方法

1. 発話音を FFT or DFT し、スペクトルを得る。

FFT: 高速 —

$$D_{DFT}(i) = DFT(f)$$

DFT: リアル —

2. 三角 or ガウジアン フィルタ を適用し、Xルスケールへ変換する

離散コサイン変換

$$MFB(j) = \sum_{i=1}^L |D_{DFT}(i)|^2 W(i)$$

Discrete Cosine Transform, DCT.

3. スペクトルへの対数を取る

(自動) 言語照合(検証), Automatic

4. スペクトルを離散コサイン変換する

Speaker
Verification

$$MFCC(r) = \sum_{j=1}^J b_j [MFB(j)] \cos\left\{\frac{r(j-0.5)\pi}{J}\right\}$$

ASV.

f : audio frame, $MFB(j)$: j 回のフレームで持つ i 段のフィルタバンプ $W(i)$ に対する計算された Xルスケル

L : トータルの DFT インデックス

スペクトル

r の MFCC 係数は $MFCC(r)$ で "もたらす" べき

一般的に、定数項や、ほんの 12 ~ 14 係数の 1 次・2 次導関数が

言語者照合において有用。

3.1.2

逆 Xル周波数ケーブル係数 (IMFCC_s)

- MFCC_s は逆周波数領域を重視している
- Xル尺度が "どうか"。
- IMFCC_s は高周波の領域を重視する

係数の抽出方法は MFCC_s とほぼ同じだが、

逆 Xルスケール フィルタを用いて実装する。

- これは (1). 周波数のドメイン (音楽的な部分, 音楽的な部分) が低周波から高周波にわたる。

- MFCC_s では相補的な情報が得られる

3.1.3

線形周波数ケーブル係数 (LFC_C)

- 全ての周波数領域を等しくモデル化する
- MFCC_s と ASV では良くはならない。
- 音声認識が大変で、評議会の例ではよく使われる。
- つまり高周波の領域で、声道の長さをモデル化するため



抽出方法

スペクトル全体を一定の解像度で見込み、

三角フィルタバンプを適用する。

3.1.4

Constant Q cepstral coefficients CQCC.

- フーリエベースの変換では、周波数のビンは規則的な間隔で配置される。
- Q値が変動しない。
- CQCCでは周波数のビンは幾何学的配置である。
 - 信号に定常Q変換(Constant Q Transform, CQT)を適用し、Q値が一定な
これが保証される。
- CQTでは、低周波の成分が周波数で決定的で、高周波の成分が時間で決定的である
[この] 音声認識特徴が好い。
 - 上記自己特徴体、言語者照合における特徴(LA, PA)がよく検出される。
他の特徴量も良好なプロマンス。
- Q値: Q-factor
Quality
品質係数
値が大きいほど
振動が安定
- ← とくに
= ?

TODD:

九月22日

抽出方法:

1. CQTを適用し、時間領域を周波数領域へ変換する。
2. 矢量を C_j , $j=1 \dots l$ とする。
3. 信号をサブフレーム化し、幾何学的間隔のCQTビンを、矩形間隔へ変換する
(DCT演算は省略する)
4. DCTを適用する

3.1.5

Linear predictive cepstrum coefficients (LPCCC)

Linear predictive Coding (LPC) を音声フレームに適用し、言者の特徴を得る。

LPCを適用すると LPC係数を得られ、次に自己再帰関数(AR)、LPCCCへ変換される。

1. 音声シグナル $s(t)$ は LPCモデルへ入力される。

$\rightarrow l$ 個の線形予測係数 $[\beta_0, \beta_1, \dots, \beta_{l-1}]$ と誤差信号 $n(t)$

2. これらの線形予測係数は l 個のLPCCC $[c_0, c_1, \dots, c_{l-1}]$ へ変換される。

$$c_j = \begin{cases} \ln(P_n) & \dots \text{if } j=0 \\ -\beta_j + \frac{1}{j} \sum_{k=1}^{j-1} \{- (j-k) \beta_k c_{j-k}\} & \dots \text{if } 1 \leq j \leq l \\ \frac{1}{j} \sum_{k=1}^l \left\{ \frac{-(j-k)}{j} \beta_k c_{j-k} \right\} & \dots \text{if } l < j < r \end{cases}$$

$j: l \rightarrow$ 線形予測係数の
インデックス
 $P_n: 誤差信号のノルム$

3.1.6

Perceptual linear prediction (PLP)

短周期のスペクトラル特徴のID。

知覚的な動特徴に基づく線形予測係数化に基づく特徴で。

人間の聴覚システムの特徴をモデル化する。

DFTによってスペクトル化されたピリオドグラム(FFT)が得られる。推定スペクトルを

計算される → 分散が大きい。

3.1.7 Power normalized cepstrum coefficients (PNCC)

- 10つ - 正規化ケプストラム係数.
- MFCCなどに比べて、112(0~15dBのSNR)の範囲で有用.
- 人間の聴覚システムに基づいて、このプロセスのシミュレートを試みる.

1. STFTする.

2. スペクトルの周波数分析はガントンフィルターバンプを適用する.

3. 112個とロバスト性を確保するために、周波数スペクトルを、非線形で時間変化を伴う連続的確率分布.

4. 上記の中間処理の後、振幅の変動の影響を軽減するため、平均パワーを最大化する.

- 指数値の $1/5$ のパワーベクトリによく、非線形処理する.

- この操作は、生物的な聴覚システムを最大限にシミュレートする.

5. DCTを適用し、PNCC特徴を得る.

- 言語識別において他の特徴量(MFCC, LFCC, etc.)と組み合わせると性能が向上する.

- 計算量は他の特徴量より大きい.

3.1.8 All-pole group delay function (APGDF)

人間の耳は音の位相情報を言語識別に利用する.

抽出された音節は複雑であるため、位相に基づいた特徴量はよく用いられる.

一方、マグニチュードベースの特徴量は人間の聴覚系に知覚されやすくなることが示されている。
(like a MFCC)

が、位相情報は発話音の重要な特徴である。黒丸Y-スケール成績が発話音を区別

する。

（金極モデルを利用して群遅延法によって音声信号が実現される。）

3.1.9 Sub-band centroid frequency coef. (SCFC)

- フィルマントベースの特徴量

- ケーブストラル特徴量の代替として研究されており。

- ケーブストラル特徴量ではなく、サブバンドの補完的な情報を得られる。

抽出:

1. K個のサブバンドがあり、最初と最後のスペクトルの周波数エンジンをマークされる。

2. 各サブバンドの加重平均周波数である。

↗ sub-band centroid frequency

3.1.10. Deep learning based feature extraction techniques.

深層学習による特徴量抽出.

2011.2~3月の Deep learning による feature extraction の発展.

通常、物体認識や画像認識等のコンピュータビジョンの文脈では、CNNs や RNNs が使われる。CNN はついつい、音声信号が適切な表現をとれば、音声を抽出せよとしている。

中間層が発話の特徴ベクトルとして抽出できる。(x-vector とか)

- d-vector

DNN のうち、出力層ではなく、最後の隠れ層の活性化関数の値の特徴ベクトルとされる。

- J-vector

extension of L-vector.

- X-vector

Time Delay Neural Network (TDNN) の出力ベクトル。

抽出した特徴は、x-vector を上回る。

3.2. ハックエンントの音叉言七.

- ASV のハックエンントは、音声の特徴とその発言者情報を入力とするが、分類モデルでいえば、
- 学習を通じて、本物の発言をもつましに発音 ~ 関する特徴(ノード)を抽出し、異なるクラスの特徴を学習する。
- 訓練されたモデルは、システム内の発言者の発言者データとのマッチングを試み、最適なモデルを決定する。

3.2.1

旧来の機械学習手法は、生成的である、分類的である。すなはち、
おりすまし検出は有用。

3.2.1.1 GMMに基づくモデル

- GMM(Gaussian Mixture Model) のアーキテクチャは、ASV の特徴量のモデル。
- この考え方、一般的な話者における形状はガウス分布によって近似されている。
- 本物か、合成かの違いを効率的にモデル化できる。

古戻りながら GMM は 特徴分布をトーランジ分布で慣用形で表す。

モデルを実装するには。

1. ハーラスターとランダムなベクトルを生成する。

2. EM(期待値最大化)アルゴリズムにより最大推定でハーラスターを推定する。

UBM(Universal Background Model) が発話関連のタスクによく用いられる。

～ Maximum a posteriori (MAP) 推定により、タスクに適するモデルを得る。

3.2.1.2 SVM によるモデル

- SVMは2次元の超平面で2つのクラスを分離する。

- クラス間マージンの最大化

- 言語検証がうまく檢出できず、良い仕事をします。

- 1つのSVMは、本物の言語の外本物の検出にも役立ちます。

- Radial Basis Function (RBF) カーネルに基づく SVMは、複数の混ざった未知のなましまじめな検出を行います。

3.2.1.3 HMM based models

- Hidden Markov Model (隠れマルコフモデル) の言語識別でよく使われる技術。

- TD-HMMシステムモデルで隠れ馬

- 豊かな数字フレームワーク、robustな特徴

3.2.1.4 k-means.

言語認識は分類問題とされています。分類問題アリス化が役立つ。

データの近いデータ同士をハッセルブルトで解決します。

3.2.2 深層学習によるアプローチ

- 深層学習では、複雑な分布構造を大きくデータセグメント化します。

- 一生の音声信号を加えて、さまざまな方法によって抽出された特徴ベクトルでも使用します。

- 單体から、フレームから、Back-end では多くの深層学習に基づく手法が用いられます。

- 生の信号を直接入力して、隠れ層から特徴量を得ることができます。

3.2.2.1 DNN によるモデル

- DNNは、隠れ層の分類能力を強化するためには、

クラス間の差別化を捉えるのに訓練されています。

- 学習後は、隠れ層から深い特徴量を抽出できます。

- 各フレームの隣接する context も一緒に供給されるとデータの前処理が大切。

- 特徴の特徴ベクトルは、音声によらず同じ次元なので、実用的。

3.2.2.2 畳み込みニューラルネットワーク (CNN) によるモデル

- CNNはデータを、人内のローカルで認識可能な部分をハーネスして重ねて使うのが適切です。本物がうまく音声の局部を捉えています。

- 畳み込み層のカーネルを用いるため、前処理がほんじで必要ない。

- CNNプロックを複数の "pooling 層" と組み合わせて、データを、人の空間サイズを縮小します。

- 隠れ層は主に活性化関数を使用し、ニューロンを最大にするか決める。

- ReLU

- SoftMax：出力層で使う。

- 隠れ層が105データの調整により信号の特徴を学習し、最終層が分類を行います。

3.2.2.3 RNN ベースのネットワーク.

- Recurrent neural network とは、音声信号の時間情報をキャプチャできる。
- トレーニングや音楽生成では、ネットワークは訓練盤を加えます。
- セル・Xカニズム (25), 複雑な過去の一連の出来事を記憶できます。
- すべての入力ベクトル x_t に対する、時刻 t の出力ベクトルを計算し、それをもとに次のラベルをつける。
- 例: ハムステップごとのラベルが y_t のとき、これらを单一のベクトル v の集合とする必要がある。

1. T を最後のタイムステップとする。ベクトル O_t を出力する。 $(v = O_t)$

2. 時刻 T のすべての出力ベクトル O_t を合計し、 T である確率を計算する。 v は得る。

3. v を attention (時刻 T のすべての出力ベクトル O_t に同じ加重の合計の平均をとることで計算される重みつき平均) と等しくする

3.2.2.4 LSTM based networks.

Long short-term memory network.

音楽を長期間保持できるなど、RNN の特徴 ver.

RNN は単純な neural net module の chain です。LSTM はこれで異なり。

4つのニューロンでそれぞれ特殊な方法で接続されていて構造をもつ。

CNN の勾配消失の欠点を克服し、過学習が走りにくくなる。

3.2.2.5 Wave-U-Net based networks

音声は長い時間で長い時間间隔があるなど、高品質の分割が必要。

リサンプリングをくり返すことで、対照的な時間領域で特徴マスクを計算して結合する。

3.2.2.6 復数モデルのブランブル

ブランブル: 複数の機械学習モデルを組み合わせる。

個々のモデルはハイアスやハイアスが高く、性能が上がりなっても。

これがデータの異なる特徴を学んでいれば、整合性が UP。

ストックキング、ベースティング、ハギングなど。