ISS-P-025

線スペクトル対を入力特徴とした最小分類誤り学習法の検討

梅﨑　直統† 　竹内　勇人† 　落合　翼†
渡辺　秀行‡ 　片桐　滋† 　大崎　美穂†
† 同志社大学 　‡ ATR

1. はじめに

音声合成も可能な音声認識器の開発を進めている．その一環として，広く用いられているメル周波数ケプストラム係数などのような聴覚モデルに基づく認識器入力特徴に代え，発話モデルに由来する線スペクトル対(LSP: Line Spectrum Pair)に基づく特徴の利用可能性を調査する[1]．認識器はプロトタイプ・状態遷移モデルをクラスモデル[2]とし，識別関数は入力 LSP ベクトルとプロトタイプ LSP ベクトルとの間の動的時間軸伸縮を伴う距離である．そこで，LSP に基づく距離の性質を精査するため，ユークリッド距離とマハラノビス距離との比較も行う．

2. 線スペクトル対

LSP は，短時間音声の線形予測分析に基づき，以下の多項式から求められる．

$$P(z)=(1-z^{-1})\prod_{i=2,4,\cdots,p}(1-2z^{-1}\cos\omega_i+z^{-2}) \quad (1)$$

$$Q(z)=(1+z^{-1})\prod_{i=1,3,\cdots,p-1}(1-2z^{-1}\cos\omega_i+z^{-2}) \quad (2)$$

$$\text{ただし，}0<\omega_1<\omega_2<\cdots<\omega_p<\pi \quad (3)$$

ここで$\{\omega_1,\cdots,\omega_p\}$は LSP の角周波数表現である．

3. プロトタイプ・状態遷移モデルに基づく識別関数

分類器の詳細は参考文献[2]を参照されたい．異なる点は，次式に示す(正規化)マハラノビス距離を用いる場合の識別関数である．

$$g_j(\boldsymbol{X};\boldsymbol{\Lambda})=\frac{1}{T}\sum_{t=1}^{T}\left\{\left(\prod_{k=1}^{d}\left(\sigma_{i(\varphi_{j,t},\theta_{j,t},t)}^{\varphi_{j,t}\theta_{j,t}}\right)^2\right)^{\frac{1}{d}}\left\|\frac{\boldsymbol{x}_t-\boldsymbol{r}_{i(\varphi_{j,t},\theta_{j,t},t)}^{\varphi_{j,t}\theta_{j,t}}}{\boldsymbol{\sigma}_{i(\varphi_{j,t},\theta_{j,t},t)}^{\varphi_{j,t}\theta_{j,t}}}\right\|^2\right\} \quad (4)$$

ここで，$i(\varphi_{j,t},\theta_{j,t},t)$は各単語，各時刻における特徴ベクトルとの距離が最小となる状態内のプロトタイプの指標である．ユークリッド距離では，分散ベクトル$\boldsymbol{\sigma}_{i(\varphi_{j,t},\theta_{j,t},t)}^{\varphi_{j,t}\theta_{j,t}}$の要素が全て 1 となる．

4. 評価実験

ETL-WD-I データセットを用いた単語音声認識実験を通して評価した．学習用および試験用には，ともに男女各 5 名の 492 単語を用いた．入力特徴は，10 次の LSP とパワー，および夫々の時間変化量から成る 22 次元ベクトルとした．状態遷移モデルは，音素用に 3 状態を，無音区間用に 1 状態とした．学習には MCE 学習法(学習エポック数: 50)を用いた．

識別関数としてユークリッド距離および(正規化)マハラノビス距離を用いた場合の認識率を表 1 に示す．

表 1. 各手法による認識率．

	ユークリッド距離	(正規化) マハラノビス距離
学習標本	94.17%	91.32%
試験標本	86.08%	90.50%

試験用データに対して，マハラノビス距離を用いた場合の認識率がユークリッド距離を用いた場合を上回っていた．また，ユークリッド距離と比較して，マハラノビス距離の学習用と試験用データに対する分類精度の差が顕著に小さかった．特徴ベクトルの各要素が持つ分布特性の利用が，未知標本耐性を向上させ得ることが示唆されている．

5. 今後の展望

本研究における MCE 学習法によって更新されたプロトタイプは，式(3)の制約条件を満たさなくなる場合がある．これを防ぐには，入力特徴を以下のように変換する．

$$\psi(x)=\beta\tan\left(\frac{\pi}{4}(x-2)\right) \quad (5)$$

$$\tilde{f}_i=\begin{cases}\psi(f_i) & (i=1)\\ \ln\left(\psi(f_i)-\psi(f_{i-1})\right) & (i=2,\cdots,p)\end{cases} \quad (6)$$

ここで$f_i=f_s\omega_i/2\pi$である$(f_s=8[\mathrm{kHz}])$．$\tilde{f}_i$を入力特徴として学習されたプロトタイプは，逆変換をかけると式(3)の制約条件を満たし，入力特徴が満たすべき範囲や順序関係を維持したプロトタイプの学習が可能となる．今後，この LSP の制約を満たす学習を実現するため，式(5)および式 (6)の変換を用いた学習と，プロトタイプから合成される音声に関する研究を行う予定である．

6. まとめ

LSP を入力特徴とする音声認識器の基本的性能を調査し，特にマハラノビス距離に基づく識別関数とする場合に，試験用データに対して高い分類精度を達成し得ることを示した．

謝辞: 本研究の一部は科研費(JP26280063)および文科省 H26 年度私学戦略的研究基盤形成支援事業「ドライバ・イン・ザ・ループ」の支援を受けて行われた．また，ETL-WD-I データベースは NII 音声資源コンソーシアムからの提供を受けた．

参考文献

[1] ITU-T, “G.729: Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)”, 2007.

[2] 松廣達也，他，大幾何マージン最小分類誤り学習法を用いた音声認識に関する実験的評価，情報処理学会関西支部支部大会講演論文集，6p，2016