ISS-A-056

CS-ACELP を用いた音声合成可能性を正則化条件とする最小分類誤り学習法の実験的評価

丸山　右京†　　梅崎　直統†　　大久保　拓海†
渡辺　秀行††　　片桐　滋†　　大崎　美穂†
† 同志社大学　　†† ATR

1. はじめに

音声パターン認識における分類器の究極の目標は，ベイズ誤り状態の達成である．しかし，学習後の分類器がベイズ誤りを偏って推定する状況を避けるのは容易ではない．最近この問題を解決するため，音声合成可能性を正則化条件とする最小分類誤り(MCE: Minimum Classification Error)学習法が提案された[1]．本稿では，そこで示唆された高いベイズ誤り推定能力を，先行研究とは異なるデータセットを用いて検証する．

2. 音声合成可能性を正則化条件とする MCE 学習法

本学習法は，線スペクトル対-共役構造代数符号励振線形予測(LSP-CS-ACELP: Line Spectral Pairs-Conjugate Structure-Algebraic Code Excited Linear Prediction)法[2]を組み込み，音素モデルを複数プロトタイプ・状態遷移モデルで構成する音声認識器(図 1)の利用を前提とする．この認識器は，分類器内のプロトタイプから得る LSP パラメータを用いて入力音声を真似る合成音声を生成できる．

基本的に，この音声認識器によるベイズ誤り推定の偏りは，状態内のプロトタイプの過不足に因る．特に，プロトタイプが多過ぎると，学習用標本に対するベイズ誤り推定値が過度に小さくなる一方で，試験用標本に対する推定値は過大となる．本学習法は，(LSP パラメータの意味で)プロトタイプと学習標本との距離を表す正則化項を設け，分類誤り数損失の最小化と同時にその正則化項の最小化も行うことで，そのベイズ誤り推定の偏りの抑制を目指す．

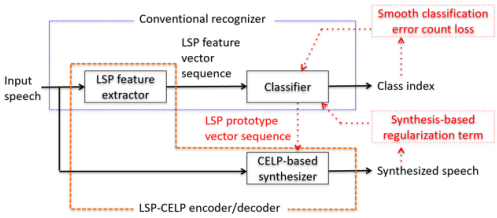

図1 採用する音声認識器の構造．

3. 評価実験

東北大-松下単語音声データベース(TMW)を用いた単語音声認識実験を用いて評価した．学習用および試験用には，共に男女各 30 名の 172 単語を用いた．入力特徴は，10 次の LSP とパワー，及び其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，無音区間用に 1 状態とした．

正則化係数βに 7 つの値を設定して得た学習法および試験用標本に対するベイズ誤り推定値を示す(図 2)．図中，緑の直線は十分大きなKを設定したセグメントK平均法を用いて 10 分割の交差検証法で求めた，このデータに関するベイズ誤り参考値(10.69%)である．βの値が小さい時ベイズ誤りは過小推定され，その値が大きい時に過大推定される一方で，$\beta = 1$の時，正確なベイズ誤り推定が実現されている(試験用標本に対する推定値=10.60%)ことがわかる．

学習後の分類器の音声合成能力を調査するため，合成音声の波形を観察した(図 3)．その結果，$\beta = 1$の場合がちょうど音声合成可能性の臨界点(学習用標本によって合成の可不可が分かれる状態)にあることがわかった．

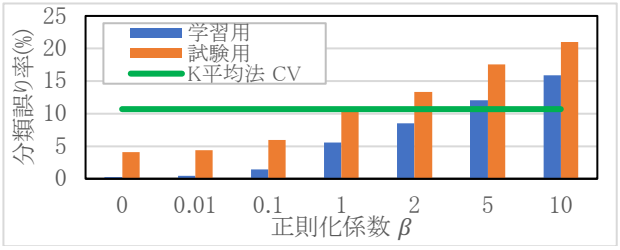

図 2　各βにおける分類誤り率(ベイズ誤り推定値)．

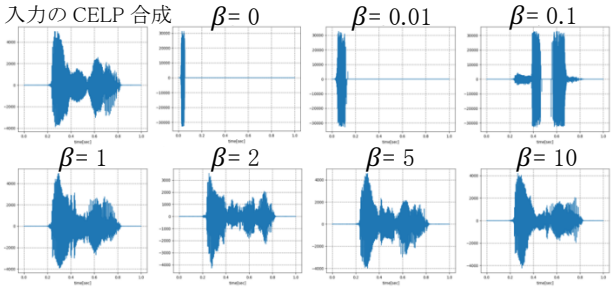

図 3　各βにおける単語“番号(ばんごう)”の合成音声．

5. まとめ

分類誤り数の最小化と音声合成能力の維持とのバランスをとることで，ベイズ誤りの正確な推定が可能であることを再確認できた．このバランスを適切に設定する方法を明らかにすることが今後の課題である．

謝辞　本研究の一部は，科研費(18H03266)の支援を受けた．またデータ TMW は NII-SRC から提供を受けた．

参考文献

[1] N. Umezaki, et al. Proc. SPML, pp. 62-72, ACM (2019 Nov).
[2] ITU-T Rec. G729: Coding of speech at 8kbits/s using Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP)