ISS-A-057

# 音声合成能力を正則化条件とする最小分類誤り学習法の音素認識による実験的評価

丸山　右京†　　片桐　滋†　　大崎　美穂†
† 同志社大学大学院　理工学研究科

## 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，実際に学習に用いることができる標本数は有限個であるので，ベイズ誤りの過小推定，即ち過学習が起きてしまう．先行研究では，この問題を回避するため，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[1]が提案された．しかし，単語音声認識以外での調査は行われておらず，分類精度や過学習の抑制効果への影響は分かっていない．本稿では，音素認識に本学習法を適用し，そのベイズ誤りの推定能力と音声合成能力との関係性を調べることを目的とする．

## 2. 音声合成能力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．

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，過学習の抑制を目指す．

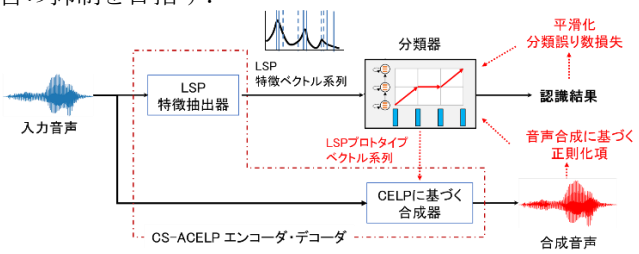


図1 採用する音声認識器の構造．

## 3. 評価実験

評価実験において TIMIT データセットをもとに音素データセットを作成した．また，そこから母音のみを抜き出したデータセットを 2 つ作成した．本稿では母音の音素で構成した 5 クラスのデータセットによる結果を掲載する．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．総話者数 486 名(男: 342 名，女: 144 名)，学習用 14488 個，試験用 749 個の標本を用いた．

正則化係数$\beta$に 6 つの値を設定して得た学習用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，緑の点線は十分大きな$K$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(39.63%)である．$\beta = 0.01$の時に試験用標本への分類誤り率が 41.12%となり，ベイズ誤りの参考値に最も近くなった．

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，どの状態においても元の音声を適切に再現できたものは無く，その一方，過学習が起きている結果もなかった．分類器の学習回数が不足し，分類誤り率の最小化を行えていない状態であることが予想され，音声合成が不十分であったものと考えられる．

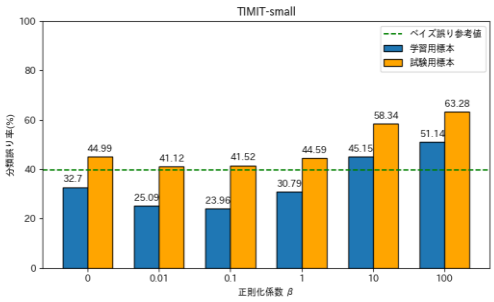


図 2　各$\beta$における分類誤り率(ベイズ誤り推定値)．

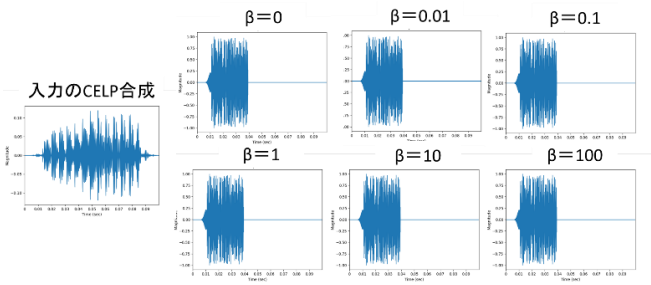


図 3　各$\beta$における音素“aa”の合成音声．

## 4. まとめ

先行研究で確認されていた分類器の分類精度と音声合成能力との関連性や過学習の抑制効果を示すことはできなかった．その原因として，分類器の学習不足が考えられるため，分類器の分類精度を向上させた後，改めて正則化の有効性を確認する必要がある．

**謝辞**　本研究は，科研費(18H03266)の支援を受けた．

## 参考文献

[1] 梅崎直統．同志社大学大学院理工学研究科修士論文．2020.

[2] ITU-T Rec. G729: Coding of speech at 8kbits/s using Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP) .