

話者適応処理が発音評定精度に及ぼす影響に関する定量的分析*

羅徳安, 喬宇, 峯松信明 (東大), 山内豊 (東京国際大), 牧野武彦 (中央大), 広瀬啓吉 (東大)

1 はじめに

近年, 音声情報処理技術の進展に伴い, 自動音声認識技術を用いた CALL (Computer-Assisted Language Learning) システムに関する研究が盛んに行われている。しかし, 学習対象言語の母語話者による音響モデルを利用して, 非母語話者である学習者の音声を認識または評価する場合, 話者性や録音環境以外に, 対象言語の習熟度に起因する誤りが起ると考えられる。

例えば音声認識に使われている話者・環境適応技術は, 発話内容を識別することを目的として, 話者性や環境雑音による影響を最小限に抑えられるように構築されている。しかし, これらの技術を発音評価にそのまま利用すると, 話者性や環境性のみならず, 習熟度に対する適応がかかり, 誤った発音を正解として判定する恐れがある。これは音韻の違いも話者性の違いもスペクトル包絡によって表現され, 話者適応は音響モデルのスペクトル包絡特性を変形するからである。

本研究は, 話者適応技術を CALL に応用する際, どのような問題が生じるのかを実験的に分析した。日本人学生による英語読み上げ音声 (ERJ) データベースを用いて, MLLR 話者適応において, 回帰木のクラスタ数の変化によって, 自動スコアと手動スコアとの相関関係の変化を調べた。これを踏まえ, 音響モデルを選択的に利用し, 音素境界検出と習熟度推定プロセスを分離する GOP (Goodness of Pronunciation) スコアの計算方法を提案し, 従来の GOP 計算方法より優位であることを実験的に示した。

2 外国語発音評価と話者適応

音声情報処理を用いた発音評価システムは, 機能的に, 主に 2 種類の技術に支えられている:

1. 発音の自動スコアリング技術 [1]
2. 発音誤り検出技術 [2]

前者は, 発音の習熟度を定量的に測定する方法である。その一つの指標として, 音響モデルの尤度比に基づく事後確率である GOP スコアが有効であることが示されている [3]。後者については, 誤りパターンを予測するネットワーク文法や, 音素 GOP スコアを用いて, 誤って発音された音声セグメントを検出する手法などが挙げられる。機能的には, 前者が能力試験, 後者は日々の学習を支援する技術である。

この二つの技術には, 一般的に学習対象言語の不特定話者音響モデルが必要である。学習者による発声をモデル化することは不可能ではないが, 発音誤りによる発声の多様性から, 高精度な音韻モデルを構築することが難しい。特定話者の音声をモデル化する場合, データ不足などの問題が考えられる。しかし, 母国語話者による不特定話者音響モデルを用いた場合でも, 特定の学習者の音声を評価する場合, 話者性や発音の習熟度によるミスマッチが当然生じる。いかに話者性のミスマッチを抑えながら, 習熟度の違いを適切に判断できるかが, CALL に求められる技術である。本研究は発音の明瞭度を表す GOP 評価とネットワーク文法による発音誤り検出の 2 つの観点から, 話者適応の問題点を調べ, これらを回避する方法を提案する。

3 適応技術の利用に関する評価実験

3.1 GOP スコアの算出

本研究では, WSJ 及び TIMIT データベースから学習した不特定話者用 HMM 音響モデルを用いて, 日本人学生による英語読み上げ音声の評定の際に参照する GOP スコアを以下のように算出する。音素 p と同定された音声セグメント

* Quantitative analysis of the adverse effect of speaker adaptation on pronunciation evaluation
by D.Luo, Y.Qiao, N.Minematsu, Y.Yamauchi, T.Mkakino and K.Hirose

Table 1 手動評価の教師間相関関係

	教師 1	教師 2	教師 3	他 2 人平均
教師 1	1	0.91	0.87	0.95
教師 2		1	0.77	0.87
教師 3			1	0.84

O に対して, $GOP(p)$ は以下の式で定義する.

$$GOP(p) = \frac{1}{D_p} \log (P(p|O)) \quad (1)$$

$$= \frac{1}{D_p} \log \left(\frac{P(O|p) P(p)}{\sum_{q \in Q} P(O|q) P(q)} \right) \quad (2)$$

$$\approx \frac{1}{D_p} \log \left(\frac{P(O|p)}{\max_{q \in Q} P(O|q)} \right) \quad (3)$$

$P(p|O)$ は観測音声 O が与えられた場合, それが音素 p による発声である事後確率である. Q は全ての音素の集合である. D_p は音素 p の持続長である. (3) 式の分子の部分は, HMM による強制アライメントによって算出できる. 分母は連続音素認識による尤度で近似的に計算できる.

3.1.1 実験データ

GOP スコアと教師採点間の相関分析では, 「日本人学生による読み上げ音声 (ERJ) データベース」から教師による評価スコアを参照し, 上級・中級・初級各 10 名 (男女各 5 名) ずつ計 30 名を選び, 一人 10 文ずつ, 計 300 文の読み上げ音声を使用した. 教師の採点は, 同データベースに記載されている, 3 名の教師による発話ごとの音素明瞭度スコアを利用する. 選定した 10 文から各話者の平均スコアを求めた. 平均スコアの教師間相関を Table 1 に示す. 以下, 3 名教師の採点の平均を手動スコアとして利用する.

3.1.2 話者適応が GOP 評価に及ぼす影響分析

WSJ+TIMIT データベースで学習した HMM 音響モデルをベースに, MLLR 話者適応による GOP スコアと手動スコアとの相関値の変化を調べた. 音響分析条件を Table 2 に示す. 評価話者による単語音声 (各話者 180 単語) を適応データとして, 教師あり MLLR 適応を行った. 未適応のモデル (クラスタ数 0 とする) と, MLLR 回帰木のクラスタ数を 1 (大局的な適応) から, 2, 4, 8, 16, 32, 64 と増やして適応したモデルを用いて, GOP スコアを求めた. GOP スコアは発

Table 2 音響分析条件

サンプリング条件	16 kHz / 16 bit
窓及び窓長	Hamming / 25 ms length
シフト長	10 ms shift
音響パラメータ	MFCC と対数パワー, Δ , $\Delta\Delta$ (計 39 次元)

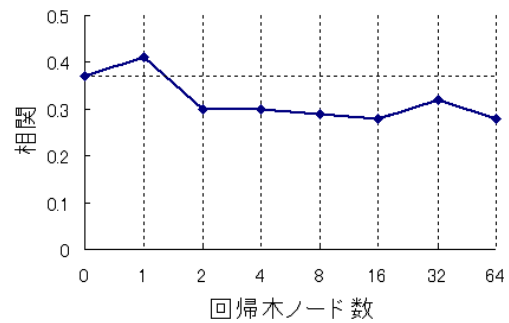


Fig. 1 MLLR 回帰木ノード数の増加による GOP・手動スコア間の相関値変化

話単位で求めることができるが, これを各話者毎に平均を求め, それをその話者のスコアとする. GOP と手動スコアの相関値を回帰木ノード数の関数として, Fig. 1 に示す. 適応していない音響モデルで求めた GOP と手動スコアとの相関が 0.37 である. 大局的適応 ($N = 1$) 時に, 相関が最も高く 0.41 となり, クラスタ数 N を 2 以上に増やすと, 相関値は下がり, 適応前のモデルよりも低い相関を示している.

3.2 発音誤り検出実験

学習者の音声と音響モデルとのミスマッチがあると, 誤り検出率が低くなることが予想される. しかし, 過適応すると, 不適切な音素を正解とみなし, この場合も誤りが検出できなくなる. ここで, 話者適応による誤り検出への影響を調べるために学習者が意図した音素系列に対して, ネットワーク文法を作成し, 種々の適応モデルを用いて, 誤り再現率の変動を調べた.

発音誤りを考慮した正解ラベリングは, 音声学者 (第 5 著者) によって行われた. 即ち ERJ データベースから各話者 1 文ずつを選んで, 音声記号で書き起こしたもの (30 名合計 30 文) を参照して [4], 音素記号に書き換えたものを正解ラベルとして利用した. 誤り検出用の文法は [5] を参照し, 日本人英語に頻出する誤りパターンを

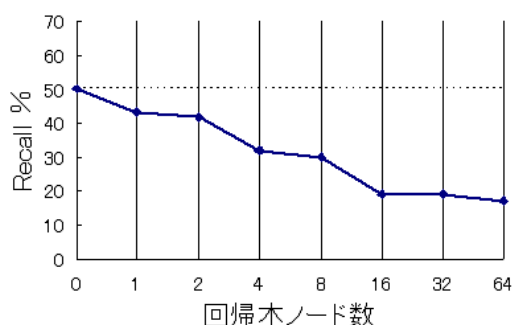


Fig. 2 話者適応による誤り再現率の変化

含めたネットワーク文法を構築した．文法には含まれていないが，実際に起きた誤りに関しては，適宜文法に追加して誤りカバー率を 100 % となるようにした．適応手法は GOP 評価と同じく，適応前の音響モデル ($N=0$)，クラスタ数を 1 から 2, 4, 16, 32, 64 に増加した時の音響モデルを用いて誤り再現率を算出した．結果を Fig. 2 に示す．話者適応の副作用として ($N = 1$ の時も含め) 誤り再現率が未適応時より低くなった．その結果，一部の誤りが検出不可能となる．

以上示したように話者適応をそのまま導入すると，発音評価，誤り検出共に精度を劣化させることが分かった．以下，GOP 算出法に着眼して，適応技術の効果的利用を検討する．

4 GOP 算出における話者適応技術の効果的な利用に関する検討

4.1 Forced aligned GOP 計算手法

3.1 で述べたように，GOP スコアは，強制アライメントと連続音素認識で求めた対数尤度の差分として，計算することができる．しかし，連続音素認識では，音素境界の検出精度が非常に低くなる．Fig. 3 は強制アライメントで検出した音素境界，ボトムアップクラスタリングによる教師なし音素セグメンテーション [6]，及び連続音素認識で検出した音素境界を示す．連続音素認識では，容易に挿入誤りが起こるため，音素境界検出は，音響モデルを参照しない教師なし音素セグメンテーションよりも低くなることも多い．

この例からも明らかなように強制アライメントによる音素境界と連続音素認識による境界は必ずしも一致しない．例えば，Fig. 4 のように，強制アライメントによる音素 p のセグメントに

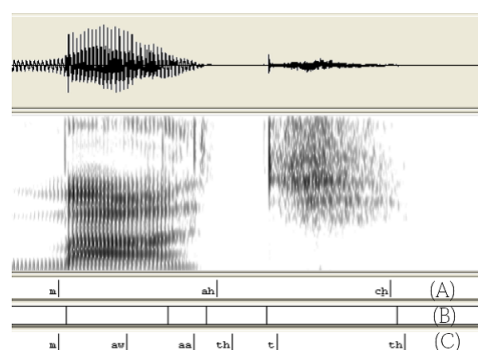


Fig. 3 (A) 強制アライメント，(B) 教師なしセグメンテーション，(C) 連続音素認識，による音素境界同定

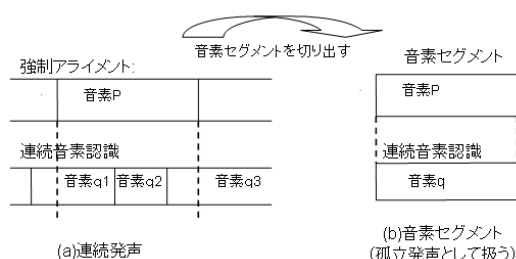


Fig. 4 Forced aligned GOP 算出手法

対して，連続音素認識では音素 $q1, q2, q3$ が対応することもある．このセグメントに対して GOP を求める場合，[3] では，(3) 式の分母として，各フレームに対して，対応する HMM 状態からの出力確率尤度を用いている．即ち音素境界のずれに対して，何ら対処していない．

そこで本研究では，連続音素認識結果は参照しない GOP スコア計算法を提案する．強制アライメントによって切り出された音素セグメントに対して，強制アライメント時とは異なる音響モデルを使って，意図された p としての尤度を求める．次に，そのセグメントに対して，尤度最大

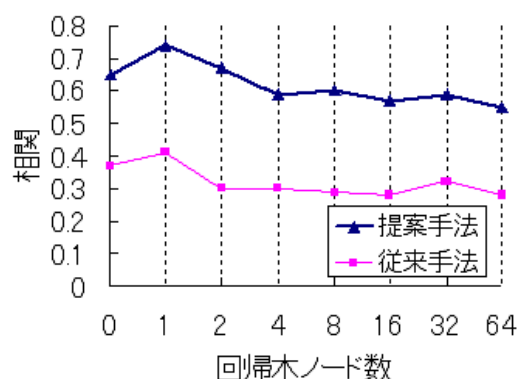


Fig. 5 Forced aligned GOP・手動スコア間の相関値変化

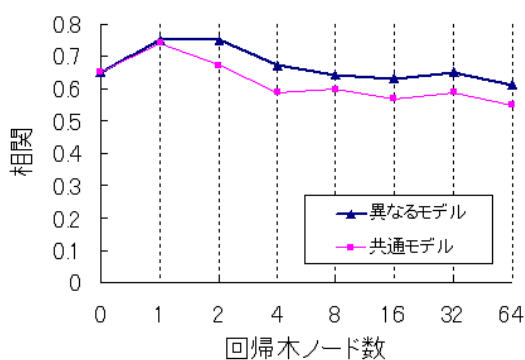


Fig. 6 異なるモデルでの Forced aligned GOP・手動スコア間の相関値変化

音素となる q を求める．その尤度を式 (3) の分母として GOP を求める．

結局提案手法は，音素境界検出と GOP 推定を分離する形となり，それぞれのステップで異なる音響モデル（または異なる適応行列による話者適応）を用いることが可能となる．例えば，音素境界推定には適応モデルを，GOP 推定には母語話者の不特定話者モデルを用いる．この手法を Forced aligned GOP (FA-GOP) と呼び，以下，その手法の有効性を実験的に示す．

4.2 実験的検証

前節と同様の条件で，まず適応していない音響モデルを用いて，強制アライメントと FA-GOP 計算を行った．話者単位での手動スコアと GOP との相関は 0.65 となり，従来の手法より，大幅に向上していることが確認された．次に，MLLR 適応における回帰木のクラスタ数を変化させて，GOP スコアと手動スコアとの相関値を調べた．まず，音素境界検出と GOP の計算を共通の音響モデルで行った結果を Fig. 5 に示す．最も相関が高いのは，クラスタ数が 1（大局適応）であり，クラスタ数が 2 以上になると，相関が下がっていく傾向が見られる．また従来の手法で求めた GOP スコア（Fig. 1）より相関値がおよそ倍になっていることも分かる．次に，音素境界検出を適応モデルで，習熟度の計算を母語話者モデル（無適応）で評価した結果を Fig. 6 に示す．相関が最も高いのは，クラスタ数が 1 と 2 のモデルで音素境界を検出する時であり，相関が 0.75 となった．クラスタ数を増加しても，共通のモデルで音素境界と習熟度を計算するより相関が高い

ことがわかる．しかし，今回の実験の場合，相関の最高値は音素境界検出と GOP 計算に共通の音響モデルを用いた場合（ $N = 1$ ）とほぼ等しく，話者適応を積極的に有効利用するには至らなかった．

5 まとめ

MLLR 適応の回帰木のクラスタ数を 1 から増やして，大局及び局所的適応を行い，自動スコアリングと誤り検出，2 つの側面から，発音評価への影響を調べた．また，音素境界と習熟度計算を分離する Forced aligned GOP 手法を提案し，従来の手法より優位であることを示した．

今後は，学習者の特性や，誤り種類などをより詳しく分析し，話者適応による影響を考察する予定である．なお本報では GOP と手動スコアの相関値を改善する技術的提案を行ったが，同様の改善は学習者に課すタスクを適切に制御することでも可能である（例えば，読み上げではなくシャドーイングを課す [7]）．このような教育的提案と本報で示した技術的提案の組合せの相乗効果についても検討予定である．

参考文献

- [1] 太田, 中川, “日本語学習者の英語語発音評価法”, 春音講論, 2-1-13, 2005.
- [2] A. Raux et al., “Automatic Intelligibility Assessment and Diagnosis of Critical Pronunciation Errors for Computer-Assisted Pronunciation Learning,” Proc. ICSLP, pp.737-740, 2002.
- [3] S.M. Witt and S.J. Young, “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning,” Speech Communications, 30(2-3):95-108, 2000.
- [4] T. Makino, “A corpus of Japanese speakers’ pronunciation of American English: preliminary research” Proc. Phonetics Teaching and Learning Conference, University College London, 2007.
- [5] S. Kohmoto, “Applied English Phonology: Teaching of English pronunciation to the Native Japanese Speaker,” Tokyo Tanaka Press, 1965.
- [6] 下村他, “制約条件つきクラスタリングによる連続音声からのイベント境界検出”, 信学技報, SP2007-12, pp.25-30, 2007.
- [7] D. Luo et al., “Automatic pronunciation evaluation of language learners’ utterances generated through shadowing,” Proc. INTERSPEECH, pp.2807-2810, 2008.