

## シャドーイング・音読発音評価を目的とした話者適応の分析と応用

羅 徳安<sup>†</sup> 喬 宇<sup>†</sup> 峯松 信明<sup>†</sup> 山内 豊<sup>‡</sup> 広瀬 啓吉<sup>†</sup>

<sup>†</sup> 東京大学 〒113-0033 東京都文京区本郷 7-3-1

<sup>‡</sup> 東京国際大学商学部 〒350-1197 埼玉県川越市の場北 1-13-1

E-mail: <sup>†</sup> {dean,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp, <sup>‡</sup> yyama@tiu.ac.jp

**あらまし** 本研究では、話者適応技術が発音評価に及ぼす作用・副作用を実験的に分析し、適応技術のシャドーイングと音読音声評価への効果的な応用を試みた。まず、日本人学生による英語読み上げ音声データベースを用いて、習熟度自動推定スコアと手動スコアとの相関分析と誤り検出実験を行った。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新しい GOP (Goodness of Pronunciation) スコアの計算方法と MLLR に制限を加える適応手法を提案し、従来の手法より優位であることを実験的に示した。さらに、日本人学習者に対して、テキスト難易度の異なるシャドーイング・文字付きシャドーイング・音読音声を収録し、自動スコアと TOEIC スコアとの相関分析を行った。実験の結果、大局的話者適応はシャドーイング音声よりも音読音声に効果が顕著である一方、シャドーイング評価は音読より学習者の総合的語学能力を反映しやすいことが示された。また、シャドーイングに基づく自動評価精度はテキスト難易度にも依存することが分かった。

**キーワード** シャドーイング, Goodness of Pronunciation, 発音評定法, 外国語発音支援

## Analysis and Utilization of Speaker Adaptation Techniques for Shadowing and Read-Speech Pronunciation Evaluation

Dean LUO<sup>†</sup> Yu Qiao<sup>†</sup> Nobuaki MINEMATSU<sup>†</sup> Yutaka YAMAUCHI<sup>‡</sup> and

Keikichi HIROSE<sup>†</sup>

<sup>†</sup> The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 105-0123 Japan

<sup>‡</sup> Tokyo International University, 1-13-1 Matobakita Kawagoe, Saitama, 350-1197 Japan

E-mail: <sup>†</sup> {dean,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp, <sup>‡</sup> yyama@tiu.ac.jp

**Abstract** In this paper, we investigate the effects and problems of MLLR speaker adaptation when applied to pronunciation evaluation on shadowing and reading aloud. Automatic scoring and error detection experiments are conducted on two databases of Japanese learners' English pronunciation. As we expected, over-adaptation causes misjudge of pronunciation accuracy. Following these experiments, two novel methods, Forced-aligned GOP scoring and Regularized-MLLR adaptation, are proposed to solve the adverse effects of MLLR adaption. Finally we examine how the automatic scores, the conventional GOP and proposed F-GOP, change according to the cognitive loads posed on learners. Experimental results show that shadowing can better reflect the learners' true proficiency than reading aloud, and by selecting native utterances of adequate semantic difficulty, the evaluation performance by shadowing is even improved.

**Keyword** shadowing, goodness of pronunciation, automatic scoring, CALL

### 1. はじめに

近年、外国語教育において、シャドーイングという学習方法が広がりを見せている。シャドーイングとは、聴取した外国語音声を即座に繰り返して発声する外国語聴取・発音訓練法である。元来、同時通訳者の訓練として広く行なわれていたが、外国語学習においてもシャドーイング学習の効果が認められるようになった[1],[2]。

しかし、シャドーイングは非常に負荷の高い訓練法

であり、シャドーイング音声は一般にかなり「崩れた」音声となる。人手でこれらを逐一評定することは膨大な時間を要するため、筆者らは音声認識技術を用いたシャドーイング発音評定の自動化を提案した[3]。

一方で、音読は従来から外国語教育に使われている手法である。シャドーイングに比べて、比較的「明瞭な」音声となり、発音評価システムでは、一般的に、この読み上げ（音読）音声を使っている。シャドーイングと比較し、音読における発音評価の向上も本研究

の目的の一つである。

自動音声認識技術を用いた発音支援システムでは、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ため、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。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。

本研究では、シャドーイングと音読音声の評価精度の向上を目的として、まず話者適応の発音評価に対する影響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、MLLR 話者適応回帰木クラス数の変化によって、習熟度自動推定スコアと手動スコアとの相関関係の変化を調べた。次に、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新しい GOP (Goodness of Pronunciation) スコアの計算方法である Forced Aligned GOP 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。更に、日本人学習者による単語読み上げ音声である産総研「Basic English」データベースを用いて、発音誤り検出性能についても調べた。分析結果に基づき、MLLR に制限を加える適応手法 (Regularized-MLLR) を提案し、誤り検出性能が向上したことを示した。最後に、テキストの難易度を 3 段階に分け、日本人学習者によるシャドーイング・テキスト付きシャドーイング・音読音声を収録し、自動評価と TOEIC との相関分析を行った。実験の結果、シャドーイングは音読より、学習者本来の総合的語学能力を反映しやすいことが分かった。また、テキスト難易度の選択もシャドーイングの自動評価に大きく影響することを実験的に示した。

## 2. 適応技術の利用に関する分析・評価

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[4]
2. 発音誤り検出技術[5]

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[4]。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は、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。

### 2.1. 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 HMM 音響モデルを用いて、シャドーイング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  $p$  と観測された音声セグメント  $O^{(p)}$  に対して、 $GOP(p)$  は以下の式によって定義する。

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

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

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

$P(p | O^{(p)})$  は観測音声  $O^{(p)}$  が与えられた場合、それが音素  $p$  による発声である確率である。 $Q$  はすべ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 Viterbi 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することができる。このように求めた平均音素 GOP を自動評定スコアの一つとして利用する。

#### 2.1.1. 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」[6]から教師間習熟度スコアの一致度の高く、かつ様々な習熟度レベルを網羅した学習者 42 名 (男女各 21 名ずつ) を選択した。各話者による 30 文読み上げ音声を評価データとして、平均音素 GOP スコアを求め、その話者の

自動評価として利用する。

### 2.1.2. 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。評価話者による 60 文読み上げ音声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 スコアは発話単位で求めることができるが、これを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、図 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 0.63 である。大局的適応 ( $N=1$ ) 時に、相関が最も高く 0.65 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、 $N$  が 4 以上になると適応前のモデルよりも低い相関を示している。

### 2.2. 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。評価データとして、日本人学習者による単語読み上げ音声データベース[7]を用いた。このデータベースには、7 名の学習者による 850 個単語の読み上げ音声収録しており、また各音声に対して専門家が忠実に書き起こした誤りラベルも含まれている。本研究では、このデータベースから、日本人学習者による典型的な誤りが多く含まれている 4 人の話者（男女各 2 名ずつ）を選び、400 個単語発声を評価データとし、残り 450 個単語発声を適応データとして利用する。誤り検出用の文法は[8]を参照し、日本人英語に頻出する誤りパターンを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 に増加した時の音響モデルを用いて誤り適合率と再現率及び両者の F-measure を算出した。結果を図 2 に示す。

回帰木のノード数  $N$  の増加につれ、適合率が増加していく。これは適応によるミスマッチを低減したことによる効果であると考えられる。しかし、再現率に関しては、 $N$  が 2 以上に増やすと未適応モデルより低くなる。これは、過適応によって誤りを正解として判断

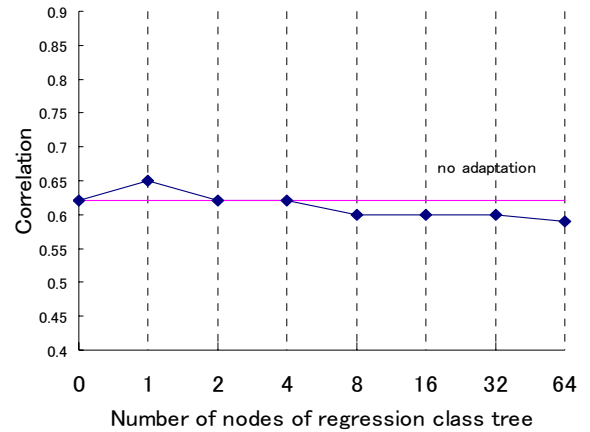


図 1: MLLR 回帰木ノード数の増加による GOP・手動スコア間の相関値変化

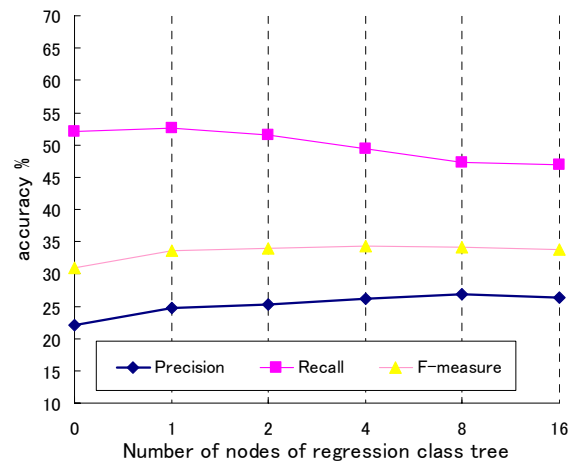


図 2: MLLR 回帰木ノード数の増加による誤り検出評価性能の変化

し、検出不可となっていることを意味する。

以上示したように話者適応をそのまま導入すると、過適応による発音評価、誤り検出共に精度を劣化させることが分かった。以下、適応技術の効果的利用を検討する。

## 3. 適応技術の効果的な利用

### 3.1. 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

2.1 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。図 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[9]、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

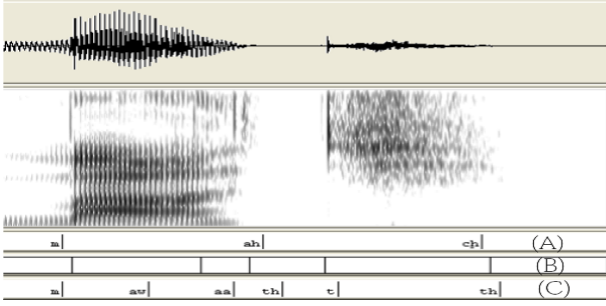


図 3: (A)強制アライメント, (B)教師なしセグメンテーション, (C)連続音素認識, による音素境界同定

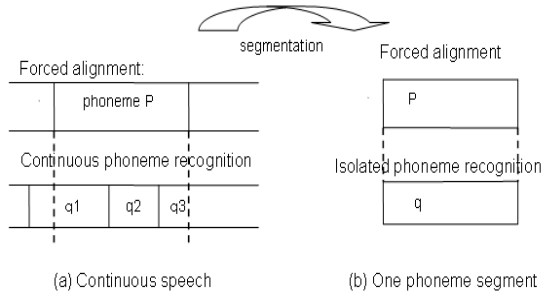


図 4: Forced aligned GOP 算出手法

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、図 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、[4] 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 Forced aligned GOP(F-GOP)を提案する。本手法では、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。

ここで 2.1.2 と同様の条件で、F-GOP 計算を以下 2 通りで計算した。F-GOP1 は境界検出と習熟度計算において、同じモデルを用いる。F-GOP2 は、境界検出において、適応モデル ( $N > 0$  の場合) を用いて、習熟

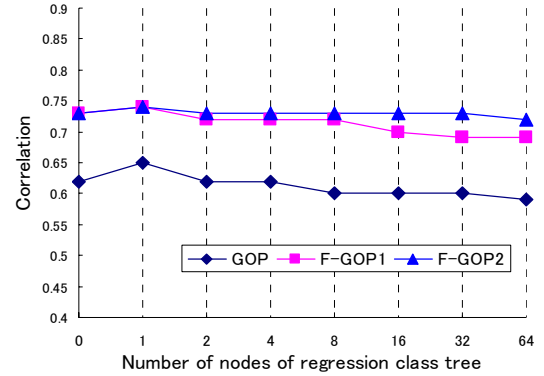


図 5: F-GOP 教師採点間の相関値変化

度計算において、未適応モデルを用いる。自動スコアと教師採点間の相関を図 5.に示す。GOP より、F-GOP のほうが有利であることが明らかである。N が 2 より大きい場合、F-GOP1 より F-GOP2 の性能が良い。これは、過適応による習熟度計算性能の低下を意味する。

### 3.2. 誤り検出における話者適応技術の効果的な利用に関する検討

2.2 節では、話者適応によるミスマッチの低減効果と過適応による再現率の低下を明らかにした。話者適応の利点を活用し、副作用を抑えるために、MLLR 話者適応に制限を加えた Regularized-MLLR を提案する。

#### 3.2.1. Regularized-MLLR (R-MLLR) の定義

誤り発音を正解として適応されないように、従来の MLLR 適応になんらかの制限を設ける必要がある。

従来の MLLR 適応では、クラスター  $r$  に対して、変換行列  $W_r$  は以下の  $Q$  関数を最大化するものとして定義される。

$$Q(M, \hat{M}) = \frac{1}{2} \sum_{r=1}^R \sum_{m_r=1}^{M_r} \sum_{t=1}^T L_{m_r}(t) \times \left[ K^{(m)} + \log |\hat{\Sigma}_{m_r}| + (o(t) - \hat{\mu}_{m_r})^T \hat{\Sigma}_{m_r}^{-1} (o(t) - \hat{\mu}_{m_r}) \right] \quad (4)$$

ここに  $M$  は HMM セットであり、 $\hat{M}$  は適応したモデルである。 $R$  は回帰木ノードの数であり、 $M_r$  はクラス  $r$  に属するガウス成分である。 $K^{(m)}$  は定数成分であり、 $L_{m_r}(t)$  は以下のように定義した確率である。

$$L_{m_r}(t) = p(q_{m_r}(t) | M, O_T) \quad (5)$$

ここに、 $O_T$  は適応データであり、 $q_{m_r}(t)$  は時刻  $t$  に  $O_T$  が出力されるガウス分布である。

ここで、 $\{W_r^{C_1}, \dots, W_r^{C_N}\}$  を  $N$  名教師の音声から求めた変換行列とし、その平均行列は  $W_r^C = (1/N) \sum_n W_r^{C_n}$

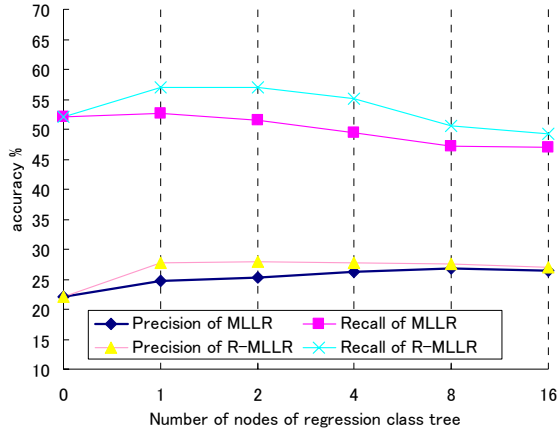


図 6: 誤り検出実験における R-MLLR の効果

とすれば, Regularized-MLLR (R-MLLR) の Q 関数は以下のように定義できる.

$$\max_{W_r} \{Q(M, \hat{M}) - \lambda \sum_{r=1}^R \|W_r - W_r^C\|_F^2\} \quad (6)$$

ここに,  $\lambda$  は話者性に依存するパラメーターである. (6)式の第 2 項は, まさに従来の MLLR の Q 関数に加えた, 過適応を防ぐための制限である.

### 3.2.2. 実験結果

本研究では, ERJ データベースに記載される 10 名の General English 話者 (男女 5 名ずつ) の音声を教師音声として, 教師平均変換行列  $W_r^C$  を求めた. 実験データと検出文法は 2.2 節と同じであり,  $\lambda$  は学習者ごとに実験的に求め, R-MLLR 適応を行った.

誤り検出結果を, 図 6. に示す. 従来の MLLR 適応に比べて, 適応率と再現率を共に向上させることができた.

以上, 読み上げ音声を対象に, その評価精度及び誤り検出精度の向上を議論したが, 筆者らは読み上げ音声と同じ GOP 手法で, 「崩れた」シャドーイング音声の自動習熟度評定を検討し, 良い性能が得られた[3]. ここで, 上述した話者適応の効果的利用は, シャドーイングにも適応できる. 次節では, 話者適応技術のシャドーイング評価への応用を実験的に検討する.

## 4. 同話者によるシャドーイングと音読の比較と自動評定

### 4.1. 実験データ

シャドーイングは音読に比べて, 認知負荷の大きいタスクである. 認知負荷を考察するために, 学習者が 3 つのモード: シャドーイング, 文字付きシャドーイング, 音読, で収録できる録音インターフェイスを導入した. さらに, テキストの難易度も考慮し, 各モードに初級・中級・上級 3 段階レベルのタスクを選べら

表 1. 被験者の TOEIC スコア

| 習熟度 | TOEIC スコア                              | 平均  |
|-----|----------------------------------------|-----|
| 上級  | 955, 926, 855, 832, 825, 792, 773, 752 | 838 |
| 中級  | 687, 686, 668, 563, 524,               | 625 |
| 初級  | 496, 425, 399, 378, 252                | 392 |

表 2. GOP スコアと TOEIC スコア間相関 (未適応モデル)

| 難易度 | シャドーイング | テキスト付きシャドーイング | 音読   |
|-----|---------|---------------|------|
| 初級  | 0.74    | 0.65          | 0.48 |
| 中級  | 0.81    | 0.68          | 0.59 |
| 上級  | 0.71    | 0.67          | 0.61 |

表 3. GOP スコアと TOEIC スコア間相関 (大局的適応)

| 難易度 | シャドーイング | テキスト付きシャドーイング | 音読   |
|-----|---------|---------------|------|
| 初級  | 0.74    | 0.68          | 0.60 |
| 中級  | 0.82    | 0.71          | 0.68 |
| 上級  | 0.70    | 0.69          | 0.67 |

れるようにした. 本研究では, 18 名の学習者に, テキスト難易度初級から上級のタスクに対して, この順で, シャドーイング・文字付きシャドーイング・音読の順で評価用音声データを収録した. 学習者の TOEIC スコアの構成を表 1. に示す.

### 4.2. 自動スコア

各タスクにおいて, 各学習者の平均音素 GOP スコアと F-GOP スコアを求めて, 自動スコアとして利用する. 自動スコアを算出するために用いた音響モデルは, 未適応モデルと音読音声の一部を用いて大局的適応を行ったモデルを利用した.

### 4.3. GOP 評価におけるシャドーイング・文字付きシャドーイング・音読の比較

未適応音響モデルで算出した GOP スコア・TOEIC スコア間の相関関係を表 2. に示す. すべての難易度のタスクにおいて, シャドーイングが文字付きシャドーイングや音読より, 良い相関を示している. また, 全体的に, テキスト難易度が中級レベルのシャドーイングタスクは, 初級と上級よりも良い相関を示している.

また, 音読データの一部を用いて, 大局的適応を行った場合の GOP スコア・TOEIC スコア間の相関関係は表 3. にまとめる. 話者適応はシャドーイング音声よりも音読のほうが効果的だが, 各難易度において, シャドーイングは依然として最も良い相関を示している.

### 4.4. F-GOP スコアと TOEIC スコア間の相関

未適応音響モデルで算出した F-GOP スコアと TOEIC

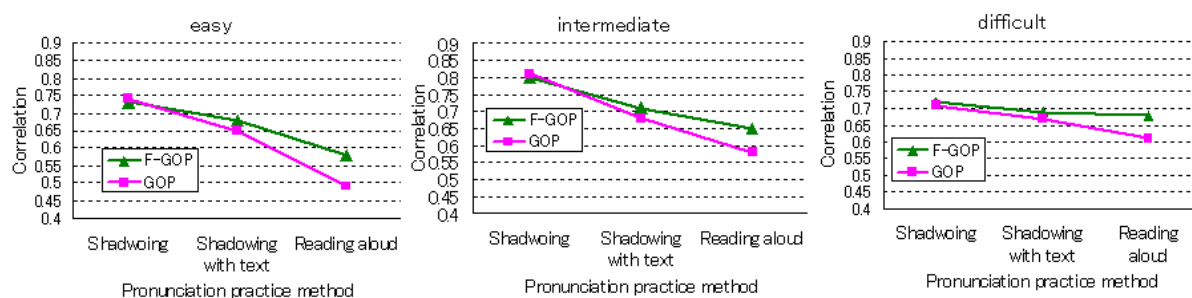


図 6: F-GOP と GOP 評価の比較 (未適応モデルを使用)

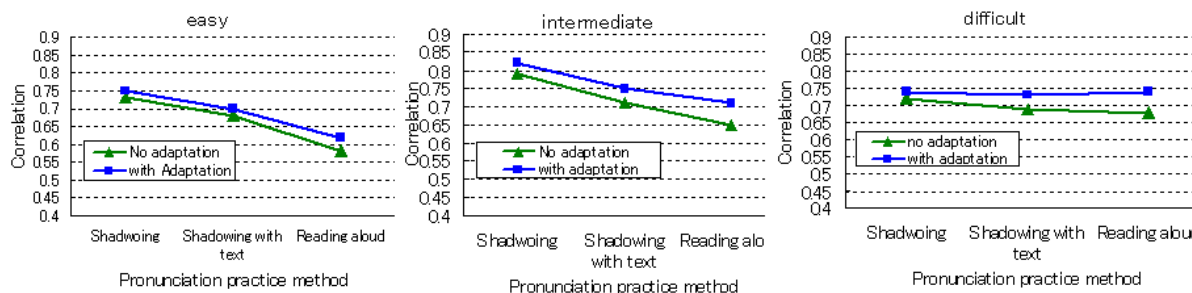


図 7: 大局的 MLLR 適応が F-GOP 評価へ及ぼす影響

スコア間の相関を調べ、GOP スコアとの比較を図 6 に示す。図 7 は、F-GOP 評価において、MLLR 適応を行うか否かの比較を示す。

図 6 に示すように、F-GOP による効果は、シャドーイング評価において、音読ほど顕著ではない。これは、シャドーイング音声に対する強制アラメントの精度は、音読音声より低いと考えられる。また図 7 に示すように、大局的に MLLR 適応は、F-GOP 評価に効果が見られる。

#### 4.5. 考察

すべての難易度タスクにおいて、シャドーイング評価は音読より良い性能を示している。認知負荷を適切に学習者に掛けているシャドーイングは、学習者本来の習熟度をその音声に反映しやすいことを実験的に示した。音読評価の場合に比べて、シャドーイング評価において、話者適応効果は小さい。これは、シャドーイング音声の発話スタイルが音読と大きく異なり、それに起因するミスマッチは、現状の話者適応ではうまく対応できないと考えられる。

#### 5. まとめ

自動スコアリングと誤り検出、2 つの側面から、話者適応が発音評価への影響を調べた。話者適応を積極的に利用する手法として、F-GOP と R-MLLR を提案し、その有効性を実験的に示した。これらの分析結果に踏まえて、同一学習者によるシャドーイング、文字付きシャドーイング、音読音声の分析・評価を行い、適切な認知負荷をかけるシャドーイングが音読より学習者の習熟度を反映しやすいことを実験的に示した。

今後は、一定の時間にわたる繰り返しシャドーイング訓練における効果を実験的に検証する予定である。

#### 文 献

- [1] 門田修平, “シャドーイングと音読の科学”, コスモピア株式会社, 2007.
- [2] 玉井健, “リスニング指導法としてのシャドーイングの効果に関する研究”, 神戸大学大学院総合人間科学研究科博士学位論文, 2001.
- [3] D.Luo et al, “Automatic assessment of utterances shadowed by learners using speech technologies,” Proc. INTERSPEECH, pp.2807-2810 (2008-9)
- [4] S.M. Witt and S.J. Young, “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning,” Speech Communications, 30 (2- 3): pp.95-108, 2000
- [5] A. Raux et al., “Automatic Intelligibility Assessment and Diagnosis of Critical Pronunciation Errors for Computer-Assisted Pronunciation Learning,” Proc. ICSLP, pp.737-740, 2002.
- [6] Minematsu et al, “English Speech Database Read by Japanese Learners for CALL System Development,” Proceedings of International Conference on Language Resources and Evaluation, pp896-903, 2002
- [7] Tanaka et al, “Acoustic models of language-indigent phonetic code systems for speech processing,” Spring meeting of the Acoustical Society of Japan, pp191-192, 2001
- [8] S. Kohmoto, “Applied English Phonology: Teaching of English pronunciation to the Native Japanese Speaker,” Tokyo Tanaka Press, 1965.
- [9] 下村直也他, “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