GOPと重回帰分析を用いたシャドーイング評価の高精度化*

☆加藤集平, 鈴木雅之, 峯松信明, 広瀬啓吉（東大）, 山内豊（東京国際大）

1 はじめに

シャドーイングは，聴取した外国語音声を即座に繰り返して発声することで発音能力と聴取能力とを同時に鍛える外国語聴取・発音訓練法である．もともとは同時通訳者の訓練として広く行われていたが，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ている [1]．シャドーイングにおいては，学習者が提示音声をそのまま真似ることは難しく，学習者自身の話し方の癖や学習者の母語に関する言語知識が無意識のうちに使われることが知られている [2]．

シャドーイングでは提示音声の発話速度に追従する必要があるため，学習者のシャドーイング音声はかなり崩れた，不明瞭なものになることが多い．そのため，シャドーイング音声を評価しようとする場合，人手で評価するには膨大な時間を要する．そこで，発音評価技術を用いてこれを自動評価する手法が提案されている [3]．[3] では，HMM 尤度比に基づいた発音評価スコアである Goodness of Pronunciation (GOP) を学習者の発話を音素単位に区切った音声セグメントごとに計算し，それらを発話した文章全体にわたって平均した GOP_{all} を，発話に対する自動評価スコアとしている．そして，GOP_{all} と学習者の TOEIC スコアの間に高い相関があることを明らかにしている．すなわち，学習者の発話に対して GOP_{all} という 1 つの自動評価スコアを計算し，その GOP_{all} を説明変数とした線形単回帰を行うことにより，学習者の TOEIC スコアを推定できることになる．

本研究ではこれを発展させ，学習者の発話に対して GOP に基づいた複数の自動評価スコアを計算し，それらを説明変数とする線形重回帰を行うことにより，学習者の TOEIC スコアを単回帰の場合よりも高精度に推定することを検討した．また，文章によって各音素の出現回数は異なるので，GOP_{all} は文章への依存度が高いスコアと言える．そのため文章によって TOEIC スコアの推定精度に大きな差が出てしまう可能性がある．そこで，文章への依存度が低いと考えられるスコアを使用して，文章によらず TOEIC スコアを高精度に推定することを検討した．

2 Goodness of Pronunciation (GOP) の計算

発音を自動評価する技術として，HMM に基づいた様々な評価法が提案されている．その 1 つに Goodness of Pronunciation (GOP) と呼ばれる HMM 尤度比に基づいた評価法があり，発音の明瞭度の指標として有効であることが示されている [4, 5]．音素 p であると観測された音声セグメント $O^{(p)}$ に対する GOP は以下の計算式により算出する．

$$GOP(p) = \frac{1}{D_p} \log \left(P(p \mid O^{(p)}) \right) \tag{1}$$

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

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

ここで，$P(p \mid O^{(p)})$ は，観測された音声 $O^{(p)}$ が音素 p による発声である事後確率である．また，Q はすべての音素の集合，D_p は音声セグメント $O^{(p)}$ の継続長である．式 (3) の分子の部分は HMM による強制 Viterbi アライメントにより算出する．分母の部分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．以下では，このようにして算出した $GOP(p)$ を元に，学習者の発声に対する自動評価スコアを計算した．

3 実験

3.1 音声収録

日本語母語話者 49 名および米語母語話者 2 名の計 51 名にシャドーイングを行わせた．シャドーイングに用いた英文は 2 種類で，それぞれ文章 A と文章 B とする．文章 A は 21 文，文章 B は 14 文からなっている．被験者のうち文章 A のシャ

*Improved automatic evaluation of shadowing utterances using GOP and multiple regression analysis. by Shuhei KATO, Masayuki SUZUKI, Nobuaki MINEMATSU, and Keikichi HIROSE (Univ. of Tokyo), Yutaka YAMAUCHI (Tokyo International Univ.)

Table 1　被験者の TOEIC スコア

文章	TOEIC スコア
文章 A	990, 990, 968, 955, 940, 895, 825, 625, 601, 592, 581, 436, 432, 427, 421, 395, 367, 308, 301, 289, 278, 275, 252, 202, 197, 158
文章 B	778, 721, 721, 707, 693, 679, 665, 636, 622, 594, 594, 580, 566, 552, 481, 424, 410, 396, 368, 325, 311, 311, 255, 226

Table 2　HMM の音響分析条件

標本化周波数	16 kHz
量子化ビット	16 bit
窓および窓長	ハミング窓／ 25 ms
シフト長	10 ms
音響特徴量	MFCC12 次元および対数パワー，それらの Δ，ΔΔ（計 39 次元）
音素の種類	ɑː, æ, ʌ, ɔː, aʊ, aɪ, b, ʧ, d, ð, ɛ, ɚː, eɪ, f, g, h, ɪ, iː, ʤ, k, l, m, n, ŋ, oʊ, ɔɪ, p, r, s, ʃ, t, θ, ʊ, uː, v, w, j, z, ʒ, sil, sp（合計 41 種類）

ドーイングを行ったのは 27 名，文章 B のシャドーイングを行ったのは 24 名であり，両方の文章をシャドーイングした被験者はいなかった．被験者の TOEIC スコアを Table 1 に示す．

3.2　音響分析条件

GOP の計算には，WSJ および TIMIT データベースから学習した，英語の音素を単位とする monophone HMM を用いた．HMM の音響分析条件を Table 2 に示す．

3.3　自動評価スコア

シャドーイング音声に対する自動評価スコアは，シャドーイング音声を音素単位に区切った音声セグメントごとに式 (3) により計算した $GOP(p)$ をもとに以下のものを定義した．なお，発話した文章には n 個（n 種類ではない）の音素が含まれ，先頭から順に $p_1 p_2 \ldots p_n$ のように並んでいるものとする．また，音素 sp および sil（どちらも無音を表す）に対する GOP は自動評価スコアの計算には用いない．

GOP_{all}（従来手法）

$GOP(p)$ を発話した文章全体にわたって平均したもの．以下の式で表される．

$$GOP_{all} = \frac{1}{n} \sum_{k=1}^{n} GOP(p_k) \tag{4}$$

GOP_{all} では文章中での出現回数が多い音素の GOP ほど影響が大きくなる．文章によって各音素の出現回数は異なるので，文章への依存度が高いスコアと言える．そのため文章によって TOEIC スコアの推定精度に大きな差が出てしまう可能性がある．

GOP_{vow} および GOP_{cons}

GOP_{vow} は，母音に対する $GOP(p)$ を発話した文章全体にわたって平均したもの，GOP_{cons} は子音に対して同様の計算を行ったものである．文章中に母音が n_{vow} 個，子音が n_{cons} 個含まれるとしたとき，それぞれ以下の式で表される．

$$GOP_{vow} = \frac{1}{n_{vow}} \sum_{p_k \in \text{母音}} GOP(p_k) \tag{5}$$

$$GOP_{cons} = \frac{1}{n_{cons}} \sum_{p_k \in \text{子音}} GOP(p_k) \tag{6}$$

文章によって各音素の出現回数は異なるので，GOP_{all} と同じく文章への依存度が高いスコアと言える．

GOP_{phone}

$GOP(p)$ を音素の種類ごとに平均したもの．例えば音素 ɑːが文章中に $n_{\text{ɑː}}$ 個含まれるとしたとき，$GOP_{\text{ɑː}}$ は以下の式で表される．

$$GOP_{\text{ɑː}} = \frac{1}{n_{\text{ɑː}}} \sum_{p_k = \text{ɑː}} GOP(p_k) \tag{7}$$

文章中の当該音素の出現回数で正規化されるため，文章への依存度が低いスコアと考えられる．そのため GOP_{phone} を説明変数とした回帰式は，文章によらず適用できる可能性がある．

3.4 TOEIC スコアの推定

3.4.1 タスク

3.3 で定義した自動評価スコアをもとに線形回帰を行い，学習者の TOEIC スコアを推定した．推定タスクは以下の 4 種類を行った．

1. 文章 closed, speaker-open（文章 A）
2. 文章 closed, speaker-open（文章 B）
3. 文章 open, speaker-open（文章 A で学習した回帰式で文章 B の被験者の TOEIC スコアを推定）
4. 文章 open, speaker-open（文章 B で学習した回帰式で文章 A の被験者の TOEIC スコアを推定）

1. および 2. は，各文章の全サンプルから 1 話者のサンプルを除外したセットで学習した回帰式を用いて，除外した 1 話者の TOEIC スコアを推定する作業を，全話者について行うものである．3. および 4. については，一方の文章のサンプルで学習した回帰式を用いて，他方の文章の話者の TOEIC スコアを推定するものである．

推定精度の評価は，回帰式による TOEIC スコアの推定値と，実際の TOEIC スコアの相関係数を求めることにより行った．

3.4.2 回帰の種類および説明変数

回帰は最小二乗法による線形単回帰または線形重回帰を行った．説明変数（の組）としては以下の 3 つを用いた．

- GOP_{all}（従来手法）
- $\boldsymbol{GOP_{v+c}} = [GOP_{vow}, GOP_{cons}]^\top$
- $\boldsymbol{GOP_{each\_v}} = [GOP_{\text{ɑː}}, GOP_{\text{æ}}, \ldots, GOP_{\text{uː}}]^\top$
  （各母音種類に対する GOP）

各子音種類に対する GOP を説明変数の組として用いることは，文章 B において「サンプル数 ≤ 子音の種類」となってしまうため行わなかった．

なお，$\boldsymbol{GOP_{each\_v}}$ に対しては，二次の正則化項を導入したリッジ線形回帰も行った．リッジ回帰の正則化パラメータ λ は，予備実験により $\lambda = 50$ に固定した．また，各説明変数は平均 0，分散 1 とする正規化を行い，バイアス項に対する重みについては正則化を行わなかった．

3.4.3 従来手法と提案手法の関係

GOP_{all} を説明変数とする単回帰で TOEIC スコアを推定する従来手法の回帰式は，各音素種類に対する GOP を用いて以下のように解釈できる．

$$\begin{aligned} \hat{TOEIC} =& wGOP_{all} + const. && (8)\\ =& w\frac{n_{\text{ɑː}}}{n}GOP_{\text{ɑː}} + w\frac{n_{\text{æ}}}{n}GOP_{\text{æ}} + \\ & \ldots + w\frac{n_{\text{ʒ}}}{n}GOP_{\text{ʒ}} + const. && (9)\\ (n =& n_{\text{ɑː}} + n_{\text{æ}} + \cdots + n_{\text{ʒ}}) \end{aligned}$$

式 (8) では，発話した文章全体から求められた GOP_{all} に対して重み w がかけられてているが，式 (9) の解釈では，各音素種類に対する GOP に対して，w に当該音素の文章中での出現割合を掛けたものが重みとしてかけられている．すなわち，GOP_{all} は，各音素種類に対する GOP に出現回数に比例した重みづけをして足しあわせたものであると解釈できる．

提案手法は，GOP_{all} が行っているように各音素種類に対する GOP に出現回数に比例した重みづけをするよりも，より高精度に TOEIC スコアを推定できるような重みづけを学習することを意図したものである．また，$\boldsymbol{GOP_{each\_v}}$ は文章への依存度が低いと考えられるスコアであるため，文章によらず高精度に TOEIC スコアを推定できる可能性がある．

4 結果

実験結果を Table 3 に示す．$\boldsymbol{GOP_{v+c}}$ については，文章 closed, speaker-open（文章 A）において GOP_{all} と比べて相関が少し下がったものの，他のタスクにおいては相関が上がった．$\boldsymbol{GOP_{each\_v}}$ については，最小二乗法による重回帰では文章 closed, speaker-open（文章 A）以外のタスクにおいて GOP_{all} と比べて相関が大きく下がってしまったものの，リッジ回帰では文章 open, speaker-open（文章 B → A）においてなお GOP_{all} と比べて相関がやや下回っているが，他のタスクにおいては相関が上がった．

5 考察

$\boldsymbol{GOP_{each\_v}}$ を説明変数とした最小二乗法による重回帰を除いては，ほとんどの場合で GOP_{all}

Table 3　TOEIC スコアの推定値と実際の TOEIC スコアの相関係数

タスク	GOP_{all}	$\boldsymbol{GOP_{v+c}}$	$\boldsymbol{GOP_{each\_v}}$	$\boldsymbol{GOP_{each\_v}}$（リッジ回帰）
文章 closed, speaker-open（文章 A）	0.787	0.766	0.818	0.796
文章 closed, speaker-open（文章 B）	0.723	0.732	0.663	0.748
文章 open, speaker-open（文章 A → B）	0.763	0.790	0.569	0.785
文章 open, speaker-open（文章 B → A）	0.817	0.824	0.229	0.796

と比べて相関が上がった．説明変数の数を増やすことで，よりよい重みづけができたものと考えられる．$\boldsymbol{GOP_{each\_v}}$ が最小二乗法の場合に相関が下がったのは，話者数に対する説明変数の数が他に比べて相対的に多く，過学習が発生したからであろう．

なお，文章 open のタスクは，文章への依存度が高いと考えられる GOP_{all} や $\boldsymbol{GOP_{v+c}}$ を説明変数とした場合でも，使用した両文章で既に 0.8 前後と高い相関が見られた．そして両文章とも，文章への依存度が低いと考えられる $\boldsymbol{GOP_{each\_v}}$ を説明変数としても大幅に相関係数が上がることはなかった．実験に用いた文章がたまたま両文章とも TOEIC スコアを高精度に推定できるような文章だったのか，どんな文章を用いても TOEIC スコアを高精度に推定できるのか，あるいはどのような文章なら TOEIC スコアを高精度に推定できるのかは今回の実験結果からは不明である．

6　おわりに

シャドーイング音声を自動評価し話者の TOEIC スコアを推定する手法の高精度化について検討した．従来の，発話から GOP_{all} という 1 つの自動評価スコアを説明変数とした線形単回帰を行う手法を発展させ，GOP に基づいた複数の自動評価スコアを計算し，それらを説明変数とする線形重回帰を行うことにより，話者の TOEIC スコアを単回帰の場合よりも高精度に推定することを検討した．また，文章への依存度が低いと考えられるスコアを用いて，文章によらず適用できる回帰式を推定することを検討した．結果としては，説明変数に用いるスコアによっては，重回帰（最小二乗法あるいはリッジ回帰）によって単回帰の場合よりも高精度に TOEIC スコアを推定することができた．一方，実験には 2 種類の文章を用いたが，文章への依存度が高いと考えられるスコアを説明変数とした場合でも，両文章間で TOEIC スコアを推定するタスクの精度に大差はなかった．そして，文章への依存度が低いと考えられるスコアを説明変数としても大幅に相関係数が上がることはなかった．

今後の課題としては，単語の難易度，強勢音節中の音素か否かなどの指標を取り入れた新たなスコアを定義し，評価のさらなる高精度化を行うことがあげられる．また，実験に用いた文章がたまたま両文章とも TOEIC スコアを高精度に推定できるような文章だったのか，どんな文章を用いても TOEIC スコアを高精度に推定できるのか，あるいはどのような文章なら TOEIC スコアを高精度に推定できるのかを調査する必要がある．そして，本研究では GOP という音素に関するスコアのみを用いて評価を行ったが，韻律を自動評価する仕組みを取り入れることも考えている．

参考文献

[1] 門田修平，“シャドーイングと音読の科学,” コスモピア，2007.

[2] P. W. Nye *et al.*, “Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English, ” *Journal of Phonetics*, 63–69, 2003.

[3] D. Luo *et al.*, “Automatic pronunciation evaluation of language learners’ utterances generated through shadowing, ” *Proc. INTERSPEECH*, 2807–2810, 2008.

[4] S. M. Witt *et al.*, “Phone-level pronunciation scoring and assessment for interactive language learning, ” *Speech Communication*, 30 (2-3), 95–108, 2000.

[5] L. Neumeyer *et al.*, “Automatic scoring of pronunciation quality, ” *Speech Communication*, 30 (2-3), 83–93, 2000.