社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

外国語学習を対象としたシャドーイング音声の自動評定法に関する検討

羅 徳安† 下村 直也† 峯松 信明† 山内 豊‡ 広瀬 啓吉†

†東京大学 〒113-0033 東京都文京区本郷 7-3-1
‡東京国際大学商学部 〒350-1197 埼玉県川越市的場北 1-13-1
E-mail: †{dean,shimo,mine,hirose}@ gavo.t.u-tokyo.ac.jp, ‡ yyama@tiu.ac.jp

あらまし 近年，英語教育において，シャドーイングが注目されている．シャドーイングとは，外国語音声を聞きながらほぼ同時にその発話を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訓練方法である．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崩れ，不明瞭になる場合が多い．したがって，シャドーイング音声を正確に評価することは非常に困難なタスクとなる．本研究では，提示音声の書き起こしが入手可能な場合の手法として，音響モデル（HMM）に基づく Goodness of Pronunciation (GOP)手法及び，書き起こしが入手不可能な場合の評価手法として，クラスタリングに基づく自動評定の 2 種類の手法を検討した．この 2 種類の手法を比較し，自動評定と手動スコア及び，自動評定と TOEIC スコアとの相関関係を算出した．実験の結果，自動評定と手動スコア及び，自動評定と TOEIC スコアとの間に良好な相関を観測することができた．

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

A Study on Automatic Scoring Methods for Language Learners' Shadowing Productions

Dean LUO† Naoya SHIMOMURA† Nobuaki MINEMATSU† Yutaka YAMAUCHI‡ and Keikichi HIROSE†

†The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 105-0123 Japan
‡Tokyo International University, 1-13-1 Matobakita Kawagoe, Saitama, 350-1197 Japan
E-mail: †{dean,shimo,mine,hirose}@ gavo.t.u-tokyo.ac.jp, ‡ yyama@tiu.ac.jp

Abstract Shadowing has been used as a method for improving speaking and listening ability that requires learners to repeat, or shadow, a presented native utterance as closely and as quickly as possible. Since learners have to follow the speaking rate of the input native utterance, especially in the case of beginners, their pronunciation often becomes inarticulate and corrupt. These features of shadowing make it very difficult to build a reliable scoring system for shadowed utterances. In this paper, we investigate the automatic pronunciation scoring methods for shadowing. HMM-based and clustering-based pronunciation evaluation techniques are proposed and the relationship between automatic scores and the learners' overall English proficiency is investigated. Experiments show that good correlations are found between the automatic scores and human scores or TOEIC overall proficiency scores.

Keyword shadowing, goodness of pronunciation, automatic scoring, articulatory effort, CALL

1. はじめに

コミュニケーション力を重視する昨今の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],[3],[4].

学習初期段階の日本人が英語を発声すると，カタカナ英語と呼ばれる発音となることがある．認知心理学的には「英単語の発音が，日本語の音韻に変換された状態で，長期記憶中の心的辞書（メンタルレキシコン）に保持されていることに起因する」と考えられている．シャドーイングは，心的辞書から語彙情報を検索する時間を十分に与えずに発声を要求するため，入力音声の音的イメージをそのまま再生させることに繋がり，母語の音韻体系に引きずられることなくスピーキング能力を向上させることができる，と考えられている．さらに，シャドーイングはリスニング能力の向上ももたらす．リスニングは「知覚」と「理解」から構成されているが，両段階において，認知資源を消費する．シャドーイングは，母語話者の発音を繰り返して聞くことで音声知覚過程を鍛え／自動化し，同時に，スピーキングを通して正確な発音（音的イメージ）を心的

辞書に定着させることで，理解の段階により多くの認知資源を割り当てられるようになる．これらの結果，リスニング能力についても，その向上が期待できる[1],[2].

このようにシャドーイングは，スピーキング／リスニング能力を同時に訓練できるため，コミュニケーション能力を重視する近年の外国語学習において広がりを見せている．学習意欲維持のためには学習者が自らの習熟度を把握し，また教師側は，学習者発声を短時間で評定し教示する必要がある．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，発音評定技術を用いた自動化が望まれるところである．

本研究では，２つの条件を考慮して，シャドーイング音声の自動評価について検討する．1 つは，提示音声の書き起こしが入手可能な場合である．もう一つは書き起こしが入手不可能な場合である（たとえば，提示音声が何語であるか不明の場合も含む).実験の結果，提示音声の書き起こしを利用する場合の手法として，ＨＭＭ音響モデルによる自動評定スコアは，教師による手動スコアとの相関は発話単位では 0.85，話者単位では 0.94, TOEIC スコアとの相関は 0.84 と良好な値を観測した．また，提示音声の書き起こしを利用しない場合の手法として，教師なしクラスタリングに基づく自動評定スコアは，HMM に基づく GOP スコアとの相関が発話単位では 0.75, 文単位では 0.87 と強い相関を得られ，手動スコアとの相関は，発話単位では 0.79，話者平均では 0.92 となり，TOEIC スコアとの相関は 0.72 と比較的良好な相関を示しており，本研究で提案した言語非依存モデルの有効性を示している．

2. HMM に基づく自動評定

2.1. Goodness of Pronunciation (GOP) 評定

既存の発音支援システムでよく使われている発音評価技術として，さまざまなＨＭＭに基づく自動評定法が提案されている．GOP(Goodness of Pronunciation)と呼ばれる HMM 尤度比ベースの評定法が，読み上げ音声に対して，発音の明瞭度の指標として有効であることは，多くの研究において示されている[5],[6].

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．

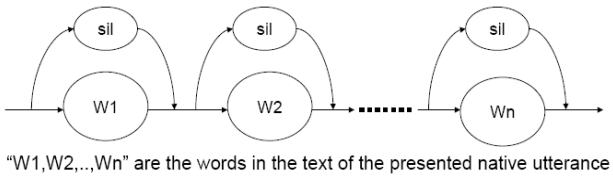

図 1: 「黙り」を検出するためのネットワーク文法

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

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

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

$P(p \mid O^{(p)})$ 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 $GOP(p)$を自動評定スコアの一つとして利用する．

2.2. 明瞭に発声した単語の検出法

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には，単語レベルの脱落や言い間違いが多く見られている．収集した学習者のシャドーイング音声データを考察した結果，2 種類の典型的な誤りパタンが現れている．一つは，学習者が聞き取れた単語のみ，はっきり復唱し，それ以外の単語に対して完全に黙っている傾向である．もう一つは，学習者が提示音声のスピードに追従するために，とにかく「音」を出しているだけであり，どの単語を発声しているのか完全に判別できなくなる傾向である．

第 1 種類の誤りに対して，本研究では図 1 で示すような単語単位のネットワーク文法を導入し，「黙り」に置き換えられた単語を検出する．これらの「黙り」を除いた後，さらに第 2 種類の誤り傾向の対策として，認識された単語セグメントの平均音素 GOP を求め, その GOP スコアが事前に設定した閾値以上の単語のみ，明瞭に発声した単語として定義する．このように「明瞭に発声した単語」の数(Number of Proficiently Pronounced Words, NPPW) を数えて，提示音声に含まれる語数で割った値を学習者の習熟度の指標として利用する．

3. クラスタリングに基づく評定法

シャドーイングは学習者にとって，非常に負荷の高いタスクであり，特に初学者の場合，シャドーイング音声は非常に崩れた／不明瞭な音声となる．外国語発音の自動評定を行なう場合，通常，母語話者の読み上げ音声から構築された HMM を用いて行なうことが多い．しかし，1)発話スタイルが読み上げ調と大きく異なる，2)任意の外国語のシャドーイング音声評定を行なうためには，提示音声及びシャドーイング音声のみを用いて評定を行なえることが望ましい．これらの点を考慮し，提示音声の書き起こしや対象言語の音響モデルなどを一切必要としないシャドーイング音声自動評定手法を提案する．

3.1. 時間制約付きボトムアップクラスタリング

連続音声から自動的に音素境界を求める研究が広く行われている．入力音声の書き起こしや対象言語の音響モデルなどの事前知識を与える場合，それを教師ありのセグメンテーションと言う．一方，教師無しセグメンテーションの研究も行われており，その場合多くは，局所的なスペクトル変化の大きい時点をセグメント境界とする方法が多い[7],[8]．これに対して筆者らはスペクトル的に類似し，かつ，隣接して存在するフレームをマージする形で纏め上げ（ボトムアップクラスタリング），音声ストリームの大局的な階層構造を抽出する形で教師無しセグメンテーションを実装した．

更にマージ対象となる 2 セグメント（クラスタ）の選択基準として種々の統計量に基づく尺度を検討し，提案手法が従来の局所的なスペクトル変化に基づく手法よりも優位であることを示した[9]．なお，本稿では実装が容易な Ward 法によるボトムアップクラスタリングに基づく手法[10]を採用する．

Ward 法はユークリッド距離に基づくクラスタリングであり，2 つのセグメントをマージした際の「群内偏差平方和の増加量」を両者の非類似度と定義している[11],[12]．この非類似度が最も低い（即ち，最も類似している）セグメント同士をマージさせ，最終的には 1 つのセグメントへと纏め上げる．今，時間的に隣り合うセグメント p，$p+1$ をマージして新しいセグメント $r\ (=p\cup(p+1))$を作ることを考える．p の偏差平方和を $E(p)$とおくと，p，$p+1$ をマージし，r を生成した際の偏差平方和の増分 $\Delta E(p,p+1)$ は

$$\Delta E(p,p+1)=E(r)-\{E(p)+E(p+1)\} \qquad (4)$$

となる．各段階でマージによる偏差平方和の増分 $\Delta E(p,p+1)$ が最小となる p と $p+1$ をマージする．以上の操作は，フレームの部分系列をより大きな纏まりとして捉えることに相当する．

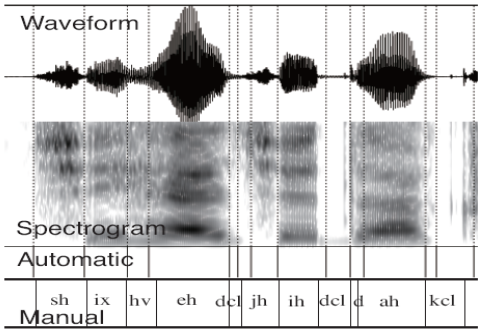

図 2: 自動セグメンテーションの一例

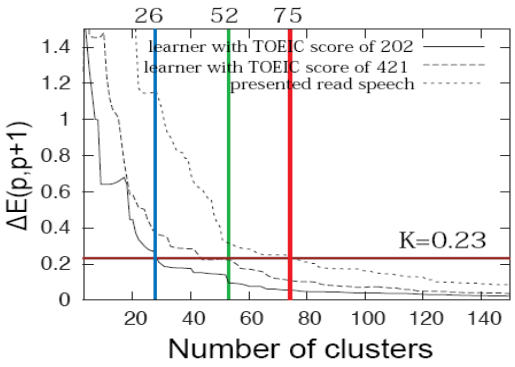

図 3: シャドーイング音声の分析結果

3.2. クラスタリングの停止条件

ある段階において，各セグメントが凡そ各音素に対応している状態を考える．この場合，次のマージ操作は異なる音素を強引にマージすることを意味する．ある話者が生成する各音素の音響的特徴を考える．この場合，任意の 2 音素間距離（重心間距離）の最小値は，凡そ話者非依存であると仮定する．その結果，どの話者の発声した音声であっても，音素に対応した形でクラスタリングされた状態に対する次のマージ操作は，比較的大きな更新コスト（群内偏差平方和の増加量）を呈するはずである（$E(p\cup p+1)>>E(p)+E(p+1)$）．

そこで，$\Delta E(p,p+1)$ に対応する閾値 K を定め，各セグメントが凡そ音素に対応したときに自動的に停止する方法を検討した[10]．具体的な自動セグメント例を図 2 に示す．

図 3 は，提示音声及び，その音声を TOEIC421 点，202 点の英語学習者がシャドーイングした発声をボトムアップクラスタリングした結果である．音声横軸にクラスタ数，縦軸に Ward 法の更新コスト，つまり $\Delta E(p,p+1)$ をプロットした．閾値 K=0.23 でクラスタリングを停止させた時のクラスタ数を見てみると，提示音声が 75 クラスタ，TOEIC421 点話者の発声が 52 クラスタ，TOEIC202 点話者の発声は 26 クラスタとなっている．同一文を発声した場合でも，個々の音を明瞭に区別（調音）して発声した場合，停止時のクラス

タ数は多く，個々の音が不明瞭に発声されれば，停止時のクラスタ数は少なくなる．

3.3. 音響事象間距離と調音努力

調音努力とは，個々の音を区別して調音するために行なうべき調音運動量と解釈される量である．例えば母音構造を考えれば，その中心には弱母音，即ち，最も脱力した状態で発声される母音が位置しており，その他の母音は，その母音を発声すべく調音努力を払って声道形状を制御して生まれる音である．事象間距離に対する考察は，読み上げ音声／話し言葉音声の間でも行なわれており，当然話し言葉の方が「なまけ」などの理由で事象間距離が小さくなる[13]．これらを考慮すると，事象間距離の大小を通して，発声時に払われた調音努力の大小を推定することは十分妥当である．

結局，適切な固定閾値の下でクラスタリングを停止させ,その時のセグメント数の大小を議論することは,その発声において払われた調音努力の大小を推定することに相当する．言い換えれば，与えられた発声に対して，どの程度「滑舌の良い」「呂律の回った」発声であったのかを推定することになる．シャドーイング音声は,習熟度が低ければ「もごもごした」音声であり,高ければ「はきはきした」音声となることを考えると,筆者らが提案する教師無しセグメンテーションは，シャドーイング音声の評定に対して相性が良いと言える．

4. シャドーイング音声の自動評定実験

4.1. シャドーイング音声の収録と手動評定

日本人学習者 27 名にシャドーイングを行なわせた．TOEIC テスト（990 点満点）における上位 7 名，中位者 9 名，下位者 11 名を対象とした．彼らの TOEIC スコアを表 1 に示す．シャドーイングは初めて，或は，ほぼ未経験の被験者を採択した．

シャドーイング用に提示した音声は，1 名の男性米語母語話者が読み上げた音声であり,全 21 文からなる平易な速読用の英文である．日本の魚に関するトピックで内容的に親密度は高いが，全ての被験者にとって初見である．平均話速は 140 語/分であった． 27 名による合計 567 発話のシャドーイング音声を実験に用いた．収録で用いた教室では空調等の定常雑音が随時発生しており，提示音声の収録環境とは異なる．

手動による評定作業は，小・中・高校，及び，大学で英語授業を実践してきた英語教育の専門家（第四著者）が代表的な話者 11 名（上級 4 名，中級 3 名，初級 4 名）による 10 文を選び，採点を行なった．各発声に対して，その単語として意図された単語発声の個数を数え上げ,その文に含まれる語数で割った値（百分率）を，その発声のスコアとした．ただし，言い直した場合に言い直す前の単語は不要語として-1 語，動詞や複数語尾の誤りは-0.5 語として評定した．1 名分（約 2 分の音声データ）の評定に 1 時間程の時間を要したことを明記しておく．

4.2. 音響分析及びクラスタリングの諸条件

ＨＭＭに基づく手法における音響分析条件は表 2.に示す．クラスタリングに基づく手法の各種の分析条件を表 3.にまとめる．スペクトル変化を捉える音響特徴量として，聴覚特性を考慮したメルケプストラムを用いた．なお本稿では，スペクトルの安定区間を纏めることでセグメント数を得ることを検討していること，そして，収録されるシャドーイング音声によってはパワーが大きく異なることから，パワー項（MCEP の 0 次項）は用いていない．クラスタリングの停止条件である閾値 K は，TIMIT train パートの全 4620 発話に対して,正解音素数と自動推定音素数との相関が 0.83 と最も高かった K=0.23 を用いた．

表 1. 被験者の TOEIC スコア

習熟度	TOEIC scores	Average
上級	990, 990, 968, 955, 940, 895, 825	938
中級	625, 601, 592, 581, 512, 436, 432, 427, 421	514
初級	395, 367, 308, 301, 289, 278, 275, 252, 202, 197, 158	275

表 2. HMM 手法における音響分析条件

サンプリング	16bit / 16kHz
窓及び窓長	ハミング / 25 ms length
シフト長	10 ms shift
音響パラメータ	MFCC12 次元と対数パワー，それらの Δ，ΔΔ（計 39 次元）

表 3. クラスタリング手法における音響分析条件

サンプリング	16bit / 16kHz
窓及び窓長	ハミング / 16 ms length
シフト長	10 ms shift
音響パラメータ	MCEP (1～12)
閾値	K=0.23

4.3. 自動評定スコア

HMM に基づく自動評定スコアは，平均音素 GOP スコアそのものと,GOP を利用して検出した NPPW を提示音声の語数で正規化したスコアを用いた．クラスタリングに基づくシャドーイング音声自動評定法は，自動停止した時のセグメント数を算出し，そして「提示音声のセグメント数」に対する「発声者が生成したセグメント数」を百分率で算出し，これをその発声の自動評定スコアとした．

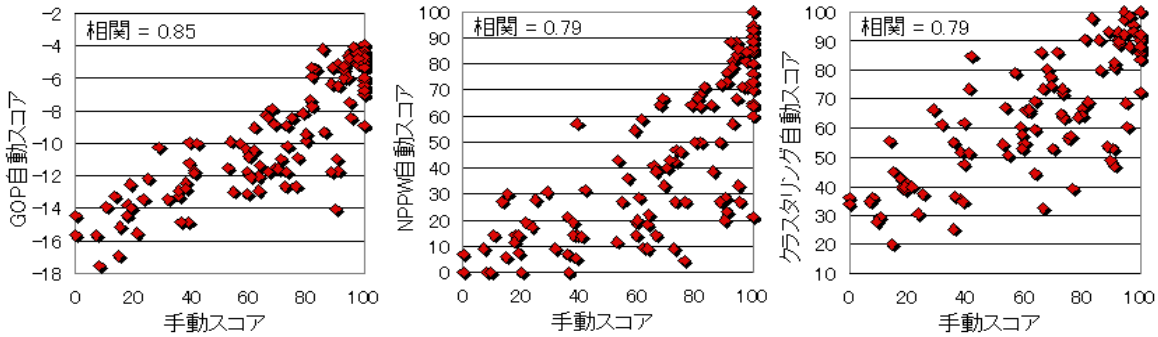

図.5: 発話単位での自動スコアと手動スコアとの相関関係

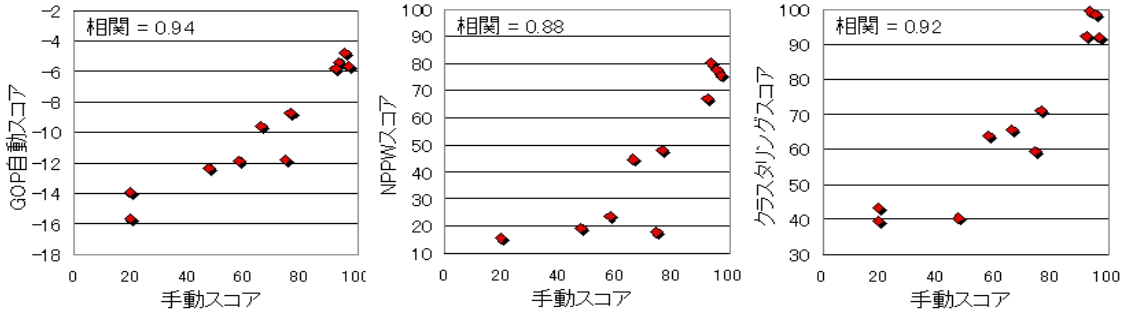

図. 6: 話者単位での自動スコアと手動スコアとの相関関係

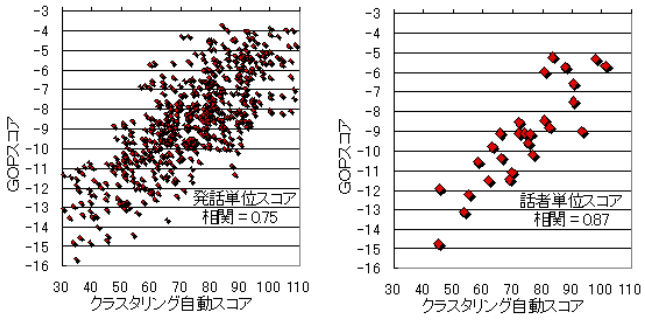

図 4:発話単位と話者単位での GOP とクラスタリングスコアとの相関関係

4.4. HMMに基づく評定法とクラスタリングに基づく評定法との比較

既に述べたように，GOP は発音の明瞭度を示す良い指標であり，また，筆者ら提案しているクラスタリングに基づく手法では，クラスタリングが停止する時のセグメント数の大小も，その発声において学習が払われた調音努力の大小を推定できる．したがって，両者は異なる手法で求めた評定基準でありながら，対等な役割を果たしていると考えられる．ここで，両者の相関関係を調べた．その結果，図 4.に示すように，文単位での相関は 0.75, 話者単位での相関は 0.87 と強い相関が観測された．

4.5. 自動スコアと手動スコアとの相関関係

手動による評定は 4.1 節で述べた方法で，11 名話者による 110 発話のシャドーイング音声対して，採点を行った．各話者の構成は表 4.に示す．

HMM に基づく評定法(GOP,NPPW)とクラスタリングに基づく評定法で評価した自動スコアと手動によるスコアとの相関を調べた。発話単位での相関関係は図 5 に示すように，GOP スコアは最も良い相関 0.85 を示し，NPPW とクラスタリングに基づく評定スコアも 0.79 と良好な相関を示している．話者単位では，図 6 に示すように，GOP スコアが 0.94，NPPW が 0.88，クラスタリング自動評定が 0.92 と非常に高い相関が観測された。

表 4. 手動評定データの話者構成

習熟度	TOEIC scores	Average
上級 4 名	968, 955, 940, 895	940
中級 3 名	432, 427, 421	427
初級 4 名	301, 252, 202, 197	238

4.6. 自動スコアと TOEIC スコアとの相関関係

27 名被験者全員による全 21 文（計 567 発話）のデータを HMM 及びクラスタリングに基づく評定法で評価した．TOEIC スコアは話者単位のスコアであるため，得られた自動評定スコアを話者ごとに平均を求め，学習者の TOEIC スコアとの相関関係を調べた．その結果，NPPW がもっとも良い相関 0.84 を示し，言語情報によらないクラスタリングに基づく評定スコアも 0.72 の比較的に良い相関を示している．自動評定スコアと TOEIC スコアとの相関関係は図 7 に示す．

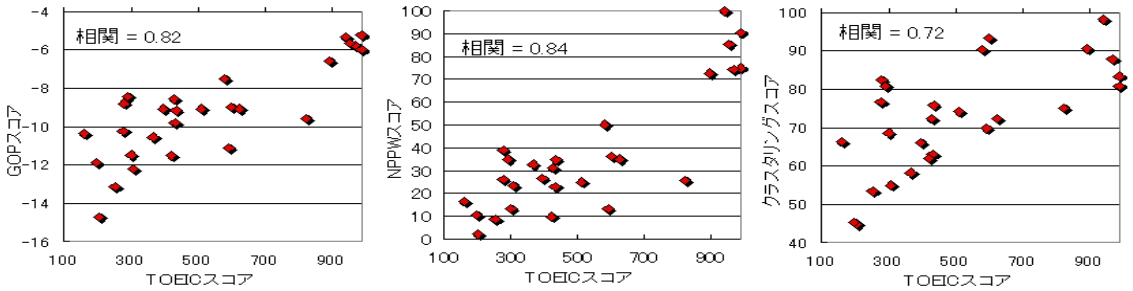

図.7: 発話単位での自動スコアと TOEIC スコアとの相関関係

5. 全体の考察

シャドーイング音声の自動スコアと手動スコア及び TOEIC スコアとの間に強い相関が観測された. 通常の読み上げ音声を評価する近年の研究例では，本研究と同様な HMM に基づく GOP 手法を用いていても, 手動採点と間の相関が本研究の結果よりかなり低い．例えば，[14]では 0.56, [15]では 0.61 と報告されている．この理由は，シャドーイングは認知的負荷を学習者に適切に掛けているタスクであるため，学習者本来の習熟度がその発声に反映され易いと考察することができる．今後は，同じ話者による読み上げ音声を収録／分析し，シャドーイング音声との違いを考察したい．

6. まとめ

シャドーイング音声の評定法として，HMM に基づく手法とクラスタリングに基づく評定法を提案した．HMM に基づく手法は，手動スコア及び TOEIC スコアとの間に，非常に高い相関関係を示している．HMM 評定法に及ばないが，言語情報に依存しないクラスタリングに基づく手法も比較的良い性能を示しており，また，HMM に基づく GOP 評定スコアとの間に高い相関が観測された．クラスタリングに基づく手法は，提示音声とシャドーイング音声のみを用いており，対象言語の事前知識は全く必要としない。

今後の課題として, 次のことが挙げられる．1）HMM に基づく技術とクラスタリングの基づく技術を融合し，より高い信頼性と実用性もつハイブリッドシステムを提案する．2）読み上げ音声，そして，オフラインの復唱音声を収録し，それらとシャドーイングとの違いを考察する．3）1 回目のみのシャドーイング音声ではなく，繰り返して練習した場合の学習者音声を分析し，習熟度の変化を考察する．4）ピッチ，イントネーションを含む韻律情報を評価に取り入れていく．5）シャドーイング一言で言っても，学習者が実際に行っていることは多様なので，異なるシャドーイングタイプをどう扱うかを検討する．

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] S.Miyake, “Cognitive processes in phrase shadowing and EFL listening,” JACET Bulletin Tokyo: Japan Association of College English Teachers. Forthcoming

[4] T.Hori, “Exploring Shadowing as a Method of English Pronunciation Training,” A Doctoral Dissertation Presented to the Graduate School of Language Communication and Culture, Kwansei Gakuin University. 2008

[5] 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

[6] L.Neumeyer et al., “ Automatic scoring of pronunciation quality,” Speech Communications, 30(2-3): pp.83-93, 2000.

[7] S. Dusan et al., “On the relation between maximum spectral transition positions and phone boundaries,” Proc. InterSpeech,pp.17–21, 2006.

[8] Y. P. Estevan et al., “Finding maximum margin segments in speech,” Proc. ICASSP, pp.937–940, 2007.

[9] Y. Qiao et al., “Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons,” Proc.ICASSP, 2008.

[10] 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

[11] 宮本定明，“クラスタ分析入門ファジィクラスタリングの理論と応用”，森北出版, 1999.

[12] C.Hervada-Sala et al., “A program to perform Ward’s clustering method on several regionalized variables,” Computers & Geosciences 30, pp.881-886, 2004

[13] M. Nakamura et al., “Acoustic and linguistic characterization of spontaneous speech,” Proc. Int.Workshop on Speech Recognition and Intrinsic Variations, pp.3–8, 2006.

[14] Abhishek Chandel et al., ” Sensei: Spoken Language Assessment for CALL Center Agents,” Proc. ASRU, pp.711-716, 2007

[15] J. Zheng et al., “Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation,” Proc. ICASSP,vol.4, pp.201-204, 2007