

シャドーイング訓練法における外国語学習者の発音分析とその自動評価*

○羅徳安, 峯松信明 (東大), 山内豊 (東京国際大), 広瀬啓吉 (東大)

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の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価することは膨大な時間を要するため、発音評価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、対象言語の音響モデルが入手可能な場合の手法として、HMM 尤度に基づく Goodness of Pronunciation (GOP) 手法と連続音素認識に基づく手法及び、音響モデル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価の3種類の手法を検討した。この3種類の手法を比較し、自動評価と手動スコア及び、自動評価と TOEIC スコアとの相関関係を算出した。実験の結果、自動評価と手動スコア及び、自動評価と TOEIC スコアとの間に良好な相関を観測することができた。

2 HMM に基づく自動評価

2.1 Goodness of Pronunciation(GOP)評価

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、隠れマルコフモデル(HMM)に基づく様々な自動評価法が提案されている。GOP と呼ばれる HMM 尤度比ベースの評価法が、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[3]。

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価の際に参照する 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 による発声である事後確率である。 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 $GOP(p)$ を自動評価スコアの一つとして利用する。

3 クラスタリングに基づく評価法

外国語発音の自動評価を行なう場合、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)シャドーイングの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価を行なうためには、提示音声及びシャドーイング音声のみを用いて評価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価手法を提案する。

* Analysis and Automatic Scoring of Language Learners' Pronunciation in Shadowing, by Dean Luo, Naoya Shimomura, Nobuaki Minematsu (The University of Tokyo), Yutaka Yamauchi (Tokyo International University) and Keikichi Hirose (The University of Tokyo).

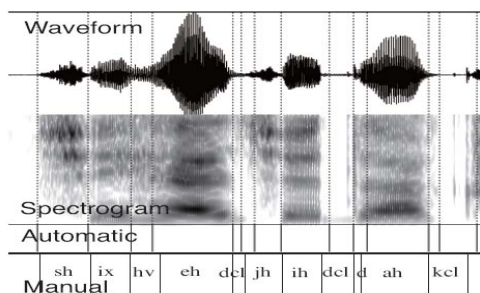


図 2: 自動セグメンテーションの一例

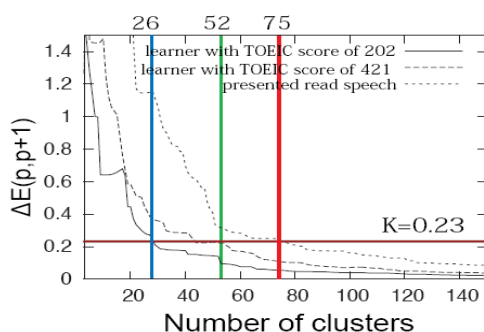


図 3: シャドーイング音声の分析結果

3.1 時間制約付きクラスタリング

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[4]。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[5]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[6]。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。

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。

そこで、更新コストに対応する閾値 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。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。

4 シャドーイング音声の自動評価実験

4.1 シャドーイング音声の収録と手動評価

日本人学習者 27 名にシャドーイングを行なわせた。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。

手動による評価作業は、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 (第三著者) 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、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 語として評価した。1 名分 (約 2 分の音声データ) の評価に 1 時間程の時間を要したことを明記しておく。

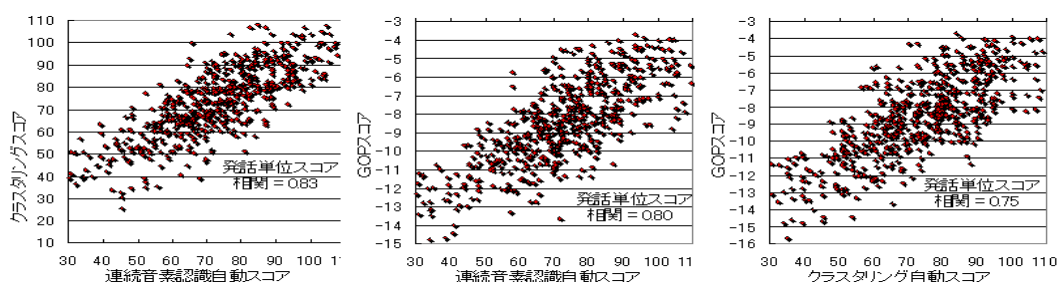


図 4: 発話単位での自動スコア間の相関関係

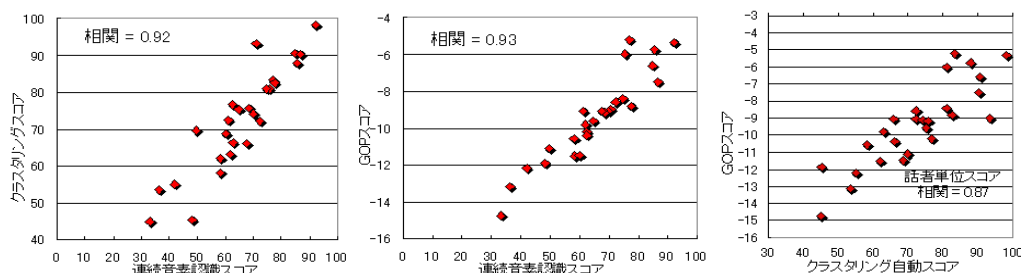


図 5: 話者単位での自動スコア間の相関関係

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. 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.

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 スコア以外にも, 連続音素認識で出力した音素数を提示音声の音素数で正規化したスコアも検討した. 後者はクラスタリング同様, 異なる音が発声の中に幾つ存在するのかわを示す指標である. 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.

4.4 音響モデル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

GOP スコア, 連続音素認識スコア及びクラスタリングに基づくスコア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, 三者の相関関係を調べた. 文単位では図 4, 話者単位では, 図 5 に示す. 特に話者単位では, 非常に高い相関が観測されている.

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,

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 , $\Delta \Delta$ (計 39 次元)

表 3. クラスタリング手法音響分析条件

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	$K=0.23$

11 名話者による 110 発話のシャドーイング音声に対して, 採点を行った. 発話単位での相関関係は図 5 に示すように, GOP スコアと連続音素認識スコアは最も良い相関 0.85 を示し, 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, 連続音素認識スコアが 0.97, 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.

4.6 自動スコアと TOEIC スコア間の関係

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

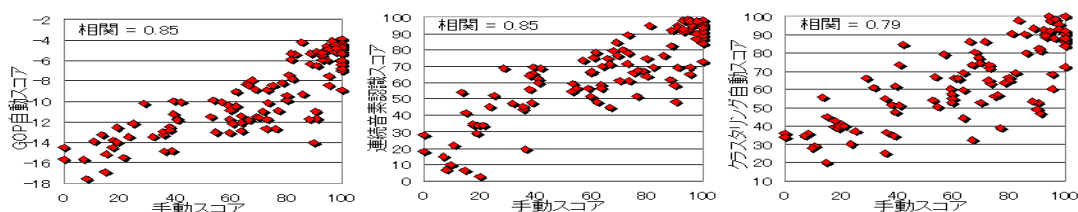


図.6: 発話単位での自動スコアと手動スコアとの相関関係

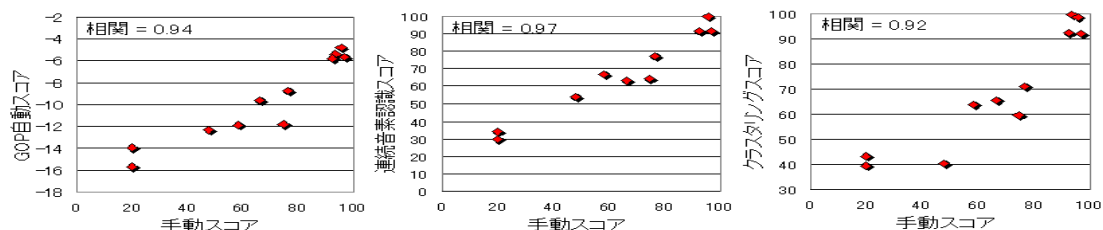


図 7: 話者単位での自動スコアと手動スコアとの相関関係

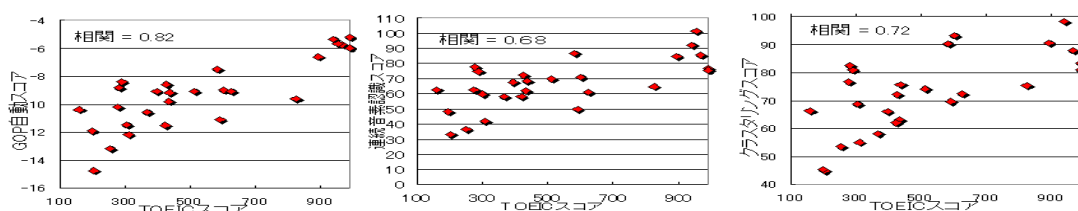


図 8: 話者単位での自動スコアと TOEIC スコアとの相関関係

TOEIC スコアとの相関関係を調べた (図 8)。その結果, GOP スコアが最も良い相関 0.82, 連続音素認識スコアが 0.68 を示し, 言語情報によらないクラスタリングに基づく評価スコアも 0.72 の比較的良い相関を示している。

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の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, [7]では 0.56, [8]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。

6 まとめ

シャドーイング音声の評価法として, HMM に基づく手法とクラスタリングに基づく評価法を提案した。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 評価法に及ばないが, 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 評価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。

今後は, 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予定である。

参考文献

- [1] 門田修平, “シャドーイングと音読の科学”, コスモビア株式会社, 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.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
- [4] Y. Qiao et al., “Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons,” *Proc. ICASSP*, 2008.
- [5] 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
- [6] 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.
- [7] Abhishek Chandel et al., “Sensei: Spoken Language Assessment for CALL Center Agents,” *Proc. ASRU*, pp.711-716, 2007
- [8] J. Zheng et al., “Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation,” *Proc. ICASSP*, vol.4, pp.201-204, 2007