教師無しセグメンテーションを用いたシャドーイング音声の自動評定に関する実験的検討*

◎下村直也, 峯松信明（東大）, 山内豊（東京国際大）, 喬宇, 朝川智, 広瀬啓吉 （東大）

1　はじめに

近年，言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，聴取した（母語話者により発声された）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（この場合，故意に delay を置いて通訳するなど，認知的により高いタスクを要求する），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]．

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，母国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 [1, 2]．

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大きい訓練であり，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．しかしシャドーイング音声は，従来の評定技術が対象としてきた比較的「奇麗な」音声とはかなり異なる．筆者らの知る限り，シャドーイング音声を対象とした自動評定手法は提案されていない．

筆者らは，笑い声や鳴き声といった「崩れた」音声をも対象としうる技術として，ボトムアップクラスタリングに基づく教師無しセグメンテーションを検討してきた [3, 6]．本稿では，この技術をシャドーイング音声の自動評定に応用した．その結果，シャドーイング対象の文を適切に選定することで，自動評定スコアと教師による手動評点スコアとの相関は 0.70 と，比較的良好な値を示した．特に TOEIC 中位者と低位者とを明確に分離する結果を示した．

2　従来のシャドーイング音声の評定方法

初期段階の英語学習者のシャドーイング音声は，非常に崩れた発声となる．耳から入る音声の知覚が十分に自動化されておらず，更には，英語の各音韻・音節を生成するための調音運動が十分に習得できていないため，時に言い黙り，言い淀みが生じる．逆に，熟練者のシャドーイング音声は調音努力が十分になされた流暢／明瞭な音声となる．このような発声の差異を捉えるシャドーイング音声評定方法が玉井によって 2 種類提案されている（音節法及びチェックポイント法 [2]）．また，これらの手法の問題点を考察し，3 つ目の手動評定方法として「全単語法」を考える．

2.1　音節法

音節法とは，英語における発話の最小単位と考えられている音節毎の正誤を判定する方法である．素材となる外国語テキストの書き起こしをもとに，1 音節語はそのままに，2 音節以上の単語は各音節毎に分け，評定を行なう．評定単位が単語より小さく，評定の信頼性が保たれる．しかし，採点者は音声を音節毎に区分化，評定する必要があり，時間的コストや体力的負担を被ることとなる．そのため，必ずしも実用性が高い方法ではないと筆者らは考える．

2.2　チェックポイント法

音節法と違って，単語毎に評定する簡便法として，チェックポイント法が提案されている．英語テキストの全単語を n 単語毎に，その単語が正しく発声できているかどうかを判定する．[2] では，全単語が 350

*Experiment study of automatic scoring of utterances generated through shadowing by using unsupervised segmentation techniques, by N. Shimomura, N. Minematsu (Univ. of Tokyo), Y. Yamauchi (Tokyo International Univ.), Y. Qiao, S. Asakawa, K. Hirose (Univ. of Tokyo)

語以上で，各文が 8 単語程度の長さで構成されている場合に n=5 を採用している．この場合，音節法との評定結果の相関として 0.89 が示されている [2].

2.3 全単語法

音節法は一見精度が高そうに見えるが，because を become とシャドーイングした場合，音節法では 50 %の正答を与えることになり，チェックポイント法では 0 %になる．because を become とシャドーイングするのは，明らかに単語を取り違えており，英語をコミュニケーション・ツールとして捉え，実践的コミュニケーション能力の養成を目的とする，近年の英語教育の方向性と乖離している．また，n の設定方法は十分に明らかとなっていない．これらを考慮し，本稿で行なう手動の評定は n=1，即ち全単語に対して「その単語が発声できているか」を判定し，手動の評定スコアとした．この場合，その単語として意図されたと思われる発声であれば，正解として判定している．

3 時間制約付きボトムクラスタリングによる教師無しセグメンテーション

3.1 時間制約付きボトムアップクラスタリング

教師無しセグメンテーションの先攻研究の多くは，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 [4, 5]．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 [3]．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 [6]，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 [6]．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 [3] 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 [7]．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，セグメント p と隣り合うセグメント $p+1$ 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

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

となる．各段階でクラスタの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．これにより，フレームの部分系列をより大きな纏まりとして捉えていくことが可能となる．

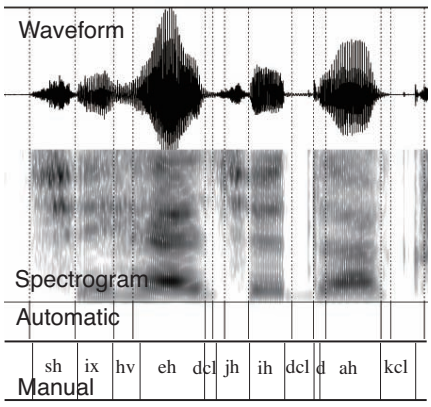

Fig. 1 自動セグメンテーションの一例

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（$E(p\cup p+1) \gg E(p)+E(p+1)$）．

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 [3]．具体的な自動セグメント例を図 1 に示す．閾値 K を固定して，様々な音声を自動セグメントすると次のことが分かる．同一文を発声した場合でも，個々の音を明瞭に区別（調音）して発声した場合のセグメント数は多く，個々の音が不明瞭に発声されれば，セグメント数は少なくなる．

3.3 音響事象群における事象間距離と調音努力

「個々の音が音響的に明瞭に区別できていない」場合，それは調音的にも区別できていないことを意味する．音響事象群に対して，全ての二事象間距離を求めて幾何学構造（距離行列）として事象群を表象する音声の構造的表象が提案されている．この場合，距離行列から構造のサイズ（構造の半径に相当する）が求まるが，この量が，凡そ調音努力に相当する定量的尺度になることが実験的に示されている [8]．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

Table 1　被験者の TOEIC スコア

熟練度別学習者数	素点	平均点
中位者 3 名	432, 427, 421	427
下位者 3 名	301, 202, 197	233

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 [9]．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非常に相性の良い技術であると言える．以下，実験的に検証する[1]．

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のよる評定

日本人英語学習者 6 名にシャドーイングを行なわせた．今回，特に低習熟度者の発声（より崩れた音声）に対する評定技術の構築を考えており，TOEIC テスト（990 点満点）における中位者 3 名，下位者 3 名を対象とした．彼らの TOEIC スコアを表 1 に示す．

シャドーイング用に提示した音声は，1 名の男性母語話者が読み上げた音声であり，全 21 文（335 単語）である．平均話速は 140 語/分であった．6 名による合計 126 発話のシャドーイング音声を実験に用いた．なお，収録に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（第三著者）によって全単語法で行なわれた．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．1 名分（約 4 分の音声データ）の評定に 2 時間程の時間を要した．

[1]なお，ボトムアップクラスタリングを行なわず，初期の $N \times N$ 距離行列（N= 総フレーム数）における構造サイズをもって，与えられた発声の調音努力は推定可能と考えられるが，本稿では固定閾値におけるセグメント数をもって調音努力と解釈した．

Table 2　音響分析条件

サンプリング	16bit / 16kHz
窓及び窓長	ハミング窓 / 16msec
シフト長	10msec
音響パラメータ	MCEP 1〜12 次元
クラスタリング停止条件	閾値 $K = 0.23$

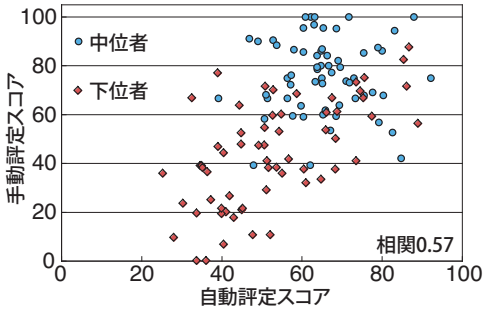

Fig. 2　自動評定スコアと手動評定スコア

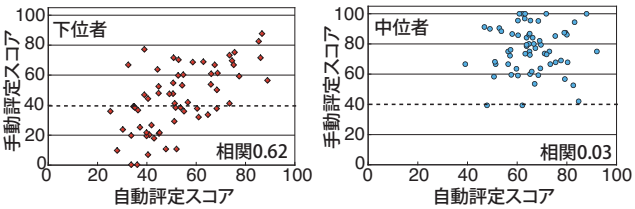

Fig. 3　各学習者グループに対する評定スコア

4.2　音響分析及びクラスタリングの諸条件

各種の分析条件を表 2 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，事前実験により，TIMIT データベース 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　各シャドーイング発声の自動評定結果

収録した 126 発声及び，提示した 21 発声を各々クラスタリングし，自動停止した時のセグメント数を算出した．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．自動評定／手動評定スコアの関係を図 2 に示す．TOEIC スコアの中位者／下位者を青／赤で示している．全体の相関は 0.57 となり，高い関係性は示されなかった．

しかし，この相関図を中位者／下位者別にプロットすると（図 3），両者の分布には大きな差があることが分かる．中位者にとっては今回のタスクの 21 文は

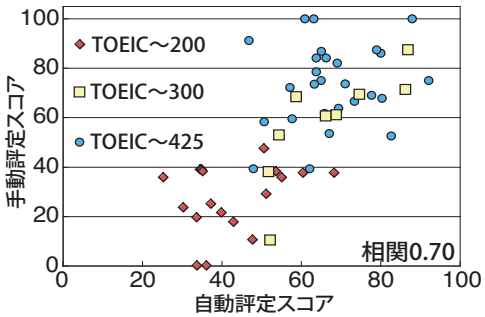

Fig. 4　低スコア文に対する手動／自動評定

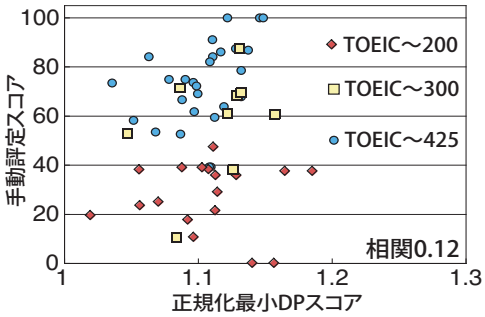

Fig. 5　正規化 DP スコアと正規化手動評定スコア

どれも似たような難易度であり，その結果手動／自動両スコアとも固まって存在する．例えば，自動評定スコアが 50 未満の発声が散見されれば確実に下位者である．下位者にとっては提示された 21 文には容易な文から困難な文まで存在し，スコアは手動／自動ともに大きくばらついている．そこで 2 人以上の下位者が手動評定スコアで 40 点以下となった文（9 文）に限定して再度分析を行なった．結果が図 4 である．相関は 0.70 となった．なお，黄は TOEIC スコア 301 点の学習者である．図よりスコア 200 点群と 425 点群とは手動／自動スコア両者において比較的明確に分離され，300 点学習者がその両者に渡って存在する様子が分かる．習熟度を自動推定する場合，当然，習熟度の高低が適切に反映されるタスクを課す必要がある．タスク設定を適切に行なうことができれば，提案手法の実用性も向上すると考えられる．

4.4　DP マッチングによる自動評定結果

提案手法は，提示音声とシャドーイング音声間で一切音響的照合は行なっていない．シャドーイングは提示音声をそのまま真似る（再生する）という側面を有しており，比較対象として，提示音声と再生音声を音響的に照合することでスコアを算出する DP マッチングも行なった．この場合，両音声間の DP スコアが小さいほど「音響的に」類似している音声となる．その結果，DP スコアと手動評定スコアとは負の相関が見られるはずである．図 5 に図 4 で用いた 9 文に対する正規化 DP スコアと手動評定スコアの関係を示す．425 点群／ 200 点群間の手動評定スコアの差は明確に現れているが，正規化 DP スコアにはその差は全く現れておらず，無相関の結果となった．DP マッチングは，話者性の違いによってスコアが大きく変動する．不一致（ミスマッチ）問題が浮き彫りになった．

5　まとめ

筆者らが提案している教師無しセグメンテーション手法が，調音努力（滑舌の良さ）に相当する定量的尺度を提供できることを鑑み，近年注目を集めているシャドーイングに着目し，その自動評定を試みた．その結果，シャドーイング対象となる文セットを適切に選択することができれば，手動の評価に沿った自動評定が可能であることを示した．その一方で，DP マッチングでは自動評定は極めて困難との結果を得た．

本提案手法は言語非依存の技術である．仮に，新たに言語が発見された場合であっても，その直後に，その言語の発音・聴取能力の自動評定が可能となる技術である．また提案手法は不一致（ミスマッチ）問題とは無縁の技術である．距離行列計算で必要なのは，同一話者内での「音と音の距離計算」のみである．DP や HMM のように異なる話者間で「音と音の距離計算」を行なえば，不可避的に不一致問題が発生する．発音評定を行なうシステムを構築する場合，低いスコアを提示された時に，それが学習者の習熟度が低いからなのか，それとも学習者の声質がシステムの学習データに合致しないのか，不明であることが多い．筆者等が提唱する音声の構造的表象を含め，教育応用には，安全かつ健全な技術構築が望まれると考える．

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] 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.
[4] S. Dusan *et al.*, “On the relation between maximum spectral transition positions and phone boundaries,” Proc. InterSpeech, pp.17–21, 2006.
[5] Y. P. Estevan *et al.*, “Finding maximum margin segments in speech,” Proc. ICASSP, pp.937–940, 2007.
[6] Y. Qiao *et al.*, “Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons,” Proc. ICASSP, 2008 (to appear).
[7] 宮本定明, “クラスター分析入門 ファジィクラスタリングの理論と応用”, 森北出版, 1999.
[8] N. Minematsu *et al.*, “Para-linguistic information represented as distortion of the acoustic universal structure in speech,” Proc. ICASSP, vol.1, pp.261–264, 2006.
[9] M. Nakamura *et al.*, “Acoustic and linguistic characterization of spontaneous speech,” Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.3–8, 2006.