母語話者シャドーイングに基づく 可解性自動計測と回帰分析による高精度化*

☆井上雄介, 椛島優, 齋藤大輔, 峯松信明 (東大)

1 はじめに

第二言語獲得の為には，スピーキング，リスニング，ライティング，リーディングの４技能全てを習得する必要があるが，特にスピーキングとリスニングにおいては，他者との音声コミュニケーションが求められる。リスニングに関しては CD 等の音声教材を用いても訓練可能であるが，スピーキングに関しては他者とのコミュニケーションを妨げる発音誤りを認識する必要があり，故に母語話者と接する機会をより多く持たなければならない。しかし，実際には自国で学ぶ学習者の多くは授業以外で母語話者と接する機会が少ないため，これを技術的に支援する対話形式の CALL（Computer-Aided Language Learning）システムが研究されてきた [1, 2, 3]。これらのシステムは発音誤りや文法誤りを自動検出し，その誤りをどのように修正すべきかといったフィードバックを返す。この時，母語話者音声で訓練した音響モデルを用いて学習者音声を評価する。これはつまり母語話者音声との比較によって学習者音声を評価していることになる。この技術により確かに外国語訛りに起因する不自然な発音を自動検出することが可能であるが，一方で外国語訛りの程度によっては，コミュニケーションが妨げられないことが知られている [4, 5, 6]。

学習者発音に対する評価として，応用言語学の分野では intelligibility と comprehensibility という指標が良く用いられる [4]。本研究では，[4] に倣ってそれぞれを以下のように定義する。intelligibility は与えられた発話に対して，単語などの言語単位でどれだけ正確に聞き取られるかを示す指標である。intelligibility の度合いは母語話者に発話を書き起こさせることにより客観的な測定が可能である。一方 comprehensibility は，与えられた発話内容の理解に対する認知負荷を示す指標であり，母語話者にアンケートを課し主観的に評価することや，意味理解テストによって評価することが多い。以上の定義から，本研究では intelligibility を了解性，comprehensibility を可解性と訳す。発話内容を正しく理解するためには，単語の同定に加えて統語構造の把握や，発話者の意図理解など高次の処理が必要であることが多く，可解性は了解性を包含する概念であると考えられる。例えばある発話のすべての単語を正しく同定できた（了解性が高い）としても，発話内容の理解に努力を要した場合には，その発話の可解性は高いとは見なされない。

[4, 5, 6] では，外国語訛りの程度によっては，了解性，及び可解性を下げないことが示されている。すなわち，ある程度の外国語訛りは母語話者の許容の範囲内であり，円滑なコミュニケーションを妨げない。学校での発音指導，及び CALL システムにおいては，可解性を著しく低下させるような発音誤りを優先的に指摘するべきである。そして技術的に実現すべきは，学習者音声の発音誤りに対する聞き手（母語話者）の許容度のモデリングである。

では可解性を下げる発音誤りを学習者自身が知るにはどうすればよいか。一般の母語話者は面と向かって学習者の発音を厳しく指摘することは少なく，婉曲的あるいは上辺だけの指摘をする場合が多い。そこで筆者らは，母語話者のより率直な可解性に関する印象を，観測可能な反応に基づき推測する手法として，母語話者シャドーイングを提案している [7]。

シャドーイングは即座の反応が求められるため，母語話者が感じた学習者音声の可解性がシャドーイングの円滑度（smoothness）に反映されると考えられる。[7] では被験者実験を通して，学習者音声の可解性を客観的に測定する本手法の有効性を示した。

本稿ではシャドーイングの円滑度に関する更なる特徴量を検討し，母語話者シャドーイング音声，及び学習者音声から得られた値と，主観スコアとの相関を計算した。さらに重回帰分析により各種特徴量から主観スコアの予測を行い，被験者間相関との比較により本モデルを学習者発話の可解性自動評価に利用することの妥当性を検討した。

2 母語話者シャドーイングコーパス [7]

ベトナム人日本語学習者の読み上げ音声を収録し，日本語母語話者に対して母語話者シャドーイング実験を行なった。学習者音声の話速が遅いと可解性は常に高くなり，発音の影響がシャドーに反映されにくいと考えられる。そこで音声収録の際には話速の統制を行った。またベトナム人学習者の日本語音声を収録するとともに，比較のため日本語母語話者の音声も収録した。以下，音声収録の概略を説明する。

テキストは中級レベルの日本語の教科書を採用した [8]。この教科書には音声 CD が付属しており，日本人ナレーターによるモデル音声が収録されている。この教科書から 10 文章を選出した。1 文章あたり平均約 16 フレーズ（文より短い文節群），合計 164 個の異なりフレーズである。この時，固有名詞を含む文章は除外した。また日本語の読みやすさを計算するツール，Jreadability [9] を用いて，これら 10 文章が同一レベルに属することを確認した。

10 文章中のそれぞれのフレーズを，6 名のベトナ

---

* Automatic measurement of comprehensibility through native speakers' responsive shadowing and its improvement based on regression analysis. by Inoue Y., Kabashima S., Saito D., and Minematsu N. (The University of Tokyo)

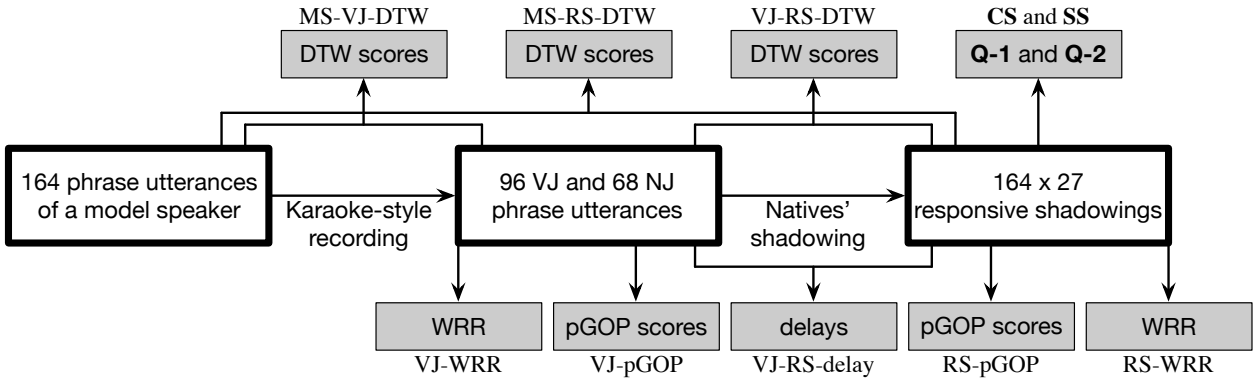

Fig. 1　母語話者シャドーイング実験の全体図

ム人（男性 3 名，女性 3 名）と 6 名の母語話者（男性 3 名，女性 3 名）に読み上げさせた。6 名のベトナム人学習者は，3 名が学習歴 3 年未満（平均 2.7 年）の中級レベル，残り 3 名が学習歴 3 年以上（平均 5.8 年）の上級レベルである。読み上げ時の話速統制には，CD モデル音声の強制アライメントによって得られた時間情報に従い文字色が変化する，カラオケスタイルの録音アプリケーションを用いた。また読み上げの際に吃ったり，言い間違えたりした場合には何度でもやり直しを許した。

最終的にベトナム人学習者 1 人辺り約 100 音声，母語話者 1 人あたり 164 音声が得られた。ベトナム人学習者の場合習熟度によって収録所要時間に差があり，得られた音声の数に差がある。これらからベトナム人日本語音声（VJ）96 音声と，日本人日本語音声（NJ）68 音声を提示音声として選択した。VJ と NJ には重複するフレーズは存在せず，164 個の異なりフレーズとなっている。

シャドーイング実験には合計 27 名の被験者（日本語母語話者）が参加した。被験者全員が VJ 96 音声，NJ 68 音声，及び解析に用いないダミー音声 36 音声の合計 200 音声をシャドーイングした。ダミー音声は，非日本語母語話者による日本語読み上げ音声コーパス Japanese Read by Foreigners (JRF) [10] から，ベトナム人による読み上げ音声を選択した。音声呈示はヘッドホンを通してランダムな順序で行われ，音声収録には単一指向性のイヤーフックマイクを用いた。

シャドーイングの際には呈示した学習者音声の訛りを真似しないように指示した。また呈示音声を単語単位で同定し，標準的な日本語でシャドーするように指示した。

1 音声のシャドーイングが終わる度に，下記 2 つの主観評価を課した。

Q-1　呈示音声がどれくらい理解し易かったか

Q-2　どれくらいスムーズにシャドー出来たか

Q-1 は可解性に関する質問であり，**Q-2** はシャドーイングの円滑度に関する質問である。どちらの質問も 7 段階評価とした。実験を開始する前に，JRF から実験に使用しないベトナム人学習者音声を選択し，シャドーイング及び主観評価を約 10 分間練習させた。本稿では **Q-1** に関する主観スコアを **CS** (comprehensibility score)，**Q-2** に関する主観スコアを **SS** (shadowability score) と呼ぶ。

3　シャドーイングの円滑度に関する特徴量

シャドーイングの円滑度を定量化するため，二つの特徴に着目した。一つは調音の正確さに関する特徴，もう一つはシャドーイングの遅れに関する特徴である。[7] では前者の特徴量として，DNN 音響モデルを用いて各時刻の音響特徴ベクトルを音素事後確率ベクトルに変換し，フレーム単位の GOP (Goodness of Pronunciation) スコアを計算した。しかし一般に母音は子音に比べ時間長が長く，発話全体で事後確率を平均する場合，母音の影響が強調される。このバイアスを防ぐため，各音素区間ごとにフレーム平均 GOP を計算し，それを全出現音素数で平均する方法をとった。これを pGOP (phoneme-based GOP) とする。本研究では調音の正確さに関する更なる特徴量として，[11] を参考に DTW (Dynamic Time Warping) 及び WRR (Word Recognition Rate) を計算した。

DTW とは、2 つの時系列に対して系列同士の累積距離が最も小さくなる対応付けを求める技術である。ここで系列同士の累積距離とは、系列を構成する要素間の局所距離の総和である。この累積距離が小さいほど 2 つの時系列の類似度が高いと言える。距離関数としては音素事後確率分布間のバタチャリヤ距離を用いる。音素事後確率ベクトル系列同士を DTW 計算することにより，話者性に対して独立に二つの音声を比較することが可能である。DTW は CD ナレーター音声，学習者音声，シャドー音声から三種類の組み合わせに対して計算した。

次に WRR は自動音声認識の精度を評価する際に広く用いられる指標で，以下のように計算される。

$$WRR = \frac{N - D - S}{N} \tag{1}$$

N は正解テキストの単語数，S は置換単語数（Substitution），D は消失単語数（Deletion）を表す。WRR は学習者音声とシャドーイング音声に対して計算した。

Table 1　各種特徴量と **CS** 及び **SS** の相関

特徴量	**CS**	**SS**	特徴量	**CS**	**SS**
RS-fGOP [7]	0.73	0.73	VJ-RS-delay*	0.59	0.69
VJ-fGOP [7]	0.63	0.50	MS-RS-DTW*	0.58	0.62
RS-pGOP	0.74	0.79	VJ-pGOP	0.58	0.44
RS-WRR	0.53	0.57	VJ-WRR	0.47	0.43
VJ-RS-DTW*	0.55	0.52	MS-VJ-DTW*	0.52	0.47

* 印は相関係数が負であることを示している。表中では比較のため絶対値を掲載している。

なお，DNN 音響モデルは KALDI toolkit [12] の CSJ (日本語話し言葉コーパス) [13] レシピに従い構築し，単語認識率計算には CSJ トライグラムを言語モデルとして用いた。

一方シャドーイングの遅れに関する特徴量として，強制アライメントにより学習者音声とそれに対応する母語話者シャドー音声それぞれの音素境界時間を取得し，対応する音素境界対の比較によりその遅れを計算した。二つの音声間の音素単位の遅れ時間の平均をシャドー音声の遅れ時間と定義する。

母語話者シャドーイング実験の全体図を Fig. 1 に示す。ただし母語話者シャドーイング音声を RS (Responsive Shadowing) とし，各特徴量の略称も示した。なお主観スコア **CS** 及び **SS** と，RS に関係する特徴量 (RS-pGOP，VJ-RS-DTW，MS-VJ-DTW，RS-WRR，VJ-RS-delay) は，提示した 164 フレーズ毎の被験者 27 名の平均値としている。

4　回帰モデルによる主観スコア予測

これまで述べた特徴量を説明変数とし，線形回帰モデルを構築することで主観スコアを予測する。これにより素の特徴量一つ一つを使う場合と比較して，高精度な予測が可能となることが期待される。

ただし説明変数間で相関係数が高い際には多重共線性が問題となることがある。そこで線形回帰モデルには Lasso を用いた。Lasso は線形回帰モデルに L1 正則化を行なったものであり，変数選択の性質を持っている。

5　実験結果と考察

5.1　各特徴量と主観スコアの相関

Table 1 に，シャドーの円滑度に関する各特徴量と，主観スコア **CS** 及び **SS** 間の相関係数を示す。ただし VJ96 フレーズについてのみ計算した。参考までに，[7] で計算した fGOP (frame-based GOP) との相関係数も示している。

VJ-RS-delay が負の相関なのは，可解性の低い音声ほどシャドーイングが遅れやすい，すなわちシャドーイング遅れ時間が大きくなるためと考えられる。また DTW スコアも負の相関となっているが，これは可解性の低い音声ほど二つの音声の音素事後確率ベクトル系列の類似度が下がり，累積距離が大きくなるためと考えられる。

Table 2　重回帰分析の結果

models	**CS**	**SS**
Lasso	0.81	0.86
inter-rater	0.66	0.59

特に注目すべきは，RS に関する特徴量が，それと対応する VJ に関する特徴量と比較してより高い相関を示している点である。特に RS-pGOP, VJ-RS-delay, MS-RS-DTW は **CS** と **SS** に対して高い相関を示しており，回帰モデル構築の際に有効な特徴量であると考えられる。Table 1 より，学習者音声の可解性が観測対象である場合，学習者音声そのものを解析するよりも母語話者シャドー音声を解析する方が有効であるということが言える。

また GOP，及び DTW に関する特徴量は，WRR に関する特徴量よりも高い相関を示した。音響モデルは通常，母語話者音声コーパスで学習を行い，母語話者音声を正しく認識するように最適化されている。つまり学習者の発音誤りに対する聞き手（母語話者）の許容度を測定するために ASR 技術を用いることが不適当である可能性がある。ASR モデルを学習者音声コーパスで学習を行い学習者音声の認識率を高めたとしても，それは聞き手の許容度を表す指標にはならず，つまり可解性の測定には適当でない。母語話者シャドーイングが利用可能である場合には，それを用いる方がより直積的でかつ適切である。

[7] では母語話者によるシャドーイング音声を効率的に集める手段として，異なる言語を学ぶ学習者間の相互シャドーイングを提案している。これは異なる言語を学ぶ学習者間でライティング能力を高めるインフラである，Lang-8 [14] の音声版である。これを実現するため，現在ベトナム人日本語学習者と，日本人ベトナム語学習者間の相互シャドーイング音声の収集を行っている。

5.2　線形回帰モデルによる予測結果

CS，**SS** に対してそれぞれ Lasso 回帰モデルを構築した。なおデータ数は VJ96 フレーズに対応する，96 組の各種特徴量と主観スコアである。データを 3 対 1 の割合で訓練データおよびテストデータに分け，訓練データの中で 3-fold のクロスバリデーションを行いハイパーパラメータを決定した。テストデータの取り方によって予測精度も変わるため，データ分割・ハイパーパラメータ調整・テストデータ予測を 1 セットとし，合計 50 セットの精度評価値の平均値を計算した。なお予測値の精度評価指標は，正解データとの相関係数である。

Table 2 に回帰分析結果と，被験者間の相関係数を示す。被験者間の相関係数の計算方法は次のように行った。まず被験者 27 名のうち 1 名と 26 名のグループに分ける。そして 26 名の **CS**，**SS** の平均値を計算し，残りの 1 名の **CS**，**SS** と相関を計算する。こ

れを 27 名の被験者が各々1 名の被験者となるよう繰り返し計算する。最後に，計算された 27 名分の相関係数の平均値が被験者間の相関係数である。Lasso モデルで予測する値 **CS**，**SS** が，被験者 27 名の平均値であることから，以上のような計算方法を採用した。

得られた被験者間相関はあまり高くない。[15] では事前に被験者に対して，comprehensiblity の各評価値に対して代表的な音声を示すことで，被験者間のコンセンサスを取っている。本研究では特にこのような操作は行っておらず，被験者間の評価基準が異なっていた可能性がある。

回帰モデルの予測結果は被験者間相関と比較して，かなり高い相関を示した。よって本モデルを評定者の代わりとして，学習者発話の可解性自動推定に用いることが可能であると考えられる。

6 結論

本研究では，母語話者シャドーイングコーパスに対してシャドーの円滑度に関する特徴量を計算し，被験者が付した主観評価スコアとの間に有意な相関があることを示した。また計算された特徴量を説明変数として主観評価スコアを予測する回帰モデルを構築し，予測精度が被験者間の相関を上回った。以上より，本モデルを学習者音声の可解自動計測に用いることの妥当性が示された。

今後の課題として，学習者間相互シャドーイングに向けて日本人ベトナム語学習者の音声を収録し，ベトナム語母語話者に対して同様の実験を行う予定である。学習者間相互シャドーイングのインフラ化が実現されれば，学習者同士が相互支援可能な新たなコミュニケーションツールとなる可能性がある。

また母語話者シャドーの崩れ（例えば pGOP）を可解性ラベルとして用いることができれば，母語話者シャドーイングは母語話者による可解性ラベリング作業だと考えられる。このコーパスを使えば，入力音声に対する可解性スコアを出力する予測モデルを構築することが出来るだろう。

さらにシャドーイングの円滑度が，intelligibility と comprehensibility のどちらに近いのかという議論も興味深い。シャドーイングは聴取しながらの繰り返し（再生）であることを考えると，円滑度は online intelligibility と解釈できる。その一方，円滑な理解が困難な音声のシャドーは崩れたり遅れたりすることも事実であり，この点からは comprehensibility に近いと考えられる。すでに筆者らは実験データに基づく検討を行っており，興味のある読者は [16] を参照していただきたい。

参考文献

[1] Reima Karhila, *et al.*, “SIAK-agame for foreign language pronunciation learning,” *Proc. INTERSPEECH*, pp. 3429–3430, 2017.

[2] Wei Li, *et al.*, “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling, ” *Proc. ICASSP*, pp. 6135–6139, 2016.

[3] Wei Li, *et al.*, “Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees, ” *Proc. INTERSPEECH,* pp. 3127–3131, 2016.

[4] Murray J. Munro and Tracey M. Derwing, “ Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, ” *Language Learning*, Vol. 45, No. 1, pp. 73–97, 1995.

[5] Murray J. Munro and Tracey M. Derwing, “The functional load principle in ESL pronunciation instruction: An exploratory study,” *System*, Vol. 34, pp. 520–531, 2006.

[6] Tracey M. Derwing and Murray J. Munro, *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research,* published by John Benjamins Publishing, 2015.

[7] Yusuke Inoue, *et al*., “A study of objective measurement of comprehensibility through native speakers’ shadowing of learners’ utterances, ” *Proc. INTERSPEECH*, accepted in 2018 (accepted).

[8] 松浦他, “日本語音読トレーニング,” アスク出版, 2014.

[9] *Jreadability*, `https://jreadability.net`

[10] Kikuko Nishina, *et al.*, “Speech database construction for Japanese as second language learning,” *Proc. O-COCOSDA*, pp. 187–192, 2002.

[11] 椛島他, “DNN-GOP と DNN-DTW に基づくシャドーイング音声自動評価の高精度化,” 春季音講論, pp. 1363–1366, 2018.

[12] Daniel Povey, *et al.*, “The KALDI speech recognition toolkit,” *Proc. ASRU*, 2011.

[13] Kikuo Maekawa, *et al.*, “Spontaneous speech corpus of Japanese,” *Proc. LREC*, pp. 947–952, 2000.

[14] *Lang-8*, `http://lang-8.com`

[15] Saito K., and Shintani N., “Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech?,” *TESOL Quarterly*, Vol. 50, No. 2, pp. 421– 446, 2016.

[16] Tasavat Trisitichoke, *et al.*, “Influence of content variations on native speakers’ performance of shadowing,” *Proc. Fall Meet. Acout. Soc. Jpn.*, 2018.