Vol.2018-SLP-123 No.4
2018/7/27

母語話者シャドーイングに基づく非母語話者音声の可解性自動計測

井上 雄介[1,a)]　椛島 優[1,b)]　齋藤 大輔[1,c)]　峯松 信明[1,d)]　金村 久美[2,e)]　山内 豊[3,f)]

概要：外国語発音学習の主目的は，母語話者に十分理解されやすい発音の獲得である。ところが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，その獲得が困難である場合が多い。また一般の母語話者が面と向かって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である場合が少なくない。そこで本研究では母語話者に対して，学習者音声のシャドーイングを課した。シャドーイングは聴取音声を出来るだけ即座に復唱する行為であり，母語話者が母語音声に対して行う場合は，聴取音声の可解性がシャドーイングの円滑度に反映されると考えられる。実験の結果，1）学習者音声の可解性，及びシャドーイングの円滑度それぞれに関する主観スコア間に強い相関が見られ，2）学習者音声の可解性に関する主観スコアは，学習者音声の GOP よりも母語話者シャドーイング音声の GOP とより強い相関を示した。この結果により，「可解性の高い発音は即ちシャドーイングしやすい発音である」と考えることの妥当性が示された。

キーワード：語学学習支援，可解性の主観評価，可解性の客観評価，母語話者シャドーイング，GOP

1. 研究背景

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ような発音誤りを把握する必要があり，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]。

このシステムは発音誤りや文法誤りを自動的に検出し，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場合が多い。つまり母語話者によるモデル音声と学習者音声との比較によって評価している。この技術により確かに外国語訛りに起因する不自然な調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]。

英語は国際的に広く使用される言語であるため，多種多様な外国語訛りが受け入れられている。また，インドやシンガポール，フィリピンなど多くの国で英語が公用語とされているが，彼らは独自の訛りで英語を話す上，それをアイデンティティと考えている場合もある。世界諸英語 [7], [8] という言葉は，英語の現状をよく特徴付けている。

しかし，多様な外国語訛りが受け入れられている英語であっても，コミュニケーションが妨げられるケースがあるのは事実である [10]。同時に多くの学習者は，母語話者に十分理解されやすい発音を獲得したいと願っている。

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

[1] 東京大学：The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–0033, Japan
[2] 名古屋経済大学：Nagoya University of Economics, 61–1, Uchikubo, Inuyama-shi, Aichi, 484–8504, Japan
[3] 創価大学：Soka University, 1–236, Tangimachi, Hachioji-shi, Tokyo, 192–8577, Japan
[a)] inoue0124@gavo.t.u-tokyo.ac.jp
[b)] kabashima@gavo.t.u-tokyo.ac.jp
[c)] dsk_saito@gavo.t.u-tokyo.ac.jp
[d)] mine@gavo.t.u-tokyo.ac.jp
[e)] kanamura@nagoya-ku.ac.jp
[f)] yutaka@soka.ac.jp

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すことで主観的に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，単語を正しく同定する必要があるため，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。学校での発音指導，及び CALL システムにおいては，了解性或い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。

では学習者が円滑な理解を妨げる発音誤りを自覚するにはどうすればよいか。[5] では機能負荷量（cognitive load）という概念を用いている。本研究では，母語話者の観測可能な反応に基づき可解性を推測する新たな手法を提案する。

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。それは初学者のやる気を維持するには良いかもしれないが，上級者はより率直な指摘を求めるだろう。

本研究では学習者音声に対する母語話者のより率直な印象を推定するため，母語話者に学習者の音声をシャドーイングさせた。シャドーイングは即座の反応を求められるため，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）として反映されると考えられる。この円滑度を定量化することで可解性を客観的に測定できるだろう。著者らが知る限り，母語話者に学習者のシャドーイングを課すことは L2 研究の中では初の試みであり，さらに可解性を客観的に測定する試みも今まで検討されていなかった。これを実現する為，本研究では学習者読み上げ音声収録，及び母語話者シャドーイング実験を行った。そしてシャドーイングの円滑度の指標として GOP（Goodness Of Pronunciation）スコア，及びシャドーイング遅れ時間を計算し，被験者が付した主観評価スコアとの比較を行った。

2. 了解性の客観的計測

[9], [10] では非母語話者音声の了解性を客観的に計測している。[9] は米国在住の移民（L1 は様々），[10] は日本人大学生の英語読み上げ音声を電話回線越しに米語母語話者に呈示した。母語話者には呈示した音声を書き起こすのではなく復唱するように指示した。そして復唱された音声は録音され，後日第三者によって書き起こされた。この書き起こし結果を元にして，発話毎に単語単位の同定率，すなわち了解性が計算された。

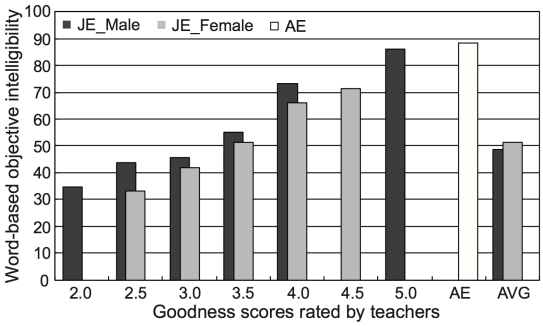

図 1 学習者習熟度毎の単語同定率 [10]

了解性は外国語訛りに対する聴取者の経験に依存すると考えられる。そこで [10] では，今まで日本人と接触経験のない米語母語話者を聴取者として採用した。また呈示音声は，ERJ(English Read by Japanese) コーパス [11] より日本人大学生 200 名に対し一人当たり平均 4 文音声を選択し，合計 800 音声とした。そして録音した英語読み上げ音声を米語母語話者 173 名に呈示した。この時，1 音声あたり平均 20 名の母語話者を割り当てた。また英語教師による主観評価の結果を元に，学習者を 7 つのグループにわけた。図 1 は各グループに対する単語単位の了解性を示している。単語単位の了解性とは単語単位の平均同定率のことであるが，母語話者の音声に対してはおよそ 90%[*1]であったのに対し，日本人英語学習者全体の単語了解性は約 50% ととても低い値となった。

[10] では聴取者に復唱という一定のタスクを課すことで客観的に了解性を測定した。しかし復唱方法の統制を取らなかった為，復唱までの遅れ時間や知識による補完などは聴取者に依存していたことが予想される。ここで復唱するまでの遅れ時間を最小化していくと復唱はシャドーイングとなり，知識による補完や推測がほぼ排除されると考えられる。またシャドーイングの場合，聞き取り難いと感じた瞬間の聴取者の心的な反応がシャドーイングの円滑度に反映されると予想される。よって筆者らはシャドーイングの円滑度は了解性というよりも可解性を表していると考えている。さらに復唱させる場合長い文は使うことができない。なぜなら復唱を始めるまでに文頭の単語を忘れてしまう可能性があるからである。しかしシャドーイングの場合は，文の長さは問題にならない。[10] では, 聴取者は日本人英語学習者と接触経験のない米語母語話者であった。本研究では学習者音声に対する経験に応じて 3 グループの母語話者を聴取者として採用した。

[*1] 連続する文音声は内容的に繋がりはなく，かつ話者も変わるという多少人工的な実験であったため，母語話者音声でも約 90%の正答率であった。

図 2　カラオケスタイル録音 Web サイト

3. シャドーイングの円滑度の定義

本研究ではシャドーイングの円滑度を定量化するため，2 つの特徴に着目した。調音の正確さに関する特徴，及びシャドーイングの遅れに関する特徴である。前者には GOP [12] を用いた。GOP は調音の正確さを表すベースラインの特徴として広く用いられており，我々の先行研究でも学習者のシャドーイングの評価に用いられてきた [13], [14], [15]。

GOP は音素事後確率ベクトル $P(c_i|o_t)$ として定義される。ここで，o_t は各時刻 t で観測される音響特徴量であり，c_i は音素 i のクラスである。[12], [13], [14] では HMM（Hidden Markov Models）音素生成モデルを用いて GOP 計算を行なっているが，近年は [15], [16] のように DNN (Deep Neural Network) 音素識別モデルが用いられている。DNN モデルは誤り検出や熟達度予測などのタスクにおいて実験的に HMM モデルよりよい結果を示している。本研究では学習者音声，及び母語話者シャドーイング音声に対して DNN に基づく GOP を計算した。まず強制アライメントによって時刻 t で意図された音素を取得し，$P(p_t|o_t)$ を発話全体で累積する。その後発話 x に対する GOP は以下のように計算される。

$$\mathrm{GOP}(x) = \frac{1}{D_x} \sum_t P(p_t|o_t), \tag{1}$$

ただし，D_x は与えられた音声のフレーム長である。

一方シャドーイングの遅れ時間は，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。

また考察の際には音声認識技術との比較のため，シャドー音声及び学習者音声に対する単語認識率を計算した。なお，DNN は KALDI[17] の CSJ(日本語話し言葉コーパス [18]) 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。

4. 母語話者シャドーイング実験

4.1 学習者読み上げ音声収録時の話速統制

本研究では対象言語を日本語，学習者をベトナム人とした。また呈示音声として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。この時，話速が遅すぎる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。

テキストは中級レベルの日本語の教科書を採用した [19]。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 [20] 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。さらに読み上げ時の話速統制のため，図 2 に示すカラオケスタイルの録音アプリケーションを用いた。

このアプリケーションでは，CD モデル音声に対する強制アライメントにより得られた時間情報に合わせて，各フレーズ中の文字色が変化する。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。そのうち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。

4.2 母語話者被験者の構成

シャドーイングを円滑に行えるかどうかは，ベトナム人日本語に対する経験に依存することは容易に予想される。

本研究では以下 3 グループの母語話者被験者を用意した。

NS-1 ベトナム人留学生との会話経験が全くない大学生
NS-2 研究室でベトナム人留学生と会話経験がある大学生
NS-3 ベトナム人留学生に日本語を日常的に教える教師

上から 17 名，5 名，5 名の日本人（20 歳以上）が被験者として実験に参加した。実験の教示として，呈示音声をシャドーする際に，決して訛りを真似ないよう指示した。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。

4.3 シャドーイングの主観評価，及び客観評価

実験には合計 27 名の被験者が参加した。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) [21] から，ベトナム人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

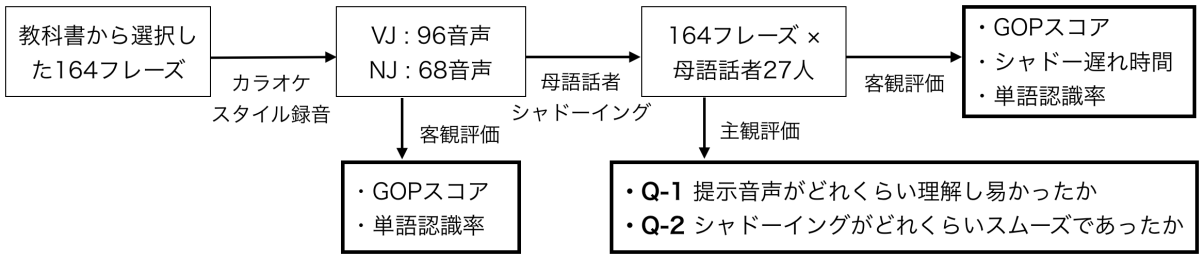

図 3　母語話者シャドーイング実験の全体図

表 1　可解性，及びシャドーの円滑度に関する主観スコアの平均値（VJ 音声に関して計算）

	NS-1	NS-2	NS-3
comprehensibility (S_C)	4.13	4.20	4.24
smoothness (S_S)	4.58	4.45	4.78

表 2　可解性，及びシャドーの円滑度に関する主観スコアの平均値（NJ 音声に関して計算）

	NS-1	NS-2	NS-3
comprehensibility (S_C)	6.53	6.75	6.40
smoothness (S_S)	6.17	6.08	5.87

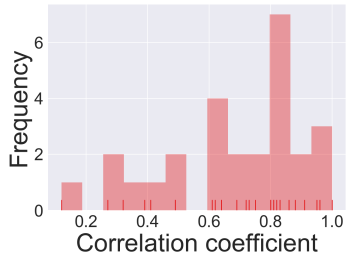

図 4　被験者毎の S_C，S_S 間の相関係数のヒストグラム

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。また 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 シャドーイングがどれくらいスムーズであったか

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。常識的に考えると，これら 2 つの値は高い相関となることが予想される。

さらに 3 節に示した方法によって，客観評価を行った。特に DNN-GOP スコア（S_G）は，シャドーイング音声だけでなく学習者音声に対しても計算した。5.2 節ではこれら 2 種類の DNN-GOP スコアの比較を行う。またシャドーイング遅れ時間はすべてのシャドーイング音声に対して計算した。本実験の全体図を図 3 に示す。

5. 結果と考察

5.1 被験者グループ毎の主観評価スコア

同じ学習者音声に対し，4.2 節に示した 3 グループ間で異なる可解性スコア（S_C），及び円滑度スコア（S_S）が得られると予想される。表 1 に，VJ に対する 2 つのスコアの平均値を被験者グループ別に示す。一元配置分散分析の結果，S_S に関して **NS-1**–**NS-3** 間，及び **NS-2**–**NS-3** 間にのみ有意差 (p<0.05) が見られた。S_C においては有意差の見られる組み合わせは存在しなかったが，**NS-1** から **NS-3** にかけて上昇する傾向があることがわかった。表 2 は NJ に関する平均値である。

5.2 二つの主観評価スコア間の相関

二つの主観評価スコア S_C，S_S 間の相関を被験者毎に計算した。その結果，平均値は 0.68 とそれほど高くない値となった。図 4 は被験者毎の S_C，S_S 間の相関係数のヒストグラムを示している。27 名の内，7 名の被験者は S_C，S_S 間の相関が低かった（平均 0.36）。この 7 名の被験者は，学習者音声の可解性，及びシャドーの円滑度に対する評価基準が異なっていたと考えられる。評価基準の統制のため，可解性とシャドーの円滑度の各スコアの基準となる音声を示し，事前にコンセンサスを取るべきであったと考えられる。しかし十分にコンセンサスを取れていたとしても，多少の評価基準の不一致は避けられないと考えられる。また残りの 20 名の相関係数の平均値は 0.79 と高い値となった。以降の節では，客観的な評価として GOP スコア S_G 及びシャドーイング遅れ時間の計算結果について議論する。その際，27 名すべての被験者から得られた結果に加えて，S_C，S_S 間の相関が比較的高かった 20 名の被験者から得られた結果についても述べる。

5.3 DNN-GOP スコアと二つの主観評価スコア

VJ 96 音声と NJ 68 音声のそれぞれに対して，27 名のシャドワーによるシャドーイング音声，S_C，及び S_S が被験者実験で得られた。そして全てのシャドーイング音声に対して S_G を計算した。さらに 164 個の呈示音声毎に，すべての被験者間の S_G，S_C，S_S の平均値を計算した。図 5 は S_C と S_G の平均値間，及び S_S と S_G の平均値間の相関を表している。赤点は VJ，青点は NJ の音声を表している。各図の **R** は VJ のみに関して計算された相関係数である。図 5 の S_G と S_S のどちらもシャドーイング音声に対して直接得られたスコアであり，高い相関が得られるのは自然であると考えられる。興味深いことに，S_C はシャドーイング音声に対するスコアではなく学習者音声に対するスコアであるにも関わらず，S_G と高い相関を示した。

5.2 節で述べた 20 名の被験者に限定して相関係数を計算

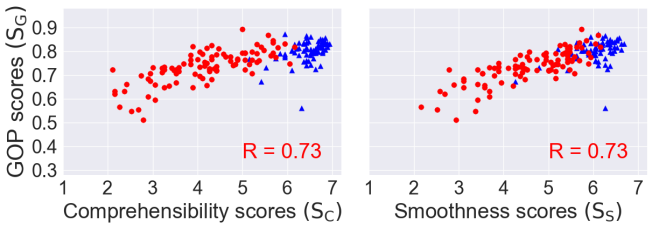

図 5　シャドー 音声の GOP と二つの主観評価スコアの相関

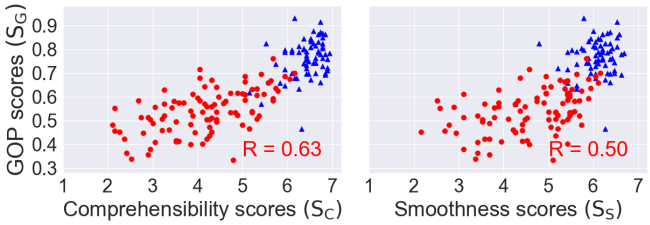

図 6　学習者音声の GOP と二つの主観評価スコアの相関

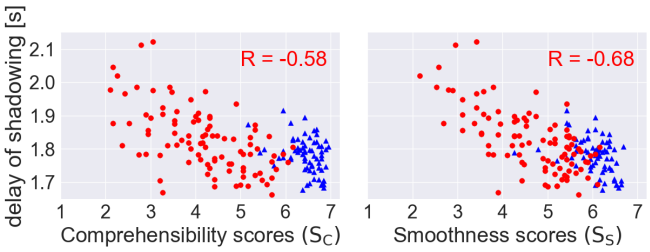

図 7　シャドーイング遅れ時間と二つの主観評価スコアの相関

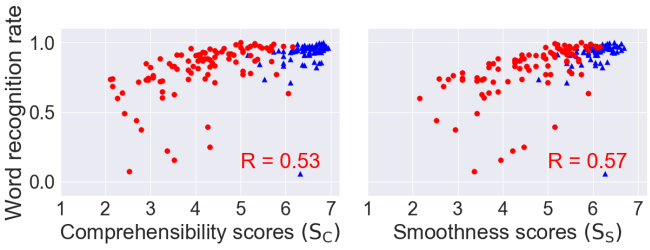

図 8　シャドー音声の単語認識率と二つの主観評価スコアの相関

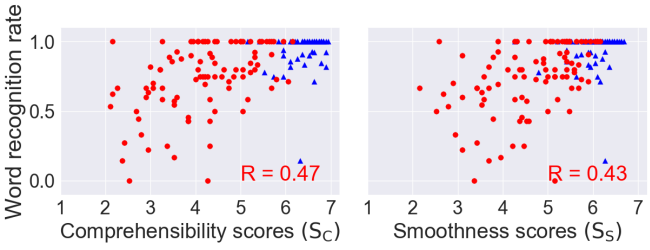

図 9　学習者音声の単語認識率と二つの主観評価スコアの相関

すると，$\mathbf{R}(S_G, S_C)$=0.75，及び $\mathbf{R}(S_G, S_S)$=0.72 であった。この結果は，呈示音声の可解性を測定するために S_G を用いることが妥当であることを示している。

S_G をシャドーイング音声ではなく，学習者音声（VJ，及び NJ）に対しても計算した。学習者音声の S_G と，2 つの主観評価スコアの相関を図 6 に示す。S_G 及び S_C は学習者音声から直接得られた値であるが，相関係数は図 5 の値よりも低い値となった。以上の結果から，GOP に基づいた学習者音声の可解性評価においては，学習者音声そのものよりも母語話者シャドーイング音声を分析した方がより適切であると考えられる。これには二つの理由があると考えられる。一つは学習者音声に対する可解性スコア S_C

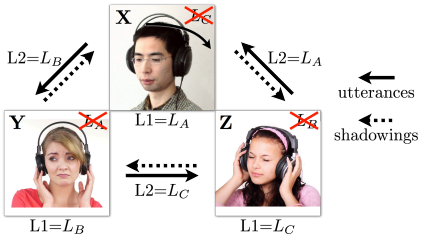

図 10　学習者間相互シャドーイングの構図

は通常聴取者に依存するが，学習者音声から計算された S_G は聴取者には全く依存しない点である。もう一つの理由は，母語話者シャドーイング音声は母語話者による発話なので，非母語話者音声に対して S_G を計算する場合と比較して，技術的により安定していると考えられる点である。

5.4　シャドーイング遅れ時間と 2 つの主観評価スコア

図 7 はシャドーイング遅れ時間と 2 つの主観評価スコア S_C，S_S の間の相関を示している。可解性の高い呈示音声ほど即座にシャドーイング出来ると考えられるため，2 つのグラフともに負の相関が得られたのは自然な結果である。絶対値は S_S に対する相関係数の方が大きい。これはシャドーイングが遅れた場合，シャドーイングのスムーズさ，つまり S_S は下がるが，常に S_C が下がる訳ではないと考えられるからである。

この節ではシャドーイング音声の S_G と学習者音声の S_G，及びシャドーイング遅れ時間を別々に分析したが，これらの値をすべて用いて回帰モデルを構築し，より高精度に自動予測することが可能である。この回帰分析は今後の課題として速やかに行うこととする。

5.5　単語認識率と 2 つの主観評価スコア

図 8 は母語話者シャドー音声，図 9 は学習者音声に対する単語認識率と，S_C，S_S の間の相関である。相関係数は図 5，図 6 より低い。教科書 CD の音声を音声認識した場合でも，天王星→天皇制，土星→怒声のように同音異義語として認識されることがあり，これは学習者音声やシャドー音声でも同様であった。これらの単語置換（同音異義語置換）は，言語モデルの学習データに起因する問題であるが，音声認識率は不可避的に言語モデルの影響を受ける。発音評価タスクを「与えられた文を，正しい調音制御・韻律制御で音声化できているかどうかを評価するタスク」と考えた場合，本来これは発声する単語列とは独立であるため，音声認識率を直接利用する場合は注意が必要である。

5.6　学習者間相互シャドーイング

本研究では，学習者音声に対する可解性自動計測のフレームワークとして，母語話者シャドーイングという新たな方法論を提案している。ところがこの方法論において，一つ重大な問題が存在する。それは学習者の数に対して十分な数の母語話者をどのように集めるかということである。

著者らはこの問題は学習者間相互シャドーイングというインフラによって解決できると考える。すべての言語学習者は学習者であると同時に少なくとも一つの言語の母語話者であり，その言語の学習者音声を母語話者シャドーすることが可能である。図 10 は学習者間相互シャドーイングの構図を示している。言語 L_A を L1 とし言語 L_B を学習しているが，言語 L_C は学習したことがない学習者 **X** が言語 L_B で文章を読み上げる。次に言語 L_B を L1 とし言語 L_C を学習しているが，言語 L_A は学習したことがない学習者 **Y** が，学習者 **X** の音声をシャドーする。同様に言語 L_C を L1 とし言語 L_A を学習しているが，言語 L_B は学習したことがない学習者 **Z** が，学習者 **Y** の音声をシャドーする。最後に学習者 **Z** の音声を学習者 **X** がシャドーする。

またこの枠組みには更なる利点もある。学習者にとって最も可解性が高い発音は，自分自身の発音であると言うことは自明である。つまりその発音が母語話者にとってどれほど聞き取り難いかを学習者は自覚できない。もし様々な言語に関して母語話者シャドーの円滑度をスコア化出来れば，学習者は自身の L2 発話の可解性と同等の可解性を持つ L1 発話を体感することができる。

6. 結論

本研究では，学習者音声の可解性自動計測の新たな手法として母語話者シャドーイングを提案し，学習者読み上げ音声収録，及び母語話者シャドーイング実験を行なった。そして GOP スコア，及びシャドー遅れ時間と，被験者が付した主観評価スコアとの比較によってこの手法の妥当性を示した。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。今後の課題として，シャドーの円滑度を定量化する更なる指標の検討，及び重回帰分析による可解性予測精度の向上を図りたい。

参考文献

[1] Reima Karhila, Sari Ylinen, Seppo Enarvi, Kalle Palomäki, Aleksander Nikulin Olli Rantula, Vertti Viitanen, Krupakar Dhinakaran, Anna-Riikka Smolander, Heini Kallio, Katja Junttila, Maria Uther, Perttu Hämäläinen, and Mikko Kurimo, “SIAK-agame for foreign language pronunciation learning,” Proc. INTERSPEECH, pp. 3429–3430, 2017.

[2] Wei Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-Hui Lee, “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling,” Proc. ICASSP, pp. 6135–6139, 2016.

[3] Wei Li, Kehuang Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-Hui Lee, “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] Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson, “ The handbook of World Englishes”, published by Wiley-Blackwell, 2009.

[8] Jennifer Jenkins, “World Englishes: a resource book for students,” published by Routledge, 2009.

[9] Jared Bernstein, “Objective measurement of intelligibility,” Proc. ICPhS, pp. 1581–1584, 2003.

[10] Nobuaki Minematsu, Kohji Okabe, Keisuke Ogaki, and Keikichi Hirose, “Measurement of objective intelligibility of Japanese accented English using ERJ database,” Proc. INTERSPEECH, pp. 1481–1484, 2011.

[11] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, “Develop of English speech database read by Japanese to support CALL research,” Proc. Int. Conf. Acoustics, pp. 557–560, 2004.

[12] Silke M. Witt and Steve J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” Speech Communication, Vol. 30, No. 1, pp. 95–108, 2000.

[13] Luo Dean, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose, “Automatic assessment of language proficiency through shadowing,” Proc. ISCLP, pp. 1–4, 2008.

[14] Luo Dean, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose, “Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences,” Proc. SLaTE, pp. 37–40, 2009.

[15] Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito, and Nobuaki Minematsu, “Automatic scoring of shadowing speech based on DNN posteriors and their DTW,” Proc. INTERSPEECH, pp. 1422–1426, 2017.

[16] Wenping Hu, Yao Qian, and Frank K. Soong, “An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners’ speech,” Proc. SLaTE, pp. 71–76, 2015.

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, Karel Veselý, “The KALDI speech recognition toolkit,” *Proc. ASRU*, 2011.

[18] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara, “Spontaneous speech corpus of Japanese, ” *Proc. LREC*, pp. 947–952, 2000.

[19] 松浦真理子，福池秋水，河野麻衣子，吉田佳世, “日本語音読トレーニング,” アスク出版, 2014.

[20] Jreadability, https://jreadability.net

[21] Kikuko Nishina, Yumiko Yoshimura, Izumi Saita, Yoko Takai, Kikuo Maekawa, Nobuaki Minematsu, Seiichi Nakagawa, Shozo Makino, Masatake Dantsuji, “Speech database construction for Japanese as second language learning,” Proc. O-COCOSDA, pp. 187–192, 2002.