Vol.2018-SLP-125 No.22
2018/12/11

# 母語話者シャドーイングに基づく学習者音声の可解性自動計測と回帰分析による高精度化

井上 雄介[1,a)]　椛島 優[1,b)]　齋藤 大輔[1,c)]　峯松 信明[1,d)]

**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，即ち可解性（comprehensibility）の高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（lack of exposure）。また一般に，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，婉曲的あるいは上品な指摘となる場合が少なくない。そこで筆者らは [1] にて，母語話者に学習者音声をシャドーイングさせる母語話者シャドーイングを導入し，学習者音声の可解性を，シャドーイング音声の崩れとして客観的に計測する妥当性・有効性を実験的に示した。本研究では母語話者シャドーイングの崩れに関する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ることを確認した。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。

**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. 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，このため，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 [2], [3], [4]。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 [5], [6], [7]。

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 [5]。本研究では，[5] 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。

[5], [6], [7] では，外国語訛りの程度によっては，了解性・可解性を下げないことが示されている。つまり，ある程度の外国語訛りは母語話者の許容範囲内であり，円滑なコミュニケーションを妨げない。学校での発音指導や CALL

[1] 東京大学：The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–0033, 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

システムにおいては，可解性を著しく低下させる発音誤りを優先的に指摘すべきである。この場合，技術的に実現すべきは，母語話者の発音モデルではなく，学習者音声に対する聞き手（母語話者）の許容度のモデルであろう。

[8] では，学習者のシャドーイング音声に対して語学教師が評価し，そのスコアを回帰モデルにより高精度に予測している。語学教師は segments（分節的な適切さ），prosody（韻律的な適切さ），及び correctness の 3 つの尺度に関して評価したが，実験の結果，correctness に関する予測精度は他の尺度と比べて低かった。シャドーイングはテキストを参照せず，呈示された音声を聴取し，ほぼ即座に再生する必要があるため，正しく認識，理解できないと崩れた発声となる。このシャドーイング特有の発声に着眼し，correctness は，「学習者音声の母語話者らしさではなく，学習者本人が聴取した音声（各単語）を同定しながらシャドーできているように聞こえるかどうか」の尺度として導入されている。つまり，外国語訛りは凡そ許容された尺度となっている。[8] で検討した回帰モデルでは，全ての説明変数は，学習者音声と呈示した母語話者音声に対する分析結果であるため，「発音の母語話者らしさ」とは異なる尺度に対する予測精度は，低くならざるを得なかった。

本研究では，学習者音声が「母語話者音声からどの程度かけ離れているのか」ではなく，「母語話者にとってどの程度聴き取り易いか」に着眼している。[5], [6], [7] が主張するように，ある程度の外国語訛りは可解性に影響を与えないとすれば，上記で説明した学習者音声や，（学習者に呈示した）母語話者音声の分析結果を用いた説明変数では精度向上が望めない。そこで [1] では，母語話者シャドーイングを提案した。シャドーイングは（単なる再生，repetition とは異なり），呈示音声に対して発話尾まで聞く（待つ）ことが許されず，呈示に対してなるべく遅れずに再生することが求められる。それ故，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）に直接的に反映されると考えられる。[1] 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。即ち，学習者音声の分析結果よりも，それを聞いた母語話者のシャドーイング音声の分析結果の方が，その母語話者が感じ取った可解性との相関は，より高くなった。

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 [1]

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，比較のため母語話者の音声も収録した。以下，収録の概略を説明する。

まず，教科書として中級レベルの音読用教科書を採用した [9]。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 [10] 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。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 音声をシャドーイングの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 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) [11] 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

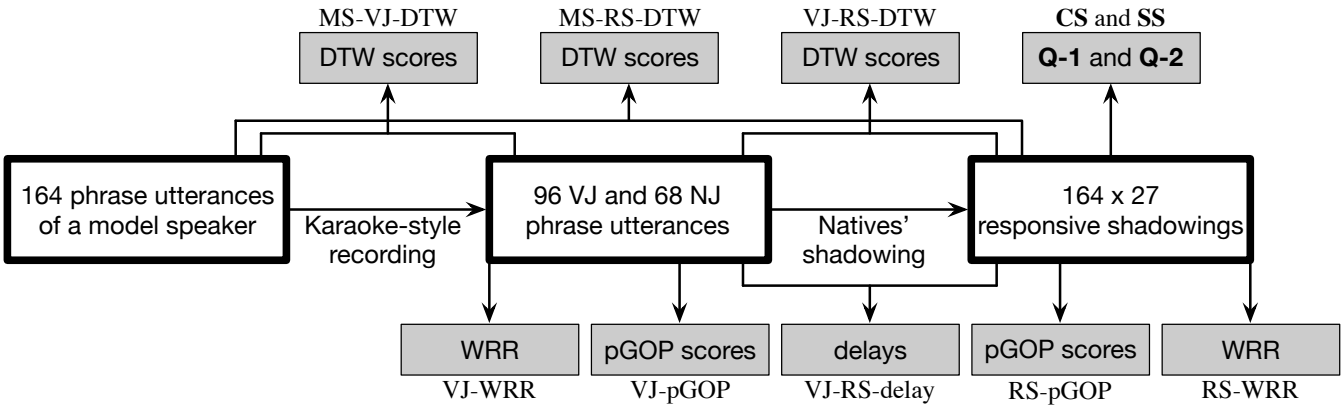


**図 1** 母語話者シャドーイング実験の全体図

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。[1] 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，[8] 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。

$$WRR = \frac{N - D - S}{N} \quad (1)$$

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。

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図 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

表 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，[1] 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

表 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 [1] | 0.73 | 0.73 | VJ-RS-delay* | 0.59 | 0.69 |
| VJ-fGOP [1] | 0.63 | 0.50 | VJ-RS-DTW* | 0.60 | 0.61 |
| RS-pGOP | 0.74 | 0.79 | VJ-pGOP | 0.58 | 0.44 |
| RS-WRR | 0.53 | 0.57 | VJ-WRR | 0.47 | 0.43 |
| MS-RS-DTW* | 0.68 | 0.71 | MS-VJ-DTW* | 0.60 | 0.51 |

* 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。

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からである。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である。

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（RS-pGOP と VJ-pGOP，RS-WRR と VJ-WRR，MS-RS-DTW と MS-VJ-DTW）。特に RS-pGOP, MS-RS-DTW，VJ-RS-delay 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。学習者音声の可解性が考察対象である場合，学習者音声そのものを分析するよりも，母語話者シャドー音声を分析する方が有効であることが再確認された。

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し，母語話者音声を正しく認識するよう最適化されている。つまり学習者の発音誤りに対する聞き手（母語話者）の許容度を計量するために ASR 技術を直接的に用いることは必ずしも適当ではない。ASR モデルを学習者音声コーパスで学習し，学習者音声の認識率を高めても，それを聞き手の許容度を表す指標とするのは難しいだろう。（母語話者である）聞き手が感じる認知負荷量については何ら計測していないからである。

音声認識は，話者が意図した単語列を，話者の音声から推定する技術である。可解性を検討する場合，話者が意図した単語列を推定するのではなく，聞き手が聴き取った単語列を推定するモデルが必要となる（許容モデル）。多様な学習者音声コーパスが整備されているが [14]，多くは，話者が意図した（読上げた）テキストが付与されており，意図推定モデルとしての ASR は構築できるが，聞き手の理解モデル・許容モデルの構築は困難であろう。母語話者が話し，母語話者が聞く場合は，話者が意図した単語列は，凡そ，聞き手が聞き取る単語列となるため，両者を同一視することができるが，学習者音声の場合，両者は異なる。聞き手の理解モデル・許容モデルの構築には，学習者音声を母語話者に書き取らせることで構築されるコーパスが必要となるが [15]，極めて開発コストが高い。母語話者によるシャドーイングは，シャドーイングというタスクを通して，学習者音声に comprehensibility というスコアを付与する・ラベル化する作業として考えられる。今後，大規模な母語話者シャドーイングコーパスも構築したい。

表 2　重回帰分析の結果

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

なお母語話者シャドーイングに基づく可解性推定は，学習者のみならず，彼らの音声を聞いてシャドーする母語話者を必要とする。母語話者によるサポーターを必要とする類似した学習フレームワークとして，Lang-8 がある [16]。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。学習者が入力した作文を，学習対象言語の母語話者が修正する。修正した母語話者が学習者として別言後を学んでいる場合は，同様に，学習言語の作文を入力すればよい。異なる言語を学ぶ学習者同士を繋いで，互いにシャドーさせる，学習者間の相互シャドーイングを [1] では提案したが，これはシャドーイング版の Lang-8 でもある。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。

表 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。[17] 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，事前に様々な習熟度の学習者音声を聴取させたが，comprehensibility のスコア付与に関し


図 2　学習者同士の相互シャドーイング

ては被験者独自の方針に任せた。被験者間で評価戦略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，極めて高い相関を示した。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。

### 5.3　学習者相互シャドーイングに基づく効果的な教示生成

今回の実験結果とは直接的な関与はないが，異なる言語を学ぶ複数の学習者をつなぎ，相互にシャドーさせる枠組みを構築できれば，従来とは大きく質の異なる教示が可能となる。これについて言及しておく。母語話者シャドーイングは，学習者自身の発声が母語話者にどのように聞かれているのか，をあぶり出す手段として提案している。Lack of exposure と言われる，外国語学習における大きな問題の解決として提案している。母語話者シャドーイングを学習者間の相互シャドーイングへと発展させる（図 2）と，外国語教育のもう一つの問題を解決する可能性が開ける。

Comprehensibility や intelligibility に基づく発音教育は，「容易に聞き取れる発音」を獲得することが目的であるが，全ての学習者にとって，自身の発音は極めて聞き取りやすい（恐らく母語話者発音より聞き取りやすい）。言い換えれば，他者が自身の発音をどう聞いているのか／感じているのか，が分からない。自身の発音の相対化が難しい。このような状況で「母語話者のような発音」を目指す教育から，「聞き取れる／理解できる発音」を目指す教育に移行すると，「自分が聴き取れるから直す必要がない」となる。母語話者にシャドーさせ，母語話者の shadowability score を返せば，他人がどう聞いているのかを（間接的に）伝えられるが，このような教示が効果的かどうか（学習者の学習動機を刺激するのか）は懐疑的である。

母語話者シャドーイングを学習者相互シャドーイングに発展させると，他者がどう聞いているのかを，より直接的に，より分かり易く呈示できると考えている。図 2 では，日本人学習者の英語音声を米国人にシャドーしてもらっているが，と同時に，様々な外国人の日本語音声を，当該学習者はシャドーしている。つまり，米国人の shadowability score と等しいシャドーイングの崩れを（当該学習者に）誘発する，外国人の日本語音声が得られる。米国人のシャドーイングを数値で示すと同時に，同レベルの shadowability score となる外国人の日本語音声を呈示できる。言い換えれば，「貴方の英語音声は，凡そ，こういう（外国人の）日本語音声に相当します」と外国語訛りの母国語音声を呈示できる。数値を使って説明されるのと，外国人の日本語音声を使って説明されるのでは，後者の方が，自身の英語音声の聞き取りやすさ／難さをより容易に感覚できることは想像に難くない。学習者自身の発音の相対化を，より直接的，かつ，感覚的に理解しやすく行なえる。このような教示が学習動機を刺激するのか，あるいは，衰退させるのか[*1]，現場に立つ教師との協議を進めていきたい。

## 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，学習者間の新たなコミュニケーションツールとなる可能性がある。

現実的な問題として，音響モデルがフレームごとに出力する音素事後確率は数千クラスに及ぶため，計算量やデータ量が非常に大きくなってしまうことが挙げられる。しかし一般には事後確率はスパース性を持つため，コンパクトなベクトル表現が可能である。[18] ではコンパクト化したベクトル表現を用いた DTW 計算の利点が報告されており，本研究における利用も検討したい。

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 [19] を参照していただきたい。

[*1] 直接的な発音相対化は，学習者の動機を損なう可能性がある。そういう場合は，（特に初級者に対しては）上品な，policitally-correct な教示が必要となるであろう。

本研究は，MEXT/JSPS 科研費 JP26118002, JP18H04107 の助成を受けました。

## 参考文献

[1] Yusuke Inoue, *et al.*, "A study of objective measurement of comprehensibility through native speakers' shadowing of learners' utterances," *Proc. INTERSPEECH*, pp.1651–1655, 2018.
[2] Reima Karhila, *et al.*, "SIAK-agame for foreign language pronunciation learning," *Proc. INTERSPEECH*, pp. 3429–3430, 2017.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] Tracey M. Derwing and Murray J. Munro, *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research,* published by John Benjamins Publishing, 2015.
[8] 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.
[9] 松浦他, "日本語音読トレーニング," アスク出版, 2014.
[10] *Jreadability*, `https://jreadability.net`
[11] Kikuko Nishina, *et al.*, "Speech database construction for Japanese as second language learning," *Proc. O-COCOSDA*, pp. 187–192, 2002.
[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] `https://en.wikipedia.org/wiki/Non-native_speech_database`
[15] N. Minematsu, K. Okabe, K. Ogaki, K. Hirose, " Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) database," *Proc. INTERSPEECH*, pp. 1481–1484, 2011.
[16] *Lang-8*, `http://lang-8.com`
[17] 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.
[18] 田中他, "音素事後確率のコンパクト化と発話比較への応用," 情報処理学会研究報告, 2018.
[19] Tasavat Trisitichoke, *et al.*, "Influence of content variations on native speakers' fluency of shadowing," *IPSJ SIG Technical Report.*, 2018.