母語話者シャドーイングとそれに基づく「聞き取り易さ」の客観的計測

峯松信明・井上雄介・椛島優・齋藤大輔（東京大学大学院）
金村久美（名古屋経済大学）山内豊（創価大学）
{mine,inoue0124,kabashima,dsk_saito}@gavo.t.u-tokyo.ac.jp,
kanamura@nagoya-ku.ac.jp, yutaka@soka.ac.jp

1　はじめに

外国語音声学習の目標は，多くの場合（母語話者のような発音ではなく），実用的に十分伝わる発音の習得であると言える。音声技術を用いて学習者音声を評価する場合，母語話者の発音モデルとの比較を通してスコア付けをすることが一般的であり (河原・峯松, 2013)，原理的に，母語話者の発音に近いほど高得点となる。本研究では，可解性（comprehensibility）の高低を技術的に自動計測するための手段として，母語話者による学習者音声シャドーイングを提案し，母語話者によるシャドーイング音声を用いた可解性の自動予測に関して，その妥当性を実験的に検証した。

2　了解性・可解性と外国語訛り

応用言語学では，学習者音声中の個々の単語が明確に聞き取れるか否かに着眼した尺度を了解性（intelligibility）と呼び，例えば，母語話者の書き取りテストで定量化される。一方，その音声の意味理解が容易か否か（理解するための認知タスクが低いか否か）に着眼した尺度を可解性と呼んでおり，主観的な定量化や，理解度テストにより定量化される (Derwing & Munro, 2015)。可解性は，各単語の知覚のみならず，統語・意味・談話解析が容易か否かも考慮され，学習者音声に対するより総体的な評価基準と言える。外国語訛りが了解性や可解性に与える影響について多くの先行研究があり，母語話者発音からのずれがあっても十分了解性・可解性は高いが (Munro & Derwing, 1995)，聴取者がその訛りに慣れていない場合は，低い了解性となる (Mineamtsu et al., 2011)。

了解性・可解性に基づく発音評価を技術的に実装する場合，母語話者発音モデルとの近接性ではなく，聴取者の許容量（に相当するもの）を計測・モデル化する必要がある。これは，学習者発音に対する調音的・音響的分析では計測できず，聴取者の認知プロセスが計測の対象となり，客観的計測そのものが困難となる。筆者らの知る限り，特に可解性に対して，聴取者の許容量を，客観的かつ容易な計測方法を通して検討した研究例はない。筆者らは予備的検討として，ストレス感の自動計測を謳っている,「感性アナライザ」（種々の機能が付与された簡易脳波計）(電通, 2016) を試用した。強い外国語訛りの音声や，雑音下音声を用いた意味理解テストを構成し，聴取者の様子を計測したが，音声以外の刺激（視覚刺激など）によってもセンサー値が変動し，安定した計測は困難であった。

3　了解性の客観的な計測と可解性計測へのアプローチ

了解性（可解性ではない）を客観的に計測する場合，学習者音声を母語話者に提示し，書き取らせる／復唱させるなどして単語単位の正解率を見ることが多い。例えば，Mineamtsu et al. (2011) では，日本人の読上げ英語音声を対象とした了解性の客観的測定が行なわれている。この場合，学習者音声を提示し，書き取らせる／復唱させる訳だが，その際に許される認知的作業量を制限するなどの条件は課していないため，どの程度努力したのか（頑張ったのか）は聴取者依存である。例えば，学習者音声の提示後に復唱させれば，呈示内容の推測に十分な時間が与えられることになる。

復唱に対して時間的制限を設け，聴取と同時に復唱を行なわせれば，それは追唱（シャドーイング）となる。時間的制約をかけることで，当該単語の同定や，更には単語間の統語的，意味的，談話的関係を捉える認知的作業に割く時間が限られてくるため，より容易に（迅速に）処理が行なえなければ，シャドーの精度に影響を与えると期待される。筆者らはこれまで，母語話者音声に対する学習者シャドーイング音声（外国語教育の現場で広く行われる通常のシャドーイング）の「崩れ」を自動的に計測する手段として，音素事後確率に基づく定量化を検討してきた (Yue et al., 2017)。本研究で

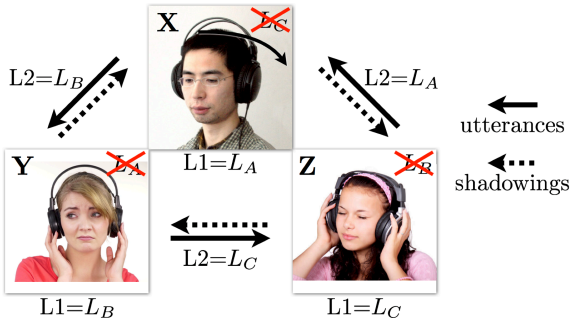

図 1: 学習者相互シャドーイング

はこの技術を母語話者による学習者音声シャドーイングに適用し，母語話者にとってのシャドーイングの容易さ（shadowability）を，可解性と解釈することの妥当性を実験データを通して論じる。

4　学習者相互シャドーイング

議論を進める前に，学習者音声の（可解性）評価のために，母語話者シャドワーを事前に用意する必要性について言及する。Lang-8 という外国語教育支援サイトでは，学習者が学ぶ言語で書いた文章を，その言語の母語話者が添削し，その母語話者がある言語を学んでいる場合は，その言語の作文を，更にその言語の母語話者が添削する，という，学習者相互依存型の支援を実現している。どの学習者も第一言語を持っており，母語話者として他者を支援し，学習者として他者から支援される訳だが，この枠組みをシャドーイングに導入する（図 1 参照）。即ち，学習者相互のシャドーイングインフラ（互いが互いをシャドーし，シャドーされる[1]）を構築すれば，学習者音声の可解性自動評価において，母語話者シャドワーが必要となる前提は，決して大きな問題ではないと考えている。

更に，学習者音声と母語話者シャドーイング音声の対が大量に入手できれば，任意の新たな学習者音声に対して，どのようなシャドーイング音声が得られるのか，それを予測する技術的枠組みも検討可能となる。この段階まで来れば，母語話者シャドワーは不要となる。

以下，母語話者シャドーイングにおけるシャドーの精度（shadowability）と，その母語話者がシャドー時に主観的に感じた可解性（comprehensibility）との関係を実験的に検討する。

5　母語話者による学習者音声シャドーイング実験

5.1　ベトナム人学習者によるカラオケ式読上げ日本語音声の収録

L2 を日本語，L1 をベトナム語として学習者音声の収録を行なった。母語話者にシャドーイングを課す場合，提示音声の話速が遅すぎると（ベトナム語訛りが強くても）可解性は高くなるため，話速を統制しつつ音声収録を行なった。具体的には中級レベルの音読用教科書 (松浦・福池・河野・吉田, 2014) に添付されている音声 CD を用い，モデル話者の話速に従って読上げテキストの文字色が変わる，カラオケ式音声収録ソフトを作成し，それを用いて学習者からの音声を集めた。

教科書から，固有名詞等のない 10 文章を抜き出し，6 名のベトナム人学習者（3 名は上級，3 名は中級）と 6 名の母語話者から，カラオケ音読ソフトを使って音声収録した。最終的に，学習者一人当り約 100 音声，母語話者一人当り 164 音声が得られた。この中から，学習者の習熟度を考慮し，ベトナム人日本語 96 音声（VJ）と，母語話者 68 音声（NJ）をシャドーイング用提示音声として用いた。これらは，フレーズ（句）を単位とした音声であり，全て，異なるフレーズである。

5.2　シャドーイング精度に関する三つの指標

VJ96 音声，NJ68 音声に対する母語話者シャドーイング音声に対して音声分析・音声認識技術を適用し，shadowability に関係すると思われる，下記の音声特徴量を抽出した。

[1]学習者が自身の母語を学ぶ外国人に音声指導することは，専門知識がなければ困難である（故に Speech-8 は存在しない）。しかし，一般の母語話者が学習者音声をシャドーする場合に，特別な専門知識が必要となることはない。

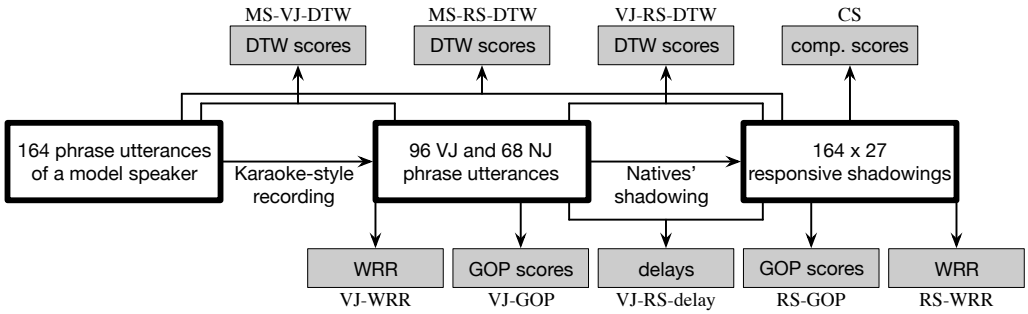

図 2: 母語話者による学習者音声シャドーイング実験と算出した各種音声特徴量

5.2.1 音素事後確率に基づく GOP スコア（DNN-GOP）

提示音声に対して単語の同定ができない場合（了解性が低い部位があると），シャドーイングが崩れる（不適切な調音制御となる）ことが多い。調音制御の正確さを示す尺度として音素事後確率を導入する。ある時刻 t のスペクトル特徴量 o_t に対して,「o_t が母語話者の発声だとするとどの音素 c_i が意図されたのか」を確率分布として表現したのが音素事後確率（$P(c_i|o_t)$）である。深層学習に基づく音声認識技術では，その front-end モジュールにて，スペクトル特徴量を音素事後確率に変換しており，このモジュールを利用する。事後確率化することで，話者や年齢など非言語的特徴を凡そ抑制できる。シャドワーに提示した音声の音素表記を用いてシャドーイング音声から音素境界を推定し，個々の音素区間に対して当該音素の事後確率を求め，それを音素数だけ累積して平均化する。即ち，学習者音声評価タスクにおいて標準的技術として使われる GOP（Goodness Of Pronunciation）スコアを，母語話者シャドーイング音声に適用する。もちろん，GOP スコアは学習者音声（即ちシャドーイングの提示音声）からも算出である（本来この目的のために提案された技術である）。

5.2.2 事後確率ベクトルに基づく発話比較（DNN-DTW）

GOP 計算では，提示音声の音素表象は用いるが，提示音声とシャドーイング音声を直接比較することはない。シャドーイング音声を音素事後確率分布の時系列とすることができるように，提示音声も同様に音素事後確率分布の系列に変換できる。この二つの系列を，時系列の自動対応付け技術である，Dynamic Time Warping（DTW）で対応づけ，両者の差異を定量化する。今回の実験の場合，1) 教科書 CD のモデル音声，2) カラオケ式収録で得られたシャドーイング対象の音声，3) 母語話者によるシャドーイング音声と 3 種類の音声があるため，任意の 2 対間で DTW 距離を算出できる。

5.2.3 シャドーイング遅れ

DNN-GOP, DNN-DTW ともにシャドー時の調音制御の乱れを定量化することを意図しているが，シャドーイングの遅れについても自動計測した。提示音声，シャドーイング音声両者から音素境界を推定し，その時間差を音素境界の数だけ累積し，平均化する。この平均遅れを，そのシャドーイング音声のシャドー遅れとして定義する。

5.3 母語話者シャドーイング実験

27 名の正常な聴力をもつ成人母語話者がシャドーイング実験に参加した。シャドーイングは web 上に構築した専用のプログラムを通して実施した。クリック後，1 秒して音声提示が始まり，被験者はシャドーする。シャドーイング音声は，イヤーフックマイクを通して収録した。シャドーに関しては,「意味を捉え，ベトナム語特有の訛りを真似せず，母語話者として日本語でシャドーするよう」指示した。また，各フレーズ音声のシャドー後に，提示された音声の理解しやすさについて 7 段階で評価させた（comprehensibility，可解性の主観的評価）。実験全体の流れと，自動計測した音声特徴量を図 2 に示す。MS は Model Speech，RS は Responsive Shadowing，CS は Comprehensibility Score の略である。また，WRR とは Word Recognition Rate を意味し，深層学習型音声認識の標準ツールキッ

表 1: 各種音声特徴と可解性（CS）との相関

VJ-GOP	VJ-WRR	MS-VJ-DTW		
0.58	0.47	-0.52		
RS-GOP	RS-WRR	VJ-RS-DTW	MS-RS-DTW	VJ-RS-delay
0.74	0.53	-0.55	-0.58	-0.59

トである KALDI を，CSJ コーパスに適用して構築された CSJ-KALDI を用いた精度を意味する。

従来学習者音声の評価に使われてきた音声特徴量は，VJ-GOP，VJ-WRR，MS-VJ-DTW である。本研究では，母語話者発音との近接性ではなく，学習者音声の可解性（CS）を主眼に置いた評価を検討している。この場合，VJ-GOP，VJ-WRR のように学習者音声から得られる特徴量と，RS-GOP や RS-WRR，更には VJ-RS-delay や MS-RS-DTW など母語話者シャドーイング（RS）があって初めて計測可能となる特徴量のどちらがより可解性と相関が高いのか，が検討の中心的課題となる。

5.4 結果と考察

VJ のフレーズ単位の 96 発声の各々について，GOP スコア（VJ-GOIP），音声認識精度（VJ-WRR），モデル音声との DTW 距離（MS-VJ-DTW）が算出される。また，27 名の被験者から CS スコアが得られる。この CS スコアの被験者間平均を，当該 VJ 音声の CS スコアと定義する。一方，母語話者シャドーイング音声の各々について，GOP スコア（RS-GOP）や，音声認識精度（RS-WRR），更には VJ 発声からの遅れ（VJ-RS-delay），VJ 発声やモデル発声からの DTW 距離（VJ-RS-DTW, MS-RS-DTW）が計測される。即ち，母語話者シャドーイング音声の各々に対して，これら五種類の音声特徴の被験者間平均が計算できる。各 VJ 発声の CS スコアに対する相関値を表 1 に示す。母語話者シャドワーに提示した VJ 音声よりも，彼らのシャドーイング音声からの音声特徴の方が，VJ 音声に対する主観的可解性と，高い相関を示していることが分かる。従来学習者音声を自動評定する場合，学習者音声を分析対象としていたが，上記の結果は，学習者音声よりもそれを聴取及びシャドーした，母語話者音声の方が分析対象としてより適切であることを意味している。評価基準を母語話者音声との近接性とすれば，学習者音声を分析対象とすべきだが，評価基準を可解性とした場合は，母語話者シャドーイング音声の方へ着眼すべきである。更に，単なる音声認識結果よりも，音素事後確率化した方が，主観的な可解性との相関がより高くなっている。

これらの音声特徴を説明変数として用い，Lasso 回帰[2]を用いて，CS スコアを予測する実験を行なった。交差検定の結果，予測値と実測値（CS 平均値）との相関は 0.81 となった。これは，26 名の被験者の CS 平均値と残り一人の被験者の CS スコアとの相関値の平均，0.66 を大きく上回った。

5.5 Shadowability は online intelligibility なのか，それ以上なのか？

学習者音声よりも，その音声を母語話者にシャドーさせ，そのシャドーイング音声の崩れ（調音的崩れ，シャドーの遅れ）を測定した方が，聴取者の主観的可解性と高い相関を示すことが実験的に示された。これら shadowability は，その測定方法を考えれば，comprehensibility よりも，時間制約付き intelligibility，即ち online intelligibility と解釈すべき指標とも言える。intelligibility と comprehensibility は個々の単語の同定のみに着眼するのか，単語間の関係の把握（即ち意味の把握）までに着眼するのかが異なる。shadowability が intelligibility と comprehensibility のどちらに近いのかを考える場合，shadowability が（提示音声の）意味の把握の容易さによって左右されるのか否か，を検討することになる。以下，筆者らの検討 (Trisitichoke et al., 2018) を紹介する。

本節では外国語訛りが混入された提示音声を用い，それによって shadowability が変わる様子を示した。一方 Trisitichoke et al. (2018) では，プロのナレータによる読上げ音声を用いており，読上げ文の内容を制御することで，意味理解の難易度が定性的に異なる文章音声に対する母語話者シャドー

[2]過学習とならないよう，正則化による制限を導入した線形回帰モデル

表 2: 6 種類の読上げ音声

A	桃太郎
B	NHK NEWS WEB EASY (NWE) 中の文章
C	NHK NWE 中の内容語のランダム列
D	NHK NWE 記事に対する元原稿
E	日経サイエンスからの記事
F	無意味モーラ（清音のみ）列

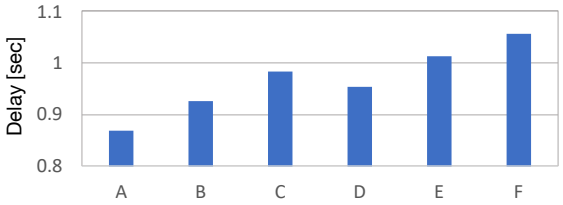

図 3: 母語話者シャドーイングの遅れ

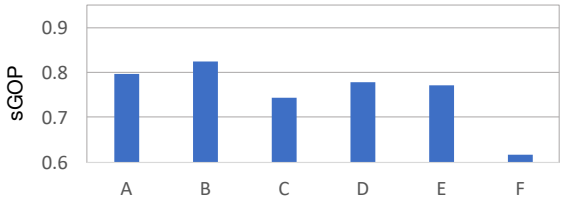

図 4: 母語話者シャドーイング音声の GOP

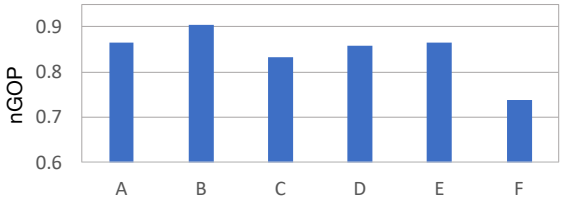

図 5: プロのナレータ読上げ音声の GOP

イングを検討している。準備した文章音声を表 2 に示す。平易な日本語文として，日本語学習者向けに用意されたニュース文（**B**）を用い，一般的な日本語文として，内容は異なるが NWE 記事に対するオリジナル原稿（**D**）を用いた。また，語彙は平易だが統語構造も意味構造もない刺激として，NWE に出現する内容語のランダム列（**C**）を用意した。**A** は，次の単語やフレーズが容易に予測できるが，日常語以外の語彙も含まれる文章を意図しており，桃太郎の読上げ音声を用いた。

詳細な実験条件や有意差検定結果は Trisitichoke et al. (2018) を参照して戴きたいが，7 名の母語話者シャドワーが示した shadowability として，遅れと GOP の（被験者間及びセット内）平均を文セット毎に図 3，図 4 に示す。更にはプロのナレータによるモデル音声に対する GOP スコアも図 5 に示す。表 2 に示した文セットの意味的難易度にほぼ従う形で，シャドーイング音声の遅れや GOP が分布している。興味深いことに，被験者の GOP スコアと非常に類似した変化パターンが，プロのナレータ音声にも観測された。ナレータ音声の収録は事前のリハーサル後に行なっており，図 5 の GOP スコアの変動は，各文章セットの意味理解の難易度に起因するものと考えられる。当然，通常の母語話者が読み上げる場合でも，意味理解の困難さが，その音声に影響を与えていることが示唆されるが，タスクを読上げから復唱（repetition）や追唱（shadowing）に変えても，同様の傾向は観測されるであろう。図 4 はシャドー時の結果である。Mineamtsu et al. (2011) では，母語話者の復唱結果を書き起こし，その書き起こし結果に基づいて了解性を算出している。書き起こしではなく，復唱音声を分析すれば，可解性と類似した計量が可能であったと推測される。タスクを復唱から追唱にすれば，より可解性へ迫ることが可能になると考えられる。

6　学習者相互シャドーイングが可能にするもの

本研究では，学習者音声を母語話者にシャドーさせ，そのシャドーイング音声の崩れを計測することで，聴取者が学習者音声に対して感じる可解性を効果的に予測できることを実験的に示した。また，シャドーイングというタスクそのものが，単に提示音声中の語の同定及びその復唱のみを意味せず，提示音声の意味理解に依存した発話活動となっている様子も実験的に示した。

本節では，この学習者相互シャドーイングを通した音声評定方式が有する，幾つかの興味深い側面について論じる。一般の外国語教育現場では，学習者の動機付けを大切にするあまり，学習意欲が阻害されないよう，各学習者への教示内容（言葉遣い）に配慮することがある。所謂，学習者（教える側にとっては，学習者は顧客となる場合もある）への忖度である。初級者に対しては有効な教示戦略であると思われるが，学習過程の進捗と共に，より厳しい，本音の教示を返す必要もあろう。学習者音声をシャドーイングする場合，忖度とは，上手なシャドーを返してあげることを意味するが，

事前に学習者の発声内容を知らない限り，このような忖度シャドーイングは不可能である。即ち，学習者間の相互シャドーイングは，母語話者が聴取時に感じる本音の「聴き取り易さ」（可解性）を直接的に返す場を学習対象言語に依らず，全外国語学習者に対して導入する，という意味を持つ。

了解性にせよ，可解性にせよ，これらは，十分聴き取り易い発音であればどれだけ母語話者発音からずれていても是とする教育戦略と関係する。さて，ある学習者の外国語発音を，最も了解性・可解性が高いと評価する聴取者は誰だろうか？筆者らが考えるに，それは学習者本人である。「十分聞き取れるから，発音を改善する必要性が理解できない」という学習者がいても不思議ではない。了解性・可解性を強調すればするほど，そのような学習者は増えるのかもしれない。学習者相互シャドーイングは，他者となる母語話者にとっての可解性を直接的に返す枠組みであるが，それ以上の機能を持つ。日本人が自身の英語音声を母語話者にシャドーしてもらうと共に，日本語学習者の日本語音声をシャドーする。二種類のシャドーイング音声を使えば，以下が可能となる。その学習者の英語音声が，（母語話者）シャドーイングをどの程度崩すのか，その「崩れ」と等価な「崩れ」をもたらす外国人の日本語音声を，その日本人英語学習者に呈示できる。即ち，自身の英語音声と（可解性という尺度において）等価な外国人の日本語音声を提示できる（外国語訛りの母語音声を通して，自身の英語音声を把握させる）。可解性がレベル C とか，30 点とか論理的に示すよりも，学習者の感覚に訴えられる，説得力のある教示となるかもしれない。なお，機械による合成音声を品質評価する場合，評価者が対象言語の母語話者なのか否かは明確に区別して行われる。その理由は，母語話者の方が不自然さに対する感度が極めて高いからである。この「不自然さに対する高い感度」をもって，自身の英語発音の現状を把握させることが可能となるだろう。理論的には学習者を，極めて甘い評価者から，極めて厳しい評価者へと変遷させることになる訳だが，まだ何ら実験的検証結果はない。現在，実験的検証に向けて，データ収集の準備を行なっている段階である。

7　まとめ

本研究では，学習者音声を母語話者にシャドーさせ，その崩れを計測することが，可解性に基づく評価を，客観的に，定量的に，かつ，容易に実現できることを示した。様々な言語の学習者を繋いで，学習者間の相互シャドーイングを音声学習インフラとして実現した場合に期待される効果についても記述した。後者については何ら実験的検証を行なっていないが，ベトナム・ハノイにて日本語学習者音声の大規模収録を計画しており，実験的検証を行なう環境を整えつつある。

本研究は，JSPS/MEXT 科研費 JP26118002, JP26240022, JP18H04107 の支援を受けた。

参考文献

Derwing, T. M., & Munro, M. J. (2015) *Pronunciation fundamentals: Evidence-based perspectives for l2 teaching and research*. Amsterdam: John Benjamins Publishing.

Mineamtsu, N., et al. (2011) “Measurement of objective intelligibility of japanese accented english using erj database.” In *Proc. INTERSPEECH* (pp. 1481–1484).

Munro, M. J., & Derwing, T. M. (1995) “Foreign accent, comprehensibility, and intelligibility in the speech of second language learners.” *Language Learning*, 45, 73–97.

Trisitichoke, T., et al. (2018) “Influence of content variations on native speakers’ performance of shadowing.” In *Proc. ASJ autumn meeting.*

Yue, Y., et al. (2017) “Automatic scoring of shadowing speech based on dnn posteriors and their dtw.” In *Proc. INTERSPEECH* (pp. 1422–1426).

電通 サイエンスジャム (2016) *Kansei analyzer*. `https://kansei-analyzer.com`.

松浦 真理子・福池 秋水・河野 麻衣子・吉田 佳世 (2014) 『日本語音読トレーニング』東京: アスク出版.

河原 達也・峯松 信明 (2013) 「音声情報処理技術を用いた外国語学習支援」 『電子情報通信学会論文誌』96: 7, 1549–1565.