Vol.2021-MUS-131 No.50
Vol.2021-SLP-137 No.50
2021/6/19

逆シャドーイング法を用いた瞬時了解度アノテーションとその高精度化に関する分析的検討

箱田 峻[1]　朱 伝博[1]　齋藤 大輔[1]　峯松 信明[1]　中西 のりこ[2]

概要：コロナウィルスの影響により諸外国との往来は激減したが，オンラインツールの発達により，依然として英語を用いた国際コミュニケーション能力の習得は重要である．英語発音の指導では，「母語話者らしい発音」の習得を目指すのではなく，「十分相手に伝わる発音」を目指すことが多い．この場合，発話の「伝わりやすさ」すなわち「了解度（intelligibility）」を，任意の発話に対して定量的にスコア化（ラベル付け）する必要がある．筆者らはこれまで，学習者音声を母語話者（相当）にシャドーさせ，そのシャドー音声の崩れを分析する逆シャドーイング法を検討してきた．逆シャドーイング法では，シャドー音声の他に（学習者が読み上げ時に参照した）原稿を見ながら行うスクリプト・シャドー音声も別途収録し，両者の posteriorgram DTW のスコアを，自動計測された了解度スコアとして用いている．本研究では，シャドー音声の崩れの様子を考慮し，また DTW 計算において，スクリプト・シャドー音声が常に規範音声として使われることを鑑み，非対称な局所パスを導入し，精度向上を検討した．更に，学習者音声やシャドー音声に対する音声認識結果（自動書き起こし）を使うことの是非についても検討した．

キーワード：英語教育，瞬時了解度，逆シャドーイング，スクリプト・シャドーイング，DTW

1. はじめに

外国語学習の発音指導において，母語話者らしい発音を学習者に求めるのはコストが高く，また，母語訛りはその学習者のお国柄（民族的な identity）を表すため，それを否定することは教育的に不適切である．即ち，相手に十分伝わる発話能力の獲得が目的となる [1–5]．ある L2 音声に対して，その伝わりやすさ（了解度，intelligibility）をスコア化する場合，二つの言語的多様性を考慮する必要がある．一つ目は，発話者，つまり学習者の多様性である．様々な母語の学習者が英語を学んでおり，その結果，多様な母語訛りが存在する [6]．もう一つは，聴取者の多様性である．ある言語を母語とする英語利用者は，どのような英語発音が聞き取りやすく／聞き取り難いのか，を考えると，了解度は聞き手の言語背景に大きく依存する [7, 8]．即ち了解度とは，話者と聴取者，両方に依存することとなり，両者に依存して活用できる簡便な計測方法が求められている．

従来，外国語教育や応用言語学の分野では，L2 音声の了解度を計測する手法として，音声を聴取者（多くは母語話者）に書き起こさせ，その書き起こしと L2 音声の原稿（スクリプト）との単語単位の一致度を算出している [9–13]．一般的に単語を単位とすることが多いが，音節や音素など細かい単位を採択する場合は，原稿や書き起こしを，音節や音素を単位として表記しなおして比較すれば良い．聴取者の多様性を考慮するならば，上記の作業を様々な言語背景の聴取者を対象として実施することになる．

しかしながら，この手法には大きな問題が存在する．まず書き起こしは一般に，提示音声の数倍の時間を要する．了解度計測は複数回聴取は許されず，また，聴取者の短期記憶の容量にも制限があるため [14, 15]，短文しか利用できない [10–13, 16]．更に，書き起こし中に発話内容を深く推測する余地を与えることとなる．以上の問題は，従来方法が聴取プロセスを「聴取の後」に観測していることがそもそもの原因であり [11]，聴取プロセスを「聴取中」のみ，観測することで解決される．

筆者らは，発音訓練として学習者に広く導入されているシャドーイングを，母語話者（相当）に導入し，彼らが L2 音声を聴取している様子をモニタリングする手法を提案した（逆シャドーイング法）[17–20]．従来法では L2 音声を書き起こし「その書き起こしがどのくらい崩れたのか」を計測するが，逆シャドーイング法では，L2 音声をシャドーし「そのシャドー音声がどのくらい崩れたのか」を計測す

[1] 東京大学
The University of Tokyo, Bunkyo, Tokyo, Japan
[2] 神戸学院大学
Kobe Gakuin University, Kobe, Hyogo, Japan

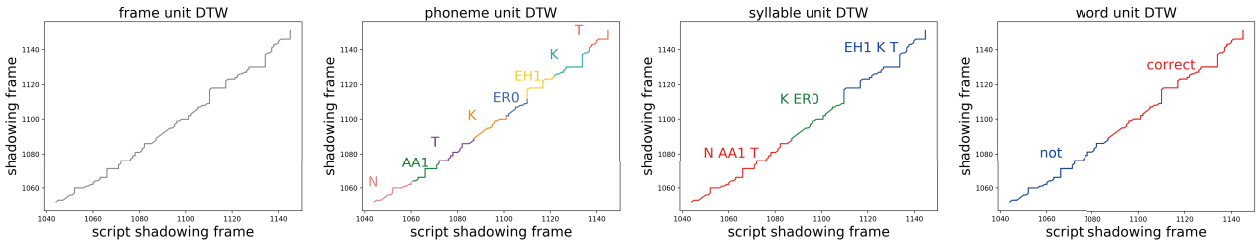
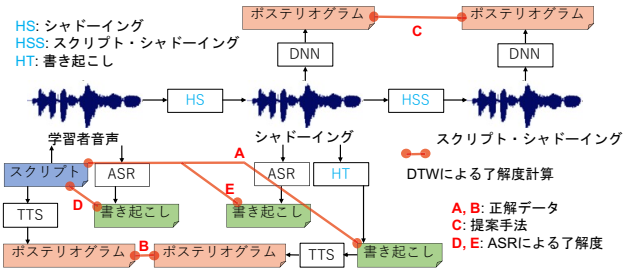


図 2　4 つの言語単位による学習者音声 “not correct” の DTW



HS, HSS, HT の H は人間（human）による作業という意図である。
図 1　本実験全体の概略図

る．この場合，推測の余地が入らず，また，30 秒程度であればシャドーし続けることは容易である．即ち，聴取しているその場での了解度，瞬時了解度をスコア化できる．本研究では，逆シャドーイング法において了解度を算出する際のアルゴリズムを工夫し，より高精度な了解度のアノテーションができるか検討した．

なお，本報告の実験の全体像を**図 1** に示す．以降の説明は適宜，この図を参照しながら行う．

2.　逆シャドーイング法

2.1　シャドーイングと逆シャドーイング

シャドーイングとは，モデル音声を聴きながら即座に復唱する訓練法のことであり，外国語学習において広く導入されている．逆シャドーイングとは，シャドーする／される関係性を逆にしたもので，L2 音声を母語話者（相当）の聴取者がシャドーする．この場合，外国語訛りまでを真似ることは求めず，学習者が意図した単語列を自身の発音で復唱することが求められる．逆シャドー音声の崩れの定量化は，シャドーの後に，（学習者が参照した）原稿・スクリプトを見ながらのシャドーである，スクリプト・シャドーも課し，二種類のシャドー音声の比較を通して定量化する [20]．スクリプト・シャドーは内容既知であり，一番流暢なシャドー音声と解釈でき，それを基準としてシャドー音声の崩れを計測する．両音声を posteriorgram-DTW 法によって比較すれば，平均 DTW スコア（調音崩れの平均に相当）が瞬時明瞭度の指標となる（**図 1C** ）．

2.2　複数粒度による逆シャドーイング分析

Posteriorgram DTW で利用する局所距離尺度としてはバタチャリヤ距離を用い，また，音声の posteriorgram 化には，WSJ-KALDI [21] のレシピを用いたフロントエンドを用いている．平均 DTW スコアはフレーム，音素，音節，単語の 4 つの単位で求めることができる．フレームは音響的分析の単位，音素は発音の最小単位，音節はリズム知覚，聞こえに関する単位，単語は音声コミュニケーションの最小単位である．DTW 処理はフレームを単位として局所距離を累積するが，構成された DTW パスを音素，音節，単語で区分（forced alignment）し，各単位でのサブスコアを求め，これらのサブスコアの（全パスに対する）平均値を計算する．各単位での DTW 平均スコアを，fDTW (frame), pDTW (phoneme), sDTW (syllable), wDTW (word) と以降，呼ぶこととする．**図 2** に，学習者発声 “not correct” に対するシャドー音声，スクリプト・シャドー音声から得られた DTW パスと，4 種類のスコア計算の様子を示す．言語単位の明示的な導入は，その単位の継続長の違いを正規化する効果を有する．例えば母音は長く，子音は短いが，fDTW は継続長の偏りを正規化していないため，母音のシャドー崩れの影響が出やすい．pDTW は音素継続長の偏りは回避できるが，例えば，長い名詞と短い名詞の差は正規化できていない．wDTW は単語毎にサブスコアを計算するので，単語単位での継続長の正規化が行われている．了解度と言っても，どの粒度で了解度を検討したいのかは教育指針に依存する．教育指針に柔軟に対応できるよう，複数粒度での分析を行なっている．

3.　書き起こしによる瞬時了解度の算出

逆シャドーを課すことで，瞬時的聴解の崩れを定量的に予測することが可能であることを説明した．本節では，得られた自動スコアに対する ground truth である，手動スコアについて説明する．従来研究では，学習者音声を書き起こし，これと学習者が参照した原稿とを比較していたが，本研究では，母語話者（相当）によるシャドー音声の手動書き起こしと，学習者の原稿とを比較する（**図 1A**）．即ち，自動スコアである「シャドー音声とスクリプト・シャドー音声の DTW スコア」に対応する手動スコアとして，「シャドー音声の書き起こしと学習者の原稿との差異」を使う．書き起こし，原稿，両方とも単語書き起こしであるが，これを発音辞書を参照して音素書き起こし化し，また，音節分かち書きすることで，音節書き起こし化できる．即ち手動スコアも，単語，音節，音素の粒度のスコアを準備できる．以上は二種類のテキスト比較に基づくスコア化である

が，このテキストを音声合成により音声化し，posteriogram DTW を通して比較することもできる（**図 1B**）．

3.1 記号的な書き起こしとその比較

シャドー書き起こしと学習者が参照した原稿は，文字列を対象とした DTW で比較する（**図 1A**）．シャドー音声中に検出された置換誤り（S），削除誤り（D），挿入誤り（I）の数を用いて計算される正解率 %Accuracy（Acc）$=(N-D-S-I)/N$（N は単語総数）を，書き起こしに基づく瞬時了解度の手動スコアとする．以上を単語，音節，音素を単位として行い，wAcc, sAcc, pAcc を計算する．

pAcc と sAcc では，wAcc よりも，シャドワーの聴解精度をより詳細に評価できる．例えば，“conception” と “consumption” を単語単位で比較する場合，両者は異なるとだけ判定されるが，音節単位で評価した場合，中間の /sep/ と /sʌmp/ が異なっているだけで，それ以外は合致する．音素単位の評価すると， /e/ が /ʌ/ と /m/ に変化しただけである．より小さい単位での評価により，より細かい比較が可能だが，後述の確率的な書き起こし（posteriorgram）を用いれば，より細かい比較が可能となる．

3.2 確率的な書き起こしとその比較

上記の/e/が/ʌ/と/m/に変化した場合では，/e/と/ʌ/との違いの方が，/m/との違いよりも小さい．これは，/e/と/ʌ/がともに母音である一方，/m/は子音であるためであるが，pAcc による記号的な評価では，音素間の音声学的差異が全て等価に扱われてしまう．そこで，シャドー書き起こしと，学習者の原稿を音声合成器（Amazon Polly [22]）により一旦音声化し，posteriorgram 化して比較する．二種類のテキストの差異の定量化を，posteriorgram を通して行うことで，より詳細な差異の定量化が可能になる（**図 1B**）．なお，posteriorgram は音素事後確率ベクトルの時系列であるので，確率的な音素書き起こしと解釈でき，逆に通常の音素書き起こしは，one-hot ベクトルだけで構成された posteriorgram と解釈できる．posteriorgram DTW を書き起こしに適用したものを S-DTW (Synthesis-based DTW) とする．S-DTW も，複数の粒度で算出できる．

4. 非対称 DTW パスの利用

筆者らの先行研究では，Posteriorgram DTW で用いる局所パスとして**図 2-1** に示す，最も一般的な，対称的な局所パスを用いていた．この場合，比較する二発声が等価に扱われる．しかし逆シャドーイング法の場合，比較対象はシャドー音声とスクリプト・シャドー音声であり，発話の崩れは常に前者にあり，後者は常に模範発声として扱われる．また，シャドー音声中に散見される（無音や繰り返しの）挿入や，発声の間延び [23] は，発話崩れとは考えずにスコア計算から除外すべき発話現象である．これらを考慮して，**図 2-2** に示す非対称局所パスも検討した．これを用いると，局所距離の累積は常にスクリプト・シャドー音声の時間軸に沿って行われ，累積回数は当該音声長と等しくなる．即ち，スクリプト・シャドー音声の各時刻を等価に扱える．なお，このままだと母音長，子音長の偏りの影響を直接的に受けることになるので，非対称パスによる DTW 結果も，音素，音節，単語を単位として再集計する．

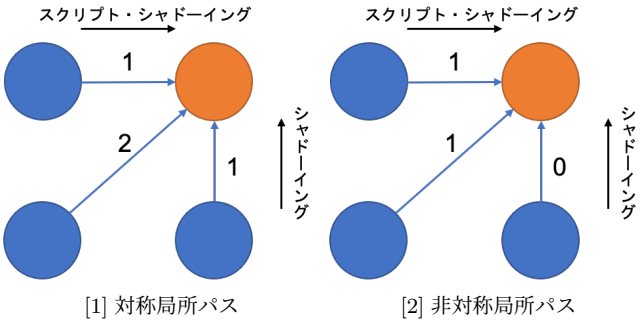


図 2 2 種類の局所パス

5. 実験

5.1 実験条件

了解度を計測する学習者音声として，30 人の日本人大学生に英語で作文させ，それを音読させて録音した．内容は学術研究から社会問題に至るまで多岐にわたる．音声長は 30〜50 秒で，文法ミスや発音間違いが一部含まれている．学習者は Versant English Test [24] を受けており，そのスコアから，幅広い習熟度を一様に網羅できる 12 名を選定した．それらに米国人，英国人の母語話者の英語音声を加え，合計 14 音声を提示刺激として用いた．なお，聴取により，音声長が約 30 秒となるよう編集した．

5.2 実験手順

14 種類の音声をシャドーするシャドワーとして，英語母語話者（N1, N2），英語上級者であるが非母語話者（NN1–NN3），英語上級者の日本人（J1, J2）の 3 グループのシャドワーが本実験に参加した．最初の二グループは日本語の知識はない．なお，NN1–NN3 の母語は，ベトナム語が一人，中国語が二人である．シャドワーは各々の音声をシャドーし，その後スクリプト・シャドーした．

シャドー音声の録音はウェブ上で行った．練習として，英検 2 級のリスニング問題用 3 音声と日本人英語 1 音声を用意した．練習は，希望に応じて複数回行わせた．実験刺激である 14 音声の方は，シャドー，スクリプトシャドー，共に一度ずつ行わせた．その後，シャドーイング音声全てを書き起こし，書き起こしによる了解度を計算した（**図 1A**）．音声合成による音声化，posteriorgram 化することによる了解度も計算した（**図 1B**）．

了解度のサンプル数を増やすため，読み上げ原稿を文ごとに区切り，posteriogram を分割し，それぞれ独立に了解

表 1 手動スコア (**A**, **B**) と自動スコア (**C**, **D**, **E**) の相関

学習者音声全体に対して集計した場合

Corr.	J1	J2	N1	N2	NN1	NN2	NN3	mean
D-**A**	0.481	0.501	0.734	0.626	0.629	0.728	0.743	0.635
E-**A**	0.874	0.690	0.889	**0.854**	0.904	0.910	0.847	0.853
C-**A***	0.814	0.920	0.886	0.847	0.920	0.923	0.914	0.889
C-**B***	**0.890**	**0.932**	**0.961**	0.835	**0.961**	**0.951**	**0.976**	**0.929**

学習者音声中の文毎に集計した場合

Corr.	J1	J2	N1	N2	NN1	NN2	NN3	mean
D-**A**	0.334	0.275	0.520	0.425	0.376	0.398	0.499	0.404
E-**A**	**0.785**	0.583	**0.869**	**0.776**	0.791	0.829	0.795	0.776
C-**A***	0.699	0.722	0.736	0.771	0.787	0.822	0.825	0.766
C-**B***	0.760	**0.805**	0.824	0.770	**0.806**	**0.860**	**0.873**	**0.814**

∗ 相関は負となるが，符号は除いている．

度を算出した．また，単語数の少ない文では， text-based DTW のスコアリングの粒度が極端に粗くなるため，8 単語以上の文のみ分析対象とし，全 64 文のうち，それ未満の 12 文は除外した．最終的に 14 のシャドー音声から，学習者発話全体に対するスコアが 14，文毎のスコアが 52，得られることになる（**図 1C**）．**図 1A**，**B** に示す手動スコアも同様，発話全体のスコア，文毎のスコアを用意した．

5.3 実験結果と考察

5.3.1 ASR による書き起こしとの比較

本実験の結果を示す前に，シャドー音声を人手で書き起こす代わりに，シャドー音声（および学習者音声）を音声認識して得られる（自動）書き起こしを用いた場合の結果を示す．シャドー音声の認識結果が十分機能すれば（**図 1E**），スクリプト・シャドー音声は不要となる。また，学習者音声の認識結果が了解度を示す指標として十分機能すれば（**図 1D**），逆シャドーそのものが不要となる．Amazon の米英語用音声認識器 [25] を用い，サブスコア計算の単位は単語であり，DTW の局所パスは対称型を用いている．

表 1 に手動スコアである **A**，**B** と，自動スコアである **C**，**D**，**E** との相関を，各学習者発話全体に対して求めた場合と，文毎に求めた場合の結果を示す．**D**-**A** の相関は低く，音声認識結果を，人間の即時聴解を模擬するために使うこと適切ではないだろう．一方で **E**-**A** の相関は高く，聴取者のシャドー音声に対する音声認識結果を使うことは一定の効果がある．L2 音声全体を用いた場合は平均値では，**C**-**A** が **E**-**A** より高いが，文毎に集計すると後者の方が高くなった．シャドワー別にみると，J2，NN3 の相関は **C**-**A** の方が高い．彼らの英語音声を聞くと，母語訛りが比較的強く，音声認識精度が十分に高くなかったことが考えられる．一方，シャドー音声とスクリプト・シャドー音声の DTW 比較は，当該話者間で発話を比較しており，外国語訛りには頑健に対応できている．国際語である英語は，世界諸英語（World Englishes）とも呼ばれるように，ある英語音声の了解度を議論する場合，様々な母語を持つ英語利用者に対して調査する必要がある．そのような状況を考えると，外国語訛りへの頑健性は大きな意味を持つ．

シャドー音声に対する wDTW はバタチャリヤ距離の平均値として計算されるが，書き起こしに対する wAcc は，単語の合致／非合致を 1/0 で計算しており，スコアリングの粒度に乖離がある．後者を posteriogram 化してスコア計算すると，この乖離は解消される（**図 1B**）．**表 1** 中，最高相関値は，多くの場合で **C**-**B** で得られている．

表 2 各種手動スコア（**A**）と各種自動スコア（**C**）の相関

シャドワー間平均，対称局所パス使用

	A	**C**	mean (発話毎)	mean (文毎)
1	pAcc	fDTW	0.841	0.726
2	pAcc	pDTW	**0.904**	**0.813**
3	sAcc	fDTW	0.844	0.751
4	sAcc	sDTW	**0.890**	**0.816**
5	wAcc	fDTW	0.848	0.726
6	wAcc	wDTW	**0.889**	**0.766**

相関は負となるが，符号は除いている．

表 3 各種手動スコア（**B**）と各種自動スコア（**C**）の相関

シャドワー間平均，対称局所パス使用

	B	**C**	mean (発話毎)	mean (文毎)
1	S-fDTW	fDTW	0.875	0.743
2	S-pDTW	pDTW	0.922(+0.018)	0.845(+0.032)
3	S-sDTW	sDTW	0.915(+0.025)	0.827(+0.011)
4	S-wDTW	wDTW	0.929(+0.040)	0.814(+0.048)

括弧内は**表 2** からの増分

表 4 各種手動スコア（**A**）と各種自動スコア（**C**）の相関

シャドワー間平均，**非**対称局所パス使用

	A	**C**	mean (発話毎)	mean (文毎)
1	pAcc	fDTW	0.897(+0.056)	0.795(+0.069)
2	pAcc	pDTW	**0.904**(+0.000)	**0.820**(+0.007)
3	sAcc	fDTW	0.893(+0.049)	0.809(+0.058)
4	sAcc	sDTW	**0.896**(+0.006)	**0.830**(+0.014)
5	wAcc	fDTW	0.863(+0.015)	0.763(+0.037)
6	wAcc	wDTW	**0.893**(+0.004)	**0.780**(+0.014)

括弧内は**表 2** からの増分

表 5 各種手動スコア（**B**）と各種自動スコア（**C**）の相関

シャドワー間平均，**非**対称局所パス使用

	B	**C**	mean (発話毎)	mean (文毎)
1	S-fDTW	fDTW	0.922(+0.047)	0.802(+0.059)
2	S-pDTW	pDTW	0.924(+0.002)	0.853(+0.008)
3	S-sDTW	sDTW	0.923(+0.008)	0.842(+0.015)
4	S-wDTW	wDTW	0.935(+0.006)	0.831(+0.017)

括弧内は**表 3** からの増分

5.3.2 複数粒度による了解度アノテーション

表 2 に，対称局所パス使用時の手動スコア（各種 Acc，**A**）と自動スコア（各種 DTW，**C**）との相関を示す。3 種類の粒度の手動スコアに対して，フレーム単位及び当該言語単位の DTW スコアとの相関を各シャドワー毎に算出し，その平均値を示している．また，約 30 秒の発話毎の集計と，文毎に集計を別々に示している．まず，**表 1** 同様，分析対象の音声長が長いほど自動スコアと手動スコアの相関は高くなっている．また，各種粒度（言語単位）に対して，フレーム単位の平均 DTW スコアを直接使うよりも，当該言語単位でサブスコアを計算し，その平均値を最終的なスコアとした方が，いずれの場合も相関が高い．L2 音声の了解度を考える場合，どの単位に基づいて検討するのかは教育方針に依存するが，いずれの方針に対しても，より適切な自動スコアが計算できている．なお，単語を単位と

して L2 音声を文単位で集計した場合の相関が低いが，これは，単語単位の text-based DTW を行うと，単語の合致／非合致を 1/0 で計算しており，文中の単語数が少ない場合，スコアリングの粒度が粗くなることが原因である．

この問題は，書き起こしと原稿との DTW をテキストで行うのではなく，音声化・posteriorgram 化して行う（S-DTW，B）ことで回避できると予想される．各種単位の S-DTW スコアを用いた場合の相関を**表 3** に示す．括弧内の数値は**表 2** からの増分である．いずれの単位でも相関は増加しておりその効果が窺われるが，単語を単位として文毎に集計した場合の増分が一番高い．

5.3.3 非対称 DTW パスによる了解度アノテーション

表 4，**5** に C における DTW に非対称局所パスを導入して，前節と同じ分析を行った結果を示す．いずれの場合も，発話毎，文毎での比較ともに，非対称局所パスの導入によって，手動スコアとの相関は高くなっており，シャドー音声の崩れの特性を適切に捉えられていると言える．また，相関の増分は fDTW での増分が顕著に高く，それ以外は凡そ，wDTW ≈ sDTW > pDTW となっている．非対称局所パスによる fDTW は**図 2** の左図において，縦方向に伸びる局所パスのスコアを無視し，局所スコアの累積は規範発声であるスクリプト・シャドーの各音声フレーム毎に行うこととなる．その結果，fDTW は，手動スコアとしてどの言語単位を使った場合でも，より類似した自動スコアを呈するようになったが，当該単位を使った DTW スコアには至っていない．また，wDTW ≈ sDTW > pDTW の傾向であるが，サブスコアを算出する区間が長いほどシャドーイングが間延びした部分の影響が大きく，非対称局所パスによるその影響の打ち消しの効果が大きいことによる．

なお，対称局所パスを用いた**表 1** では，文毎での比較の場合，wDTW を用いた平均相関（C-A）が，シャドー音声の音声認識結果を用いた平均相関（E-A）を下回っていたが，**表 4** において非対称局所パスを導入することで，wDTW の平均相関が上回る結果となった．最終的にテキスト比較においても posterior-gram DTW を導入することで（**表 5**），音素，音節，単語いずれの言語単位に対しても，発話全体の場合は 0.92 以上の相関を示し，文単位でも 0.83 以上の相関を示すことができた．シャドー音声の手動書き起こしを用いなくても，シャドー音声とスクリプトシャドー音声を確率的に自動音素書き起こしすることで，高い精度で瞬時了解度のアノテーションは取得できる．

6. まとめと今後の課題

本報告では，逆シャドーイング法による瞬時了解度のアノテーションを，複数の言語単位に基づく粒度で定量化し，さらに非対称な局所パスを用いた DTW によりアノテーションの高精度化を分析的に検討した．その結果，DTW を教育方針に応じた粒度で，非対称局所パスを用いて行うことで，手動書き起こしによる正解ラベルとより高い相関を持つ了解度アノテーションを実現できた．同時に L2 音声やシャドーイング音声に音声認識器を適用し，自動で書き起こすことによってシャドーイングやスクリプト・シャドーイングを代替できないか検討した．音声認識器によるアノテーションは，シャドーイング音声を自動で書き起こし，学習者の原稿と比較することで，手動による了解度とある程度の相関を示したが，シャドワー自身の訛りの影響を受け，精度があまり上がらないものも見られた．

今回はアノテーションを複数粒度で行い，局所パスを変えるというアプローチで高精度化を図ったが，精度のより一層高いアノテーションを行うには，いくつかの手法が考えられる．一つは，アノテーション粒度の階層化である．今回，複数粒度によるアノテーションとして，言語単位に応じてサブスコアを求めたが，そのサブスコアはその区間において fDTW をすることによって求められる．例えば wDTW におけるサブスコアの計算を pDTW の結果を使うといったように，ある言語単位におけるサブスコアの計算を，それより細かい言語単位を粒度とした DTW で改めて行うことで，サブスコア内でも継続長の正規化がなされ，よりアノテーションの高精度化が期待できる．二つ目として，posteriorgram 以外の特徴量（特に韻律）の使用が考えられる．逆シャドーイング法は，シャドワーの即時聴解の様子をシャドー音声とスクリプト・シャドー音声の差異を通して観測しているが，両者の差異は，調音構造の崩れ（posteriorgram DTW に相当）以外にも，韻律の崩れとしても観測される．これらを併用することで，瞬時了解度アノテーションの更なる高精度化を検討したい．

参考文献

[1] Munro, M. J. and Derwing, T. M.: Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Language learning*, Vol. 45, No. 1, pp. 73–97 (1995).
[2] Murphy, J. M.: Intelligible, comprehensible, non-native models in ESL/EFL pronunciation teaching, *System*, Vol. 42, pp. 258–269 (2014).
[3] Derwing, T. M. and Munro, M. J.: *Pronunciation fundamentals: evidence-based perspectives for L2 teaching and research*, John Benjamins (2015).
[4] Liontas, J. I., Association, T. I. et al.: *The TESOL Encyclopedia of English Language Teaching*, John Wiley & Sons (2018).
[5] Levis, J.: Revisiting the intelligibility and nativeness principles, *Journal of Second Language Pronunciation*, Vol. 6, No. 3, pp. 310–328 (2020).
[6] Smith, B.: *Learner English: A teacher's guide to interference and other problems*, Ernst Klett Sprachen (2001).
[7] Zhu, C., Lin, Z., Minematsu, N. and Nakanishi, N.: Analyses on instantaneous perception of Japanese English by listeners with various language profiles, *Proc. The Phonetic Society of Japan General Meeting*, pp. 26–31 (2020).

[8] Zhu, C., Minematsu, N. and Nakanishi, N.: Objective and semi-automatic measurement of smoothness of instantaneous understanding of L2 English speech (to appear), *Proc. Pronunciation in Second Language Learning and Teaching Conference* (2020).
[9] Derwing, T. M. and Munro, M. J.: Accent, intelligibility, and comprehensibility: Evidence from four L1s, *Studies in second language acquisition*, pp. 1–16 (1997).
[10] Bernstein, J.: Objective measurement of intelligibility, *Proc. ICPhS*, pp. 1581–1584 (2003).
[11] Minematsu, N., Guo, C. and Hirose, K.: CART-based factor analysis of intelligibility reduction in Japanese English, *Proc. EUROSPEECH*, pp. 2069–2072 (2003).
[12] Minematsu, N., Okabe, K., Ogaki, K. and Hirose, K.: Measurement of objective intelligibility of Japanese accented English using ERJ database, *Proc. INTERSPEECH*, pp. 1481–1484 (2011).
[13] Kang, O., Thomson, R. I. and Moran, M.: Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension, *Language Learning*, Vol. 68, No. 1, pp. 115–146 (2018).
[14] Alison, M., Rebecca, A., Catherine, S. and Paula, W.: Exploring the Relationship Between Modified Output and Working Memory Capacity, *Language Learning*, Vol. 60, No. 3, pp. 501–533 (2010).
[15] Munro, M. J. and Derwing, T. M.: Foreign accent, comprehensibility and intelligibility, redux, *Journal of Second Language Pronunciation*, Vol. 6, No. 3, pp. 283–309 (2020).
[16] Field, J.: Intelligibility and the listener: the role of lexical stress, *TESOL quarterly*, Vol. 39, No. 3, pp. 399–423 (2005).
[17] Inoue, Y., Kabashima, S., Saito, D., Minematsu, N., Kanamura, K. and Yamauchi, Y.: A study of objective measurement of comprehensibility through native speakers shadowing of learners' utterances, *Proc. INTERSPEECH*, pp. 1651–1655 (2018).
[18] Trisitichoke, T., Ando, S., Inoue, Y., Saito, D. and Minematsu, N.: Influence of content variations on smoothness of native speakers' reverse shadowing, *Proc. ICPhS* (2019).
[19] Ando, S., Lin, Z., Trisitichoke, T., Inoue, Y., Yoshizawa, F., Saito, D. and Minematsu, N.: A Large Collection of Sentences Read Aloud by Vietnamese Learners of Japanese and Native Speaker's Reverse Shadowings, *Proc. O-COCOSDA*, pp. 1–6 (2019).
[20] Lin, Z., Takashima, R., Saito, D., Minematsu, N. and Nakanishi, N.: Shadowability annotation with fine granularity on L2 utterances and its improvement with native listeners' script-shadowing, *Proc. INTERSPEECH*, pp. 3865–3869 (2020).
[21] Povey, D., Ghoshal, A., Boulianne, G., Glembek, L. B., Goel, N., Hannemann, M., Motlíček, P., Y., Q., P., S., Silovský, J., Stemmer, G. and Veselý, K.: The KALDI speech recognition toolkit, *Proc. ASRU* (2011).
[22] Amazon: *Amazon Polly*. `https://aws.amazon.com/polly/`.
[23] Shi, S. and Minematsu, N.: A corpus-based analysis of shadowing speech: case of L2 English by Japanese learners, *Proc. ISAPh*, No. 3, pp. 34–37 (2016).
[24] VERSANT: *Versant English Test*. `https://www.pearson.com/english/versant/tests.html`.
[25] Amazon: *Amazon Transcribe*. `https://aws.amazon.com/transcribe/`.