L2 音声を対象とした瞬時的了解度計測のための シャドーイング音声コーパスの構築 *

峯松信明，朱伝博，梱原卓也，箱田崚，齋藤大輔（東大），中西のりこ（神戸学院大）

1　はじめに

様々な L2 音声コーパスが構築されている [1] が，評定ラベルを付与する言語単位としては，発話全体（即ち学習者），各文音声，文音声中の各単語，各音節，各音素など，様々な粒度がある。[2] では発話全体を対象とし，非専門家である母語話者を複数募り，評定スコア平均をラベルとしている。[3] では音素を対象とし，専門家である音声学者が L2 音声を音声記号で詳細に書き起こしている。一方，応用言語学の分野では，L2 音声の評定対象は「母語話者発音からの逸脱ではなく，その L2 音声はどのくらい聴取者に伝わるのか」である，という主張がある [4]。この場合，評定ラベルは学習者の発音習熟度以外にも，聴取者の言語背景や発話内容など複数の要因に依存する。これは，評定ラベルは文脈（状況）依存となることを示唆する。更に，聞いているその場で（瞬時的に）聴取者に伝わったかどうかを細かな粒度で観測すること自体，容易なことではない。脳計測や瞳孔サイズの計測を通して認知的負荷量の時系列計測が行われているが，得られた時系列データは「話者が意図した単語を聴取者が正しく知覚した」ことを直接的に示す訳ではない。筆者らは [5, 6] において，聴取者に L2 音声をシャドー＋スクリプトシャドーさせることで，上記の問題は凡そ解決可能であることを示している。本報では，この手法に基づいて日本人英語音声に対して瞬時的了解度ラベリングを行うべく，シャドー音声コーパスを構築したので，その様子を報告する。

2　シャドーイングによる瞬時的了解度計測

ある L2 発声が提示され，その場で瞬時的に正しく聞き取られたかどうか（**図 1** 参照）はどのように計測すべきだろうか？応用言語学の分野では，読み上げ L2 音声を聴取者に書き取らせ，「正解書き取り率」を了解度（intelligibility）としているが，記憶容量に限界があるため，提示 L2 音声長は短くせざるを得ない。また，書き取り中に文を推測・再構成することも起こりえる [4]。音声認識を導入し，人間の聴取者の代行をさせることもあるが（**図 1** 参照），音声認識は図中の W を予測する技術であり，W^h を予測する技術ではない。W^h を予測するためには，人間と同じように聞き取り，そして聞き誤るよう音声認識器を訓練する必要がある。人間と機械の聞き取りの違い（W^h と W^m の違い）は [7, 8] などで報告されている。

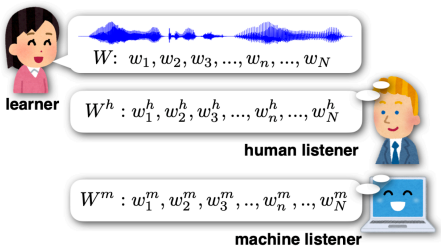

Fig. 1　人間と機械による瞬時的な単語同定

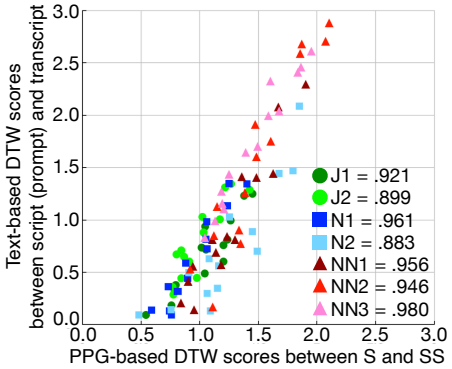

Fig. 2　S, SS による瞬時的了解度の計測 [6]

筆者らの先行研究 [9] では，「学習者の読み上げ音声が聴取者に伝わった」という命題を，「聴取者の心的プロセス（心的辞書検索）の結果，学習者が意図した単語が検索された」，即ち「読み上げ音声中の各単語を，聴取者が想起できた」と解釈した。そして [5, 6] では，瞬時的に想起できたか否かを古典的な聴取実験タスクであるシャドーイングを通して分析した。シャドーイングは外国語教育において「学習者がモデル音声を聞きながら復唱するタスク」として使われているが，本来は，聴取者の心的辞書検索過程を分析する目的で導入された実験タスクである [10]。[5, 6] では，母語話者（相当）に L2 音声をシャドーさせ，意図された単語を瞬時に想起できたか否かを分析した。

シャドー音声（S）と，直後に収録されたスクリプト・シャドー音声（SS，学習者の読み上げ原稿を見ながらのシャドーイング）を音素ポステリオグラム（PPG）に変換し，PPG-based DTW で比較した。SS は内容が聞き取れたシャドーイングに相当し，DTW 比較は，どこでシャドーが崩れたのか（どこで聞き取りにつまずいたのか）を検出することに相当する。**図 2** は PPG-based DTW スコアと，S の書き起こしと対応する読み上げ原稿との差異（瞬時了解度）の関係である。聴取者の言語背景（N：母語話者，J：日本人，NN：非母語話者）によらず，相関は非常に高い。

---

*Development of a shadowing speech corpus for instantaneous intelligibility measurement on L2 speech, N. Minematsu, C. Zhu, T. Kunihara, R. Hakoda, D. Saito (UTokyo), N. Nakanishi (KGU)

3　シャドーイングコーパスの構築

3.1　コーパスの構築の目的

異なる言語背景の聴取者から，ある規模の L2 音声に対する（単語や音素を単位とした）瞬時了解度ラベルを取得できれば，彼らのシミュレータ（virtual shadower，L2 音声とその原稿を入力して，瞬時的了解度を単語／音素単位で出力する）が構築可能となると思われる。以下，3 名のシャドワーを対象とした S+SS 音声コーパス構築について述べる。なお，教育的には**図 1** における W^h（ある話者が L2 音声を聞いた場合に聞き取った単語列）を予測することが望ましい。しかしここでは，聴取者が想起した単語（w_i^h）は予測せず,「W と W^h はどこが異なるのか」を予測することをタスクとしている。

3.2　L2 音声

K 大学の学部 1，2 年生 270 名を対象として英作文を課した。テーマは大学生活から社会問題まで多岐に渡る。その後，作成した英文章を各自の読みやすい話速で音読させた。音声の収録は，本課題用の収録 web（Speech Saver [11]）を通して収録し，読み上げた英文章も提出させた。音声サンプル数は約 8,700 である。英文章には教師側の校正は入っておらず，文法誤りがあった場合も，その通りに音読している。シャドワーから SS 音声を収録する場合，句単位でテキスト表示する必要があるため，L2 音声を原稿を使って強制切り出しする必要がある。この作業は WSJ-KALDI と CMU 発音辞書を用いて行った。ここで，シャドワーにとって未知語となる語が含まれる英文章は除外し，自動計測された雑音レベルを使って雑音が強い音声も排除した。更に，文間ポーズが 2 秒以上ある場合は，2 秒に短縮した。シャドワーに提示する音声を約 30 秒とするため，音声区間頭から音声区間尾までが約 30 秒となる連続する文を検索したところ，3,860 音声区間が候補として得られた。これら音声区間のテキストに対して tf-idf を用いてテキスト間距離を求め，クラスタリングした。テキスト内容が重ならず，また，話者の重なりも少なくなるよう音声区間を選定し，最終的に 753 個の音声を選定した。なお，音読時に，用意した英文を微修正して読み上げる例が散見されたため，選定された音声区間の原稿は大学院生 10 名によって精査され，音声と原稿との一致度を高めた。

3.3　シャドワー

L2 音声の了解度は聴取者の言語背景に依存するので [5]，以下の 3 名をシャドワーとして採択した。S1：英語圏に 6 年の留学経験を持ち，英語教育を専門とする日本人留学生。日本人英語は聞き慣れている。S2：日本語を学ぶ（中級レベル）ロシア系米国人の大学学部生。専門は応用言語学である。S3：日本語を知らない中国系米国人。音楽を専門とする専攻を修了したばかりである。いずれの参加者も外国語学習・教育に強い関心を持ち，本プロジェクトの狙いを理解した上で参加している。即ち，L2 音声をシャドー，スクリプトシャドーし，両者を比較することで瞬時的了解度を計測する方法論に賛同し，自身の S，SS 音声を通して，L2 音声に，瞬時的了解度ラベルが付与されることを十分理解した上で参加している。

3.4　S+SS 音声収録

シャドーイング練習として，英検 2 級リスニング課題のモノローグ音源と日本人英語音声を幾つかシャドーさせた。3 人とも母語話者の英語は流暢にシャドーできる。練習の後，S, SS 音声の収録となった。音声収録は収録用 web を構築し，同一のヘッドセットマイクロホンを 3 人に提供した。シャドーを 2 回，その後，スクリプトシャドーを 1 回収録した。S 音声の収録は各回で再収録は許可していないが，SS 音声は基準となる音声であるため，読み間違いなどがあれば再収録を許可した。また，L2 音声が聞き取り難いこと以外の要因による発話の乱れ，技術的な不具合があった場合は，その旨，クリックして記録を残させた。原稿執筆時点で S1，S2, S3 より，560, 88, 211 の L2 音声に対する S+SS 音声を得た。

4　おわりに

瞬時的了解度ラベリングを L2 音声に対して行うため，異なる言語背景を持つ 3 人のシャドワーを対象とした，S+SS 音声コーパス収集について報告した。S1 の音声を使った virtual shadower 構築については，初期検討結果を [12] で報告している。

参考文献

[1] Non-native speech database: https://en.wikipedia.org/wiki/Non-native_speech_database

[2] K. Saito *et al.*, “Acoustic characteristics and learner profiles of low-, mid-and high-level second language fluency,” *Applied psycholinguistics*, 39, 3, 593–617, 2018

[3] T. Makino *et al.*, “English read by Japanese phonetic corpus: An interim report,” *Research in Language*, 10, 1, 79–95, 2012

[4] M. J. Munro *et al.*, “Foreign accent, comprehensibility and intelligibility, redux,” *Journal of Second Language Pronunciation*, 6, 3, 283–309, 2020

[5] C. Zhu *et al.*, “Multi-granularity analysis of online intelligibility of L2 speech based on reverse shadowing,” *Proc. Acoust. Soc. Jpn. Spring Meeting*, 3-2-17, 2021

[6] 箱田他，“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”，情報処理学会研究報告，2021-SLP-137，50，1–6，2021

[7] T. M. Derwing *et al.*, “Does popular speech recognition software work with ESL speech?,” *TESQL Quarterly*, 34, 3, 592–603, 2000

[8] 峯松他，“逆シャドーイングに基づく瞬時的明瞭度の自動計測と音声認識精度の比較〜音声認識は人間の瞬時的聴解にどこまで迫れるのか?〜”，外国語教育メディア学会，全国大会予稿集 2021

[9] 峯松他，“米語母語話者を対象とした日本人英語の聞き取り調査”，信学技報，SP2004-142，31–36，2005

[10] W. D. Marslen-Wilson, “Speech shadowing and speech comprehension,” *Speech Communication*, 4, 55–73, 1985

[11] Speech Saver: https://noriko-nakanishi.com/speech/

[12] C. Zhu *et al.*, “Automatic prediction of instantaneous intelligibility of L2 speech in sequence,” *Proc. Autumn Meeting of ASJ*, 1-3-16, 2021