自動採点スピーキングテスト SJ-CAT の能力推定の検証

今井新悟（筑波大学） 赤木彌生（山口大学） 石塚賢吉（株式会社ドワンゴ）
伊東祐郎（東京外国語大学） 菊地賢一（東邦大学） 篠崎隆宏（東京工業大学）
中園博美（島根大学） 中村洋一（清泉女学院短期大学） 西村隆一（和歌山大学）
本田明子（立命館アジア太平洋大学） 家根橋伸子（東亜大学） 山田武志（筑波大学）
imai.shingo.fu@u.tsukuba.ac.jp

1 はじめに

言語能力の 4 技能「読む，聞く，書く，話す」の測定において，「書く」「話す」の産出能力のテストの実施は難しい．中でも，スピーキングテストは，テスター（評定者）を養成し，確保し続けることが必要であり，多大な時間とコストがかかる．評定者による対面形式に代わるスピーキングの CBT（Computer based Test）がいくつか存在するが，コンピュータを介して音声を録音し，後で人が評定する仕組みが一般的である．対面で行うテストに比べて時間とコストの削減はある程度できるものの，テスターの養成と確保という根本的な課題の解決にはならない．これを解決するには，人を介さない自動採点のシステムが必要である．英語においては Versant™ English Test[1]と Speech Rater™ [2]が開発されている．前者はテストとして利用できる唯一のものであるが，自由回答（Open Questions）形式が 2 問出題されるものの，それは自動採点の対象になっていない[1]．後者は正式なテストとしてリリースされておらず，開発が継続中である．本研究で開発した SJ-CAT（Speaking Japanese Computerized Adaptive Test）は日本語では初のスピーキング自動採点テストであり，自由回答形式も含む．

本稿では，SJ-CAT のシステムの説明に続き，自動採点と教師による採点の比較，SJ-CAT と人が採点する他のスピーキングテストとの比較を行い，自動採点によって，人による評価に近い評価が得られたことを示す．

2 SJ-CAT の概要

SJ-CAT は非母語話者を対象とした日本語のスピーキングテストであり，インターネットでアクセスできる点，音声認識技術を用いている点，項目応答理論の多段階反応モデルを用いている点，アダプティブテストである点，自動採点である点が特色である．WEB 上で公開しており[3]，ユーザー登録をすると受験できる．テストは 2 つのセクション，4 種類の問題で構成される．Section1 は文読み上げ問題と選択肢読み上げ問題，Section2 は文生成問題と自由発話問題からなる．

文読み上げ問題では，画面上に「おじさんとおじいさんが来ました」のような 1 文が現れ，それを読み上げる音声が聞こえる．その後，その文を読み上げることが課題となる．

選択肢読み上げ問題では，静止画や動画で場面が提示される．例えば，2 人が話をしている場面の映像を示され，「2 人は何をしていますか.」という質問が流れ，「a. ご飯を食べています.」「b. 話をしています.」「c. 本を読んでいます.」という画面上の選択肢から正解を選んで読み上げる．

[1] http://www.disc.co.jp/sp/versant/index.html
[2] https://www.ets.org/research/topics/as_nlp/speech/
[3] https://www.sj-cat.org/

文生成問題では，例えば箱を開けている映像が流れ，「何をしていますか.」という質問が聞こえる．5 秒間考えたのち，「箱を開けています.」のように文で応答する．

自由回答問題は例えば，「次の質問に，30 秒ぐらいで答えてください．消費税が上がることに賛成ですか，反対ですか．その理由も言ってください.」という質問に対して，5 秒間考えたのち，30 秒程度（録音時間制限は 40 秒）で応答するものである．

テスト終了と同時に画面上に採点結果が，セクション 1 が 25 点満点，セクション 2 が 75 点満点，計 100 点満点で示される．この得点は後述する能力推定値を換算して算出している．テスト時間は，最初の録音音量の設定，各セクション 2 問ずつの練習問題を含めて，15 分から 20 分程度である．

3　採点の方法

自動採点の方法は，いくつかの提案を行ってきたが（[2][3]など），現在のバージョンでは音声認識器として，Julius と T3 デコーダを併用した以下の方法を採用している．

文読み上げ問題では，両認識器に正解文を含む多数の文を単語として登録しておき，第 1 段階として，単語（正解文）が認識されなければ 0 点とし，それ以外は次の 8 次元の音響特徴量を用いてサポートベクター回帰 (SVR) を用いて採点する．

①発音を評価するため：Julius と T3 の単語音響尤度のフレーム平均

②モーラの自然さを評価するため：母語話者の回答 (10 人分の平均) における各音素の発音タイミングと受験者の発話における各音素の発音タイミングの差である，発音タイミング距離

③アクセントおよびイントネーションを評価するため：母語話者のピッチパターンとの差である基本周波数パターン距離

④流暢さを評価するため：以下の 4 種類のスピーキングレート指標

(S1) 音素数／発話全体の長さ

(S2) 音素数／音声区間の長さ

(S3) 息継ぎ区間の長さ／発話全体の長さ

(S4) $\sum_{k}^{n}(S2-1/\text{音素}_k\text{の長さ})^2$　／音素数

S3 はポーズの長さであり，この値が小さいほど高評価となることを仮定している．S4 は音素内の発話の速さを示す．

選択肢読み上げ問題では，第 1 段階として，Julius と T3 のいずれでも選択肢にある文が認識されなければ 0 点，いずれかで不正解の選択肢が認識されれば 1 点，両方で不正解の選択肢が選択されれば，正解ではないが発音が良いと仮定して 2 点，いずれかで正解の選択肢が認識されれば，発音は悪いが正解を認めて 2 点，両方で正解の選択肢が認識されれば，第 2 段階として文読み上げ問題と同じ 8 次元の音響特徴量で採点する．

文生成問題では，スピーキングレート指標の S1, S2 で流暢さを評価するほか，Julius が認識したキーワード，T3 が認識したキーワード，キーワードスポッティングによるキーワードの有無を特徴量とした計 5 次元の SVR で採点する．キーワードはあらかじめサンプリングした文生成問題 35 問×191 人分の音声データを文字化し，問題ごとに高頻度に現れる語を抽出した．キーワードとのマッチングをし，発話文の内容を評価する．

自由回答問題でもキーワード一覧を作成し，第 1 段階で Julius の認識文と T3 の認識文のいずれにもキーワードがなければ 0 点とし，それ以外は第 2 段階でスピーキングレート指標の S1，S2 で流暢さを評価するほか，発話量と語彙多様性[4]の特徴量を含む 4 次元の SVR で採点する．

（発話量）音素数／録音時間

（語彙多様性）異なり語数／sqrt（2×延べ語数）

発話量が多いほど高く評価されるが，同じ発話量であっても，同じ表現を繰り返すよりも，豊富な語彙を用いる方が高く評価される．

4　能力推定と出題の方法

以上の手法で各問題の採点が行われ，0 から 4 までの連続値を得る．それを 0 点から 4 点の 5 段階の離散値に変換して，項目応答理論段階反応モデル[5]を使って，能力値を 1PL モデル，ベイズ EAP によって推定する．本テストはアダプティブであり，各問題の採点により，次に出題される問題が変わる．テストの開始時は最初の 2 問の回答による暫定能力値によって 3 問目に出題する問題を選択する．その後はベイズ推定による能力値の事後分散の期待値が最小となる問題を選択する[6]．終了条件（現行設定は，推定誤差が 0.5 未満になる，あるいは 1 セクションでの出題数が 12 問に達する）を満たすとテストが終了する．

5　SJ-CAT の検証

5.1　自動採点と教師による採点の比較

自由回答問題を 5 人，その他の問題を 3 人の日本語教師が採点した．自動採点では，自由回答問題と文読み上げ問題各 81 人分の音声データ，選択肢読み上げ問題と文生成問題各 114 人分の音声データを学習データとして SVR で学習し，各問題について 20 人分の音声データで評価した．

表 1　自動採点と教師の採点の相関

	相関	RMSE
自由回答	0.91	0.63
選択肢読み上げ	0.89	0.64
文読み上げ	0.77	0.49
文生成	0.70	1.25

表 1 の通り，強い相関が得られた．文読み上げ問題という制限の強い問題よりも，自由度の高い自由回答問題の方の相関が高くなるというのは当初の予想に反していた．当初は，文の空所に適当な語を入れて読み上げる穴埋め形式の問題もあった．穴埋めに使われた語を評価したが，この形式は相関が低くなったため廃止した．このように回答が固定され，また短くなると相関が低くなる傾向を示した．短答・固定回答では，音声認識の成否が大きく採点結果に影響し，音声認識が成功しない場合に採点が不安定になると考えられる．一方，自由回答では音響特徴量がより強く採点に影響し，音声認識の失敗による影響が少ないため，相対的に安定した採点ができたと考えられる．

今後，評価者を 6 人～8 人に増やした採点結果を用い，自動採点アルゴリズムも変えてさらに検証する予定である．

5.2　SJ-CAT と JSST の比較

6 大学の日本語学習者に SJ-CAT と JSST（アルク社）[4]を受験してもらい，その結果を比較した．JSST は電話で受験するもので，出題数は 10 問で，回答が録音される．3 人の評定者が採点し，10 段階のレベルで初級から上級までを評価する．「～した時のことについて話してください」というような質問に対し，45 秒または 60 秒で回答する．

有効なデータを得た受験者は 178 人で，原則として同日に両テストを受験した．ただし JSST の録音ができていないため，後日 JSST を再受験した者（約 1 割）がいた．これを含め，両テストの受験の順番を変えて，カウンターバランスを取った．SJ-CAT を先に受験した人，JSST を先に受験した人，それぞれ 89 人ずつであった．

表 2　テスト結果の基本統計量　（n=178）

	最小値	最大値	mean	SD
SJ-CAT	4	94	61.7	18.0
JSST	1	9	5.9	1.5

[4] http://www.alc.co.jp/jsst/

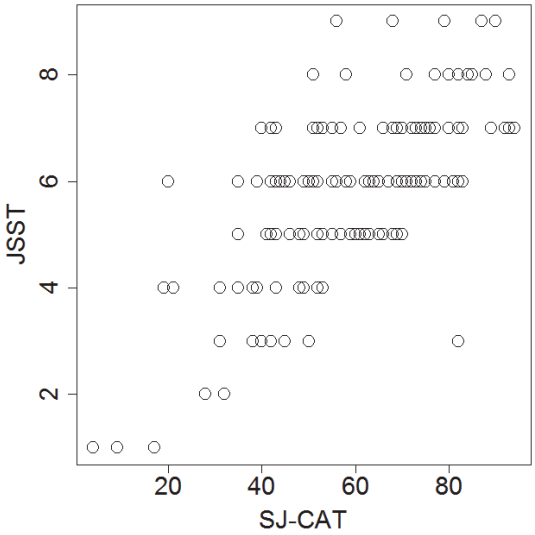

図 1　SJ-CAT と JSST の散布図

両テスト間に中程度の相関 r=0.651（p<0.001）があった．図 1 によると，下位レベルの受験者が少なかったことが分かる．これにより相関が押し下げられた可能性がある．レベルが比較的分散している A 大学分 60 人分（SJ-CAT：mean=55.4, SD=18.9，JSST：mean=5.1, SD=1.6）では，r=0.812 (p<0.001)となり，強い相関が認められた．

また，JSST の中級のレベル 6, 7 において SJ-CAT の分散が大きいことが分かる．人による評価でも中級レベルが最も難しいと思われる．個々の回答音声データを分析し，両テストの評価のずれの原因を探ることが今後の課題である．

6　おわりに

本稿では，SJ-CAT の自動採点を評価するために，2 つの検証を実施した．その結果，問題タイプにより違いはあるものの，人と機械の採点に高い相関が認められたことから，個々の問題項目の自動採点の信頼性はある程度確保できたと言える．JSST との比較においても高い相関を示していることから，総合的評価において，併存妥当性を認めることができ，人による評価の代替になりうる可能性を示している．機械による評価は条件が同じ（音声データが同じ）であれば常に同じ結果が出るが，人の場合は条件が同じでも評価がぶれることがある．この点ではむしろ機械による評価が優位とも言えよう．

謝辞

本研究は科学研究費基盤（A）(22242014)「コンピュータ自動採点日本語スピーキングテストの実用化と妥当性の検証」の助成を受けた．開発に関わった多数の大学院生，協力者に感謝する．

参考文献

[1] Pearson Education, *Versant™ English Test: Test Description and Validation Summary*, 2011. http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf (2017.1.15 アクセス)

[2] Y. Ono, M. Otake, T. Shinozaki, R. Nisimura, T. Yamada, K. Ishizuka, Y. Horiuchi, S. Kuroiwa, S. Imai. “Open answer scoring for S-CAT automated speaking test system using support vector regression,” Proc. APSIPA ASC 2012, Dec. 2012.

[3] H. Lu, T. Yamada, S. Imai, T. Shinozaki, R. Nisimura, K. Ishizuka, S. Makino, N. Kitawaki. "Automatic scoring method for open answer task in the SJ-CAT speaking test considering utterance difficulty level," Proc. APSIPA 2014, WA1-1-3, pp. 1-5, Dec. 2014.

[4] 田島ますみ, 深田淳, 佐藤尚子.「語彙多様性を表す指標の妥当性に関する研究－日本人大学生の書き言葉コーパスの場合－」中央学院大学社会システム研究所紀要 9(1), pp. 51-62, 2008.

[5] F. Samejima, “Estimation of Latent Ability Using a Response Pattern of Graded Scores,” Psychometric Monograph, 17. Psychometric Society, 1969.

[6] 菊地賢一.「段階反応モデルに基づく汎用的適応型テストシステムの開発」日本行動計量学会大会発表論文抄録集, pp. 362-363, Aug. 2005.