日本語スピーキングテスト SJ-CAT における 項目応答理論に基づく能力値推定の検証 *

☆小野友暉，山田武志（筑波大），△菊地賢一（東邦大），△今井新悟，牧野昭二（筑波大）

1 はじめに

総合的な日本語能力の評価のためには，リーディングテストとリスニングテストのみならず，発話能力をテストするスピーキングテストの実施が必要不可欠である．従来方式のスピーキングテストは，訓練された評定者が解答発話を聞いて採点するという主観評価方式であり，多大な費用と時間がかかってしまうという問題がある．従って，自動採点方式のスピーキングテストの開発が急務である．

現在，日本語スピーキングテスト SJ-CAT（Speaking Japanese Computerized Adaptive Test）[1] の開発が進められている．SJ-CAT の問題は，2 つのセクションに分けられる．Section 1 は文読み上げ問題と選択肢読み上げ問題，Section 2 は文生成問題と自由発話問題からなり，各問題には多数の設問が用意されている．各問題の概要を Table 1 に示しておく．SJ-CAT の最大の特徴は，項目応答理論 [2][3] に基づくアダプティブテストにより受験生の能力値を逐次的に推定することである．これにより，受験者の能力値を短い時間で正確に推定することができる．

SJ-CAT のテストの流れを Fig. 1 に示す．まず，現時点の推定能力値に応じて，設問プールから設問の選択を行う．受験者は，出題された設問に口頭で解答する．次に，採点アルゴリズム [4]-[7] を用いて，解答発話のスコアを推定する．このスコアと現時点の推定能力値から，受験者の能力値を再推定する．収束条件を満たすまで，上記を繰り返す．

これまでに，我々は個々の解答発話を採点する採点アルゴリズムを提案し，有効性の検証を行った [4]-[7]．さらに本稿では，能力値推定の精度を SJ-CAT のプロトタイプシステムを用いて検証する．

Table 1 問題の種類と概要

問題の種類	問題の概要
文読み上げ	指定された文を読み上げる問題
選択肢読み上げ	設問に対する解答を候補から選択して読み上げる問題
文生成	設問に対する短い応答文を考え発話する問題
自由発話	指定のテーマに沿って 30 秒程度発話する問題

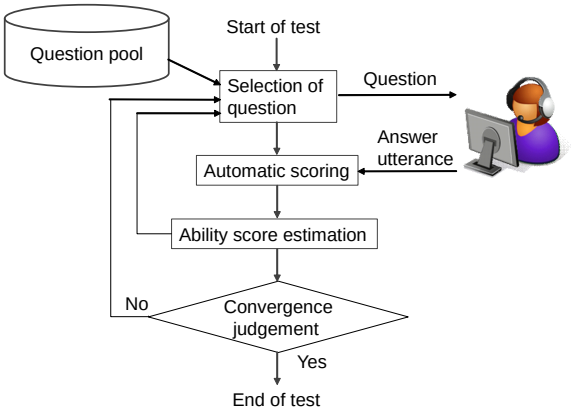

Fig. 1 SJ-CAT のテストの流れ

2 SJ-CAT における能力値推定

2.1 項目応答理論に基づく能力値推定

SJ-CAT では，項目応答理論に基づく段階反応モデル [8] を用いて能力値推定を行う．SJ-CAT の設問プールに存在する各設問には，式 (1) のようにモデル $p^*_{jk}(\theta)$ が設定されている，

$$p^*_{jk}(\theta) = \begin{cases} \frac{1}{1+e^{-1.7a(\theta-b_{jk})}} & k = 1,2,3,4 \\ 1 & k = 0 \end{cases} \quad (1)$$

これは能力値 θ の受験者が，設問 j においてスコア k $(k = 0,1,2,3,4)$ 以上を獲得する確率を表している．ここで，a は定数，b_{jk} は設問の難易度を表している．これらのパラメータは周辺最尤法を用いて事前に求めておく．このモデルを用いると，能力値 θ の受験者が設問 j においてスコア k を獲得する確率 $p_{jk}(\theta)$ は，以下のように表される．

$$p_{jk}(\theta) = \begin{cases} p^*_{jk}(\theta) - p^*_{jk+1}(\theta) & k = 0,1,2,3 \\ p^*_{jk}(\theta) & k = 4 \end{cases} \quad (2)$$

受験者の能力値推定はベイズ推定により行われる．能力値推定のアルゴリズムの詳細を以下に示す．

*Item Response Theory-Based Ability Estimation in SJ-CAT Japanese Speaking Test. by Yuki ONO, Takeshi YAMADA(Univ. of Tsukuba), Kenichi KIKUCHI(Toho Univ.), Shingo IMAI, Shoji MAKINO(Univ. of Tsukuba)

1. $n-1$ 番目の設問に解答した時点での，受験者の能力値分布 (事前分布) を $h_{n-1}(\theta)$ とする．なお，事前分布の初期値は適切に設定する必要がある．
2. 採点アルゴリズムにより，n 番目の設問 j のスコアが u と推定されたとする．
3. 式 (2) を用いて $p_{ju}(\theta)$ を求める．
4. 次式により，能力値分布を更新する．

$$h_n(\theta) = \frac{h_{n-1}(\theta)p_{ju}(\theta)}{\int_{-4}^{4} h_{n-1}(\theta)p_{ju}(\theta)d\theta} \tag{3}$$

5. $h_n(\theta)$ の標準偏差が，閾値 ϵ より小さくなるまで上記を繰り返す．
6. $h_n(\theta)$ の平均を，受験者の能力値として出力する．

2.2 項目応答理論に基づく設問選択

SJ-CAT では，受験者の能力値分布 (事後分布) の分散の期待値が最も小さくなるような設問を選択する．これは，受験者の能力値推定の収束を早めるためである．設問選択のアルゴリズムを以下に示す．

1. $n-1$ 番目の設問に解答した時点での，受験者の能力値分布 (事前分布) を $h_{n-1}(\theta)$ とする．
2. 設問プールにおける，まだ出題していない設問の集合を J とする．設問 $j \in J$ において，能力値の事後分布 $h(\theta)_{n-1}p_{ju}(\theta)$ の分散 σ_{ju}^2 $(u = 0, 1, ..., 4)$ を求める．
3. j を出題したとき，スコアが u となる確率 q_{ju} $(u = 0, 1, ..., 4)$ を次式により求める．

$$q_{ju} = \int_{-4}^{4} h_{n-1}(\theta)p_{ju}(\theta)d\theta \tag{4}$$

4. 事後分布の分散の期待値 $\sum_{u=0}^{4} \sigma_{ju}^2 q_{ju}$ が最も小さい設問 $j \in J$ を次に出題する設問として決定する．

2.3 自動採点アルゴリズム

本節では，SJ-CAT のプロトタイプシステムに採用されている採点アルゴリズムの概要を述べる．採点アルゴリズムは，解答発話から特徴量を抽出し SVR（Support Vector Regression）[9] を用いてスコア推定を行う．各問題に対して抽出する特徴量を，以下に示す．

文読み上げ問題と選択肢読み上げ問題については，発音の良さ，発音のタイミング，基本周波数などの音響特徴量を使用している．これらの特徴量を用いて推定したスコアと，複数の評定者によるスコアの平均を比較したところ，文読み上げ問題では相関係数は 0.77，RMSE（Root Mean Square Error）は 0.49 だった．また選択肢読み上げ問題では，相関係数は 0.89，RMSE は 0.64 だった．

文生成問題については，上記で用いられている音響特徴量に加え，キーワード判定に対応する特徴量が用いられている．自由発話問題については，さらに語彙量，発話量に対応する特徴量が追加されている．これらの特徴量を用いて推定したスコアと，複数の評定者によるスコアの平均を比較したところ，文生成問題では相関係数は 0.70，RMSE は 1.25 だった．また，自由発話問題では相関係数は 0.91，RMSE は 0.63 だった

3 SJ-CAT における能力値推定の精度の検証

3.1 実験条件

SJ-CAT における能力値推定の精度の検証を行った．本実験では，次の 3 つのシステムを用意した．

- *proposed system*: 2 章で述べたシステム．
- *proposed system with human raters*: *proposed system* と同様だが，採点アルゴリズムによる推定スコアの代わりに，評定者が採点したスコアを用いる．自動採点アルゴリズムにおいて推定エラーを避けることは難しく，この推定エラーは能力値推定に悪影響を及ぼす．この影響を調査するため，*proposed system* と本システムを比較する．
- *proposed system with human raters and all the questions*: *proposed system with human raters* と同様だが，能力値推定のために設問プールのすべての設問を用いる．このシステムにより推定した能力値を真の能力値とみなす．

実験条件を Table 2 に示す．本実験では，留学生 20 名の解答発話を使用した．設問プールに含まれる設問数は，Section 1 においては 31 問，Section 2 においては 43 問である．*proposed system with human raters* と *proposed system with human raters and all the questions* における評定者は 5 名（自由発話問題のみ 8 名）であり，各評定者のスコアの平均を用いた．推定能力値と真の能力値の差と，収束条件を満たすまでに出題される設問数にはトレードオフの関係がある．そこで，2.1 節のステップ 5 における収束条件の閾値 ϵ を $0.3, 0.4, ..., 0.7$ の範囲で変化させ，実験を行った．

Table 2　実験条件

受験者	留学生 20 名
設問プール	Section 1：31 問 (選択肢読み上げ 17 問, 文読み上げ 14 問) Section 2：43 問 (文生成 34 問, 自由発話 9 問)
評定者の人数	5 名 (自由発話のみ 8 名)

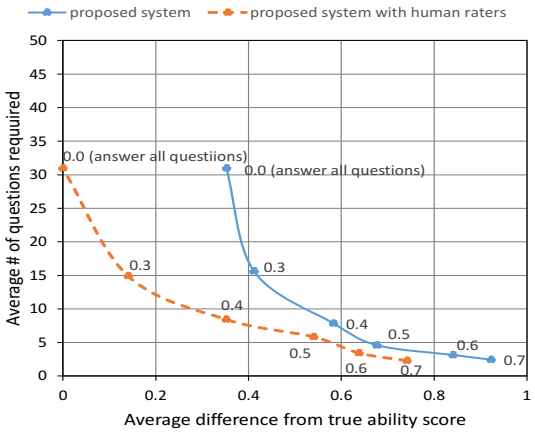

Fig. 2　Section 1 における能力値の推定誤差と出題設問数

3.2　実験結果

閾値を変化させたときの，能力値の推定誤差と出題された設問数の関係を Fig. 2 と Fig. 3 に示す．ここで，Fig. 2 は Section 1，Fig. 3 は Section 2 である．縦軸は，収束条件を満たすまでに出題された設問数（受験者 20 名の平均）である．横軸は，能力値の推定誤差（受験者 20 名の平均）である．閾値が 0.0 のときは，すべての設問が出題されることとなる．

図より，どちらのセクションにおいても，閾値を 0.4 より小さくすると急速に出題設問数が増加することから，閾値を 0.4 以上に設定するのが妥当であるといえる．また，例えば閾値を 0.5 に設定したとき，Section 1 においては *proposed system* と *proposed system with human raters* の間で出題設問数と能力値の推定誤差ともに大きな差がないことが分かる．一方，Section 2 では *proposed system* と *proposed system with human raters* の間で出題設問数には大きな差がないものの，能力値の推定誤差には大きな差があることが分かる．これは，Section 2 は Section 1 と比べ採点アルゴリズムの推定精度が低いことが原因だと考えられる．

3.3　設問スキップによる能力値の推定誤差の改善

前節の実験結果から，採点アルゴリズムにおけるスコアの推定誤差が，受験者の能力値推定に悪影響

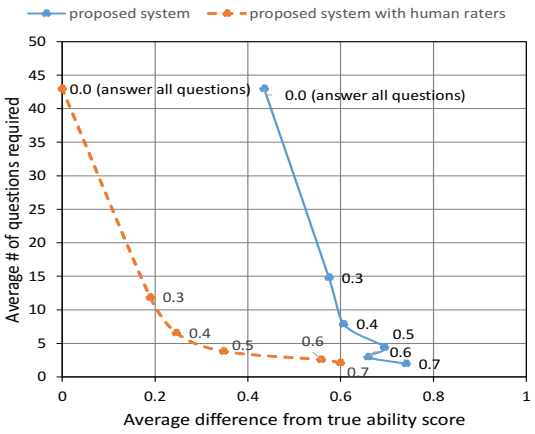

Fig. 3　Section 2 における能力値の推定誤差と出題設問数

を及ぼすことが確認された．そこで，推定誤差の大きい解答発話を能力値推定に用いないようにすることにより，能力値推定に及ぼす影響を緩和することを考える．この効果を検証するため，次のシステムを用意した．

- *proposed system with problem skip*: *proposed system* と同様だが，各設問において採点アルゴリズムによる推定スコアと評定者が採点したスコアの間に 1 点以上の差があるとき，その設問を能力値推定に用いず次の設問を出題する．

閾値を変化させたときの，能力値の推定誤差と出題された設問数の関係を Fig. 4 と Fig. 5 に示す．ここで，Fig. 4 と Fig. 5 は，Fig. 2，Fig 3 に *proposed system with problem skip* を追加したものである．図より，同数の設問を出題した場合 *proposed system with problem skip* は *proposed system* と比べ，能力値の推定精度が向上することが分かった．今後は，これを自動化する方法について検討する．

4　おわりに

本稿では，SJ-CAT における受験者の能力値推定について検証を行った．実験により，出題設問数という観点からは閾値を 0.4 以上に設定することが妥当であるということが分かった．また，個々の解答発話の採点精度が，受験者の能力値推定に影響を及ぼすことが確認された．この問題を解決するためには，採点精度が低い解答発話を能力値推定に用いないようにすることが有効であることが分かった．

謝辞　本研究は科研費 (22242014) の助成を受けた．本研究をご支援いただいた SJ-CAT プロジェクトのメンバーに深く感謝する．

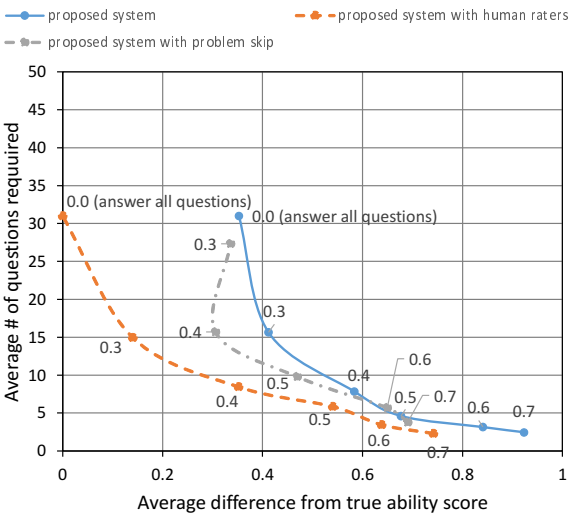

Fig. 4　Section 1 における設問スキップを加えた能力値推定精度と出題設問数

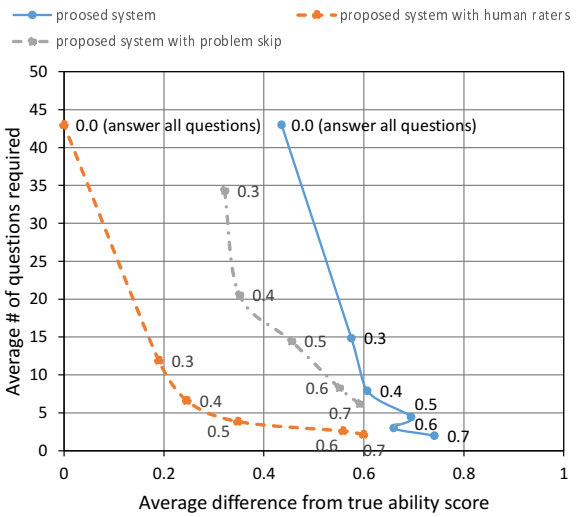

Fig. 5　Section 2 における設問スキップを加えた能力値推定精度と出題設問数

参考文献

[1] 今井新悟, “Speaking Japanese Computerized Adaptive Test 開発の目的・方法と構成,” 日本行動計量学会 41 回大会, SC1-2, 2013.

[2] F. M. Lord, “Application of Item Response Theory to Practical Testing Problems,” 1980.

[3] 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.

[4] N. Okubo, Y. Yamahata, T. Yamada, S. Imai, K. Ishizuka, T. Shinozaki, R. Nisimura, S. Makino, N. Kitawaki, “Automatic Scoring Method Considering Quality and Content of Speech for SCAT Japanese Speaking Test,” Proc. Oriental COCOSDA 2012, pp. 72–77, Dec. 2012.

[5] 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.

[6] 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.

[7] 西村竜一, 栗原理沙, 篠崎隆宏, 石塚賢吉, 山田武志, 今井新悟, 河原英紀, 入野俊夫, “日本語スピーキングテスト S-CAT における並列セグメンテーションを用いた自動採点の検討,” 日本音響学会秋季研究発表会, pp. 397–398, Sep. 2012.

[8] F. Samejima, “Estimation of Latent Ability Using a Response Pattern of Graded Scores,” VA: Psychometric Society, 1969.

[9] C.-C. Chang, and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.