SJ-CAT における 項目応答理論に基づく能力値推定の精度改善 *

☆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

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 のテスト問題は，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，各問題には多数の項目（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高精度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，項目プールから項目の選択を行う．受験者は，出題された項目に口頭で解答する．次に，採点アルゴリズム（例えば [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，受験者の能力値を更新する．以上の処理を収束条件を満たすまで繰り返す．

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]．また，SJ-CAT における受験者の能力値推定について検証を行い，個々の解答発話に対する採点誤りが能力値推定精度と出題項目数に悪影響を及ぼすことを確認した [8]．本稿では，この問題を解決するため，採点の信頼度が顕著に低い解答発話を自動で検出し，能力値推定に用いないようにする手法を提案する．提案手法では，採点アルゴリズムと同じ特徴量を用いた確率推定付きのクラス分類により解答発話に対する採点信頼度を求める．採点信頼度が閾値を下回った場合はその項目をスキップし，上回った場合は採点アルゴリズムにより求めたスコアを用いて能力値を更新する．提案手法の有効性を実験により検証する．

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	項目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	項目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

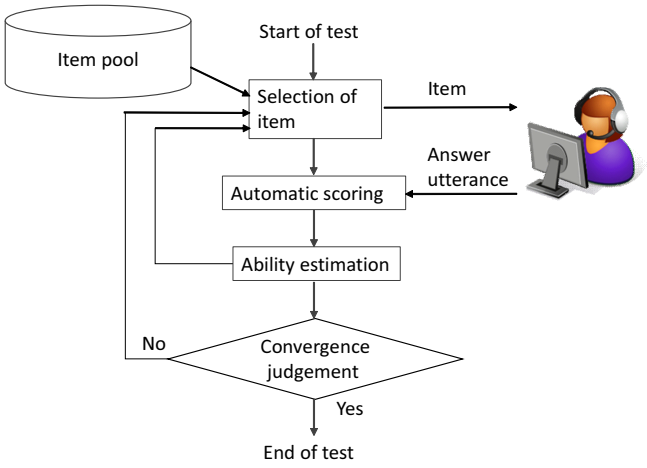

Fig. 1 SJ-CAT のテストの流れ

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 [2][3][9] を用いて能力値推定を行う．SJ-CAT の項目プールに存在する各項目には，式 (1) のようにモデルが設定されている，

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

$p^*_{jk}(\theta)$ は能力値 θ の受験者が，項目 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

* Accuracy Improvement of 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)

している．ここで，a は識別力（ただし，1 パラメータモデルでは定数），b_{jk} は項目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 の受験者が項目 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．

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

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．

1. $n-1$ 番目の項目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 番目の項目 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項目選択

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項目を選択する．これは，受験者の能力値推定の収束を早めるためである．項目選択のアルゴリズムを以下に示す．

1. $n-1$ 番目の項目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. 項目プールにおける，まだ出題していない項目の集合を J とする．項目 $j \in J$ において，能力値の事後分布 $h_{n-1}(\theta)p_{ju}(\theta)$ の分散 σ^2_{ju} $(u = 0, 1, ..., 4)$ 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 $\sum_{u=0}^{4} \sigma^2_{ju}q_{ju}$ が最も小さい項目 $j \in J$ を次に出題する項目として決定する．

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）[10] 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．

Section 1 の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，5 名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 であった．また選択肢読み上げ問題では，相関係数は 0.89，RMSE は 0.64 であった．

Section 2 の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．これらの特徴量を用いて推定したスコアと，5 名の評定者によるスコアの平均を比較したところ，相関係数は 0.70，RMSE は 1.25 であった．また自由発話問題については，文生成問題で用いられている特徴量に加え，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，8 名の評定者によるスコアの平均を比較したところ，相関係数は 0.91，RMSE は 0.63 であった．

3 提案手法

上述したように，採点アルゴリズムによる解答発話に対する採点の誤りは受験者の能力値推定に悪影響を及ぼす [8]．本稿では，この問題を解決するため，採点の信頼度が顕著に低い解答発話を自動で検出し，能力値推定に用いないようにする手法を提案する．本手法は 2.1 節におけるステップ 2 の後に適用する．

まず，項目選択アルゴリズムにより選択された項目に対して受験者が解答を行う．その解答発話に対し採点アルゴリズムと同じ特徴量を用いて，0，1，…，4 の 5 クラスで確率推定付きのクラス分類を行う．なお，本稿ではクラス分類器として SVM（Support Vector Machine）を用いた．クラス分類により，最も所属している確率が高いクラスに属する確率値と，2 番目に所属している確率が高いクラスに属する確率値を得る．次に，これら 2 つの確率値の対数尤度比を求め，これを解答発話に対する採点信頼度とする．あらかじめ閾値 δ_1（文生成問題用），δ_2（自由発話問題用）を定めておき，採点信頼度がその閾値を下回った場合は 2.1 節におけるステップ 1 に戻る（すなわち選択された項目をスキップし能力値分布を更新しないようにする）．採点信頼度が閾値を上回った場合は，2.1 節のステップ 3 に進む（すなわち採点アルゴリズムにより求めたスコアを用いて能力値分布を更新する）．

4　提案手法の有効性の検証

4.1　実験概要

提案手法の有効性の検証を行うため，次のシステムを用意した．なお，本実験は採点アルゴリズムにおける採点精度の影響がより大きいと考えられる Section 2 を対象とする．

- *conventional system*：2 章で述べたシステム．
- *conventional system with human raters*：*conventional system* と同じだが，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．ここで，項目プールのすべての項目を用いて推定した場合の能力値を真の能力値とみなす．
- *proposed system*：3 章で述べた提案手法を実装したシステム．ここで，採点信頼度の算出に用いる SVM のカーネルは RBF であり，採点アルゴリズムの学習に用いたデータにより学習した（SVM パラメータはグリッドサーチにより最適化した）．

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．項目プールに含まれる項目数は計 43 問である．*conventional system with human raters* における評定者は文生成問題は 5 名，自由発話問題は 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（推定誤差）と，収束条件を満たすまでに出題された項目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．また，*proposed system* における採点信頼度の値は文生成問題と自由発話問題共に 0.1 から 0.5 の範囲に集中していることが分かった．そこで，採点信頼度の閾値 δ_1，δ_2 はそれぞれ $0.1, 0.2, ..., 0.5$ の範囲で変化させることにした．

Table 2　実験条件

受験者	留学生 20 名
項目プール	43 問 (文生成 34 問，自由発話 9 問)
評定者の人数	文生成 5 名，自由発話 8 名

4.2　実験結果

収束条件の閾値 ϵ を変化させたときの，能力値の推定誤差と出題された項目数の関係を Fig. 2 に示す．ここで，縦軸は収束条件を満たすまでに出題された項目数，横軸は能力値の推定誤差である（いずれも受験者 20 名の平均）．

Fig. 2 より，*conventional system with human raters*（青点線）は他のシステムと比較してより少ない出題項目数で高精度に能力値を推定できているということが分かった．*conventional system*（青実線）は解答発話に対する採点誤りの影響により，能力値推定精度が大きく低下してしまっている．*proposed system*（赤実線）は上記の 2 つのシステムの間に曲線が位置しており，*conventional system* と比較して出題項目数にはあまり変化はないが，能力値推定精度が向上していることが確認できる．

最後に採点信頼度の閾値について考察する．Fig. 2 に示されている *proposed system* の閾値は，最も良い結果が得られた $\delta_1 = 0.1$，$\delta_2 = 0.2$ の場合である．これは，採点信頼度が顕著に低かった項目のみをスキップした場合に相当する．一方，閾値を大きくした場合は採点誤りがなかった項目を数多くスキップしてしまうことになり，このことは出題項目数の増加と能力値推定精度の低下をまねくことが分かった．

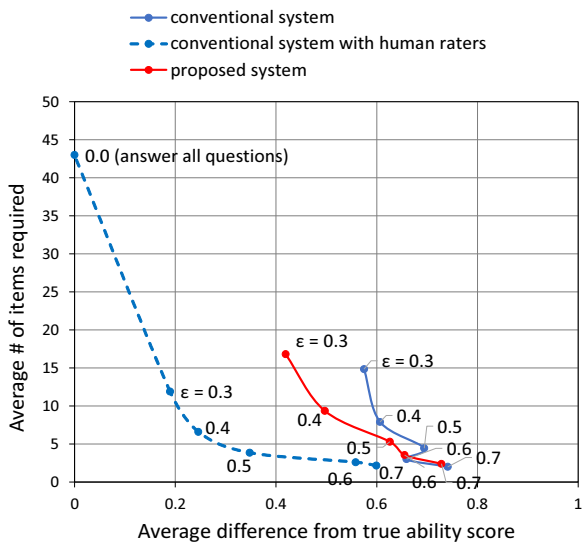

Fig. 2　能力値の推定誤差（横軸）と出題項目数（縦軸）

5　おわりに

本稿では，SJ-CAT における能力値推定精度を改善するため，採点信頼度が顕著に低い解答発話を自動で検出し，能力値推定に用いないようにする手法を提案した．実験を行った結果，提案手法は，すべての解答発話を用いて能力値推定を行う従来手法と比べて，出題項目数を増加させることなく能力値推定精度を改善できることが確認された．

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．

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] 小野友暉, 山田武志, 菊地賢一, 今井新悟, 牧野昭二, “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,” 日本音響学会秋季研究発表会, pp. 253–256, Sep. 2016.

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

[10] 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.