日本語スピーキングテストS-CATの文読み上げ問題における発話の冗長性・不完全性を考慮した自動採点の検討*

☆山畑勇人，盧昊，山田武志，△今井新悟 (筑波大)，
△石塚賢吉 (株式会社ドワンゴ)，牧野昭二，北脇信彦 (筑波大)

1 はじめに

現在，日本語学習者の日本語能力を自動で評価するテストとして，J-CAT（Japanese Computerized Adaptive Test)[1] が運用されており，国内外で広く利用されている．J-CAT ではウェブブラウザを用いてインターネット上で受験ができる．また，項目応答理論 [2] を用いたアダプティブテストであるため，テストの所要時間の短縮と，採点精度の向上を同時に実現している．現在のところ，J-CAT は聴解，語彙，文法，読解のセクションからなり，発話能力の評価は行われていない．そこで，J-CAT への導入を目指し，自動採点形式のスピーキングテストである S-CAT（Speaking section of J-CAT）が開発されており [3]，これまでにいくつかの自動採点手法が提案されている [4–7].

S-CAT には，文読み上げ，選択肢読み上げ，文生成，自由発話の 4 つの問題が設定されている．本稿では，解答者が読み上げ対象文を読み上げる，文読み上げ問題を対象とする．文読み上げ問題において，評定者（日本語教師）は採点ガイドラインに基づいた 0〜4 の 5 段階絶対評価で総合点をつける．その自動採点には，解答音声から抽出する特徴量の選定，特徴量から総合点を推定するための推定モデルの構築が必要になる．評定者は，発音などの個別要因を考慮して総合点を評価すると考えられることから，我々は主観評価実験を行うことにより，総合点と個別要因の関係を調査した．その結果，文読み上げ問題においては，発音と流暢さが総合点に特に寄与していることが分かった．そして，これらの個別要因を考慮した自動採点手法を提案し，一定の有効性を確認した [6, 7].

しかし，次に述べる 2 つの性質がある発話に対しては，推定誤差が大きいという問題がある．1 つ目は，読み上げ対象文中の単語などを繰り返し発話してしまうといった，言い直しが存在する場合である．2 つ目は，発話が完結していない，つまり制限時間内に読み上げ対象文を最後まで発話できなかった場合や，音節などを局所的に読み誤ってしまった場合である．

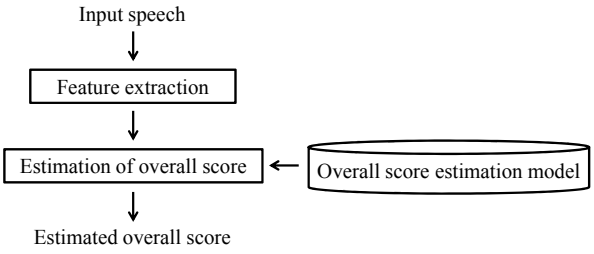

Fig. 1 Overview of the proposed method.

本稿では，これら 2 つの性質をそれぞれ発話の冗長性，及び不完全性と呼ぶ．また，従来手法では，推定モデルの構築のために各個別要因の主観スコアが必要であり，学習データの量を容易に増やすことができないため，ロバストな推定モデルを構築しにくいという問題がある．

以上より，本稿ではこれらの問題に対処できる特徴量及び推定モデルを導入する．そして，実験によりその有効性を示す.

2 発話の冗長性・不完全性を考慮した自動採点

2.1 提案手法のアプローチ

発話の冗長性・不完全性を考慮した自動採点手法のフローチャートを Fig.1 に示す．提案手法では，解答音声から抽出した特徴量を用いて総合点を直接推定する．従って，従来手法とは異なり，学習データの量を比較的容易に増やすことができる．特徴量としては，従来手法において発音と流暢さの主観スコアを推定するために使用した計 5 種類の特徴量に，発話の冗長性・不完全性に対処するための 3 種類の特徴量を追加した計 8 種類を使用する．総合点推定モデルには，RBF カーネルを用いた SVR（Support Vector Regression）[8] を採用する．小野らは，S-CAT の自由発話問題の自動採点に SVR を初めて導入し，その有効性を示している [4].

*Automatic scoring method for sentence-read-aloud task in S-CAT Japanese speaking test considering the redundancy and incompleteness of utterance. by Yuto YAMAHATA, Hao LU, Takeshi YAMADA, Shingo IMAI (Univ. of Tsukuba), Kenkichi ISHIZUKA (DWANGO Co., Ltd), Shoji MAKINO, Nobuhiko KITAWAKI (Univ. of Tsukuba)

2.2 提案手法の特徴量

従来手法では，個別要因として設定した発音と流暢さの主観スコアを推定するために，それぞれ 2 種類と 3 種類の特徴量を使用した．これらの特徴量は評定者による総合点と強い相関があったことから，同様に使用することとする．

従来手法において，発音スコアの推定には以下の 2 種類の特徴量を使用した．

x_1：GOP_{all}

x_2：articulation rate

GOP_{all} は，読み上げ対象文への音素アライメントと連続音素認識結果への音素アライメントを用いて，各音素毎に GOP（Goodness Of Pronunciation）[9] を計算し，その平均をとったものである．なお，GOP とは 2 つのアライメント間のフレーム対数尤度差である．本稿では，対象が日本語であるため連続音節認識（音節制約つき連続音素認識）を用いた．articulation rate[10] は音素レベルの調音速度である．発音が良いほど x_1，x_2 共に大きくなる．

また従来手法において，流暢さスコアの推定には以下の 3 種類の特徴量を使用した．

x_3：発話区間における無音区間の割合

x_4：発話区間における無音区間の回数

x_5：発話区間における音節長の変動係数
（音節長の標準偏差を音節長の平均で割った値）

流暢さは，発話中に言い淀みや音節長の間延びが存在する場合に低い評価になっていた．そこで，x_3 は言い淀みの程度，x_4 は言い淀みの回数，x_5 は音節長のばらつきの程度に着目している．x_3，x_4，x_5 の値が大きくなるほど，発話が流暢でなくなると考えられる．

提案手法では x_1〜x_5 の 5 種類の特徴量に加え，発話の冗長性・不完全性に対処するために，以下の特徴量 x_6，x_7，x_8 を導入する．

x_6：読み上げ対象文の音節数に対する
連続音節認識で出力された音節数の割合

x_7：読み上げ対象文の母音における GOP の最小値

x_8：読み上げ対象文の子音における GOP の最小値

ここで，x_6 は発話の冗長性と不完全性の両方に対処するための特徴量である．言い直しがある場合，実際に発話した音節数は読み上げ対象文の音節数に比べて多くなる．よって，x_6 の値は大きくなると想定される．逆に，発話が完結していない場合，x_6 の値は小さくなると想定される．また，x_7，x_8 は，発話中の局所的な読み誤りに対処するための特徴量である．読み誤りが存在する個所は，発音が著しく悪い個所と見なすことができる．従って，局所的な発音の良し悪しを評価することに対応する．読み誤りが存在する場合，x_7，x_8 の値は特に小さくなると想定される．

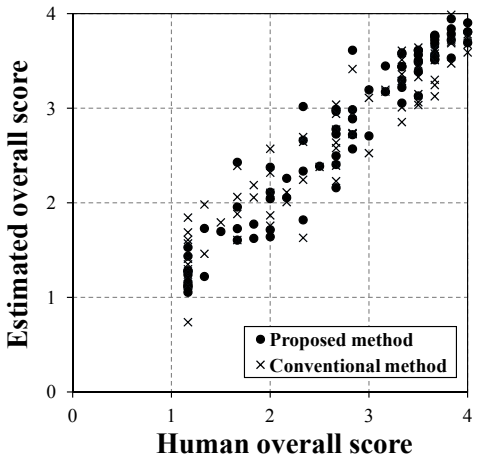

Fig. 2 Relationship between the human overall score and the overall score estimated by the conventional and proposed methods in the closed test.

2.3 従来手法との比較による有効性の評価

従来手法と同条件で総合点を推定する実験を行い，推定精度を比較する．従来手法では，20 名の解答者が 4 設問に解答した計 80 個の解答音声サンプルを学習データとした．まず，別途主観実験により得た発音及び流暢さの主観スコアから総合点を推定するモデルを線形重回帰により構築した．さらに，特徴量 x_1 と x_2 から発音，x_3〜x_5 から流暢さの主観スコアを推定するモデルを線形重回帰により構築した．そして，評価データから抽出した x_1〜x_5 と 3 種類（発音スコア，流暢さスコア，総合点）の推定モデルを用いて総合点を推定した．一方，提案手法では，従来手法と同じ学習データから抽出した x_1〜x_8 から総合点を推定するモデルを SVR により構築した．そして，評価データから抽出した x_1〜x_8 と総合点推定モデルを用いて総合点を推定した．なお，特徴量抽出の際，音声認識器には Julius[11]，音響モデルには日本人の音声から学習された IPA の不特定話者 PTM トライフォンモデル [12] を用いた．

まずクローズドテストについて述べる．従来手法と提案手法の両方において，学習データを評価データとして使用し，総合点を推定した．評定者による総合点と推定した総合点の関係を Fig.2 に示す．なお，評定者による総合点は評定者 6 名の平均値を用いた．Fig.2 中の×のプロットが従来手法によるものであり，評定

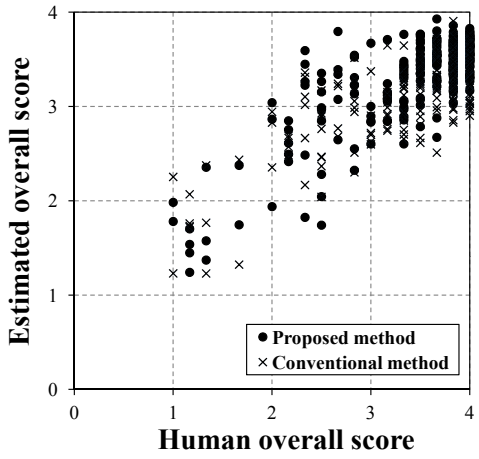

Fig. 3 Relationship between the human overall score and the overall score estimated by the conventional and proposed methods in the speaker and question open test.

者による総合点と推定した総合点の相関係数は 0.95, RMSE は 0.31 である．●のプロットが提案手法によるものであり，相関係数は 0.97，RMSE は 0.24 である．このことから，提案手法において推定精度が向上していることが分かる．

次に，解答者と設問の両方をオープンにしたテスト（以下，解答者・設問オープンテストと呼ぶ）について述べる．評価データには，学習に用いたものと異なる解答者 20 名と設問 13 問を用いた．このうち無音であったり，雑音が大きい解答音声サンプルを除き，計 255 サンプルを使用した．これらのサンプルに対し，従来手法と提案手法の両方で総合点の推定を行った．評定者による総合点と推定した総合の関係を Fig.3 に示す．従来手法における相関係数は 0.79, RMSE は 0.51 である．また，提案手法における相関係数は 0.81，RMSE は 0.43 である．これより，解答者・設問オープンテストにおいても推定精度の向上を確認できる．特に，冗長性や不完全性が多く存在する，評定者による総合点が低いサンプルにおいて，推定誤差が小さくなっていることが分かる．以上から，提案手法の有効性が確認された．

3 設問の違いが推定精度に及ぼす影響の調査

3.1 調査の方法

2.3 節の実験において，提案手法による推定精度には，クローズドテストと解答者・設問オープンテストで大きな差があった．そこで，設問の違いが推定精度に及ぼす影響を，クロスバリデーションを行うことにより調査する．

Table 1 Training and test sets.

	Question number	
Set	Training	Test
Set 1	5 6 7 8 9 10 11 12 13 14 15 16 17	1 2 3 4
Set 2	1 6 7 8 9 10 11 12 13 14 15 16 17	2 3 4 5
⋮	⋮	⋮
Set 17	4 5 6 7 8 9 10 11 12 13 14 15 16	17 1 2 3

文読み上げ問題には設問が 17 問ある．これらを 13 問と 4 問に分け，それぞれ学習用と評価用とした．設問の分け方を Table 1 に示す．設問番号の連続する 4 問が評価用となるように分けたため，分け方は 17 パターン存在する．なお，解答者は学習用が 169 名，評価用が 20 名であり，これらは固定とした．以上から，学習用の解答音声サンプルは 2197（169 名× 13 問）個である．なお，これらのサンプルはクローズドテストでも使用する．また，評価（解答者・設問オープンテスト）用の解答音声サンプルは 80（20 名× 4 問）個である．これらのうち無音であったり，雑音が大きい解答音声サンプルは除いている．17 パターンの各セットにおいて，学習用の解答音声サンプルを用いて，2.3 節と同様の手順で総合点推定モデルを構築し，クローズドテストと解答者・設問オープンテストを行った．

3.2 実験結果と考察

まず，クローズドテストについて述べる．評定者による総合点と推定した総合点の相関係数，RMSE を Fig.4 に示す．横軸が各セットと平均を示している．縦軸は相関係数と RMSE の値であり，黒いバーが相関係数，折れ線が RMSE を示している．Fig.4 より，どのセットにおいても安定して高い推定精度が得られていることが分かる．これより，学習に使用した設問であれば，提案手法により良好な精度で総合点を推定可能であることが確認できた．

次に，解答者・設問オープンテストについて述べる．評定者による総合点と推定した総合点の相関係数，RMSE を Fig.5 に示す．クローズドテストに比べて，推定精度に大きなばらつきがみられることが分かる．解答者・設問オープンテストにおいて推定精度が特に低かったのは，Set4 から Set7 までの 4 セットであった．これらのセットには共通して以下の設問が含まれており，この設問に対する推定精度は他の設問と比較して特に低かった．

- 親戚のおじさんやおじいさんやおばさんやおばあさんが集まった

この設問は，他の設問と比べ，評価のために着目す

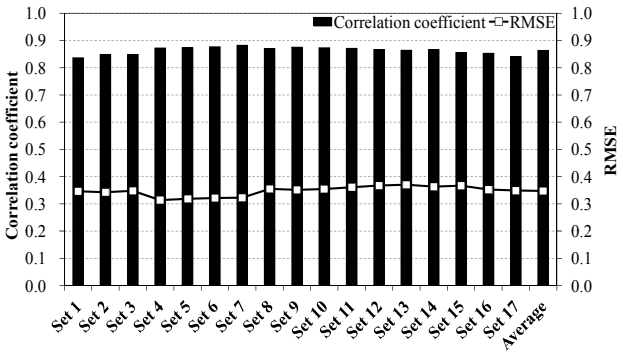

Fig. 4 Correlation coefficient and RMSE between the human overall score and the overall score estimated by the proposed method on each test set in the closed test.

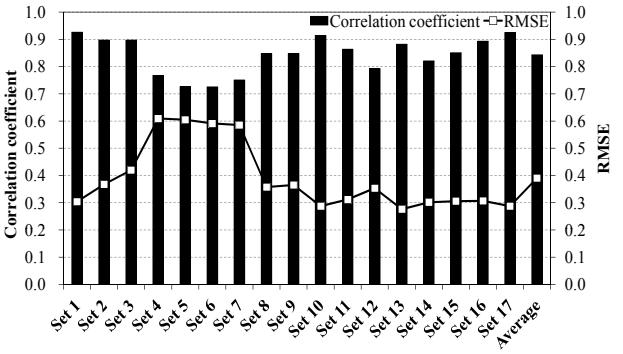

Fig. 5 Correlation coefficient and RMSE between the human overall score and the overall score estimated by the proposed method on each test set in the speaker and question open test.

べき点が特に明確である．つまり，この設問においては,「じ」と「じい」,「ば」と「ばあ」のような直音と長音を適切に区別し発話できているか，ということに着目して総合点の評価が行われている．他の設問は，発話全体に着目して総合点の評価が行われる傾向があったため，この特有の特徴が原因で推定精度が低くなったと考えられる．このように他と明らかに異なる特徴がある設問に対しては，個別に推定モデルを構築するなど，その扱い方を検討する必要がある．

4 おわりに

本稿では，S-CAT における文読み上げ問題を対象として，発話中の言い直しなどの冗長性，発話が完結していないなどの不完全性を考慮した自動採点手法を提案した．従来手法との比較を行った結果，クローズドテストと解答者・設問オープンテストの両方で推定精度が向上した．一方，クローズドテストに比べ，解答者・設問オープンテストでは推定精度が大きく低下したため，設問の違いが推定精度に及ぼす影響をクロスバリデーションにより調査した．その結果，他とは明らかに異なる特徴がある設問に対しては，個別に推定モデルを構築するなど，扱い方を検討する必要があることが分かった．

今後の課題としては，特有の特徴をもつ設問の扱い方の検討，及び提案手法を実際のシステムに組み込むことで挙動や実用性を確認することが挙げられる．

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

参考文献

[1] J-CAT, http://www.j-cat.org/.

[2] M. Du. Toit(ed.), “IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT,” Scientific Software International, 2002.

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

[4] 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, pp. 1–4, 2012.

[5] 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, 3-Q-17, pp. 397-399, 2012

[6] N. Okubo, Y. Yamahata, T. Yamada, S. Imai, K. Ishizuka, T. Shinozaki, R. Nisimura, S. Makio, N. Kitawaki, “Automatic Scoring Method Considering Quality and Content of Speech for SCAT Japanese Speaking Test,” Proc. OCOCOSDA2012, pp. 72–77, 2012.

[7] 山畑勇人, 大久保梨思子, 山田武志, 今井新悟, 石塚賢吉, 篠崎隆宏, 西村竜一, 牧野昭二, 北脇信彦, “日本語スピーキングテスト SCAT における文読み上げ・文生成問題の自動採点手法の改良, ”日本音響学会春季研究発表会, 1-Q-52a, pp. 465–468, 2013.

[8] D. Basak, S. Pal, D. Chandra Patranabis, “Support Vector Regression,” Neural Information Processing-Letters and Reviews., Vol. 11, No. 10, pp. 203–224, 2007.

[9] L. Neumeyer, H. Franco, V. Digalakis, M. Weintraub, “Automatic scoring of pronunciation quality,” Speech Commun., Vol. 30, pp. 83–93, 2000.

[10] 中村直生, 中川聖一, “日本人の英語発音評価法,” 電子情報通信学会技術研究報告. SP, 音声 102(107), pp. 13–18, 2002.

[11] 河原達也, 李晃伸, “連続音声認識ソフトウエア Julius,” 人工知能学会誌, Vol. 20, No. 1, pp. 41–49, 2005.

[12] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, K. Shikano, “Free software toolkit for Japanese large vocabulary continuous speech recognition,” Proc. ICSLP2000, pp. 476–479, 2000.