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*

☆山畑勇人，大久保梨思子，山田武志，△今井新悟，△石塚賢吉（筑波大）
篠崎隆宏（千葉大），西村竜一（和歌山大），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，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．そこで我々は，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 SCAT（Speaking section of J-CAT）を開発している．SCAT には，文読み上げ，選択肢読み上げ，空所補充，文生成，自由発話の 5 つのタスクが設定されている．この順に解答の自由度が高くなり，自動採点も難しくなる．これは，読み上げのように発話内容が既知である問題であれば，発話音声そのものを評価すればよいが，文生成以降の問題は発話内容が未知であるために，発話音声に加え発話内容の評価も必要になるためである．

自動採点の実現のためには，日本語教師による採点（以下，総合点と呼ぶ）を，解答音声から抽出した特徴量を用いて推定する必要がある．なお SCAT における総合点は，採点ガイドラインに基づいた 0〜4 の 5 段階絶対評価で採点される．また，発音や聴解力などの個別要因に対する採点は行われていない．そこで我々はまず，文読み上げ，及び文生成問題を対象として，総合点に影響を及ぼす個別要因を主観評価実験により調査した．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 [2]．本稿では，提案手法の推定精度をさらに向上させるために，設問の難易度と解答者の日本語習熟度をより幅広く設定した主観評価実験を実施し，その結果に基づいて推定モデルと特徴量の改良を行う．

2 提案手法の概要

提案手法のフローチャートを Fig.1 に示す．提案手法では，まず解答音声から特徴量を抽出し，それを用いて各個別要因の主観値（以下，個別要因スコアと呼ぶ）を推定する．次に，推定された個別要因スコア（以下，個別要因推定スコアと呼ぶ）を用いて総合点を推定する．なお，個別要因はタスクの種類に応じて事前に決定する．後者の推定に用いる総合点推定モデルは，個別要因スコアに基づいて予め決定する．提案手法の特徴は，個別要因推定スコアから総合点を推定することにある．これは，特徴量から直接総合点を推定するよりは容易であると考えられる．また，受験者に対し総合的な採点結果だけでなく，発話能力を改善するためのアドバイスを示すような応用も可能である．

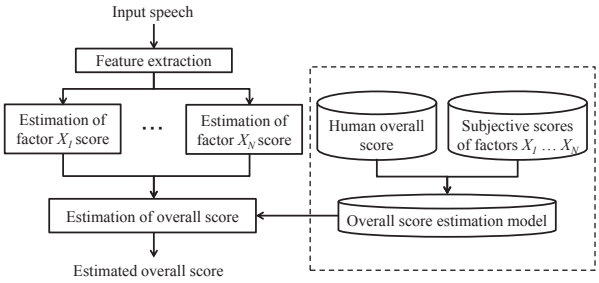

Fig. 1 Overview of the proposed method

3 文読み上げ問題の自動採点手法の改良

3.1 個別要因の設定

文読み上げ問題は発話内容が既知とみなせるため，発話内容ではなく発話音声に関する要因が総合点に影響を及ぼすと考えられる．そこで文献 [3, 4] を参考に，発音（X_1），イントネーション（X_2），アクセント（X_3），流暢さ（X_4）を設定した．なお，流暢さは，言い淀みの有無等の時間軸方向のスムーズさに着目する．文献 [2] においては，要因としてラウドネス（X_5）も設定しているが，実験の結果，総じて高得点であったため本稿では除外した．

3.2 主観評価実験

主観評価実験を行うことで，各個別要因と総合点との関係を調査する．被験者は日本人学生 20 名である．被験者は，防音室内でヘッドホンにより留学生の解答音声サンプルを受聴し，3.1 節の 4 つの要因をそれぞれ評価した．文読み上げ問題における解答音声サンプル数は，4 つの設問に対して留学生 20 名が発話した計 80 個である．文読み上げ問題の各設問が，項目応答理論の困難度 [5] に従い 4 段階の難易度に分けられるため，設問は各難易度から 1 問ずつ選択した．また，留学生は，日本語習熟度の高い者から低い者までが満遍なく含まれるように選択した．評価尺度は一般の日本人が日常会話で使用する標準的な日本語を基準とした 0〜4 の 5 段階絶対評価尺度である．最終的な個別要因スコアは被験者 20 名の平均を用いる．総合点は日本語教師 6 名の平均である．文献 [2] の実験条件と比べて，設問数，被験者数，日本語教師数を増加させた．また，留学生の習熟度，設問の難易度をバランスよく設定している．

3.3 実験結果と考察

主観評価実験の結果，各要因間には相関の強い組が存在した．そこで，日本語教師による総合点を目的

*An improvement of automatic scoring method for sentence-reading-aloud/sentence generation tasks in SCAT Japanese speaking test. by Yuto YAMAHATA, Naoko OKUBO, Takeshi YAMADA, Shingo IMAI, Kenkichi ISHIZUKA (Univ. of Tsukuba), Takahiro SHINOZAKI (Chiba Univ.), Ryuichi NISHIMURA (Wakayama Univ.), Shoji MAKINO, Nobuhiko KITAWAKI (Univ. of Tsukuba)

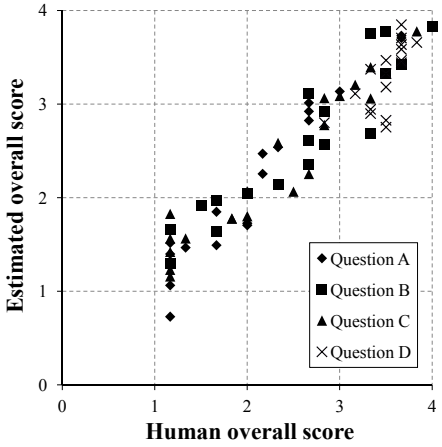

Fig. 2 Estimation result

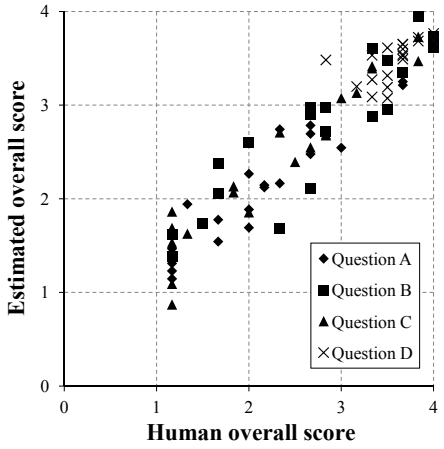

Fig. 3 Estimation result

変数，各要因スコアを説明変数として，ステップワイズ法により変数選択を行った．その結果，文献 [2] と同様に発音（X_1）及び流暢さ（X_4）が選択され，式（1）の関係が得られた．

$$\text{総合点} = 0.48X_1 + 0.47X_4 + 0.47 \qquad (1)$$

これが，図 1 における総合点推定モデルである．なお，設問の区別はせず，全ての設問のデータを用いて 1 つのモデルを構築した．これは，設問に依存せず，どのような設問に対しても適用できる共通のモデルを構築するためである．

次に，発音（X_1）スコアと流暢さ（X_4）スコアを式（1）に再度代入して総合点を推定した．教師による総合点と，推定された総合点の関係を Fig.2 に示す．相関係数は 0.96，RMSE は 0.27 であり，非常に高い精度で総合点を推定できることが分かる．また，設問毎に大きな差が見られないことから，様々な難易度の設問に対してロバストな推定ができる見込みが得られた．なお，この場合の相関係数，及び RMSE の値が，提案手法における推定精度の上限にあたる

3.4 各個別要因スコアの推定に用いる特徴量の検討

これまでに我々は，発音（X_1）スコアの推定に用いる特徴量として，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の対応関係に注目した 2 種類の特徴量を提案した [2]．本稿では，それらの特徴量の他に，文献 [6] に記載されている GOP_{all}，GOP_{vow}，GOP_{cons}，及び文献 [7] に記載されている articulation rate を新たに加え，比較検討を行った．検討の結果，GOP_{all} と articulation rate を用いた場合に最も良い結果となることが分かった．そこで，GOP_{all} と articulation rate を発音（X_1）スコア推定のための特徴量として使用する．

x_{1a}：GOP_{all}
x_{1b}：articulation rate

なお，GOP_{all} は，上述したアライメント間の尤度比，articulation rate は音素レベルの調音速度である．発音が良いほど，x_{1a}，x_{1b} 共に大きくなる．

流暢さ（X_4）スコアの推定に関しては，読み上げ対象文への音素アライメント結果を用いて算出される，以下の特徴量 x_{4a}，x_{4b} を用いる [2]．

x_{4a}：発話区間における無音区間の割合
x_{4b}：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）

x_{4a} は言い淀みの程度，x_{4b} は音節長のばらつきの程度に着目している．また，3.2 節の解答音声サンプルを観察した結果，言い淀みの回数が多いサンプルほど，流暢さ（X_4）スコアが低くなっていることが分かった．そこで本稿では，以下の特徴量を追加する．

x_{4c}：発話区間における無音区間の回数

x_{4c} は，言い淀みの回数に着目した特徴量である．x_{4a}，x_{4b}，x_{4c} 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．

3.5 提案手法の有効性の評価

発音（X_1）スコア，及び 3.2 節の音声サンプルから算出した x_{1a}，x_{1b} の値を用いた重回帰により，式（2）の発音推定モデルを得た．流暢さに対しても，同様の手順で，式（3）の流暢さ推定モデルを得た．

$$X_1 = 0.74x_{1a} + 21.7x_{1b} + 0.91 \qquad (2)$$

$$X_4 = -2.47x_{4a} - 1.48x_{4b} - 0.10x_{4c} + 3.76 \qquad (3)$$

なお，特徴量抽出の際，音声認識器には Julius[8]，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 [9] を用いた．

3.2 節の解答音声サンプルから算出した各特徴量の値を，式 (2),(3) にそれぞれ代入し，主観値である各要因スコアを推定した．その結果，発音（X_1）では，主観値と推定値の相関係数が 0.92，RMSE が 0.34 となり，流暢さ（X_4）では，相関係数が 0.92，RMSE が 0.40 となった．比較として，文献 [2] の特徴量を用いて発音（X_1）及び流暢さ（X_4）スコアの推定モデルを同様に構築し，各要因スコアを推定した．その結果，発音（X_1）では，相関係数が 0.84，RMSE が 0.46，流暢さ（X_4）では，相関係数が 0.89，RMSE が 0.47 となった．これにより，推定精度が向上していることが分かる．

次に，各要因推定スコアを式（1）に代入して総合点を推定した．教師による総合点と，推定した総合点の関係を Fig. 3 に示す．相関係数は 0.95，RMSE は 0.30 である．各要因スコアを代入した場合（Fig. 2）に迫る高い推定精度が得られた．

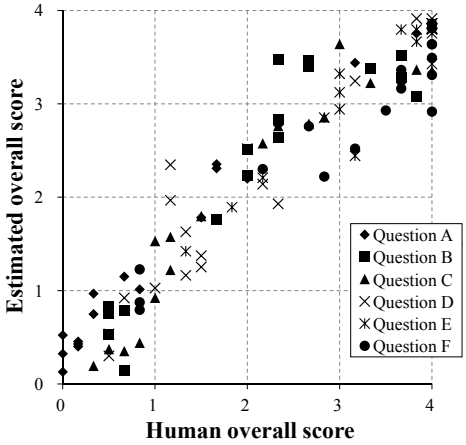

Fig. 4 Estimation result

4 文生成問題の自動採点手法の改良

4.1 個別要因の設定

文生成問題は，受験者の発話内容が未知の問題である．従って，発話内容の良し悪しも総合点に影響する．そのため文献 [2] と同様に，3.1 節で設定した 4 つの要因に加え，発話内容に関する要因として聴解力（X_6），表現力（X_7），文法力（X_8），語彙力（X_9）を設定した．

4.2 主観評価実験

3.2 節と同じ条件で主観評価実験を行った．なお，文生成問題においては，6 つの設問を用いている．文生成問題の各設問が，項目応答理論の困難度 [5] に従い 6 段階の難易度に分けられるため，設問は各難易度から 1 問ずつ選択した．

4.3 実験結果と考察

3.3 節と同様に，日本語教師による総合点を目的変数，各要因スコアを説明変数として，ステップワイズ法により変数選択を行った．その結果，聴解力（X_6）と語彙力（X_9）が選択され，式（4）の総合点推定モデルが得られた．

$$\text{総合点} = 0.25X_6 + 1.05X_9 - 1.22 \qquad (4)$$

この結果は，個別要因として発音（X_1），聴解力（X_6），表現力（X_7）が選択された文献 [2] とは異なるものとなった．しかし，要因間の相関をみると，表現力（X_7）と語彙力（X_9）には非常に強い相関があることが分かった．従って，ワンセンテンスで解答を行う文生成問題において，表現力（X_7）と語彙力（X_9）は同一の事象を反映していると考えられる．また，式（4）の要因に，さらに発音（X_1）を追加してモデルを構築しても，精度は向上しなかった．よって，文生成問題の総合点には，発音（X_1）はそれほど寄与しないと考えられる．

主観評価実験により得られた，聴解力（X_7）スコア，及び語彙力（X_9）スコアを式（4）に代入し，総合点を推定した．教師による総合点と，推定された総合点の関係を Fig.4 に示す．相関係数は 0.95，RMSE は 0.42 と，文読み上げ問題同様，非常に高い精度が得られた．こちらも設問毎に大きな差は見られない．

4.4 各個別要因スコアの推定に用いる特徴量の検討

聴解力（X_6）スコア，及び表現力（X_7）スコアを推定するための特徴量として，以下の特徴量 $x_{6\cdot9a}$，$x_{6\cdot9b}$ を用いる．

$x_{6\cdot9a}$：解答に含まれるべきキーワード数に対する，抽出されたキーワード数の割合

$x_{6\cdot9b}$：重要キーワードの音素列と，連続音素認識で得た音素列の各部分区間との正規化編集距離の最小値に基づく類似度

$x_{6\cdot9a}$ は，これまでに我々が提案した，ディクテーションとキーワードスポッティングの認識結果を利用した特徴量である [2]．$x_{6\cdot9b}$ は，本稿で新たに提案する特徴量である．文生成問題では，自動採点を行うことを見越して解答に含まれるべきキーワードが設定されている．これが重要キーワードである．重要キーワードの有無は，採点結果に多大な影響を及ぼす．また，文末語が正しく発話できているか否かが，各要因の評価に影響を及ぼしていることが分かっている．そこで，想定される複数の文末語を文末キーワードとする．$x_{6\cdot9a}$ はディクテーションの認識結果に，これらのキーワードがどの程度含まれているのか，ということに注目している．ここで，重要キーワードは文末キーワードよりも大きな重みでカウントしている．

また，キーワードを完璧に発話できていなくても，数音素違いの誤りであれば，聴解力（X_6）スコアは高くなる傾向があった．これは，聴解力（X_6）が設問を正しく理解しているかを評価する要因であるため，キーワードに近い発話をしている場合は，質問の意図については正しく理解していると捉えられるからだと考えられる．そこで，キーワードを数音素違いで発話した場合を考慮するために，$x_{6\cdot9b}$ を導入する．なお，$x_{6\cdot9b}$ において，重要キーワードが複数存在する場合は，各キーワードごとに算出し重要キーワードの数で平均する．

4.5 提案手法の有効性の検証

各個別要因スコアと 4.2 節の音声サンプルから算出した $x_{6\cdot9a}$，$x_{6\cdot9b}$ の値の間には非線形な関係があった．そこで，シグモイド曲線を用いてモデリングを行い，式（5）の聴解力（X_6）推定モデル，式 (6) の語彙力（X_9）推定モデルを得た．

$$X_6 = \frac{3.90}{1 + exp(2.86 - 5.33x_{6\cdot9a} - 4.72x_{6\cdot9b})} \qquad (5)$$

$$X_9 = \frac{3.62}{1 + exp(1.23 - 3.80x_{6\cdot9a} - 1.15x_{6\cdot9b})} \qquad (6)$$

ここで，特徴量抽出のための音声認識器は Julius[8] である．音響モデルは日本語話し言葉コーパス（CSJ）[10] で学習したトライフォンモデルを留学生の解答音声で適応したものを用いた．ディクテーションにおける言語モデルは，留学生の解答音声の書き起こし，ウェブテキスト，新聞記事の融合モデルである．キーワードスポッティングでは，連続音素認識をガーベージとした．

4.2 節の音声サンプルから算出した特徴量 $x_{6\cdot9a}$，$x_{6\cdot9b}$ を，式（5），(6）にそれぞれ代入し，主観値である各要因スコアを推定した．その結果，聴解力（X_6）では，主観値と推定値との相関係数が 0.82，RMSE

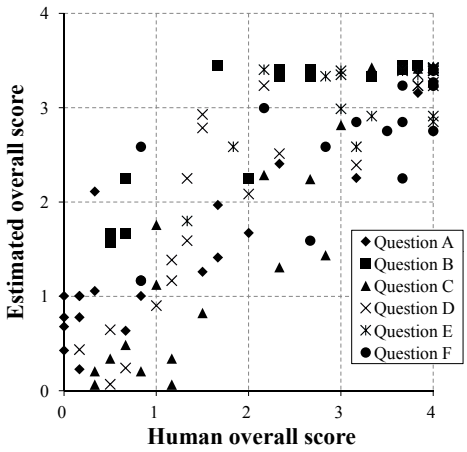

Fig. 5 Estimation result

が 0.70 となり，語彙力（X_9）では，相関係数が 0.84，RMSE が 0.55 となった．比較として，文献 [2] の特徴量を用いて聴解力（X_6）及び語彙力（X_9）スコアの推定モデルを同様に構築し，各要因スコアを推定した．その結果，聴解力（X_6）では，相関係数が 0.79，RMSE が 0.74，語彙力（X_9）では，0.84，RMSE が 0.56 となった．これにより，特に聴解力 (X_6) スコアにおいて推定精度が向上していることが分かる．

次に，各要因推定スコアを式（4）に代入して総合点を推定した．教師による総合点と，推定した総合点の関係を Fig. 5 に示す．相関係数は 0.85，RMSE は 0.74 である．各要因スコアを代入した場合（Fig. 4）に比べ，多少のばらつきが見られるものの，0.80 以上の強い相関が確認された．

5 オープンテストによる評価

3 章と 4 章では，推定モデルの構築と評価に同一のデータを用いるクローズドテストによって，有効性を検証した．本節では，提案手法の頑健性を評価するために，解答者をオープンにする場合，及び設問をオープンにする場合の 2 種類のオープンテストを行う．解答者オープンテストでは，新たな留学生 10 名が，3.2 節及び 4.2 節の設問に対し解答した音声サンプルを使用する．設問オープンテストでは，3.2 節及び 4.2 節の留学生の内の 10 名が，新たに選択した文読み上げ問題 4 問，文生成問題 6 問に対し解答した音声サンプルを使用する．それぞれのテストにおいて，各要因スコアから総合点を推定した場合，及び各要因推定スコアから総合点を推定した場合について検討を行う．

文読み上げ問題において，各オープンテストにおける，日本語教師による総合点と推定された総合点の相関係数と RMSE を Table 1 に示す．また，文生成問題における同様の結果を Table 2 に示す．Table 1，2 より，両方の問題において，各要因スコアから総合点を推定した場合は，高い精度で総合点を推定可能なことが分かる．これより，式（1）,（4）の総合点推定モデルが，解答者と設問の違いに頑健であることが確認された．一方で，各要因推定スコアから総合点を推定した結果をみると，ほぼすべての場合で推定精度が若干低くなっている．これは，各要因スコアを推定する際の推定誤差の影響を受けているためである．推定精度向上のためには，特徴量，及び各要因推定モデルのさらなる改良が必要である．

Table 1 Correlation coefficients and RMSE in sentence-reading-aloud task

	From subjective scores	From estimated scores
Closed	R:0.96, RMSE:0.27	R:0.95, RMSE:0.30
Open(speaker)	R:0.87, RMSE:0.44	R:0.87, RMSE:0.44
Open(question)	R:0.91, RMSE:0.42	R:0.84, RMSE:0.52

Table 2 Correlation coefficients and RMSE in sentence generation task

	From subjective scores	From estimated scores
Closed	R:0.95, RMSE:0.42	R:0.85, RMSE:0.74
Open(speaker)	R:0.88, RMSE:0.62	R:0.77, RMSE:0.84
Open(question)	R:0.97, RMSE:0.46	R:0.78, RMSE:0.67

6 おわりに

本稿では，SCAT における文読み上げ問題，及び文生成問題を対象とした自動採点手法の改良を行った．個別要因や特徴量の再検討を行った結果，クローズドテストにおいては，両方の問題で高い推定精度が得られた．また，解答者オープンテスト及び設問オープンテストを行った結果，提案した総合点推定モデルは，解答者と設問の違いに対し頑健であることが示唆された．その一方で，各要因スコアの推定において，さらなる改良の必要があることが分かった．今後の課題としては，オープンテストの結果を詳しく分析することで，特徴量，及び各要因推定モデルのさらなる改良を行うことが挙げられる．

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] N. Okubo, *et al.*, “Automatic Scoring Method Considering Quality and Content of Speech for SCAT Japanese Speaking Test,” Proc. OCOCOSDA2012, pp.72–77, 2012.

[3] 藤代昇丈, 宮地功, “ブレンド型授業による英語の音読力と自由発話力に及ぼす効果,” 日本教育工学会論文誌, 32(4), pp.395–404, 2009.

[4] “Versant English Test,” http://www.versant-test.co.uk/pdf/ValidationReport.pdf.

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

[6] 加藤 他, “GOP と重回帰分析を用いたシャドーイング評価の高精度化,” 日本音響学会講演論文集, 3-11-18, pp.417–420, 2012.

[7] 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.

[8] 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.

[9] T. Kawahara, *et al.*, “Free software toolkit for Japanese large vocabulary continuous speech recognition,” Proc.ICSLP2000, pp.476–479, 2000.

[10] T. Kawahara, *et al.*, “Benchmark test for speech recognition using theCorpus of Spontaneous Japanese,” Proc.SSPR2003, pp.135–138, 2003.