日本語スピーキングテストにおける文章読み上げ問題の自動採点の検討*

☆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

これまでに，J-CAT（Japanese Computerized Adaptive Test）[1] と呼ばれる，日本語学習者の日本語能力を自動で評価するテストが運用されており，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，文生成，空所補充，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．

本研究では，受験者が指定文を読み上げる，文読み上げ問題を対象とする自動採点手法を提案する．文読み上げ問題における日本語教師による採点は，0〜4 点の 5 段階絶対評価尺度を用いた総合的な印象評価のみによって行われている．以下，これを総合点と呼ぶ．この評価の際には発音のような個別要因を意識していると考えられる．そこで，まず総合点に影響を及ぼす要因を主観評価実験を行うことで確認する．次に，各要因を推定するための特徴量を提案する．最後に，提案手法の有効性を評価する．

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，それを用いて各要因の値を推定する．次に，各要因の推定値を用いて総合点を推定する．なお，前者の推定対象となる要因は事前に決定する．後者の推定に用いる総合点推定モデルは，人間の主観値に基づいて決定する．

提案手法の特徴は，各要因の推定値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．

3 文読み上げ問題の採点に影響を及ぼす要因

3.1 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 [2, 3] 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），ラウドネス（X_5）を要因として設定した．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．

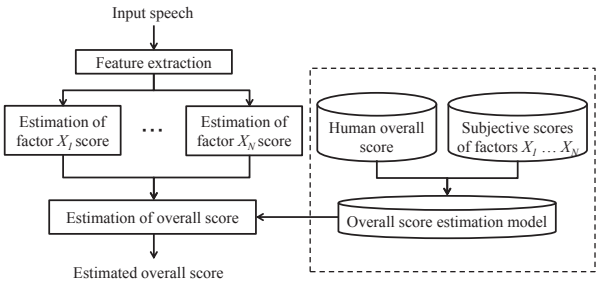

Fig. 1 Overview of the proposed method

3.2 主観評価実験

主観評価実験を行うことで各要因と総合点との関係を調査する．被験者は大学院生 5 名であり，防音室内でヘッドホンにより音声サンプルを受聴し，上述した全ての要因についてそれぞれ評価した．音声サンプルは，3 つの設問に対して留学生 20 名が発話した計 60 個である．各要因の評価尺度は，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する，0〜4 の 5 段階（非常に悪いから非常に良い）である．各要因の主観値は 5 名の評価点の平均である．また，総合点は日本語教師 3 名の平均である．

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することが分かった．そこで，日本語教師による総合点を目的変数，各要因を説明変数として，ステップワイズ法により変数選択を行った．その結果，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．

$$\text{総合点} = 0.27X_1 + 0.42X_4 + 1.30 \qquad (1)$$

これが，Fig. 1 の総合点推定モデルである．なお，設問の区別はせず，すべ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使用できる共通のモデルを構築するためである．

次に，3.2 節の実験で得た各要因の主観値を，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. 2 に示す．相関係数は 0.89，RMSE は 0.30 であり，高い精度で総合点を推定できることが分かる．また，設問毎の差もそれほどみられない．

4 提案手法の有効性の評価

4.1 各要因を推定するための特徴量

発音（X_1）の自動評価には，音声認識によって得られる音響尤度を利用した手法が提案されている [4].

---

*A study on automatic scoring method for reading task 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)

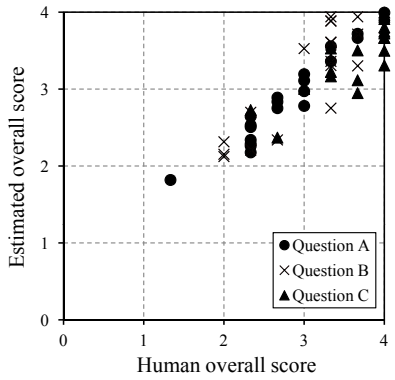

Fig. 2 Relationship between the subjective overall score and the overall score esttimated from the subjective scores of each factor

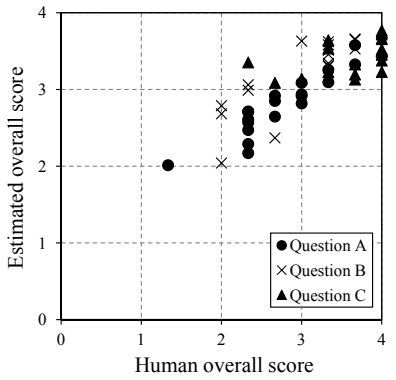

Fig. 3 Relationship between the subjective overall score and the overall score estimated from the estimated scores of each factor

しかし，音響尤度は個人差や周囲環境の影響を受けてしまう．それらの影響を除去するために，音響尤度に対し正規化を行うことは容易ではない．

そこで，発音（X_1）の主観値を推定するために，音声サンプルに対し，指定文（読み上げ対象の文）へのアライメントと連続音素認識結果へのアライメントを行い，両者の対応を比較することで算出される特徴量 x_{1a}，x_{1b} を提案する．

x_{1a}：発話区間において，両者の音素が一致しているフレームの割合
x_{1b}：発話区間において，両者の音素が一致しておらず，かつ両者の音響尤度の差が閾値以上であるフレームの割合

ここで，発話区間とは，指定文へのアライメントにおいて，発話の前後の無音部を除いたフレームの集合を表す．また，発話区間中であっても，無音が対応するフレームは特徴量 x_{1a}，x_{1b} の計算に使用しない．x_{1a} は一定レベルの発音をしている部分，x_{1b} は発音が特に悪い部分に着目している．予備実験により，x_{1b} の閾値を 2.25 に設定した．指定文を上手に発話している場合，x_{1a} の値は大きく，x_{1b} の値は小さくなると想定される

流暢さ（X_4）は，時間軸方向のスムーズさに着目している．3.2 節の音声サンプルを観察した結果，流暢さの主観値が低いデータにおいて，発話中の言い淀みが長く，その回数も多い，また，音節が間延びする傾向がみられた．そこで，流暢さ（X_4）の主観値を推定するために，音声サンプルに対し，指定文へのアライメントを行うことで算出される特徴量 x_{4a}，x_{4b} を提案する．

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，x_{4b} は音節長のばらつきに着目している．各特徴量の値が大きくなるほど，発話が流暢でなくなると想定される．

4.2 提案手法の有効性の評価

発音（X_1）の主観値，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．

$$X_1 = \quad 1.03x_{1a} - 8.90x_{1b} + 3.05 \qquad (2)$$

$$X_4 = -4.77x_{4a} - 2.39x_{4b} + 4.67 \qquad (3)$$

なお，音声認識器は Julius[5]，音響モデル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を用いた．

3.2 節の音声サンプルから算出した特長量を，式（2），（3）にそれぞれ代入し，各要因の主観値を推定した．その結果，発音（X_1）では，主観値と推定値との相関係数が 0.78，RMSE が 0.48 となった．流暢さ（X_4）では，相関係数が 0.80，RMSE が 0.60 となった．

次に，各要因の推定値を式（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 0.83，RMSE は 0.39 である．各要因の主観値を代入した場合（Fig. 2）に比べ，推定精度はやや劣るものの，それに近い精度が得られた．

5 おわりに

本稿では，SCAT における文読み上げ問題を対象として，総合点に影響を及ぼす要因を考慮した採点手法を提案し，その有効性を評価した．今後の課題として，各推定モデルの性能向上，及び従来手法との比較が挙げられる．

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（22242041）の助成を受けた．

参考文献

[1] J-CAT, http://www.j-cat.org/.
[2] 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), 395–404, 2009.
[3] “Versant English Test,” http://www.versanttest.co.uk/pdf/ValidationReport.pdf.
[4] 大田圭, 中川聖一, “日本人の英語文発話の発音評価法,” 日本音響学会春季研究発表会, pp.247-248, 2006.
[5] 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.