日本語スピーキングテストS-CATの自由発話問題における発話文の難易度を考慮した自動採点の検討*

☆盧昊，山畑勇人，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，日本の留学生総数は 13 万人を越え，また海外で日本語を学習している人は 400 万人近くに達し [1]，日本語能力測定に対する需要が益々高まっている．これまでに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)[2] が開発され，国内外で広く利用されるに至っている．現在のところ，J-CAT は聴解，語彙，文法，読解の能力を測定するセクションから構成され，発話能力の評価は行っていない．そこで我々は，J-CAT における自動採点形式のスピーキングテストとして S-CAT(Speaking section of J-CAT) の開発を進めている [3–7].

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，発話音声に加え発話内容の評価も必要になるので，自動採点も難しくなる．本稿では，発話内容に対する制約が最も少ない自由発話問題を対象とする．自由発話問題とは，あるテーマについて受験者の考えを発話したり，システムの提示する広告やグラフから内容を読み取って発話するものである．自由発話問題において，評定者による採点は，採点ガイドラインに基づき，流暢さ，表現力，内容，正確さの 4 種類の項目に対する 0〜4 点の 5 段階絶対評価尺度で行われている．各採点項目の定義を Table 1 に示す．自動採点を実現するためには，採点に寄与する特徴量を選定し，また特徴量から各項目のスコアを推定するモデルを構築する必要がある．

S-CAT の自由発話問題を対象とする自動採点手法として，西村らは，音素セグメンテーションによる音素区間情報を特徴量とした自動採点手法を提案した [4]．これは，異なる 2 つの音響モデルでパラレル動作する音声認識器 Julius[8] により，解答発話に対する認識結果の音素セグメンテーションを行う．その結果から各音素区間の開始時間と継続時間に着目し，開始時間の一致した数やその差の最大値など，計 14 次元の特徴量を抽出し，SVR で推定モデルを構築する．また小野らは，認識結果に基づく特徴量と音響特徴量を併用した自動採点手法を提案した [5]．受験者の発話音声と発話内容を総合的に考慮するため，OpenSmile[9] を用いて抽出した 383 次元の音響特徴量と，2 つの音声認識器 Julius と T3[10] を用いて抽出した認識結果に基づく 7 次元の特徴量を併用し，SVR で推定モデルを構築する．音響特徴量としては音声の対数パワーやゼロ交差率などを用いている．また，認識結果に基づく特徴量としては語彙多様性や解答に含まれる重要キーワード数などを用いている．

Table 1　Definition of scoring items in open-answer task.

流暢さ	発話がなめらかか
表現力	語彙・表現が豊富に使われているか
内容	タスクが遂行できているか 必要な情報を伝えているか
正確さ	文法が正しいか 語彙が適切に使われているか

上記 2 つの採点手法は比較的高い推定精度を達成している．しかし，問題点として自由発話問題の 4 つの採点項目に対して同じ特徴量を用いていることが挙げられる．受験者の発話能力の異なる側面を精密に捉え，4 つの採点項目それぞれの採点に寄与する特徴量を見出すことにより，更に推定精度を上げることができると考えられる．流暢さの評価は一般的に発話音声の音響的な特徴で行われる．一方，表現力，内容，正確さの評価は発話内容を考慮しなければならない．内容と正確さについては，僅かな認識誤りも許容できないので，その評価は難しい課題である．それに対して，表現力は主に語彙や表現の豊富さに注目するので，多少の認識誤りがある中でも，正しく認識された箇所 (信頼度の高い箇所) を用いることによりその評価は可能であると考えられる．

以上より，本稿では表現力の評価を対象とし，発話文の難易度を考慮した自動採点手法を提案する．そして，実験によりその有効性を検証する．

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

Table 2　Correlation coefficients between rater' Expression score and text statistics.

テキスト統計量	主観採点との相関係数
総文字数 (ひらがな)	0.73
総単語数	0.76
異なり単語数	0.75
総文節数	0.72
総文数	0.11
一文当りの文節数	0.58
一文当りの文字数	0.52
一文当りの単語数	0.57
構文木節点数の最大値	0.64
構文木節点数の平均値	0.57
構文木深さの最大値	0.62
構文木深さの平均値	0.53

2　提案手法

2.1　テキスト統計量の導入

日本語文章の難易度判定問題において，テキスト統計量の有効性が確認された [11]．テキスト統計量とは，文，文節，単語，文字など，文章の構成要素についての統計量である．文章の難易度は語彙や表現の豊かさを反映して決定されると考えられることから，同様に表現力をテキスト統計量により推定することを試みる．

各テキスト統計量と表現力主観採点の相関を調査した．S-CAT のプロトタイプを用いた模擬テストで収録した留学生の解答サンプル 70 個を選定し，これらのサンプルに対して行われた日本語教師 8 名による表現力採点の平均値を主観採点として使用した．音声認識器による誤認識の影響を排除するため，テキスト統計量は解答音声の人手による書き起こし文から抽出した．テキスト統計量を抽出するために，日本語形態素解析ツール MeCab と構文解析ツール CaboCha を用いた．

各テキスト統計量と表現力主観採点の相関係数を Table 2 に示す．Table 2 の結果から，総文数以降のテキスト統計量は主観採点との相関が比較的低いことが分かる．自由発話のような話し言葉の特徴として，文の長さは相対的に短く，かつ使われる文法は比較的シンプルである．それ故，これらのテキスト統計量は表現力採点にさほど寄与しないことが想定される．実験結果に基づき，以下に示す 4 種類のテキスト統計量を表現力採点の特徴量として選定する．

総文字数 (ひらがな)：
解答発話を平仮名で表示した時の文字数．

総単語数：
解答発話の中に出現した単語数．

異なり単語数：
解答発話の中に出現した単語の種類の数．

総文節数：
解答発話を文節に分解した時の文節数．

また，解答発話における全体の語数に対して異なる語がどの程度を占めるのかを評価するため，以下の語彙多様性を特徴量として使用する [5]．

語彙多様性：
$\frac{W_{unq}}{\sqrt{2W_{tot}}}$ で定義され，W_{unq} は異なり単語数，W_{tot} は総単語数である．

しかし，上記のテキスト統計量をそのまま表現力自動採点に利用するには問題が存在する．実際のテキスト統計量は音声認識器による認識結果から算出されるので，認識精度に大きく影響される．自由発話の大語彙連続音声認識では，認識対象語彙が膨大であり，混同しやすい単語の組が増加するため，誤認識が起きやすくなる．さらに，受験者の日本語発音の不自然さや，話し言葉で含まれるフィラー，言いよどみなども，認識精度に悪影響をもたらす．

そこで，誤認識の影響を低減するために，認識結果の単語信頼度をテキスト統計量の計算に導入する．音声認識器 Julius では，認識結果と共に，単語ごとの信頼度 [12] が出力される．信頼度の値は 0〜1 の範囲で，数値が高いほど，その認識結果の一位候補の単語に近い候補がないことを示す．総文節数以外のテキスト統計量を計算する際には，この信頼度を重み付けることにより，明瞭な発話内容を重視してカウントすることとする．Table 3 に示すように，認識結果の信頼度を導入することにより，各テキスト統計量は主観採点との相関が高くなることを確認した．

2.2　単語関連度の導入

前述のテキスト統計量は解答発話の量的な特徴を表現することができるが，具体的な発話内容を考慮していない．解答発話の表現力が高く評価されるのは，豊富な語彙を使っているだけでなく，設問の主旨に合う適切な語彙を使っているときである．このことを考慮するために，従来手法 [5] では，予め問題毎に重要キーワードを手動で設定し，それらが受験者の解答に含まれる数の多少により，受験者が使用した語彙の適切さを評価している．しかし，この手法では，設定

Table 3　Correlation coefficients between rater' Expression score and text statistics weighted by word-confidence score for LVCSR outputs.

	テキスト統計量	主観採点との相関係数
信頼度重み付けなし	総文字数 (ひらがな)	0.65
	総単語数	0.68
	異なり単語数	0.66
信頼度重み付けあり	総文字数 (ひらがな)	0.69
	総単語数	0.71
	異なり単語数	0.69

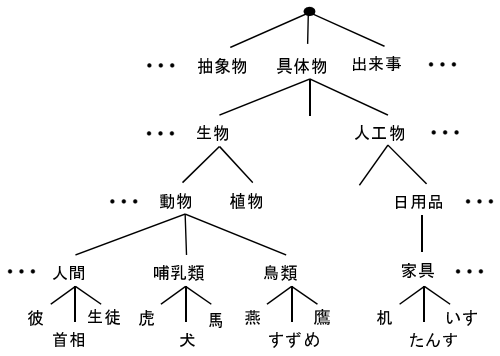

Fig. 1　Hierarichical structure of thesaurus.

したキーワード以外の語彙が評価されないという問題がある．そこで，シソーラスを用いて計算された単語関連度から受験者の解答発話に含まれる単語間の関連性を測定し，語彙の適切さを評価する手法を提案する．シソーラスとは，単語の上位・下位関係，同義・類義関係などによって単語を分類し，体系づけた辞書である．自然言語処理分野でよく使われる日本語のシソーラスとして日本語大シソーラス [13] がある．これは Fig.1 のような階層構造を持っており，シソーラス階層距離を計算することにより，単語間の関連性を評価することができる．2 つの単語のシソーラス階層距離は，式 (1) により計算される．

$$D_{i,j} = \frac{d_c \times 2}{d_i + d_j} \tag{1}$$

d_i，d_j，d_c は，単語 i，単語 j，共通上位ノード c の深さである．$D_{i,j}$ の値は 0〜1 の範囲で，数値が高いほど，2 つの単語が高い関連性を持っていることを示す．

単語関連度の具体的な算出手順について述べる．1 つの解答発話に対して形態素解析を行い，出現した全ての名詞を取り出す．その中から代名詞，数詞，接頭語，接尾語を取り除き，文章の表現に貢献する単語にのみ着目する．これらの単語を対象とし，単語ペア毎のシソーラス階層距離を計算し，式 (2) により定義される解答発話の単語関連度を求める．

$$Rel = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i}(D_{i,j} \cdot w_i \cdot w_j) \tag{2}$$

N は単語の総数であり，w_i，w_j は音声認識器 Julius が出力した単語 i，j の信頼度である．誤認識の影響を低減するために，シソーラス階層距離に認識結果の単語信頼度を重み付けている．

Table 4　Experiment conditions.

データセット	留学生 101 名 × 10 問 = 1010 解答 学習セット：81 名 × 10 問 = 810 解答 評価セット：20 名 × 10 問 = 200 解答
音声認識器	大語彙連続音声認識エンジン Julius
音響モデル	日本語話し言葉コーパスで学習した トライフォンモデルを留学生発話で適応
言語モデル	新聞記事及びウェブテキスト 90% +留学生発話書き起こし文 10%
形態素解析ツール	MeCab
構文解析ツール	CaboCha
推定モデル	SVR(RBF カーネル)

3　提案手法の有効性の検証

3.1　実験条件

提案手法の有効性を検証するために，従来手法 [4, 5] に準ずる条件で実験を行った．実験条件を Table 4 に示す．S-CAT のプロトタイプを用いた模擬テストにより収集した解答発話から留学生 101 名のデータを使用した．自由発話問題は 10 問あるので，総解答数は計 1,010 である．このうち，81 名分のデータ (解答数 810) を推定モデルの学習に，残りの 20 名分 (解答数 200) を評価に用いた．推定モデルの学習には，収集した解答発話に対する日本語教師 8 名の表現力採点の平均を使用した．推定モデルとしては，RBF カーネルを用いた SVR(Support Vector Regression) を採用した [5]．各テキスト統計量を抽出するために，形態素解析ツール MeCab，構文解析ツール CaboCha を用いた．SVR には，libsvm ツールキット [14] を用いた．

3.2　実験結果と考察

評定者による表現力主観スコアと提案手法により推定したスコアの関係を Fig.2 に示す．相関係数は 0.92，RMSE は 0.56 であり，高い推定精度が得られた．な

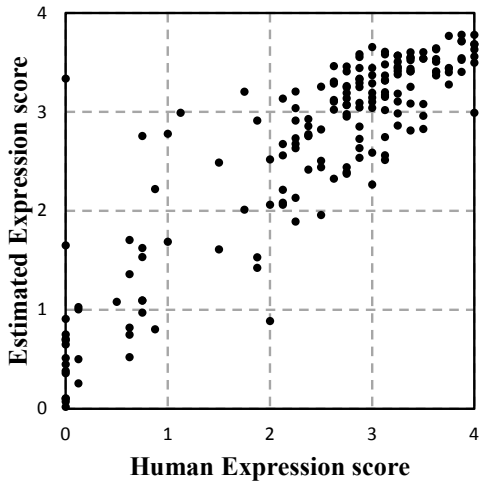

Fig. 2　Relationship between the human Expression score and the Expression score estimated by the proposed method.

お，若干条件が異なるものの，従来手法 [4] の相関係数は 0.74，従来手法 [5] の相関係数は 0.87 であった．しかし，Fig.2 に示すように，主観スコアの低いサンプルに対して高い点数で推定される傾向がある．これらのサンプルの特徴として，日本語で発話していない，発音の明瞭性が低い，設問をうまく理解できておらず解答として相応しくないなどが挙げられる．

4　おわりに

本稿では，S-CAT の自由発話問題において，採点項目の 1 つである表現力の自動採点手法を提案した．語彙の豊富さと適切さを評価するために，認識結果の信頼度を重み付けたテキスト統計量とシソーラスによる単語関連度を導入した．提案手法の有効性を評価した結果，表現力主観スコアとの相関係数は 0.92，RMSE は 0.56 であり，少数の特徴量で高い推定精度が得られた．

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 J-CAT プロジェクトのメンバーに深く感謝する．

参考文献

[1] 2012 年海外日本語教育機関調査結果, http://www.jpf.go.jp/j/about/press/dl/0927.pdf
[2] J-CAT, http://www.j-cat.org/
[3] 今井新悟，“Speaking Japanese Computerized Adaptive Test 開発の目的・方法と構成,” 日本行動計量学会大会第 41 回大会，SC1-2，2013.
[4] 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.
[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, pp. 1-4, 2012.
[6] N. Okubo, Y. Yamahata, 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. 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] 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] F. Eyben, M. Wollmer, B. Schuller, “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor,” Proc. ACM Multimedia, pp. 1459-1462, 2010.
[10] P. Dixon, D. Caseiro, T. Oonishi, S.Furui, “The Titech Large Vocabulary WFST Speech Recognition System,” IEEE ASRU, pp. 443-448, 2007.
[11] 山村毅，“日本語文章の難易度判定におけるテキスト統計量の有効性,” 電子情報通信学会論文誌 D，Vol. J96-D，No. 8，pp. 1952-1955，2013.
[12] 李晃伸，河原達也，鹿野清宏，“2 パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法,” 情報処理学会研究報告，2003-SLP-49-48，2003.
[13] 山口翼，“類義語検索大辞典 日本語大シソーラス,” 大修館書店，2003.
[14] C.C. Chang, C.J. Lin, “LIBSVM : a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.