日本語スピーキングテスト SJ-CAT における 低スコア解答発話の検出の検討*

☆小野友暉，山田武志，△今井新悟，牧野昭二（筑波大）

1 はじめに

総合的な日本語能力の評価のためには，リーディングテストとリスニングテストのみならず，発話能力をテストするスピーキングテストの実施が必要不可欠である．従来方式のスピーキングテストは，訓練された評定者が解答発話を聞いて採点するという主観評価方式であり，多大な費用と時間がかかってしまうという問題がある．従って，自動採点方式のスピーキングテストの開発が急務である．

SJ-CAT（Speaking Japanese Computerized Adaptive Test）[1] は，コンピュータによる自動採点方式の日本語スピーキングテストとして開発が進められている．現在 SJ-CAT のプロトタイプが稼働中であり，間もなく本運用を開始する予定である．SJ-CAT は項目応答理論 [2] に基づいたアダプティブテストである．これは視力検査のように，受験者が高いスコアで設問に解答した場合はより難易度の高い設問を，そうでない場合はより難易度の低い設問を出題する方式のテストである．これを繰り返すことによって，最終的に受験者の能力値に収束する．アダプティブテストでは全ての設問に解答する必要はないので，受験時間の短縮が可能になる．また，受験者の能力に近い設問が集中的に出題されるため，能力の判定をより高い精度で行うことができる．

SJ-CAT は，Table 1 に示す文読み上げ問題，選択肢読み上げ問題，文生成問題，自由発話問題の 4 種類の問題から構成されている．各解答発話に対して，複数の日本語教師が 0〜4 点（値が大きいほど高評価）の 5 段階絶対評価尺度でつけたスコアの平均値を採点アルゴリズムにより推定する．SJ-CAT の採点アルゴリズム [3]–[6] では，解答発話から抽出した発音やアクセント，発話内容などの特徴量をもとに，SVR（Support Vector Regression）を使ってスコアを推定する．特徴量抽出のためには受験者の解答発話を音声認識する必要があるが，この処理は受験者ごとに並列に行うので，処理負荷が大きくなってしまうという課題がある．この課題は計算リソースを増やすことにより解決できるが，実際には導入コスト・維持コストがかかるため難しい．よって，この課題を解決することが求められている．

Table 1 SJ-CAT における問題構成

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	選択候補から解答を読み上げる問題
文生成	設問に対し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た 30 秒程度の発話をする問題

現在，SJ-CAT では全ての解答発話に対して採点アルゴリズムでの処理が行われている．ここで，解答発話の中にはスコアが顕著に低い（1 点未満である）ものが一定の割合で含まれていることに注目する．SJ-CAT はアダプティブテストであるため，これらの解答発話を採点アルゴリズムにより厳密に採点する必要は必ずしもないと考えられる．次に出題する設問の難易度を決定するためには，スコアが顕著に低いということが分かれば十分である．従って本稿では，スコアが顕著に低い解答発話を，採点アルゴリズムでの処理を行う前に低計算量で検出する手法を提案する．提案手法で検出した低スコア解答発話は，採点アルゴリズムでの処理を行わないので，処理負荷を削減できると考えられる．

2 解答発話のスコア分布の調査

提案手法の効果を調査するために，SJ-CAT のプロトタイプの受験者（留学生）による解答発話のスコア分布を調査した．調査対象の解答発話を Table 2 に示す．解答発話計 8,000 サンプルの中から，スコアが 1 点未満のものを抽出したところ，その全体に占める割合は約 9%（718 サンプル）であった．そのうち，0 点のものは約 4%（315 サンプル），0 以上 1 点未満のものは約 5%（403 サンプル）であった．また問題の種類ごとに 1 点未満の解答発話の割合を調査したところ，文読み上げ問題は約 2%（31 サンプル），選択肢読み上げ問題は約 9%（156 サンプル），文生成問題は約 12%（422 サンプル），自由解答問題は約 10%（109 サンプル）となった．問題の種類ごとに 1 点未満の解答発話が全体に占める割合を比較すると，

*Detection of the answer utterances with a low score in SJ-CAT Japanese speaking test by Yuki ONO, Takeshi YAMADA, Shingo IMAI, Shoji MAKINO (University of Tsukuba)

Table 2　調査対象の解答発話

問題の種類	文読み上げ	選択肢読み上げ	文生成	自由発話
解答者	男性 31 名，女性 69 名，母語数 13			
解答発話数	1700 サンプル	1800 サンプル	3500 サンプル	1000 サンプル
設問数	17 問	18 問	35 問	10 問
評価者数 （日本語ネイティブの日本語教師）	6 名	6 名	6 名	8 名

文読み上げ問題は他の 3 つの種類の問題よりも少なかった．これは，文読み上げ問題は提示された文を読み上げるという問題であり，無解答の発話が比較的少なくなるからである．他の 3 つの種類の問題は受験者が適切な解答を考えて発話する必要があるため，文読み上げ問題と比べてスコアが 1 点未満の解答発話が多くなると考えられる．以上より，スコアが 1 点未満の解答発話を検出できれば，採点アルゴリズムで処理する解答発話数を 1 割程度削減できることが分かった．

なお，上記のプロトタイプの受験は監督者付きのコントロールされた環境下で行われたが，実際には監督者のいない自宅や教室など様々な状況で受験が行われる．このような状況下では，1 点未満の解答発話の割合はさらに増えると予想される．よって，計算リソースの削減効果はプロトタイプの場合より高くなると考えられる．

次章では，スコアが 1 点未満の解答発話を実際に検出するための手がかりと，それに対応する特徴量を探る．

3　低スコア解答発話の分析と特徴量

3.1　低スコア解答発話の分析

Table 2 の解答発話を実際に聴取し，低スコアとなった要因を調査することにより，自動検出するための手がかりを探った．ここで，低スコア解答発話とはスコアが 1 点未満の解答発話とし，それ以外のものは高スコア解答発話とする．

調査の結果，低スコアとなった要因は，主に以下の 4 つであることが分かった．

(1) ノイズのみが含まれている．
(2) フィラーのみの発話をしている．
(3) 間違った解答をしている．
(4) 解答と直接関係ない発話をしている．

要因 (1) では，受験者は全く発話しておらず，ノイズのみが含まれている．これは，受験者が設問を理解できなかったことや，適切な解答を見出せなかったことが原因であると考えられる．なお，ノイズには発話者の息を吐く音や環境音など様々なものがある．要因 (2) では,「えー」や「あー」といったフィラーのみが含まれている．その原因は要因 (1) と同じであると考えられる．要因 (3) は，例えば「曇りです」と解答すべきところを「晴れです」と解答してしまったものである．また,「よろしいですか」を「よろしいんでいますか」のように，日本語として誤った表現で解答してしまったものも含まれる．要因 (4) は,「すみません分かりません」とだけ発話しているものや，監督者と会話しているだけのものなどである．つまり設問に対する解答とは直接関係ない内容を発話しているものとなっている．

3.2　特徴量

上述した 4 つの要因に基づき，低スコア解答発話を検出するための特徴量を考察する．要因 (1) (2) の解答発話には，音声区間が存在しない，あるいは極端に短いという特徴がある．一方，要因 (3) (4) の解答発話は発話内容に関連するため，音声認識を行わないで正確に検出することは難しい．しかし，解答者が自信を持って発話していないことにより，音声区間が短くなるなどの特徴があることが分かった．以上より，低スコア解答発話を，音声区間・非音声区間に関する特徴量（以下，VAD 特徴量と呼ぶ）を使って検出することにする．音声区間・非音声区間の情報は，解答発話のパワー分析などによって低計算量で求めることができる．

次に，各要因に対応した VAD 特徴量を決定する．なお，本稿ではまず自由発話問題を対象とする．以下では，音声区間とは発話者が発話している区間，非音声区間とは音声区間の間の発話のない区間．発話区間とは発話を開始してから終了するまでの区間（すなわち最初の音声区間が開始してから最後の音声区間が終了するまでの区間）である．また，解答時間（解答のために与えられる時間）は設問によらず固定である．

まず，要因 (1) (2) について述べる．高スコア解答発話は，要因 (1) (2) の解答発話と比較して発話

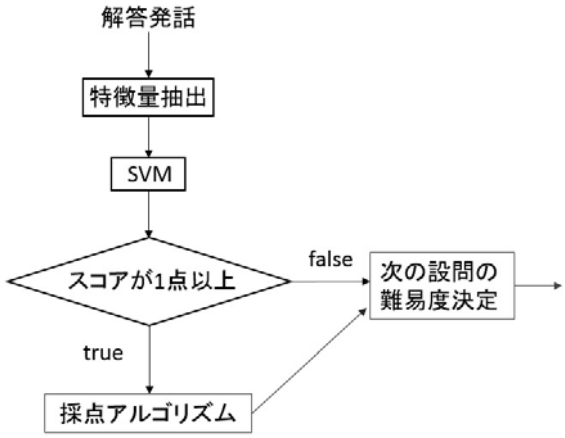

Fig. 1　提案手法の処理フロー

している時間が長く，また個々の音声区間が長い傾向にあった．よって,「音声区間の長さの合計」と「音声区間の長さの最大値」を特徴量として用いることにする．次に，要因（3）（4）について述べる．高スコア解答発話は基本的に流暢であり，要因（3）（4）の解答発話と比較して非音声区間の数が少なく，またその長さも短い傾向があった．よって,「非音声区間の出現回数」と「非音声区間の長さの変動係数（非音声区間長の標準偏差を非音声区間長の平均で割った値）」を特徴量として用いることにする．さらに，高スコア解答発話は，解答時間の最初から最後まで解答している傾向があったので,「発話区間の長さ」と「発話区間の終了時間」も特徴量として用いる．以上より，提案手法で用いる VAD 特徴量は以下の 6 次元となる．

- 音声区間の長さの合計
- 音声区間の長さの最大値
- 非音声区間の出現回数
- 非音声区間の長さの変動係数
- 発話区間の長さ
- 発話区間の終了時間

次章では，これらの VAD 特徴量を用いた提案手法について述べる．

4　提案手法

提案手法の処理フローを Fig. 1 に示す．まず，入力された解答発話から特徴量を抽出する．特徴量としては，前節で述べた 6 次元の VAD 特徴量を用いる．これらの特徴量を用いて，SVM（Support Vector Machine）による高スコアクラス，低スコアクラスの 2 クラスの識別を行う．解答発話が低スコアクラスと識別された場合，直接次の設問の難易度を決定する．解答発話が高スコアクラスと識別された場合，採点アルゴリズムによる処理が行われ，そのスコアによって次の設問の難易度を決定する．

次章では，実験により提案手法の有効性を検証する．

Table 3　実験条件

解答発話数	自由発話問題から 160 サンプル（学習用 80 サンプル，評価用 80 サンプル）
VAD の方法	人手による書き起こし文を用いる場合，およびパワーベースの VAD 手法 [7] を用いる場合の 2 通り
識別方法	SVM（libSVM[8] を使用）

5　提案手法の有効性の検証

5.1　実験条件

実験条件を Table 3 に示す．SVM の学習データ，評価データには自由発話問題の異なる 80 サンプルの解答発話をそれぞれ用いた．それぞれのデータには，要因（1）（2）の低スコア解答発話が 20 サンプル，要因（3）（4）の低スコア解答発話が 20 サンプル，高スコアの解答発話が 40 サンプル含まれる．また，高スコア解答発話は 1 から 4 までのスコアをとるが，スコアが均等に分布するように 40 サンプルを選択した．本実験では VAD 特徴量を求める際に，人手による時間情報付きの書き起こし文を用いる場合，パワーベースの VAD 手法 [7]（CENSREC-1-C のベースライン手法）を用いる場合を比較する．

5.2　実験結果と考察

提案手法により，正しく識別できた解答発話の割合を Table 4 に示す．Table 4 より，人手による書き起こし文を用いる場合，低スコア解答発話は 92.5%，高スコア解答発話は 88.5%の割合で正しく識別できていることが分かる．ここで，高スコア解答発話を低スコア解答発話として誤って識別するというエラーは，設問の難易度の決定に影響を及ぼすので，これを防ぐための特徴量を見出す必要がある．

一方，パワーベースの VAD 手法を用いる場合は，人手による書き起こし文を用いる場合よりも識別率が低下した．この理由としては，音声区間が正しく検出されなかったことや，ノイズを音声区間と誤って検出してしまったことが挙げられる．この問題については，VAD の精度を改善することにより対処する必要がある．

Table 4　実験結果

	提案手法（人手による書き起こし文）	提案手法（パワーベースの VAD 手法）
解答発話全体	90.0%（72/80）	68.8%（55/80）
低スコア解答発話	92.5%（37/40）	85.0%（34/40）
要因（1）（2）	100.0%（20/20）	80.0%（16/20）
要因（3）（4）	85.0%（17/20）	90.0%（18/20）
高スコア解答発話	88.5%（35/40）	52.5%（21/40）

6　おわりに

本稿では，スコアが顕著に低い解答発話を VAD 特徴量を用いて検出する手法を提案した．提案手法では，解答発話から抽出した VAD 特徴量を用いて，SVM による高スコアクラス，低スコアクラスの 2 クラスの識別を行う．VAD 特徴量については，受験者の解答発話を分析することにより決定した．提案手法を SJ-CAT の自由発話問題に適用した結果，人手による書き起こし文を用いて VAD 特徴量を求める場合は，90.0%の割合で正しく識別できていることが分かった．一方，パワーベースの VAD 手法を用いて VAD 特徴量を求める場合は識別率が低下した．今後，識別率を改善するために，新たな VAD 特徴量を導入し，また VAD 自体の精度を改善する必要がある．

謝辞　本研究は科研費（26244026）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] 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. O-COCOSDA2012, pp. 72–77, Dec. 2012.

[4] 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. APSIPA2014, WA1-1-3, pp. 1–5, Dec. 2014.

[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. APSIPA2012, Dec. 2012.

[6] 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.

[7] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, S. Nakamura, “CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments,” Acoustical Science and Technology, Vol. 30, No. 5, pp. 363–371, Sep. 2009.

[8] 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.