様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１６　日現在

機関番号：　１２１０２
研究種目：　基盤研究(A)（一般）
研究期間：　2014～2016
課題番号：　２６２４４０２６

研究課題名（和文）コンピュータ自動採点日本語スピーキングテストの実用化と妥当性の検証

研究課題名（英文）Practical application and validation of a computerized automatic scoring Japanese speaking test

研究代表者

今井　新悟（IMAI, Shingo）

筑波大学・人文社会系・教授

研究者番号：５０３４６５８２

交付決定額（研究期間全体）：（直接経費）　　29,000,000円

研究成果の概要（和文）：日本語学習者の日本語スピーキング能力の測定を自動で行う適応型テストシステムSJ-CAT（Speaking Japanese Computerized Adaptive Test）を開発し、インターネット上で公開した。SJ-CATは、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2セクションで構成され、音声の特徴量（韻律、音響尤度、スピーキングレートなど）やキーワードなどで評価する。項目応答理論の段階反応モデルで日本語スピーキング能力を点数化する。訓練された人間が評定を行うスピーキングテストとSJ-CATを比較し、実用化に十分な相関を得た。

研究成果の概要（英文）：We developed a testing system called SJ-CAT (Speaking Japanese Computerized Adaptive Test), which is accessible on the Internet. The test automatically measures the speaking ability of non-native speakers of Japanese language. SJ-CAT consists of four types of questions, i.e., reading a sentence, reading a correct sentence from three choices, making a sentence, and expressing one's opinion. The system evaluates one's speaking ability based on acoustic feature value (e.g. prosodic patterns, acoustic likelihood, and several kind of speaking rates) and keywords. Scores are calculated by means of a polytomous Item Response Model. Comparison between SJ-CAT and another speaking test, which is evaluated by trained human raters, showed high correlation, which indicates the practicality of SJ-CAT.

研究分野：　日本語教育

キーワード：　スピーキングテスト　CAT　日本語能力　自動採点

1．研究開始当初の背景

言語能力の 4 技能「読む、聞く、書く、話す」の測定において、「書く」「話す」の産出能力のテストの実施は難しい。中でも、スピーキングテスは、テスター（評定者）を養成し、確保し続けることが必要であり、多大な時間とコストがかかる。評定者による対面形式に代わるスピーキングの CBT（Computer based Test）がいくつか存在するが、コンピュータを介して音声を録音し、後で人が評定する仕組みが一般的である。対面で行うテストに比べて時間とコストの削減はある程度できるものの、テスターの養成と確保という根本的な課題の解決にはならない。

2．研究の目的

上記の課題を解決するには、人を介さない自動採点のシステムが必要である。英語においては英語においては Versant™ English Test[1]と Speech Rater™ [2]が開発されている。前者はテストとして利用できる唯一のものであるが、自由回答（Open Questions）形式が 2 問出題されるものの、それは自動採点の対象になっていない[1]。後者は正式なテストとしてリリースされておらず、開発が継続中である。本研究では自由回答形式も含んだ日本語で初のスピーキング自動採点テストを開発することを目的とした。

3．研究の方法

問題作成後、プレテストを実施して、音声回答サンプルを収集した。サンプル数は 80 問×192 人=15,360 であった。これを文字おこしし、各種情報のラベルを付けた。これを言語モデルの材料とした。人による採点基準を策定し、各音声サンプルを複数のトレーニングされた数教が採点した。その結果と自動採点の比較を行った。さらに、SJ-CAT と人が採点する他のスピーキングテストとの比較を行った。

4．研究成果

(1) 要旨

日本語学習者の日本語スピーキング能力を自動で測定する適応型テストシステム SJ-CAT (Speaking Japanese Computerized Adaptive Test) を開発し、インターネット上で公開した。SJ-CAT は、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 2 セクションで構成され、音声の特徴量（韻律、音響尤度、スピーキングレートなど）やキーワードなどで評価する。項目応答理論の段階反応モデルで日本語スピーキング能力を点数化する。訓練された人間が評定を行うスピーキングテストと SJ-CAT を比較し、実用化に十分な相関を得た。

(2)SJ-CAT の概要

SJ-CAT は非母語話者を対象とした日本語のスピーキングテストであり、インターネットでアクセスできる点、音声認識技術を用いている点、項目応答理論の多段階反応モデルを用いている点、アダプティブテストである点、自動採点である点が特色である。WEB 上で公開しており、ユーザー登録をすると受験できる。テストは 2 つのセクション、4 種類の問題で構成される。Section 1 は文読み上げ問題と選択肢読み上げ問題、Section 2 は文生成問題と自由発話問題からなる。

①文読み上げ問題項目：

本問題では、まず画面に「今日はいい天気ですね」のような文章が表示され、続いて日本語母語者による読み上げの例の音声が再生される。読み上げの例の音声の再生の後、受験者が読み上げる。文生成問題を読み上げる時間は問題により異なり、10 秒のものと 15 秒のものがある。文読み上げ問題では、発音・イントネーションが自然であるかなどを評価する。項目プールには文読み上げ問題は 14 問用意されている。アダプティブテストであるので、これらの問題項目からコンピュータが適宜選んで出題する。（以下の問題項目にいても同様。）

②選択肢読み上げ問題項目：

選択肢読み上げ問題では、まず動画や静止画と音声で何らかの場面が再生された後に、場面の内容に関する 3 つの選択肢が受験者に提示される。受験者は 5 秒間考えたのち、3 つの選択肢のうち、正しいものを選んで 10 秒または 15 秒で読み上げる。選択肢読み上げ問題では、受験者が提示された場面の状況を理解して正しい選択肢を選ぶことができたか、また、発音・イントネーションが自然であるかなどを評価する。選択肢読み上げ問題は 15 問用意されている。

③文生成問題項目：

文生成問題では、まず動画や静止画と音声で何らかの場面が再生された後に、場面の内容に関する質問の音声が流れる。受験者は 5 秒間考えたのちに、質問の内容に 10 秒以内で答える。例えば箱を開けている場面の映像が流れ、「何をしていますか？」という質問の音声が流れる。そして受験者は、「箱を開けています」のような文で回答する。文生成問題は、正しい文を発話しているか、また、発音・イントネーションが自然であるかなどを評価する。文生成問題は 34 問用意されている。

④自由発話問題：

自由発話問題では、まず「次の質問に 30 秒くらいで答えてください。宝くじで 1 億円当たったら、あなたは何をしますか？」というような音声が流れる。受験者は 5 秒間考えたのち、30 秒程度（録音時間制限は 40 秒）で答える。自由発話問題は 10 問用意されている。回答の「流暢さ」「正確さ」「内容」「表現力」を評価する。

---

[1] http://www.disc.co.jp/sp/versant/index.html

[2] https://www.ets.org/research/topics/as_nlp/speech/

テスト終了と同時に画面上に採点結果が、セクション 1 が 25 点満点、セクション 2 が 75 点満点、計 100 点満点で示される。この得点は後述する能力推定値を換算して算出している。テスト時間は、最初の録音音量の設定、各セクション 2 問ずつの練習問題を含めて、15 分から 20 分程度である。

(3) 採点の方法

自動採点の方法は、いくつかの提案を行ってきたが([2,3]など)、現在のバージョンでは音声認識器として、Julius と T3 デコーダを併用した以下の方法を採用している。

文読み上げ問題では、両認識器に正解文を含む多数の文を単語として登録しておき、第 1 段階として、単語（正解文）が認識されなければ 0 点とし、それ以外は次の 8 次元の音響特徴量を用いてサポートベクター回帰（SVR）を用いて採点する。

①発音を評価するため：Julius と T3 の単語音響尤度のフレーム平均

②モーラの自然さを評価するため：母語話者の回答（10 人分の平均）における各音素の発音タイミングと受験者の発話における各音素の発音タイミングの差である、発音タイミング距離

③アクセントおよびイントネーションを評価するため：母語話者のピッチパターンとの差である基本周波数パターン距離

④流暢さを評価するため：以下の 4 種類のスピーキングレート指標を使う。

(S1) 音素数／発話全体の長さ
(S2) 音素数／音声区間の長さ
(S3) 息継ぎ区間の長さ／発話全体の長さ
(S4) $\sum_{k}^{n}$ (S2 − 1/音素 $_k$の長さ)^2／音素数

S3 はポーズの長さであり、この値が小さいほど高評価となることを仮定している。S4 は音素内の発話の速さを示す。

選択肢読み上げ問題では、第 1 段階として、Julius と T3 のいずれでも選択肢にある文が認識されなければ 0 点、いずれかで不正解の選択肢が認識されれば 1 点、両方で不正解の選択肢が選択されれば、正解ではないが発音が良いと仮定して 2 点、いずれかで正解の選択肢が認識されれば、発音は悪いが正解を認めて 2 点、両方で正解の選択肢が認識されれば、第 2 段階として文読み上げ問題と同じ 8 次元の音響特徴量で採点する。

文生成問題では、スピーキングレート指標の S1、S2 で流暢さを評価するほか、Julius が認識したキーワード、T3 が認識したキーワード、キーワードスポッティングによるキーワードの有無を特徴量とした計 5 次元の SVR で採点する。キーワードはあらかじめサンプリングした文生成問題 35 問×191 人分の音声データを文字化し、問題ごとに高頻度に現れる語を抽出した。キーワードとのマッチングをし、発話文の内容を評価する。

自由回答問題でもキーワード一覧を作成し、第 1 段階で Julius の認識文と T3 の認識文のいずれにもキーワードがなければ 0 点とし、それ以外は第 2 段階でスピーキングレート指標の S1、S2 で流暢さを評価するほか、発話量と語彙多様性[4]の特徴量を含む 4 次元の SVR で採点する。

発話量：音素数／録音時間
語彙多様性：異なり語数／sqrt（2×延べ語数）

発話量が多いほど高く評価されるが、同じ発話量であっても、同じ表現を繰り返すよりも、豊富な語彙を用いる方が高く評価される。

(4) 能力推定と出題の方法

以上の手法で各問題の採点が行われ、0 から 4 までの連続値を得る。それを 0 点から 4 点の 5 段階の離散値に変換して、項目応答理論段階反応モデル[5]を使って、能力値を 1PL モデル、ベイズ EAP によって推定する。本テストはアダプティブであり、各問題の採点により、次に出題される問題が変わる。テストの開始時は最初の 2 問の回答による暫定能力値によって 3 問目に出題する問題を選択する。その後はベイズ推定による能力値の事後分散の期待値が最小となる問題を選択する[6]。終了条件（現行設定は、推定誤差が 0.5 未満になる、あるいは 1 セクションでの出題数が 12 問に達する）を満たすとテストが終了する。

(5) SJ-CAT の検証

(5-1) 自動採点と教師による採点の比較

自由回答問題を 5 人、その他の問題を 3 人の日本語教師が採点した。自動採点では、自由回答問題と文読み上げ問題各 81 人分の音声データ、選択肢読み上げ問題と文生成問題各 114 人分の音声データを学習データとして SVR で学習し、モデル構築のためのデータを提供した被験者とは別の被験者 20 名分の音声データを使用し、評価を行った。表 1 に各採点機能が行った採点結果と、日本語教員が行った評価の平均との積率相関係数と RMSE(Root mean square error) を示す。

表 1　自動採点と教師の採点の相関

	相関	RMSE
自由回答	0.91	0.63
選択肢読み上げ	0.89	0.64
文読み上げ	0.77	0.49
文生成	0.70	1.25

各問題を比較したとき、受験者に最も長い時間発話してもらう自由発話問題の採点機能が、最も相関係数の値が高くなった。受験者には長めに発話してもらったほうが、初心者と上級者の差が明確に表れ、システムでの評価が容易になっているものと考えられる。

文読み上げ問題にもう少し長めの文を読む問題を追加したり、文生成問題も少し長めの文で回答するような内容にしたりすることで、より高い精度で採点ができる可能性がある。また RMSE が最も大きいものが文生成問題となった。本システムでは、文生成問題の回答を評価する指標として、応答音声にキーワードが含まれるかどうかと、スピーキングレートしか見ていない。文法が正しいかどうかや、発話の終わり方が自然かどうかなどを評価できるようにすれば、さらに採点の精度を上げることができる可能性がある。実際に、文生成問題の採点機能にいくつかの特徴量を追加することで、採点精度を向上することができたという報告もなされている[7] が、現状では SJ-CAT で使用されている文生成問題の一部のみを対象とした実装と評価しかなされていないため、SJ-CAT のシステムへの組み込みは今後の課題とした。

問題の種類により精度に差はあるものの、構築した採点機能による受験者の応答音声の採点結果と日本語教員による採点結果との間に相関があることが確認できた。

文読み上げ問題という制限の強い問題よりも、自由度の高い自由回答問題の方の相関が高くなるというのは当初の予想に反していた。当初は、文の空所に適当な語を入れて読み上げる穴埋め形式の問題もあった。穴埋めに使われた語を評価したが、この形式は相関が低くなったため廃止した。このように回答が固定され、また短くなると相関が低くなる傾向を示した。短答・固定回答では、音声認識の成否が大きく採点結果に影響し、音声認識が成功しない場合に採点が不安定になると考えられる。一方、自由回答では音響特徴量がより強く採点に影響し、音声認識の失敗による影響が少ないため、相対的に安定した採点ができたと考えられる。

(5-2) SJ-CAT と JSST の比較

6 大学の日本語学習者に SJ-CAT と JSST（アルク社）[3]を受験してもらい、その結果を比較した。JSST は電話で受験するもので、出題数は 10 問で、回答が録音される。3 人の評定者が採点し、10 段階のレベルで初級から上級までを評価する。「～した時のことについて話してください」というような質問に対し、45 秒または 60 秒で回答する。

有効なデータを得た受験者は 178 人で、原則として同日に両テストを受験した。ただし JSST の録音ができていないため、後日 JSST を再受験した者（約 1 割）がいた。これを含め、両テストの受験の順番を変えて、カウンターバランスを取った。SJ-CAT を先に受験した人、JSST を先に受験した人、それぞれ 89 人ずつであった。

[3] http://www.alc.co.jp/jsst/

表 2　テスト結果の基本統計量　（n=178）

	最小値	最大値	mean	SD
SJ-CAT	4	94	61.7	18.0
JSST	1	9	5.9	1.5

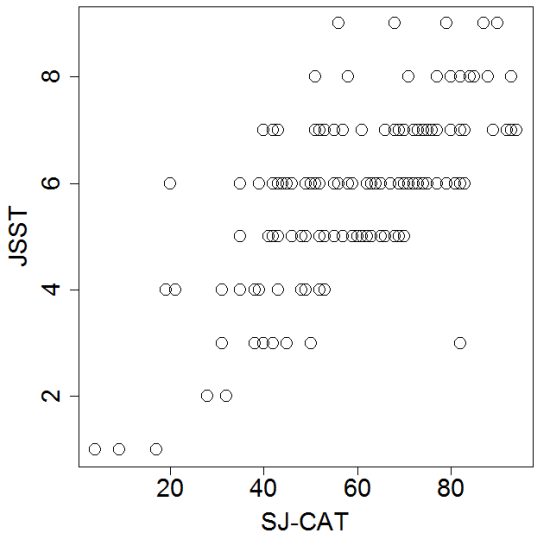

図 1　SJ-CAT と JSST の散布図

両テスト間に中程度の相関 r=0.651 (p<0.001) があった。図 1 によると、下位レベルの受験者が少なかったことが分かる。これにより相関が押し下げられた可能性がある。レベルが比較的分散している A 大学分 60 人分（SJ-CAT：mean=55。4、 SD=18.9, JSST：mean=5.1、 SD=1.6）では、r=0.812 (p<0.001) となり、強い相関が認められた。

また、JSST の中級のレベル 6，7 において SJ-CAT の分散が大きいことが分かる。人による評価でも中級レベルが最も難しいと思われる。個々の回答音声データを分析し、両テストの評価のずれの原因を探ることが今後の課題である。

(6)まとめ

本稿では、SJ-CAT の自動採点を評価するために、2 つの検証を実施した。その結果、問題タイプにより違いはあるものの、人と機械の採点に高い相関が認められたことから、個々の問題項目の自動採点の信頼性はある程度確保できたと言える。JSST との比較においても高い相関を示していることから、総合的評価において、併存妥当性を認めることができ、人による評価の代替になりうる可能性を示している。機械による評価は条件が同じ（音声データが同じ）であれば常に同じ結果が出るが、人の場合は条件が同じでも評価がぶれることがある。この点ではむしろ機械による評価が優位とも言えよう。

謝辞

開発にご協力いただいた大勢の大学院生、学習者、協力者に感謝します。

＊本報告は、「言語処理学会第 23 回年次大会」（2017 年 3 月）での研究発表予稿集および「第 232 回自然言語処理研究会」（2017 年 7 月）で発表予定の原稿に基づいて作成した。

＜引用文献＞

[1]Pearson Education, *Versant™ English Test: Test Description and Validation Summary*, 2011. http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf

[2]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, Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1-4, 2012.

[3]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, Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1 - 5, 2014,

[4]田島ますみ、深田淳、佐藤尚子、語彙多様性を表す指標の妥当性に関する研究－日本人大学生の書き言葉コーパスの場合－、中央学院大学社会システム研究所紀要 9(1)、pp. 51-62、2008 年

[5]F. Samejima, *Estimation of Latent Ability Using a Response Pattern of Graded Scores*, Psychometric Monograph, 17. Psychometric Society, 1969.

[6]菊地賢一、段階反応モデルに基づく汎用的適応型テストシステムの開発、日本行動計量学会大会発表論文抄録集 33, pp. 362-363, 2005 年

[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 年

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 8 件）

① 今井新悟 、コンピュータ適応型テスティング〔実践編〕、日本言語テスト学会誌、20 周年特別号、査読なし、招待論文、2016、pp. 186-190

② 中村洋一、Item writing and task design、日本言語テスト学会誌、20 周年特別号、査読なし. 招待論文 2016、pp. 69-72

③ 伊東祐郎 、ことばの能力を測るということ、 日本語学、査読なし、招待論文、33(12)、2014、pp.6-15

〔学会発表〕（計 27 件）

①小野友暉、山田武志、菊地賢一、今井新悟、牧野昭二、SJ-CAT における項目応答理論に基づく能力値推定の精度改善、日本音響学会春季研究発表会、2017 年 3 月 16 日、明治大学生田キャンパス、神奈川県、川崎市

②加藤拓、篠崎隆宏、半教師あり DNN を用いた日本語スピーキングテスト音声の認識、日本音響学会春季研究発表会、2017 年 3 月 15 日、明治大学生田キャンパス、神奈川県、川崎市

③今井新悟、赤木彌生、石塚賢吉、伊東祐郎、菊地賢一、篠崎隆宏、中園博美、中村洋一、西村隆一、本田明子、家根橋伸子、山田武志「自動採点スピーキングテスト SJ-CAT の能力推定の検証」言語処理学会第 23 回年次大会、2017 年 3 月 15 日、筑波大学、茨城県、つくば市

④今井新悟、SJ-CAT (Speaking Japanese Computerized Test)の開発、早稲田大学 CCDL 研究所第１回シンポジウム、外国教育研究におけるスピーキングとライティングの自動採点・評価、招待講演、2016 年 3 月 19 日、早稲田大学、東京都、新宿区

⑤伊東祐郎、これからの日本語教育と日本語力評価、ベオグラード大学日本語教育 40 周年記念プログラム、招待講演、2016 年 3 月 15 日、ベオグラード大学、ベオグラード、セルビア共和国

⑤ 伊東祐郎、日本語教育における評価、インドネシア日本語教育学会国際セミナー、招待講演、2015 年 8 月 21 日、ウダヤナ大学、バリ、インドネシア

⑥L. Hao, T. Yamada, S, Imai, T. Shinozaki, R. Nisimura, K. Ishizuka, S. Makino, and N. Kitawaki, Automatic scoring method for open answer task in the

SJ-CAT speaking test considering utterance difficulty level, Asia-Pacific Signal and Information Procedings of Association Annual Summit and Conference, Dec. 9th-12th, 2014, Sokha Angkor Resort Hotel, Siem Reap, Cambodia

〔図書〕（計 2 件）

①今井新悟 、外語教学与研究出版社、項目応 答理論とアダプティブテスト－J-CATによる評価－、徐敏民・近藤安月子主編、日语教学研究／日本学研究丛书、2016、pp.507-527

②今井新悟、くろしお出版、J-CAT Japanese Computerized Adaptive Test、李在鎬編、日本語教育のための言語テストハンドブック、2015、pp. 67-85

〔その他〕

ホームページ等

https://www.sj-cat.org/

６．研究組織

(1) 研究代表者

今井 新悟 (IMAI, Shingo)
筑波大学・人文社会系・教授
研究者番号：５０３４６５８２

(2) 研究分担者

伊東 祐郎 (ITO, Sukero)
東京外国語大学・大学院国際日本学研究院・教授
研究者番号：５０２４２２２７

中村 洋一 (NAKAMURA, Yoichi)
清泉女学院短期大学・国際コミュニケーション科・教授
研究者番号：７０３２６８０９

菊地 賢一 (KIKUCHI, Kenichi)
東邦大学・理学部・教授
研究者番号：５０２７０４２６

赤木 彌生 (AKAGI, Yayoi)
山口大学・大学教育機構留学生センター・准教授
研究者番号：３０３４６５８０

中園 博美 (NAKASONO, Hiromi)
島根大学・外国語教育センター・准教授
研究者番号：４０３１４６１１

本田 明子 (HONDA, Akiko)
立命館アジア太平洋大学・言語教育センター・准教授
研究者番号：８０３３１１３０

家根橋 伸子 (YANEHASI, Nobuko)
東亜大学・人間科学部・教授
研究者番号：８０６０９６５２

西村 竜一 (NISIMURA, Ryuichi)
和歌山大学・システム工学部・助教
研究者番号：００３７９６１１

山田 武志(YAMADA, Takeshi)
筑波大学・システム情報系・准教授
研究者番号：２０３１２８２９

篠崎 隆宏 (SHINOZAKI, Takahiro)
東京工業大学・大学院総合理工学研究科・准教授
研究者番号：８０４４７９０３

(3) 研究協力者

石塚賢吉 (ISHIZUKA, Kenkichi)