

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 5 月 23 日現在

機関番号：12102
 研究種目：基盤研究（A）
 研究期間：2010～2012
 課題番号：22242014
 研究課題名（和文） 音声認識技術を応用したコンピュータ自動採点日本語スピーキングテストの開発
 研究課題名（英文） Development of a Computer Automated Scoring Test of Spoken Japanese Using Speech Recognition Techniques
 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

研究成果の概要（和文）：日本語学習者のための日本語スピーキング能力をコンピュータ上で自動採点するテストシステムを開発した。インターネットを介して受験でき、時間と場所の制約を受けずに受験が可能である。音声認識技術を使い、受験者の発話から特徴量を抽出することにより、自動採点を実現している。項目応答理論を用い、受験者の能力に適合した難しさの問題を出題するアダプティブテストとなっており、少ない問題数で能力の判定ができる。

研究成果の概要（英文）：We have developed a computer speaking test for Japanese learners, which automatically evaluates speaking ability on computers. It will be accessible on the internet anytime, anywhere. The automatic scoring system is implemented through speech recognition techniques, which obtains acoustic features from the utterance. The system is a computerized adaptive test based on Item Response Theory, which makes it possible to evaluate the speaking ability with relatively fewer test items by adjusting to the ability of test takers and to the difficulty of the test items.

交付決定額

(金額単位：円)

	直接経費	間接経費	合計
2010年度	14,600,000	4,380,000	18,980,000
2011年度	14,800,000	4,440,000	19,240,000
2012年度	6,500,000	1,950,000	8,450,000
年度			
年度			
総計	35,900,000	10,770,000	46,670,000

研究分野：人文学

科研費の分科・細目：言語学・日本語教育

キーワード：スピーキングテスト、音声認識、自動採点、アダプティブテスト

1. 研究開始当初の背景

留学生 30 万人計画・経済連携協定など国の後押しもあり、日本国内で就学・就労する外国人が急増するなか、日本語能力を判定するためのテストは数が少なく、特にスピーキング能力を測定するテストがほとんどない。それは、コミュニケーション能力の測定としてのスピーキングテストの必要性は自明であるものの、スピーキングテストにおいては相

当数のテスター（評定者）が必要となること、および客観的評価による得点の不変性を保証することが難しいことによる。そのため、大規模テストにおいての実施は大変難しい。このジレンマを解決する方策として、スピーキングテストを人間の評定者ではなく、コンピュータを用いて即時・自動判定をするアイデアが古くからあった。ただし、現在、言語テストの分野では、「コンピュータによる」

「コンピュータを用いた」テストと銘打つテストがいくつかあるが、それらはコンピュータに音声録音し、それを人間が評定する仕組みであり、コンピュータによる自動判定ではない。そのため従来のテスト同様、多数の評定者と時間が必要となる。当然コストも高くなり、受験料も極めて高くなってしまふ。現在世界で唯一、Versant(TM) (旧 PhonePass) という電話を通して自動判定するシステムがあり、英語・スペイン語・アラビア語のテストがある。日本語については未だ実現に至っていない。このような状況のなか、日本語のコンピュータによる完全自動採点スピーキングテストは時代の要求が高く、早急な開発が望まれている。

2. 研究の目的

日本語学習者のための日本語スピーキング能力をコンピュータ上で自動採点するテストシステムを以下のようにして構築する。

(1) 項目応答理論を用い、不変的な評定尺度を実現する。(2) これまで開発してきた J-CAT (Japanese computerized adaptive test) のソフトウェアをスピーキングテスト用に改変して、能力レベルに適合した困難度レベルの問題が自動的に出題されるアダプティブ (適応型) テストとする。(3) 音声認識技術を用いて、音声の認識、採点を自動で行うシステムとする。音声認識エンジンには Julius および T3 を採用してカスタマイズする。

3. 研究の方法

(1) 問題アイテム作成

スピーキングテストの質問となる問題アイテムの作成基準を策定し、問題アイテムを作成する。音声・文字に加えて視覚情報 (グラフ、絵、動画など) を組み合わせる方法とその妥当性について研究する。

(2) 音声認識用モデル

音声認識に必要な言語モデルと音響モデルを構築または既成のモデルを使う場合の最適化の方法を研究する。音声認識エンジン Julius および T3 を併用、チューニングして非母語話者音声の音声認識精度を高める。

(3) 音声サンプル収集用テストシステム構築
問題アイテムを画面に表示し、マイクを使って回答させ、その音声を収集するシステムを構築する。

(4) 回答音声サンプル収集

学習者に受験してもらい、回答音声サンプルを収集する。

(5) 教師による採点

採点基準を作成し、回答音声サンプルを複数の教師が採点する。

(6) パラメータ推定

項目応答理論に基づき、識別力および困難度パラメータを推定する。

(7) 自動採点システムの構築

自動採点システムに使用できる指標を研究する。音響特徴量、正答とのパターンマッチングなどを想定する。

(8) 能力値の推定

項目応答理論に基づき、自動採点システムによる採点結果を使って、能力値を推定する。

(9) アダプティブシステム

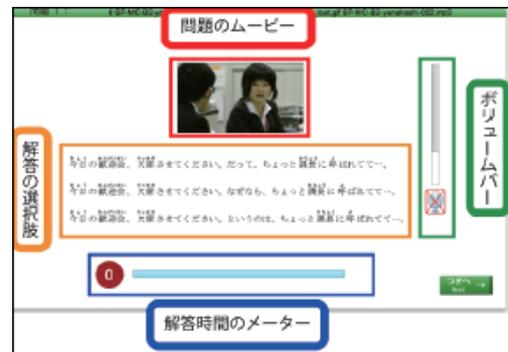
能力推定値と項目困難度から適切な項目を選択・出題する方法を研究する。

(10) 自動採点システムの評価

自動採点システムによる採点結果と教師による採点結果を付きあわせて、自動採点システムの精度を検証する。

4. 研究成果

日本語学習者のための日本語スピーキング能力をコンピュータ上で自動採点するテストシステムを構築した。



画面の例

まず、コンピュータテストというメディアに鑑み、適切な問題形式について検討し、回答の長さについては、単語レベル、文レベル、文章レベルとした。形式としては、空所補充、文読み上げ、選択肢読み上げ、文生成、自由回答の5形式とした。空所補充は文の一部を空所として提示して適切な語を入れる、語レベルで回答する形式である。文読み上げは、音素バランス文を主体とした文を画面上に提示し、それを読み上げる文レベルの形式である。選択肢読み上げは、音声、文、画像等の刺激提示の内容を理解し、それへの応答としてふさわしい文を3通りの文から選んで読み上げる文レベルで回答する形式である。文生成は、音声、文、画像等の刺激提示の内容を理解し、文レベルで回答する形式である。自由回答は指定されたトピックについて30秒程度で意見を言う、文章レベルの形式である。

これらの問題形式については、真正性を重視し、視覚情報 (イラスト、写真、動画など) を使用する方法とその妥当性について検討した。問題作成の課題として、単語・文の短答式であっても、正答パターンが多数になる

場合があること、さらにパターンマッチングさせるための回答例を網羅的に挙げきれない場合があることが明らかになった。解決策として、正答パターンが少なくなるようにアイテムを修正することとしたが、実際に音声サンプルを採録してみると、想定外の回答があり、回答例を網羅することの限界が明らかになった。このため、音声認識のキーワードスポッティングの手法を導入し、回答の完全一致ではなく、キーワードを含む部分一致の方法と音響特徴の組み合わせによる評価を行うこととした。キーワードは、問題作成者があらかじめ作成した模範回答から半自動で抽出した上、問題作成者による目視で修正して、キーワードおよび準キーワードとして選定した。

問題をインターネット上で配信し、回答を音声で入力し、オンラインで音声認識をして、認識結果を収集できるシステムを作製した。音声認識エンジンのJuliusおよびT3を組み込み、無音部分のタイミングにより発話終了を検知できるようにした。このシステムを使って、母語話者・非母語話者、男性・女性のバランスを考慮して回答サンプルを収集した。得られたデータの書き起こし・ラベル付けを行い、音響モデル構築のためのデータを作成した。

録音した回答を複数日本語教員が0~4点で採点した。採点にあたり、評価者がサンプルを採点し、合議しながら、評価基準を策定した。例えば、文読み上げ問題における評価基準は以下の通りである。

- 0点：発話なし。または、音声はあるが、意味不明。または回答と全く関係のない発話。
- 1点：例文の語を使って発話しているが、完結していない。または、例文の語を使って発話しているが、発音が悪くて、発話の意味が分からない。
- 2点：例文を読み上げているが、発音が非常によくない。
- 3点：例文を読み上げている。かつ、発音にやや難があるが、一般の日本人が少し努力すればすべて理解できる。
- 4点：例文を読み上げている。かつ、発音に母語の影響がわずかに残るが全くコミュニケーションの妨げにならず、発音・イントネーションが自然である。

自由回答問題については、流暢さ、正確さ、内容、表現力について各4点、計16点満点とした。

これら基準に基づき、90問×192人分、計17280個の音声回答データについて採点した。(無回答で0点となったものも含む。)採点は日本語教師が1データ当たり6人、自由回答問題については8人がそれぞれ採点した。全採点をまとめたあと、分散を求め、特に分散が異常に大きいもの、明らかに外れ値があ

るものなど、採点ミス、点数記述ミスなどが疑われるものについては、採点者に差し戻して確認を求め、採点の信頼性を高めた。

システムによる評価(スコア推定器)のための特徴量として、キーワードマッチングの他、単語音響尤度、発話タイミング距離、スピーキングレート、発話量、基本周波数パターン距離、語彙多様性、等などを検討し、問題形式ごとに最適な組み合わせを採択した。

単語音響尤度は音響尤度のフレーム平均(音響尤度/フレーム数、1フレーム=10msec)を用いた。文全体がはっきりと発音されているかどうかを評価していると仮定している。

発話タイミング距離は、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回答における各音素の発音タイミングの差である。発話の自然さを評価していると仮定している。

スピーキングレートには数種類あり、それぞれ、発話区間長(間の無音区間を含む発話開始点から発話終了点までの長さ)に対する音素数、音声区間(有音区間)長に対する音素数、発話区間長に対する無音区間(息継ぎや次の発話を考えている時間)の長さ、および音素数に対する音節の時間長を用いた。スピードレートは流暢さを評価すると仮定している。

発話量は、録音時間に対する音素数である。制限時間内でどれくらいたくさん話しているかを評価している。たくさん話せる受験者の方が能力が高いと仮定している。

基本周波数パターン距離は日本語話者と受験者の回答の、平均を揃えた基本周波数パターン間の距離である。韻律の類似度を測っている。

語彙多様性は、異なり語数/SQRT(2×延べ語数)で求める。能力が高い受験者の方が語彙が多様になると仮定している。

他に、編集距離についても検討した。これは2種類の音声認識器による認識結果の距離である。発音がよければ認識結果の一致度が上がるため、発音の評価になっていると仮定している。検討の結果、編集距離については使用しないこととなった。

特徴量からの評価スコアの推定方法としては、重回帰モデルを用いた手法が提案されている。しかし、重回帰モデルは過学習に弱く使用する特徴量を少数に限定する必要がある。そこで過学習に強いと期待されるサポートベクター回帰モデルの使用を検討した。自由発話問題を対象とした事前実験により、使用する特徴量数が少ない場合サポートベクター回帰は重回帰と同程度の性能を示し、特徴量を増やすに従い、より高い性能が得られることが確認された。本システムでは、全

ての問題形式においてサポートベクター回帰モデルを使用することとした。前述の通り、録音した回答を複数の日本語教師が採点した。スコア推定器では、その平均を目的変数としている。

さらに以下の各方法を取り入れることにより、システムによる評価の精度を向上させた。有音部と無音部を識別する Voice Activity Detection を組み込んだ。音声認識エンジン Julius とともに T3 を併用し、それぞれの認識結果を相互に参照するようにした。集積した音声データおよび音声書き起こしデータを基に音響モデルおよび言語モデルを作成した。言語モデルについてウェブテキストモデル、『日本語話し言葉コーパス(CSJ)』配布モデル、音声書き起こしモデルとその融合モデルを音響モデルを使って評価した。問題タイプごとにシステムによる評価項目(=特徴量)を最適化させた。

事前実験により、システムの特徴量による評価と人による評価の平均を比較して、0.7 から 0.9 弱の実用化レベルの相関があることを確認した。

項目応答理論多値モデル(Graded Response Model)を用いてテストスコアを算出するアルゴリズムを組み込んだテストシステムを構築した。予め、困難度のパラメータを付けた問題項目をプールしておく。受験者の能力の推定を逐次行いながら、その能力値に近い困難度の問題項目を項目プールから選び出して出題するアダプティブテストとすることにより、出題数を抑えつつ、能力判定の精度を確保できるようにした。

システムを使用してパイロットテストを行って点数推定のチューニングを行なうこと、実運用に耐えうるサーバー等の構築を行った後に一般公開すること、および、問題項目プールの継続的な拡充が今後の課題である。

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 2 件)

- ① 赤木彌生・今井新悟、J-CAT 日本語テストの運用 2010、大学教育、査読有、第 8 号、2011、78-87

[学会発表] (計 22 件)

- ① Naoko Okubo, Yuto Yamahata, Takeshi Yamada, Shingo Imai, Kenkichi Ishizuka, Takahiro Shinozaki, Ryuichi Nisimura, Shoji Makino and Nobuhiko Kitawaki, "Automatic scoring method considering quality and content of speech for SCAT Japanese speaking test," Proc. Oriental COCOSDA,

2012 年 12 月 10 日, University of Macau, 中国

- ② 今井新悟・石塚賢吉・山田武志・篠崎隆宏・西村竜一・酒井たか子・菊地賢一、コンピュータによる日本語スピーキング自動採点システムー日本語教師による採点との相関による検証ー、日本言語テスト学会全国研究大会、2012 年 10 月 27 日、専修大学
- ③ 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、2012 年秋季研究発表会、2012 年 9 月 21 日、信州大学
- ④ 今井新悟・酒井たか子・伊東祐郎・赤木彌生・中園博美・本田朋子・家根橋伸子・山田武志・西村竜一・篠崎隆宏・中村洋一・菊地賢一・石塚賢吉、スピーキング自動採点テストのための問題アイテム作成、日本語教育国際研究大会、2012 年 8 月 18 日、名古屋大学
- ⑤ Youichi Nakamura, Shingo Imai, Sukero Ito, Kenichi Kikuchi, Yayoi Akagi, Hiromi Nakasono, Akiko Honda, Ryuichi Nishimura, Takako Sakai, Takahiro Shinozaki and Takeshi Yamada, A Report on Japanese Computerized Adaptive Test, Korean English Language Association, 2012 年 5 月 26 日、高麗大学、韓国
- ⑥ Yutaka Ono, Misuzu Otake, Takahiro Shinozaki, Ryuichi Nisimura, Takeshi Yamada, Kenkichi Ishizuka, Yasuo Horiuchi, Shingo Kuroiwa and Shingo Imai, Open Answer Scoring for S-CAT Automated Speaking Test System Using Support Vector Regression, APSIPA 2012, 2012 年 4 月 4 日, Loews Hollywood Hotel、米国
- ⑦ 今井新悟、J-CAT (Japanese Computerized Adaptive Test in a New Era of Testing)、第 7 回日本語教育研究発表会、2010 年 10 月 9 日、University of Malaya、マレーシア
- ⑧ 菊地賢一・今井新悟・中村洋一・平村健勝、「日本語コンピュータ適応型テスト J-CAT について」、日本テスト学会第 8 回大会、2010 年 8 月 31 日、120-121、多摩大学

[図書] (計1件)

- ① 今井新悟・赤木彌生・中園博美、ココ出版、J-CATオフィシャルガイド：コンピュータによる自動採点日本語テスト、2012年、163ページ

[その他]

ホームページ

<http://www.intersc.tsukuba.ac.jp/~imai/j-cat-project/reference.html>

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・留学生日本語教育センター・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

酒井 たか子 (SAKAI TAKAKO)

筑波大学・人文社会系・教授

研究者番号：40215588

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

菊地 賢一 (KIKUCHI KENICHI)

東邦大学・理学部・准教授

研究者番号：50270426

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

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

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

篠崎 隆宏 (SHINOZAKI TAKAHIRO)

千葉大学・大学院融合科学研究科・助教

研究者番号：80447903

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

家根橋 伸子 (YANEHASHI NOBUKO)

東亜大学・人間科学部・准教授

研究者番号：80609652

(3)研究協力者

石塚賢吉 (ISHIZUKA KENKICHI)

筑波大学・大学院システム情報工学研究科・大学院生

ファムターンソン (PHAM THANH SON)

筑波大学・大学院システム情報工学研究科・大学院生