自動採点スピーキングテスト：SJ-CAT
AUTOMATIC SCORING SPEAKING TEST: SJ-CAT

今井新悟　（筑波大学）
Shingo Imai, University of Tsukuba

概要：SJ-CAT（Speaking Japanese Computerized Test)は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である。著者らが開発した J-CAT ®（Japanese Computerized Adaptive Test）のプログラムを基に作られたスピーキングテストである。

テストの構成は、文読み上げ問題、選択肢読み上げ問題、文生成問題、自由回答問題の４形式である。文読み上げ問題は、画面上に提示された文を読み上げる。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。文生成は、音声、文、画像等の刺激提示の内容を理解し、文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 45 秒以内で意見を言うものである。

キーワード：スピーキングテスト、アダプティブテスト、自動採点、日本語能力

1 開発の目的

言語能力はコミュニケーション能力として、スピーキング能力がまず問われるべきであり、それこそがテストの中心となるべきではある。しかし、受容能力のテストに対して産出能力のテストは実施は格段に難しい。例えば、大規模テストでスピーキングテストを実施するには、多くのテスター（評定者）を養成し、確保し続けることが必要である。そのためには多大な時間とコストがかかる。この実施可能性に配慮して、コンピュータを用いて即時・自動判定をするシステムを開発することとした。「コンピュータによる」「コンピュータを用いた」スピーキングテストとされるテストがいくつか存在しているが、その多くはコンピュータを介して音声を録音し、それを人間が評定する仕組みである。受験会場に集まって対面で行うテストに比べて移動の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。これを解決するには、人を介さない自動採点のシステムが必要である。現在、Versant(TM)（旧 PhonePass）と ETS の Speech Rater がある。SJ-CAT は日本語では初のシステムであり、Versant では実現されていない自由発話タイプの評価もできる。また、Speech Rater では、採点に重回帰を使っており、点数の根拠の説明はしやすい。SJ-CAT でも当初は重回帰を使っていたが、その後、より採点の精度が高くなるサポートベクター回帰に変えている。

2 開発の方法

開発に係る要点は以下の通りである。

（１）問題アイテム作成基準を策定し、問題アイテムを作った。

（2）コンピュータ上でのプレテスト用ソフトを開発してプレテストを実施した。
（3）採点基準を策定し、複数の教師による採点を行った。1 回答あたり、6 人、自由回答問題については 8 人がそれぞれ採点した。
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なく、発音・イントネーションが自然である。
（4）採点結果を見て、明らかに外れ値があるものなど、単純な記述ミスなどが疑われるものについては、採点者に確認をしながら、採点データを確定させた。
（5）J-CAT を基に、スピーキングテスト用にインターフェース、管理者画面を変更した。
（6）J-CAT で採用した項目反応理論の２値の採点アルゴリズムを多値モデルに拡張した。
（7）音声処理と採点アルゴリズムを最適化して、自動採点の処理速度及び精度を向上させた。
（8）キーワードによる内容の特徴量、語彙多様性による言語的特徴量、及び音響特徴量による採点アルゴリズムを考案した。
（9）項目応答理論の部分採点モデルを用いて、多段階の採点を実現した。
（１０）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）テストを実現した。
（１１）自動採点の点数と教師による採点の相関を目安に、自動採点のチューニングを行った。
（１２）自動採点の点数と教師による採点の相関が理論上の上限近くまでに達し、自動採点の実用化にめどをつけた。
（１３）Deep Learning を使って、音声認識精度を向上させた。
（１４）HTML5 に対応させ、インターフェースを改良した。

3 音声認識・採点システム

音声認識エンジン Julius を使用している。有音部と無音部を識別する Voice Activity Detection を組み込んだ。無音区間の検出により、無駄な認識・採点を省くこと、また、言語モデルの内容とサイズをテストに合わせて調整することにより、認識・採点速度を速めた。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を取り入れることにした。キーワードは、模範回答および受験者の回答を文字起こししたものから抽出して、選定した。キーワードのマッチングによる評価は発話の内容を評価していると仮定している。その他の採点の指標は以下の通りである。

語彙多様性：「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。

すべての問題形式において、音声認識（ディクテーション）のみに頼った場合、それが失敗した場合の採点に与える影響が大きすぎるため、以下のような音響特徴量も使うことにより、頑健性を高めた。

単語音響尤度：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：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（印象）を評価していると仮定している。

スピーキングレート：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：録音時間に対する音素数である。制限時間内の発話量を評価している。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：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。発話の自然さ（印象）を測っていると仮定している。

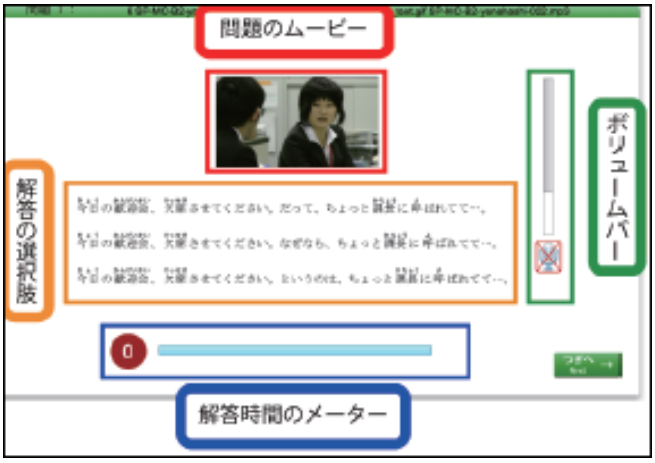

図 1 ：画面の例

4 システムの評価

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、問題形式ごとに 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。

今後は、問題項目の妥当性の検証を重ね、追加問題を作成し、問題項目プール（問題項目のデータベース）を拡充するとともに、公開して、有効性・頑健性を検証する。

謝辞：
本研究開発は発表者以外の多くの分担者・協力者の協力によって実施されている。
現在の中心メンバーは以下の通り（50 音順）である。

赤木彌生（山口大学）・石塚賢吉（ドワンゴ）・伊東祐郎（東京外国語大学）・菊地賢一（東邦大学）・篠崎隆宏（東京工業大学）・田藤千弘（和歌山大学院生）・中園博美（島根大学）・西村竜一（和歌山大学）・本田明子（立命館アジア太平洋大学）・家根橋伸子（東亜大学）・山田武志（筑波大学）・盧昊（筑波大学院生）

本研究は以下の補助金を受けている。
2014-2016 年度　科学研究費補助金　基盤研究（A）　26244026　「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」

2010-2012 年度　科学研究費補助金　基盤研究（A）22242014　「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」