Vol.2017-NL-232 No.9
2017/7/20

日本語スピーキングテストSJ-CATの開発

石塚 賢吉[1,a)]　菊地 賢一[2]　篠崎 隆宏[3]　西村 竜一[4]　山田 武志[5]　今井 新悟[5]

概要：本論文では、日本語学習者の日本語スピーキング能力の測定をインターネット上で実施できる適応型テストシステム SJ-CAT の開発について述べる。SJ-CAT のテスト問題は、日本語教員が作成した (1) 文読み上げ問題、(2) 選択肢読み上げ問題、(3) 文生成問題、(4) 自由発話問題の 4 種類の問題で構成されており、音声の特徴量（キーワード、韻律、音響尤度、スピーキングレートなど）と得点との対応関係を表現するモデルを使用して採点を行う。そして、項目応答理論に基づく段階反応モデルで受験者の総合的な日本語スピーキング能力を測定する。本論文では、訓練された人間が評定を行う日本語スピーキングテストの結果と SJ-CAT の結果を比較する被験者実験を行う。被験者実験の結果、両者にある程度の相関があり、SJ-CAT により受験者の日本語スピーキング能力を測定できることを確認した。

キーワード：スピーキングテスト、項目応答理論、アダプティブテスト、音声認識

Development of SJ-CAT（Speaking Japanese Computerized Adaptive Test）

Kenkichi Ishizuka[1,a)]　Kenichi Kikuchi[2]　Takahiro Shinozaki[3]　Ryuichi Nishimura[4]
Takeshi Yamada[5]　Shingo Imai[5]

1. はじめに

グローバル化に伴い、非母語話者の言語能力を測定するテストの需要が高まっているが、言語能力の 4 つの技能「読む、聞く、書く、話す」のうち、特に「話す能力」を測定するテストの実施には大きなコストがかかる。非母語話者のスピーキング能力を適切に測定するためには、テスターを養成し、確保し続ける必要があるためである。そこで、音声認識技術を用いて非母語話者のスピーキング能力を測定するシステムに関する研究が行われている [1][2]。ただし、これまでの研究は、文章読み上げや文生成など個別のタスクの能力を測定する研究例が多く、総合的なスピーキング能力を測定するシステムを開発することを目指す研究例はまだ少ないといえる。そこで現在、日本語スピーキングテストである SJ-CAT（Speaking Japanese Computerized Adaptive Test）の開発が行われている。なお、音声認識を用いた唯一の英語スピーキング能力テストとして Versant[3] が運用されているが、長めの自由発話を扱う自由発話問題を受験者の能力測定に利用していない。本論文では、30 秒程度の発話を扱う自由発話問題も含むテストで日本語の総合的なスピーキング能力の測定を行う SJ-CAT の開発について述べる。そして、人間が評定を行う日本語スピーキングテストの結果と SJ-CAT の結果を比較する被験者実験を行い、SJ-CAT により受験者の日本語スピーキング能力を測定できることを確認する。

[1] 株式会社ドワンゴ
DWANGO Co., Ltd.
[2] 東邦大学
Toho University
[3] 東京工業大学
Tokyo Institute of Technology
[4] 和歌山大学
Wakayama University
[5] 筑波大学
University of Tsukuba
[a)] ken57.jcom@gmail.com

2. SJ-CAT の概要

2.1 システムの構成

SJ-CAT はインターネットから受験可能な、自動採点スピーキングテストとなっている。また SJ-CAT はコンピューターアダプティブテストとなっており、項目応答理論に基づいて逐次的に推定される受験者の能力に応じて、問題プールから問題を選択しながら受験者に出題する。SJ-CAT のシステムは、テストクライアントと採点・能力推定サーバ、問題プールで構成されている。テストクライアントは HTML5 で実装されており、受験者が使用している Web ブラウザ上で問題を提示し、受験者の応答音声を録音して音声を採点・能力推定サーバへ送信する。そしてテスト完了後に、受験者に点数を提示する。採点・能力推定サーバでは、テストクライアントから受験者の応答音声を受け取り、音声認識器の Julius[4] と T^3[5] を用いて採点し、項目応答理論に基づき受験者の日本語スピーキング能力を推定する。そして、能力に応じた問題を問題プールから選んで出題する。

2.2 テストの構成

SJ-CAT のテストは、文読み上げ問題と選択肢読み上げ問題が出題されるセクション 1 と、文生成問題と自由発話問題が出題されるセクション 2 の 2 つのセクションで構成される。SJ-CAT のテストの点数は、セクション 1 が 25 点満点、セクション 2 が 75 点満点、合計 100 点満点で示される。

2.2.1 セクション 1 で出題される問題

(1) 文読み上げ問題

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。文読み上げ問題は 14 問用意されている。

(2) 選択肢読み上げ問題

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の例を 図 1 に示す。この例では、2 人が話をしている場面の映像が流され、「2 人は何をしていますか？」という質問の音声が流れた後、「ご飯を食べています。」「話をしています。」「本を読んでいます」という選択肢が提示されている。

図 1 選択肢読み上げ問題の例

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。

2.2.2 セクション 2 で出題される問題

(1) 文生成問題

本問題では、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。

例えば文生成問題では、箱を開けている場面の映像が流され、「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。

(2) 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。自由発話問題では、回答の「流暢さ」「正確さ」「内容」「表現力」を評価する。

3. SJ-CAT による受験者の能力推定

SJ-CAT は、項目応答理論に基づく段階反応モデル [6] を用いたアダプティブテストとなっている。本章では、SJ-CAT のテストによる受験者の日本語スピーキング能力推定の流れについて説明する。まず、SJ-CAT で用意している各問題には、下記のようなモデルが設定されている。

$$p^*_{j,k}(\theta) = \begin{cases} \dfrac{1}{1+e^{-1.7a_j(\theta-b_{j,k})}} & (k=1,2,3,4) \\ 1 & (k=0) \end{cases} \quad (1)$$

これは能力値 θ の受験者に問題 j を出題したとき、k 点以上のスコアとなる確率を表している。SJ-CAT では、各問題が 0 点から 4 点の 5 段階のスコアで採点される。点数の定義は問題の種類ごとに異なり、文読み上げ問題では表 1 のような基準で評価される。なお、a_j は問題の識別力、$b_{j,k}$ は問題の困難度を表すパラメータであり、事前に行う被験者実験で収集した回答パターンをもとに、周辺最尤法を用いて求める。a_j と $b_{j,k}$ の算出については 5 章で述べる。そして、能力値 θ の受験者が設問 j においてスコア k を獲得する確率を下記の式で表す。

$$p_{j,k}(\theta) = \begin{cases} p^*_{j,k}(\theta) - p^*_{j,k+1}(\theta) & (k=0,1,2,3) \\ p^*_{j,k}(\theta) & (k=4) \end{cases} \quad (2)$$

受験者の能力値推定はベイズ推定により行われる。なお、以降で説明する能力推定の計算は、θ の値の範囲を区間 $[-4,4]$ とし、この範囲を 20 の区間に分割する 21 の離散点での近似計算として実装されている [6] 。受験者がテストを開始して、結果を確認するまでの流れを図 2 に示す。

3.1 固定問題の出題

セクション開始直後は能力値が不明であるため、最初の 2 問は固定の問題を出題する。最初の 2 問は、セクション中の易しい問題と難しい問題を 1 問ずつ選んでいる。受験者が出題された問題に答えると、4 章で説明する採点機能で応答音声が採点される。

3.2 事前分布の初期値の算出

固定問題 2 問の応答音声の採点が完了すると、スコアを 0

表 1　文読み上げ問題の採点基準

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。

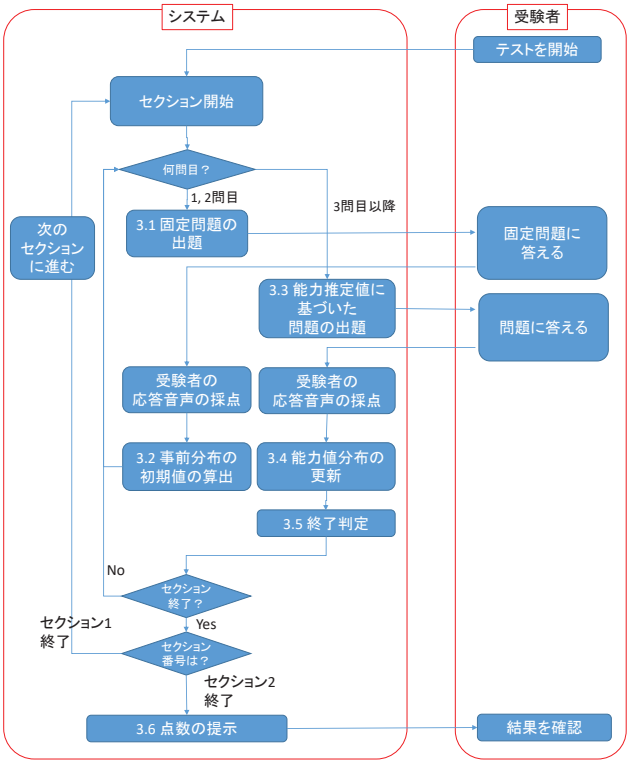

図 2　SJ-CAT のテストの流れ

点から 4 点の離散値に変換し、ベイズ推定による能力値推定に使用する事前分布の初期値を求める。ここでは、J-CAT (Japanese Computerized Adaptive Test)[7] で使用している事前分布の初期値を決めるアルゴリズムを多値のスコアに対応できるように拡張する。SJ-CAT では、固定問題の数を N、固定問題のスコアの合計を $S(S=0,1,...,8)$ としたとき、以降で説明する能力値推定の計算で使用する事前分布の初期値 $h_0(\theta)$ を、下記の μ を中心とし、分散を 1 とした正規分布としている。

$$\mu = \begin{cases} -1 & (S=0) \\ \log\left(\dfrac{S}{4N-S}\right) & (0<S<4N) \\ 1 & (S=4N) \end{cases} \quad (3)$$

3.3 能力値分布に基づいた問題の出題

3 問目以降の問題は、受験者の能力値分布（事後分布）の分散の期待値が最も小さくなるように問題プールから問題を選択する。問題プール中のまだ出題されていない問題の集合を J とし、問題 $j \in J$ を出題したときの受験者の能力値分布（事後分布）の分散の期待値 e_j は下記の式で求める。

$$e_j = \sum_{u=0}^{4} \sigma^2_{j,u} q_{j,u} \quad (4)$$

なお、$q_{j,u}$ は問題 j を受験者に出題したときに、スコアが $u(u=0,1,...,4)$ となる確率で、固定問題を除く $n-1$

問目の問題に回答した時点での、受験者の能力値分布（事前分布）を $h_{n-1}(\theta)$ としたとき、下記の式で求める。

$$q_{j,u} = \int_{-\infty}^{\infty} h_{n-1}(\theta) p_{j,u}(\theta) d\theta \quad (5)$$

$p_{j,u}$ は式 (2) で求める。また、$\sigma_{j,u}^2$ は問題 j を出題し、スコアが u となるときの受験者の能力値分布（事後分布）の分散である。問題 $j \in J$ のうち、e_j が最も小さくなる問題 j を出題する。受験者が出題された問題 j に答えると、4 章で説明する各問題の採点機能で応答音声が採点される。

3.4 能力値分布の更新

3 問目以降は、採点機能により算出された問題 j のスコアを、0 点から 4 点の離散値に変換したスコア u に基づいて能力値分布の更新を行う。なお、固定問題を除く $n-1$ 問目の問題に回答した時点での、受験者の能力値分布（事前分布）を $h_{n-1}(\theta)$ としたとき、n 問目に回答した受験者の能力値分布（事後分布）$h_n(\theta)$ は次式により求める。

$$h_n(\theta) = \frac{h_{n-1}(\theta) p_{j,u}(\theta)}{\int_{-\infty}^{\infty} h_{n-1}(\theta) p_{j,u}(\theta) d\theta} \quad (6)$$

3.5 終了判定

$h_n(\theta)$ の標準偏差が閾値未満になるか、受験者が答えた問題の数 n が閾値を超えるとセクション終了とし、次のセクションに進む。$h_n(\theta)$ の標準偏差が閾値以上であれば 3.3 節に戻り、次の問題を出題する。

3.6 点数の提示

セクション 2 が終了すると、セクションごとの能力値分布（事後分布）$h_n(\theta)$ の平均値 μ_n に従って、テスト結果を受験者に提示する。なお SJ-CAT では、μ_n の値をそのまま一般の受験者やテスト利用者に提示しても意味が伝わりにくいと考え、J-CAT[7] と同様の式でセクションの合計点が 100 点になるように換算して提示する。

$$\text{セクション 1 の点数} = (\text{セクション 1 の} \mu_n \times 15 + 50) \times 0.25 \quad (7)$$

$$\text{セクション 2 の点数} = (\text{セクション 2 の} \mu_n \times 15 + 50) \times 0.75 \quad (8)$$

なお、式 (7)(8) の μ_n にかかる係数の 15 は、J-CAT の受験者の回答データを使用したシミュレーションを行い、点数の範囲が 0 点から 100 点にできるだけ満遍なく収まるように調整して決められた値である。また、各セクションの点数が定められた範囲からはみ出した場合は、端数を切り捨てる。

4. 採点機能

本章では、音声認識器を使用して受験者の応答音声の採点を行う機能について説明する。SJ-CAT では、受験者による応答音声を受け取ると、まず CENSREC-1[8] の Voice Activity Detection (以下 VAD) により音声区間以外を除去する。そして、4 つの問題の種類ごとに用意された採点機能を使用して、応答音声の採点を行う。

採点機能は、音声認識器により受験者が適切な回答を行っているかを判定し、次のフェーズに進むかどうかを決める「音声認識フェーズ」と、音声特徴量をもとに応答音声の流暢さや自然さを評価する「音声特徴量による採点フェーズ」の 2 つのフェーズで構成されている。SJ-CAT では、音声認識器として Julius と T^3 を採用している。なお、Julius では音素アライメントを行い、音声特徴量による採点フェーズで使用する。音声認識器で使用している音響モデルは、日本語話し言葉コーパス（CSJ）[9] で学習した音響モデルを 5 章で説明する被験者実験により収集した日本語学習者の音声のサンプルで音響モデル適応したものを使用している。

4.1 文読み上げ問題採点機能

4.1.1 音声認識フェーズ

文読み上げ問題採点機能では、受験者の応答音声を受け取ると、まず Julius と T^3 で音声認識を行う。Julius と T^3 には、文読み上げ問題で使用している 14 種の読み上げ文を、それぞれ一つの「単語」として登録しており、孤立単語として認識する。なお、Julius と T^3 では N-best 解を出力する。Julius と T^3 の認識結果について、「C: 第一候補が正解文」「I: 第二候補以下だが N-best 解中に正解文がある」「W:N-best 解中に正解文がない」の条件に応じて、表 2 の採点を行う。

表 2 認識結果に基づく文読み上げ問題の採点

		T^3 の認識結果		
		C	I	W
Julius の認識結果	C	4.1.2 節へ	4.1.2 節へ	1 点
	I	4.1.2 節へ	0 点	0 点
	W	1 点	0 点	0 点

4.1.2 音声特徴量による採点フェーズ

本フェーズでは、受験者の応答音声から抽出した下記の 8 次元の音声特徴量をもとに、Support Vector Regression (以下 SVR) で 0.0 から 4.0 点の点数で採点する。

音素発話長差分距離

受験者の読み上げ音声の音素ごとの長さをもとに、発話の日本語としての自然さを評価する特徴量である。本距離

を評価するために、まず日本語母語話者 10 人による読み上げ音声をサンプルとして用意する。読み上げ文に含まれる音素の番号を $n(n = 1, ..., N)$ とし、日本語母語話者 j が読み上げた音声の音素ごとの長さを $l_{j,n}$、受験者 t の読み上げ音声の各音素の長さを $l_{t,n}$ としたとき、j と t の音素発話長差分距離 $d_l(j,t)$ を下記の式で表す。

$$d_l(j,t) = avg(|(l_{j,n+1} - l_{j,n}) - (l_{t,n+1} - l_{t,n})|) \quad (9)$$

日本語母語話者 10 人のサンプル音声と受験者 t の音声の $d(j,t)$ を比較し、最も小さいものを受験者の回答音声の音素発話長差分距離の特徴量として採用する。音素発話長差分距離が大きいほど、日本語として違和感のある読み上げになっているものと仮定している。

基本周波数パターン差分距離

受験者の読み上げ音声の韻律の自然さを評価する特徴量で、フレーム単位での日本語母語話者と受験者の応答音声の基本周波数パターンの回帰直線の傾きを比較するものである [1]。SJ-CAT では、応答音声の基本周波数パターンを抽出するために get_f0s[10] を使用する。なお、受験者の応答音声と、日本語話者の応答音声の発話長が違う場合は、音素ごとに線形伸縮してフレーム間の対応付けを行う。$i(i = 1, ..., I)$ 番目のフレームにおける日本語母語話者 j の音声の基本周波数を $f_j(i)$、受験者 t の音声の基本周波数を $f_t(i)$ としたとき、基本周波数パターン差分距離 $d_f(j,t)$ は下記の式で表される。

$$d_f(j,t) = avg(|(f_j(i+1) - f_j(i)) - (f_t(i+1) - f_t(i))|) \quad (10)$$

SJ-CAT では、日本語母語 10 名と受験者 t の音声の $d_j(j,t)$ を比較し、最も小さいものを受験者の音声の基本周波数パターン差分距離の特徴量として採用する。

スピーキングレート

受験者の発話の流暢さを評価するために、4 種類のスピーキングレート指標を算出する。まず、録音した音声から CENSREC-1 の VAD と Julius の音素アライメント情報をもとに音声区間を検出する。そして、発話開始から終了までの区間のうち、音声区間と認識されなかった部分を息継ぎ区間とする。スピーキングレート計算のための音声区間、息継ぎ区間の概念図を図 3 に示す。

そして、発話に含まれる音素数を発話全体の長さで割ったものをスピーキングレート S_1、音声区間の長さで割ったものをスピーキングレート S_2 とする。

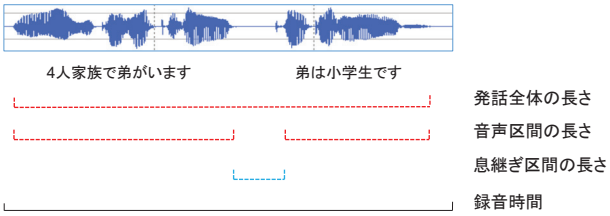

図 3　音声区間・息継ぎ区間の概念図

$$S_1 = \frac{\text{音素数}}{\text{発話全体の長さ}} \quad (11)$$

$$S_2 = \frac{\text{音素数}}{\text{音声区間の長さ}} \quad (12)$$

息継ぎ区間の長さを発話全体の長さで割ったものをスピーキングレート S_3 とする。S_3 の値が大きいほどたどたどしい発話となっており、評価が下がることを想定した指標である。

$$S_3 = \frac{\text{息継ぎ区間の長さ}}{\text{発話全体の長さ}} \quad (13)$$

また、音素を発話する時の速さの分散である S_4 もスピーキングレートの指標として採用している。

$$S_4 = \frac{1}{\text{音素数}} \sum_{k}^{n} \left(S_2 - \frac{1}{\text{音素}_k\text{の長さ}} \right)^2 \quad (14)$$

単語音響尤度のフレーム平均

発音を評価するために、Julius と T^3 で認識された正解文の単語音響尤度を、音声区間の長さで割ったものを、単語音響尤度のフレーム平均の特徴量として用いている。なお、文読み上げ問題採点機能では読み上げ文を孤立単語として認識するため、ここでの単語音響尤度は文全体の音響尤度である。

ここまでで説明した文読み上げ問題で使用する 8 つの特徴量を表 3 に示す。

4.2　選択肢読み上げ問題採点機能

4.2.1　音声認識フェーズ

文読み上げ問題と同様に、選択肢読み上げ問題で使用している 45(15 問× 3 択) 文を、Julius と T^3 にそれぞれ一つの「単語」として登録し、孤立単語として音声認識する。Julius と T^3 の認識結果の「C: 第一候補が正解文」「I: 第二

表 3　文読み上げ問題の特徴量

1	音素発話長差分距離
2	基本周波数パターン差分距離
3	スピーキングレート S_1
4	スピーキングレート S_2
5	スピーキングレート S_3
6	スピーキングレート S_4
7	Julius の単語音響尤度のフレーム平均
8	T^3 の単語音響尤度のフレーム平均

候補以下だが N-best 解中に正解文がある」「W:N-best 解中に正解文がない」に応じて、表 4 の条件で採点を行う。Julius と T^3 の認識結果のどちらかの第一候補が正解文で、もう一方の N-best 解中に正解文が含まれている場合、文読み上げ問題と同様の音声特徴量で採点する。

表 4　認識結果に基づく選択肢読み上げ問題の採点

		T^3 の認識結果		
		C	I	W
Julius の認識結果	C	4.1.2 節と同様	4.1.2 節と同様	2 点
	I	4.1.2 節と同様	4.2.2 節へ	4.2.2 節へ
	W	2 点	4.2.2 節へ	4.2.2 節へ

4.2.2　正解の選択肢が認識されなかった場合の採点

Julius と T^3 の認識結果の第一候補が両方とも選択肢中の文でない場合、0 点とする。認識結果のどちらか一方の第一候補が選択肢中の不正解の文である場合、1 点とする。両方で同じ不正解の文が認識された場合、正解ではないが発音は良いとみなして 2 点を与える。

4.3　文生成問題採点機能

4.3.1　音声認識フェーズ

文生成問題採点機能では、受験者の応答音声を受け取ると、下記 3 つの音声認識器を使用して音声認識を行う。

(1) ディクテーション用の言語モデルを使用する Julius
(2) ディクテーション用の言語モデルを使用する T^3
(3) キーフレーズスポッティングモデルを使用する T^3

(1),(2) で使用している言語モデルは、5 章で説明する被験者実験で収集した文生成問題への応答音声の書き起こし文書から生成した言語モデルであり、文生成問題全体で共通のモデルを使用している。また、この応答音声の書き起こし文書から、問題ごとに正解となる文の中に高頻度で現れるフレーズを抽出し、キーフレーズのリストを作成する。キーフレーズスポッティングモデルは、文生成問題のキーフレーズのリストの両端をガベージモデルで囲んだものとなっており、問題ごとに生成している。

(1), (2), (3) のモデルのいずれかの認識結果に文生成問題のキーフレーズが含まれていた場合は、特徴量による採点にすすむ。含まれていない場合は 0 点とする。

4.3.2　音声特徴量による採点フェーズ

本フェーズでは、受験者の応答音声から抽出する表 5 の 5 次元の音声特徴量をもとに、SVR で 0.0 から 4.0 の点数を算出する。

表 5　文生成問題で使用する特徴量

1	Julius によるキーフレーズ抽出の成否　(1 or 0)
2	T^3 によるキーフレーズ抽出の成否　(1 or 0)
3	キーフレーズスポッティングによるキーフレーズ抽出の成否 (1 or 0)
4	スピーキングレート S_1（4.1.2 節と同様）
5	スピーキングレート S_2（4.1.2 節と同様）

4.4　自由発話問題採点機能

4.4.1　音声認識フェーズ

自由発話問題採点機能では、受験者の応答音声を受け取ると、Julius と T^3 を使用して音声認識を行う。Julius と T^3 では、5 章で説明する被験者実験で収集した自由発話問題への応答音声の書き起こし文書から生成した言語モデルと様々なコーパスをもとに生成した汎用的な言語モデルを融合したものを使用している。

また、文生成問題と同様に、応答音声の書き起こし文書から問題ごとに高頻度で現れる語を抽出し、問題内容に関連するキーワードのリストを作成する。Julius と T^3 の認識結果のどちらにもキーワードが含まれていなければ、0 点とする。キーワードが含まれていれば、音声特徴量による採点フェーズに進む。

4.4.2　音声特徴量による採点フェーズ

本フェーズでは、受験者の応答音声から抽出する表 6 の 4 次元の音声特徴量をもとに、SVR で 0.0 から 4.0 の点数で算出する。語彙多様性 [11] とは、認識文に含まれる単語の異なり語数と述べ語数をもとに下記の式で計算したものである。

$$\text{語彙多様性} = \frac{\text{異なり語数}}{\sqrt{2 \times \text{述べ語数}}} \tag{15}$$

単位時間あたりの発話量は認識文に含まれる音素数を録音時間で割ったものである。

$$\text{単位時間あたりの発話量} = \frac{\text{音素数}}{\text{録音時間}} \tag{16}$$

表 6　自由発話問題で使用する特徴量

1	語彙多様性
2	単位時間あたりの発話量
3	スピーキングレート S_1（4.1.2 節と同様）
4	スピーキングレート S_2（4.1.2 節と同様）

5. 被験者実験

5.1 採点機能の検証

SJ-CAT の各採点機能が、受験者の応答音声を適切に採点できるかどうかを確かめるために被験者実験を行う。

5.1.1 採点機能のモデル構築

(1) 文読み上げ問題と選択肢読み上げ問題

文読み上げ問題全問について 81 名、選択肢読み上げ問題全問について 114 名の被験者が答えた音声データを日本語教員 3 名が 0 点から 4 点の 5 段階で採点する。そして、音声データから求めた特徴量と、日本語教員 3 名の採点結果の平均との対応関係をもとに SVR のモデルを構築する。

(2) 文生成問題

文生成問題全問について、被験者 114 名が答えた音声データを日本語教員 3 名が 0 点から 4 点の 5 段階で採点する。なお、ここで、収集した音声データの書き起こし文書をもとに言語モデルを構築して、文生成問題採点機能の Julius と T^3 のディクテーション用の言語モデルとして使用する。また、問題ごとに正解となる文に高頻度で現れる語を抽出し、キーフレーズのリストを作成する。そして、音声データから求めた特徴量と、日本語教員 3 名の採点結果の平均との対応関係をもとに SVR のモデルを構築する。

(3) 自由発話問題

自由発話問題全問について、被験者 81 名が答えた音声データを日本語教員 5 名が回答の「流暢さ」「正確さ」「内容」「表現力」を 0 点から 4 点の 5 段階で採点する。なお、ここで収集した音声データの書き起こし文書をもとに言語モデルを構築し、様々なコーパスをもとに構築した汎用的な言語モデルと 0.9999:0.0001 で融合したものを自由発話問題採点機能の Julius と T^3 の言語モデルとして使用する。また、問題ごとに高頻度で現れる語を抽出し、キーワードのリストを作成する。そして、音声データから求めた特徴量と、日本語教員 5 名の「流暢さ」「正確さ」「内容」「表現力」の採点結果の平均との対応関係をもとに SVR のモデルを構築する。

また、CSJ で学習した音響モデルを、ここで収集した音声データをもとに音響モデル適応したものを SJ-CAT の音響モデルとして使用している。

5.1.2 日本語教員による採点結果との比較検証

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、評価を行う。表 7 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ピアソンの積率相関係数 r[12] と RMSE(Root mean square error) を示す。

結果を見ると、「文読み上げ問題」、「選択肢読み上げ問題」、「自由発話問題」については、強い相関がみとめられる。また、「文生成問題」については、ある程度の相関がみとめられる。

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、最も相関係数の値が低く、RMSE が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 [13] 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、まずは本論文で説明した採点機能を使用する。

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ため、次節ではシステム全体として受験者の総合的なスピーキング能力を測定することができるかを確かめる。

表 7　採点機能による採点と日本語教員による採点との間の相関係数 r と RMSE

	相関係数 r	RMSE
文読み上げ問題	0.77	0.49
選択肢読み上げ問題	0.89	0.64
文生成問題	0.70	1.25
自由発話問題	0.91	0.63

5.2 能力推定の検証

SJ-CAT が受験者の総合的なスピーキング能力を適切に測定することができるかを確かめるため、人間のテスターが評

定を行う日本語スピーキングテストである JSST(Japanese Standard Speaking Test)[14] との結果を比較する被験者実験を行う。

5.2.1 各問題のパラメータ推定

5.1.1 節で収集した、各問題について被験者が獲得した点数のパターンをもとに、項目応答理論に基づく能力推定で使用する各問題の識別力と困難度のパラメータ値を求める。SJ-CAT では、EasyEstGRM[15] を使用して周辺最尤法によりそれぞれのパラメータ値の推定を行う。

5.2.2 JSST の結果との比較検証

6 大学の日本語学習者に SJ-CAT と JSST を受験してもらい、その結果の比較を行う。JSST は電話で受験するテストとなっており、まず受験者の回答が録音される。録音された音声は 3 人のテスターにより採点され、1 から 10 の 10 段階のレベルで評価される。JSST の問題は、「〜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 秒から 60 秒で回答するようなものとなっており、10 問出題される。

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 を再受験したものが一割ほどいた。なお、89 人が SJ-CAT を先に受験し、残りの受験者は JSST を先に受けている。

表 8 に SJ-CAT のセクションごとの点数と最終的な点数、JSST の点数の最大値、最小値、平均、標準偏差を示す。また受験者の SJ-CAT のセクションごとの結果とセクションの合計の結果、JSST の結果とのピアソンの積率相関係数 r を表 9 に示す。表 9 の SJ-CAT(セクション合計) と JSST の結果の相関係数を見ると、テスト間にある程度の相関がみとめられる。また、セクション 1 の結果とセクション 2 の結果を合計した結果が、セクション単体の相関係数より高くなっており、性質の違う問題が含まれる 2 つのセクションの結果が寄与し、受験者の総合的なスピーキング能力をより正しく測定することができていると考えられる。

6 大学の受験者全員分の SJ-CAT の結果と JSST の結果の散布図を図 4 に示す。散布図を見ると、今回の被験者実験では下位レベルの受験者が少なく中・上級者に偏っていることがわかる。また、受験者のレベルの偏りが少なかった A 大学の 60 人分の受験者の散布図を図 5 に示し、受験者のレベルが上級者に偏っている B,C 大学の 40 人分の受験者を追加して 100 人分にしたデータの散布図を図 6 に示す。そして、A 大学の受験者 60 人分と A・B・C 大学を合わせた受験者 100 人分、全 6 大学の受験者 178 人分の SJ-CAT と JSST の点数の平均、標準偏差、ピアソンの積率相関係数 r をそれぞれ表 10,11 に示す。

表 11 に示される A 大学の受験者 60 人分の SJ-CAT と JSST の結果の相関係数を見ると、両テスト間に強い相関がみとめられる。また、表 10,11 を見ると、B・C 大学の 40 人分の受験者を追加したとき、全 6 大学の受験者を合わせたときに、受験者のレベルが上級者に偏って分散が小さくなり、SJ-CAT と JSST のテスト間の相関も弱くなっていることがわかる。このことから、被験者実験のために集めた受験者のレベルの偏りの影響で相関が低くなっているとみられ、初級者の受験者を増やしてレベルの偏りを解消した状態で検証を行えば、もっと強い相関が確認できる可能性がある。

今回の検証で集めた受験者のレベルの偏りの影響で相関が低くなっている可能性があるため、今後は初級者の受験者を増やして再度検証を行うことが望ましいが、人間のテスターが採点するテストと SJ-CAT の結果の間にはある程度の相関がみとめられ、SJ-CAT により受験者のスピーキング能力を測定できることが確認できたといえる。

表 8 SJ-CAT の結果と JSST の結果

	最小値	最大値	平均	標準偏差
SJ-CAT(セクション 1)	4	24	15.8	2.8
SJ-CAT(セクション 2)	0	75	45.9	16.6
SJ-CAT(セクション合計)	4	94	61.7	17.9
JSST	1	9	5.9	1.5

表 9 SJ-CAT の結果と JSST の結果との相関係数 r

	JSST の結果との相関係数 r
SJ-CAT(セクション 1)	0.46
SJ-CAT(セクション 2)	0.63
SJ-CAT(セクション合計)	0.65

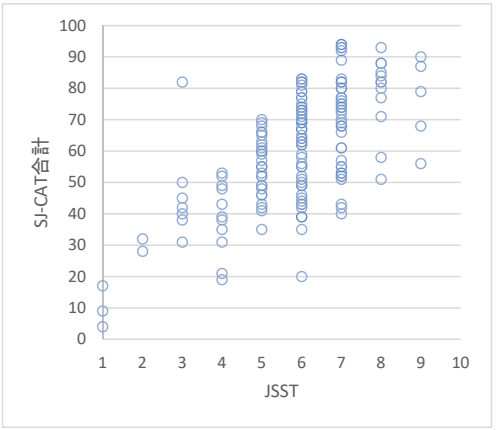

図 4 全 6 大学の SJ-CAT と JSST の結果の散布図

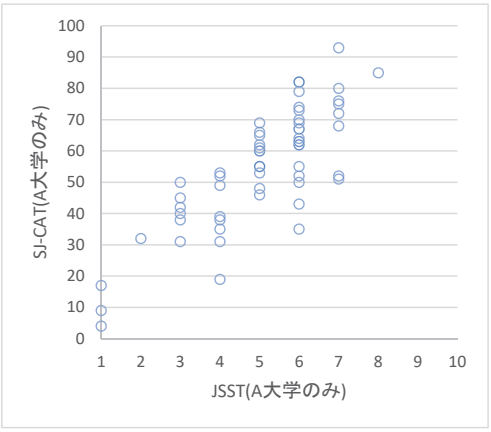

図 5　A 大学のみの受験者の SJ-CAT と JSST の結果の散布図

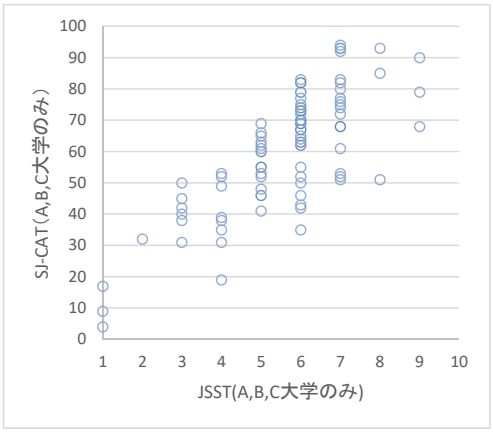

図 6　A・B・C 大学のみの受験者の SJ-CAT と JSST の結果の散布図

表 10　A 大学と A・B・C 大学、全 6 大学の平均点と標準偏差の比較

	SJ-CAT 平均	SJ-CAT 標準偏差	JSST 平均	JSST 標準偏差
A 大学	55.4	18.8	5.1	1.6
A・B・C 大学	61.4	18.4	5.6	1.6
全 6 大学	61.7	17.9	5.9	1.5

表 11　A 大学と A・B・C 大学、全 6 大学の相関係数 r の比較

	SCAT と JSST の相関係数 r
A 大学	0.81
A・B・C 大学	0.77
全 6 大学	0.65

6. おわりに

本論文では、日本語の総合的なスピーキング能力の測定を行う SJ-CAT を開発し、まず SJ-CAT の各問題の採点機能が、受験者の応答音声を適切に採点できるかどうかを確かめるための被験者実験を行った。実験の結果、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。また、人間が評定を行う日本語スピーキングテストの結果と SJ-CAT の結果の比較を行う被験者実験も行った。被験者実験の結果から、両テストの結果の間にある程度の相関がみとめられ、SJ-CAT により日本語学習者の総合的なスピーキング能力を測定できることを示した。今後は、特に精度が低かった文生成問題の採点機能に、採点に利用する特徴量を追加して精度が改善するか確かめる。また、実験で集めた受験者のレベルが中・上級者に偏っていることが原因となり、今回の検証で確認できた相関係数が低くなった可能性があるため、今後は初級者の受験者を増やして再度検証を行うことが望ましい。

謝辞　本研究は科学研究費基盤 (A)(22242014) の助成を受けた。研究にかかわった多数の協力者に感謝する。

参考文献

[1] 加藤圭介, 野沢和典, 山下洋一: 英語学習者の文発声における韻律自動評定, 情報処理学会研究報告, 音声言語情報処理, 124, pp.223-228 (2003).
[2] 近藤悠介: 日本人英語学習者の短い発話を自動採点するシステムの実現可能性の検討, 情報処理学会第 77 回全国大会, pp.4497-4498 (2015).
[3] Peason Education: Versant English Test, Test Description and Validation Summary (2011). 入手先 ⟨http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf⟩
[4] 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).
[5] 大西 翼, Dixon Paul R, 古井 貞熙: 音声認識技術の実用化への取り組み：8. WFST に基づく T^3 音声認識デコーダ, 情報処理, 51, 11, pp.1440-1448 (2010).
[6] 菊地賢一: 汎用的コンピュータ適応型テストシステムの開発, 日本テスト学会誌, 9, pp.69-75 (2013).
[7] 今井新悟: J-CAT(Japanese computerized adaptive test) の得点と Can-do スコアの関連づけ, ヨーロッパ日本語教育 14　第 14 回ヨーロッパ日本語教育シンポジウム報告論文集, pp.140-147 (2009).
[8] CENSREC-1: 入手先 ⟨http://www.slp.cs.tut.ac.jp/CENSREC/⟩
[9] 南條浩輝, 河原達也, 篠崎隆宏, 古井貞熙: 音声認識のための音響モデルと言語モデルの仕様, CSJ 付属ドキュメント.
[10] get_f0s: 入手先 ⟨https://ja.osdn.net/projects/galateatalk/downloads/22206/get_f0s-0.1.tar.gz⟩
[11] 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).
[12] 山上暁, 倉智佐一: 心理統計法, 北大路書房 (1991).
[13] 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).
[14] JSST: 入手先 ⟨http://www.alc.co.jp/jsst/⟩
[15] EasyEstGRM: 入手先 ⟨http://irtanalysis.main.jp/⟩