英検 公益財団法人 日本英語検定協会

英検 後援:文部科学省

英検 4 級 5 級スピーキングテストに関する Sinewave 社(iFlytek 社)との自動採点 共同研究レポート

2018 年 10 月

公益財団法人　日本英語検定協会

目次

1.1 研究概要

本研究においては、公益財団日本英語検定協会が株式会社 Sinewave(技術提供 iFlytek)*1 と共同で、2016 年度に実施された英検 4 級 5 級スピーキングの過去受検データを元に一部のデータで機械学習を行い、残りのデータで自動採点と採点者による採点との一致率を含めた各種指標を算出しました。今後弊協会が実際に 4 級 5 級スピーキングテストを含む英検や、その他の試験に自動採点を用いた運用を進めるにあたって参考になる点、留意すべき点などが明らかになりました。

1.2 研究背景

公益財団法人日本英語検定協会は 1964 年のスタート以来、日本の英語力向上に寄与すべく学習者及び指導者の皆さまに寄り添って参りました。昨今では 2020 年の東京オリンピック・パラリンピックの開催を前に、英語力向上のための様々な取り組みが国を挙げて推進されています。現行の学習指導要領においても、小中高を通じて、コミュニケーション能力を育成し、「聞く」「話す」「読む」「書く」の 4 技能をバランスよく育成することが目指されています。4 つの技能をバランスよく指導・学習していただくことを目標に、弊協会は 2016 年度の実用英語技能検定(以下、英検)1 級・準 1 級・2 級へのライティングタスク導入、4 級 5 級へのスピーキングテスト導入を皮切りに、2017 年度の準 2 級・3 級へのライティングタスク導入を含め、4 技能化を進めて参りました。「話す」、「書く」にそれぞれ対応するスピーキング、ライティングタスクは、「聞く」、「読む」にそれぞれ対応するリスニング、リーディングとは異なり、受検者の回答を予め決められた基準に基づき採点をする必要があります。

今後さらに 4 技能化の流れが進むにつれて、より効率的にかつ精度の高い採点を行う必要性が生じることが予想されます。本研究は今後の自動採点導入に向けた基礎的な検証という位置づけとして、運用時の課題点やより望ましい運用の仕組みを探るため、現在実施・運営中の英検 4 級 5 級スピーキングテストを対象に、株式会社 Sinewave(技術提供 iFlytek)と共同で行われました。

2.1 英検試験概要

初級段階から上級段階まである英語学習を継ぎ目なくサポートするため、英検は 5 級から 1 級まで 7 つの級を設定し、作成・実施・運営を行っております。また 2016 年度からは一般財団法人日本生涯学習総合研究所と共同で、国際基準規格の CEFR と関連性をもたせたユニバーサルなスコア尺度「Common Scale for English (CSE)」の正式な運用をスタートし

ました。

2018 年度現在、1 級から 3 級までの 5 つの級は「聞く」「話す」「読む」「書く」の 4 つの技能を測定しており、4 級と 5 級は「聞く」「話す」「読む」の 3 つの技能を測定しています。このたび「大学入試英語成績提供システム」へ参加するのに伴い、従来通り一次試験合格者のみが二次試験を受検する「従来型」に加えて高校 3 年生を対象に一次試験の結果によらず二次試験を受検する「**英検 2020　2 days S-Interview**」、「**英検 2020　1 day S-CBT**」、年齢制限なく一次試験の結果によらず二次試験を受検する「英検 CBT」が新たに設けられました。

2.2 英検 4 級 5 級スピーキングテスト概要

4 技能のバランスの取れた英語の指導・学習を促進するため、2016 年より、4 級 5 級にスピーキングテスト*2 が導入されました。1 級から 3 級までの二次試験とは異なり、スピーキングテストの受検は任意とし、級認定とは切り離して考えることで、より多くの受検者にスピーキングを体験していただける仕組みといたしました。また 1 級から 3 級までの二次試験とは異なり、4 級 5 級スピーキングは自宅 PC やスマートフォンなどを用いて受検することができます。

4 級 5 級スピーキングテストは 1) 英文の音読タスク、2)英文に関連する質問に答えるタスク、3)イラストに関する質問に答えるタスク(4 級のみ)、4)受検者自身に関する質問に答えるタスクの 4 つの種類で構成されています。

2.3 自動採点に関する先行研究

Bennett and Zhang (2016)によると自動採点とは、「自由記述の機械的な採点を用いるものであり、かつ通常事前には全ての正しい解答が未知であるため、事前情報とのマッチングによる採点とは異なるもの」と定義されています。また、自動採点の対象とするタスクや対象となる受検者から予想される解答の性質によって用いる自動採点の方法も異なります。また英語試験においては数語程度のもの、いくつかの文からなる文章、エッセイ等や、回答が予測しやすい発話、予測しにくい発話を含むものまで様々なものが開発され、PTE Academic(Pearson,2011)や TOEFL iBT(Attali and Burstein, 2005)等大規模な試験においても運用がされています。

Cohen and Wollack (2006)によると、各英語試験における自動採点はまず人間の採点者による一定数の採点結果を用いて機械学習が行われます。各自動採点のアルゴリズムは受検者の回答から抽出する特徴量(例：単語数や品詞)やそれぞれの特徴量に与えられる重みや

機械学習の方法などが異なっています。自動採点を導入する利点としては、採点時のコストや採点時間の軽減が期待できることが挙げられています。課題点としては自動採点アルゴリズムが機械学習後、一度自動採点のアルゴリズムが確定してしまうと、ある特定の解答に対して本来取るべきスコアを取っていることを示すフィードバックが自動では行われない点が挙げられています。具体的には、最初の機械学習時点で与えられた回答パタンにないイレギュラーな回答は人間の採点者による採点結果とは異なるものが付与されてしまう可能性が示されています。そのため、自動採点のみを運用するのではなく、人間の採点者と組み合わせることで、人間の採点者による採点エラーや自動採点による採点エラーを防ぐことが出来るとされています。

3.1 検証方法

本章では自動採点の精度を検証するために英検 4 級 5 級のスピーキングテストの音声データを用いて人間の採点者による採点結果と自動採点による採点結果との比較を行いました。上記 Cohen and Wollack (2006)で提案されている一般的な機械学習による精度検証の手法に倣い、人間の採点者による採点結果と音声データの一部を学習用データとして、Deep Learning を用いて学習を行い RNN（再帰型ニューラルネットワーク）を構築します。構築後に機械学習に使用しなかった音声データを検証用データとして自動採点させました。なお、対象となるスピーキングテストに関して 4 級は全 5 問、5 級は全 4 問から構成されており、問題ごとに機械学習を行いました。全体の流れを図 1 に示します。

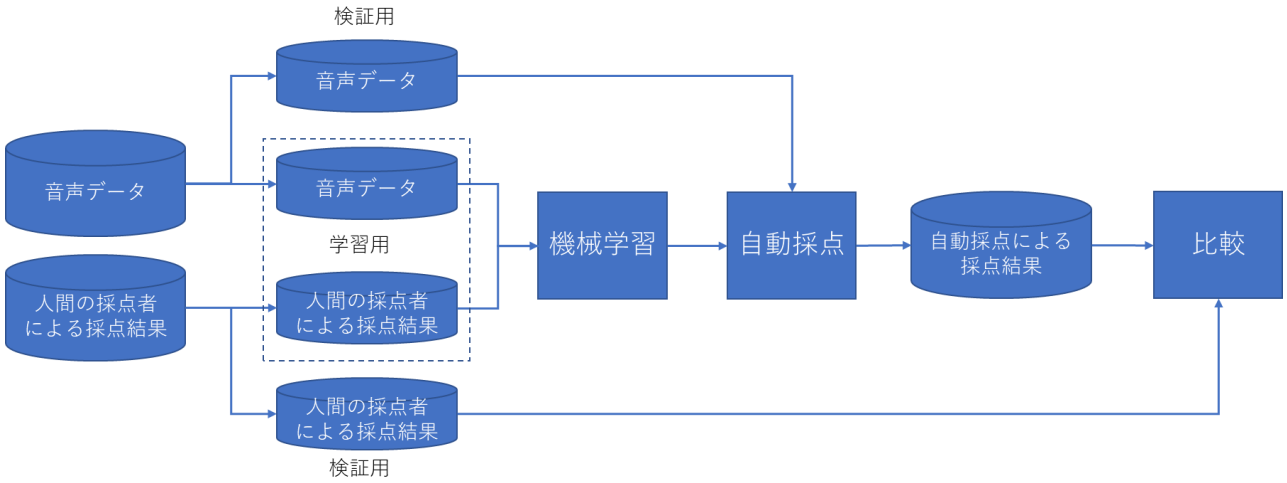


図 1　検証までの流れ

3.2 検証対象

今回使用した 4 級 5 級スピーキングテストのデータ件数は表 1 の通りとなります。4 級 5 級それぞれ 2016 年度に使用された 9 種類の問題セットのデータを用意しました。複数の問題セットを使うことにより問題ごとの採点精度のばらつきを検証しました。問題セットご

とに 300 人分のデータを学習用データとして使用し、残りを検証用データとしました。なお各問題の内容は 2.2 でも記載されていますが、Q1 は英文の音読タスク、4 級 Q2～Q4 および 5 級 Q2～Q3 は英文に関連する質問に答えるタスクとイラストに関する質問に答えるタスク、4 級 Q5 および 5 級 Q4 は受検者自身に関する質問に答えるタスクとなっています。

表１　検証データ

			Q1		Q2		Q3		Q4		Q5	
		人数	学習用データ	検証用データ	学習用データ	検証用データ	学習用データ	検証用データ	学習用データ	検証用データ	学習用データ	検証用データ
4級	set1	1294	300	994	300	994	300	994	300	994	300	994
	set2	567	300	267	300	267	300	267	300	267	300	267
	set3	570	300	270	300	270	300	270	300	270	300	270
	set4	1073	300	773	300	773	300	773	300	773	300	773
	set5	602	300	302	300	302	300	302	300	302	300	302
	set6	630	300	330	300	330	300	330	300	330	300	330
	set7	739	300	439	300	439	300	439	300	439	300	439
	set8	718	300	418	300	418	300	418	300	418	300	418
	set9	740	300	440	300	440	300	440	300	440	300	440
5級	set1	1465	300	1165	300	1165	300	1165	300	1165	—	—
	set2	523	300	223	300	223	300	223	300	223	—	—
	set3	523	300	223	300	223	300	223	300	223	—	—
	set4	1086	300	786	300	786	300	786	300	786	—	—
	set5	575	300	275	300	275	300	275	300	275	—	—
	set6	594	300	294	300	294	300	294	300	294	—	—
	set7	637	300	337	300	337	300	337	300	337	—	—
	set8	627	300	327	300	327	300	327	300	327	—	—
	set9	623	300	323	300	323	300	323	300	323	—	—

3.3 検証結果

自動採点の採点結果を人間の採点者の採点結果と比較し、点差別の件数を表２－１にまとめました。今回検証時に音声データの音響的特性として自動採点が困難なデータがあり、それを「異常」データとして表に含めました。異常データは「SN 比が低い」「振幅が極端に大きいまたは小さい」「無音」のデータが該当し、全体の 10%程度存在していました。それらの「異常」データは自動採点の対象外としました。本検証では各問題について自動採点の採点結果を人間の採点者の採点結果が完全一致または 1 点差以内を「一致」とみなしました。

表２－１の各問題セット・各問題ごとに一致した件数から一致率を算出し表２－２にまとめました。なお「異常」データは計算には含めていません。Q1 の英文の音読タスク形式の問題に対しては級や問題に関係なく約 98%の一致率を達成し、Q2 以降の質問に答えるタスク形式の問題についても平均して約 95%の一致率となることが分かりました。各問題について人の採点者の採点結果と自動採点の採点結果の分布を図２に示しました。人の採点者による採点結果と自動採点の採点結果に相関性のある結果となっています。

表２－１　人間の採点者の点数と自動採点の点差

			人間の採点者の点数と自動採点の点数の点差																																		
			Q1							Q2							Q3							Q4							Q5						
		検証用データ	0	1	2	3	4	5	異常	0	1	2	3	4	5	異常	0	1	2	3	4	5	異常	0	1	2	3	4	5	異常	0	1	2	3	4	5	異常
4級	set1	994	517	401	19	0	0	0	57	679	230	15	2	0	0	68	638	253	29	6	0	0	68	484	389	50	2	1	0	68	350	470	86	19	1	0	68
	set2	267	126	111	1	1	0	0	28	146	79	8	2	0	0	32	193	36	6	0	0	0	32	133	88	13	1	0	0	32	84	124	25	2	0	0	32
	set3	270	127	103	5	0	0	0	35	162	59	10	0	1	0	38	183	48	1	0	0	0	38	122	104	6	0	0	0	38	155	48	18	10	1	0	38
	set4	773	372	329	10	0	0	0	62	495	183	16	3	0	0	76	511	161	21	4	0	0	76	358	311	27	1	0	0	76	389	243	49	15	1	0	76
	set5	302	131	134	9	0	0	0	28	184	82	4	0	0	1	31	205	57	9	0	0	0	31	158	104	9	0	0	0	31	139	107	18	6	1	0	31
	set6	330	160	137	5	0	0	0	28	195	88	9	1	0	0	37	225	60	7	1	0	0	37	168	111	13	1	0	0	37	146	122	24	1	0	0	37
	set7	439	225	192	6	1	0	0	15	267	115	27	3	0	0	27	331	75	6	0	0	0	27	230	172	9	1	0	0	27	113	232	65	2	0	0	27
	set8	418	214	168	5	1	0	0	30	309	63	7	2	0	0	37	299	74	6	2	0	0	37	189	165	23	3	1	0	37	85	206	82	7	1	0	37
	set9	440	239	173	6	0	0	0	22	313	90	8	2	0	0	27	324	80	8	0	1	0	27	259	135	18	1	0	0	27	282	104	17	9	1	0	27
5級	set1	1165	514	507	33	1	0	0	110	646	355	25	3	0	0	136	715	286	28	0	0	0	136	873	116	27	7	6	0	136	—	—	—	—	—	—	—
	set2	223	114	66	2	0	0	0	41	114	37	5	1	0	0	66	132	22	1	2	0	0	66	96	44	10	6	1	0	66	—	—	—	—	—	—	—
	set3	223	117	72	2	1	0	0	31	123	45	3	2	0	0	50	122	45	4	2	0	0	50	139	24	8	1	1	0	50	—	—	—	—	—	—	—
	set4	786	373	304	19	0	0	0	90	499	170	11	1	0	0	105	436	219	26	0	0	0	105	416	234	28	3	0	0	105	—	—	—	—	—	—	—
	set5	275	118	121	3	0	0	0	33	172	39	1	2	0	0	61	158	48	7	1	0	0	61	122	58	25	8	1	0	61	—	—	—	—	—	—	—
	set6	294	122	124	11	0	0	0	37	168	69	4	1	0	0	52	156	68	17	1	0	0	52	129	84	25	4	0	0	52	—	—	—	—	—	—	—
	set7	337	173	133	5	0	0	0	26	275	30	2	2	0	0	28	271	34	4	0	0	0	28	191	94	19	5	0	0	28	—	—	—	—	—	—	—
	set8	327	155	130	6	0	0	0	36	159	100	9	0	0	0	59	191	74	1	2	0	0	59	171	84	9	4	0	0	59	—	—	—	—	—	—	—
	set9	323	166	115	7	0	0	0	35	238	31	6	0	0	0	48	211	61	3	0	0	0	48	156	84	31	4	0	0	48	—	—	—	—	—	—	—

表２－２　一致率

		Q1		Q2		Q3		Q4		Q5	
		1点差以内	2点差以上	1点差以内	2点差以上	1点差以内	2点差以上	1点差以内	2点差以上	1点差以内	2点差以上
4級	set1	0.98	0.02	0.98	0.02	0.96	0.04	0.94	0.06	0.89	0.11
	set2	0.99	0.01	0.96	0.04	0.97	0.03	0.94	0.06	0.89	0.11
	set3	0.98	0.02	0.95	0.05	1.00	0.00	0.97	0.03	0.88	0.13
	set4	0.99	0.01	0.97	0.03	0.96	0.04	0.96	0.04	0.91	0.09
	set5	0.97	0.03	0.98	0.02	0.97	0.03	0.97	0.03	0.91	0.09
	set6	0.98	0.02	0.97	0.03	0.97	0.03	0.95	0.05	0.91	0.09
	set7	0.98	0.02	0.93	0.07	0.99	0.01	0.98	0.02	0.84	0.16
	set8	0.98	0.02	0.98	0.02	0.98	0.02	0.93	0.07	0.76	0.24
	set9	0.99	0.01	0.98	0.02	0.98	0.02	0.95	0.05	0.93	0.07
5級	set1	0.97	0.03	0.97	0.03	0.97	0.03	0.96	0.04	-	-
	set2	0.99	0.01	0.96	0.04	0.98	0.02	0.89	0.11	-	-
	set3	0.98	0.02	0.97	0.03	0.97	0.03	0.94	0.06	-	-
	set4	0.97	0.03	0.98	0.02	0.96	0.04	0.95	0.05	-	-
	set5	0.99	0.01	0.99	0.01	0.96	0.04	0.84	0.16	-	-
	set6	0.96	0.04	0.98	0.02	0.93	0.07	0.88	0.12	-	-
	set7	0.98	0.02	0.99	0.01	0.99	0.01	0.92	0.08	-	-
	set8	0.98	0.02	0.97	0.03	0.99	0.01	0.95	0.05	-	-
	set9	0.98	0.02	0.98	0.02	0.99	0.01	0.87	0.13	-	-

図 2　各問題の得点分布

・4 級 5 級 Q1：英文の音読タスク
・4 級 Q2～Q4 および 5 級 Q2～Q3：英文に関連する質問に答えるタスクとイラストに関する質問に答えるタスク
・4 級 Q5 および 5 級 Q4：受検者自身に関する質問に答えるタスク

3.4 考察

本検証の結果を問題別にみた際、英文の音読タスク形式の問題に対しては級や問題に関係なく約 98%の一致率となることが分かりました。音読タスク形式の問題に比べて、採点の難易度の高い質問に答えるタスク形式についても平均して一致率が約 95%となったことから、これらの一致率から自動採点によって人間の採点者と遜色ない採点ができると考えられます。4 級 Q5 の受検者自身に関する質問に答えるタスクにおいては、他の問題に比べ一致率が下がっていますが、精度については学習データ数を増やしたり学習アルゴリズムの改善や学習モデルの改善を行ったりすることで、向上することが見込めます。

今回、機械学習では 300 件の学習データを用いましたが、一致率の向上とともに必要とする学習データ数の軽減についても今後の課題と考えられます。また、約 10%程度のデータは異常データとなっていますが、CBT 試験や試験会場以外の環境で録音する以上、必ず異常データは発生するものと考えられ、これらのデータについては音響的な問題から自動採点ではなく人間の採点者が採点を行うべきであると考えられます。従って自動採点を実運用に移した際、人間の採点者が全く採点を行う必要がなくなるものではないとも考えられますが、異常データを自動検知することで公平な採点が可能になります。今回の 4 級 5 級スピーキングテストに関する自動採点の精度検証結果から、同様のタスクを採用している 3 級以上のスピーキングテストに対しても同様の精度検証を行って参ります。また今回の結果から効率と精度の面から、より望ましい自動採点の運用方法について引き続き検討を行う予定です。

4.1 追記

*1 Sinewave 社(iFlytek 社)について

～　株式会社サインウェーブ Sinewave Inc.　～

音声認識/音声合成の研究開発ベンチャー企業として 2010 年 4 月設立。iFlytek 社と 2016 年 5 月に資本業務提携し英語教育事業に参入。日本における英語スピーキング自動採点システムの研究開発を行う。

～　iFlytek Co., Ltd.　～

中国科学技術大学発ベンチャー企業として 1999 年設立。音声認識/音声合成では世界トップレベルの技術力を有し、今では AI 分野に進出し世界有数の AI 企業となる。2007 年より中国国内の高校や大学の入試試験の英語スピーキング試験の自動採点を行い、これまで 2000 万人以上の採点を行ってきた実績がある。

*2 4級5級スピーキングテストについて

4級スピーキングテスト問題見本

Ken's Dream

Ken is in the soccer club at his junior high school. Ken plays soccer after school. Ken wants to become a famous soccer player.

※上記四角の枠内が受験者に画面上で提示される情報です。

[質問・満点解答例]
（下記質問の前に、パッセージ（英文）の黙読・音読タスクが課されます。）

No. 1 Please look at the passage. What does Ken want to become?
—— He wants to become a famous soccer player.

No. 2 When does Ken play soccer?
—— He plays after school.

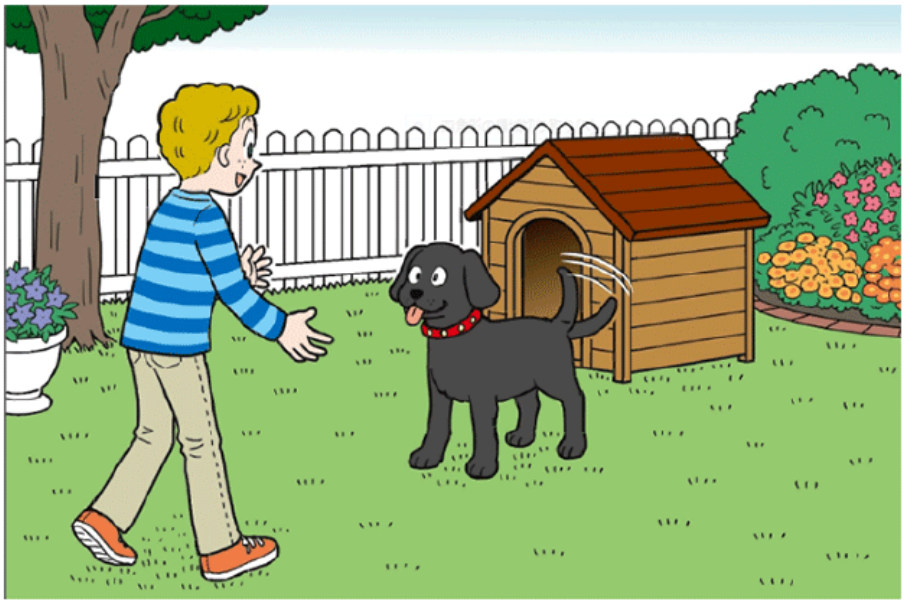
No. 3 Please look at the picture. What is the girl doing?
—— She is reading a book.

No. 4 Do you like to play sports?
Yes. → **What sport do you play?**
—— I play tennis.

5級スピーキングテスト問題見本

Sam's Pet

Sam is 10 years old, and he has a dog. The dog's name is Lisa.　Lisa is black.　Sam likes Lisa.



※上記四角の枠内が受験者に画面上で提示される情報です。

[質問・満点解答例]

（下記質問の前に、パッセージ（英文）の黙読・音読タスクが課されます。）

No. 1　Please look at the passage. How old is Sam?

—— He is 10 years old.

No. 2　What color is Lisa?

—— She is black.

No. 3　What animal do you like?

—— I like lions.

4.2 参考資料

Benett, R.E.& Zhang,M. (2016). Validity and Automated Scoring. In Drasgow, F. (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp.142-173).New York. Routledge.

Cohen, A.S. & Wollack,J.A. (2006). Test Administration, Security, Scoring, and Reporting. In B. Brennan (Ed.), *Educational measurement*(pp.355-386). Westport,CT. American Council on Education & Praeger.

Pearson.(2011). Pearson Test of English Academic: Automated Scoring.

Attali,Y. & Burstein, J. (2005).Automated Essay Scoring With E-rater v.2.0.ETS Research Report NO.RR-04-45.