

令和 6 年 6 月 5 日現在

機関番号：14301

研究種目：若手研究(B)

研究期間：2017～2023

課題番号：17K13498

研究課題名（和文）英語スピーキングの自動評価に向けた評価指標と評価目的別スコアつき発話データの構築

研究課題名（英文）Developing Evaluation Criteria and Annotated Speech Corpus for Automated Assessment of English Speaking Skills

研究代表者

金丸 敏幸（Kanamaru, Toshiyuki）

京都大学・国際高等教育院・准教授

研究者番号：70435791

交付決定額（研究期間全体）：（直接経費） 3,000,000円

研究成果の概要（和文）：本研究は、英語スピーキングの自動評価に向けた評価付きデータの作成を目的としていた。初年度は既存研究の調査と評価指標の整理、途中年度で発話データの符号化環境を整備し、スコア付とベクトル化の研究を行った。しかし、発話データの収集を行う段階となって、新型コロナウイルスの感染が拡大したため、発話の実データが収集できなかった。さらに生成AIの登場により、当初の研究目的の意義が大きく低下したため、それまでの評価指標の研究に基づき、生成AI時代の英語教育研究と対面教育の脳科学的研究に取り組むこととなった。英語教育における生成AIの活用や国際シンポジウムでの発表などで、本研究は一定の成果を挙げることができた。

研究成果の学術的意義や社会的意義

英語スピーキングの評価基準に関する分析に基づき、どのような点を指導、改善すれば英語運用能力の向上させることができるのかについての研究を行ってきた。当初は、自動評価用のシステムの学習に用いる発話データを構築する予定であったが、コロナ禍を挟んで生成AIの急速な発展により、学習用データそのものの必然性が低下した。そこで、評価のあり方を中心とする研究を進めることとなった。その研究の知見を活かし、生成AI時代に必要となる英語教育についての考察を深めることができた。これらの考察は、今後の英語教育に対する指針を示すことができた点で、学術的、社会的に果たした意義は大きいと考えられる。

研究成果の概要（英文）：This research aimed to create evaluation criteria and an annotated speech corpus for automatic evaluation for English speaking. In the first year, existing research was surveyed, and evaluation criteria were organized. In the two to three years, an environment for encoding speech data was prepared, and research on scoring and vectorization was conducted. However, at the stage of collecting speech data, the spread of COVID-19 prevented the collection of actual speech data. Furthermore, with the emergence of generative AI, the significance of the original research objective was greatly diminished. Based on the evaluation criteria research up to that point, the focus shifted to research on English education in the era of generative AI and the neuroscientific study of in-person education. This research achieved certain results through the application of generative AI in English education and presentations at international symposiums.

研究分野：英語教育

キーワード：自動評価 評価基準 生成AI

### 1. 研究開始当初の背景

従来、画像の識別や文章の分類といった高度な知的作業をコンピュータに行わせるには、人間がその判断基準(指標)を定義し、プログラムとして記述するしかなかった。そのため、指標の設計と指標の重要性の調整(重みづけ)に膨大な手間を必要としていた。しかし、1990年代の後半からの機械学習技術の発展によって、指標間の重みづけを自動的に学習させることが可能となった。機械学習では、正解である分類結果に最も近づくように、指標の重みづけを統計的に計算して、分類のためのモデルを生成する。つまり、機械学習とは、分類先のグループが持つ指標に合わせた分類出力ができるように指標間の重みづけを計算することである。2012年に発表されたDeep Learningでは、複数の層(中間層)を用いて、正確な分類モデルを学習することに成功している。

機械学習を用いた分類問題の応用先の一つがライティングの自動評価である。原理としては、次のようなものである。まず、出力すべき分類先をライティングのスコアとする。次に、ライティングデータを様々な指標に基づいて数値化し、複数の数値を持つベクトルデータに変換する。学習用の正解データとして、スコアの付与された大量のベクトル化ライティングデータを用意し、指標に応じたスコアを出力するモデルを学習する。学習されたモデルが自動評価システムとなり、新規のライティングデータを入力として与えると、指標に応じたスコアが出力される仕組みである。ライティングの自動評価に関する研究では、大きく分けて、どのような指標を用いるのかという指標の設計の問題と、どのように調整を行うのかという機械学習手法の問題が扱われている。

本研究の開始までに、学習者コーパス研究の知見に基づいて、Type Token Ratio(TTR)やTransitional Signals(TS)の使用割合などを指標として用いると、スコアとの相関が高くなることを明らかにしてきた。米国のEducational Testing Serviceは、TOEFL iBT®のライティング試験の評価にe-rator®と呼ばれる自動評価システムを導入している(Monaghan & Bridgeman, 2005)。このライティングの自動評価の仕組みを応用することで、原理的にはスピーキングの自動評価を行うことが可能となる。しかし、ライティングと異なり、スピーキングの自動評価では次の二つの点が未解決であった。一つは、音声という物理現象の扱いである。ライティングでは、すでに文字として入力さや文法間違いといった表層上のエラーを除けば、ベクトル化は容易であった。一方、スピーキングでは、そもそも英語として発音が認識されなければ、コンピュータで扱うためのデータ化そのものができない。この点をどのように評価するかが問題となる。もう一つは、発話が時間という次元を持つために、同じ発音、内容であっても、どの程度の時間で話されるかによって、評価が異なるという点である。したがって、発話の評価では、これらの観点を総合的に評価するためのCAF(Complexity, Accuracy, Fluency)指標が用いられることが一般的である。ただし、CAF指標がどのような目的の評価に用いられるかといった点が研究者によって異なるため、特定の発話データを一定の基準で比較することが難しいという現状があった(Zechner, et al., 2015)。

以上の点を考慮すると、機械学習を用いたスピーキングの自動評価のためには、共通となる指標の整備と、評価目的に応じた指標とその重みづけがなされたデータが必要であった。研究開始時点では、これらのデータを整備するための研究は行われていなかった。

### 2. 研究の目的

本研究では、スピーキングの自動評価に向けた基礎データとして、評価目的に応じたスコアと指標ごとの数値データを付与した発話データの整備することを目的とした。スピーキングの自動評価では、どのような発話がどのような評価を受けるべきかといった客観的なデータが十分に整備されていない。そこで、本研究では、評価目的と評価指標の対応関係を明確にし、その目的に応じたスコア付きのスピーキングデータを用意することで、この分野における基礎資料を提供しようと試みた。単なる自動評価の実現ではなく、当該分野の基礎資料を整備する点が本研究の独創的な狙いであった。これにより、評価目的と指標ごとの比較や、指標間の調整の優劣を競うことによって、さらに評価精度を高めるといった当該研究分野の活性化を見込んでいた。

スピーキングの自動評価が実用化されれば、小型の音声記録装置を用いて、録音した発話を大量に評価することができるため、現行の試験においても、同様の規模と費用でスピーキングの公正な評価が可能となる。また、人間による評価と異なり、同一の客観的な評価指標に基づいて評価するため、評価結果の公平性も担保される。したがって、ライティングの自動評価と同様、スピーキングの自動評価研究の発展はこれからの英語教育にとっても急務の課題であった。

### 3. 研究の方法

本研究では、研究段階を四つに分けて遂行した。まず、一年目では、英語スピーキングにおける評価指標の先行研究を整理し、先行研究で扱うスピーキング評価の目的の分類とその評価指

標に関するメタ分析研究を実施する。これにより、スピーキング評価で用いられる指標のリストと、特定の評価目的に対して有効な指標とそうでない指標を、それぞれ明らかにする。まず、関連する論文を収集し、必要と思われるデータを採取する。これらの研究データを収集した後、研究ごとに効果量を再計算し、スピーキングの評価指標として、有効であると認められる指標を評価の目的ごとに分類する作業を行う。評価の目的ごとに指標をまとめた後は、これらを基に人手で評価ができるように評価基準表（ルーブリック）を整備する。

二年目と三年目は、リスト化された指標を自動評価に利用できるように数値化、客観化できる手順を設定する。指標を客観的に処理できるように、音声処理技術や音声認識技術の導入を図る。発話データの符号化技術については、音声認識エンジン Julius を用いた実装を行う。音声認識エンジンの出力結果を元に、初年度に実施したスピーキングの評価指標を対応させ、モデル音声に対して適切な指標値が出力されるように調整を行う。モデル音声については、英語話者の聞き取り易さ（Intelligibility）を反映するために、日本人英語学習者の発話と Intelligibility を対応させた ERJ コーパスを参考として、英語の音声面での特性を評価に反映する。同時に、英語スピーキング能力を測る検定試験のスコアを持つ参加者の協力を得て、スコア別の発話データの収集を行う。これらのデータに対し、公開されている試験の指標に添ったスコアの付与と、自動評価に利用できるようになった指標に基づいてデータの数値化を行う。同じく初年度にまとめたルーブリックに従い、各発話データに対して人手によりスコアを付与する。

最終年度には、これらのデータを用いた分類実験を行って指標間の調整を行う。調整の結果に基づいて指標の見直しや再度の数値化を行い、スコアと指標ごとの数値データを付与した発話データを公開用に整備する。評価目的に応じたスコアを出力する自動評価モデルの実験では、外部のクラウドサービスを利用して、自動評価モデルを構築する。自動評価モデルの調整ができ、評価指標に基づいたスコアと評価値の付与ができるようになった後は、テストデータを作成してスピーキングの自動評価の実験を行う。自動評価を行う機械学習のモデル出力と、人手によるルーブリックに基づくスコアの比較を行い、スコア出力の正解率を確認し、モデルの正確さを確認する。これを何度か繰り返し、指標の見直しや再度の指標値の出力を行う。最後に、評価目的ごとのスコアと評価指標ごとの数値データを付与した発話データのセットを整備し、研究目的のデータ公開準備を進める。

#### 4. 研究成果

初年度は、英語スピーキングの自動評価に向けた評価付きデータの作成に向けて、関連研究の収集およびデータ抽出を行い、外国語学習に関する評価と指標についての整理を行った。また、整理した結果に基づいた研究発表とシンポジウムでのパネリストとしての発表を行った。具体的には、これまで行われてきたスピーキング指導と評価に関する論文を収集し、それらの研究で使用されている評価の記述を抜き出し、項目、観点、レベル、評価の粒度といった分類軸ごとに整理を行った。これにより、現状のスピーキングにおける評価の特徴が明らかとなり、今後の発話データへの評価を行う下地を整えることができた。初年度の研究により、現在、大きな注目を集めている人工知能や機械学習によるスピーキング指導やスピーキングの自動評価についての動向を整理し、今後の展望を中心とした研究発表を行った。外国語教育メディア学会などでのシンポジウムや研究発表を通じて、自動評価や言語処理に関する意見交換や情報収集を積極的に行い、今後の研究活動の発展に向けた取り組みを進めることができた。

二年目は、英語スピーキングの自動評価に向けた評価付きデータの作成に向けて、(1)発話データの符号化技術に関する課題と、(2)発話データの収集、および収集したデータのスコア付与とベクトル化に関する課題、の二つの研究課題を並行して実施する予定であったが、研究スケジュールの都合により、発話データの収集を実施することができなかった。しかしながら、データのスコア付与とベクトル化に関する課題について、順調に進めることができ、これらの成果に基づいた招待講演とシンポジウムでのパネリストとしての発表を行った。具体的には、発話データを符号化する環境整備として音声認識エンジンである Julius のインストールや調整を行った。また、音声分析のため、音響分析ソフトである Praat のインストールや調整も行った。これにより、音声データの分析準備が完了し、昨年度の発話データ評価基準をもとにして、符号化された音声データの評価分析を行うことのできる準備が整った。また、初年度の文献研究により、人工知能や機械学習によるスピーキングの自動評価についての概要と問題点を整理し、今後の課題を中心とした研究発表を行った。東北大学での招待講演や京都大学でのシンポジウム講演を通じて、自動評価や言語処理に関する意見交換や情報収集を積極的に行い、三年目以降の研究活動の発展に向けた取り組みを進めることができた。

三年目は、英語スピーキングの自動評価に向けた評価付きデータの作成に向けて、前年度に引き続き、発話データの収集および収集したデータのスコア付与とベクトル化に関する課題を実施する予定であったが、実験協力者を十分に集められず、発話データの収集についてはさらに次年度も引き続き行うこととした。現在までのところ、集まったデータに対して、データのクリーニングと問題点の洗い出しを行い、次年度のデータ収集後、スムーズに実験に移行できるよう準備を整えた。具体的には、録音環境および機材の整備、またソフトウェアによるノイズリダクションのパラメータ調整などを行い、その値を記録している。また、音声分析のために準備した音響分析ソフトである Praat により、音声データの分析を試行し、発話データ評価基準をもとにして、符号化された音声データの評価分析の見通しを立てるなどの予備実験を実施した。一方で、

最終年度の機械学習の実験環境を整備するために、実験プラットフォームの選定やテスト試行を行って、最終年度にスムーズに実験を開始できるための準備を進めた。また、三年目も引き続き、文献研究を実施し、人工知能や機械学習によるスピーキングの自動評価についての概要と問題点を整理している。学会等で自動評価や言語処理に関する意見交換や情報収集を積極的に行い、計画最終年度に向けて研究を進めることができた。

当初の最終年度に新型コロナウイルスの感染拡大が起こり、対面でのデータ収集が行えない事態が続き、当初の研究計画を二年間延長することとなった。しかしながら、延長した研究期間においても追加のデータ収集が一切行えず、十分な実験データを用意することができなかった。そのため、当初予定していた実装モデルとスコアの調整についての実験も実施することができなかった。さらに、2022年11月に公開されたOpenAI社のChatGPTによって、生成AIによる自動評価の研究が急速に進み、本来の研究目的である機械学習のための評価指標の構築、ならびに自動評価の評価値の付与についても、その研究的意義が大きく低下することとなった。しかし、これまでの文献調査の成果を活かし、生成AI時代に必要な英語運用能力の評価に関する研究を進める方へ舵を切ることができた。

最終的に、生成AIを活用した英語教育に関する研究を展開し、その成果に基づいて多くの講演を実施することとなり、本研究は一定の成果を挙げることができたと言える。また、生成AIを用いた英語教育では、人間による教育の意義について再検討することが必要であるという結論が得られたため、社会性に基づく英語教育という観点から、Neural Couplingという対面、対人教育時に観察される脳波の同期現象に着目した脳科学との国際シンポジウムの開催と発表を行った。当初の目的の延長として、生成AI時代を見据えた英語教育について、知見を重ねることができた点が本研究の大きな成果であったと言える。

5. 主な発表論文等

〔雑誌論文〕 計3件（うち査読付論文 0件／うち国際共著 0件／うちオープンアクセス 0件）

1. 著者名 金丸敏幸	4. 巻 73
2. 論文標題 生成AIにできること, できないこと	5. 発行年 2024年
3. 雑誌名 英語教育	6. 最初と最後の頁 44-45
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 金丸敏幸	4. 巻 72
2. 論文標題 ChatGPTはどうやって言葉を学習しているのか?	5. 発行年 2023年
3. 雑誌名 英語教育	6. 最初と最後の頁 34-35
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 金丸敏幸	4. 巻 66
2. 論文標題 英語の「やり取り」はAIロボットでできる?	5. 発行年 2018年
3. 雑誌名 英語教育	6. 最初と最後の頁 30
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

〔学会発表〕 計8件（うち招待講演 4件／うち国際学会 1件）

1. 発表者名 金丸敏幸
2. 発表標題 Prospects of Foreign Language Education Based on Social Cognitive Systems
3. 学会等名 Symposium on Social Neuroscience for Education and Development (国際学会)
4. 発表年 2024年

1. 発表者名 金丸敏幸
2. 発表標題 大規模言語モデルから見直すUsage-based Model的言語観
3. 学会等名 日本認知言語学会第24回全国大会
4. 発表年 2023年

1. 発表者名 金丸敏幸
2. 発表標題 ChatGPTが語学（英語）教育に与えるインパクト
3. 学会等名 【第63回】大学等におけるオンライン教育とデジタル変革に関するサイバーシンポジウム「教育機関DXシンポ」（招待講演）
4. 発表年 2023年

1. 発表者名 金丸敏幸
2. 発表標題 オンライン教材ことはじめ 企画から開発、運用に求められること
3. 学会等名 JACET第3回ジョイントセミナー（招待講演）
4. 発表年 2021年

1. 発表者名 金丸敏幸
2. 発表標題 英語スピーキング自動評価の試み
3. 学会等名 シンポジウム「大学の総合力を活かした外国語教育：スキルと教養」
4. 発表年 2019年

1. 発表者名 金丸敏幸
2. 発表標題 外国語教育から見た自動採点の現状と展望
3. 学会等名 みちのく情報伝達学セミナー（招待講演）
4. 発表年 2018年

1. 発表者名 金丸敏幸，山下仁司，東淳一
2. 発表標題 人工知能によるスピーキングの自律学習支援と自動評価の可能性
3. 学会等名 外国語教育メディア学会第57回（2017）全国研究大会
4. 発表年 2017年

1. 発表者名 金丸敏幸
2. 発表標題 授業内で活用できる自然言語処理技術
3. 学会等名 外国語教育メディア学会（LET）関東支部第39回大会（招待講演）
4. 発表年 2017年

〔図書〕 計2件

1. 著者名 Akira Tajino ed.	4. 発行年 2019年
2. 出版社 A Systems Approach to Language Pedagogy	5. 総ページ数 149
3. 書名 Springer Singapore	

1. 著者名 Kanamaru, T., & Pearce, D. R.	4. 発行年 2018年
2. 出版社 Routledge	5. 総ページ数 244
3. 書名 Ch. 7: Map grammar and ICT applications: Tajino, A. (Ed.) A new approach to English pedagogical grammar: The order of meaning.	

〔産業財産権〕

〔その他〕

-

6. 研究組織		
氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計1件

国際研究集会 Symposium on Social Neuroscience for Education and Development	開催年 2024年～2024年
--	--------------------

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------