様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 6 月 24 日現在

機関番号：34315
研究種目：若手研究（B）
研究期間：2009～2010
課題番号：21720209
研究課題名（和文）　英語学習者のための韻律自動評価システムの構築とその評価

研究課題名（英文）　The development and the evaluation of automatic speech scoring system for learners of English

研究代表者
近藤　悠介（KONDO YUSUKE）
立命館大学・言語教育センター・講師
研究者番号：80409739

研究成果の概要（和文）：
本研究では、学習者の発話に表れる特徴と評定者による評価の関係を検証し、発話に表れる特徴から評定者による評価を予測するというモデルを用いて自動評価システムの構築を試みた。評定者による評価を統計的有意に予測する特徴量（話す速さに関する指標および等時性に関する指標）を発見し、これらをもとに自動評価システムを構築し、システムが送出する評価の予測可能性を検証したところ、かなりの程度正確に評定者による評価を予測していることが分かり、システムの実用性が示唆された。

研究成果の概要（英文）：
The purpose of this study is to build an automatic L2 speech evaluation system which predicts evaluation scores given by experienced teachers who are L2 users. The predictability of the evaluation scores by speech characteristics of learner performance data is investigated in two types of speech. Based on the results of the pilot studies, an automatic L2 speech evaluation system was constructed, and its reliability was examined. The results indicate that the system is capable of delivering reliable scores in L2 assessment of read-aloud speech.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	600,000	180,000	780,000
2010 年度	600,000	180,000	780,000
年度			
年度			
年度			
総　計	1,200,000	360,000	1,560,000

研究分野：応用言語学
科研費の分科・細目：言語学・外国語教育
キーワード：教育評価・測定、自動評価システム

１．研究開始当初の背景

一般に、第二言語のスピーキング能力の評価において、受検者はある話題について議論したり、数枚の絵を描写したり、決められたシチュエーションでロール・プレイを行い、そのパフォーマンスは、訓練を受けた評定者によって評価される。評定者間の一貫性および評価値の信頼性を得るために、何人もの評定者が、繰り返し録音あるいは録画されたもので受検者のパフォーマンスを評価するものが多く、スピーキング能力の測定は相当な時間を要することが実情である。このような背景から、本研究はスピーキング能力の中でも特に韻律を評価対象とした自動評価システムの構築を目指すものである。

スピーキング能力の自動評価システムに

関する研究では、学習者のスピーキングの特徴を捉え、分析する必要があるため、大量のデータが必要である。そのため、音声コーパスの研究と大きくかかわっている。第二言語学習者の音声コーパスの構築、韻律および発音の自動評価システムの構築、さらに、発音および韻律の自動学習システムに関する研究は、Cucchiarini, C., Strik, H., and Boves, L. (2000).や Deterding, D., and Ling, L. E. (2005). など海外でも多く行われており、また、国内においても音声工学の分野で、近藤、前木、白井、匂坂(2003)、前木、内藤、近藤、白井、匂坂(2003)、Muto, M., Sagisaka, Y., Naito, T., Maeki, D., Kondo, A., and Shirai, K. (2003).などで行われている。匂坂らの韻律自動評価システムの構築に関する研究では、学習者がある一定の文を読み、それを評定者が評価し、その評価値を客観的に測定できる学習者の特徴（評価の対象となる文章を学習者が読んだ際のポーズの時間長さや話す速さなど）で予測しようという試みがなされている。すなわち、評定者による主観的評価値を基準変数にし、客観的に測定できる学習者の値を予測変数にした重回帰モデルである。本研究においても匂坂らの研究に基づき重回帰モデルを採用する。

このモデルにおいて基準変数である、評定者が付与する評価値は、予測変数である客観的測定値を精査する際の基準となる値である。そのため、この評価値の信頼性は十分に検討されるべきものである。しかしながら、これまで音声工学の分野で行われた研究において、評価値は単独の評定者によるものが使用されることが多く、評価項目、評価基準が十分に議論されることはなかった。さらに、学習者が読み上げる文は短文で、評定者が評価する際に学習者の発話の差異を発見することが困難であると推測される。

本研究における自動評価システムの構築のために使用するデータは、アジア人英語学習者 111 人がイソップ童話の『北風と太陽』を音読したデータである。113 単語、5 文からなる文章で、熟達度の差異が音読に表れると考えられて選択された。『北風と太陽』は第二言語としての英語の音声データベースの読み上げ文としていくつかの研究で採用されていたことも理由のひとつである。このデータでは、すべての学習者の音読に訓練を受けた評定者により評価値が付与されている。評価においては、総勢 11 人の英語教師が参加し、Common European Framework of References (CEFR)を評価基準とし、CEFR の基準ですでにレベル分けされた学習者の映像を視聴し、学習者の特徴について議論した。また、評価項目は、八代, 荒木, 樋口, 山本, コミサロフ. (2001)から音読の評価において適切な項目（読む速さの適切さ、ポーズの回数の適切さなど 14 項目）を選出した。この過程は評定者の議論により評価値の信頼性、評定者間の評価の一貫性を高めようとするものであった。さらに評価値は、項目応答理論に基づいて分析され、測定概念にそぐわない評価項目、また、評価に一貫性のない評定者は削除された。この過程の効果は、平成 19-20 年度科研費に助成を受けて本申請者が共同研究で行った Nakano, Kondo, Tsubaki, and Sagisaka. (2008)において報告された。この研究では、評定者が行った評価についての議論の前後で信頼性の変化を一般化可能性理論に基づいて分析を行った。結果としては議論の後で、推定される項目による誤差が大幅に減少していることが認められ、評定者が評価について議論することおよび明確な評価基準を参照することによって、評価の信頼性が向上することが統計的に確認された。

客観的に測定できる学習者の特徴を用いて、訓練を受けた評定者の評価値を予測する予測式を求めることが自動評価システム構築の第一段階である。つまり、$\hat{Y}=X1+X2\ldots Xn$ において実際の評価値に近い値 $\hat{Y}$ を予測する客観的測定値を探索することである。平成 19-20 年度科研費の助成を受けて行った一連の研究（Kondo, Tsutsui, and Nakano. (2008), Kitagawa, and Kondo. (2008), Nakano, Kondo, Kondo, Tsutsui, Tsubaki, Nakamura, Sagisaka, and Nakano. (2007), Kitagawa, Kondo, and Nakano. (2007), 近藤, 筒井, 中野, 鍔木, 中村, 匂坂. (2007)）では、学習者の発話の韻律に関する特徴（話す速さ、ポーズの時間長、強勢音節と非強勢音節との比など）、母音の質、統語的な理解が現れると考えられるポーズの位置などと評価値との相関が検討され、話す速さと強勢音節と非強勢音節との比のふたつから得る予測値が実際の評価値と最も相関が高いことが分かった（重相関係数.69）。本研究の第一の目的は、項目応答理論、一般化可能性理論に基づき評価値を再度検討し、また評定値との相関関係から客観的測定値を精査し、より精度の高い予測式を求めることである。

本研究で使用する音読データは、Hidden Markov Model Toolkit を利用して音素アライメントを行った。本研究では本申請者が TIMIT Acoustic-Phonetic Continuous Speech Corpus や VoxForge で公開されている音響モデルを基に、収集済みの音読データを用いて、音響モデルを第二言語学習者の連続音声の認識に適応させる。本研究の第二の目的は第二言語学習者の連続音声が正確に認識できる音響モデルを構築することである。この音響モデルを利用して音声認識システムを開発し、その結果をスクリプト言語(Perl)で計算し、受検者に結果を示すシステムを開発する。

２．研究の目的

本研究の目的は、英語学習者のための発話自動評価システムを構築し、実装、評価することである。このシステムは、学習者の発話に表れる客観的特徴を用いて経験のある英語教師が付与する評価値を予測するものである。

３．研究の方法

第一に、本研究で評価基準とする CEFR（Council of Europe, 2001）が設定するレベルの本研究における応用可能性をヨーロッパ言語ポートフォリオ（ELP: Little, 2002）を用いて検討する。CEFR および ELP は主にヨーロッパの外国語学習者を対象に作成されたもので、他の環境で外国語を学ぶ学習者の評価における適用可能性が検証されなければならない。第二に、CEFR を基づいて行った評定者の訓練の効果を検討する。評価における評定者の一貫性、信頼性を検討するため、分析には一般化可能性理論および多相ラッシュ・モデルを用いる。本研究は第二言語使用者を評定者として採用しており、評定者の一貫性、信頼性を検証する必要がある。第三に、自然発話および読み上げ文における評価において評定者、評価項目の信頼性を検討し、自然発話および読み上げ文において学習者の発話に表れる特徴と評定者による評価の関係を探り、この結果から英語学習者のための発話自動評価システムを構築し、このシステムが算出する評価値の信頼性を検証する。

４．研究成果

CEFR の日本での適用可能性は、ELP の Can-do 項目を使用し検証した。全体的に見ると CEFR が設定する 6 段階のレベルを使用して日本人英語学習者の発話能力を評価できると判断できる結果が得られた。いくつかの項目で教員・受講者間で困難度に対する認識の違いがみられ、また、CEFR が設定するレベルから逸脱するような困難度を持つ項目もみられたが、それらは学習者の発話能力そのものに関連するものではなく、翻訳や文化差といった周辺的な問題であり、CEFR のレベルおよび ELP は全体として日本の英語学習環境において適用可能であると判断した。評定者訓練の効果の検証では、一般化可能性理論に基づく分析において、評価項目に関する分散推定値が訓練後に大幅に減少することが分かった。また、多相ラッシュ・モデルに基づく分析においても評価項目の中に一貫性の低い項目があることが分かった。さらに、評定者の一貫性および厳しさにおいては訓練の前後で大きな変化は確認されなかった。本研究の評定者は経験のある日本人英語教師であるため、訓練の前後で一貫性や厳しさの変化は確認されなかったと考えられる。この結果は Weigle（1998）の結果が示すものと同様である。さらに、評定者は訓練を受けることにより評価項目に関する理解を共有するため、一般化可能性理論に基づく分析においてその分散推定値が大幅に減少したと考えられる。本研究の評定者の評価の一貫性を Kim（2009）の英語母語話者の評価を比較した場合、本研究の評定者はほぼ同等の一貫性を示しており、このことからも本研究の評定者の評価の信頼性、一貫性という観点から評定者としての適格性が実証された。

自己紹介発話において、学習者の発話の特徴量から評定者による評価値の予測を試みる研究では、CEFR に基づいて訓練を受けた第二言語使用者である評定者が、第二言語学習者の自己紹介発話を評価した。八代、荒木、樋口、山本、コミサロフ（2001）からこの評価に使用できる評価項目を評定者の議論のうえ、24 項目選出した。評価の信頼性を高めるため、多相ラッシュ・モデルに基づいて分析し、評価の一貫性という観点から、多相ラッシュ・モデルが算出する一貫性に関する指標（infit）を用い、評価項目および評定者の一貫性を検証した。分析の過程で一貫性のない評定者はいなかったが、4 つの評価項目が一貫性の低い項目として検知された。「パラ言語的な手掛かりの使用」、「自信の程度」、「緊張の程度」、「外国語訛りの程度」が一貫性の低い項目として検知されたが、最初の 3 項目に関しては自己紹介発話を音声のみで評価する場合には不適切な項目であり、一貫性なく使用されたと推測できる。「外国語訛りの程度」に関しては、現在の英語教育を取り巻く環境を鑑み、発話能力の熟達度とは関係が希薄なものと推測できる。これら 4 項目を除外したのちの分析では一貫性の低い評価項目および評定者は検知されなかった。このような結果は、本研究に評定者として参加した英語教師の経験や知識および評定者訓練の実施に起因するものである。

上記の評価で算出された評価値と発話に表れる学習者の特徴の関係を検証したところ、話す速さとフィラーの数が、評価値を統計的有意に予測する変数であることが分かった。本研究の自己紹介発話において測定した指標には語彙の豊富さ、文法的正確さ、統語的な複雑さを示す指標も含まれていたが、これらと評価値の相関係数は総じて高くなく、重回帰分析において統計的有意に評価値を予測するものはなかった。反対に学習者の発話の時間制御に関する指標（無音ポーズの数など）は評価値との相関が高いことが分かった。この結果により、評定者が受検者に付与する評価値はコンピュータで計測可能な時間制御に関する指標でかなりの程度予測することが可能であり、自動評価システムにおける第二言語学習者の発話評価の可能性

が示唆された。

対象を読み上げ文とした研究では、多相ラッシュ・モデルに基づいた読み上げ文での評価の分析と自己紹介発話の評価の分析を一貫性という観点から比較した場合、読み上げ文の評価の方が平均して評定者が一貫した評価ができていると言える。このことから読み上げ文の評価の方が自己紹介発話に比べより信頼性の高い評価が得られると判断できる。

自己紹介発話において評価を統計的有意に予測する特徴は話す速さの指標およびフィラーの数であった。第 5 章の実験結果を鑑みると、第二言語の発話評価において主要な予測変数は、文法的な正確さや語彙の豊富さではなく、発話における時間制御に関する特徴であると言える。この結果をもとに、5 つの実験を行い、読み上げ文における評価と発話における特徴の関係を検証し、評定者による評価を統計的有意に予測する特徴を探索した。これらの実験では評価と高い相関を示す特徴もいくつか発見されたが、話す速さの指標、等時性に関する指標（弱母音の時間的長さの平均と強母音時間的長さの平均の比）が統計的有意に評価を予測する変数であることが検証された。

自動評価システムの構築および評価に関する研究では、システムが送出する評価の予測精度は、受検者の自己評価とシステムが算出する評価の関係、評定者による評価とシステムが送出する評価の関係の 2 点により検証した。両者ともに妥当な結果が得られ、本システムがかなりの程度正確に評定者による評価を予測していることが検証され、システムの実用性が示唆された。

本研究の結果を解釈する際にはいくつかの限界を考慮する必要がある。本研究では評定者の訓練に使用する基準に関して十分な議論を行わず、CEFR を採用した。しかしながら、現在使用可能な言語の熟達度に関する基準において CEFR ほど豊富な資料を用意しているものがないことが理由として挙げられる。次に、評定者の適格性についての問題である。本研究では、日本人英語教師がアジアの様々な言語を母語として持つ学習者を評価したが、評定者と被評定者が共通の母語を持つ場合とそうでない場合においては評価に違いが現れる可能性がある。しかし、Kim（2009）が示すように英語母語話者と非英語母語話者の評価においても大きな違いが見られないことから、評定者が英語教育に十分な知識を持つ者であること、また、明確な基準を用いて訓練を行っていること、評価の分析を行い、評定者の一貫性、信頼性を検討していることから、本研究で得られた評価値は十分に信頼性の高いものと考えられる。評価に関するもうひとつの問題は、本研究の評価項目と評価に関して評定者による議論と統計に基づく判断のみにより決定していることである。しかしながら、North & Schneider（1998）が指摘しているように、現在のところ、理論的にも実証的にも妥当な第二言語の熟達度に関するモデルは存在しない。したがって、英語教師の経験と統計的な分析に基づいて評価項目およびタスクの種類を検討した。さらに、本研究で用いた学習者の発話誘出タスクは、自己紹介発話と物語の読み上げ文の 2 種類である。言語テストの分野ではタスクの種類や長さにより学習者の発話が異なることが分かっており、さまざまなタスクを用いて学習者の発話の特徴に関して今後検討されなければならない。最後に、本研究で提案した自動評価システムの採点方法は大きく既知のデータに依存していることを指摘する。本研究で採用した採点方法では、新たな受検者の発話の特徴量とすでに 3 段階のレベルに分けられた発話データの特徴の平均値との距離を計算し、最も近い平均値を持つレベルがその受検者の評価となる。この方法では、既存のデータが変われば、それぞれの平均値も変化し、新たな受検者の評価も変わることとなる。しかしながら、新たな受験者を募り行ったシステムが送出する評価の検証では、システムが送出する評価は評定者による評価をかなりの程度正確に予測していることが検証されたことから、本研究で構築された自動評価システムが送出する評価は信頼性があるものと言える。

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 5 件）

① Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics, 14 (2)*, 1-23, 査読有.

② Kondo, Y., Tsutsui, E., & Nakano, M. (2010). A tentative method of reforming your assessment of English abilities into international standards such as Common European Framework of Reference (CEFR) (1): The eligibility of raters and rater training effect in L2 performance assessment. *Proceedings of the 15th Conference of Pan-Pacific Association of Applied Linguistics*. 436-443, 査読有.

③ Kondo, Y., Tsutsui, E., & Nakano, M. (2010). Bridging the Gap between L2 Research and Classroom Practice (2): Evaluation of Automatic Scoring System for L2 Speech. *Proceedings of INTERSPEECH*

2010 Satellite Workshop on Second Language Studies. CD-ROM, 査読有.

④ Kondo, Y., & Nakano, M. (2009). Construction and implementation of automatic L2 speech evaluation system. *Proceedings of the 14th Conference of Pan-Pacific Association of Applied Linguistics*, 33-38, 査読有.

⑤ 近藤 悠介, 中野 美知子. (2009). 英語学習者のための音読自動評価システムの構築. 第 21 回日本音声学会全国大会予稿集, 87-92, 査読無.

〔学会発表〕(計 2 件)

① 大和田 和治, 近藤 悠介, 中野 美知子, 筒井 英一郎. (2010 年 9 月 8 日). ニューラルテスト理論の英語教育における利用. 2010 年度 ICT 調査研究特別委員会企画. 大学英語教育学会第 49 回全国大会. 宮城大学. 仙台.

② 上田 倫史, 見上 晃, 中野 美知子, 近藤 悠介, 筒井 英一郎. (2009 年 9 月 4 日). ICT 活用授業と授業評価～1 対 1 対応の遠隔授業と多地点遠隔授業、英語発信力の自動判定、個人差を考慮できる ICT 活用. 2009 年度 ICT 調査研究特別委員会企画. 大学英語教育学会第 48 回全国大会. 北海学園大学. 北海道.

６．研究組織

(1)研究代表者

近藤　悠介（KONDO YUSUKE）
立命館大学・言語教育センター・講師
研究者番号：80409739