

科学研究費助成事業 研究成果報告書

平成 26 年 6 月 12 日現在

機関番号：32689

研究種目：基盤研究(C)

研究期間：2011～2013

課題番号：23520721

研究課題名(和文) 英語学習者の総合的な発話能力を自動評価するシステムの構築とその評価

研究課題名(英文) The development of an automated scoring system for English learners' speech in discourse completion task and its evaluation

研究代表者

近藤 悠介 (Kondo, Yusuke)

早稲田大学・オープン教育センター・助教

研究者番号：80409739

交付決定額(研究期間全体)：(直接経費) 3,600,000円、(間接経費) 1,080,000円

研究成果の概要(和文)：本研究の目的は、スマートフォンやパソコンで使われるようになってきている音声認識技術を利用し、英語学習者のスピーキング能力を自動的に採点するシステムを開発することである。本研究では、設定された状況において適切な発話が求められる課題を用いて英語学習者のスピーキング能力の自動採点を試みた。音声認識技術は外国語の音声の認識において話されている状況や内容によって精度がかなり下がる。本研究で作成した課題のうち、認識の精度が高い課題においては、自動採点システムが算出するスコアは、実際に運用できる精度のものが得られた。

研究成果の概要(英文)：The purpose of this study is to develop an automated speech scoring system for learners of English, applying the speech recognition technology that recently used on personal computers and smart phone. The system automatically scores learners' speech elicited by Discourse Completion Task where learners are required to speak in a situation. The degree of accuracy of speech recognition depends heavily on a situation and contents of speaking. In the tasks with the high degree of accuracy of recognition, the system produces reliable scores, which means that the system edges closer to practical use.

研究分野：人文学

科研費の分科・細目：言語学・外国語教育

キーワード：自動採点システム 発話能力 英語学習者 国際情報交流(台湾) 言語テスト

1. 研究開始当初の背景

コンピュータを用いた英語の能力の評価・測定は言語テストに大きな影響を与えたが、客観的に評価することが難しい発話能力に関しては、いまだ開発過程にあると言える。英語学習者の発話能力を自動的に採点するシステムは、現在実用化されているもので、Pearson VUE が開発した Versant™ や Educational Testing Service が開発した SpeechRater™ がある。これら 2 つの発話自動採点システムは発話自動採点システムが抱える大きな問題に対し異なる解決方法を採用している。

発話自動採点システムが抱える問題とは、使用されるタスクおよび評価の観点が音声認識の技術に大きく制限されていることである (Xi, 2010, p. 294)。音声認識の技術では、一般的に発話の自由度が上がるにつれてその認識精度は落ちる。この問題の解決方法として Versant™ では、テキストで提示した文を読み上げたり、提示された語の反対語を言うタスクを用い、発話の自由度を制限している。Versant™ は、その構成概念を Facility in L2 とし、Facility は「日常的な事柄について話されている言葉を理解する能力および分かりやすい言葉を使って母語話者同士の会話と同じペースで適切に話すことができる能力」と定義している (Bernstein, Van Moere, & Cheng, 2010, p. 358)。つまり、Versant™ は、測定しようとする能力を間接的に測定していることとなる。一方で、SpeechRater™ は一般的な発話能力を評価するために、自然に近い発話を測定の対象としているが、報告されている単語認識率は、受検者によって 10% から 80% と大きく異なるが、全体として約 50% である。話された単語の約半分が正しく認識されているという状況で算出された評価が妥当であると言えることはできない。

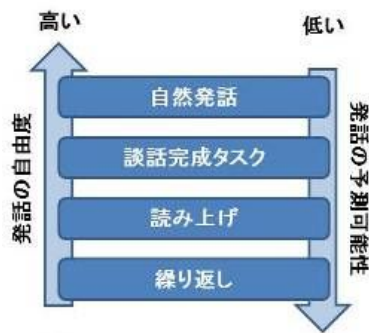


図1. 発話の自由度と予測可能性

図1が示すように、学習者の発話は発話を誘出するタスクによってその自由度と予測可能性は相反する。自然発話のような完全に自由な発話では、発せられる単語、文の予測は難しく、現在の音声認識の技術ではその精度は低い。反対に提示された音声を繰り返すだ

けのタスクにおける発話では、学習者の発話はほぼ一定のものとなり、現在の音声認識の技術を用いてかなりの程度正確に認識できるが、学習者の評価は、母音や子音の質、抑揚など限定的なものとなる。学習者の発話を多面的に捉え、自動的に評価するシステムの構築は、音声認識の技術およびタスクによる評価の違いを考慮し、学習者に発話の自由度が異なるタスクを与え、学習者の発話の様々な側面を採点し、総合的な評価を送出することが妥当である。

さらに、複数のタスクを用いて学習者の発話能力を自動的に採点するシステムを開発しようとする、新たな問題も発生する。発話自動採点システム開発では、既存の発話データにおいてその評価と発話の特徴量の関係を検証し、予測方法を選択する。つまり、開発の最初の段階では、評価値が付与された膨大な量の発話データが必要だということである。

2. 研究の目的

前述の発話自動採点システムの開発に関する問題を鑑み、本研究では英語学習者の音声認識率の向上、開発段階において評定者にかかる負担を考慮し、談話完成タスク (Discourse Completion Task: DCT) を用いて学習者の発話を自動的に採点するシステムの開発を目的とした。

3. 研究の方法

本研究で収集したデータ

本研究のデータは早稲田大学オープン教育センターが提供する Tutorial English において使用されている教科書を参照して作成された DCT を用いて収集した。以下に例を示す。

You cannot help her because you're busy now.

A: Could you help me with my homework?

B: () .

ある状況での適切な発話が学習者に求められるタスクである。発話データの提供者は録音専用のウェブサイトにアクセスし個別にデータを提供した。

Tutorial English は 6 レベルが設定されており、本研究ではその下から 2 番目のレベルである Basic を参照しタスクを作成した。144 項目の DCT を作成し、日本人英語学習者および台湾人英語学習者から約 26000 発話を収集した。この発話データを用いて、音声認識器の作成に必要な音響モデルおよび言語モデルを作成した。音響モデルの作成には、母語話者の約 6000 発話を含む TIMIT Acoustic-Phonetic Continuous Speech Corpus および日本人英語学習者による約 80000 の発話が収録された ERJ データベース

を利用した。また、研究代表者が平成 21 年度～22 年度に科学研究費補助金若手研究(B)「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」の助成を受けて収集した約 100 人のアジア人英語学習者の読み上げ音声も適宜利用した。

英語学習者の音声認識率の向上
 上記データの中から日本人大学生約 80 人より提供された発話データを使用した。DCT のひとつの項目に対し 78 人から 86 人分の発話が提供された。発話の総数は 10617 であった。
 音響モデルの作成において、母語話者の発話を含むモデル、非母語話者のみを使用して作成したモデル、男女混合モデル、男女別モデル、話者適応の技術を用いた認識結果を検討した。

評定者にかかる負担の軽減
 自動採点システム開発の目的のひとつに評定者にかかる負担の軽減がある。システム開発の初期段階で評価値が付与された発話データが必要であり、評定者による評価が必要であるということである。システム開発の上で評定者による評価値は不可欠であるが、本研究はその負担を軽減するために評価値が付与された少量のデータを用いて、評価値が付与されていない発話データの評価値を予測する方法を提案した。

4. 研究成果

英語学習者の音声認識率の向上
 母語話者の発話を利用してモデルを作成し、男女別、話者適応後の認識率の変化の一例を以下に示す。

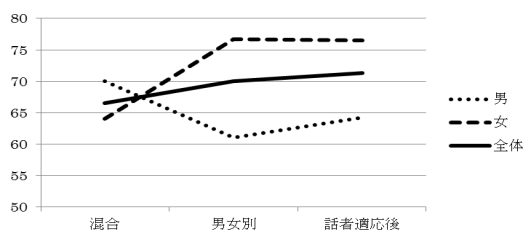


図 2: モデル別の平均認識率の変化

男女別モデルにすると男性の認識率は下がり、女性の認識率は上昇する。話者適応の効果はかなり低いことが分かった。

また、非母語話者の発話をモデルの作成に利用した場合と母語話者の発話を利用した場合では、非母語話者の発話を利用した場合の方が若干ではあるが高い認識率が得られた。
 また、認識率にかんして、モデルの違いは多少の影響を与えるが、タスクの種類が認識率に与える影響の方が大きいことが分かった。

評定者にかかる負担の軽減
 本研究ではアジア人英語学習者の読み上げ音声を利用した。この発話データはすでに A、

B、C の 3 段階の評価値が付与されている。さらに、自動採点システム開発のために使用されたデータであるため、評価値をよく予測できる発話の特徴量であると検証された 2 つの指標（話す速さの指標およびリズムに関する指標）が測定されている。A、B、C の 3 段階からランダムに 10 ずつのデータを選び、2 つの指標を用いて EM アルゴリズムに基づき、残りのデータの評価値を予測した。図 3 に実際のデータ、図 4 に EM アルゴリズムによる予測を示す。図中の縦軸は話す速さの指標、横軸はリズムに関する指標である。また、図中の ○ は A、● は B、× は C の評価が付与されたデータを示す。

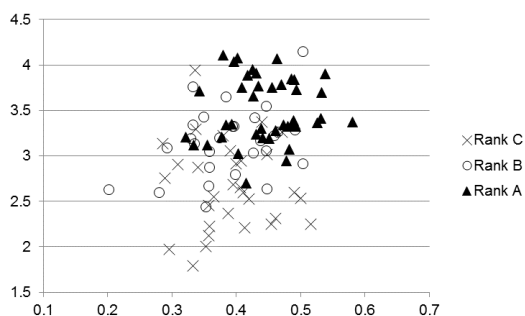


図 3: 評定者による評価

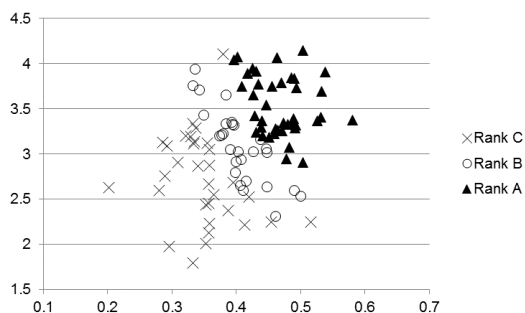


図 4: EM アルゴリズムによる予測

評定者による評価と EM アルゴリズムによって予測された評価値の相関係数は .64 と高い相関は得られなかった。

自動採点システムの開発

Higgins, Xi, Zechner, Williamson (2011) を参考に、人間が聞き取れない発話を排除するフィルターを開発した。音力、ピッチ、音響モデルのスコア、単語数、話す速さの指標を用いて、人間が聞き取れるか（採点可能かどうか）をサポート・ベクター・マシン (SVM) を用いて識別した。識別率は 84% であり、SVM により採点可能と判断された発話データの中から、単語認識率が高いタスクのデータを選び、自動採点を試みた。これらのタスクでは、発話の自由度がある程度制限されているため、自動採点の精度は実際に運用できる段階であった。

5. 主な発表論文等

〔雑誌論文〕(計8件)

Kondo, Y. and Ishii, Y. (2014). Bridging the Gap Between Second Language Acquisition Research and the Development of Automated Scoring System for Second Language Speech. In Tsai, R., C-H., and Redmer, G. (2014). *Language, Culture, and Information Technology*. 149-164, Taipei: Bookman Books. 査読無.

石井雄隆, 近藤悠介 (2014) 「英語学習者を対象とした自動採点システム?課題と展望」『メソドロジー研究部会報告論集』第4号, 1-11. 査読無.

Kondo, Y. (2013). Applying Naïve Bayes Classifier to the score prediction in L2 performance assessment. *Proceedings of the 18th Conference of Pan-Pacific Association of Applied Linguistics*. 査読無.

Kondo, Y. (2013). A method for reducing burden imposed on human raters in the construction of automated scoring systems for second language learners' speech. *Selected Papers of the 17th Conference of Pan-Pacific Association of Applied Linguistics*. 54-61. 査読有.

Nakano, M., Kondo, Y., Owada, K., Ueda, N. and Yoshida, S. (2013). English Language Education as a Lingua Franca in Asia. *The Asian Conference on Education 2012 Official Conference Proceedings*. 1368-1389. 査読無.

近藤 悠介, 中野 美知子. (2013). 談話完成タスクにおける第二言語の音声認識率向上の検討. *情報処理学会第75回講演論文集*. 第4巻. 405-406. 査読無.

Kondo, Y. (2012). World Englishes and learning pronunciation of English: Does vowel quality really matter? *Information Communication Technology Practice & Research 2011 2011年度授業実践報告書*. 大学英語教育学会 JACET-ICT 調査研究特別委員会. 査読有.

Kondo, Y., Kitagawa, A., and Nakano, M. (2011). Does vowel quality really matter? *Proceedings of the 16th Conference of Pan-Pacific Association of Applied Linguistics*. 74-79. 査読無.

〔学会発表〕(計5件)

近藤 悠介, 石井 雄隆. (2014年3月). 英語学習者のための発話自動採点システムの開発. 外国語教育メディア学会メソドロジー研究部会 2013年度第3回研究会, 東京.

近藤 真理子, DETEY Sylvain, 鄭 秋豫, 近藤 悠介, 鐔木 元. (2012年6月). アジア言語話者の英語発話コーパスの構築と研究への応用. 日本音声学会第325回研究例会, 東京.

Kondo, Y. (2012, Feb.). Does vowel quality really matter? PAAL International Symposium of World Englishes. Singapore.

近藤 悠介. (2011, Oct.). 音読と談話完成タスクを用いた英語学習者の発話自動採点の試み. 2011 Waseda Workshop on Language And Speech Science (LASS) and language education. 東京.

Nakano, M., Tsutsui, E., Kondo, Y., and Ueda, N. (2011, Aug.). A study of assessment of speech production and self-assessment based on Common European Framework of Reference (CEFR). The 16th World Congress of Applied Linguistics. Beijing, China.

6. 研究組織

(1)研究代表者

近藤 悠介 (KONDO, Yusuke)

早稲田大学・オープン教育センター・助教
研究者番号: 80409739

(2)研究分担者

中野 美知子 (NAKANO, Michiko)

早稲田大学・教育・総合科学学術院・教授
研究者番号: 70148229

吉田 諭史 (YOSHIDA, Satoshi)

早稲田大学・教育・総合科学学術院・教授
研究者番号: 00608838