英語らしさと正確さを考慮した英語発話のレベル判定の試み

Determination of the level of English speech in terms of Englishness and accuracy

追手門学院大学　心理学部教授
井佐原　均

通商産業省工業技術院電子技術総合研究所、郵政省通信総合研究所、独立行政法人情報通信研究機構、国立大学法人豊橋技術科学大学を経て、現職。産業日本語研究会世話人会代表

1　はじめに

社会のグローバル化に伴い、英語をはじめとする外国語教育の重要度は増している。我が国でも英語教育における情報通信技術（ICT）の活用が進みつつある。教育機関における追加の語学学習や社会人のリカレント教育においては学習支援ツールが有効であり、人工知能技術を用いた外国語学習支援のニーズが高まってきている。英語の４技能のうち、「読む」「聞く」「書く」に関しては自動学習支援の目途が立っているが、「話す」に関しては、既存の音声認識システムが学習者の音声を認識するためには作られていないことから、学習者の「話す」能力を判定することが難しい。本稿では学習者の発話を単語やテキストとしての正確さだけではなく、「英語らしさ」をも考慮した語学能力判定システムを提案する。これにより、不明瞭や不正確な学習者の発話がどの程度「正しい」かを判定することができ、リピーティングや自由対話による学習支援システムを用いての「話す」技能の向上を支援することが期待される。

教育における ICT の活用においては、履修管理や教材の管理・提供といった授業支援だけではなく、教育や学習の過程を直接支援する e-Learning 教材やアプリの開発が行われている。英語をはじめとする外国語教育においても、授業で使用するもの、個人で使用するものなど､様々な学習ツールが公開されている｡このようなツールを用いた学習は、提示する評価結果の正確さは人手による評価に劣るという課題はあるが、学習者の都合のよい時間に利用でき、結果の判定が即時に行われ、短時間に多くの反復演習が可能であり、学習に関わるコストが抑えられる等の利点がある。

英語教育においては４技能を評価することが必要となる。「読む」と「聞く」に関しては理解度テストによって評価が可能であり、「書く」に関しては近年発展著しいテキスト処理技術によって、内容面にも踏み込んだ評価が可能となろう。「話す」に関しても、発音の優劣を評価する自動採点システムをはじめとした発音訓練（CAPT: Computer Assisted Pronunciation Training）に関する研究が 1980 年代から数多く行われてきた。1990 年代に音声認識技術が大きく進展したが、これに伴い発音訓練に関する研究も大きく進展した。これは近年の深層学習の活用により、さらに精度向上が可能となった。このように「話す」に関しては音声認識システムが用いられるが、現状の音声認識システムは母語話者などの「正しい」発声を対象としており、学習者の発声の認識には適さない。また音声認識システムの「性能が向上」したために、不確かな発話でも正しく認識できるようになり、学習者の発声の正確さが音声認識の正解率と対応しない場合もある。

その一方で、人間同士の対話においては、母語話者同士であっても、必ずしもすべての発話が音韻的に正しく発声されているわけではなく、聞き手もすべての語を正しく認識しているわけではない。

人間は音声をどのように認知し、それを単語として認識しているのであろうか。人間教師は学習者の「正しくない」発声から学習者の意図した発話を認識したり、発話の「正しくない」度合いを評価したりできる。音声に

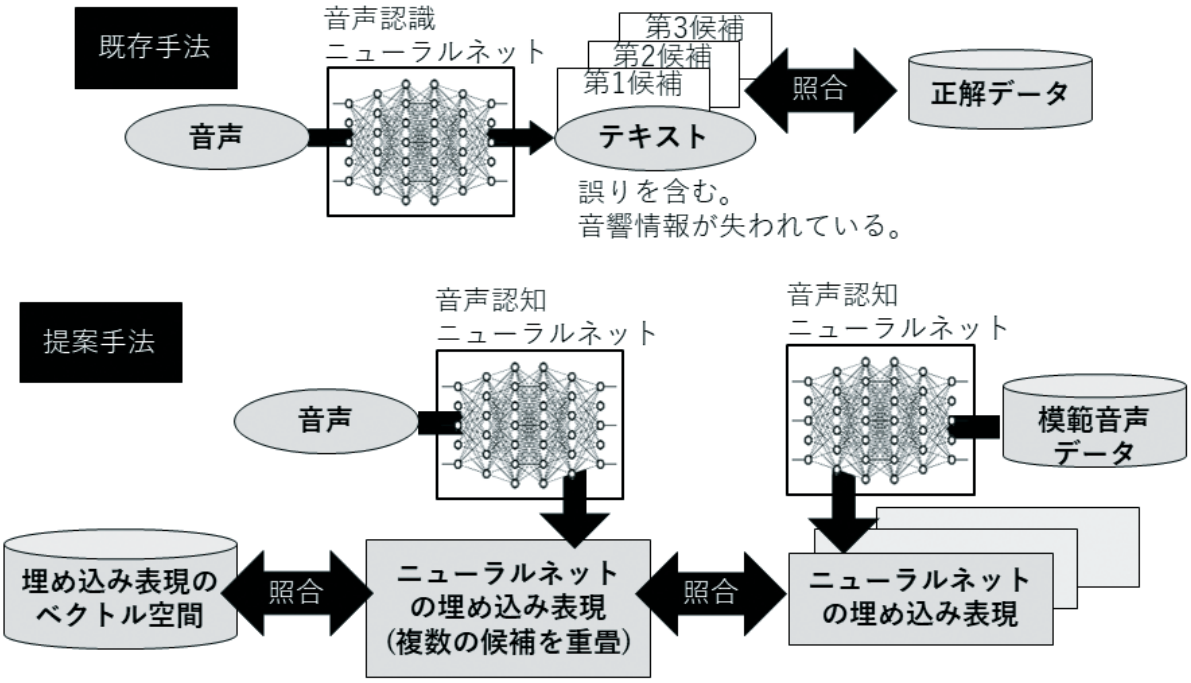

図 1　埋め込み表現を用いた音声の照合

対する人間の認知・認識機能を検討し学習者の発声に適用することにより、人間教師と同じように認識や評価ができるであろう。

2 音声認知の提案（図 1）

音声認識システムの研究開発や、音声認識ステムを用いた英語学習システムの開発・公開が行われてきた。しかしながら、音声認識ステムは本来、目的とするタスクの実現に向けて、音声入力をできるだけ正しくテキスト化するものである。いわゆる巨大 IT 企業も参入しており、アマゾン、マイクロソフト、IBM、グーグルなどが、それぞれ、AWS Comprehend、Microsoft Speech Services、Watson Speech to Text、Google Speech to Text といった音声認識サービスを公開している。音声処理と言語処理を統合した試みとしては、近年の音声認識システムでは、ある時点で文字化された出力がその後の文脈によって修正されるようになってきている。しかしこれは文脈によって音声認識結果を修正するものであり、学習者の個々の発声を評価することには逆行する。

また音声認識のテキスト化の精度には限界がある。人が音声でコミュニケーションを行う場合、相手の発話をその音響信号から完璧に文字列化して正しく認識しているわけではない。たとえば「私は本を買った」という発話が音響的には「私和音多かった」としか認識できないような場合も、我々は的確に理解し、反応できる。音声認識の学習データを増やしてテキスト化の精度を上げるという手法には限界がある。音声を用いた実用システムの実現のためには、完璧に認識された結果のテキストを目指すだけではなく、応答を選択するのに必要十分な表現を得るという考え方が必要となる。語学教育においても同様であり初学者の誤った発声を母語話者の発声であるかのように認識したのでは、学習者の誤りや習熟度を適切には判断できない。

ここでは、音声認識の出力として正確な書きおこしテキストを出力するのではなく、最終的に認知精度が上がるような埋め込み表現を得る。すなわち、音声処理を後段の照合過程に最適化する。これは深層学習における目的関数を認識精度（正しい書きおこしとの差分）ではなく、照合性能を表す指標に置き換えることによって実現できると考えられ、現在主流のニューラルネットワークを用いた end-to-end（E2E）音声認識との親和性が高い。

語学教育への適用を前提とすれば、学習者の不正確な発声を正解に向けて正しく認識するのではなく、学習者の英語レベルを踏まえて、「正しく」認知することが重

要である。語学の初学者など、広いレベルの学習者を対象とし、その能力を自動評価するためには、学習者の不正確な音声を「評価」する機構が必要である。既存の音声認識技術は音声（音響信号）をテキスト（単語列）化するものである。本提案では音声をその特徴を表現した埋め込み表現に変換する。この時点では単語と対応（単語を認識）していないため、音声認識ではなく音声認知と呼ぶ。

ここでは認知の軸として、機械翻訳の評価と同様に流暢度と忠実度の二つを考える。流暢度では発話に含まれる個々の単語を判定・評価するのではなく、全体としての英語らしさを判定する。学習者がどの程度英語的な発声をしているかを評価する。一方、忠実度では学習者が適切な語を発声しているか、あるいは意図した語を発声しているかを評価する。

流暢度においては、英語らしさを判定するわけであるが、ここでは英語としての語や文法の正しさについては判断せず、音の特徴について評価する。発声した音声に含まれるリズム、強弱、イントネーションを評価する。誤解を恐れずに言えば、「タモリの物まね外国語」や「中川家の中国人料理人の中国語」に対して、それぞれの言語での発話としては高い流暢度の値を与えるといった手法である。

近年、精度向上が著しい自然言語処理では入力文の意味や、単語の意味、分野の知識を埋め込み表現で表現し、それらの間の距離計算によって、情報検索をはじめとする様々なサービスを実現している。本提案では、流暢度の判定のために、発話を認識してテキスト化するのではなく、発話を深層学習の埋め込み表現に変換し、埋め込み表現のレベルで認知する。学習者の発話から得られた埋め込み表現と、評点付きの発話データの埋め込み表現と比較することにより、学習者の発話の英語らしさを判定する。一方、忠実度に関しては、既存の音声認識手法を用いて、音声をテキスト化したものに対して適切性を判定する。

3 埋め込み表現による音声認知

音声認識技術の発展に伴い、発音の誤りを指摘するシステム（Mis-pronunciation detection）や発音の良し悪しのスコアを自動的に付けるシステム（Pronunciation scoring）の研究が始まった。教師の発声した単語と学習者の発声した単語の Dynamic Time Warping（DTW）に基づく手法に始まり、音声認識で用いられる Hidden Markov Model（HMM）を用いて母語話者のデータから学習された HMM と学習者発声を照合する自動採点手法が提案された。このような従来型の音声認識技術に基づく手法では、MFCC や韻律 prosody、音響パワー intensity、リズム rhythm などのハンドクラフト特徴量が用いられているが近年は特徴量抽出や系列のモデル化を含めて最適化を行う深層学習を用いた E2E 手法が広く検討されている。

E2E 手法としては、注意機構とマルチエンコーダを用いた手法や自己教師あり学習（Self-supervised Learning: SSL）に基づく手法があるが、以下の理由により、SSL に基づく事前学習モデルを用いている。

（1）音声波形から直接特徴量を抽出することができる。
（2）タスク依存の学習データや学習方法を用いないため、音声波形を表現する汎化性の高い特徴量を抽出できる。
（3）NLP 分野の様々なタスクにおいて優位性が十分に証明されている。

実験では公開されている英語発話データに加え、独自に作成した発声レベル得点付き日本人英語話者データを用いて実験を行っている。

4 応用

英語学習へのさらなる応用として、リピーティング教材や自由対話学習システムへの応用が考えられる。

a) リピーティングへの適用

機械翻訳システムや音声認識システムの技術を用いたリピーティング学習システムが開発され、企業での使用を通しての実証が行われている。リピーティング学習システムでは英語テキスト、そのテキストを音読した模範音声、学習者の音声を利用可能である。既存のシステムではたとえば学習者の音声は汎用の音声認識システムでテキスト化され、英語テキストと比較される。両者の差分は全文一致、単語誤り率、文字誤り率で評価される。しかしながら、音声認識システムが学習者の不正確な音声に十分に対応していないために精度は限定的であった。

ここで提案した手法は、元の英語テキストの入力を想定していないが、試験的に行ったこれまでの実験では流暢度の判定におけるテキストの利用の効果はあまり現れなかった。しかし模範音声の繰り返しを想定するリピーティングやシャドーイングにおいては効果的であるとも考えられ、これらを実証する実験を検討している。少なくとも本手法は音声認識システムによるテキスト化を介さないため、学習者の能力をより的確に判断できる可能性がある。

b) 自由対話への適用

既存の「話す」技能の学習アプリには人工知能との自由対話ができるとするものがある。実際に自由対話を実現するためには学習者の発話を理解する必要があるが、学習者の発話のように不正確な発話であった場合には、現状の応答選択の精度は低く、柔軟な自由対話は実現できていない。

対話における応答生成の手法として、FAQ 検索のチャットボットなどでは、入力と FAQ の質問部分とを照合し、対応する応答部分を生成するものがある。本手法を応用して、模範音声との照合と同様に、埋め込み表現レベルでの照合（検索）を行うことにより、応答生成の精度を向上できよう。音声認識の性能を改善する手法として、複数の認識結果候補を利用する方法があるが、埋め込み表現を用いることにより、埋め込み表現の空間で複数候補の重ね合わせができ、音声認知に適した手法を実現できる。このように複数の候補をまとめてひとつの埋め込み表現にできることが、表層単語ベースの手法にない強みである。

また、音響情報を含む埋め込み表現同士のベクトル空間での距離の照合を行うことにより、文字レベルでの照合ではできなかった判断が可能となる。たとえば、発話が不明瞭あるいは不正確であっても、その音声から予想される語がひとつしかなければ、特に母語話者であれば、相手がその語を発話したものと認識できる。音声入力から得られた埋め込み表現の照合において、ベクトル空間上での語の埋め込み表現の分布を考慮することにより、このような照合が可能となる。

5 おわりに

ここで提案したシステムの目的は音声認識システムではなく、学習者の音声の評価である。音声認識システムは正しく発声された音を音声認識し、何らかのタスクを行う。学習システムは正しくない音声の正しくない程度を判定することが必要である。実用の観点に立てば、誤った発音であっても誤解されない発音であれば、急いで訂正する必要はないという場合もあろう。埋め込み表現での照合により、誤解される誤りと誤解されない誤りを識別できるかもしれない。正しい発音、間違った発音、理解できる発音を弁別できる手法を開発することが今後の課題である。