HMM に基づく日本人英語音声合成における 中学生徒の英語音声を用いた評価*

◎ 高道 慎之介, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲 (奈良先端大)

1 はじめに

英語等の第二言語学習における早期の言語経験は、当該言語の知覚に強く影響する [1]。英語のスピーキングにおける韻律的特徴は、発話の流暢性を決定する強い要因 [2] であり、特に日本語母語話者は、日本語のモーラ拍と英語の強勢拍の違い [3] を考慮した学習を必要とする。近年の CALL 関連技術の発展により英語学習の自動化・個別化が進む中、我々は、学習者の発話した英語音声を利用して、発話のリファレンスとなる音声を合成する方法 [4] を提案している。具体的には、任意のテキストから音声を合成する手法の一つである HMM 音声合成 [5] を用いて、学習者による英語音声の韻律的特徴を補正することで、学習者自身の声質で自然な英語音声を合成する。

本稿では、英語学習経験の比較的浅い中学生徒を学習者と想定して、合成音声を評価する。実験的評価の結果、英語学習年数 1〜9 年の全生徒に対して、提案している合成法による合成音声の自然性が、通常の HMM 学習法と比較して改善することを示す。

2 学習者の音声を補正して構築する HMM 音声合成器

HMM 音声合成 [5] の学習部では、事前収録した音声の音声特徴量（パワー、スペクトル、基本周波数、及び非周期成分）、及び継続長を、対応するテキストのコンテキスト依存 HMM でモデル化する。生成時には、入力テキストに対応する HMM から音声特徴量を生成し、任意のテキストに対する音声を合成する。また、ある話者の音声特徴量をモデル化する HMM を、目標話者の音声特徴量を用いてモデル適応 [6] することで、目標話者の HMM を構築可能である。

[4] では、モデル適応技術を利用して学習者の声質で自然な英語音声を合成する（Fig. 1）。学習者の音声特徴量を用いて、英語母語話者の音声特徴量で事前学習した HMM の韻律を保持しながら適応することで、英語母語話者の韻律を反映した HMM を構築可能である。

3 テキストコーパスと音声合成器の自動構築

音声合成器の構築手順を Fig. 2 に示す。まず、用意したテキストから学習者に読み上げさせるテキストコーパスを作成し、音声を収録する。次に、HMM 構築に使用する音声特徴量を選択し、HMM を構築する。

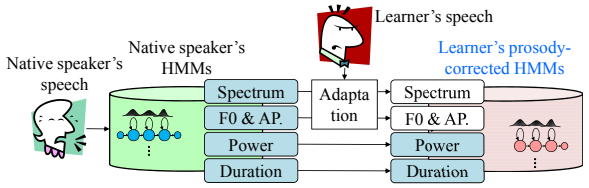

Fig. 1 学習者の音声を補正して構築する HMM 音声合成器（F_0 と AP. はそれぞれ、基本周波数と非周期成分を示す。）

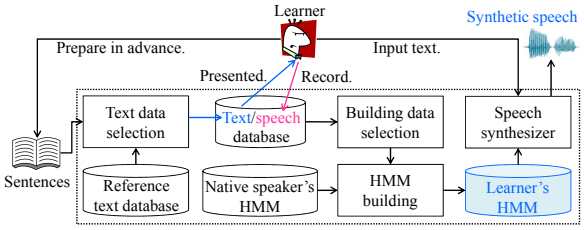

Fig. 2 学習者の発話を利用して音声合成器を構築するシステム

3.1 テキストコーパスの構築

学習者に提示する英語テキストコーパスを構築する。頑健な音声合成器の構築には、音素バランスのとれたテキストコーパスの使用が望ましい。しかしながら、既存の英語コーパスは英語母語話者 [7] 又は大学生・大学院生 [8] の言語経験（例えば、単語・フレーズのボキャブラリ）を対象とするため、言語経験の浅い学習者に対して不適切である。また、学習者の慣れ親しんだテキストの使用により、より高品質な音声合成器を構築可能である。

そこで本稿では、学習者の履修する英語授業で使用される教科書を元にテキストコーパスを自動構築する。まず、教科書の文章に対して構文解析を行い音素を決定する。ただし、構文解析のエラーを緩和するため、テキストに含まれる固有名詞（登場するキャラクタの名前など）の一部を対応する代名詞に置き換える。次に、その文章から固定文数の文セットを複数回サンプリングし、既存の音素バランス文との音素ヒストグラム距離が最小となる文セットを、テキストコーパスとする。音声収録時には、文セットを順に一文ずつ学習者に提示して音声を収録する。

* An Evaluation of HMM-Based English-Read-by-Japanese Speech Synthesis Using English Speech Read by Japanese Junior High School Students by TAKAMICHI, Shinnosuke, TODA, Tomoki, SAKTI, Sakriani, NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

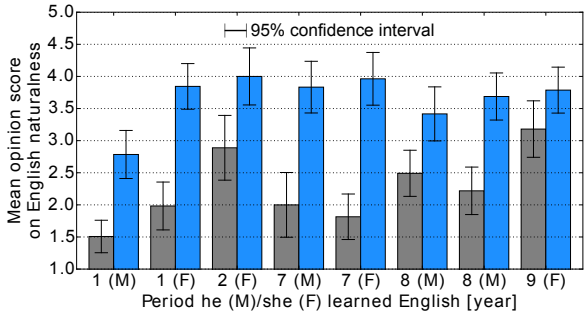

Fig. 3　合成音声の英語の自然性に関する 5 段階 MOS 評価の結果（灰色は “w/o mod”、青色は “w/ mod” の評価結果を示す。）

3.2　音声合成器の構築

全文の収録が完了した後、収録音声の音声特徴量を用いて音声合成器を学習する。ただし、発音等の誤りが収録音声に含まれる場合、HMM の学習精度が低下する傾向にある。そこで収録音声のうち、HMM の構築時に相対的に尤度の低い音声を除いて、HMM を構築する。

4　実験的評価

4.1　実験条件

学習者は、中学 1 年生の男女各 4 名であり、小学校入学（7〜9 年）または中学校入学（1〜2 年）前後から英語学習を開始した生徒である。英語母語話者は、ARCTIC speech database [7] に含まれる男女各 1 名とする。音声分析合成器、音声特徴量、音響モデル、及び英語母語話者の HMM の学習データは [4] と同様だが、スペクトル特徴量には 39 次までのメルケプストラム係数を使用し、生成時には、変調スペクトルを考慮したポストフィルタ [9, 10] を使用する。テキストコーパスは、z 会 New TREASURE Stage 1 [11] の本文約 900 文から構築した 500 文とする。音素バランスのリファレンスとする英語コーパスは、ARCTIC speech database とする。

学習者の負担を考慮し、収録時間を約 2 時間（休憩を含む）に限定し、時間内で可能な限りの文数を収録した。8 名中 3 名が 500 文全て、最少で 400 文の収録を完了した。400 文のうち、50 文を評価データとし、（1）残る 350 文を用いて、通常の HMM 学習で構築した音声合成器（“w/o mod”）と（2）350 文のうち尤度の低い文を除いた 100 文を用いて、2 節の手法で構築した音声合成器（“w/ mod”）による合成音声を比較する。評価は、大学院生 10 名による、合成音声の英語の自然性に関する 5 段階 MOS 評価である。評価者には、全学習者の合成音声をランダムに提示した。

4.2　評価結果

評価結果を Fig. 3 に示す。全学習者の合成音声において、補正による自然性改善効果が得られることが確認できる。次に、補正後のスコアに対する、学習年数の依存性を考察する。学習年数 9 年の学習者の補正後スコアと比較して、多くの学習者が同等のスコアを獲得しており、学習年数への依存性が小さいことが確認できる。しかしながら、学習年数 1 年の学習者のスコアが大きく低下している事が分かる。これは、この学習者による収録音声のほとんどで発音間違いが認められ、HMM の学習精度が低下したためだと思われる。

5　まとめ

本稿では、言語経験の比較的浅い中学生徒を学習者と想定して、提案している合成法による合成音声を評価した。実験的評価の結果、英語学習年数 1〜9 年の全生徒に対して、合成音声の自然性が改善することを示した。

今後は、学習年数に関する詳細な分析、学習効果の確認、シャドーイング法の利用による発音誤り緩和、発音誤りを考慮した HMM 構築及びテキストコーパスと HMM のオンライン構築を行う。また、本稿では手作業により固有表現を置換したが、固有表現抽出器（例えば、[12]）を利用した自動置換も調査する。

謝辞　本研究の一部は，JSPS 科研費 26・10354 の助成を受け実施した．また、本研究の音声収録は奈良学園登美ヶ丘中学校・高等学校の川口 典子教諭の協力の下で実施した。氏に感謝する。

参考文献

[1] P. K. Kuhl, *Nature Reviews Neuroscience*, Vol. 5, pp. 831–843, 2004.
[2] R. L. Allington, Fluency: The neglected reading goal, *Reading Teacher*, 36 (6), pp. 556–561, 1983.
[3] E. Grabe *et. al., Laboratory Phonology 7*, pp. 515–546, 2002.
[4] 大島 他, 情処研報, Vol. 2015-SLP-105, No. 3, pp. 1–6, 2015.
[5] K. Tokuda *et. al., Proc. of IEEE*, Vol. 3, pp. 1315–1318, 2013.
[6] J. Yamagishi *et. al., IEICE Trans. Inf. and Syst.*, Vol. 90, No. 2, pp. 533–543, 2007.
[7] J. Kominek *et al., Tech. Rep. CMU-LTI-03-177*, 2003.
[8] N. Minematsu *et al., Proc. INTERSPEECH*, pp. 1481–1484, 2011.
[9] S. Takamichi *et al., Proc. ICASSP*, pp. 290–294, 2014.
[10] S. Takamichi *et al., Proc. GlobalSIP*, pp. 710–714, 2014.
[11] New TREASURE English Series Stage 1, *z* 会出版編集部, 2009.
[12] Stanford Named Entity Recognizer (NER), `http://nlp.stanford.edu/software/CRF-NER.shtml`