話者適応に基づく日本人英語発話の認識、合成 *

☆上田怜奈, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

英語を学習する日本人にとって、英語の読み書きに比べ会話は苦手意識を感じることが多くある。その理由としては、発話困難な慣れない発音が英語には多く存在すること、日本語に比べて英語は発話速度が速いこと、アクセント位置がわからないなどの要因が挙げられる。これらの問題点はネイティブの発話を模倣し反復学習することで改善が見込まれるが、学習者にとって大きな負担となってしまう。そこで本研究では、日本人英語発話 (ERJ: English Read by Japanese) を声の特徴は残しつつネイティブのような発話に変換することを目的とし、音声認識、合成を用いた変換を試みる。認識部では日本人英語発話データを使い学習した HMM(Hidden Markov Model) 音声認識モデルを作成、また合成部では、複数人のネイティブ話者発話を学習データとし、同じ日本人話者発話を適応データとし HMM 音声合成モデルを作成した。非母国語の音声認識は発音誤りなどが原因で認識誤りの増加が想定される。そのため本研究では、複数認識候補に対して合成音声を用いて学習した HMM で認識を行うことで最終的な一つの出力を得るという手法をとっている。

2 HMM 音声合成における話者適応

HMM音声合成において、大量の不特定話者データを学習データとし、平均声モデル [1] を作成すれば、少量の適応話者のデータを用いて、適応モデルを作ることができる。適応法としては MLLR [2] が知られているが、ここでは CSMAPLR [3] を用いる。平均声モデルの i 番目の確率分布の平均ベクトルを μ_i、共分散行列を Σ_i とすると、適応モデルの平均ベクトル $\overline{\mu_i}$、共分散行列 $\overline{\Sigma_i}$ は以下のように求められる。

$$\overline{\mu_i} = H\mu_i + b \tag{1}$$

$$\overline{\Sigma_i} = H\Sigma_i H^T \tag{2}$$

ここで、H は適応行列、b はバイアスベクトルであり、回帰クラス木のノードごとに推定される。このように線形回帰的に適応したのち、事後確率を最大化するようにパラメータを更新していく。

3 日本人英語発話の認識、合成

本研究ではＨＭＭ音声認識とＨＭＭ音声合成の適応技術を組み合わせた手法を提案する。提案法の概要を Fig. 1 に示す。

今回、一つ目の認識部において、学習データは日本人英語発話を使用し、ラベリングには英語音素を使用している。非母国語話者のデータに対して、母国語英語音素体系を使用しているため、発音誤りによるミスマッチが起こってしまう。そこで、複数認識候補に対して合成音声を用いて学習した HMM で認識を再度行い最終的に一つの合成音を選び出し出力とする。Fig. 1 のように一回目の認識では ERJ の肉声を使い学習した HMM を用いる。そして複数認識候補からなる辞書を作成し、その辞書を二回目の認識で用いる。二回目の認識では学習データは ERJ を適応させた合成音を用いる。このとき入力音声は一回目の認識での入力と同じものを使い辞書内の候補単語の中から最終的に一つの合成音の出力を得る。

4 評価実験

4.1 実験条件

一回目の認識部の学習､テストデータには 20 リスト、合計 1,000 単語から成る PB(phonetically balance) word list [5] を使用し日本人男性一名の英語発話を収録、学習データは 1,000 単語 (11~20 リスト、一回目発話)、テストデータは 250 単語 (11~15 リスト、二回目発話) を使用した。サンプリング周波数は 16kHz、フレームシフトは 5ms、ＨＭＭ状態数は 5 状態、特徴量には 12MFCC+12ΔMFCC+12power を使用した。合成部の学習データには CMU_ARCTIC 音声データベース中の男性 2 名、女性 2 名それぞれ 1,132 文を使用し、適応データについては認識部と同じ男性の英語発話 50 文を使用した。サンプリング周波数は 48kHz、音声分析合成系には STRAIGHT [4] を使用し、対数パワー、MFCC24 次元、対数 F0、5 周波数帯域の ap 成分、それらの動的特徴量 (Δ,ΔΔ) を特徴量として使用した。二回目の認識部においては、学習データは合

*Recognition and Synthesis for English-Read-By-Japanese based on speaker adaptation, by Reina Ueda, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

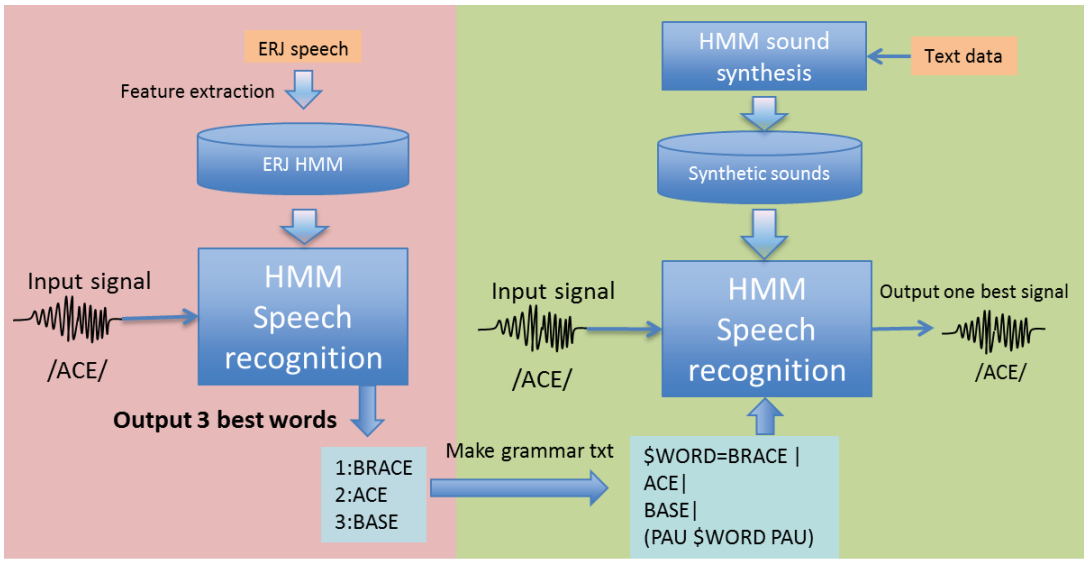

Fig. 1　日本人英語発話の認識、合成

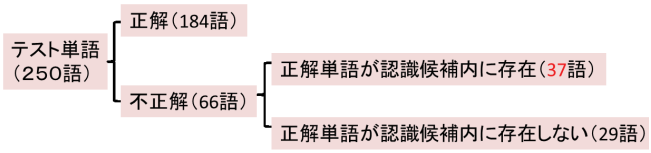

Fig. 2　ERJ HMM による認識結果

成音 1,000 単語 (11~20 リスト)、テストデータは一回目の認識部の中で、認識に失敗しかつ正解単語が上位 3 位以内に入っているものを対象とする。その他の条件は一回目の認識部と同じであるとする。

4.2　実験

一回目の認識部は 1,000 単語タスクで行った。その後認識候補を最大 3 単語とし、誤ったものを二つ目の HMM を用いて認識する。一回目の認識での認識結果を Fig. 2 に示す。今回合成音を使った認識実験では不正解単語の中で、認識候補内に正解単語が存在するものを対象とした。HMM 混合数を 1~4 に変化させたときの認識結果を Fig. 3 に示す。

Fig. 3 より、最も正解率が高いのは混合数が 2 のときで 60% である。そして、最も正解率が悪かったのは混合数が 1 のときで 43% であった。これらの正解率をより向上させるには、肉声音と適応した合成音との距離を更に小さくする必要があり、改善策としてはもう少し適応データを増やす、合成の際の学習データの話者を男性のみにするなどが挙げられる。

5　おわりに

今回は適応した合成音声を用いて HMM を学習した。合成音は大量に作成することができるので、精度の高い認識器が構築できる可能性がある。今後はこ

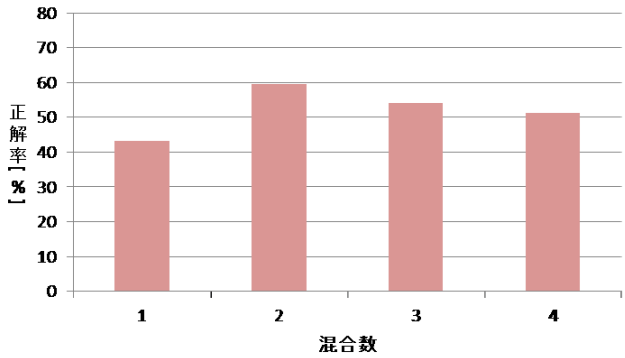

Fig. 3　合成音 HMM における単語正解率の混合数による変化

の合成音の利点を応用した研究を進めたいと考えている。

参考文献

[1] J. Yamagishi and T. Kogayashi,“Average-voice-based speech synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training,” IEICE TRANSACTIONS on Information and Systems, Vol. E90-D, No. 2, pp. 533-543, 2007.

[2] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, Vol. 9, No. 2, pp. 171-185, 1995.

[3] Yamagishi et al., Proc. of IEEE, Vol. 17, No. 1, pp. 66-83, 2009.

[4] H. Kawahara et al., Speech Commun. Vol. 27, No. 3-4, pp. 187-207, 1999.

[5] Egan, J.P. “Articulation testing methods,” The Laryngoscope, Vol. 58, No. 9, pp. 955-981, 1948.