日本人英語音声合成における話者性を保持した韻律補正法と英語習熟度が与える影響*

☆大島悠司, 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

声質変換 [1] や HMM 音声合成 [2] を用いた日英間クロスリンガル音声合成は，同言語間の場合と比較して，話者性の低い英語音声を合成する傾向にある．これに対し我々は，日本人英語（ERJ: English Read by Japanese）[3] の利用，また，日本人英語の韻律誤りに対する韻律補正法により，話者性を強く反映しつつ自然性を改善する手法を提案している [4]．しかしながら，日本人英語の自然性低下の要因である音韻誤りが考慮されていないため，得られる自然性改善効果は限定される．また，日本人話者の英語習熟度に対する補正法の影響も十分に調査されていない．

本稿では，音韻補正法として，話者性に対する影響が小さいと予想される無声子音スペクトルに対する置換処理を提案するとともに，英語習熟度に対する韻律・音韻補正法の効果の違いを調査する．実験的評価結果により，韻律補正法は英語習熟度に依存せず，話者性を保持しつつ自然性を改善できることを示す．また，音韻補正法により，英語習熟度の低い話者の自然性を改善できることを示す．

2 モデル適応による韻律補正法

Fig. 1 に韻律補正法の手順を示す．目標とする日本語母語話者の話者性を反映した英語音声合成用 HMM を構築するために，目標話者の日本人英語音声を用いて， 英語母語話者の HMM を適応する．日本人英語音声の自然性を劣化させる要因とし て，継続長およびパワーに着目し，状態継続長と対数パワー以外に対するモデルパラメータのみを適応することで，英語母語話者の韻律を考慮した日本人英語の HMM を構築する．

3 無声子音スペクトル置換による音韻補正法

Fig. 2 に音韻補正法の手順を示す．提案法では，英語母語話者のスペクトルパラメータを部分的に使用することで，日本人英語の音韻を補正する．その際に，話者性を保持するために，話者依存性が小さいと予想される無声子音に対してのみ，補正処理を適用する．

まず，英語母語話者の HMM と韻律補正された日本語母語話者の HMM から，それぞれ音声パラメータを生成する．パラメータ生成時には，継続長モデルの尤度最大化により状態継続長を決定したのち，HMM の尤度最大化によりパラメータを生成する [5]．ここで，各 HMM は同一の継続長モデルを有するため，生成パラメータは時間的に対応付けられていることに注意する．次に，日本語母語話者のスペクトルパラメータ系列の中で，無声子音に対応するフレームのみを，英語母語話者のスペクトルパラメータに置換する．その際に，置換後のスペクトルと元の有声／無声情報の不一致により生じる音質劣化を回避するために，無声子音のフレームにおける英語母語話者の F_0 が有声である場合，当該フレームを置換しない．

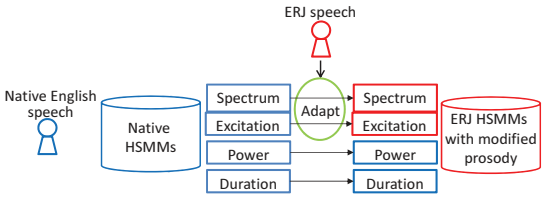

Fig. 1 モデル適応による韻律補正法

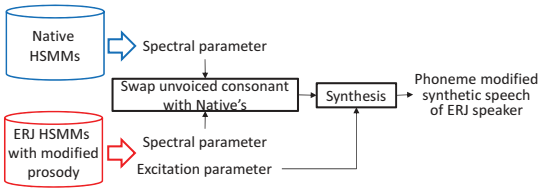

Fig. 2 無声子音スペクトル置換による音韻補正法

4 実験的評価

4.1 実験条件

学習データとして，CMU ARCTIC 音声データベース [6] 中の英語母語話者の男女各 1 名による A セット 593 文を用いる．適応データには，日本人学生による読み上げ英語音声データベース [3] 中の最高（“High”）もしくは最低（“Low”）英語習熟度スコアを持つ男女計 4 名による TIMIT[7] 60 文，及び，留学経験の無い男子大学院生（“Monolingual”）と留学経験のある男子大学生（“Bilingual”）による CMU ARCTIC 音声データベース A セット 60 文を用いる．音声分析合成系には STRAIGHT[8] を使用し，対数パワー，1 次から 24 次のメルケプストラム係数，対数 F_0，5 周波数帯域における平均非周期成分，及びそれらの 1 次と 2 次の動的特徴量を音声パラメータとして用いる．HMM は 5 状態 left-to-right 型の HSMM[9] とし，対数パワーとメルケプストラム係数は同一ストリームで学習する．モデル適応は CSMAPLR+MAP[10] を利用し，回帰行列には静的特徴量及び，1 次と 2 次の動的特徴量に対応したブロック対角行列を用いる．ただし適応時には，適応データの話者と同じ性別の英語母語話者のデータで学習された HMM を用いる．被験者は，英語母語話者 6 名である。

以下の手法による合成音声を用いて，話者性及び自然性に関する主観評価を実施する．

- HMM+VC： 英語母語話者の話者依存 HSMM の出力音声パラメータに対して，GMM 声質変換を適用して得られる合成音声 [11]
- Adapt: 全モデルパラメータを適応させた適応 HSMM による合成音声
- Dur.+Pow.： 提案法として対数パワーと状態継続長以外を適応させた適応 HSMM による合成音声
- Dur.+Pow.+UVC: 提案法として Dur.+Pow に音韻補正を施した合成音声
- Native： 英語母語話者の話者依存 HSMM による合成音声

評価結果は，英語習熟度（“High”と “Low”）毎に計算する．ただし，“Monolingual”と “Bilingual”はそ

*Prosody Correction Preserving Speaker Individuality in English-Read-By-Japanese Speech Synthesis and Effects of English Proficiency Level, by OSHIMA, Yuji, TAKAMICHI, Shinnosuke, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

れぞれ，“Low”と “High”に属するものとする．

韻律補正法を用いた話者性の評価では，日本語母語話者の日本人英語分析合成音声をリファレンスとした 5 段階 DMOS（Degradation Mean Opinion Score）評価を実施する．評価する手法は，“HMM+VC”，“Adapt”，“Dur.+Pow.”の 3 つである．自然性の評価では，英語音声の自然性に関する 5 段階 MOS（Mean Opinion Score）評価を実施する．評価する手法は，“HMM+VC”，“Adapt”，“Dur.+Pow.”，“Native”の 4 つである．ただし，“Monolingual”と “Bilingua”の適応データについては，[4] で評価しているため，本稿では使用しない．

音韻補正法を用いた話者性の評価では，日本語母語話者の日本人英語分析合成音声をリファレンスとしたプリファレンステスト (XAB テスト) を実施する．評価する手法は、“Dur.+Pow.”、“Dur.+Pow.+UVC”の 2 つである。自然性の評価では，英語音声の自然性に関するプリファレンステスト (AB テスト) を実施する．評価手法は，“Dur.+Pow.”，“Dur.+Pow.+UVC”，“Native”の 3 つである。

4.2 実験結果

4.2.1 韻律補正法の効果

図 3，図 4 にそれぞれ，韻律補正法に対する話者性と自然性に関する主観評価結果を示す．GMM 声質変換を利用した手法 “HMM+VC”に着目すると，“Low”において，全モデルパラメータを適応した手法 “Adapt”と比較して，話者性が大きく劣化する傾向が見られる．“High”においても，劣化の程度は小さくなるが，同様の傾向が見られる．提案法の継続長およびパワーを補正した “Dur.+Pow.”に関しては，英語習熟度に関係なく “Adapt”と同等の話者性を保っていることが分かる．一方，自然性に関しては，“Dur.+Pow.”は “HMM+VC”と同等となる．“Adapt”と比較すると，自然性が改善されており，“Low”に対する改善効果が著しい．

以上の結果から，英語習熟度の高低に関わらず，提案法は頑健に動作することを確認でき，継続長及びパワー補正により，日本人英語の話者性を保持しつつ，自然性の高い英語音声を合成できることが分かる．

4.2.2 音韻補正法の効果

図 5，図 6 にそれぞれ，音韻補正法に対する話者性と自然性に関する主観評価結果を示す．“Low”に対して，“Dur.+Pow.+UVC”は “Dur.+Pow.”と比較して，話者性を同等程度に保持しつつ自然性を改善できることが分かる．また，“High”に対しては，“Dur.+Pow.+UVC”は “Dur.+Pow.”と同等の自然性および話者性を保持できることが分かる．

5 おわりに

本稿では，モデル適応による日本人英語合成音声の韻律補正法を英語習熟度の異なる複数話者に適用し，話者性と自然性に関する主観評価を実施した．主観評価結果より，本提案法は英語習熟度に依存せず，話者性を保持しつつ自然性を改善できることを示した．また，部分的な音韻補正法の提案も行い，英語習熟度の低い話者に対して，自然性を改善できることを示した．今後は，明瞭度に関する評価および目標話者毎の音韻誤りに基づく最適な補正法を検討する必要がある．

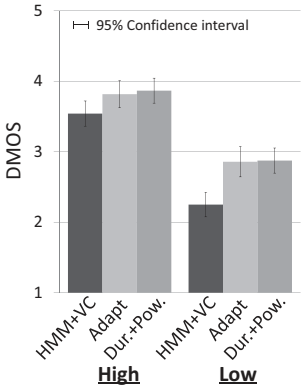

Fig. 3 韻律補正法に対する話者性に関する主観評価結果

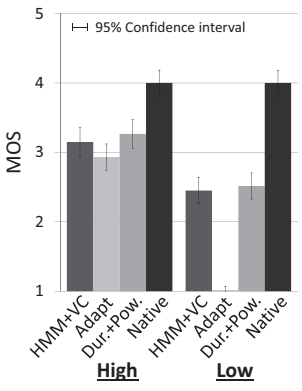

Fig. 4 韻律補正法に対する自然性に関する主観評価結果

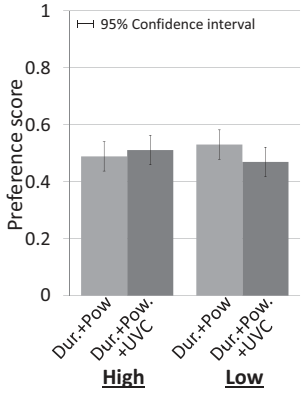

Fig. 5 音韻補正法に対する話者性に関する主観評価結果

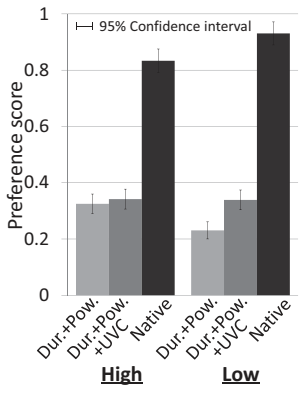

Fig. 6 音韻補正法に対する自然性に関する主観評価結果

謝辞 本研究の一部は，(独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」および JSPS 科研費 26280060 の助成を受け実施したものである．

参考文献

[1] T. Toda *et al., IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007.
[2] K. Tokuda *et al., Proc. IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.
[3] N. Minematsu *et al., Proc. ICA*, Vol. 1, pp. 557–560, 2004.
[4] 大島 他, 信学技報, Vol. 114, No. 365, SP2014-111, pp. 63–68, 2014.
[5] K. Tokuda *et al., Proc. ICASSP*, Vol. 3, pp. 1315–1318, 2000.
[6] J. Kominek *et al., Tech Report.*, CMU-LTI-03-177, 2003.
[7] J. Garofolo *et al., Tech Report.*, NISTIR 4930, NIST, 1993.
[8] J. Yamagishi *IEEE Trans. ASLP*, Vol. 17, No. 6, pp. 1208–1230, 2009.
[9] H. Kawahara *et al., Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[10] H. Zen *IEICE Trans. Inf. and Syst.*, Vol. 90, No. 5, pp. 825–834, 2007.
[11] N. Hattori *et al., Proc. INTERSPEECH*, pp. 2769–2772, 2011.