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*

☆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

クロスリンガル音声合成は，ある言語の発話者の話者性を異言語の合成音声において反映させる技術であり，特に日本では，日英間における合成技術の需要が高い．これまでに，声質変換技術 [1] や HMM 音声合成 [2] における話者適応技術 [3] などにおいて，英語を母国語とする話者の音声に対して，バイリンガル音声や日本語音声といった自然性の高い音声データを活用した話者変換処理を施す手法が広く研究されている．これらの手法で合成される英語音声は，比較的自然性は高いものの，話者性は同言語間における合成音声と比較して大きく劣化する傾向にある．これに対して，できる限り話者性を強く反映させた英語音声を合成するためには，日本人英語（ERJ: English Read by Japanese）[4] を直接利用する合成法が有効であると考えられる．しかしながら，日本人英語の不自然な韻律（リズム，抑揚）及び発音誤り [5] により，得られる合成音声の自然性は低下する．

本稿では，日本人英語音声の話者性を保持しながら自然性を改善する手法として，HMM 音声合成におけるモデル適応技術を用いた日本人英語の韻律補正法を提案する．提案法では，英語母語話者の HMM を日本人英語を用いて適応することで，話者性を強く反映した英語 HMM を構築する．その際に，適応する HMM パラメータを限定することで，日本人英語における韻律の不自然性を低減する．実験的評価から，提案法により，日本語母語話者の話者性を保持しつつ自然性を改善できることを示す．

2 HMM 音声合成における適応技術

HMM 音声合成では，モデル適応技術を用いることで，ある話者の HMM から目標話者の HMM を構築することができる．本稿では，モデル適応法の一つとして，CSMAPLR[6] を用いる．適応元モデルの i 番目の分布（状態出力分布や継続長分布）における平均ベクトルを $\boldsymbol{\mu}_i$，共分散行列を $\boldsymbol{\Sigma}_i$ としたとき，適応後の平均ベクトル $\hat{\boldsymbol{\mu}}_i$ と共分散行列 $\hat{\boldsymbol{\Sigma}}_i$ は次式で計算される．

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{A}\boldsymbol{\mu}_i + \boldsymbol{b} \tag{1}$$

$$\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{A}\boldsymbol{\Sigma}_i\boldsymbol{A}^T \tag{2}$$

ここで，適応行列 $\boldsymbol{A}$ とバイアスベクトル $\boldsymbol{b}$ は回帰パラメータであり，複数の分布が属する回帰クラスごとに推定される．CSMAPLR では，回帰クラスを木構造で表現し，親ノードにおける推定結果を子ノードにおける事前分布として伝搬させながら，適応データが与えられた際の回帰パラメータの最大事後確率推定を行う．なお，HMM 音声合成では，スペクトルパラメータ，音源パラメータ，状態継続長が正規分布でモデル化されており，それら全てに対して適応処理が行われる．これにより，声質などの分節的特徴のみでなく韻律的特徴も適応することが可能となる．

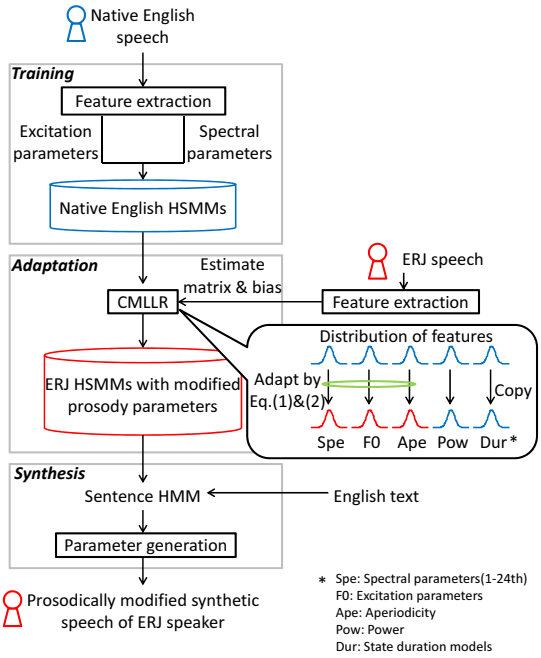

Fig. 1 英語母語話者の英語音声を用いた韻律補正処理の流れ

3 韻律補正を考慮した日本人英語音声に対する適応法

目標とする日本語母語話者の話者性を強く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．その際に，通常の適応処理のように，全 HMM パラメータを適応すると，日本人英語音声の特徴を全てモデル化した HMM となるため，得られる合成音声は日本人英語音声と同様，不自然なものとなる．本稿では，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，これらの音声パラメータに対しては英語母語話者 HMM の特徴を活用することで，日本人英語の韻律を補正する手法を提案する．

Fig. 1 に提案法の手順を示す．まず，英語母語話者の英語音声を用いて，英語母語話者の話者依存 HMM を学習する．観測データとして用いる音声パラメータは，対数パワー，スペクトルパラメータ，音源パラメータであり，各パラメータに対する出力分布と状態継続長分布が得られる．次に，目標話者の日本人英語音声を適応データとして，HMM の適応処理を行う．この時，対数パワーと状態継続長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．得られた適応 HMM から，目標話者の英語音声を合成する．

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

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 [7] 中の男性英語母語話者による 1000 文を使用し，評価データとして残りの 132 文を使用する．サンプリング周波数は 16 kHz とする．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．

目標話者とする日本語母語話者は 20 代男性 1 名とし，上述の男性英語母語話者の英語音声をリファレンスとしたシャドーイング法により，日本人英語音声を収録する．適応データは学習データと同じ 1000 文とし，回帰行列にはブロック対角行列を用いる．

提案法による韻律補正の効果を評価するために，次の HMM から生成した音声を用いて，自然性及び話者性に関する評価を行う．

- Native： 英語母語話者の話者依存 HMM
- ERJ： 日本語母語話者の話者依存 HMM（1000 文を用いて学習）
- VC+HMM：従来法 [9] に基づき，英語母語話者の英語音声から日本語母語話者の日本人英語音声へと声質変換を行い，得られた変換音声を用いて学習された話者依存 HMM
- Adapt： 全モデルパラメータを適応させた適応 HMM
- Dur.： 状態継続長以外を適応させた適応 HMM
- Dur.+Pow.： 対数パワーと状態継続長以外を適応させた適応 HMM

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）評価を行う．話者性の評価では，日本語母語話者の英語分析合成音声をリファレンスとして，5 段階 DMOS（Degradation MOS）評価を行う．被験者は，日本語母語話者 6 名である．

4.2 実験結果

Fig. 2 に自然性に関する評価結果を示し，Fig. 3 に話者性に関する評価結果を示す．“ERJ” と比較し，全モデルパラメータを適応した “Adapt” は自然性および話者性が劣化しているが，提案法による継続長補正 “Dur.” を行うことで，“ERJ” と同等の話者性を保持しつつ，自然性を改善できることが分かる．なお，継続長およびパワーを補正した “Dur.+Pow.” に関しては，自然性は “Dur.” と同等であり，話者性が “ERJ” よりも劣化してしまうことから，継続長のみを補正する “Dur.” の方が有効であることが分かる．なお，従来法である “VC+HMM” に関しては，自然性および話者性共に “ERJ” よりも劣化することが分かる．

以上の結果から，提案法により継続長補正を行うことで，日本人英語の話者性を保持しつつ，より自然性の高い英語音声を合成できることが分かる．

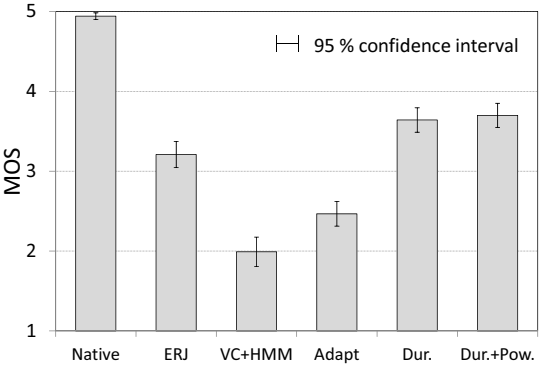

Fig. 2 自然性に関する主観評価結果

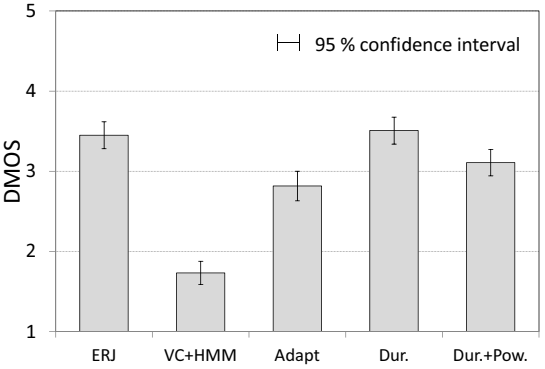

Fig. 3 話者性に関する主観評価結果

5 おわりに

本稿では，日本人話者の話者性を保持した英語音声合成手法として，日本人英語音声を用いた HMM 音声合成モデル適応における韻律補正法を提案した．実験的評価結果から，継続長の補正により，日本語母語話者の話者性を保持しつつ，自然性を改善できることを示した．今後は，評価者の母国語の影響 [10]，シャドーイング法による収録の影響 [11]，日本語母語話者の英語習熟度の影響 [4] について調査する．

謝辞 本研究の一部は，JSPS 科研費 24240032 の助成を受け実施したものである．

参考文献

[1] M. Abe *et al., Proc. ICASSP88*, Vol. 1, pp. 655–658, 1988.
[2] K. Tokuda *et al., Proc. of IEEE*, Vol. J96-D, No. 7, pp. 1549–1565, 2013.
[3] C.J. Leggetter and P.C. Woodland, *Computer Speech and Language.,* Vol. 9, No. 2, pp. 171–185, 1995.
[4] N. Minematsu *et al., Educational Technology.*, Vol. 27, No. 3, pp. 259–272, 2004.
[5] S. Kohmoto *et al.,* “Applied English Phonology -TEACHING OF ENGLISH PRONUNCIATION TO THE NATIVE JAPANESE SPEAKER-,” TANAKA Press, 1965.
[6] J. Yamagishi *et al., Proc. of IEEE*, Vol. 17, No. 1, pp. 66–83, 2009.
[7] J. Kominek and A. Black, *Tech Report.*, CMU-LTI-03-177, 2003.
[8] H. Kawahara *et al., Speech Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.
[9] 加島 他, 情処研報, Vol. 2008-MUS-74, No. 1, pp. 121–126, 2008.
[10] K. Tokuda and A. Black, 音響誌, Vol. 62, No. 6, pp. 466–472, 2006.
[11] D. Luo *et al., Proc. of SLaTE.*, CD-ROM, 2009.