一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2014-112(2014-12)

HMMを用いた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†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630–0192 奈良県生駒市高山町 8916–5
E-mail: †{oshima.yuji.ok0,shinnosuke-t,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

あらまし　日本語母語話者の話者性を反映した英語音声合成用音響モデルを構築するためには，日本人英語（ERJ: English Read by Japanese）に含まれる話者性を最大限に活用する枠組みが有効であると考えられる．しかしながら，日本人英語の利用により，言語体系の違いに起因する日英間の韻律の相違や発音誤りの影響を強く受けるため，自然性が大きく劣化するという問題が生じる．これに対し本報告では，主に韻律の相違に着目し，英語母語話者の韻律を用いて日本人英語音声の韻律を補正することで，日本語母語話者の話者性を保持したまま，より自然な英語音声を合成する手法を提案する．HMM 音声合成におけるモデル適応法を応用することで，英語母語話者の継続長とパワーの特徴を，日本人英語の合成音声に反映させる．実験的評価結果から，提案法により，日本語母語話者の話者性を保持しつつ，より自然性の高い英語音声を合成できることを示す．

キーワード　日本人英語，HMM 音声合成，韻律補正，話者性，モデル適応

Prosody Correction Preserving Speaker Individuality in English-Read-By-Japanese Speech Synthesis Based on HMM

Yuji OSHIMA†, Shinnosuke TAKAMICHI†, Tomoki TODA†, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology,
Tatayama-cho 8916–5, Ikoma, Nara, 630–0192 Japan
E-mail: †{oshima.yuji.ok0,shinnosuke-t,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

Abstract　To build an English acoustic model that well captures speaker individuality of each Japanese speaker, a framework using English-Read-by-Japanese (ERJ) voices is effective as it enables to directly model speaker-dependent acoustic characteristics. However, naturalness of English speech synthesized by such an ERJ acoustic model is significantly degraded as it is directly affected by prosodic differences and pronunciation errors often caused by differences of a language system between Japanese and English. To synthesize more natural English speech while preserving speaker individuality of individual Japanese speakers, we propose a technique to correct prosody of ERJ voices based on that of a native English speaker. The duration and power of the native English speaker are effectively used to develop the ERJ acoustic model for each Japanese speaker by using model adaptation techniques in HMM-based speech synthesis. The experimental results show that our proposed method is capable of significantly improving naturalness of ERJ synthetic speech while preserving its speaker individuality.

Key words　English-Read-by-Japanese (ERJ), HMM-based speech synthesis, prosody correction, speaker individuality, model adaptation

1. まえがき

情報通信技術や交通の発達に伴って，急速な国際化が進み，音声合成の多言語化の需要が増加している．近年では，より柔軟な多言語音声合成の実現を目指して，クロスリンガル音声合成の研究が盛んに行われている [1]．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反映させる技術であり，話者性による情報源の特定を促し，円滑なコミュニケーションを促進する役割を担う．特に日本では，日英間における合成技術の需要が高く，応用例としては，発話者自身の声で翻訳できる音声翻訳システム，声優本人の声での吹き替え映画，自身の声を聞きながら発音練習が可能な CALL システムなどが考えられる．

これまでに，統計的声質変換技術 [2] や隠れマルコフモデル（HMM: Hidden Markov Model）に基づく音声合成技術 [3] における話者適応技術 [4] において，英語を母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．統計的声質変換に基づく手法の例として，英語母語話者の話者依存 HMM の出力音声パラメータに対して，教師無し話者適応処理が可能な混合正規分布モデル （GMM: Gaussian Mixture Model） に基づく一対多話者変換法 [5] を適用する手法 [6] がある．また，話者適応に基づく手法の例として，バイリンガル話者による各言語の音響モデル間で状態の対応付けを行い，別言語で推定された話者適応パラメータを英語音響モデルに適用する手法 [7] がある．近年では，素片選択処理を用いて別言語の音声波形をフレーム単位で英語音声にマッピングし，得られた英語音声から HMM を学習する手法 [8] も提案されている．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劣化する傾向がある [6]．

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手法として，日本人英語（ERJ: English Read by Japanese）[9] を利用する合成法が有効であると考えられる．ERJ を用いた手法の例として，英語母語話者と日本語母語話者の英語音声間で GMM を学習し，日本語話者の声質に変換された英語母語話者音声を HMM の学習に利用する手法 [10] が挙げられる．しかしながら，日英間の言語構造の違いにより，日本人英語は英語母語話者による英語と大きく異なる音響的特徴を持つ．例えば，声の強弱や長短，高低，リズムやイントネーションに対応する韻律特徴に着目すると，日本語ではモーラ拍リズムに対し，英語は強勢拍リズムで韻律が変化する [11]．また，日本人英語では，母音の挿入，置換，脱落などの発音誤りが頻繁に生じる [12]．文法的に正しい発話内容であったとしても，英語独自の発音や抑揚を再現できていない日本人英語では，英語母語話者には半分程度しか通じないという報告もある [13]．このような特徴を持つ日本人英語を直接モデル化する枠組みを用いた場合，話者性は保持される一方で，得られる合成音声の自然性は著しく劣化する．

本報告では，コミュニケーションの円滑化や言語教育支援を目的とし，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 に対し，日本人英語を用いて適応処理を施すことで，話者性を強く反映した英語音声合成用 HMM を構築する．その際に，適応する HMM パラメータを限定し，英語母語話者らしい韻律を一部合成音声に反映させることで，日本人英語の話者性を保持しつつ，韻律の不自然性を低減する．実験的評価結果から，日本人英語音声合成において，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．

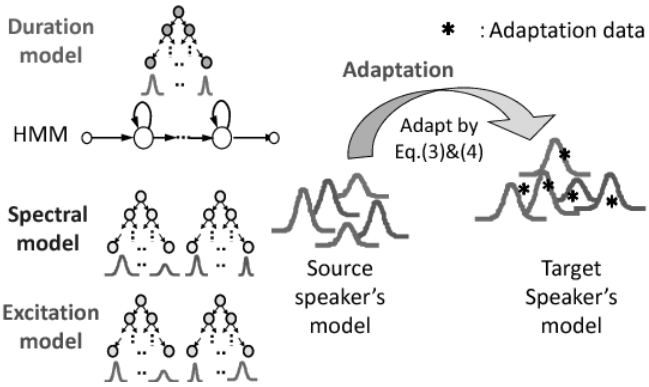

図 1　HMM 音声合成におけるモデル適応
Fig. 1　Model adaptation in HMM-based speech synthesis.

2. HMM 音声合成における適応技術

図 1 に HMM 音声合成におけるモデル適応の概要図を示す．

2.1 HMM 音声合成

HMM 音声合成では，音声のスペクトルパラメータ，音源パラメータ，状態継続長を，HMM に基づく統一的な枠組みでモデル化する．ここで，音声特徴量はアクセント型や前後の音素の影響を受けて変動するため，これらの変動要因を考慮したコンテキスト依存音素モデルを利用する．コンテキスト数は膨大であるため，各コンテキストに対する質問で構成される決定木によるクラスタリングを行い，同一クラスに属するモデル間でモデルパラメータを共有する [14]．クラスタリングは各 HMM 状態および各音声パラメータに対して別々に行われる．クラス c の出力確率分布 $b_c(\boldsymbol{o}_t)$ は，次式で表される．

$$b_c(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{1}$$

ただし，$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta\Delta\boldsymbol{c}_t^\top\right]^\top$ は，時刻 t における静的特徴量 $\boldsymbol{c}_t$ とその一次と二次の動的特徴量 $\Delta\boldsymbol{c}_t$，$\Delta\Delta\boldsymbol{c}_t$ の結合ベクトルを表し，$\mathcal{N}(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ は，平均 $\boldsymbol{\mu}_c$，共分散行列 $\boldsymbol{\Sigma}_c$ を持つ正規分布を表す．また，クラス c の状態継続長分布 $p_c(d_c)$ は，次式で表される．

$$p_c(d) = \mathcal{N}\left(d_c; \mu_c, \sigma_c^2\right) \tag{2}$$

ここで，d は状態継続長を表す．

合成時には，入力テキストのコンテキストに対応するクラスを HMM 状態および音声パラメータ毎に決定し，文 HMM を形

成するための出力確率分布および状態継続長分布が決定される．状態継続長分布の尤度最大化に基づき，HMM 状態系列を決定した後に，静的・動的特徴量間の明示的な制約条件下で HMM 尤度を最大化するような音声パラメータ系列を生成する [15]．生成された音声パラメータ系列から，ボコーダを用いて，音声波形を合成する．

2.2 HMM に基づくモデル適応

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 を構築できる．予め学習しておいた適応元モデルと目標話者の適応データを用いて，適応元モデルのパラメータを変形することで，目標話者へと適応されたモデルが得られる．本稿では，適応元である英語母語話者の HMM に対して，CSMAPLR（Constrained Structural Maximum *a Posteriori* Linear Regression）と MAP（Maximum *a Posteriori*）推定を組み合わせた CSMAPLR + MAP [16] を用いて，モデル適応処理を施す．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，CSMAPLR 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．

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{A}\boldsymbol{\mu}_i + \boldsymbol{b} \tag{3}$$

$$\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{A}\boldsymbol{\Sigma}_i\boldsymbol{A}^T \tag{4}$$

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 MAP 推定を行う．なお，木構造として，適応元である英語母語話者の HMM におけるコンテキストクラスタリングの決定木を用いる．CSMAPLR で適応されたモデルパラメータを事前分布のハイパーパラメータとして利用することで，さらに MAP 推定を施す．

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, P(\boldsymbol{o}|\boldsymbol{\lambda})\, P(\boldsymbol{\lambda}) \tag{5}$$

ここで，$\boldsymbol{o}$ は適応データ，$P(\boldsymbol{o}|\boldsymbol{\lambda})$ は HMM の尤度関数，$P(\boldsymbol{\lambda})$ は共役事前分布を表し，$\hat{\lambda}$ は MAP 推定されたモデルパラメータセットを表す．

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，分節的特徴のみでなく韻律的特徴も同時に適応可能になる．

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．しかしながら，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，

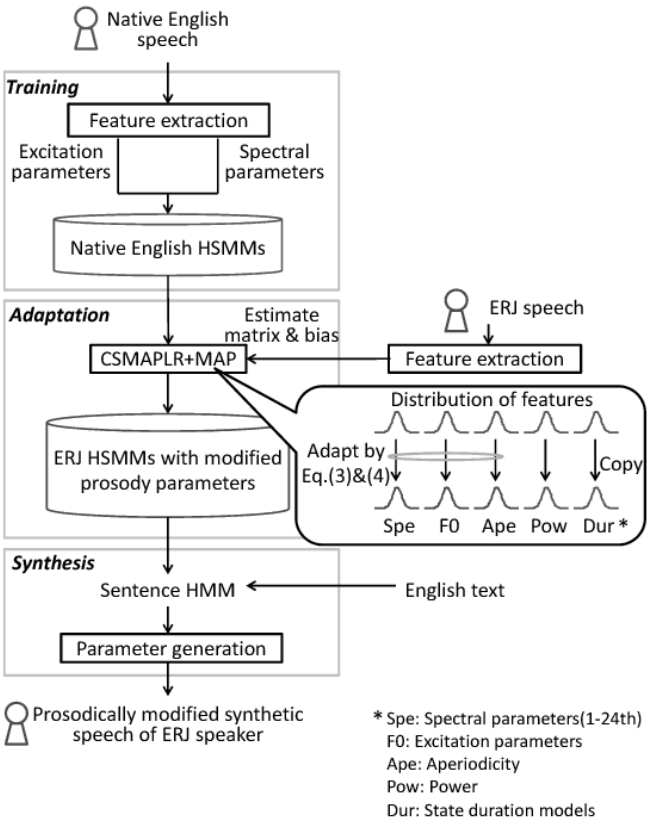

図 2　英語母語話者の英語音声を用いた韻律補正処理の流れ

Fig. 2　Flow of the prosody correction processing using English speaker's speech.

得られる合成音声は日本人英語音声と同様，不自然な音声となる．本報告では，日本人英語音声の自然性を大きく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．

図 2 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，英語母語話者に対する話者依存 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，スペクトル包絡パラメータ，音源パラメータであり，各パラメータに対する出力確率分布と状態継続長分布が得られる．次に，目標日本人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．日本人英語の音声パラメータとして，英語母語話者の英語音声の分析時と同様のパラメータを利用する．適応時には，スペクトル包絡パラメータと音源パラメータに対するモデルパラメータには通常の適応処理を施すが，対数パワーと状態継続長に対するモデルパラメータには適応処理を施さず，英語母語話者の話者依存 HMM のパラメータをそのまま保持する．本適応処理により，英語母語話者のパワーと継続長が反映された日本人英語 HMM が構築される．スペクトル包絡および音源に関しては，目標日本語母語話者の話者性が色濃く反映される一方で，自然性に大きな影響を与えるパワーと継続長は英語母語話者の特徴が反映されるため，本 HMM を用いることで，目標日本語母語話者の話者性を出来る限り保持したまま，自然性が改善された日本人英語音声を合成できると期待される．

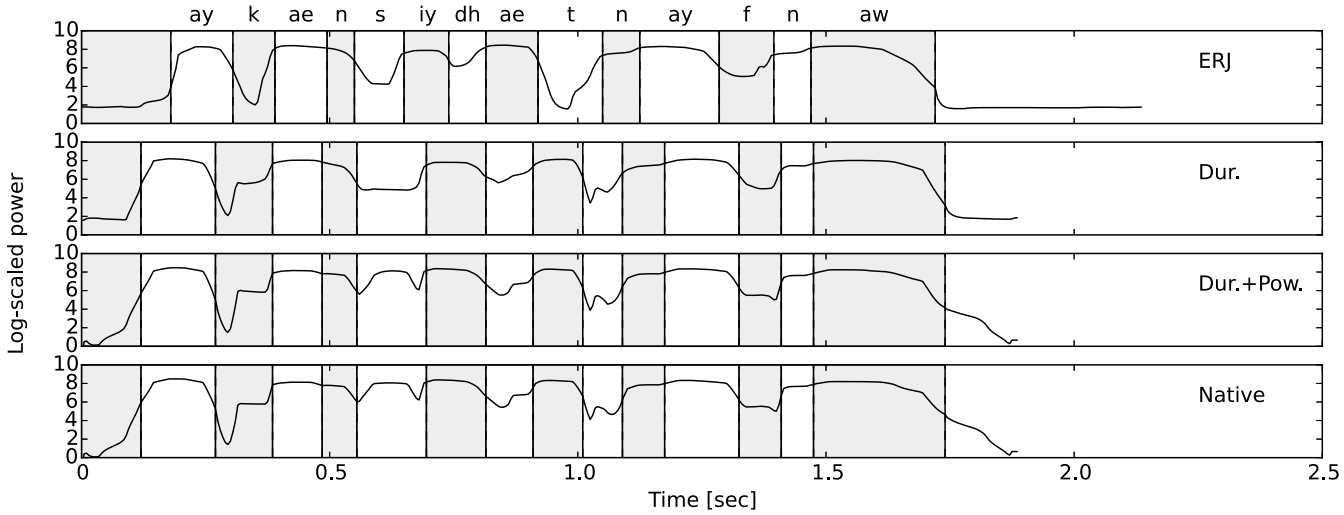

図 3　各モデルによる英語合成音声のパワー軌跡の例 （Monolingual）．発話内容は “I can see that knife now.”

Fig. 3　Example of power trajectories of English speech samples synthesized by individual models for a sentence “I can see that knife now.” （Monolingual）.

提案法では，パワーおよび継続長に関しては，他の音声パラメータと比較して，話者性に与える影響は少ないが，自然性に対しては無視できない影響を与えると仮定している．この仮定については，次節の実験的評価において検証する．また，従来法である [6]〜[8] などでは，異言語間で話者変換を行うため，パワーと継続長のみでなく F_0 パターンに関しても英語母語話者の特徴を用いることになるが，提案法では，F_0 パターンに関しては日本人英語の特徴がそのまま反映されるため，目標日本語母語話者の話者性をより強く反映した合成音声が得られると考えられる．

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 [17] 中の男性英語母語話者 1 名による A セット 593 文を用いる．評価データは同 B セット 50 文とする．一方で，目標話者は，20 代男性の日本語母語話者 2 名とする．内一人は，留学経験の無い大学院生であり，日本の標準的な英語教育を受けてきた話者であり，以下では “Monolingual” とする．もう一人は，1 年間オーストラリアへの留学経験のある大学生であり，より英語の習熟度が高い話者であり，以下では “Bilingual” とする．上記二名の日本語母語話者が発声した ARCTIC 音声データベース中の A セット 593 文を適応データとして使用する．学習データ，評価データ，および，適応データのサンプリング周波数は 16 kHz である．

音声パラメータの分析には STRAIGHT 分析 [18] を使用し，スペクトル特徴量として，対数パワーおよび 1 次から 24 次のメルケプストラム係数を用いる．音源特徴量として，対数 F_0 および 5 周波数帯域における平均非周期成分 [19] を使用する．フレームシフトは 5 ms とする．これらの音声パラメータに 1 次と 2 次の動的特徴量を加えたものを観測ベクトルとし，5 状態 left-to-right 型の隠れセミマルコフモデル（HSMM: Hidden Semi-Markov Model）[20] の学習を行う．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モデル化する．CSMAPLR 適応時に用いる回帰行列は，静的特徴量，1 次の動的特徴量，および，2 次の動的特徴量に対応したブロック対角行列とする．なお，予備実験の結果，英語母語話者依存 HSMM を目標日本語話者に適応した際に，合成音声の状態継続長が著しく不自然になるという傾向が見られた．そこで，英語母語話者の HMM を学習する際の状態継続長に対するコンテキストクラスタリングにおいて，ルート部分で当該音素がポーズか否かで大きく二分した後で，他の質問による通常のクラスタリングを行う．本処理で得られるコンテキストクラスタリングの決定木に基づき，回帰パラメータを推定することで，上記の問題が大きく緩和されることを確認済みである．

提案法による韻律補正の効果を評価するために，以下に示す合成音声を用いて，話者性及び自然性に関する主観評価を実施する．

- Native： 英語母語話者の話者依存 HSMM による合成音声
- ERJ： 日本語母語話者の話者依存 HSMM（593 文を用いて学習）による合成音声
- HMM+VC： 従来法 [6] （ただし，一対多話者変換ではなく ERJ を用いた一対一話者変換を使用） に基づき，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 に基づく統計的声質変換を適応して得られる合成音声
- 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.： 提案法として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.： 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

話者性の評価では，日本語母語話者の日本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．自然性の評価では，英語音声の自然

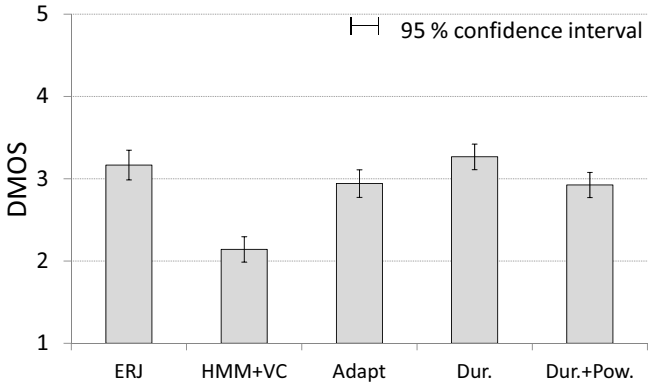

図 4　話者性に関する主観評価結果 (Monolingual)
Fig. 4　Result of opinion test on individuality (Monolingual).

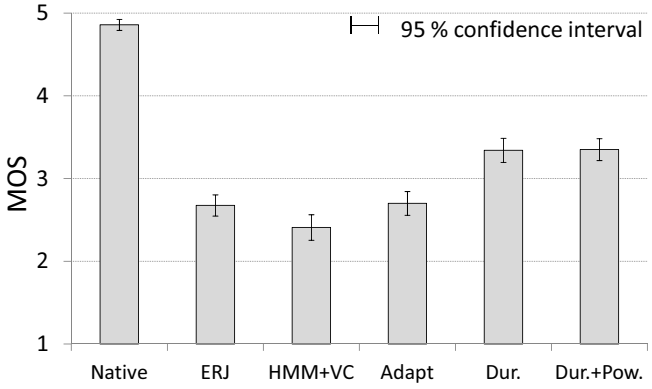

図 5　自然性に関する主観評価結果 (Monolingual)
Fig. 5　Result of opinion test on naturalness (Monolingual).

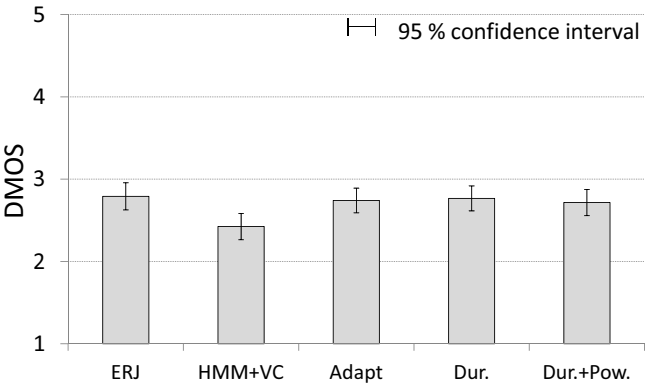

図 6　話者性に関する主観評価結果 (Bilingual)
Fig. 6　Result of opinion test on individuality (Bilingual).

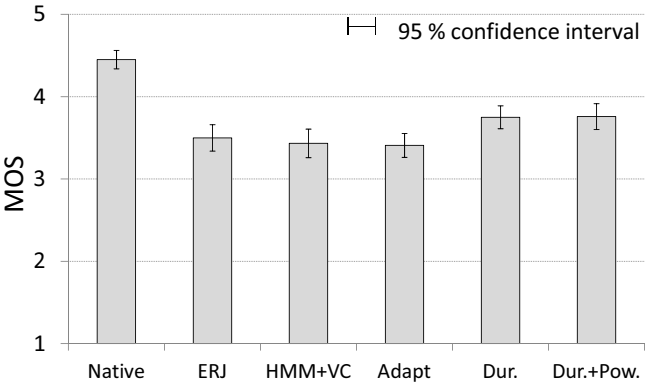

図 7　自然性に関する主観評価結果 （Bilingual）
Fig. 7　Result of opinion test on naturalness （Bilingual）.

性に関する 5 段階 MOS（Mean Opinion Score）評価を実施する．被験者は，日本語母語話者 6 名である．

4.2 実験結果

図 3 に，各手法によるパワー軌跡の例を示す．“Native”と比較し，“ERJ”は音素継続長が異なっており，パワー軌跡についても子音のパワーが小さいなど異なる特徴が見られる．これに対し，“Dur.”では音素継続長が補正されていることが分かる．また，これに伴い，パワー軌跡も “Native”のものに近づく傾向が見られる．さらに，“Dur.+Pow.”により，“Native”と同等のパワー軌跡が得られることが分かる．

図 4，図 5 にそれぞれ，“Monolingual”における話者性と自然性に関する評価結果を，図 6，図 7 にそれぞれ，“Bilingual”における話者性と自然性に関する主観評価結果を示す．まず，統計的声質変換を利用した手法 “HMM+VC”に着目すると，“Monolingual”において，“ERJ”と比較して，話者性が大きく劣化する傾向が見られる．また，自然性に関しても若干劣化する傾向が見られる．“Bilingual”においても，劣化の程度は小さくなるものの，話者性に関して同様の傾向が見られる．これらは，声質変換処理による品質劣化によるものと考えられる．次に，全モデルパラメータを適応した手法 “Adapt”に着目すると，概ね “ERJ”と同等の自然性および話者性が得られることが分かる．これに対して，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．“Monolingual”と “Bilingual”の両話者において改善傾向が見られるが，“Monolingual”における改善傾向が著しいことが分かる．一方で，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.”と同等だが，話者性が “Monolingual”において “ERJ”よりも劣化する傾向が見られる．この原因を調査するために，図 8 に “Monolingual”および “Bilingual”に対する “Dur.+Pow.” による合成音声と “Native”による合成音声のスペクトログラムの一部を示す．図から，特に無声子音部/s/において，“Bilingual”と “Native”では比較的滑らかなスペクトル遷移が確認できるが，“Monolingual”に関しては多少の不連続部分が確認できる．合成音声を聴取すると，“Monolingual”の場合は，パワー補正により，/s/の部分において雑音が強調されたような音声となる．これらのことから，適切なスペクトル構造が得られていない音素に対してパワー補正を施すと，悪影響が生じることが分かる．

以上の結果から，提案法による継続長補正により，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．本補正処理は，英語習熟度が低い話者に対して大きな自然性改善効果をもたらすものであり，英語習熟度が高い話者に対しても有効であることが分かる．一方で，パワー補正に関しては，英語習熟度が低い話者に対しては，パワーとスペクトル構造との不一致により，悪影響が生じ得ることが分か

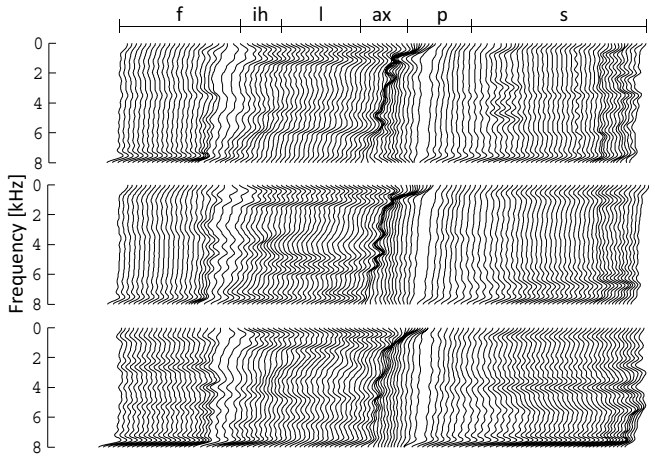

図 8　各種合成音声のスペクトログラム（発話文中の “Philip’s.”という単語に対応）の例．上から，“Dur.+Pow.（Monolingual）”，“Dur.+Pow.（Bilingual）”，“Native”を表す．

Fig. 8　Example of spectrograms of synthetic speech samples for a word fragment “Philip’s.” From top to bottom:“Dur.+Pow.（Monolingual）”，“Dur.+Pow.（Bilingual）” and “Native.”）

る．このため，パワーの補正処理に応じて，スペクトル構造も補正する必要があることが分かる．

5. ま と め

本報告では，日本語母語話者の話者性を保持しつつ，より自然な英語音声を合成する手法として，日本人英語音声を用いたHMM 音声合成モデル適応に基づく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．一方で，英語習熟度が低い話者に関しては，パワー補正が悪影響をもたらすことが分かった．今後は，評価者の母国語の影響 [21] や日本語母語話者の英語習熟度の影響 [9] について調査するとともに，スペクトル構造の補正処理についても検討を進める．

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．

文　献

[1] K. Oura, K. Tokuda, J. Yamagishi, S. King, and M. Wester, “Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis,” Proc. ICASSP, pp.4594–4597, Texas, USA, Mar. 2010.

[2] T. Toda, A.W. Black, and K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” IEEE Trans. ASLP, vol.15, no.8, pp.2222–2235, Nov. 2007.

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden markov models,” Proc. IEEE, vol.101, no.5, pp.1234–1252, Apr. 2013.

[4] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training,” IEICE Trans. Inf. and Syst, vol.90, no.2, pp.533–543, Feb. 2007.

[5] T. Toda, Y. Ohtani, and S. Kiyohiro, “One-to-many and many-to-one voice covnersion based on eigenvoices,” Proc. ICASSP, pp.1249–1252, Hawaii, USA, Apr. 2007.

[6] N. Hattori, T. Toda, H. Kawai, H. Saruwatari, and K. Shikano, “Speaker-adaptive speech synthesis based on eigenvoice conversion and language-dependent prosodic conversion in speech-to-speech translation,” Proc. INTERSPEECH, pp.2769–2772, Florence, Italy, Aug. 2011.

[7] H. Liang, Y. Qian, F.K. Soong, and G. Liu, “A cross-language state mapping approach to bilingual (Mandarin-English) TTS,” Proc. ICASSP, pp.4641–4644, Las Vegas, NV, Mar. 2008.

[8] Y. Qian, J. Xu, and F.K. Soong, “A frame mapping based HMM approach to cross-lingual voice transformation,” Proc. ICASSP, pp.5120–5123, Prague, May. 2011.

[9] N. Minematsu, Y. Tomiyama, K. yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, “Development of English Speech Database Read by Japanese and Americans for CALL System Development,” Japan journal of educational technology, vol.27, no.3, pp.259–272, Dec. 2003.

[10] S. Kajima, A. Iida, K. Yasu, Y. Aikawa, T. Arai, and T. Sugawara, “Development of a Japanese and English Speech Synthesis System Based on HMM Using Voice Conversion for the People with Speech Communication Disorder,” SIG-SLP, vol.2008, no.12, pp.121–126, Feb. 2008.

[11] P. Ladefoged, A Course in Phonetics, Third Edition, Harcourt Brace Jovanovich College Publishers, 1993.

[12] S. Kohmoto, Applied English phonology : teaching of English pronunciation to the native Japanese speaker, Tanaka Press, Tokyo, Japan, 1965.

[13] N. Minematsu, K. Okabe, K. Ogaki, and K. Hirose, “Measurement of Objective Intelligibility of Japanese Accented English Using ERJ (English Read by Japanese) Database.,” Proc. INTERSPEECH, pp.1481–1484, Florence, Italy, Aug. 2011.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based speech synthesis,” Proc. EUROSPEECH, pp.2347–2350, Budapest, Hungary, Apr. 1999.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech Parameter Generation Algorithms for HMM-based Speech Synthesis,” Proc. ICASSP, vol.3, pp.1315–1318, Istanbul, Turkey, Jun. 2000.

[16] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis,” IEEE Trans. ASLP, vol.17, no.6, pp.1208–1230, Jun. 2009.

[17] J. Kominek and A.W. Black, “CMU ARCTIC databases for speech synthesis CMU Language Technologies Institute,” Technical report, Tech Report CMU-LTI-03-17, 2003.

[18] H. Kawahara, I. Masuda-Katsuse, and A. deCheveigné, “Restructuring Speech Representations Using a Pitch-adaptive Time-frequency Smoothing and an Instantaneous-frequency-based F0 Extraction: Possible Role of a Repetitive Structure in Sounds,” Speech Commun., vol.27, no.3–4, pp.187–207, Apr. 1999.

[19] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” MAVEBA, pp.59–64, Firenze, Italy, Sep. 2001.

[20] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Hidden Semi-Markov Model Based Speech Synthesis System,” IEICE Trans., Inf. and Syst., E90-D, vol.90, no.5, pp.825–834, 2007.

[21] K. Tokuda and A. Black, “Speech synthesis research in a new age of cooperation and competition–The Blizzard Challenge,” J. ASJ, vol.62, no.6, pp.466–472, Jun. 2006.