韻律・音韻の部分補正に基づく話者性を保持した日本人英語音声合成と英語習熟度が与える影響

大島 悠司[1]　高道 慎之介[1]　戸田 智基[1]　Sakriani Sakti[1]　Graham Neubig[1]　中村 哲[1]

概要：声質変換や HMM 音声合成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）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．しかしながら，評価者の母語と発話者の英語習熟度に対する補正法の影響の違いは十分に調査されておらず，また，日本人英語の自然性低下の要因である音韻誤りについても考慮されていない．本稿では，評価者の母語と発話者の英語習熟度が韻律補正の効果に与える影響を調査するとともに，新たに無声子音スペクトル置換に基づく音韻補正法を提案する．実験的評価により，（1）パワー補正による自然性の改善効果は，英語母語話者による評価において顕著に見られること，（2）英語習熟度に関わらず，韻律補正法により自然性が改善すること，（3）音韻補正法も自然性改善に有効であることを示す．

キーワード：日本人英語, HMM 音声合成, 韻律補正, 音韻補正, 話者性

English-Read-By-Japanese Speech Synthesis Preserving Speaker Individuality Based on Partial Correction of Prosody and Phonetic Sounds and Effects of English Proficiency Level on Its Performance

Yuji Oshima[1]　Shinnosuke Takamichi[1]　Tomoki Toda[1]　Sakriani Sakti[1]　Graham Neubig[1]
Satoshi Nakamura[1]

Abstract: Cross-lingual speech synthesis for generating naturally sounding English speech uttered by Japanese speakers based on voice conversion and HMM-based speech synthesis tends to cause the degradation of speaker individuality in synthetic speech compared to intra-lingual speech synthesis. To address this issue, we have proposed an ERJ(English Read by Japanese) speech synthesis method to preserve speaker individuality in synthetic speech and a prosody correction method to improve its naturalness. However, their effectiveness has never been evaluated by native listeners: the effects of each speaker's English proficiency level on their performance have never been evaluated; and incorrect phonetic sounds of ERJ have never been addressed. In this paper, we evaluate these points by applying the proposed method to multiple speakers with various English proficiency levels and also propose a correction method of some incorrect phonetic sounds based on spectrum swapping for unvoiced consonants. The experimental results demonstrate that (1) the effectiveness of power correction is well confirmed by native listeners; (2) the naturalness of ERJ synthetic speech is successfully improved over various English prociency levels by the prosody correction method; and (3) the proposed phonetic sound correction method is also effective for further improving its naturalness.

Keywords: English-Read-by-Japanese (ERJ), HMM-based speech synthesis, prosodical correction, phonetic correction, speaker individuality

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，音声翻訳システム，海外映画の吹き替えやCALL システム [1] への応用が期待される．

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 [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 [5], [6], [7] が広く研究されている．これらの手法は，比較的高い自然性を持つ英語音声を合成できる一方で，同一言語間における合成音声と比較すると，話者性の劣化を招く傾向がある [5].

これに対し我々は，日本人英語（ERJ: English Read by Japanese）[8] を利用したモデル構築，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 [9]．しかしながら，本提案法の評価条件は，日本語母語話者から成る評価者と少数発話者のみに留まっており，評価者の母語と発話者の英語習熟度による影響が調査されていない．また，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．

本稿では，評価者の母語と発話者の英語習熟度が韻律補正の効果に与える影響を調査するとともに，無声子音スペクトル置換に基づく音韻補正法を提案する．実験的評価により，（1）パワー補正による自然性の改善効果は，英語母語話者による評価において顕著であること，（2）英語習熟度に関わらず，韻律補正法により自然性が改善すること，（3）音韻補正法も自然性改善に有効であることを示す．

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．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 [10]．コンテキストクラスタリングによるクラス 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．

HMM 音声合成では，モデル適応技術を用いることで，

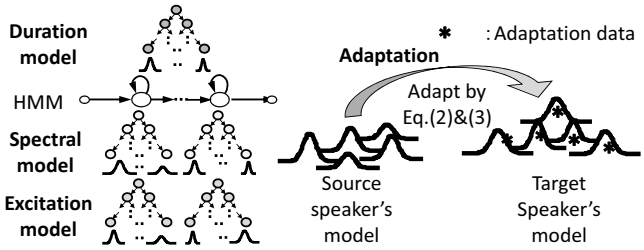

図 1　HMM 音声合成におけるモデル適応
Fig. 1　Model adaptation in HMM-based speech synthesis.

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．適応後の平均ベクトル $\hat{\boldsymbol{\mu}}_c$ と共分散行列 $\hat{\boldsymbol{\Sigma}}_c$ は次式で計算される．

$$\hat{\boldsymbol{\mu}}_c = \boldsymbol{A}\boldsymbol{\mu}_c + \boldsymbol{b} \tag{2}$$

$$\hat{\boldsymbol{\Sigma}}_c = \boldsymbol{A}\boldsymbol{\Sigma}_c\boldsymbol{A}^T \tag{3}$$

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．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．

パラメータ生成時には，入力テキストを解析することで得られるコンテキストに基づき，文 HMM を構築する．その後，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，静的・動的特徴量間の明示的な制約の下で，HMM の尤度最大化によりパラメータを生成 [11] し，ボコーダに基づく波形生成処理を経て音声が合成される．

3. 日本人英語音声合成における韻律補正法と音韻補正法

3.1 モデル適応による韻律補正法

図 2 にモデル適応による韻律補正法の概要を示す．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．本手法では，日本人英語音声の自然性を劣化させる要因として，継続長及びパワーに着目し，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．本適応法により，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の合成が可能になる [9].

[1] 奈良先端科学技術大学院大学　情報科学研究科

表 1　評価に用いる手法
Table 1　Synthetic speech samples used for evaluation.

手法名	学習データ	適応データ	韻律補正	音韻補正
ERJ	日本人英語	–	なし	なし
HMM+VC[2]	英語母語話者英語	–	–	–
Adapt	英語母語話者英語	日本人英語	なし	なし
Dur.	英語母語話者英語	日本人英語	状態継続長	なし
Dur.+Pow.	英語母語話者英語	日本人英語	状態継続長，対数パワー	なし
Dur.+Pow.+UVC	英語母語話者英語	日本人英語	状態継続長，対数パワー	無声子音スペクトル
Native	英語母語話者英語	–	–	–

3.2　無声子音スペクトル置換による音韻補正

図 3 に無声子音スペクトル置換に基づく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．音高や母音は話者性知覚に強く影響する [12] 一方で，無声子音の話者依存性は小さいと予想される．そのため，日本人英語の無声子音の置換により，話者性を保持しつつ自然性を改善できると考えられる．

まず，英語母語話者 HMM と韻律補正された日本人英語 HMM から，それぞれ音声パラメータを生成する．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うち，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．置換の際に，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，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．

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 [13] 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．評価データは同 B セット 50 文とする．学習データ，評価データ，及び，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 [14] 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 及び 5 周波数帯域における平均非周期成分を用いる．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 HSMM[15] 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学習する．モデル適応は CSMAPLR+MAP[16] を利用し，回帰行列には静的特徴量，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．

[2] 従来法 [5]（ただし，一対多話者変換ではなく日本人英語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用

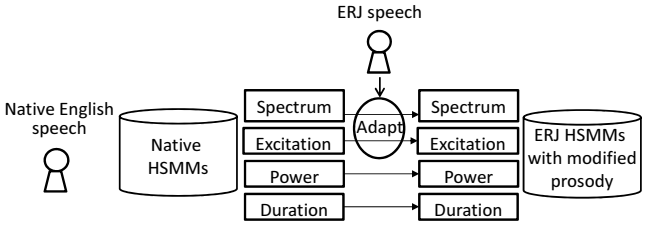

図 2　モデル適応による韻律補正の概要
Fig. 2　An overview of the prosody correction method based on model adaptation technique.

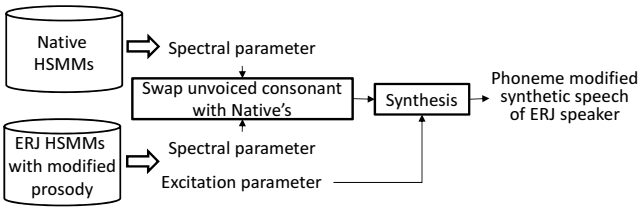

図 3　無声子音スペクトル置換による音韻補正の概要
Fig. 3　An overview of the phoneme correction method based on spectrum swapping of the unvoiced consonants.

提案法による韻律補正の効果を評価するために，表 1 に示す手法による合成音声を用いて，話者性，自然性及び明瞭性に関する主観評価を実施する．

4.2　モデル適応による韻律補正の効果

4.2.1　韻律補正法における評価者の母語の影響

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る（“Monolingual”）．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，英語習熟度が高い話者である（“Bilingual”）．上記 2 名が発声した 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．

話者性の評価では，目標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．評価する手法は，“ERJ”，“HMM+VC”，“Adapt”，“Dur.”，“Dur.+Pow.”の 5 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，“ERJ”，“HMM+VC”，“Adapt”，“Dur.”，“Dur.+Pow.”，“Native”の 6 つである．なお，各評価は，目標話者毎に作成した実験セットを用いて，日本語及び英語母語話者各 6 名により実施する．

図 4 と図 5 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評価結果を示す [3]．まず，補正無しの手法（“ERJ” と “Adapt”）における評価者の母語の影響に着目する．図 4 の (a) と (b) の比較から，話者性のスコアは，異なる母語を持つ評価者間で同程度である一方で，図 5 の (a) と (b) の比較から，英語母語話者による自然性のスコアは，日本語母語話者によるスコアと比較して，大きく減少する傾向が見られる．次に，英語母語話者のパワーを反映した手法（“HMM+VC” と “Dur.+Pow.”）に着目すると，図 5 に示す自然性に関する評価結果において，英語母語話者による評価では，日本語母語話者による評価と比較して，相対的なスコアの上昇がみられる．これらの結果は，英語発話のリズムおよび強勢に対して，英語母語話者は日本語母語話者よりも過敏であるためだと考えられる．

なお，両母語話者による自然性に関する評価において，“Dur.” と “Dur.+Pow.” は他の手法よりも高いスコアを獲得している．また，話者性に関する評価においても，“Dur.” と “Dur.+Pow.” は “ERJ” と同等の話者性を保持している．このことから，提案する韻律補正法の有効性が確認できる．

以上の結果から，日本語母語話者と英語母語話者との間には評価結果に違いが生じており，英語母語話者の方がより英語発話の韻律に対して敏感であることが確認でき，また，提案法による継続長及びパワー補正により，日本語母語話者の話者性を保持しつつ，英語母語話者にとって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 韻律補正法における発話者の英語習熟度の影響

適応データは，日本人学生による読み上げ英語音声データベース [8] 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[17] 60 文とする．ただし，本稿の英語習熟度は，データベース中で定義されている複数の基準（音素生成，リズム等）における評定点の平均を指す．評価法は 4.2.1 節と同様（ただし，話者性の評価では，目標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する点のみ異なる）であり，話者性の評価では “HMM+VC”，“Adapt”，“Dur.+Pow.” の 3 つ，自然性の評価では “HMM+VC”，“Adapt”，“Dur.+Pow.”，“Native” の 4 つの手法を評価する．なお，各評価は，全日本語母語話者の音声を含んだ実験セットを用いて，英語母語話者 6 名により実施する．

[3] ただし，図 4 (a) と図 5 (a) は [9] の再掲である．

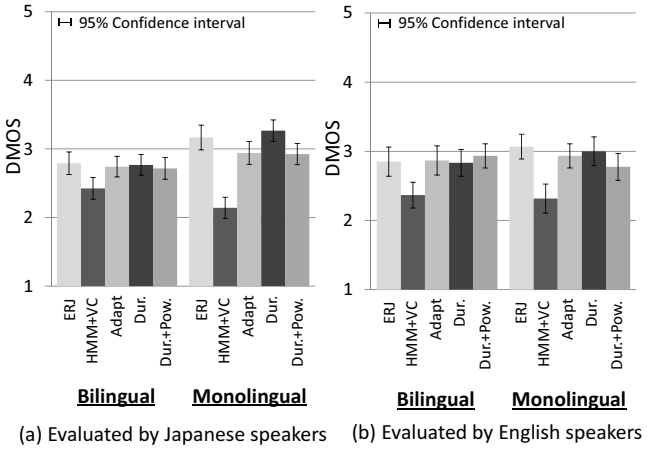

図 4 韻律補正法に対する話者性に関する主観評価結果
Fig. 4 Results of subjective evaluation of individuality for prosody correction method.

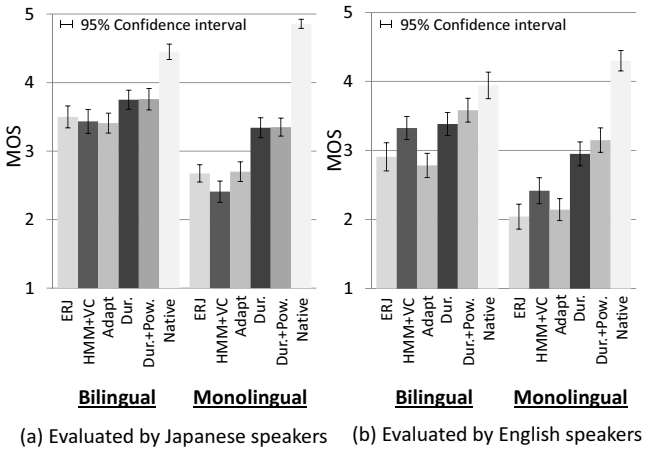

図 5 韻律補正法に対する自然性に関する主観評価結果
Fig. 5 Results of subjective evaluation of naturalness for prosody correction method.

図 7 に，英語習熟度（“High” と “Low”）毎に集計した，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．まず話者性に関する評価結果において，GMM 声質変換を利用した手法 “HMM+VC” に着目すると，“Low” に対して，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．一方で，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．

次に，自然性に関する評価結果を見ると，“HMM+VC” と比較し，“Adapt” は”Low” において大幅な劣化を生じさせることが分かる．これに対し，“Dur.+Pow.” は，韻律補正により自然性劣化を防ぐことが可能であり，英語習熟度に関係なく “HMM+VC” と同等の自然性が得られることが分かる．

以上の結果から，英語習熟度に関わらず，提案法が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，

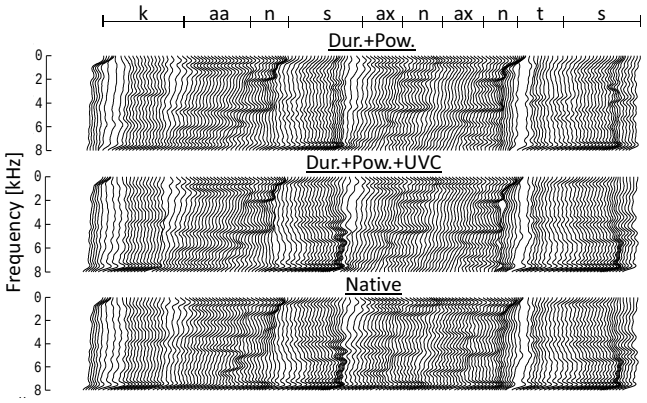

図 6　各種合成音声のスペクトログラム（発話文中の “consonants” という単語に対応）の例.

Fig. 6　Example of spectrograms of synthetic speech samples for a word fragment “consonants”

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．また，補正効果は，英語習熟度の低い話者において特に有効であることが分かる．

4.3　無声子音スペクトル置換による音韻補正の効果

適応データは， 4.2.1 節の “Monolingual” 及び “Bilingual” による ARCTIC 音声データベース中の A セット 60 文，並びに，4.2.2 節の “High” もしくは “Low” に属する 4 名による TIMIT 60 文とする．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．評価する手法は，“Dur.+Pow.”，“Dur.+Pow.+UVC”，“Native” の 3 つである．各評価は，全ての日本語母語話者の音声を含んだ実験セットを用いて，英語母語話者 6 名により実施する．ただし，評価結果は英語習熟度毎に計算し，“Monolingual” と “Bilingual” はそれぞれ，“Low” と “High” に属するものとする．

図 6 に，各手法によるスペクトルグラムの例を示す．図から，“Native” と比較し，“Dur.+Pow.” では無声子音部分（/s/など）において，特に高周波数領域におけるスペクトル包絡の形状が大きく異なることが分かる．これは，パワー補正を実施した際に，異音を生じさせる要因となる．これに対し，“Dur.+Pow.+UVC” では，“Native” 同様のスペクトル包絡形状が得られるため，パワー補正による悪影響を緩和することができ，自然性の向上が期待できる．

図 8 に，英語習熟度（“High” と “Low”）毎に集計した，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” に着目すると，“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” において，

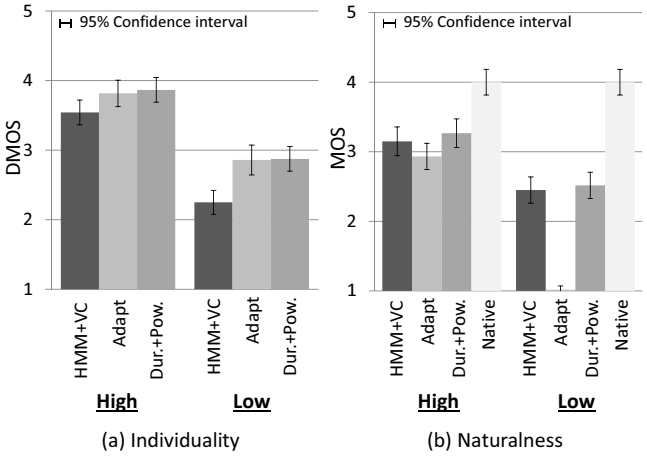

図 7　英語習熟度毎に計算した主観評価結果（韻律補正法）

Fig. 7　Results calculated in each English proficiency level (prosody correction method).

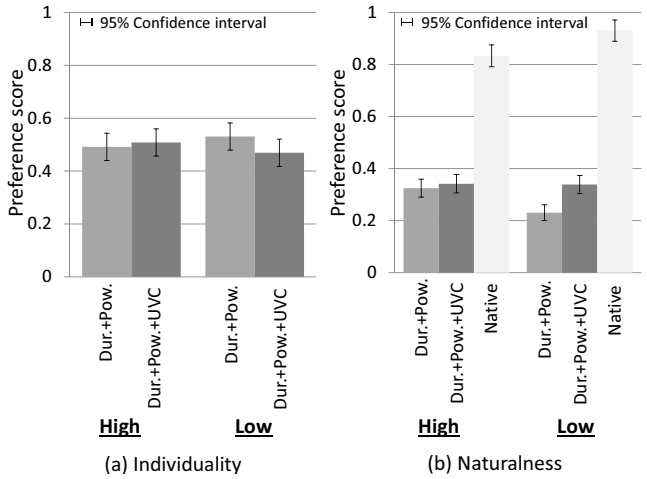

図 8　英語習熟度毎に計算した主観評価結果（音韻補正法）

Fig. 8　Results calculated in each English proficiency level (phoneme correction method).

“Dur.+Pow.+UVC” は “Dur.+Pow.” と同等の自然性及び話者性を保持できることが分かる．なお，“Dur.+Pow.” と “Dur.+Pow.+UVC” に対し，t 検定を行ったところ，“Low” の自然性のみ有意差が確認された（$p < .01$）．

以上の結果から，韻律補正法と同様に，提案した音韻補正法も自然性改善に有効であり，特に英語習熟度の低い話者において有効であることが分かる．

4.4　明瞭性に関する評価

提案法に対して，明瞭性に関する書き取り試験を実施する．評価データは SUS[18] 50 文とし，評価する手法は，“HMM+VC”，“Dur.+Pow.+UVC”，“Native” の 3 つである．なお，各評価は，全日本語母語話者の音声を含んだ実験セットを用いて，英語母語話者 6 名により実施する．ただし，評価結果は英語習熟度毎に計算し，“Monolingual” と “Bilingual” はそれぞれ，“Low” と “High” に属するものとする．

図 9 に，英語習熟度（“High” と “Low”）毎に集計し

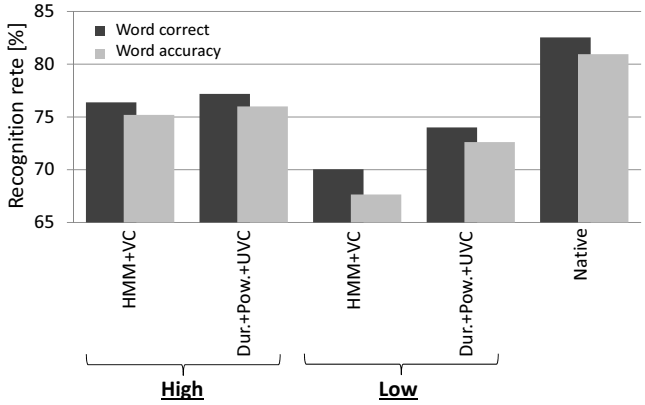

図 9　英語習熟度毎に計算した明瞭性に関する書き取り試験結果
Fig. 9　Results of dictation test on intelligibility calculated in each English proficiency level.

た，書き取り試験結果を示す．“Low”に着目すると，“Dur.+Pow.+UVC”は “HMM+VC”と比較して，明瞭性が改善していることが分かる．これは，“HMM+VC”に対し、“Dur.+Pow.+UVC”が持つ音韻補正効果により，無声子音音素の明瞭性が回復したためだと考えられる．また，“High”においては，“Dur.+Pow.+UVC”は “HMM+VC”と同等の明瞭性で，“Low”よりも高い明瞭性が得られている．なお，“Dur.+Pow.+UVC”は，“Native”と比較すると，単語正解精度の劣化を “High”において約 5 %，“Low”において約 8 %にとどめることができる．

5. おわりに

本稿では，日本人英語音声合成における話者性を保持した自然性改善を目的として，モデル適応による韻律補正法に対して，評価者の母語と発話者の英語習熟度が与える影響について調査し，また，子音スペクトル補正による音韻補正法を提案した．実験的評価により，（1）パワー補正による自然性の改善効果は，英語母語話者による評価において顕著であること，（2）英語習熟度に関わらず，韻律補正法により自然性が改善すること，（3）音韻補正法も自然性改善に有効であることを示した．今後は，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．

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] 高道慎之介，大島悠司，戸田智基，Graham, N.，Sakriani, S.，　中村哲：日本人英語のための音声合成技術を用いた英語学習支援の検討，教育システム情報学会，Vol. 29, No. 5, pp. 111–116 (2015).
[2] Toda, T., Black, A. W. and Tokuda, K.: Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory, *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235 (2007).
[3] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K.: Speech synthesis based on hidden Markov models, *Proc. IEEE*, Vol. 101, No. 5, pp. 1234–1252 (2013).
[4] Yamagishi, J. and Kobayashi, T.: 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 (2007).
[5] Hattori, N., Toda, T., Kawai, H., Saruwatari, H. and Shikano, K.: Speaker-adaptive speech synthesis based on eigenvoice conversion and language-dependent prosodic conversion in speech-to-speech translation, *Proc. INTERSPEECH*, pp. 2769–2772 (2011).
[6] Liang, H., Qian, Y., Soong, F. K. and Liu, G.: A cross-language state mapping approach to bilingual (Mandarin-English) TTS, *Proc. ICASSP*, pp. 4641–4644 (2008).
[7] Qian, Y., Xu, J. and Soong, F. K.: A frame mapping based HMM approach to cross-lingual voice transformation, *Proc. ICASSP*, pp. 5120–5123 (2011).
[8] Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M. and Makino, S.: Development of English Speech Database Read by Japanese to Support CALL Research, *Proc. ICA*, Vol. 1, pp. 557–560 (2004).
[9] 大島悠司，高道慎之介，戸田智基，Graham, N.，Sakriani, S.，　中村哲：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，信学技報，Vol. 114, No. 365, pp. 63–68 (2014).
[10] 吉村貴克，徳田恵一，益子貴史，小林隆夫，　北村正：HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，信学論，Vol. J83-D2, No. 11, pp. 2099–2107 (2000).
[11] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T.: Speech Parameter Generation Algorithms for HMM-based Speech Synthesis, *Proc. ICASSP*, Vol. 3, pp. 1315–1318 (2000).
[12] Kitamura, T. and Akagi, M.: Speaker Individualities in Speech Spectral Envelopes, *Proc. ICSLP*, Vol. 3, pp. 1183–1186 (1994).
[13] Kominek, J. and Black, A. W.: CMU ARCTIC databases for speech synthesis CMU Language Technologies Institute, Technical report, CMU-LTI-03-177 (2003).
[14] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: 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 (1999).
[15] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: Hidden Semi-Markov Model Based Speech Synthesis System, *IEICE Trans., Inf. and Syst., E90-D*, Vol. 90, No. 5, pp. 825–834 (2007).
[16] Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S. and Renals, S.: Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis, *IEEE Trans. ASLP*, Vol. 17, No. 6, pp. 1208–1230 (2009).
[17] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G. and Pallett, D. S.: DARPA TIMIT acoustic-phonetic continous speech corpus, Technical report, NISTIR 4930, NIST, Gaithersburg, MD (1993).
[18] Benoît, C., Grice, M. and Hazan, V.: The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences, *Speech Communication*, Vol. 18, No. 4, pp. 381–392 (1996).