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
SP2015-73(2015-12)

音声波形加工に基づく非母語音声の継続長補正法の評価と分析

倶羅　真也† 　高道慎之介† 　戸田　智基†,†† ニュービッグ　グラム† 　サクティ　サクリアニ†
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
†† 名古屋大学情報基盤センター教育情報メディア研究部門　〒 565–0456 愛知県名古屋市千種区不老町
E-mail: †{kura.shinya.ke9,shinnosuke-t,neubig,ssakti,s-nakamura}@is.naist.jp, ††tomoki@icts.nagoya-u.ac.jp

あらまし　言語学習等への応用を見据え，非母語話者により発話された音声に対して，発話リズムを補正する技術が研究されている．代表的な手法として，母語話者の音声を参照音声として，動的時間伸縮を用いて音声特徴量系列を変形することで，補正音声を合成する枠組みが提案されている．柔軟な発話リズム補正が可能となる一方で，音声分析合成処理による音質劣化は避けられない．本稿では，音声分析合成処理による音質劣化を回避する補正法として，動的時間伸縮と音声波形加工処理による発話リズム補正法を提案する．時間伸縮規則を獲得する際には，統計的声質変換を適用することで，話者性の差異が音声スペクトルパラメータに与える影響を緩和するとともに，非母語話者音声に見られる音素挿入誤りにも対処する．日本語母語話者による英語音声（日本人英語）を用いた実験的評価から，提案法により，音質劣化を緩和しつつ，より自然な継続長を持つ英語音声を合成できることを示す．また，時間伸縮により生じる局所的な音質劣化の影響について，音声スペクトルパラメータの変調スペクトルを用いて分析する．
キーワード　非母語音声，継続長補正，動的時間伸縮，波形加工，変調スペクトル

Evaluation and Analysis of Duration Correction for Non-Native Speech Based on Waveform Modification

Shinya KURA†, Shinnosuke TAKAMICHI†, Tomoki TODA†,††, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Nara Institute of Science and Technology　Takayama-cho 8916–5, Ikoma, Nara, 630–0192 Japan
†† Information Media Division, Information Technology Center, Nagoya University　Furo-cho, Chikusa-ku Nagoya, Aichi, 464–8601 Japan
E-mail: †{kura.shinya.ke9,shinnosuke-t,neubig,ssakti,s-nakamura}@is.naist.jp, ††tomoki@icts.nagoya-u.ac.jp

Abstract　There are several attempts at correcting durational patterns of non-native speech towards language learning. One of the typical approaches modifies a speech parameter sequence with Dynamic Time Warping (DTW) using native speech as the reference, generating corrected speech from the modified speech parameter sequence. Although this approach makes it possible to flexibly modify durational patterns of non-native speech, quality of the corrected speech significantly degrades due to the use of analysis-synthesis process to generate the corrected speech. In this report, we propose a method for correcting durational patterns using direct waveform modification for performing DTW. In calculating a temporal warping function, statistical voice conversion is effectively used to reduce an adverse effect caused by speaker differences. Moreover, phoneme insertion often observed in non-native speech is also handled. We conducted an experimental evaluation using English speech read by Japanese, demonstrating that the proposed method was capable of flexibly modifying durational patterns while avoiding quality degradation caused by the analysis-synthesis process. Furthermore, waveform segments suffering from quality degradation caused by temporal warping was analyzed using the modulation spectrum of spectral parameters.
Key words　non-native speech, correction of durational patterns, dynamic time warping, waveform modification, modulation spectrum

1. まえがき

情報通信技術や交通の発達に伴う急速な国際化が進み，言語教育支援の需要が高まっている．特に，日本においては，日本語母語話者に対する英語学習支援の重要性が注目されている [1]．日本語と英語では，言語体系が大きく異なるため，日本語母語話者による英語（日本人英語）音声は，独自の特徴を持ちやすい．例えば，日本語は，モーラ単位の発話等時性を持つモーラ拍言語 (mora-timed language）である一方，英語は，ストレス単位の等時性を持つ強勢拍言語（stress-timed language）である [2]．それ故に，声の強弱，発話リズム，イントネーション等の韻律特徴が両言語間で大きく異なるため [3]，日本人英語音声の韻律特徴は，英語独自のものとは大きく異なる場合が多々見られる．また，分節的特徴においても，音素セットの違いのみならず，日本語の音節は，母音もしくは子音と母音の対により構成されるのに対し，英語の音節は，母音に子音が後続したり，複数の子音が連続したりするといった違いがある．そのため，日本人英語音声には，子音の直後に母音を付け加えて発音する「挿入」，日本語に存在しない音素を日本語の音素に置き換えて発音する「置換」，必要な音素が脱落する「削除」といった音素誤りが生じやすい [4]．これらの要因により，日本人英語音声の明瞭性は大きく低下し，また，この明瞭性低下の問題について，日本人英語の発話リズムが大きな要因であることも知られている [5]．なお，英語に限らず，その他の言語においても，発話リズムが明瞭性低下をもたらすという知見が得られている [6]．

日本語母語話者に対する英語学習支援を行う上で，両言語間の言語体系の違いに着目することは有効であり，音声合成技術を活用する手法が研究されている [7]．例えば, [8] では，日本語のアクセント型発音教育の CALL システムにおいて，韻律の誤りを修正した学習者本人の音声のフィードバックを提示する技術が提案されている．また, [9] では，隠れマルコフモデル (Hidden Markov Model: HMM) に基づくテキスト音声合成技術 [10] を用いて，事前収録した日本人英語音声から，より自然な発話リズムを持つ英語音声を合成する手法が提案されている．HMM 音声合成の適応技術 [11] を用いて，英語母語話者の継続長およびパワー変化の特徴を日本人英語音声へと反映させることで，学習者の個人性をできる限り保持しつつ，より自然な英語音声を合成可能となる．この手法の効果は，英語初学者の音声を用いた場合においても確認されている [12]．

一方で，実際に発話された日本人英語音声に対して変形処理を施すことで，より自然な英語音声を生成する手法も研究されている．テキスト音声合成のように複雑な処理を必要としないため，品質の高い英語音声の生成が可能となる．代表的な手法として，非負値時空間分解を用いて，発話された非母語音声の発話リズムを補正する手法が提案されている [13]．非母語音声のスペクトルパラメータ系列を，各音素を表す基底スペクトルと時間変化情報を表す補間重み系列に分解し，同一文を発話した母語音声の時間変化情報を転写することで，補正音声が合成される．音声特徴量の変形による柔軟な発話リズム補正が可能となるが，一方で音声の分析再合成処理を必要とするため，それに伴う品質劣化は避けられない．

本稿では，非母語音声に対する高音質な発話リズム補正処理として，非母語音声波形に対する直接的な変形処理手法について検討する．非母語音声と同一文を発話した母語音声を用いて，非母語音声に対して動的時間伸縮 (Dynamic Time Warping: DTW) [14] を施すことで，時間伸縮規則を推定する．その際には，統計的声質変換技術 [15] を用いて，話者の違いに起因する音響特徴量の差異を緩和する．次に，時間伸縮規則に基づいて非母語音声波形に対し WSOLA (Waveform Similarity based OverLap-Add) [16] を施すことで，非母語音声の発話リズムを補正する．日本人英語音声を用いた実験結果から，分析再合成を用いる手法と比較して，提案法により高音質な補正処理が可能であることを示す．さらに，時間伸縮規則により局所的に生じる自然性劣化の影響を，変調スペクトル [17] を用いて分析するとともに，自然性劣化箇所の自動検出を試みる．なお, [18] および [19] においても，非母語音声波形に対する変形処理に基づく発話リズム補正法が提案されているが, [18] は，音素挿入誤りへの対処について考慮されておらず, [19] は，統計的声質変換については考慮していない．一方で，本稿では，これらを考慮した枠組みを提案するとともに，時間伸縮に基づく発話リズム補正処理のさらなる高品質化を目指し，自然性劣化をもたらす要因を調査する．

2. DTW を用いた非母語音声の継続長補正

提案法の処理手順を図 1 に示す．非母語音声と母語音声の同一発話文を用いて，母語音声の時間長に一致するように非母語音声を時間伸縮する．

2.1 時間伸縮規則推定部

メルケプストラムひずみを距離尺度とした DTW により時間伸縮規則を推定する．その際に，非母語話者と母語話者の話者性の差異がメルケプストラムに与える影響を緩和するため，混合正規分布モデル （Gaussian Mixture Model: GMM） に基づく声質変換法を用いる．

2.1.1 GMM 声質変換を用いた話者性の影響緩和

非母語話者と母語話者による事前収録音声から成るパラレルデータから，GMM のパラメータセット $\boldsymbol{\lambda}$ を推定する．

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \prod_{t} P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) \tag{1}$$

ここで，$\boldsymbol{X}_t$，$\boldsymbol{Y}_t$ はそれぞれ，フレーム t における非母語音声および母語音声の音声特徴量であり，本稿では，メルケプストラム係数及びその 1 次の動的特徴量の結合ベクトルとする．$\boldsymbol{X}_t$ と $\boldsymbol{Y}_t$ の対応付けは，DTW による時間フレームアライメントにより行う．結合確率密度関数 $P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}\right)$ は，次式で表される．

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(ZZ)}\right) \tag{2}$$

ただし，

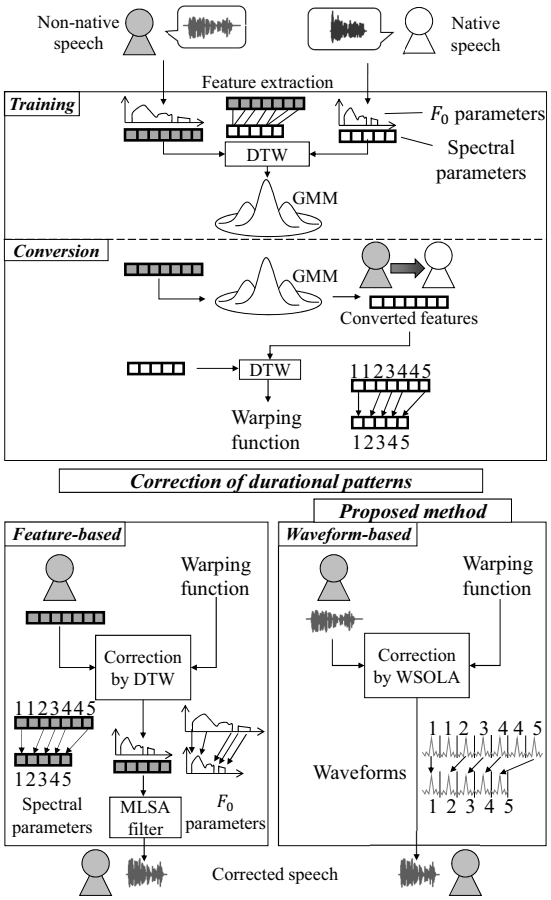

図 1　発話リズムの補正処理手順．下部左 (Feature-based) は，従来法である音声特徴量系列に対する補正処理を示し，下部右 (Waveform-based) は提案法である波形加工に基づく補正処理を示す．

Fig. 1　Procedure for correcting durational patterns. A lower left box ("feature-based") shows a conventional method using analysis-synthesis process, and a lower right box ("waveform-based") shows the proposed method using direct waveform modification process.

$$\boldsymbol{Z}_t = \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix}, \boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(ZZ)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \tag{3}$$

である．ここで，GMM パラメータセットは M 個の混合要素を有し，インデックス m の混合要素は，混合重み α_m，平均ベクトル $\boldsymbol{\mu}_m^{(Z)}$，共分散行列 $\boldsymbol{\Sigma}_m^{(ZZ)}$ から成る正規分布で表される．$\boldsymbol{\mu}_m^{(Z)}$ は，入出力音声の特徴量の平均ベクトルである $\boldsymbol{\mu}_m^{(X)}$ と $\boldsymbol{\mu}_m^{(Y)}$ から成る．また，$\boldsymbol{\Sigma}_m^{(ZZ)}$ は，入出力音声の特徴量の共分散行列である $\boldsymbol{\Sigma}_m^{(XX)}$ 及び $\boldsymbol{\Sigma}_m^{(YY)}$ と，相互共分散行列である $\boldsymbol{\Sigma}_m^{(YX)}$，$\boldsymbol{\Sigma}_m^{(XY)}$ から成る．

次に，非母語音声のメルケプストラム系列から母語話者のメルケプストラム系列へ変換する．非母語話者と母語話者の特徴量系列ベクトルを，それぞれ $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ と $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とする．ここで T はフレーム数である．変換された静的特徴量系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ は，次式で表される．

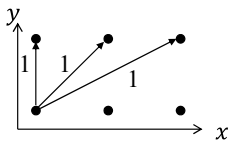

図 2　DTW における遷移コスト．x は非母語話者の時間フレームを表し，y は母語話者の時間フレームを表す．

Fig. 2　Transtion cost in DTW, where x and y show time-frame of non-native speech and that of native speech, respectively.

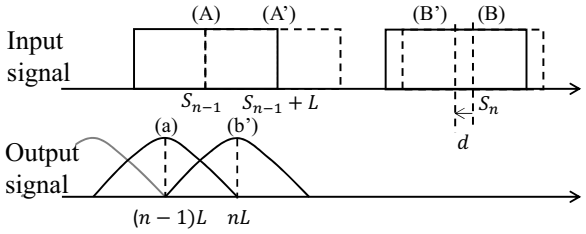

図 3　WSOLA の手順

Fig. 3　Procedure of WSOLA.

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{4}$$

ここで $\boldsymbol{W}$ は，静的特徴量系列をその 1 次の動的特徴量の結合ベクトル系列に拡張する行列である．条件付確率密度関数 $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ は，結合確率密度関数の式 (2) から解析的に導出される．なお，過剰な平滑化による変換スペクトルパラメータの品質劣化を緩和するため，系列内変動を考慮する．

2.1.2　DTW を用いた規則推定

GMM 声質変換により変換されたメルケプストラム系列と，母語音声のメルケプストラム系列に対して，DTW を行うことで，非母語音声に適用する時間伸縮規則を推定する．母語音声のメルケプストラム系列長は固定とし，変換後の系列とのメルケプストラム距離を最小とする時間伸縮規則を求める．本稿で用いる遷移コストを図 2 に示す．

2.2　継続長補正部

推定された時間伸縮規則に基づき，非母語音声に対して，WSOLA を用いた継続長変換を施すことで，発話リズム補正処理を行う．WSOLA の手順を図 3 に示す．ここでは，局所的に継続長を短くする（すなわち，発話リズムを速くする）際の処理を表しており，波形セグメント (A) と波形セグメント (B) の接続処理を行う．波形セグメント (A) の中心時間を S_{n-1} とし，波形セグメント (B) の中心時間を S_n とする．ここで，S_n は，時間 S_{n-1} における時間伸縮規則に基づき決定される．また，各セグメント長は $2L$ とし，波形セグメント (A) を L だけシフトさせた波形セグメントを (A') とする．接続による自然性劣化を抑えるために，波形セグメント (B) の中心時間をシフトさせ，波形セグメント (A') と最も類似度が高くなるシフト長 d を探索する．得られたシフト長に基づき，$S_n - d$ を中心とする波形セグメント (B') が決定する．波形セグメント (A) の後半部（$S_{n-1} \sim S_{n-1} + L$）と波形セグメント (B') の前半部（$S_n - d - L \sim S_n - d$）に対して，窓掛けを施して重畳加算を行うことで，波形セグメント (A) と波形セグメント (B') を接続する．図において，(a) と (b') は窓掛けされた波形セグメント (A) と (B') に対応する．次に，波形セグメント (B') を波形

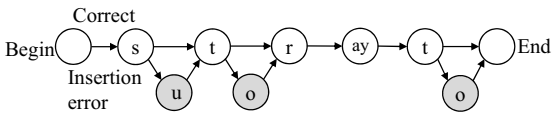

図 4　音素挿入誤りを考慮した音素ネットワークの例．発話単語は “straight”．

Fig. 4　An example of the phoneme network considering the phoneme insertion errors for a word fragment “straight.”

セグメント (A) として，この手順を音声波形の終端まで繰り返す．本稿では，シフト長 d を決定する際の類似度として，次式で計算される正規化相互相関係数 $r(d)$ を用いる．

$$r(d) = \frac{\Sigma_t \left[(x(t) - \mu_x) \cdot (y(t-d) - \mu_y)\right]}{\sqrt{\Sigma_t \left[x(t) - \mu_x\right]^2} \cdot \sqrt{\Sigma_t \left[y(t-d) - \mu_y\right]^2}} \qquad (5)$$

ここで，$x(t)$，$y(t-d)$ は，類似度計算を行う各波形セグメントの振幅値を表す．μ_x および μ_y は，各波形セグメントにおける振幅値の平均を表す．また，シフト長の探索幅は 10 ms，セグメント長は $2L$ =20 ms とし，窓掛けには窓長 20 ms の三角窓の片側 10 ms を用いる．

2.3　HMM 学習を用いた音素誤りの検出

DTW による時間伸縮に基づく発話リズム補正処理では，対象とする 2 つの音声の発話内容が同一であるという仮定を置く．そのため，日本人英語の様に，音素挿入が生じる非母語音声を対象とする際には，何らかの対応が必要となる．DTW の際に許容するスキップフレーム数を大きく設定する方法も考えられるが，日本人英語において誤って挿入される母音は継続長が長いため，十分な効果が得られるとは限らず，また，他の箇所においても不適切な時間伸縮を許容する可能性が高まり，悪影響が生じる場合がある．そこで，本稿では，音素挿入箇所を検出し，前もって除外する処理を施す．日本人英語の音素誤りの規則性は古くから研究されており [4]，図 4(注1) の様に，発話テキストの音素列から音素挿入誤りを含む音素ネットワークを規則的に構築可能である．この音素ネットワークを用いて，非母語音声に対して HMM による状態アライメントを行うことで，誤って挿入された音素を自動検出する [20]．検出された音素に対しては，継続長を 0 にするように時間伸縮規則を設定することで，挿入音素を削除する．

2.4　考　　察

日本人英語においては，音素挿入の他にも，音素置換，音素削除が生じやすい．これらについては，提案法による発話リズム補正のみでは対処できない．また，挿入音素の削除処理においても，挿入音素の先行音素と後続音素を直接接続するため，調音結合の影響が考慮されておらず，不自然な補正音声が生成される場合がある．なお，継続長を大幅に短くするような時間伸縮を用いた際にも，同様の悪影響が生じやすい．

本稿では，品質劣化を伴わない補正処理の実現に向けて，時間伸縮に伴う局所的な品質劣化をもたらす要因に関する分析を行う．時間伸縮により生じる不自然な調音結合を捉える特徴量として，音声スペクトルパラメータ系列の変調スペクトル [17] を用いる．本稿では，メルケプストラム係数の時系列から計算される対数パワースペクトルを，変調スペクトルとする．変調スペクトルと品質の関係を調査するとともに，局所的な品質劣化箇所の自動検出についても試みる．

(注1)：図 4 は，音素挿入のみを考慮した音素ネットワークである．

3.　実験的評価

3.1　実 験 条 件

日本語母語話者の英語音声に対する発話リズム補正処理において，提案法を評価する．英語母語話者の音声には，CMU ARCTIC 音声データベース [21] 中の男性話者 1 名による 100 文を用いる．また，日本語母語話者の英語音声として，留学経験のない日本の標準的な英語教育を受けてきた男子大学院生 2 名による同一発話内容の 100 文を用いる．各日本語母語話者において，これらの英語母語話者音声 100 文および日本語母語話者音声 100 文を用いて，GMM を学習する．この内，各日本語母語話者に対して 25 文を選択し，計 50 文を評価データとして使用する．

音声のサンプリング周波数は 16 kHz，音声分析時のフレームシフトは 5 ms とする．高品質な音声分析合成系として，STRAIGHT [22] を用いる．スペクトルパラメータとして，0 次から 24 次のメルケプストラム係数を用い，音源パラメータとして，F_0 および 5 周波数帯域における平均非周期成分 [23] を使用する．GMM の学習時には，短時間波形パワーに基づき抽出された有音フレームのみを対象とし，1 次から 24 次のメルケプストラム係数およびその動的特徴量を用いる．メルケプストラム係数に対する GMM の混合数は 64 とし，非周期成分に対する GMM の混合数は 16 とする．学習された GMM に基づき，全フレームに対する変換処理を行う．その際に，短時間波形パワーは保存する．0 次から 24 次までのメルケプストラム係数および動的特徴量に対する距離尺度に基づき，DTW を行うことで，継続長補正処理のための伸縮規則を決定する．

音素挿入検出に用いる音響モデルは，5 状態 left-to-right 型の話者依存モノフォン HMM とする．観測特徴量は非母語音声の 0 次から 24 次までのメルケプストラム係数，及び，その 1 次，2 次の動的特徴量とする．HMM の出力確率密度関数は 16 混合 GMM とする．発話テキストに対して，Flite [24] によるテキスト解析を行い，音素系列を決定した後，音素挿入規則 [4] に基づき，音素ネットワークを構築する．音素挿入を含まない音素系列を用いて，モノフォン HMM を学習した後に，音素ネットワークに基づく Viterbi アライメントを実施し，音素系列を更新する．これらを複数回実施し，最終的な HMM を得る．

提案法の有効性を調査するために，以下の音声を用いて音質および継続長の自然性に関する主観評価を実施する．

Native： 英語母語話者の自然音声
ERJ： 日本語母語話者の自然音声
Feat： 音声分析再合成による補正音声（従来法）
Wave： 波形加工による補正音声（提案法）

“Feat” に関しては，推定された伸縮規則に基づき，非母語話者のスペクトルパラメータ系列および音源パラメータ系列を伸縮させた後に，音声合成処理を行う．音質評価として，5 段階オピニオンスコアを用いたオピニオン評価を実施し，平均オ

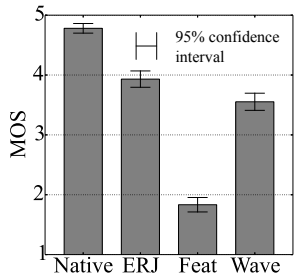

図 5　音質に関する主観評価結果
Fig. 5　5-point MOS scores on speech quality.

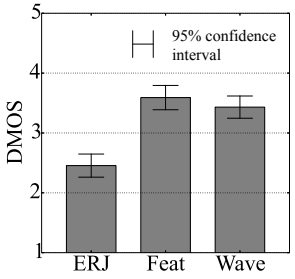

図 6　継続長の自然性に関する主観評価結果
Fig. 6　5-point DMOS scores on duration naturalness.

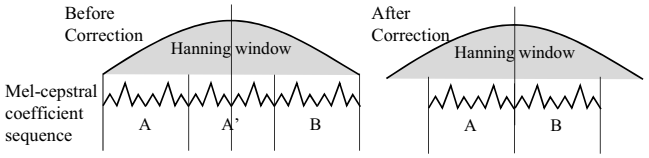

図 7　変調スペクトル分析に用いるメルケプストラムセグメントの例
Fig. 7　An example of mel-cepstral coefficient segments used for modulation spectrum analysis.

ピニオンスコア（Mean Opinion Score: MOS）を求める．一方で，継続長の自然性に関しては，“Native”をリファレンスとした 5 段階 DMOS（Degradation MOS）評価を行う．つまり，“Native” に対する各音声の発話リズムの類似度を測ることに相当する．被験者は，日本語母語話者 6 名である．ただし，この実験では HMM アライメントによる音素挿入検出は実施しない．また，波形加工による発話リズム補正の効果に着目するため，時間伸縮関数は “Feat” および “Wave” で同一のものを用いる．

3.2　音質と継続長の自然性に関する実験結果

図 5 に，音質に関する評価結果を示し，図 6 に継続長の自然性に関する評価結果を示す．図 5 から，“ERJ” と比較して “Feat” の音質が顕著に劣化していることが分かる．これに対して，音声分析再合成を行わない “Wave” では，音質劣化を大幅に低減できることが分かる．このことから，音声分析再合成による音質劣化は大きく，提案法による品質改善効果は大きいことが分かる．一方，図 6 から，“Feat” および “Wave” ともに，“ERJ” と比較し，自然性が顕著に改善していることから，発話リズム補正の効果が認められる．また，“Wave” は “Feat” と同等の自然性改善効果が得られている．以上のことから，提案法（“Wave”）による発話リズム補正により，高い音質を保持したまま，非母語話者音声の自然性を大きく改善できることが分かる．

4.　時間伸縮に伴う局所的な品質劣化の分析

4.1　分析データセットの作成

評価に使用する音声として，男性日本語母語話者 1 名により，誤り音素を意図的に多く含むように発話された CMU ARCTIC 音声データベース中の 50 文を用いた．サンプリング周波数や音声特徴量などは，3.1 節と同様とした．この 50 文に対して，

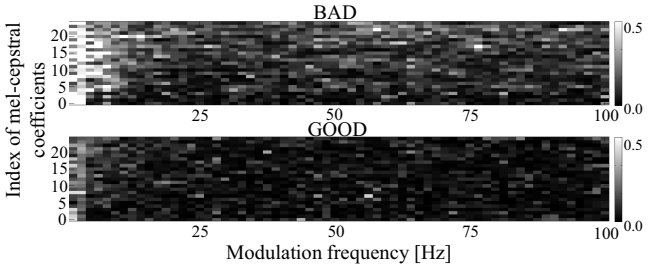

図 8　補正前と補正後におけるメルケプストラム係数の変調スペクトルの分布間距離 (KL ダイバージェンス)
Fig. 8　KL divergence of Gaussian distribution of the modulation spectrum of corrected mel-cepstral coefficients from that of original mel-cepstral coefficients.

音素挿入を考慮した音素ネットワークを用いて HMM を学習し，挿入音素の検出を行った結果，検出された挿入音素箇所は 108 箇所であった．また，男性英語母語話者 1 名の同一発話 50 文に対して，音素挿入を考慮しない HMM 学習を行った．各話者に対する HMM 状態アライメントの結果から各音素の継続長を求めた後に，対応する音素における継続長比に応じた線形伸縮として，時間伸縮規則を決定した．得られた時間伸縮規則に基づき，WSOLA による発話リズム補正処理を行った．ここで，日本人英語音声における挿入音素に対しては，時間伸縮倍率を 0 とすることで，除去処理を行った．

次に，補正音声において，時間伸縮に伴う局所的な品質劣化が生じている箇所の抽出を行った．まず，挿入音素が削除された 108 箇所に対して，男性日本語母語話者 1 名による主観評価を行い，品質劣化の有無を判定した．結果，品質劣化が生じていると判定された箇所は，38 箇所であった．さらに，挿入音素削除箇所以外に対しても，局所的な品質劣化が生じている箇所についての判定を行った．結果，50 文中で新たに検出された箇所は 22 箇所であった．また，局所的な品質劣化が生じていない箇所として，先の挿入音素が削除された 108 箇所の内，上述の 38 箇所を除く 70 箇所の中から，ランダムに 38 箇所を抽出した．挿入音素の削除が行われていない箇所からも，上述の 22 箇所を除外して，ランダムに品質劣化が生じていない箇所 22 箇所を抽出した．

上記の処理により，局所的な品質劣化が生じている箇所として計 60 箇所，局所的な品質劣化が生じていない箇所として計 60 箇所を抽出した．

4.2　変調スペクトルによる比較

4.1 節で抽出した計 120 箇所に対して，変調スペクトルに基づく分析を行う．継続長補正前後の非母語音声に対して，メルケプストラムの変調スペクトルを計算する．図 7 に変調スペクトルの分析区間を示す．補正前の音声に対しては，挿入音素の中央時刻に対応する時間フレームを中心として窓掛けを行い，変調スペクトルを計算する．補正後の音声については，挿入音素が削除された時刻に対応する時間フレームを中心として窓掛けを行い，変調スペクトルを計算する．ただし，補正後の音声のメルケプストラムは，WSOLA を施した音声波形に対する STRAIGHT 分析により抽出する．窓関数は 128 点のハニング

窓を使用し，離散フーリエ変換のタップ長は 128 点である．

局所的な品質劣化が生じている 60 箇所 (“BAD”) において，補正前後の音声の変調スペクトルを，各々多次元正規分布でモデル化する．共分散行列は対角成分のみを用いる．補正前 (自然音声) の変調スペクトルに対する多次元正規分布から，補正後 (継続長補正音声) の変調スペクトルに対する多次元正規分布への KL ダイバージェンスを求める．同様に，局所的な品質劣化が生じていない 60 箇所 (“GOOD”) に対しても，KL ダイバージェンスを求める．図 8 に示した結果から，“BAD” の方が KL ダイバージェンスが大きくなる傾向が見られ，特に低域変調周波数帯域において，“BAD” と “GOOD” の間で KL ダイバージェンスの違いが顕著である．

4.3 変調スペクトルによる識別

4.1 節で抽出した計 120 箇所を用いて，局所的な品質劣化が生じている箇所の識別実験を行う．識別器として，サポートベクターマシンを用いる．素性として，メルケプストラム系列の変調スペクトル，F_0 系列の変調スペクトル，非周期成分系列の変調スペクトルを用いる．計 120 箇所のデータセットに対して，6 分割交差検定を 10 回実施する．交差検定の結果得られた平均識別率は 76.83 %である．チャンスレベル 50 %と比べて有意に高い識別率が得られており，変調スペクトルが局所的な品質劣化を捉える特徴量として有効であることが分かる．

5. ま と め

本報告では，非母語音声として日本人英語音声に着目し，DTW および WSOLA に基づく発話リズム補正法を提案した．実験的評価結果から，提案法により，音声分析再合成処理の使用を回避することで，音質劣化を最小限に抑えつつ柔軟な継続長変換を実現することができ，自然性を大幅に改善可能な発話リズム補正処理が実現できることが分かった．また，時間伸縮に伴い生じる局所的な自然性劣化に着目し，音声スペクトルパラメータ系列の変調スペクトルに基づく分析を行った．その結果，変調スペクトルは，局所的な自然性劣化を捉えることが分かった．今後は，局所的な自然性劣化を緩和する後処理手法の検討や，発話リズム補正処理の改善に取り組む．

謝辞 本研究の一部は，JSPS 科研費 26280060 の助成を受け実施した．

文 献

[1] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, “Development of English speech database read by Japanese to support call research.,” Proc. ICA, vol.1, pp.557–560, Kyoto, Japan, Apr. 2004.

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

[3] N. Minematsu, C. Guo, and K. Hirose, “CART-based factor analysis of intelligibility reduction in Japanese English.,” Proc. INTERSPEECH, pp.2069–2072, Geneva, Switzerland, Sep. 2003.

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

[5] R.L. Allington, Fluency: The Neglected Reading Goal, vol.36, The Reading Teacher, 1983.

[6] H. Quené and L.E.V. Delft., “Non-native durational patterns decrease speech intelligibility,” Speech Communication, vol.52, no.11-12, pp.911–918, Mar. 2010.

[7] A.W. Black, “Speech synthesis for educational technology.,” Proc. SLaTE, pp.104–107, Farmington, U. S. A, Oct. 2007.

[8] ジャンドランフレデリック，廣瀬啓吉，峯松信明，“アクセント型 CALL システムのための音声合成による矯正フィードバック生成,” 電子情報通信学会技術研究報告，vol.102，no.618，pp.1–6，Jan. 2003.

[9] Y. Oshima, S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Non-native speech synthesis preserving speaker individuality based on partial correction of prosodic and phonetic characteristics,” Proc. INTERSPEECH, pp.299–303, Dresden, Germany, Sep. 2015.

[10] 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.

[11] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training.,” IEICE Trans. Inf. and Syst., vol.E90-D, no.2, pp.533–543, Feb. 2007.

[12] 高道慎之介，戸田智基，G. Neubig，S. Sakti，中村 哲，“HMM に基づく日本人英語音声合成における中学生徒の英語音声を用いた評価,” 日本音響学会 2015 年秋季講演論文集，pp.1231–1232，Sep. 2015.

[13] S. Hiroya, “Non-negative temporal decomposition of speech parameters by multiplicative update rules.,” IEEE Trans. ASLP, vol.21, no.10, pp.2108–2117, Oct. 2013.

[14] L. Juang and B.-H. Rabiner, Fundamentals of Speech Recognition, Prentice Hall Inc., 1993.

[15] 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.

[16] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech.,” Proc. ICASSP, vol.2, pp.554–557, Minneapolis, U.S.A, Apr. 1993.

[17] S. Takamichi, T. Toda, A.W. Black, and N. Satoshi, “A postfilter to modify the modulation spectrum in HMM-based speech synthesis.,” Proc. ICASSP, vol.2, pp.290–294, Florence, Italy, May 2014.

[18] K. Cho and J.G. Harris, “Towards an automatic foreign accent reduction tool,” Proc. Speech Prosody, vol.2, pp.569–572, Dresden, Germany, May 2006.

[19] K. Hirose, F. Gendrin, and N. Minematsu, “A pronunciation training system for Japanese lexical accents with corrective feedback in learner’s voice,” Proc. EUROSPEECH, pp.3149–3152, Geneva, Switzerland, Sep. 2003.

[20] Y. Tsubota, T. Kawahara, and M. Dantsuji, “Recognition and verification of English by Japanese students for computer-assisted language learning system.,” Proc. ICSLP, pp.1205–1208, Denver, U. S. A, Sep. 2002.

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

[22] H. Kawahara, I. Masuda-Katsuse, and A.D. Cheveigné, “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.

[23] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation,” Proc. INTERSPEECH, pp.2266–2269, Pittsburgh, U. S. A, Sep. 2006.

[24] A.W. Black and K.A. Lenzo, “Flite: a small fast run-time synthesis engine,” 4th ISCA Workshop on Speech Synthesis, p.204, Perthshire, Scotland, Aug. 2001.