音声波形加工に基づく非母語音声の継続長補正による品質劣化の分析 *

☆倶羅 真也, 高道 慎之介 (奈良先端大), 戸田 智基 (名大/奈良先端大),
ニュービッグ グラム, 中村 哲 (奈良先端大)

1 はじめに

言語学習等への応用を見据え，我々は，音声波形の直接加工に基づく非母語音声の継続長補正法 [1] を提案している．この手法は，同一発話文の母語音声の継続長に一致するように非母語音声波形を時間伸縮させることで，音声分析合成処理による品質劣化を回避した高音質な補正処理を可能にする．一方で，母語音声と非母語音声間において音素系列が同一であると仮定するため，非母語音声において生じ得る音素の挿入，置換，削除誤りにより，品質劣化が生じる．

本稿では，音声波形の時間伸縮においても対応できる可能性が高い挿入誤りについて対処するとともに，変調スペクトル (MS: Modulation Spectrum) [2] を用いた局所的な自然性劣化の自動検出法について検討する．分析結果から，MS を用いた自然性劣化箇所の検出が有効であることを示す．

2 波形加工に基づく非母語音声の継続長補正

非母語話者と母語話者の同一発話音声を用いて，母語音声の継続長と一致するように非母語音声を時間伸縮する．まず，メルケプストラムひずみを距離尺度とした動的時間伸縮 (Dynamic Time Warping: DTW) [3] により，非母語音声に適用する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 [4] により非母語話者と母語話者の話者性の差異による影響を緩和する．次に，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 (Waveform Similarity OverLap-Add) [5] を用いた継続長補正を施す．これにより，波形の切り出しおよび重畳加算のみに基づく高品質な補正処理を実現する．

3 挿入音素の検出及び削除

日本人英語の音素挿入箇所を自動検出し，時間伸縮処理により削除する処理を施す．日本人英語の音素挿入の規則 [6] に基づき，発話テキストの音素列から音素挿入を含む音素ネットワークを規則的に構築する．この音素ネットワークを用いて，非母語音声に対して隠れマルコフモデル (Hidden Markov Model: 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 [7]．検出された挿入音素に対しては，継続長を 0 とするように時間伸縮規則を設定することで，当該箇所を削除する．

4 品質劣化箇所を捉える特徴量

挿入音素を削除する際，挿入音素の先行音素と後続音素の音声波形を直接接続するため，適切な調音結合が実現されず，不自然な補正音声が生成される場合がある．例として，補正音声の品質劣化箇所と，自然音声の当該箇所のスペクトログラムを Fig. 1 に

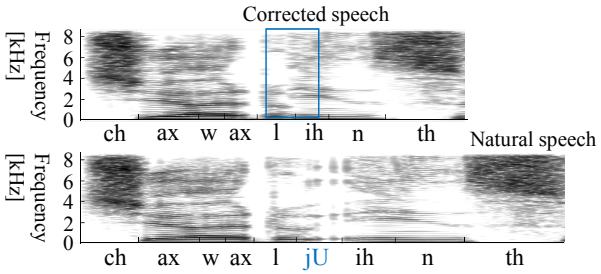

Fig. 1 品質劣化箇所のスペクトログラムの例

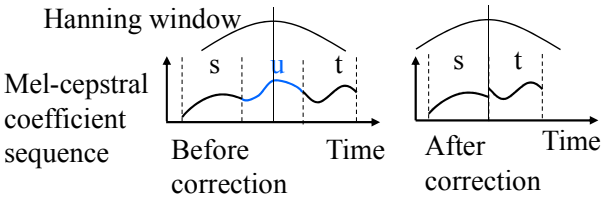

Fig. 2 MS 分析分析区間の例

示す．挿入音素 “jU” を削除した箇所において，自然音声と比較してスペクトログラムの不自然な遷移が生じていることが分かる．なお，この問題は，挿入音素を削除した箇所のみで生じるわけではなく，他の継続長補正箇所においても生じ得る．

本稿では，この問題の対処に向けて，時間伸縮に伴う局所的な品質劣化箇所を自動検出するために，MS に着目する．まず，継続長補正前後の非母語音声に対して，STRAIGHT 分析により抽出したメルケプストラムの MS を計算する．補正前後の MS を比較することで，品質劣化箇所の分析及び検出に取り組む．Fig. 2 に MS 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，MS を計算する．補正後の音声に対しては，削除された挿入音素の先行音素の終了時刻に対応する時間フレームを中心とし，MS を計算する．

5 実験的評価

5.1 実験条件

本稿では，日本人英語音声の補正処理を対象とする．分析に用いる英語母語話者の音声には，CMU ARCTIC 音声データベース [8] 中の男性話者 50 文を用いる．また，日本人英語音声として，男子大学院生 1 名による，誤り音素を多く含んだ同 50 文を用いる．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．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．日本人英語音声に対しては，音素挿入を考慮した HMM 学習を行う．一方で，英語母語話者音声に対しては，音素挿入を考慮しない HMM

* Analysis of quality degradation caused by duration correction of non-native speech using direct waveform modification. by KURA, Shinya, TAKAMICHI, Shinnosuke (NAIST), TODA, Tomoki (Nagoya Univ./ NAIST), NEUBIG, Graham, NAKAMURA, Satoshi (NAIST)

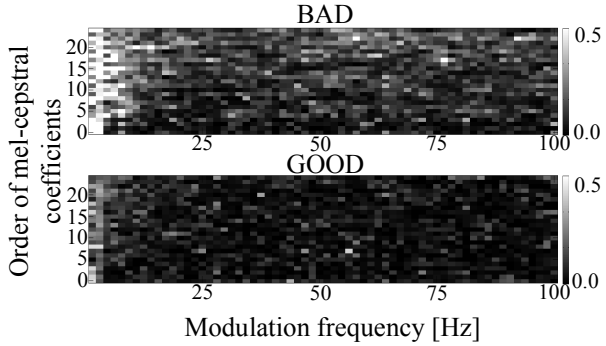

Fig. 3　補正前と補正後におけるメルケプストラム係数の MS の分布間距離 (KL ダイバージェンス)

学習を行う．観測特徴量は 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．自然性劣化の検出に用いる MS は，128 点のハニング窓により窓かけされた音声パラメータセグメントに対して，128 点のフーリエ変換を施して計算する．

品質劣化箇所の分析を行うため，データセットを作成する．まず，日本人英語音声に対して，挿入音素の検出を行う．結果，得られた挿入音素箇所は 108 箇所である．本実験では，両言語の同一発話音声に対する 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する．日本人英語音声の挿入音素に対しては，時間伸縮倍率を 0 とすることで削除する．得られた時間伸縮規則を用いて，日本人英語音声の継続長を補正する．

次に，補正音声において品質劣化箇所の分析を行う．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 名による聴取を通して，品質劣化の有無を判定する．結果，品質劣化が生じていると判定された箇所は，38 箇所である．また，挿入音素以外の箇所についても同様に聴取を行い，品質劣化の有無を判定する．結果得られた品質劣化箇所は 22 箇所である．加えて，品質劣化が生じていない箇所として，挿入音素箇所からランダムに 38 箇所，挿入音素以外の箇所から 22 箇所をランダムに抽出する．結果得られた品質劣化が生じている箇所 60 箇所 (“BAD”) と，品質劣化が生じていない箇所 60 箇所 (“GOOD”) に対して，分析および識別実験を行う．

5.2　品質劣化の分析

“BAD” の 60 箇所において，継続長補正前後の音声の MS を抽出し，各々多次元正規分布でモデル化する．補正前 (自然音声) の MS に対する多次元正規分布から，補正後の MS 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，“GOOD” の 60 箇所に対しても，KL ダイバージェンスを求める．

各メルケプストラム係数の各変調周波数成分において求められた KL ダイバージェンスを Fig. 3 に示す．“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る．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．

5.3　品質劣化の識別

次に，抽出したデータセットを用いて，サポートベクターマシンによる品質劣化の識別実験を行う．素性として，メルケプストラム系列の MS，F_0 系列の MS，非周期成分 [9] 系列の MS，時間伸縮規則を用いる．MS については，継続長補正前後の差分を素性とし，時間伸縮規則については，WSOLA によって削除された挿入音素の継続長を素性とする．全ての素性に対して正規化を行い，線形カーネルに基づく 6 分割交差検定を 10 回実施し，平均識別率を求める．まず，全ての素性を用いた識別率を得た後，識別率への影響が小さい特徴量から順に取り除き，残った素性で再度識別率を計算する．この手順を，素性が一つになるまで繰り返す．

Table 1　品質劣化箇所の識別結果

素性	識別率
(1) メルケプストラム系列の MS	69.10 %
(2) (1) + 非周期成分系列の MS	71.67 %
(3) (2) + F0 系列の MS	72.50 %
(4) (3) + 時間伸縮規則	76.83 %

結果を，Table 1 に示す．メルケプストラム系列の MS が，素性の中で識別率が最も高く，音声の品質劣化を最も捉えやすいことを示す．さらに，他の素性も併用することで識別率が向上し，全ての素性を用いることで 76.83 %の識別率が得られる．このことから，チャンスレベル 50 %と比べて有意に高い精度で，品質劣化箇所を検出できることが分かる．

6　まとめ

本稿では，日本人英語の継続長補正を対象として，波形加工に基づく補正処理における音素挿入誤りへの対応と，それにより生じ得る局所的な品質劣化の分析および識別を行った．その結果，スペクトル包絡の変調スペクトルが，品質劣化を捉える特徴量として有効であることが示された．今後は，品質劣化緩和のための波形加工処理について検討する．

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

参考文献

[1] 倶羅 他, 音講論（春）, 1–2–8, 2015.
[2] R. Drullman *et al., J. Acoust. Soc. of America*, Vol. 95, pp. 2670–2680, 1994.
[3] L. Rabinar *et al.*, “Fundamentals of Speech Recognition,” Prentice Hall Inc., 1993.
[4] T. Toda *et al., IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007.
[5] W. Verhelst *et al. Proc. of ICASSP*, Vol. 2, pp. 554–557, 1993.
[6] S. Kohmoto, “Applied English phonology: teaching of English pronunciation to the native Japanese speaker,” Tanaka Press, 1965.
[7] Y. Tsubota *et al., Proc. of ICSLP*, pp. 1205–1208, 2002.
[8] J. Kominek and A. W. Black, *Tech Report*, CMU-LTI-03-177, 2003.
[9] Y. Ohtani *et al., Proc. of INTERSPEECH*, pp. 2266–2269, 2006.