非母語話者の音声に対する継続長補正の評価*

☆倶羅 真也, 高道 慎之介, 戸田 智基 (奈良先端大), 亀岡 弘和 (NTT／東京大),
Graham Neubig, Sakriani Sakti, 中村 哲 (奈良先端大)

1 はじめに

我々はこれまでに，HMM（Hidden Markov Model）音声合成 [1] を用いた韻律補正法を提案しており，話者性を保持したまま日本語母語話者による英語発話の自然性を改善できることを報告している [2]．本手法により，任意のテキストに対応した日本語母語話者の英語音声が合成可能となるが，その音質は HMM 音声合成の性能に限定され，十分なものとは言い難い．

本稿では，より品質の高い日本語母語話者による英語音声を合成する手法として，日本語母語話者が発話した英語音声に対して直接変換処理を施すことで，自然性を改善する手法を検討する．日本語母語話者英語音声と同一発話内容の英語母語話者英語音声が入手可能である場合を想定し，日本語母語話者音声に対して動的時間伸縮（DTW： Dynamic Time Warping）[3] を施すことで，継続長を補正する．その際には，声質変換技術を用いて，話者の違いに起因する音響特徴量の差異を補正する．また，継続長補正処理として，ボコーダに基づく特徴量ベースの手法と，ボコーダを用いない波形ベースの手法を適用する．実験的評価結果から，波形ベースの手法は特徴量ベースの手法と比較し，音質劣化を大幅に低減しつつ継続長補正が可能であることを示す．

2 DTW を用いた非母語話者音声の継続長補正

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

2.1 伸縮規則の推定

伸縮規則の推定のために，スペクトルひずみを距離尺度として用いて DTW を行う．その際に，スペクトルパラメータは話者の影響を強く受けるため，十分な精度で伸縮規則を推定することは困難となる．そこで本稿では，GMM （Gaussian Mixuture Model） に基づく声質変換 [4] を用いて，スペクトルパラメータにおける話者性の影響を緩和することで，伸縮規則の推定精度の向上を図る．まず，非母語話者音声と母語話者音声のスペクトルパラメータを用いて DTW を行い，時間的に対応付けられたパラレルデータを作成する．次に，そのパラレルデータを用いて GMM を学習し，得られた GMM に基づき非母語話者のスペクトルパラメータを母語話者のスペクトルパラメータへと変換する．その後，変換スペクトルパラメータと母語話者のスペクトルパラメータ間において，再度伸縮規則を推定し，パラレルデータを更新する．この処理を繰り返し行うことで，伸縮規則の推定精度を向上させる．

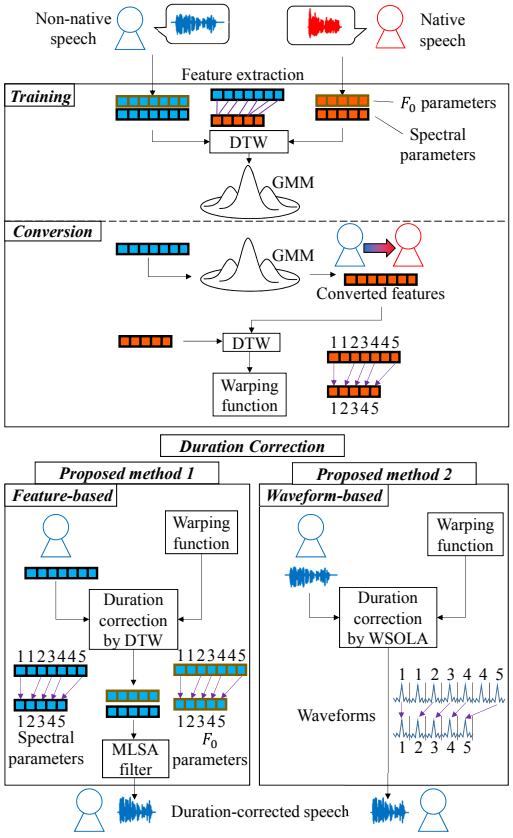

Fig. 1 提案法の処理手順

2.2 継続長の補正

推定された伸縮規則に基づき，非母語話者音声に対して継続長補正処理を行う．

特徴量ベースの補正法： 非母語話者音声に対して音声分析処理を行うことで，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抽出する．推定された伸縮規則に基づいて，両パラメータ系列を伸縮させた後に，ボコーダによる音声合成処理を行うことで，継続長補正された非母語話者音声を生成する．

波形ベースの補正法： 推定された伸縮規則に基づき，非母語話者の音声波形に対して直接補正処理を施す．補正処理には，音声波形の連続性を考慮して，短時間波形の切り出し，間引き，繰り返し，シフト，重畳加算を行うことで時間軸伸縮を行う WSOLA（Waveform Similarity Based OverLap-Add）[6] を用いる．本手法は，ボコーダの使用に伴う音質劣化を回避することができるため，高い音質を保持した継続長補正処理が実現できると期待される．

*An Evaluation of Duration Correction for Non-Native Speech. by KURA, Shinya, TAKAMICHI, Shinnosuke, TODA, Tomoki (NAIST), KAMEOKA, Hirokazu (NTT/The University of Tokyo), NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

3　実験的評価

3.1　実験条件

本稿では，日本語母語話者による英語発話を想定して，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 [7] 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 100 文を用いる．GMM の学習データとして，各日本語母語話者に対して，英語母語話者音声 100 文および同一発話内容の日本語母語話者音声 100 文を使用する．その中から，各日本語母語話者につき 25 文を選択し，合計 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．音声分析合成系として，高品質なボコーダである STRAIGHT[8] を用いる．スペクトル特徴量として，0 次から 24 次のメルケプストラム係数を用い，音源特徴量として，F_0 および 5 周波数帯域における平均非周期成分 [9] 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て，DTW を行う．最終的に得られた GMM に基づき，全フレームに対する変換処理を行い，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て 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．

提案法による継続長補正の効果を確認するために、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： 特徴量ベースの補正法による補正音声
Wave： 波形ベースの補正法による補正音声

音質評価には，5 段階 MOS（Mean Opinion Score）評価法を用いる．継続長の自然性に関する評価には，“Native” をリファレンスとした 5 段階 DMOS（Degradation MOS）評価法を用いる．被験者は，日本語母語話者 6 名である．

3.2　実験結果

Fig. 2 に，音質に関する評価結果を示し，Fig. 3 に継続長の自然性に関する評価結果を示す．Fig. 2 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．この原因として，ボコーダによる音声分析再合成処理の影響が考えられる．一方で，分析再合成を行わない “Wave” では，音質低下を大幅に緩和できていることが分かる．なお，補正処理前の自然音声 “ERJ” と比較すると，“Wave” による継続長補正音声は，若干の音質劣化が見られる．この原因として，WSOLA による処理のひずみのみでなく，音質の評価に継続長の自然性が影響を与えた可能性が考えられる．これは，自然音声にも拘わらず，“ERJ” が “Native” よりも大きく音質が劣化している結果からも推察される．

一方，Fig. 3 から，提案法により継続長の自然性が顕著に改善できることが分かる．また，“Feat” のスコアと比較して “Wave” のスコアは若干低下しているものの，有意な差はなく同等の自然性を持つ継続長を実現できる．このことから，ボコーダを使用せずとも，波形上での処理で十分な継続長補正が可能であることが分かる．

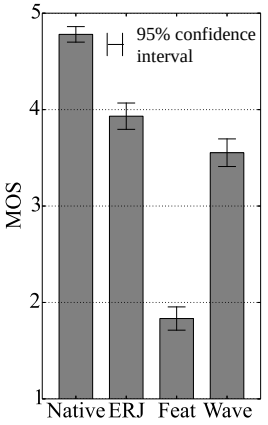

Fig. 2　音質に関する主観評価結果

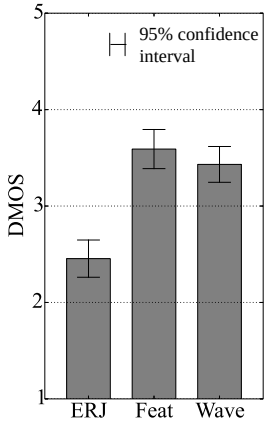

Fig. 3　継続長の自然性に関する主観評価結果

以上の結果から，提案法による継続長補正により，非母語話者英語音声の自然性を改善可能であり，また，ボコーダによる分析再合成処理の回避により，音質低下を緩和できることが分かる．

4　おわりに

本稿では，非母語話者音声の自然性を改善する手法として，動的時間伸縮に基づく継続長補正法を提案した．実験的評価結果から，提案法による自然性を大幅に改善することが可能であり，また，ボコーダによる分析再合成処理を回避することで音質低下を大幅に緩和可能であることが明らかになった．今後は，音韻誤りの考慮，パワー・F_0 における補正の影響，また，初等教育中の話者への適応を検討する．

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

参考文献

[1] K. Tokuda *et al., Proc. of IEEE*, Vol. J96-D, No. 7, pp. 1549–1565, 2013.
[2] 大島 他，信学技報，Vol. 114，No. 365，SP2014-112，pp. 63–68，2014.
[3] L. Rabinar *et al.*, “Fundamentals of Speech Recognition,” Prentice Hall, 1993.
[4] T. Toda *et al. IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007.
[5] 今井 他，信学論 (A)，Vol. J66-A，No. 2，pp. 122–129，1983.
[6] W. Verhelst *et al. Proc. of ICASSP*, Vol. 2, pp. 554–557, 1993.
[7] J. Kominek and A. W. 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] 大谷 他，信学論，Vol. J91-D, No. 4, pp. 1082–1091, 2008.