

高品質声質変換のための特徴量分析再訪*

☆ 須田仁志, 小谷岳, 高道慎之介, 齋藤大輔 (東大)

1 はじめに

声質変換は入力音声の音響特徴量が所望の特性に近づくよう音声を変換することをいう。声質変換を実現するには、入力音声から音響特徴量を分析し、何らかの変換モデルを用いて特徴量を変換し、その特徴量から目標音声を合成する。本稿では特に声質変換のうち、言語内容を変えずに話者情報のみを変換する話者変換を扱う。

声質変換研究においてはその変換モデルに関する検討が多くなされてきた [1–3]。一方で分析合成に着目した議論がなされることは少ない。しかし声質変換はどのような変換モデルを用いるにしろ最終的には音声合成が必要となるため、分析および合成時のパラメータ選択は変換音声の品質に少なからず影響を与えると考えるのが自然である。そこで本稿では、分析法として WORLD [4] (D4C edition [5]) を、特徴量としてメルケプストラム係数を、合成法として差分スペクトル補正 [6] を用いた一対一声質変換を対象に、さらに高品質な声質変換のための特徴量分析および合成パラメータ選択法を検討する。

本稿では分析再合成のみで変換を行わない場合、存在するパラレルデータを用いて理想的な特徴量変換が行えた場合、そして実際に混合ガウスモデル (GMM: Gaussian Mixture Model) を用いて変換した場合のそれぞれについて、最適なパラメータの組み合わせを検討した。

2 実験条件

主観実験では、合成音声の自然性に対する評価を A/B テストにより行った。データセットとして男女それぞれ 2 話者の読み上げた ATR 日本語音素バランス 503 文 [7] のうち A セット 50 文を用いた。サンプリング周波数は 22050 Hz とした。被験者はどの実験においても 23 名とし、各被験者は 10 問ずつ回答した。なお各グラフにおけるエラーバーは 95% 信頼区間を表す。

3 分析再合成に関わるパラメータの検討

本節では変換音声との比較として、変換を行わず分析再合成を行った場合の品質を評価する。本節自体は最終的な変換音声の品質評価にはあたらないが、変換の有無による最適パラメータの差異に関して討究を行うことができる。

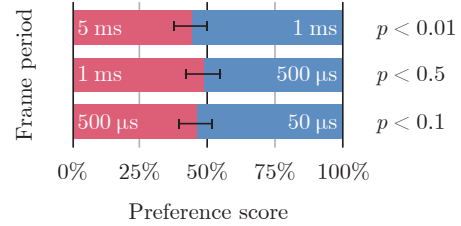


Fig. 1 分析合成時におけるフレームのシフト長による合成音声の自然性の比較。

3.1 分析フレームのシフト長

音声の分析時には分析フレームの間隔 (シフト長) を任意に選ぶことができる。シフト長を短くすることで時間軸に対してより緻密な分析合成が行えるため、合成品質の向上が期待できる。

3.2 メルケプストラム係数の次数

WORLD ボコーダでは合成時のスペクトル包絡としてパワースペクトルを用いる。しかし声質変換において GMM などに変換モデルを構築する場合パワースペクトル自体を扱うことは難しいため、特徴量としてメルケプストラム係数を用いることが一般的である。メルケプストラム係数の次数が高いほど緻密にスペクトル包絡を表現することができる。

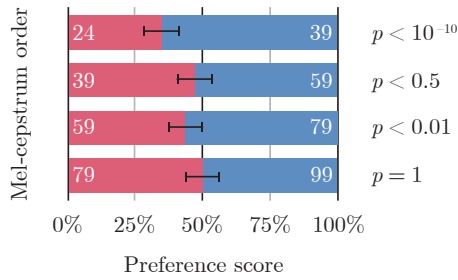
3.3 主観評価実験

ここでは分析再合成における分析フレームのシフト長とスペクトル包絡の圧縮の効果を自然性評価により検討する。

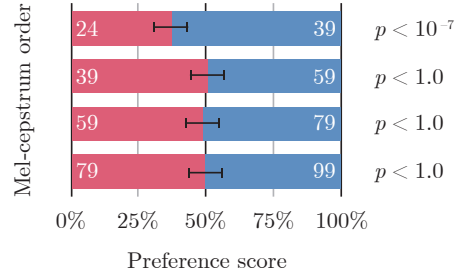
まずシフト長を 5 ms から 50 μ s まで 4 通りに選び、再合成音声に関して自然性の評価を行った。ここで 50 μ s は 1 サンプルに相当する。スペクトル包絡はメルケプストラム係数に圧縮せずそのままとした。評価結果を Fig. 1 に示す。全体としてシフト長を短くするほど合成音声の自然性が高くなるが、1 ms より短い場合は有意差は現れなかった。この結果から分析再合成においては 1 ms が適当なシフト長であると考えられる。

またシフト長を 1 ms もしくは 50 μ s とし、それぞれにおいてメルケプストラム係数を 24 次から 99 次まで 5 通りに選び、自然性の評価を行った結果を Fig. 2 に示す。次数が高いほど合成音声の自然性も高まるが、シフト長が 1 ms では 79 次元が、50 μ s では 39 次元が明確な品質向上の限界となった。

*A Revisit to Feature Analysis for High Quality Voice Conversion, by Hitoshi SUDA, Gaku KOTANI, Shinnosuke TAKAMICHI, Daisuke SAITO (the University of Tokyo).



(a) フレームのシフト長が 1 ms の場合



(b) フレームのシフト長が 50 μ s の場合

Fig. 2 分析合成時におけるスペクトル包絡の圧縮次数による合成音声の自然性の比較.

3.4 本節のまとめ

本節では WORLD 分析合成を用いて分析再合成を行った場合の合成音声品質の評価を行った。フレームのシフト長は短いほど、メルケプストラムの次数は高いほど合成音声の自然性は高くなるが、どちらも一定以上では有意な影響を与えなくなる。

シフト長が短いほど品質が向上するのは、合成時にフレーム間での特徴量の補間が少なくなるためと考えられる。またメルケプストラム係数の次数が高いほど品質が向上するのは、もとのスペクトル包絡を再現する能力が高くなるためと考えられる。

4 変換に関わるパラメータの検討

本節では声質変換を想定して、実際に存在するパラレルデータを用いて差分スペクトル補正を行った際の合成音声に対して評価を行う。すなわち理想的な特徴量変換が行えたと仮定して合成を行った際の合成品質について評価する。ここでは入出力話者をともに男性とし、基本周波数の変換は行わない。

4.1 アフィン DTW

声質変換に際してパラレルデータを用いる場合、発話内容が同一であっても話者ごとに話速などが異なる。そのためそれぞれの特徴量系列に対して時間的なアラインメントを取る必要があり、これには動的時間伸縮 (DTW: Dynamic Time Warping) を用いることが一般的である。DTW では 2 話者のメルケプストラム係数系列の二乗誤差が最小となるように経路を探索するが、メルケプストラム係数には言語情報のみならず話者の情報も含まれるため、単純

な二乗誤差規準での比較では話者性の違いがアラインメントの精度に影響する。

本稿ではこの話者性の違いによる影響を軽減するため、入力話者のメルケプストラム係数にアフィン変換をかけて DTW を行う。アフィン変換を行う場合は係数行列の推定が必要となるため、初期値を恒等変換として DTW を行い、その結果からアフィン変換のパラメータを推定する。この操作を反復することで話者性を吸収した DTW を行うことができ、この手法をアフィン DTW と呼ぶことにする。

なおアフィン変換によるメルケプストラム係数の変換は、混合数 1 の GMM による声質変換と等価である。したがって、アフィン DTW は大域的な声質変換と DTW を繰り返していることに相当する。

4.2 WORLD 差分スペクトル補正

差分スペクトル補正は声質変換を対象とした波形生成法で、変換前と変換後の特徴量から推定したフィルタを入力音声に畳み込むことで変換音声を作成する手法である。ボコーダを用いて合成する場合と比べ、より自然な音声を合成することができる。

差分スペクトル補正においては、メルケプストラム係数から Padé 近似した MLSA (Mel Log Spectral Approximation) フィルタを用いる手法が一般的である [8]。本稿ではこの近似を行わず、メルケプストラム係数から得たスペクトル包絡の実ケプストラムを用いて最小位相フィルタを生成する手法も検討する。この手法は WORLD の合成部において用いられている手法であり、これを利用して本手法を実装した。ここでは合成法の名前として、前者を MLSA、後者を WORLD と呼ぶ。

4.3 主観評価実験

ここでは、波形合成法、メルケプストラム係数の次数、シフト長の最適な組み合わせを自然性評価により検討する。合成法は MLSA と WORLD、メルケプストラム係数は 24 次と 79 次、シフト長は 5 ms と 1 ms のそれぞれ 2 通りを候補とした。時間的なアラインメントは用いるメルケプストラム係数の次数に関わらず、24 次メルケプストラム係数でのアフィン DTW の結果を用い、反復は 4 回 (アフィン変換を 3 回, DTW を 4 回) 行った。データセットとなる 50 文のうち、アフィン DTW におけるアフィン変換係数の推定には 40 文を用い、評価には 10 文を用いた。MLSA における Padé 近似の次数は 4 とした。

まずメルケプストラム係数の次数とシフト長の各組み合わせにおける合成法の評価結果を Fig. 3 に示す。どの組み合わせにおいても WORLD の有効性が示された。

また合成法とシフト長の各組み合わせにおけるメルケプストラム係数の次数の評価結果を Fig. 4 に示

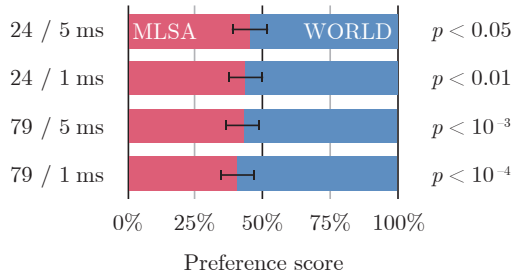


Fig. 3 合成法の違いによる合成音声の自然性の比較.

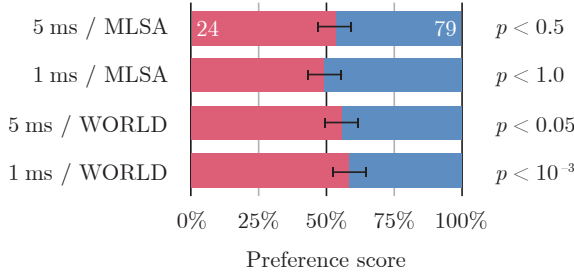


Fig. 4 メルケプストラム係数の次数の違いによる合成音声の自然性の比較.

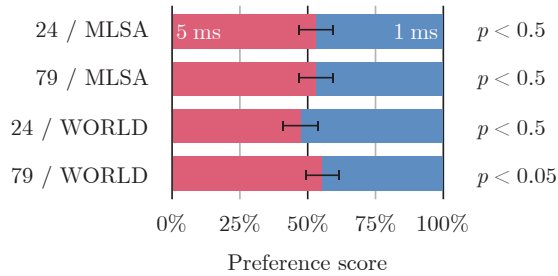


Fig. 5 分析シフト長の違いによる合成音声の自然性の比較.

す. MLSA では有意差が見られず, WORLD ではシフト長に関わらず 24 次において自然性が高い.

さらに合成法とメルケプストラム係数の次数の各組み合わせにおけるシフト長の評価結果を Fig. 5 に示す. WORLD 合成法と 24 次メルケプストラム係数の組み合わせでは, 有意差がないものの 1 ms のほうが適しているといえる.

4.4 本節のまとめ

本節では理想的な特徴量変換が可能であったと仮定して, 分析合成時パラメータの組み合わせによる合成音声の自然性を主観実験によって評価した. 評価結果より, 合成手法に WORLD 合成法を, 特徴量に 24 次メルケプストラム係数を, フレームのシフト長に 1 ms を選ぶ組み合わせにおいて, 最も自然性の高い変換音声の合成が可能となるといえる.

WORLD 合成法が有効なのは, Padé 近似を介する MLSA に比べて特徴量からよくフィルタを再現することができるためと考えられる. またメル

ケプストラム係数の次数に関しては, 完全な特徴量変換ができないことを考えれば, 複雑な特徴量で緻密に推定し誤差を生むよりも次数の低い曖昧な特徴量を用いて変換するほうが聴覚上優れているということになる. ここでは理想的な特徴量変換であると仮定したが, 時間的な構造のミスマッチが残っておりこのような結果になったと考えられる.

5 統計的変換に関わるパラメータの検討

本節では GMM により変換モデルを学習し声質変換を行ううえで最適な条件を主観評価実験により検討する. ここでは特徴量を 24 次メルケプストラム, シフト長を 1 ms で固定し, 波形合成法のほか特徴量生成法に関して種々の組み合わせを検討した.

5.1 動的特徴量と系列内変動

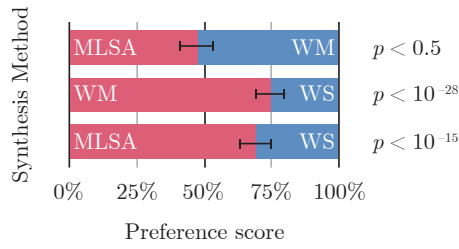
GMM 声質変換において最小平均二乗誤差規準で特徴量を生成する場合, 各フレームが独立に変換されるため隣接した特徴量の連続性が保持されない. そこで静的な特徴量にその動的成分を連結した静的動的特徴量を用いることで, 特徴量の不自然な遷移を抑制することができる [9]. また同様に発話内の生成特徴量の分散が自然音声よりも小さくなる過平滑化についても, 発話内の分散を補償するようにパラメータ生成を行うことで改善できる [9].

5.2 逆フィルタの近似

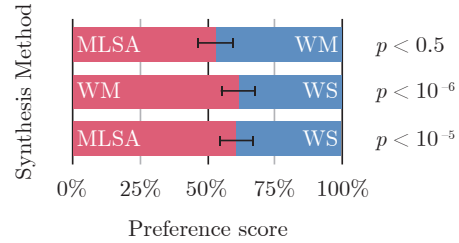
差分スペクトル補正は, 入力音声に対して入力話者の声道特性の逆フィルタと出力話者の声道特性の順フィルタを順次適用している操作と解釈できる. MLSA を用いる場合はメルケプストラム領域での差によって上述のフィルタを実現しており, 両フィルタに用いるメルケプストラム係数の次数は同一と仮定している. 一方で WORLD を用いる場合はスペクトル領域での除算によって実現するため, 分母にあたる入力話者の声道特性フィルタに関しては変換モデルに制限されない特徴量を採用できる. すなわち, 低次のメルケプストラム係数から復元されたスペクトル包絡以外にも, 分析時のスペクトル包絡を直接用いることができる.

5.3 主観評価実験

特徴量生成については静的特徴量のみを用いた場合, 静的動的特徴量を用いた場合, 静的動的特徴量を用い系列内変動を考慮した場合の 3 通りとした. また合成法については, MLSA, 逆フィルタとして出力特徴量と同じ次数のメルケプストラム係数を用いた WORLD, 分析時のスペクトル包絡を用いた WORLD の 3 通りとした. ここでは WORLD の両合成法をそれぞれ WM および WS と表記する. GMM の混合数は動的特徴量を用いない場合は 64, 用いる場合は 128 とした. また入力と出力の自己共



(a) 自然性



(b) 話者性

Fig. 6 GMM 声質変換を行った際の合成手法による合成音声の主観評価結果.

分散および相互共分散は対角行列を仮定した.

合成法の違いによる変換音声の主観評価結果を Fig. 6 に示す. ここでは静的動的特徴量を用い, また系列内変動を考慮して変換を行った. 自然性と話者性の両方に同様の傾向が見られ, MLSA および WM の間には性能の有意差がなく, またどちらの手法にも WS の性能は及ばなかった.

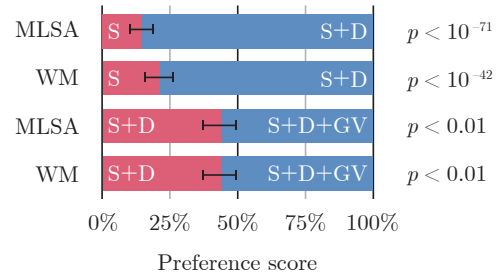
また用いた特徴量の違いによる変換音声の主観評価結果を Fig. 7 に示す. 自然性と話者性のどちらにおいても, また MLSA および WM のどちらの合成法においても動的特徴量および系列内変動の有効性が示された.

5.4 本節のまとめ

本節では統計的声質変換における合成法および特徴量による効果を検討した. 特徴量の選択においては, 動的特徴量および系列内変動の有効性が差分スペクトル補正による声質変換でも示された. 合成法においては, 従来法の MLSA と逆フィルタにメルケプストラム係数を用いた WORLD では同程度の性能を確認できたが, 逆フィルタに分析時スペクトル包絡を選ぶことは品質の向上につながらなかった. 差分フィルタをメルケプストラム係数から生成する場合は入力音声の特徴量に表現されない高次の特徴量が合成音声に引き継がれるが, 逆フィルタに分析時スペクトル包絡を選ぶ場合はこの高次の特徴量が失われ, 合成音声の品質が低下したと考えられる.

6 おわりに

本稿では, より高品質な声質変換を実現するため, 特徴量分析および波形合成時のパラメータ選択法



(a) 自然性



(b) 話者性

Fig. 7 GMM 声質変換を行った際の特徴量による合成音声の主観評価結果. S は静的特徴量のみの場合, S+D は静的動的特徴量を用いた場合, S+D+GV は S+D に加え系列内変動を考慮した場合を示す.

を, 主観評価を交えて検討した. また新たな差分スペクトル補正における合成法として WORLD 差分スペクトル補正を提案した. 今後は, GMM 声質変換におけるシフト長やメルケプストラム係数の次数による効果や, 基本周波数変換法の比較も検討する.

謝辞

本研究の一部は, JSPS 科研費 16H06681 およびセコム科学技術支援財団の助成を受け実施した.

参考文献

- [1] Y. Stylianou, *et al.*, IEEE Trans. on Speech & Audio Process., 6 (2), pp. 131–142, 1998.
- [2] R. Takashima, *et al.*, In Proc. IEEE Spoken Language Technology Workshop, 2012.
- [3] S. Desai, *et al.*, In Proc. ICASSP, 2009.
- [4] M. Morise, *et al.*, IEICE Trans. on Information & Systems, E99-D (7), pp. 1877–1884, 2016.
- [5] M. Morise, Speech Communication, 84, pp. 57–65, 2016.
- [6] K. Kobayashi, *et al.*, In Proc. INTERSPEECH, 2014.
- [7] 匂坂芳典, 浦谷則好, 音響誌, 48 (12), pp. 878–882, 1992.
- [8] 今井聖 他, 信学論 A, J66-A (2), pp. 122–129, 1983.
- [9] T. Toda, *et al.*, IEEE Trans. on Audio, Speech, & Language Process., 15 (8), pp. 2222–2235, 2007.