特定話者の同一文発話間におけるスペクトル特徴量変動とその予測*

☆犬飼辰夫, 戸田智基, Graham Neubig, Sakriani Sakti, 中村哲 (奈良先端大)

1 はじめに

統計的手法に基づく声質変換 [1, 2] は，言語情報を保持したまま所望の非言語情報を変換する技術であり，様々な応用が期待されている．声質変換において変換モデルの学習や変換精度の評価は通常，変換音響特徴量が目標音響特徴量に近づくほど良いという考えに基づき行われる．しかし仮に同一話者が同一文を発声した音声においても，音響特徴量は必ずしも一致しない．そのため，同一文発話間における音響特徴量の変動量を調査することは，声質変換精度の評価をより正確にするためにも有用である．

本稿では音響特徴量としてスペクトル特徴量 (メルケプストラム) に着目し，特定話者による同一文発話間における変動量 (メルケプストラム歪) について調査する．さらに韻律がスペクトル特徴量に与える影響 [3] に着目し，同一文発話間において各種韻律特徴量の変動量からメルケプストラム歪を予測する手法を提案する．実験的評価により提案法の有効性を示す．

2 同一文発話間のスペクトル特徴量変動

2.1 特定話者による同一文発話データ

声質変換処理の中で特にリアルタイム変換処理においては，韻律特徴量に対する複雑な変換は行わず，F_0 範囲の調整といった簡易的な処理がしばしば行われる．この場合，理想的な変換結果は，目標話者が元話者音声の韻律を真似て発声した際の音声となる．そこで本稿では，特定話者による同一文発話として，他の複数話者による同一文発話の韻律を参照して発声された音声データを取り扱う．その際に，特定話者が他の話者の韻律を真似やすくするために，参照される話者の音声に対して分析合成処理を施し，F_0 範囲を調整したものを参照音声として用いる．具体的には，対数 F_0 の平均値及び分散値が特定話者のものと一致するように線形変換を施す．

2.2 スペクトル特徴量変動

同一文発話対において計算されるメルケプストラム歪の頻度分布を Fig. 1 に示す．ここでは，男性話者一名がある一文に対し，24 名の別話者による同一文発話を参照として 8 回繰り返し発話した計 192 発話の音声データを用いる．全ての同一文発話対に対し，波形パワーにより自動抽出される有音フレーム系列において動的時間伸縮 (dynamic time warping: DTW) を行うことで，歪計算を行う．Fig. 1 から同一文発話対においても，メルケプストラム歪は 0 dB とならず，平均 4.4 dB，標準偏差 0.38 dB 程度のスペクトル変動が生じる．また，参考までに，同じ話者が別話者の音声を参照せずに，自身の韻律で同一文を 200 回繰り返し発話した音声データに対する結果も，同図に示す．メルケプストラム歪の平均は 3.9 dB, 標準偏差は 0.35 dB 程度である．別話者の韻律を真似ることで韻律特徴量の違いが増し，結果，スペクトル変動も増すことが分かる．

3 スペクトル特徴量変動の予測

声質変換における変換モデルの学習や変換精度の評価において，特定話者による同一文発話間における

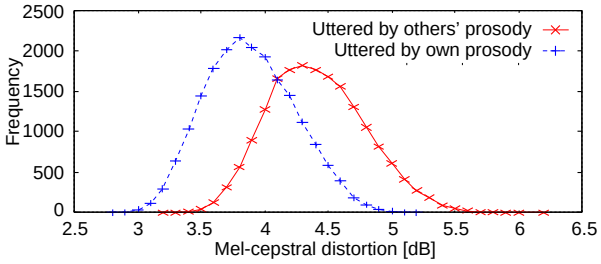

Fig. 1 Histogram of mel-cepstral distortion between utterances of the same sentence

メルケプストラム歪を予測することができれば，より良い学習尺度や評価尺度が得られると期待される．そこで，同一文発話間における韻律特徴量の変動から，メルケプストラム歪を予測する手法を提案する．

3.1 予測モデル

同一文発話の音声データ対において，次節で述べる複数の韻律変動パラメータを抽出し，重回帰モデルを用いて発話間のメルケプストラム歪を予測する．

3.2 韻律変動パラメータ

継続長の変動を捉える発話長歪及び DTW 歪，F_0 の変動を捉える有声／無声不一致率及び F_0 歪，パワーの変動を捉えるパワー歪を韻律変動パラメータとして用いる．なお，これらのパラメータは 0 以上の値をとり，韻律変動が無い場合は 0 となる．

発話長歪

発話間における話速の違いを表わすため，次式に示す発話長歪を用いる．

$$D_{\mathrm{dur}} = \log N_l - \log N_s \tag{1}$$

ここで，N_l は発話長が長い方の発話における有音フレーム数，N_s は発話長が短い方の発話における有音フレーム数である．

DTW 歪

DTW により得られる伸縮関数に対し，一方の発話の各フレームにおいて，前後 1 フレームを用いて回帰直線を計算することで，伸縮関数の傾きを求める．もう一方の発話に対しても同様の処理を行う．各発話の有声フレーム数を N_1 及び N_2 とし，フレーム t における傾きを $a_1(t)$ 及び $a_2(t)$ とする．発話長の比を基準として，DTW 歪を次式により定義する．

$$D_{\mathrm{DTW}} = \frac{1}{2N_1}\sqrt{\sum_{t=1}^{N_1}\left(a_1(t) - \frac{N_2}{N_1}\right)^2} + \frac{1}{2N_2}\sqrt{\sum_{t=1}^{N_2}\left(a_2(t) - \frac{N_1}{N_2}\right)^2} \tag{2}$$

有声/無声不一致率

発話間における有声/無声情報の違いを表すため，DTW により対応付けられたフレーム間における有声/無声不一致率を次式にて求める．

$$D_{\mathrm{U/V}} = \frac{1}{N}\sum_{t=1}^{N} e(f(t)) \tag{3}$$

---

* Spectral parameter variation between utterances of the same sentence by a single speaker and its prediction
by INUKAI, Tatsuo, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

Table 1　話者毎の発話対の数

文	J01	J02	J03	J04	J05	J06	合計
男性話者	18336	20706	19900	19900	19900	19900	118642
女性話者	1225	1225	1225	1225	0	0	4900

Table 2　実験結果

学習話者	女性	女性	男性	男性	女性	男性
評価話者	女性	女性	男性	男性	男性	女性
重回帰モデル	文依存	文非依存	文依存	文非依存	文非依存	文非依存
相関係数	0.83	0.71	0.77	0.82	0.69	0.60
RMSE	0.22	0.28	0.22	0.27	0.68	0.42

ここで，N は DTW における伸縮関数上でのフレーム対の総数を表し，$f(t)$ は t 番目のフレーム対を表す．また $e(\cdot)$ はフレーム対に対し，有声/無声が一致したら 0，不一致なら 1 を返す関数である．

F_0 歪

発話間における F_0 の違いを表すため，DTW により対応付けられたフレーム間において，F_0 歪を次式にて求める．

$$D_{F_0} = \frac{1}{N_v}\sqrt{\sum_{t=1}^{N_v}\left(\log(F_0^{(1)}(t)) - \log\left(F_0^{(2)}(t)\right)\right)^2} \quad (4)$$

ここで，N_v は DTW における伸縮関数上での有声フレーム対の総数を表し，$F_0^{(1)}(t)$ 及び $F_0^{(2)}(t)$ は t 番目の有声フレーム対における各発話の F_0 を表す．これに加え，発話内における対数 F_0 差の絶対値の最大値 $D_{F_0}^{(\max)}$ 及び最小値 $D_{F_0}^{(\min)}$ も用いる．

パワー歪

発話間におけるパワーの違いを表すため，DTW により対応付けられたフレーム間において，パワー歪を次式にて求める．

$$D_{\mathrm{pow}} = \frac{1}{N}\sqrt{\sum_{t=1}^{N}\left(p^{(1)}(t) - p^{(2)}(t)\right)^2} \quad (5)$$

ここで，N は DTW における伸縮関数上でのフレーム対の総数を表し，$p^{(1)}(t)$ 及び $p^{(2)}(t)$ は t 番目のフレーム対における各発話の正規化パワー (dB 値) を表す．これに加え，発話内における正規化パワー差の絶対値の最大値 $D_{\mathrm{pow}}^{(\max)}$ 及び最小値 $D_{\mathrm{pow}}^{(\min)}$ も用いる．

4　実験的評価

4.1　実験条件

男女各一名を話者とし，2.1 節で述べた方法で，ATR 音素バランス文 J セット内の文を発話したものをデータとして用いる．25 名の分析合成音声の韻律を参照しつつ，男性話者は 6 文 (J01 から J06) 各々において約 200 発話行い，女性話者は 4 文 (J01 から J04) 各々に対して 50 発話行う．実験に用いた発話対の数を Table 1 に示す．スペクトル特徴量として STRAIGHT 分析により抽出された 1 次から 24 次のメルケプストラム係数を用いる．サンプリング周波数は 16 kHz，シフト長は 5 ms とする．

重回帰モデルの予測精度を評価するために，予測されるメルケプストラム歪と実測のメルケプストラム歪の間において，相関係数と二乗平均平方根誤差 (Root Mean Square Error: RMSE) を求める．評価は 5 分割交差検定で行う．各話者に対し，同一文発話内でモデル学習及び予測を行う場合 (文依存) と，全発話でモデル学習及び予測を行う場合 (文非依存) の評価を行う．話者依存性を調査するために，男性話者のデータで学習した文非依存モデルを用いた女性話者のデータ予測，及びその逆の処理における評価も行う．また，個々の韻律変動パラメータの有効性を調査するために，各話者の文非依存モデルにおいて，相関係数の平均が最大となるように韻律変動パラメータを一つずつ追加した際の評価も行う。

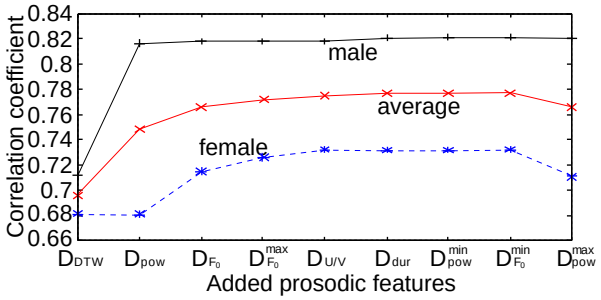

Fig. 2 Correlation coefficients when adding prosodic features one-by-one

4.2　実験結果

全ての韻律変動パラメータを用いた際の結果を Table 2 に示す．文依存モデルを用いることで，RMSE で 0.22，相関係数で約 0.8 程度の精度でメルケプストラム歪を予測できることが分かる．文非依存モデルを使用すると，RMSE については増加する傾向が見られる．また，異なる話者のモデルを用いた結果から，提案する予測処理は話者依存性が強いことが分かる．

Fig. 2 に韻律変動パラメータ追加による相関係数の変化を示す．両話者ともに，DTW 歪が予測に大きく寄与している．一方で，F_0 歪及びパワー歪における最大値と最小値，有声/無声不一致率，発話長歪は予測にほとんど寄与しない．また，男性話者ではパワー歪が予測に大きく寄与しているのに対し，女性話者ではあまり寄与していない等，話者間で異なる傾向が見られる．

5　まとめ

特定話者が発声した同一文発話間におけるスペクトル特徴量変動について調査した．韻律がスペクトル特徴量に与える影響に着目して，韻律特徴量変動からスペクトル特徴量変動を予測する手法を提案し，その有効性を示した．

謝辞　本研究の一部は，科研費補助金若手研究 (A) による支援を受けた．

参考文献

[1] Stylianou *et al.*, *IEEE Trans. Speech & Audio Process.*, 6(2), pp. 131–142, 1998.
[2] Toda *et al.*, *IEEE Trans. Speech & Audio Process.*, 15(8), pp. 2222–2235, 2007.
[3] 峯松 他, 音響誌, Vol.55, No.3, pp. 165–174, 1999.