同一文発話間におけるスペクトル特徴量変動予測の評価*

☆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る．その評価指標及び学習指標には，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，両者の間の距離尺度 (メルケプストラムひずみ等) がよく用いられる．そのため，同一文を複数回発話した際に生ずる発話間のスペクトル特徴量変動や，韻律の変動がもたらすスペクトル特徴量変動 [3] 等，同一話者内においても生じ得るひずみについては考慮されていない．これに対し著者らは，これらの要因がもたらすスペクトル変動量は許容できるという考え方に基づき，特定話者による同一文発話間におけるスペクトル変動量についての調査，及び韻律変動からのスペクトル特徴量変動量の予測に取り組んでおり，ごく少数の話者に対する結果を報告している [4]．本稿では，より多くの話者に対して，韻律変動からのスペクトル特徴量変動量の予測精度を調査する．また話者非依存の予測モデルについての評価も行う．

2 スペクトル特徴量変動の予測

本節では，同一話者による同一文発話音声対において，韻律変動を表すパラメータから発話間のメルケプストラムひずみを予測する手法について述べる．

2.1 予測モデル

韻律変動パラメータから発話間のメルケプストラムひずみを予測するために，重回帰モデルを用いる．なお，韻律変動パラメータやメルケプストラムひずみは各話者の個人性の影響を大きく受けると考えられる．この影響を抑え，不特定話者に対応した予測モデルを構築するために，それぞれの話者に対して，発話対から得られた各特徴量 (韻律変動パラメータ及びメルケプストラムひずみ) を標準化 (平均 0，分散 1 になるように正規化) した後に，全話者に対するデータを用いて重回帰モデルを学習する．

2.2 韻律変動パラメータ

継続長の変動を捉える発話長ひずみ及び時間軸伸縮 (Dynamic Time Warping: DTW) ひずみ，F_0 の変動を捉える有声／無声不一致率及び F_0 ひずみ，パワーの変動を捉えるパワーひずみ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．

発話長ひずみ

発話間における話速の違いを表わすため，次式に示す発話長ひずみを用いる．

$$D_{\mathrm{dur}} = |\log N_1 - \log N_2| \tag{1}$$

ここで，N_1 及び N_2 は各発話の有声フレーム数である．

DTW ひずみ

DTW により得られる時間軸伸縮関数に対し，一方の発話の各フレームにおいて，前後 1 フレームを用い回帰直線の傾きを計算することで，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に計算する．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，DTW ひずみを次式にて求める．

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

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．

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

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 ひずみ

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 ひずみを次式にて求める．

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

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．

パワーひずみ

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，パワーひずみを次式にて求める．

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

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．

3 実験的評価

3.1 実験条件

男性 4 名及び女性 1 名の計 5 名を話者とし，ATR 音素バランス文セットのサブセット J 内の文を発話した音声をデータとして用いる．各話者は 25 名の

---

* An evaluation of prediction of intra-speaker spectral parameter variation between utterances of the same sentence by INUKAI, Tatsuo, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, and NAKAMURA, Satoshi (Nara Institute of Science and Technology)

Table 1: メルケプストラムひずみの予測結果

話者	男性 1	男性 1	男性 2	男性 2	男性 3	男性 3	男性 4	男性 4
重回帰モデル	文依存	文非依存	文依存	文非依存	文依存	文非依存	文依存	文非依存
相関係数	0.77	0.82	0.74	0.72	0.76	0.72	0.72	0.76

話者	女性 1	女性 1	全話者平均	全話者平均	不特定話者	不特定話者	–	–
重回帰モデル	文依存	文非依存	文依存	文非依存	文非依存 正規化無し	文非依存 正規化有り	–	–
相関係数	0.83	0.71	0.76	0.75	0.70	0.73	–	–

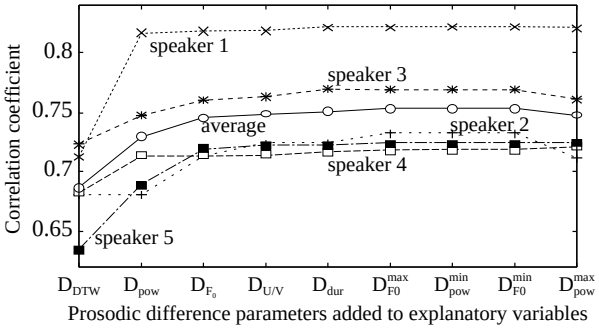

Fig. 1: Correlation coefficients when adding prosodic features one-by-one.

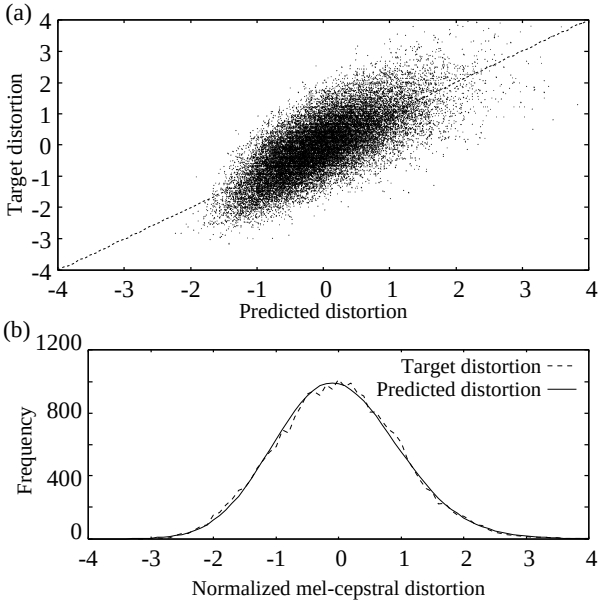

Fig. 2: Comparison between target mel-cepstral distortion and that predicted by speaker-independent model: (a) scatter diagram and (b) histograms.

分析合成音声を聴取し，韻律を参照しつつ発話を行う．男性話者 1 は 6 文 (J01 から J06) 各々に対して約 200 発話行い，その他の話者は 4 文 (J01 から J04) 各々に対して 50 発話行う．スペクトル特徴量として STRAIGHT 分析 [5] 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ひずみと実測のメルケプストラムひずみの間において相関係数を求める．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．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．さらに，不特定話者モデルによる予測についても評価する．その際には，正規化有り／無しの両手法について結果を比較する．なお，不特定話者モデルの評価時には，各話者の発話対の数を等しくする (4900 対に統一)．

3.2 実験結果

全韻律変動パラメータを用いた際の結果を表 1 に示す．文依存モデルによる予測では相関係数 0.76 程度，文非依存モデルによる予測では相関係数 0.75 程度の精度で，メルケプストラムひずみを予測できており，提案法における文依存性は低いことが分かる．また，図 1 に，特徴量の有用性の順に韻律変動パラメータを一つずつ追加して予測した際における相関係数の変化を示す．全ての話者において DTW ひずみが予測に大きく寄与しており，さらに，F_0 ひずみおよびパワーひずみを併用することで，全ての韻律変動パラメータを使用した際と同等の予測精度が得られることが分かる．

不特定話者モデルの評価結果についても表 1 に示す．正規化を行わない際には，特定話者モデルと比較し，相関係数が平均で 0.05 低下している．これに対して，正規化を施すことで，話者依存性を低減することができ，相関係数で 0.73 程度の予測精度が得られることが分かる．正規化を行った際の，メルケプストラムひずみの予測値と実測値の関係を図 2(a) に示す．この図から，概ね良好に予測できていることが分かる．また，図 2(b) に実測値と予測値のヒストグラムを示す．ここで，予測値のヒストグラムは，重回帰モデルによる予測値を平均，学習時に得られる平均平方根誤差を分散とした正規分布により，各発話対に対する予測分布が与えられるとした時の結果を表す．図 2(b) から，予測値のヒストグラムは実測値のヒストグラムを良くモデル化できていることが分かる．

4 まとめ

本稿では，同一文における韻律変動パラメータからスペクトル特徴量変動量を予測する手法に対し，複数話者に対して実験的評価を行った．その結果，話者毎に変動量の正規化を行うことで，不特定話者への対応が可能であることが分かった．今後，統計的手法に基づくスペクトル特徴量変換において，スペクトル特徴量変動量の予測値を考慮した評価尺度および学習尺度の設計に取り組む予定である．

謝辞 本研究の一部は，JSPS 科研費 22680016 の助成を受け実施したものである．

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.
[4] 犬飼 他, 信学技報, SP2012-74, pp. 13–18, 2012.
[5] Kawahara *et al.*, Speech Commun., 27(3–4), pp. 187–207, 1999.