社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

同一文発話間における話者内スペクトル特徴量変動とその予測

犬飼　辰夫† 　戸田　智基† 　グラムニュービッグ† 　サクリアニサクティ† 　中村　哲†

† 奈良先端科学技術大学院大学 〒 630-0192 奈良県生駒市高山町 8916-5
E-mail: †{tatsuo-i,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

あらまし　統計的声質変換技術におけるスペクトルパラメータ変換処理おいて，その評価指標や変換モデルの学習指標として，しばしば変換パラメータと目標パラメータ間の距離尺度（例えばメルケプストラムひずみ）が用いられる．しかしながら，同一話者が同一文を発話した際においても，スペクトルパラメータは変動するため，スペクトルパラメータ間の距離は零とはならない．また，特にリアルタイム変換処理においては，複雑な変換処理を韻律パラメータに対して施すことは困難であるため，しばしば入力音声の韻律特徴を保持した変換音声が生成される．同一話者が同一文を異なる韻律特徴で発話した音声を生成することが目標となるが，この際に許容される発話間のスペクトルパラメータ間の距離についても考慮されていない．本報告では，スペクトルパラメータとしてメルケプストラムに着目し，同一話者が同一文を発話したときに生じるその変動量について調査する．また，発話間の韻律特徴の違いから，メルケプストラムの変動量を予測する手法を提案し，実験的評価結果から，その有効性を示す．
キーワード　声質変換，学習・評価指標，スペクトル変動，同一文発話，韻律変動

Intra-speaker spectral parameter variation between utterances of the same sentence and its prediction

Tatsuo INUKAI†, Tomoki TODA†, Graham NEUBIG†, Sakriani SAKTI†, and Satoshi NAKAMURA†

† Nara Institute of Science and Technology Tatayama–cho 8916-5, Ikoma, Nara, 630-0192 Japan
E-mail: †{tatsuo-i,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

Abstract　In spectral conversion of statistical voice conversion technologies, distance measures between the converted and target parameters, such as mel-cepstral distortion, are often used as evaluation/training metrics. However, even if the same speaker utters the same sentence, the spectral parameters of those utterances vary, and therefore, a distance between them still exists. Moreover, in real-time conversion procedure, converted speech keeping original prosodic features of input speech is often generated due to an essential difficulty of complex conversion of those features in real time. In such a case, an ideal sample of converted speech will be a speech sample uttered by a target speaker imitating prosody of the input speech but a spectral variation caused by such a prosodic change is not considered in the current evaluation/training metrics. In this report, we investigate an intra-speaker spectral variation between utterances of the same sentence focusing on mel-cepstrum as a spectral parameter. Moreover, we propose a method for predicting it from prosodic parameter differences between those utterances and conduct experimental evaluations to show its effectiveness.
Key words　voice conversion, training/evaluation metric, spectral variation, utterances of the same sentence, prosodic variation

1. ま え が き

声質変換とは，入力された発話の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 [1]．声質変換技術の中でも，統計的手法に基づく声質変換 [2], [3] は，少量の学習データから自動的に抽出される変換規則を用いて，入力音声の音響特徴量を変換する．統計処理に基づく複雑な変換処理が可能となるため，発声障害者補助 [4] やサイレント音声通話 [5] など，

多岐にわたる応用例に対して適用できる．また，特に人対人のコミュニケーション応用を目指し，リアルタイム変換処理 [6] も実現されており，今後さらなる発展が期待される．

代表的な声質変換の応用例である話者変換 [7] では，元話者の音声から目標話者の音声への変換が行われる．入力として言語情報を一切使用しない枠組みにおいては，主にスペクトルパラメータの変換に対して，統計的手法に基づく複雑な変換処理が行われる．この時，その評価指標や変換規則の学習指標として，基本的に変換スペクトルパラメータが目標音声のスペクトルパラメータに近づくほど良いという考えに基づき，変換パラメータと目標パラメータ間の距離尺度（例えばメルケプストラムひずみ）がしばしば用いられる．

従来の距離尺度に基づく評価指標や学習指標は有用ではあるものの，同一文発話間におけるスペクトル特徴量変動については考慮されていない．例えば，同一話者が同一文を複数回発声した音声は，声質変換において理想的な出力結果とみなせるが，スペクトルは発話毎に変動するため，スペクトルパラメータ間の距離は零とはならない．また，声質変換処理の中でも特にリアルタイム変換処理においては，処理遅延を抑えるため，韻律特徴量に対する複雑な変換処理は行わず，F_0 範囲の調整のみといった簡易的な変換処理がしばしば行われる．この場合，入力音声の韻律特徴を保持した変換音声が生成されるため，声質変換の理想的な出力結果は，目標話者が自身の F_0 範囲において元話者音声の韻律を忠実に真似て発声した際の音声となる．すなわち，同一話者が同一文を異なる韻律特徴で発話した音声を生成することが目標となるが，従来の評価指標・学習指標では，韻律の差異がスペクトルパラメータに与える影響 [8] は考慮されていない．

本報告では，スペクトルパラメータとして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（メルケプストラムひずみ）について調査する．また，入力音声の韻律特徴を保持した変換処理において，許容すべきスペクトル変動量を考慮した評価・学習指標の確立を目指して，同一発話間における韻律特徴の変動を表すパラメータから，メルケプストラムひずみを予測する手法を提案する．実験的評価結果から，提案法により，特定話者用予測モデルを用いた際には，高い予測性能が得られることを示す．

以下，2. で同一話者による同一文発話間のメルケプストラムひずみについて調査し，3. で韻律特徴量変動からのメルケプストラムひずみの予測法について述べる．4. で実験的評価について述べ，最後に 5. で本報告の結論と今後の課題を述べる．

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

同一話者が同一文を複数回発声する際に，発話毎にどの程度メルケプストラムが変動するかを調査するために，特定話者による同一文発話音声の収録する．話者は男性話者 1 名とし，ある 1 文を 200 回繰り返し発話する．

また，同一話者が同一文を異なる韻律特徴で発話した際のメルケプストラム変動を調査するために，同じ男性話者が他の複数話者 (参照話者) による同一文発話の韻律を真似しながら発声した音声を収録する．その際には，参照話者の韻律を真似しやすくするために，参照話者の音声に対して分析合成処理を施し，次式に従い F_0 範囲を調整したものを参照音声として用いる．

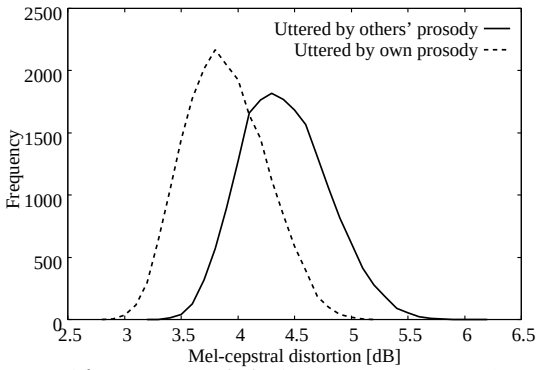


図 1: 同一話者による同一文発話間におけるメルケプストラムひずみのヒストグラム
Fig. 1 Histogram of mel-cepstral distortion between utterances of the same sentence uttered by the same speaker.

$$\log \hat{F_0} = \frac{\sigma^{(s)}}{\sigma^{(r)}} \left(\log F_0 - \mu^{(r)} \right) + \mu^{(s)} \tag{1}$$

ここで $\log F_0$ は調整前の対数 F_0 を表し，$\log \hat{F_0}$ は調整後の対数 F_0 を表す．$\mu^{(s)}$ 及び $\sigma^{(s)}$ は繰り返し発話を行う男性話者の対数 F_0 の平均及び標準偏差を表し，$\mu^{(r)}$ 及び $\sigma^{(r)}$ は参照話者の対数 F_0 の平均及び標準偏差を表す．参照話者は男性話者 12 名と女性話者 12 名の計 24 名である．上述のものと同じ 1 文を用いて，個々の参照話者の分析合成音声を参照にして，8 回 (全参照話者を通して合計 192 回) 繰り返し発話する．

2.2 スペクトル特徴量変動

個々の同一文発話対において，発話単位で正規化された波形パワーにより有音フレーム系列を自動抽出し，動的時間伸縮 (dynamic time warping: DTW) を行うことで，メルケプストラムひずみを計算する．STRAIGHT 分析 [9] に基づき抽出される 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 とし，分析シフトは 5 ms とする．

参照話者の韻律を参照せずに自身の韻律で発話した音声データ 200 発話セットと，参照話者の韻律を真似して発話した音声データ 192 発話セットの各々において，全ての発話対に対して計算したメルケプストラムひずみの頻度分布を図 1 に示す．同一文発話対においても，メルケプストラムひずみは 0 dB とならず，自身の韻律で発話した際においても，平均 3.9 dB，標準偏差 0.35 dB 程度のスペクトル変動が生じる．また，参照話者の韻律を真似して発話した際には，スペクトル変動はより大きくなる傾向があり，平均 4.4 dB，標準偏差 0.38 dB 程度の変動が生じる．

以上の結果から，声質変換において，同一話者による同一文発話音声を変換音声の目標とするのであれば，必ずしも変換スペクトルと目標スペクトル間のメルケプストラムひずみを零にする必要はないことが分かる．また，元話者と目標話者の同一文発話において，変換音声と目標音声の韻律が大きく異なる場合には，変換スペクトルと目標スペクトル間のメルケプストラ

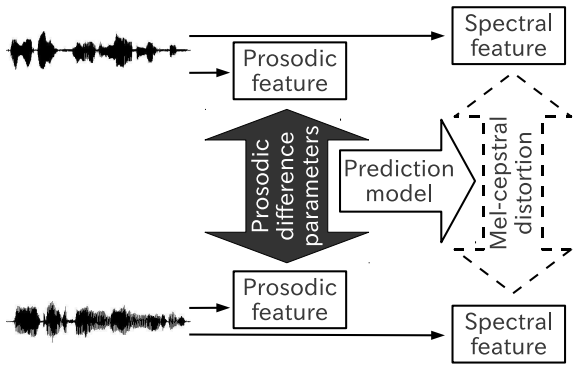


図 2: 韻律特徴量変動からのメルケプストラムひずみ予測処理
Fig. 2 Prediction procedure of mel-cepstral distortion from prosodic difference parameters.

ムひずみは，より大きな値をとっても良いと考えられる．

3. スペクトル特徴量変動の予測

統計的手法に基づく声質変換におけるスペクトルパラメータ変換処理において，目標音声と目標話者が変換音声の韻律を模擬して発話した音声 (すなわち，理想的な変換音声) との間で生じ得るメルケプストラムひずみを予測することができれば，より良い評価指標および学習指標が得られると期待される．例えば，個々の発話対に応じて，許容されるメルケプストラムひずみを変化させることも可能となる．

上記の予測処理において，予測に使用できる主な音響特徴量は，目標音声のスペクトル特徴量および韻律特徴量と，変換音声の韻律特徴量（元音声の韻律特徴量に大きく依存）となる．本報告では，予測処理の構築に向けた第一歩として，同一文発話間における韻律特徴量の変動から，メルケプストラムひずみを予測する手法を提案する．処理の流れを図 2 に示す．まずメルケプストラムひずみを測定する同一文発話の音声データから韻律特徴量を抽出し，発話間の韻律の違いを捉える韻律変動パラメータを求める．そして，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．

3.1 予測モデル

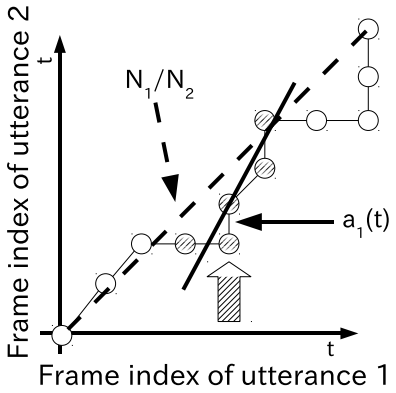
同一文発話の音声データ対において，次節で述べる複数の韻律変動パラメータを抽出し，次式に示す重回帰モデルを用いて発話間のメルケプストラムひずみを予測する．

$$\hat{m}_{\mathrm{i,j}} = \mathbf{a}^{\top}\mathbf{p}_{\mathrm{i,j}} + \mathrm{c} \tag{2}$$

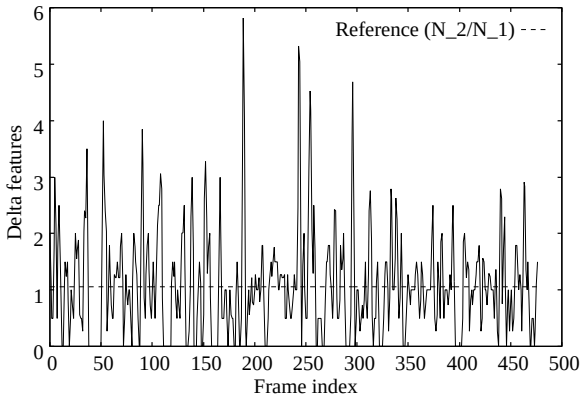
ここで，$\hat{m}_{\mathrm{i,j}}$ は発話 i と発話 j 間のメルケプストラムひずみの予測値であり，$\mathbf{p}_{\mathrm{i,j}}$ は発話 i と発話 j 間の韻律変動パラメータベクトルを表す．また，$\mathbf{a}$ は係数ベクトル，c は定数項を表し，最小二乗誤差基準により最適化する．

3.2 韻律変動パラメータ

韻律変動パラメータとして，継続長の変動を捉える発話長ひずみ及び DTW ひずみ，F_0 パターン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 (二つの発話の韻律が完全に一致する場合) は 0 となる．



(a) Time warping function



(b) Dynamic feature sequence

図 3: 時間軸伸縮関数と動的特徴量系列
Fig. 3 Time warping function and its dynamic feature sequence.

3.2.1 発話長ひずみ

発話間における平均的な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．

$$D_{\mathrm{dur}} = \log N_l - \log N_s \tag{3}$$

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り，発話単位で正規化されたパワーに基づき自動的に抽出する．

3.2.2 DTW ひずみ

発話間における局所的な話速の違いを表わすため，DTW により得られる時間軸伸縮関数の傾きの値を用いる．発話間の局所的な話速が等しい際には，時間軸伸縮関数は各発話長の比を傾きとする直線となる．そこで，局所的な話速の違いを捉える DTW ひずみとして，次式により，時間軸伸縮関数の傾き (動的特徴量) と発話長の比の違いを求める．

$$D_{\mathrm{DTW}} = \frac{1}{2N_1}\sqrt{\sum_{t=1}^{N_1}\left(a_1(t) - \frac{N_2}{N_1}\right)^2} + \frac{1}{2N_2}\sqrt{\sum_{t=1}^{N_2}\left(a_2(t) - \frac{N_1}{N_2}\right)^2} \tag{4}$$

ここで N_1 及び N_2 は各発話の有音フレーム数を表す．また，$a_1(t)$ 及び $a_2(t)$ は各発話を時間軸の基準と見なした時のフレーム t における回帰直線の傾き (一次の動的特徴量) であり，フ

表 1: 話者毎の発話対の数
Table 1　Number of utterance pairs in each speaker.

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

レーム t の前後 1 フレームに存在する対応点を用いて，次式により計算する．

$$a(t) = \frac{\sum_{(x,y)\in P(t)} (x-\overline{x})(y-\overline{y})}{\sum_{x\in X(t)} (x-\overline{x})^2} \tag{5}$$

ここで $P(t)$ はフレーム t の前後 1 フレームに存在する対応点を表し，x と y はその x 座標（時間軸の基準としている発話のフレーム番号）と y 座標（もう一つの発話のフレーム番号）を表す．また，$\overline{x}, \overline{y}$ は，$P(t)$ 内の x 座標，y 座標それぞれの平均値である．図 3 に，時間軸伸縮関数の一例と，その動的特徴量系列を示す．

3.2.3　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．

$$D_{\mathrm{U/V}} = \frac{1}{N}\sum_{t=1}^{N} e(f(t)) \tag{6}$$

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．

3.2.4　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．

$$D_{F_0} = \frac{1}{N_v}\sqrt{\sum_{t=1}^{N_v} \left(\log(F_0^{(1)}(t)) - \log\left(F_0^{(2)}(t)\right)\right)^2} \tag{7}$$

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．

3.2.5　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．

$$D_{\mathrm{pow}} = \frac{1}{N}\sqrt{\sum_{t=1}^{N} \left(p^{(1)}(t) - p^{(2)}(t)\right)^2} \tag{8}$$

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．

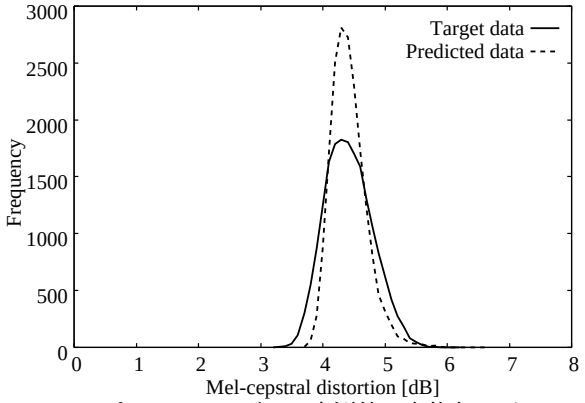


図 4: メルケプストラムひずみの実測値と文依存モデルによる予測値のヒストグラム
Fig. 4　Histograms of target mel-cepstral distortion and that predicted by sentence-dependent model.

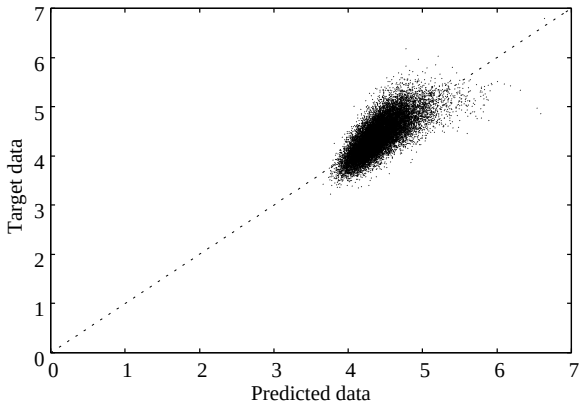


図 5: メルケプストラムひずみの実測値と文依存モデルによる予測値の散布図
Fig. 5　Scatter diagram of target mel-cepstral distortion and that predicted by sentence-dependent model.

4. 実験的評価

4.1 実験条件

男女各一名を話者とし，2.1 で述べた方法で，ATR 音素バランス文セット [10] のサブセット J 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) 各々に対して約 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表 1 に示す．スペクトル特徴量として STRAIGHT 分析 [9] 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．F_0 パターンの抽出においても，STRAIGHT 分析の F_0 抽出法 [11] を使用する．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

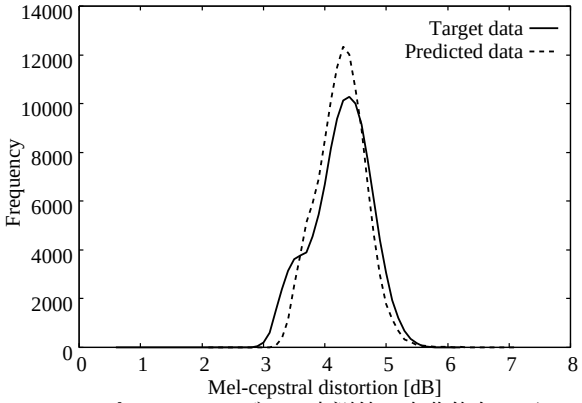


図 6: メルケプストラムひずみの実測値と文非依存モデルによる予測値のヒストグラム

Fig. 6 Histograms of target mel-cepstral distortion and that predicted by sentence-independent model.

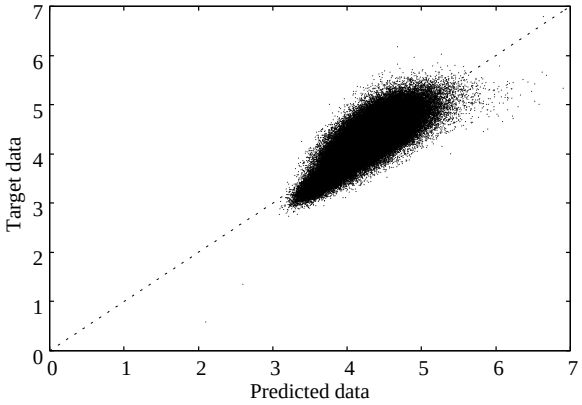


図 7: メルケプストラムひずみの実測値と文非依存モデルによる予測値の散布図

Fig. 7 Scatter diagram of target mel-cepstral distortion and that predicted by sentence-independent model.

パラメータの有効性を調査するために，各韻律変動パラメータ間の相関係数を求めるとともに，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。

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表 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ひずみを予測できることが分かる．図 4 に，男性話者の J01 文に対するメルケプストラムひずみの実測値と文依存モデルによる予測値のヒストグラムを示す．予測を点推定で行っているため，予測値の分布の分散は小さくなる傾向が見られるが，分布形状は概ね一致していることが分かる．また図 5 に，男性話者の J01 文に対するメルケプストラムひずみの実測値と文依存モデルによる予測値の散布図を示す．目標値であるメルケプストラムひずみの実測値が小さい際には，予測誤差が小さくなる傾向が見られる．

文非依存モデルを使用すると，表 2 から，RMSE については文依存モデルより増加する傾向が見られる．これは，他の文と比べると，メルケプストラムひずみの実測値の分布が大きく異なる文があり，その影響により全体としての分散が大きくなるためであると考えられる．図 6 に，男性話者の全 6 文に対するメルケプストラムひずみの実測値と文非依存モデルによる予測値のヒストグラムを示す．文非依存モデルにおいても，分布形状は比較的類似していることが分かる．また図 7 に，メルケプストラムひずみの実測値と文非依存モデルによる予測値の散布図を示す．文依存モデルを用いた際の結果 (図 5) と同様の傾向が見られる．

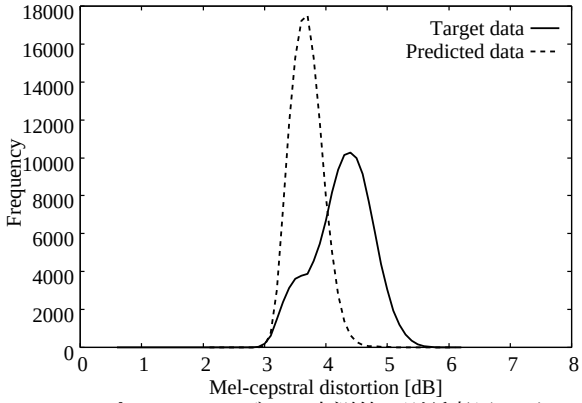


図 8: メルケプストラムひずみの実測値と別話者用モデルによる予測値のヒストグラム

Fig. 8 Histograms of target mel-cepstral distortion and that predicted by mismatched speaker-dependent model.

異なる話者のモデルを用いると，表 2 から，RMSE が大幅に増加し，相関係数も大幅に低下することが分かる．図 8 に，男性話者におけるメルケプストラムひずみの実測値と，女性話者用のモデルを用いた予測値のヒストグラムを示す．分布形状が異なるのみでなく，平均値も大きくずれる傾向があり，十分な予測性能が得られないことが分かる．また，図 9 に，男性話者におけるメルケプストラムひずみの実測値と，女性話者用のモデルを用いた予測値の散布図を示す．同一話者モデルを用いた際の結果 (図 7) と比較し，明らかに予測精度が劣化していることが分かる．これらのことから，提案した予測処理は話者依存性が強いといえる．

各韻律変動パラメータ間の相関係数の話者平均を表 3 に示す．F_0 ひずみとパワーひずみにおいて，その最大値との間には相関係数で 0.6 程度の関連性がみられるが，概ね相関係数は低い傾向にあり，各韻律変動パラメータは異なる特徴を捉えていると考えられる．図 10 に韻律変動パラメータを一つずつ追加した際における相関係数の変化を示す．両話者共に，DTW ひずみが予測に大きく寄与している．一方で，F_0 ひずみ及びパワーひずみにおける最大値と最小値や発話長ひずみは予測にほぼ寄与しない．また，男性話者ではパワーひずみ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など，話者間で重要な韻律変動パラメータは異なる傾向が見られる．

5. まとめ

従来の統計的声質変換の評価・学習指標では，同一話者による同一文発話間におけるスペクトル特徴量の変動や，韻律の違いにより生じるスペクトル特徴量変動の増加は考慮されていなかった．本報告では，より適切な評価・学習指標の確立を目指す上で，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．また，韻律の変動がスペクトル

表 2: メルケプストラムひずみの予測結果
Table 2　Prediction results of mel-cepstral distortion.

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

表 3: 韻律変動パラメータ間の相関
Table 3　Correlation coefficients between individual prosodic difference parameters.

	D_{DTW}	$D_{U/V}$	D_{F_0}	$D_{F_0}^{max}$	$D_{F_0}^{min}$	D_{pow}	D_{pow}^{max}	D_{pow}^{min}
D_{dur}	0.28	0.06	0.11	0.03	0.03	0.10	0.04	0.03
D_{DTW}	–	0.44	0.36	0.27	0.08	0.40	0.30	0.07
$D_{U/V}$	–	–	0.41	0.18	0.21	0.48	0.29	0.09
D_{F_0}	–	–	–	0.57	0.29	0.40	0.20	0.12
$D_{F_0}^{max}$	–	–	–	–	0.04	0.27	0.24	0.07
$D_{F_0}^{min}$	–	–	–	–	–	0.11	-0.00	0.03
D_{pow}	–	–	–	–	–	–	0.63	0.24
D_{pow}^{max}	–	–	–	–	–	–	–	0.11

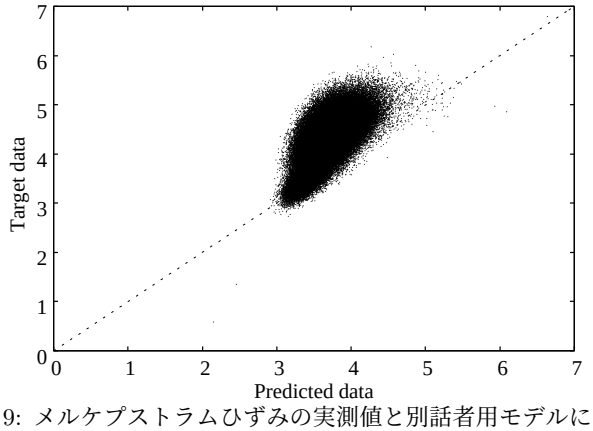


図 9: メルケプストラムひずみの実測値と別話者用モデルによる予測値の散布図
Fig. 9　Scatter diagram of target mel-cepstral distortion and that predicted by mismatched speaker-dependent model.

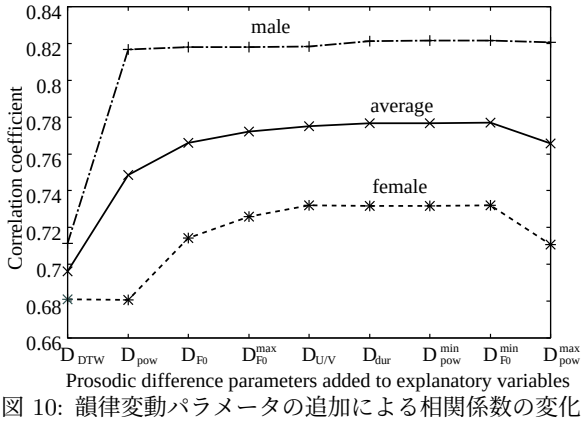


図 10: 韻律変動パラメータの追加による相関係数の変化
Fig. 10　Correlation coefficients when adding prosodic difference parameters one by one to explanatory variables in prediction.

特徴量変動に与える影響を考慮するために，発話間の韻律特徴量変動を捉える各種パラメータを用いて，重回帰モデルによりスペクトル特徴量変動を予測する手法を提案した．実験的評価結果から, 同一話者内であれば，異なる文に対しても同一のモデルにより良好な予測結果が得られることを示した．今後は予測値を考慮した評価指標の作成を行う予定である．

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．

文　献

[1] Y. Stylianou, “Voice transformation: a survey,” Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on IEEE, pp. 3585–3588, Apr. 2009.

[2] Y. Stylianou, O. Cappe, E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. SAP, Vol. 6, No. 2, pp. 131–142, Mar. 1998.

[3] T. Toda, A. W. Black, K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.

[4] H. Doi, K. Nakamura, T. Toda, H, Saruwatari, K. Shikano, “Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models,” IEICE Trans. Inf.& Syst., Vol. E93-D, No. 9, pp. 2472–2482, Sep. 2010.

[5] T. Toda, M. Nakagiri, K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” IEEE Trans. ASLP, vol. 20, No. 9, pp. 2505–2517, Nov. 2012.

[6] T. Toda, T. Muramatsu, H. Banno, “Implementation of computationally efficient real-time voice conversion,” Proc. of INTERSPEECH, Sept. 2012.

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” J. Acoust. Soc. Jpn. (E), Vol. 11, No. 2, pp. 71–76, 1990.

[8] 峯松信明, 中川聖一, “F_0 変化に伴うスペクトル変動に対する分析とモデル化,” 音響誌, Vol. 55, No. 3, pp. 165–174, 1999.

[9] H. Kawahara, I. Masuda-Katsuse, A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneousfrequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Commun., Vol. 27, No. 3–4, pp. 187–207, Oct. 1999.

[10] 阿部 匡伸, 匂坂 芳典, 梅田 哲夫, 桑原 尚夫, “研究用日本語音声データベース利用解説書 (連続音声データベース),” ATR Technical Report, TR-I-0166, ATR 自動翻訳通信研究所, Sep. 1990.

[11] H. Kawahara, H. Katayose, A. deCheveign , and R.D. Pateterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity,” Proc. Eurospeech, vol.99, pp.2781-2784, Sep. 1999.