基本周波数変形を考慮したスペクトル変換手法の検討

A method of spectrum conversion considering F0 transformation

床 建吾

Kengo Toko

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本研究では，基本周波数の変形を考慮したスペクトル変換手法を提案する．現在，高品質なテキスト音声合成方式を実現するためには大量の音声データが必要となる．そこで，基本周波数の変形を考慮したスペクトル変換を音声合成方式に導入することで，少量のデータでも高品質な音声合成が可能となると考えられる．評価実験の結果，提案手法によるスペクトル変換の有効性と変換規則生成法の改良の必要性が示された．

1 はじめに

近年，テキスト音声合成の研究ではより肉声に近い自然な音声を作り出すことが一つの目標になっている．一般に，大量の音声データがあれば，数多くの音声素片を用いた音声素片接続により自然な音声を合成できることが知られている．主な方式としては波形合成方式 [1] がある．しかし，このような高品質な音声合成システムが誰でも簡単に実用化できるわけではない．原因の一つとして，大量の音声データを収集することが時間や費用の面で困難であるという問題がある．これに対し，隠れマルコフモデル（HMM）を用いた HMM 音声合成方式 [2] は少量の音声データでも滑らかな韻律性を持つ音声を合成することが可能である．しかし，統計的な処理による過剰な平滑化の影響により，合成される音声はこもった音声になってしまう．

そこで本研究では，小規模な音声データベースでも高品質な音声が合成可能な方式の実現を目指し，基本周波数の時間変化パターン（F0 パターン）に基づいたスペクトルの変換による新たな音声合成方式を提案する．本方式では，変換後のスペクトルが所望の F0 パターンに対応したスペクトルになるように，まず音声データに対して F0 パターンに基づくクラスタリングによって音声データを分類する．次に分類した各クラスタ間でスペクトルを変換する規則を学習する．この変換規則を変換対象のスペクトルに適用することで，スペクトルの変換を行う．これらの処理により，類似した特徴を持つ波形間でのスペクトル変換が実現できるため，所望の F0 パターンの特徴を持ったスペクトルを得られると考えられる．そのため，データベース内に存在しない F0 パターンに対応したスペクトルを補うことができ，小規模なデータベースでも高品質な音声合成が可能になると考えられる．本報告では F0 パターンに基づくスペクトル変換手法の概要と性能評価について述べる．

2 提案手法

提案手法の概要を図 1 に示す．提案手法は大きく「F0 パターンに基づく音声データのクラスタリング」と「スペクトル変換規則の生成」の 2 つの過程に分けることができる．以下ではそれぞれの過程について説明する．

2.1 F0 パターンに基づく音声データのクラスタリング

F0 パターンに対応したスペクトル変換規則を生成するためには，まず音素データベース内の音素ごとに音声データを F0 パターンに基づいて分類する必要がある．提案手法ではこれを k-means 法（k-平均法）によって行い，F0 パターンを上昇しているもの（上り調子），あまり変化していないもの（平坦），下降しているもの（下り調子）の 3 種類と仮定し，分類を行う．この際，データ数によっては 1 回の分類に膨大な計算量が必要となる可能性がある．そこで，本研究では以下に示す方法で対処する．

1. 1 音素につき，各音韻環境ごとに k-means 法による分類を行う．
2. 各音韻環境のクラスタリング結果（3 つのクラスタ中心）を集め，それらを k-means 法によって分類する．

これにより，対処前と比べて 1 回の分類に必要な計算量が減少し，なおかつ対処前と同程度の分類精度が得られると考えられる．

2.2 スペクトル変換規則の生成

音声データのクラスタリング結果をもとに，スペクトル変換規則の生成を行う．本研究では変換規則の生成に GMM(Gaussian Mixture Model) に基づく声質変換 [3] を用いる．GMM に基づく声質変換では入力話者と出力話者のスペクトルを用いて GMM 学習を行っているが，本研究では異なる F0 パターン同士のスペクトルを用いて GMM 学習を行う．GMM に基づく声質変換では，入力特徴量 $\boldsymbol{X}$ と出力特徴量 $\boldsymbol{Y}$ に対して GMM のパラメータを推定し，以下の式で表される条件付確率密度関数を求めることで，入出力の対応関係

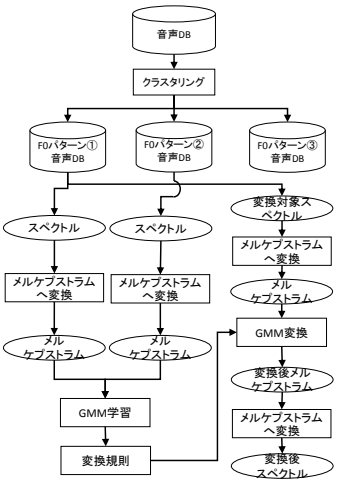

図 1: 提案手法の概要

を確率的にモデル化を行う．

$$P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(X,Y)}) = \sum_{\boldsymbol{m}} \gamma_{\boldsymbol{m}}^{(X)} \mathcal{N}(\boldsymbol{Y};\boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)},\Sigma_{\boldsymbol{m}}^{(Y|X)})$$

ここで，$\boldsymbol{m}$ は分布系列 $\{m_1,\ldots,m_T\}$，$\gamma_{\boldsymbol{m}}^{(X)}$ は混合重み，$\boldsymbol{\mu}_{\boldsymbol{m}}^{(Y|X)}$ は平均ベクトル，$\Sigma_{\boldsymbol{m}}^{(Y|X)}$ は共分散行列を表し，$\boldsymbol{\lambda}^{(X,Y)}$ はこれらのパラメータの集合である．変換の際には，この条件付確率密度関数を最大化する出力特徴量を求めることにより変換を行う．本研究では学習のための特徴量として音素データベースから取り出したメルケプストラムとその動的特徴量の結合ベクトルを入出力それぞれの特徴ベクトルとして用いる．そのため，変換規則の学習の際にはスペクトルをあらかじめメルケプストラムに変換しておくことが必要となる．

3　評価実験

スペクトル変換規則の性能を評価するために，目標のメルケプストラムに対して変換前後のそれぞれでメルケプストラム距離を計算した．この際，変換後のスペクトルが変換前と目標のどちらにより近いかを判断するため，目標のメルケプストラム同士でもメルケプストラム距離を計算し，その最小値と平均値を算出した．もし変換後のスペクトルが変換前よりも目標のスペクトルに近ければ，変換後と目標との間のメルケプストラム距離の値は目標同士のメルケプストラム距離の値により近い値になっているはずである．また，メルケプストラム距離の計算結果をもとに変換前後と目標のそれぞれのスペクトルを用いた対象音素の音声を作成した．実験条件に関しては，対象音素を/A/とし，F0 パターンは“上り調子”，“平坦”，“下り調子”の 3 つと仮定して実験を行う．変換規則の学習とテストに用いるデータはそれぞれ対象音素の各 F0 パターンにおいて音韻環境ごとに 90% と 10% を用いる．また，GMM 変換においては 0–24 次のメルケプストラムとその Δ を入力特徴量として用い，混合数は 32 とした．使用した音声データの詳細は表 1 の通りである．

メルケプストラム距離の結果を図 2 に示す．目標最小と目標平均の項目はそれぞれ目標のメルケプストラム同士のメルケプストラム距離における最小値と平均値を表し，縦軸がメルケプストラム距離の値，横軸が入出力それぞれのスペクトルが対応している F0 パターンの全組合わせとなっている．変換前後で比較すると，どの F0 パターンの組み合わせについても変換後の方が変換前に対して距離が 21%〜23% 減となっており，提案手法による変換が有効に働いていると考えられる．また，合成した音声については変換後の方が変換前に対して明瞭性が向上したものとなっていた．したがって，提案手法による変換で合成音の品質の向上が期待できると考えられる．一方で，変換後の音声は前後の音韻環境の特徴が欠落してしまっていることも判明した．そのため，前後の音韻環境の特徴が欠落しないための工夫が必要となる．

表 1: 音声データの詳細

話者数	女性話者 1 名
データ量	約 20 時間
サンプリング周波数	22.05 kHz
音声分析・合成	STRAIGHT [4]
分析時のフレームシフト	5 ms
メルケプストラム	0–24 次

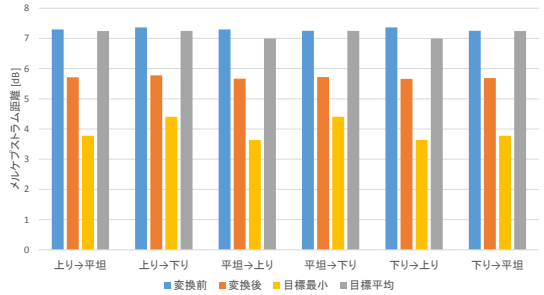

図 2: 平均メルケプストラム距離

4　まとめ

本報告では，基本周波数の変形を考慮したスペクトル変換手法を提案した．評価実験の結果，メルケプストラム距離や合成音の明瞭性という点で提案手法の有効性が示されたが，変換後の音声は前後の音韻環境の特徴が欠落することがわかった．今後は前後の音韻環境の特徴の欠落を防ぐため，変換対象と目標のスペクトルの差分を学習する手法を検討する．

参考文献

[1] Alan W. Black, *et al.*, Proc. of EUROSPEECH, pp. 581–584, 1995.
[2] Heiga Zen, *et al.*, ICASSP'96, pp. 1039–1064, 2009.
[3] Y. Stylianou, *et al.*, Proc. of EUROSPEECH, pp. 447–450, Sept. 1995.
[4] H. Kawahara, *et al.*, Speech Communication 27, pp. 187–207, 1999.