F-22 HMM歌声合成における音韻・音高の同時モデル化

Simultaneous modeling of phonetic and pitch parameters in HMM-based singing voice synthesis

石川 ちさと† 伊藤 正典† 酒向 慎司† 宮島 千代美† 徳田 恵一† 北村 正†
Chisato ISHIKAWA Masanori ITO Shinji SAKO Chiyomi MIYAJIMA Keiichi TOKUDA Tadashi KITAMURA

1. はじめに

我々は，これまでに提案してきたHMMに基づくテキスト音声合成法 [1] を応用し，任意の楽譜から歌声を合成する手法について検討してきた．HMMに基づく音声合成手法では，音声データを統計的にモデル化し，接続が滑らかな合成音声が得られる他に，話者への適応，多様な声質を実現可能といった利点がある．それにより，マルチメディアコンテンツ制作分野への応用や，エンターテイメントの場への技術の提供が期待できる．

本研究では，MSD-HMMによりスペクトルと基本周波数を同時にモデル化 [2] することで，自然性の高い歌声合成が可能となっている．また，歌声は，通常の発音とは異なり，音長や音高の変化が大きな表現の要因として含まれるため，それらを考慮して歌声モデルを作成する必要があると考える．そこで，歌声データベースを用いて，楽譜から得られる情報を利用することで，音の高さ，長さを考慮したモデル化を行っている．本システムでは，楽譜情報に基づき学習を行い，また楽譜情報を入力として自動的に歌声を合成することにより，効率的な歌声合成が可能となる．

2. 歌声合成システム

本研究で構築する歌声合成システムを図1に示す．本システムは，学習，合成の2つのパートからなり，それぞれの過程において，楽譜情報から求められる歌詞，音高，音長情報を入力としている．

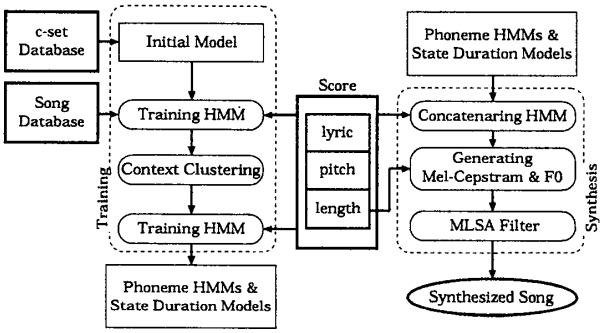

図 1: HMM歌声合成システム

2.1 モデルの学習

本研究ではメルケプストラムと基本周波数の系列を特徴パラメータとして歌声HMMを学習する．メルケプストラムの分析条件を表1に，基本周波数の分析条件を表2にそれぞれ示す．

以上の分析を行い，更にメルケプストラムと基本周波数のそれぞれの動的特徴と2次動的特徴を求める．これら全てのベクトルを一つに結合し，これを特徴量として学習を行う．

†名古屋工業大学 知能情報システム学科，
Dept. of Computer Science, Nagoya Institute of Technology

学習データ	会話文データ	歌声データ
データ数	3000文 (男性20名，各150文)	60曲 (男性1名)
サンプリング周波数	16kHz	
フレーム周期	5ms	
分析窓長	25ms	
窓関数	Blackman窓	
分析	24次メルケプストラム分析	

表 1: データベースの分析条件（メルケプストラム）

まず，会話文データベースを利用して，当該音素の初期モデルを作成し，HMMの学習を行う．次に，歌声データベースを用いて，音階，音長を考慮した歌声モデルとして再学習する．その後，コンテキストクラスタリングによって，スペクトル，基本周波数，継続長の各状態を共有する．

2.2 歌声合成

歌声は，楽譜データを入力として合成される．

まず，楽譜に基づき，歌詞，音階，音長に依存したモデルを選択し，それらを連結する．また，楽譜から得られるモーラ長から音素継続長を決定し，状態継続長分布より，各状態の継続長を求める．そして，パラメータ生成アルゴリズム [3] によってスペクトルと基本周波数が生成される．最後に，MLSAフィルタ [4] によって歌声音声が合成される．

3. 歌声モデル

音声を構成する基本的な要素として，スペクトルと基本周波数が挙げられる．これらはアクセントや声の高さなどの様々な要因によって影響を受ける．このような変動要因をまとめてコンテキストと呼ぶ．歌声の場合，これらとは異なる変動要因があることを仮定して，モデルの作成に楽譜情報を利用する．

本研究では，コンテキストとして，音素，音高（楽譜の音階），0.1秒単位のモーラ長の3点を考慮する．

また，それぞれのコンテキストは，該当音素と，その先行，後続を含めた3音素を1つとして (triphone)，1モデルの単位とする．これらの情報は楽譜データから一意に決定される．

以上の条件で，スペクトル，基本周波数，状態継続長を，それぞれHMMでモデル化する．

学習データ	会話文データ	歌声データ
データ数	3000文 (男性20名，各150文)	60曲 (男性1名)
サンプリング周波数	16kHz	
フレーム周期	5ms	
分析窓長	25ms	
分析ツール	STRAIGHT	
出力の上限	300Hz	370Hz
出力の下限	70Hz	

表 2: データベースの分析条件（基本周波数）

4. モデルのクラスタリング

モデル数はコンテキスト要素の増加により指数的に増加する．そのため有限なデータでモデルを学習する際，学習データの不足から，信頼性の低いモデルが構築される問題がある．そこで，本研究では，MDL 基準を用いた決定木に基づくクラスタリング [5] を行う．この手法では，学習データ不足のモデルは，類似したコンテキストに対応するモデルを用いることにより，モデルの信頼性の低下を防ぐことができる．

クラスタリングはコンテキストに対する質問により，木を作成する．このときに用いる質問の概要を以下に示す．

- 音素に関する質問
  - ある音素の種類（破裂音，摩擦音など）に該当するか
  - ある音素に該当するか
- 音高に関する質問
  - ある音高よりも高い音声か
- 音長に関する質問
  - ある長さよりも長い音声か

これらの質問を先行音素，当該音素，後続音素それぞれに対して適用する（図 2）．

スペクトルと基本周波数のモデルは，それぞれ各状態毎に決定木を作成し，状態継続長のモデルは全ての状態をまとめて一つの決定木を作成する．

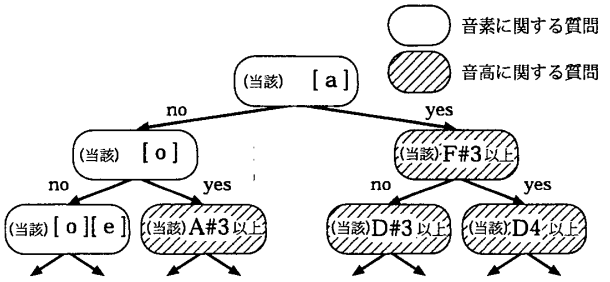

図 2: メルケプストラムに対する決定木（一部）

5. 歌声合成実験

5.1 実験条件

本研究では音声データベースとして，歌声データベースと通常会話文データベースを用いた．

歌声データベースは楽譜データ（MIDI データ）と音声データから構成されている．この歌声データベースには，音素に関する境界情報が無く，また，音素の出現頻度のバランスが考慮されていない為，初期モデルの学習では通常会話文データとして，ATR 日本語音声データベースの c-set を用いた．

使用した HMM は，5 状態 left-to-right モデルである．

学習データ中に存在した音階の幅は，G2 から E4 であった．音長は，モーラ長で 0.1 秒単位で考えたところ，0 秒から 2.5 秒であった．以上の 2 つのコンテキストと音素について，先行，当該，後続を考慮し，直積をとったものをラベルとした．

歌声合成実験は以下の 2 通りの方法で基本周波数系列を生成し，それぞれの合成音を比較した．

手法 1 学習したモデルからパラメータを生成する．

手法 2 楽譜データの音高情報より，基本周波数に相当するパラメータを計算する（従来法）．

また，これらの手法で合成した歌声を対比較試験により比較した．比較試験では，歌声 8 曲からそれぞれ 2 箇所切り取り，16 個の歌声データを用意した．試験では被験者 10 人により評価した．

5.2 実験結果

各手法により生成した基本周波数パターンを図 3 で比較する．手法 1 では，手法 2 に比べ，より自然性のある基本周波数パターンが得られた．しかし，手法 1 では音程が外れて合成される事があった．これはクラスタリングの際に異なる音高のモデルと共有されたためと考えられる．手法 1 の基本周波数は手法 2 のものよりも全体的にわずかに値が低かったが，これはデータベースの歌い手の特徴が現れたものと思われる．

また，対比較試験の結果を図 4 に示す．手法 1 では，手法 2 に比べ話者性が強く現れていた事が分かる．

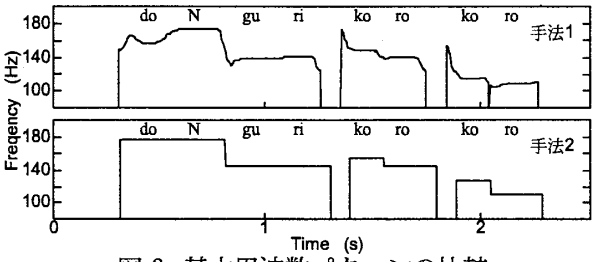

図 3: 基本周波数パターンの比較

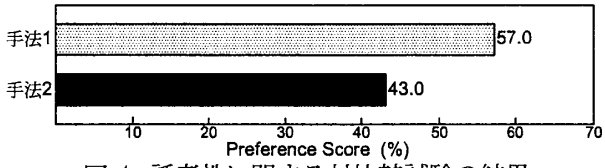

図 4: 話者性に関する対比較試験の結果

6. むすび

本研究では HMM に基づく音声合成システムを応用した歌声合成システムを構築した．音高，音長情報を考慮した歌声のモデル化と，スペクトルに加え基本周波数を HMM によって同時にモデル化した事により，話者の自然性が向上した．しかし，音程の外れた歌声が合成されることがあった．

今後の課題として，音程の問題の改善，学習時のクラスタリングにおける分割基準の検討，パワー制御に関する検討などが挙げられる．

参考文献

[1] 益子 貴史, 徳田 恵一, 小林 隆夫, 今井 聖,『動的特徴量を用いた HMM に基づく音声合成』, 信学論, J79-D-II, 12, pp.2184-2190, 1996.

[2] 徳田 恵一, 益子 貴史, 宮崎 昇, 小林 隆夫,『多空間上の確率分布に基づいた HMM』, 信学論, J79-D-II, 7, pp.1579-1589, 2000.

[3] K.Tokuda, T.Kobayashi and S.Imai,『Speech parameter generation from HMM using dynamic featuers』, Proc.of ICASSP, pp.660-663, 1995.

[4] 今井 聖, 住田 一男, 古市 千恵子,『音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ』, 信学論 (A), J66-A, 2, pp.122-129, Feb.1983.

[5] J.J.Odell,『The Use of Context in Large Vocabulary Speech Recognition』, PhD dissertation, Camblidge University, 1995.