2003－MUS－51　(13)
2003／8／5

隠れマルコフモデルに基づいた歌声合成システム

酒向 慎司[†]　宮島 千代美[†]
徳田 恵一[†]　北村 正[†]

隠れマルコフモデルに基づく音声合成方式を歌声合成に拡張することにより構築した歌声合成システムについて述べる．本システムでは，歌い手の声の質と基本周波数パターンに関する特徴をモデル化するため，スペクトルと基本周波数パターンを HMM により同時にモデル化している．特に，自然な歌声を合成する上で重要な要素となる音符の音階や音長の基本周波数パターンへの影響を精度良くモデル化するため，楽譜から得られる音階と音長を考慮したコンテキスト依存モデルを構築している．これらのモデルに対して決定木によるコンテキストクラスタリング行うことで，未知の楽曲からの歌声合成が可能となっている．実験から，歌い手の特徴を再現し，なめらかで自然性の高い歌声の合成が可能であることを示す．

A Singing Voice Synthesis System Based on Hidden Markov Model

Shinji Sako,[†] Chiyomi Miyajima,[†] Keiichi Tokuda[†]
and Tadashi Kitamura[†]

We describe a singing voice synthesis system by applying HMM-based speech synthesis technique. In this system, a sequence of spectrum and F0 are modeled simultaneously in a unified framework of HMM, and context dependent HMMs are constructed by taking account of contextual factors that affects singing voice. In addition, the distributions for spectral and F0 parameter are clustered independently by using a decision-tree based context clustering technique. Synthetic singing voice is generated from HMMs themselves by using parameter generation algorithm. In the experiments, we confirmed that smooth and natural-sounding singing voice is synthesised. It is also maintains the characteristics and personality of the donor of the singing voice data for HMM training.

1. はじめに

今日，様々なテキスト音声合成システムが開発され，人々の身近なところで利用されつつある．また，品質向上に留まらず，個人性や感情といった豊かな表現を可能とするための研究が各所で進められており，コンピュータとの対話手段の他にも様々な用途が期待されている．例えば，コンピューターによる自然な歌声の合成は，エンターテイメント，アミューズメント分野への応用を考えることができ，歌声を合成する試みは，これまでにも幾つか提案されている[1),2)]．

このような歌声合成システムを構築するために，テキスト音声合成を利用することが考えられる．これまでに提案されてきた音声合成システムの多くは単位選択という方式に分類される．これは音素や音韻といった音声単位ごとに分類した波形データを，合成したいテキストに従ってつなぎ合わせることで音声を合成する手法である．発声された音声波形を利用できるため，クリアな合成音を得やすいというメリットがある一方，接続部分の歪みが生じやすい，多様な声質や発話スタイルなどを得ようとすると，膨大な波形データを必要とするなどの問題がある．それに対して，隠れマルコフモデル (Hidden Markov model: HMM) に基づいた音声合成手法[3)] では，これらの問題を解決するため，音声認識の分野で広く利用されている HMM を利用し，HMM 自身から音声パラメータを生成する．この手法の特長として，動的特徴量を考慮したパラメータ生成アルゴリズム[4)] によって，滑らかに変化する音声パラメータが得られるほか，モデルパラメータの変換により，別の話者への適応や，多様な声質や感情を表現した音声の合成に柔軟に対応できるなどの点があげられる[5)~7)]．

本論文では，我々がこれまでに提案してきた HMM に基づいた音声合成手法を拡張した歌声合成手法について検討し，歌い手の話者性や歌い方の特徴を再現

† 名古屋工業大学 大学院工学研究科
Department of Computer Science and Engineering, Nagoya Institute of Technology

可能な歌声合成システムを構築することを目的とする．本システムの大きな特徴は，システムのすべてのモデルパラメータを学習データ提供者の歌声により自働学習する点にある．音の高さや長さは，合成時に楽譜から一意に定めることもできるが，そこから合成される歌声は単調で機械的なものになり，歌声としての魅力に欠けるものである．実際の歌声では，声質のほか声の高さやそれらの時間的な構造などにより，それぞれの歌い手独自のスタイルが存在している．そこで，自然の歌声にあるような音の高さの変化を再現するため，声質を表すスペクトル情報と高さを表す基本周波数を，可変次元に対応した多空間上の確率分布に基づく HMM (Multi-Space probability distribution HMM; MSD-HMM)[8] を用いてモデル化している．

さらに，より精密なモデル化を行うために，コンテキスト依存モデルを学習する．歌声は，通常の会話やテキストの読み上げなどの場合と比較して，発声する音の高さや時間的な長さ，または声の強弱などの変動の様子が大きく異なることから，歌声に特化したモデル化手法が必要となる．本手法では，楽譜から得られる歌詞のほか，音高と音長をコンテキストとして考え，前後の環境を考慮したそれらの組み合わせについてモデルを分類したコンテキスト依存モデルとしている．これらのモデルに決定木によるコンテキストクラスタリング[9] を適用することにより，未知の楽譜に対しても自然な歌声の合成を可能としている．このようにして学習データに基づいて得られた歌声モデルは，楽譜上では表現できない歌い手の持つ様々な特徴を備え，合成時にそれらを再現することが可能となる．

実験では，童謡など 60 曲を収録して構築した歌声データベースから歌声モデルを学習し，本論文では，特に未知の楽曲から自然な歌声の合成が可能であることを示す．

以下，本稿は次のように構成されている．2 節では HMM に基づいた歌声合成システム，3 節でデータベースの収録と整備，4 節で実験及び合成された歌声の評価と考察を述べ，最後に 5 節でまとめる．

2. HMM 歌声合成システム

本研究で提案する歌声合成システムは，大きく分けて学習部と，歌声合成部の 2 つに分けられる．学習部では，初期モデルを元に楽譜情報と歌の波形データから成る歌声データベースを用いて歌声モデルを学習し，合成部では，合成したい歌の楽譜情報を入力として，学習部で得られた歌声モデルから歌声を合成する．なお，各部で利用される楽譜データとして，MIDI[10] を利用している．

2.1 学 習 部

学習部では，音声合成用のモデルとして，スペクトルパラメータ，基本周波数，および継続長を HMM によって音素単位でモデル化する．音声のスペクトルパラメータは，連続 HMM によってモデル化することができるが，基本周波数は有声区間では連続値をとり，無声区間では値を持たない可変次元の時間系列信号であるため，通常の連続 HMM や離散 HMM で直接モデル化することはできない．そこで，可変次元に対応した多空間上の確率分布に基づく MSD-HMM[8] を用いて，スペクトルパラメータとしてメルケプストラム[11] を多次元ガウス分布，基本周波数の有声音を 1 次元空間，無声音を 0 次元空間のガウス分布として単一の枠組みの中で同時にモデル化する．

また，歌声に表れる声の特徴には，様々な要因によって影響をうけ変動していると考えられる．例えば，同じ音階の声であっても，広い範囲では楽曲のジャンルやテンポ，局所的には前後の歌詞や音階などによって，異なる特徴を持っていると考えられる．テキスト音声合成においても，テキストから得られる言語的な情報が音声パラメータに影響を与えていると考え，それらの要因をコンテキストと呼び，コンテキストを考慮したモデル化が行われている．

コンテキストに依存したモデル化を行うことで，精度の高い歌声モデルを得ることができるが，コンテキストの種類に応じてその組み合わせの数も莫大となってしまう．また，すべてのコンテキストの組み合わせに対応したモデルについて，充分な学習を行うためには，あらゆるパターンを網羅したデータベースが必要となってしまうため，現実的ではない．

この問題に対する優れた解決法として，コンテキストクラスタリングによってモデル間でパラメーターを共有させる手法がある[12]．これは，二分木を用いて，モデルの集合を木構造に分割することで，類似したコンテキストの組み合わせごとにモデルパラメータをクラスタリングする手法である．木の各ノードには，コンテキストを二分する質問があり，各リーフノードには，特定のモデルに相当するモデルパラメータがある．任意のコンテキストの組み合わせは，ノードにある質問に沿って木を辿ることで，何らかのリーフノードに到達でき，該当するモデルを選択することができる．

図 1 に，学習部の概要を示す．まず，初期モデルとして，テキスト読み上げ文による音声データベースから音素単位のモデルを作成する．次に歌声データベースを用いて楽譜情報を考慮した学習を行う．ここでは，歌声のモデル化に効果的なコンテキストとして以下にあるものを考え，それぞれ当該および前後の環境に依存した歌声モデルを学習する．

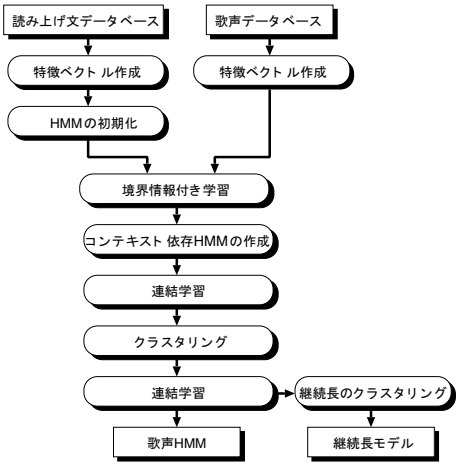

図 1 歌声システムの学習部
Fig. 1 Training part of the system

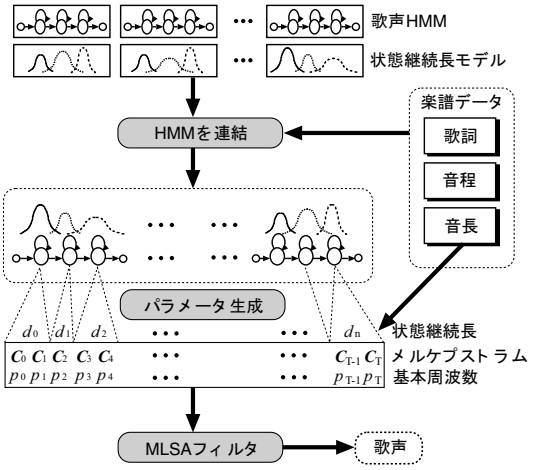

図 2 歌声合成システムの合成部
Fig. 2 Synthesis part of the system

- **歌詞**：音素名をコンテキストし，母音・子音・有声音など音素に関する質問を適用．
- **音高**：当該・先行・後続の音符の MIDI 音階値をコンテキストとし，音階の高低に関する質問を適用．
- **音長**：各音符の長さを 100ms 単位で表したものをコンテキストとし，音符の長さ関する質問を適用．

さらに，これらのコンテキストに基づいて MDL 基準を用いた決定木によるコンテキストクラスタリングを行い，あらゆるコンテキストの組み合わせに対応した歌声モデルを参照可能とする．一方，音素に対応する各歌声モデル内部の時間構造を表す状態継続長モデルは，HMM の各モデルの状態継続長を多次元ガウス分布でモデル化し，そのモデルパラメータは HMM の連結学習時に作られるトレリス上で求める[13]．

2.2 合 成 部

図 2 に示す合成部では，楽譜情報（歌詞つき MIDI データ）を入力として歌声を合成する．まず，楽譜から得られる，歌詞，音高，音長情報に基づいて，歌声モデルから対応するモデルを選択する．次に，楽譜から与えられた各音符の長さを制約として，音符内の音素継続長及び音素内部の状態継続長を，各モデルの状態継続長分布に基づいた尤度最大化基準により決定する．得られた状態系列から，パラメータ生成アルゴリズムによってメルケプストラムと基本周波数パラメータの列を生成する[4]．最後に，生成されたパラメータに基づいて MLSA フィルタを励振させることで，歌声を合成する[14]．

3. 歌声データベース

統計的アプローチによる音声のモデル化には，データベースが不可欠である．これまで，テキスト音声合成に利用可能な読み上げ文章などの音声データベースは様々なものが整備されているが，現在のところ歌声に関しては適切なものが入手できないため，新たに歌声データベースの整備を行った．

3.1 データの収録

童謡を中心とした 60 曲を用いて，男声 1 名の歌い手による歌声を収録した．収録の際は，歌い手は楽曲の MIDI 演奏をヘッドホンでモニターしながら，DAT デッキを用いて歌声を録音した．なお，MIDI データについては，WWW 上で公開されているものを収集した．

収録された歌声は，DAT-Link+を用いて，サンプリング周波数 16kHz, サンプルサイズ 16bit，モノラル音声のデータとして計算機へ取り込んだ．

3.2 データの整備

様々なコンテキストに基づいて歌声モデルを学習する場合，コンテキスト情報の信頼性は，合成音の品質に大きく影響を与えると考えられるため，テキスト音声合成においても，正確なコンテキスト情報の整備が重要な要素となっている．本手法では楽譜情報をコンテキスト要因の一つとして扱うが，実際の収録データには，歌詞の読み誤りや歌声と楽譜の音階が一致しないなどの誤りが含まれている．

予備的な実験から，誤ったコンテキスト情報を含むデータから学習を行った場合，部分的に音程を外した歌声が合成されるなどすることがわかっている．そこで，高精度な歌声のモデル化を行うため，歌声データベースの整備として以下の作業を行った．

- MIDI データの編集

学習時に利用するため，主旋律データを作成する

表 1 データベースの分析条件 (メルケプストラム)
Table 1 Experimental condition for Mel-cepstral analysis

学習データ	歌声
データ数	60 曲 (男声 1 名，約 72 分)
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
分析次数	24 次

表 2 データベースの分析条件 (基本周波数)
Table 2 Experimental condition for F0 analysis

学習データ	歌声
サンプリング周波数	16kHz
フレーム周期	5ms
分析窓長	25ms
上限/下限	370Hz / 70Hz

とともに，音符ごとに歌詞情報を付加する．

- 歌声に合わせた MIDI データの修正
  実際の歌声と伴奏に使用した MIDI データの間の，歌詞，旋律，音程の誤りについて，MIDI データを歌声に合わせて修正する．
- 音素境界ラベリング
  不特定話者の音素 HMM を用いて，歌声の音素境界の Viterbi アライメントを求め，音素境界を手作業で修正する．

4. 実　　験

前節で作成した歌声データベースを用いて歌声モデルを学習する．さらに，学習したモデルに対して学習データに含まれない楽曲を入力として，歌声を合成した．

4.1 学習データの作成

まず，収録した 60 曲の歌声についてメルケプストラム分析と基本周波数抽出を行い，HMM の学習データを作成した．基本周波数の抽出には TEMPO[15] を用いた．メルケプストラム分析，基本周波数抽出に関する分析条件をそれぞれ**表 1**，**表 2** に示す．

得られた分析データから，0〜24 次のメルケプストラム係数ベクトルと基本周波数値をフレーム毎の静的特徴量とし，これに前後のフレームから計算される動的特徴量を加えたものを歌声モデルの学習データとした．t 番目のフレームのメルケプストラムの静的特徴をそれぞれ $\boldsymbol{c}_t$ としたとき，その動的特徴量 $\Delta\boldsymbol{c}_t$ および 2 次動的特徴量 $\Delta^2\boldsymbol{c}_t$ は以下の式 1,2 から計算した．

$$\Delta\boldsymbol{c}_t = \frac{1}{2}(-\boldsymbol{c}_{t-1} + \boldsymbol{c}_{t+1}) \tag{1}$$

$$\Delta^2\boldsymbol{c}_t = \frac{1}{4}(\boldsymbol{c}_{t-1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t+1}) \tag{2}$$

基本周波数 p_t についても同様に $\Delta p_t, \Delta^2 p_t$ を求め，メルケプストラムと基本周波数の 2 つのストリームからなる学習ベクトルの次元数は合計 78 次元となる．

4.2 歌声モデルの学習

歌声データから抽出されたメルケプストラムと基本周波数を MSD-HMM によってモデル化する．HMM は単混合 5 状態の left-to-right モデルとし，音素はポーズと無音を含んだ 36 種類とした．

まず，音素バランスの考慮されたテキスト読み上げ文データベースから，初期モデルを作成する．ここでは楽譜に関する情報は考慮せず，音素単位で HMM の学習を行った．この初期モデルから歌声データを用いて，楽譜情報に依存したコンテキスト依存モデルを学習した．前節で述べたとおり，歌詞から得られる音素の他，MIDI データの音階表現値を利用した音高と，当該音素を含むモーラの時間長を 100ms 単位で分類した音長のコンテキストについて先行，当該，後続を考慮し，さらに MDL 基準に基づいたコンテキストクラスタリングを行い各モデルの状態を共有化した．

なお，メルケプストラム，基本周波数，継続長の各モデルにコンテキストクラスタリングを行う際に，以下の 2 種類の手法を検討した．

手法 A: 各モデルで，2.1 節で述べたすべてのコンテキストを適用する．

手法 B: 基本周波数モデルに関してのみ，当該音高別にクラスタリングを行う．

これは，当該音高をコンテキストとした場合 (手法 A)，異なる音高のデータが一つのクラスタに分類され，正しい音高が再現できなくなる可能性があるためである．

4.3 歌 声 合 成

学習データに含まれていない楽曲を用いて歌声を合成する．まず，基本周波数のモデル化による自然性を確認するため，以下の二つの手法から基本周波数系列を生成した．

手法 1: 楽譜から，直接音階に相当する基本周波数系列を生成．

手法 2: 学習した基本周波数モデルから生成 (提案法)．

なお，手法 1 の音階に相当する基本周波数は，以下の式 3 から求める．p は MIDI 規格の音階を表す数値であり，$p = 57$ が男声の基本周波数 110Hz に相当する「ラ」の高さを表現している．

$$F_p = 110(2^{\frac{1}{12}})^{p-57} \tag{3}$$

4.4 評価と考察

クラスタリングによって，メルケプストラム，基本周波数 (手法 A) のモデルから構築された決定木の一

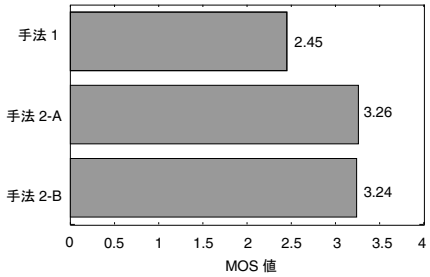

図 5 各手法の自然性についての評価
Fig. 5 Evaluation of naturalness for each methods

部を**図 3** に示す．主に，メルケプストラムでは音素に関する質問が多く適用されており，クラスタリング時に音高を固定しない場合 (手法 A) の基本周波数モデルでは，音高に関する質問が多く適用されていることがわかる．

また，**図 4** に生成された基本周波数 (手法 1, 2) とスペクトル系列の一部を示す．手法 1 では音符単位に階段状に変化する平坦な基本周波数パターンが生成されているが，手法 2 では HMM の内部状態に沿って複雑に変化する基本周波数パターンが生成されている．また，楽譜上の音階と比較して，必ずしもそれに一致せず，全体的にやや低くなっていることがわかる．これは学習データの歌い手が実際の楽譜よりも低く歌う傾向があり，その特徴がモデル化に表れていることが考えられる．

4.5 主観評価試験

合成された歌声の品質を評価するため，手法 1，手法 2 についてクラスタリング時の手法 A，B を考慮した 3 通りの方法から，学習データに含まれない 10 曲を合成した．各曲から 4 小節程度に分割した合計 32 のサンプルを切り出し，被験者ごとにランダムに選択した 20 サンプルを用いて，主観評価試験を行った．各被験者は，各サンプルの自然性について 5 段階で評価を行った．10 人の被験者から得られた MOS 値を**図 5** に示す．

試験の結果から，手法 2 の基本周波数モデルから生成した場合では，手法 1 の楽譜の旋律から生成した場合と比べ，高いスコアを得ている．これは図 4 の結果と同様に，HMM 内部の各状態ごとに，コンテキストに応じた変化をもつことが自然性の向上に大きく寄与していると考えられる．

また，基本周波数のクラスタリング手法による違い (手法 A, B による) に関しては，とくに明確な差が得られていない．本来，音高を固定しないでクラスタリングを行う場合には，学習データが正しい音高のクラスタに集まることは保証されず，合成される歌声の音階がずれる可能性を持つことになる．今回のサンプルでは，そのようなケースが見られなかったのは，音高を固定しないクラスタリングにおいても，音高に関する質問が他のコンテキスより多く適用され，音高に応じたクラスタの分割が精度良く行えていることが原因の一つとして考えられる．

なお，非公式な結果として，本システムから合成された歌声を受聴した者の多くが，合成された歌声がデータベースの歌い手によるものと同定でき，また，当人の歌い方の特徴を感じ取ることができたことを付け加えておく．

5. ま と め

HMM に基づいた音声合成手法を拡張し，歌声合成システムを構築した．本システムでは，MSD-HMM によりスペクトルと基本周波数パターンを同時にモデル化することで，歌い手の特徴を再現した歌声合成が可能である．また，コンテキストクラスタリングにおいて，楽譜から得られる音高や音長ををを利用することにより，歌声の精密なモデル化が可能であることを示した．

本研究では，新たに童謡 60 曲からなる歌声データベースの収録と整備を行い，それを用いた実験から，学習データに含まれない楽曲に対しても自然な歌声を合成可能であり，また，合成された歌声には，学習データの話者性が再現できていることを確認した．

一方，曲によってはパワーの変動が不安定になるという問題も見られたため，今後の課題として，音の強弱に関するコンテキストの導入が考えられる．また，合成音に対して様々なフィルタ処理を行い，より実用性の高い歌声合成システムを開発することなどが挙げられる．

謝辞 卒業研究を通して，本実験で使用したデータベースの収録，整備や実験にあたられた伊藤正典氏，石川ちさと氏の両名に感謝いたします．

参 考 文 献

1) 吉田由紀, 中嶌信弥: 歌声合成システム CyberSingers, 情報処理学会研究報告 音声言語情報処理, Vol. 25, No. 8 (1995).
2) Macon, N. W., Jensen-Link, L. J., Oliverio, J. and Clements, M. A.: A Singing voice synthesis system based on sinusoidal modeling, *Proc. of ICASSP*, Vol. 1, pp. 434–438 (1997).
3) 吉村貴克, 徳田恵一, 小林隆夫, 北村正: HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化, 信学論 (D-II), Vol. J83-D-II, No. 11, pp. 2099–2107 (2000).
4) Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T.: Speech Parameter Generation Algorithm for HMM-based Speech, *Proc. of*

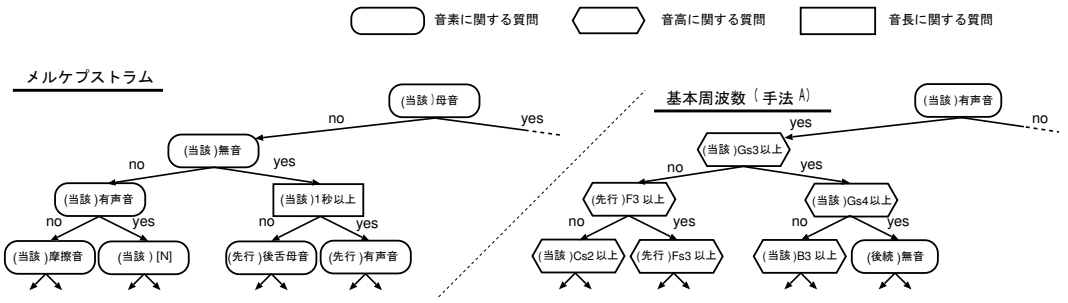

図 3 クラスタリングによって作成された各モデルの決定木
Fig. 3 Examples of decision trees

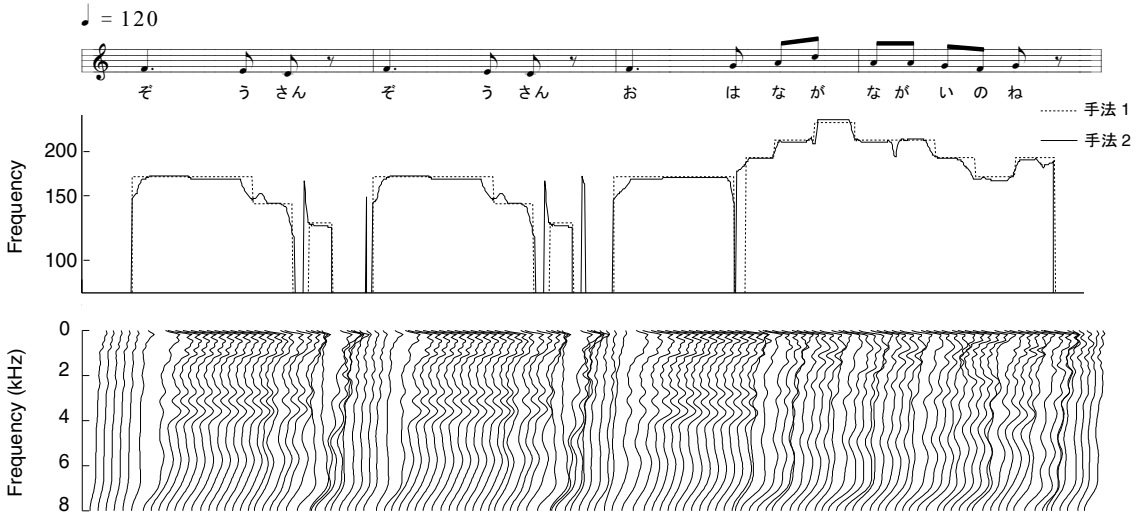

図 4 合成されたスペクトルと基本周波数パターン
Fig. 4 Example of generated spectra and F0 pattern

ICASSP, Vol. 3, pp. 1315–1518 (2000).

5) Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T.: Speaker Adaptation for HMM-based Speech Synthesis System Using MLLR, *Proc. of The Third ESCA/COCOSDA workshop on Speech Synthesis*, pp. 273–276 (1998).

6) Yoshimura, T., Tokuda, K., Masuko, T. and Kobayashi, T.: Speaker Interpolation in HMM-based speech synthesis system, *Proc. of EUROSPEECH*, Vol. 5, pp. 2523–2526 (1997).

7) Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K. and Kitamura, T.: Eigenvoices for HMM-based speech synthesis, *Proc. of ICSLP*, pp. 1269–1272 (2002).

8) 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 (D-II), Vol. J79-D-II, No. 7, pp. 1579–1589 (2000).

9) 篠田浩一, 渡辺隆夫: 情報量基準を用いた状態クラスタリングによる音響モデルの作成, 信学技報, Vol.SP96-79, pp. 9–16 (1996).

10) MIDI Manufactures Association. `http://www.midi.org/`.

11) 徳田恵一, 小林隆夫, 斉藤博徳, 深田俊明, 今井聖: メルケプストラムをパラメータとする音声のスペクトル推定, 信学論 (A), Vol. J74-A, No. 8, pp. 1240–1248 (1991).

12) Odell, J. J.: *The use of context in large vocabulary speech recognition*, PhD Thesis, Cambridge University (1995).

13) 吉村貴克, 徳田恵一, 益子貴史, 小林隆夫, 北村正: HMM に基づく音声合成のための状態継続長モデルの構築, 信学技報, Vol. DSP98-85, No. 262, pp. 45–50 (1998).

14) 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), Vol. J66-A, No. 2, pp. 122–129 (1983).

15) Kawahara, H., Masuda, I. and Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time -frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,, *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207 (1999).