Vol.2012-MUS-94 No.7
Vol.2012-SLP-90 No.7
2012/2/3

ラップスタイル歌声合成の検討

才野 慶二郎[†1] 大浦 圭一郎[†2] 橘 誠[†1]
剣持 秀紀[†3] 徳田 恵一[†2]

ラップのような短時間のうちに音高などの特徴が大きく変動するスタイルの歌い方は，それを適切に表現するための記譜法が確立されておらず，従来のように五線譜基づく合成の仕組みではユーザが直観的にそのスタイルの歌声を再現することが難しかった．本稿では，ラップスタイルの歌唱のための記譜法を定義し，それを用いて HMM 歌声合成の枠組みでラップスタイルの歌声合成を行った．その結果得られた合成音声はラップ特有のグリッサンド技法によるピッチ変動の現象を含むものになっていることが確認された．また，合成時に得られる対数基本周波数系列を素片接続型の歌声合成器に与えてラップスタイルの歌声を合成することも試みた．

Rap-style Singing Voice Synthesis

Keijiro Saino,[†1] Keiichiro Oura,[†2]
Makoto Tachibana,[†1] Hideki Kenmochi[†3]
and Keiichi Tokuda[†2]

This paper addresses rap-style singing voice synthesis. Since it has not been very clear how to write a musical score for rap-style songs, existing singing voice synthesis systems based on musical scores are not suitable for synthesizing them with an intuitive input. Here a new type of musical score specialized for a rap-style is defined. An HMM-based singing voice synthesis system is used to realize an automatic synthesis of realistic rap-style singing. Glissando phenomenon which is special for the style could be found in synthesis results. It was also tried to apply pitch parameters generated from the HMMs to a sample-concatenation-based singing voice synthesis system.

1. はじめに

人が実際に歌を歌うことなく任意の歌詞とメロディを持った歌声を人工的に作ることのできる歌声合成技術は，楽曲制作のプロセスに新たな可能性をもたらした．それは単純に人間の歌い手なしにボーカルトラックを作成することを可能にしたというだけではなく，ユーザの操作次第で合成音声に非常に多様な歌唱表現を与えることが可能になことから，音楽的な表現そのものの可能性を大きく広げた．現在，歌声合成技術は幅広くエンドユーザの利用するところとなっており[1)]，同時にさらなる表現ポテンシャルの向上が切望されている．

従来は，歌声合成のための入力としての譜面情報はピアノロールや五線譜のような形式で与えることが想定されてきた．しかしながら，既存の譜面情報で適切に記述することが可能であるかどうか不明確なジャンルやスタイルの歌声に対してはその的確な入力方法が存在せず，合成を行うことが難しかった．そこで本稿では，音高などの特徴が短時間に激しく変動することで知られ，特別な記譜法の確立されていないラップスタイルの歌声をユーザの直観的な入力により合成することのできるシステムを目指し，ラップスタイル歌唱のための特別な記譜法を提案する．歌声合成のためのアプローチとしてはルールベースのものとコーパスベースのものが考えられる．ルールベースの手法は歌唱スタイルのように個人ごとに異なる特徴を再現するためにはその都度ルールを人手で設定する必要があり，労力がかかる．一方，コーパスベースの手法であれば，適切な譜面を用意すれば自動的に元歌唱者のスタイルを再現する合成音声が得られる．そこで本稿ではコーパスベースのシステムである HMM 歌声合成システム[2)] をラップスタイルの歌声合成のために使用する．さらに，素片接続型の歌声合成システムである VOCALOID[3)] にこの手法により出力された音高情報を与えて合成することも試みる．

以下，2 章でラップスタイルの歌唱について考察した上で，ラップのための特別な記譜法を定義する．3 章で HMM 歌声合成システムについて述べ，4 章で音声合成実験およびその主観評価結果について記述する．5 章で HMM 合成の中間出力結果を素片接続型歌声合成器へ与えた合成を行う．最後に 6 章で全体のまとめと今後の課題について述べる．

†1 ヤマハ株式会社 研究開発センター
Corporate Research and Development Center, Yamaha Corporation
†2 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
†3 ヤマハ株式会社 yamaha+推進室
yamaha+ Division, Yamaha Corporation

Vol.2012-MUS-94 No.7
Vol.2012-SLP-90 No.7
2012/2/3

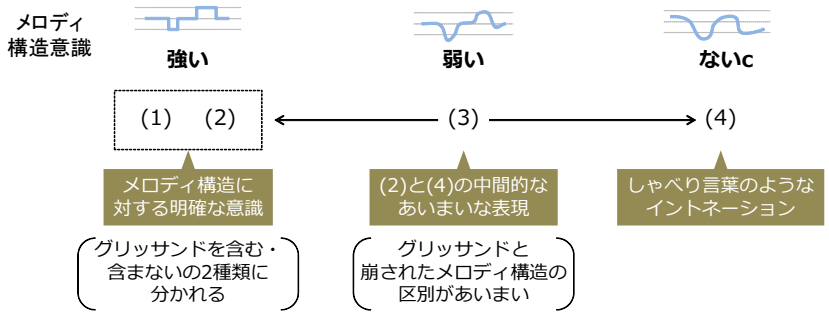

図 1 ラップサブスタイルの分類例
Fig. 1 An example of classification of substyles of rap.

2. ラップスタイル歌唱

ラップの歌唱スタイルは歌い手や楽曲のジャンルなどによって様々であり[4]，画一的に定義することは難しく，ラップと呼ばれるすべての歌唱スタイルを網羅する考察を行うことは現実的ではない．そのため本稿では，ラップスタイルについての研究の最初のステップとして，エイベックス・マネジメント社所属の音楽グループ m.o.v.e[5] のメンバーで作詞およびラップを担当する motsu 氏のラップを考察の対象とした．以降，本稿においてラップとは motsu 氏のものを指すものとする．

2.1 ラップスタイル歌唱のサブスタイル

本稿では,「メロディ構造の意識」という観点においてラップスタイルの中のサブスタイルの存在を考える（なおここでは特定の音階に則った離散的な音高の時間変化をメロディと呼ぶ）．図 1 にその分類の例を示す．本稿では歌唱者が意識の中で特定のメロディをどの程度明確に持ちそれに従うように歌うかによって，(1)・(2) から (3) を経て (4) までの連続的なスタイルに分類されるものと仮定した．(1) と (2) の違いは，グリッサンド（目標音高からスタートして明確な目標終了音高なしに歌唱の音高を降下もしくは上昇させる技法）を含むかどうかである．以上の分類のうち，motsu 氏によるラップスタイルの歌声で，(1), (2), (4) にあたる motsu 氏の歌声の対数基本周波数をプロットしたものを図 2 に示す．このうち (1), (2) においては motsu 氏自身により提示された歌唱の際に意識するメロディ構造も同時に示している．(1) では子音の発音に依存する局所的な沈み込みを除いて対数基本周波数が意識下のメロディ構造にほぼ従っている．また (2) では意識下のメロディ構造に従いながらも，motsu 氏の提示したグリッサンドを意識したことを意味する矢印の箇所で対数基本周波数の降下および上昇が確認できる．一方 (4) では，特定の音高に合わせることなくしゃべり言葉のようなイントネーションの音高変化になっている．このように，意識下のメロディ構造の存在を譜面に取り入れるか入れないかによってラップスタイル歌唱のための記譜法のデザインもまったく異なったものにならざるを得ないため，(1) から (4) までのスタイルが混在した歌唱を記譜の対象と考えることは大変難しい．そこで本稿ではラップスタイルの中でもある程度メロディ構造に従う歌い方である (1), (2) のみをラップ用譜面による表現の対象とした．

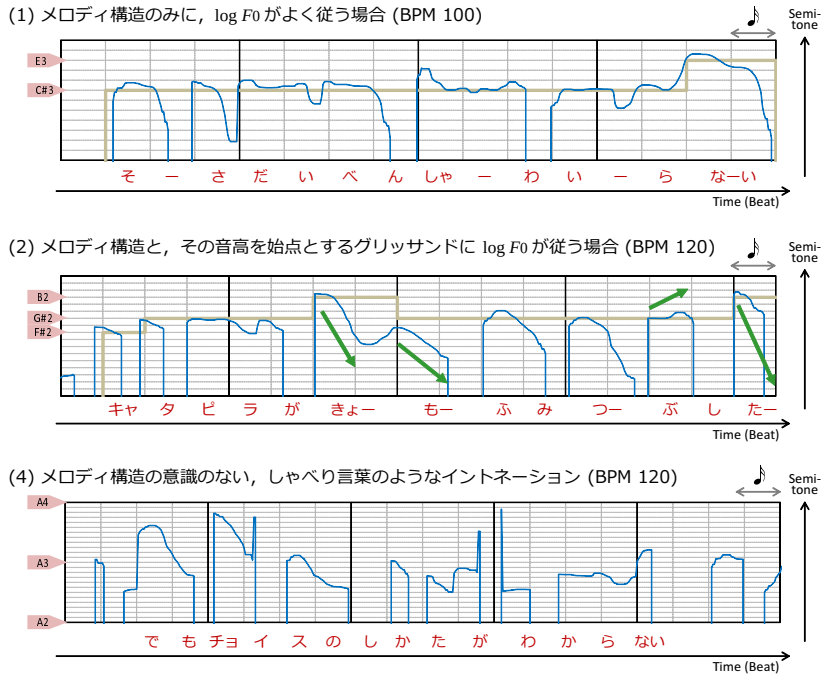

図 2 motsu 氏によるラップ歌唱の対数基本周波数軌跡の例
Fig. 2 Examples of log F_0 series of rap-style singing voice by motsu.

2.2 ラップスタイル歌唱のための記譜法の定義

歌声合成のために使用することを前提として，(1) および (2) のラップスタイル歌唱をよ

Vol.2012-MUS-94 No.7
Vol.2012-SLP-90 No.7
2012/2/3

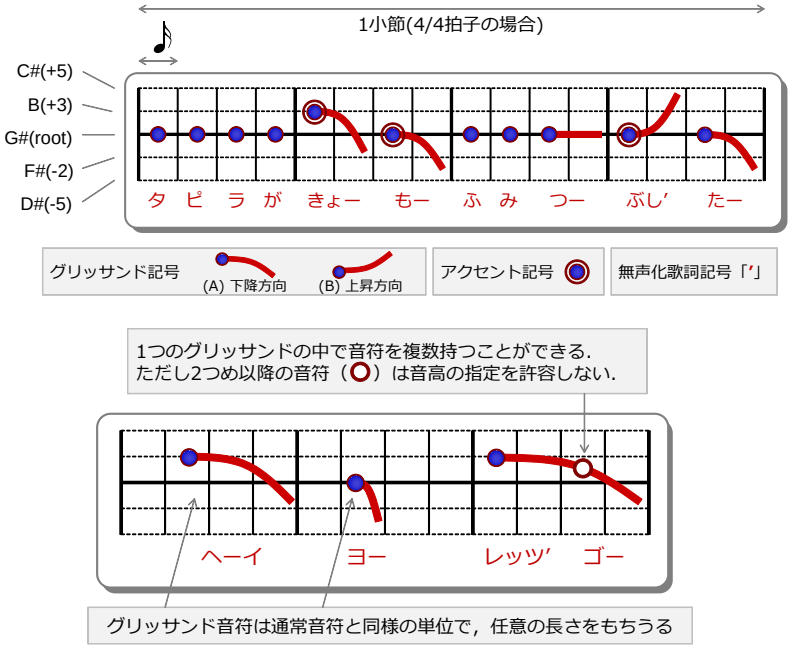

図 3 本稿で定義したラップ歌唱記譜法
Fig. 3 The defined musical notation rules.

く表現できるような記譜法を定義する．ここでは楽曲を，motsu 氏の楽曲で，その中でも典型的な，短調で，スウィングのリズムを含まない（譜面の音価どおりのリズムで発音される）ものとする．2.1 節の内容と motsu 氏の歌唱者自身としての意見を踏まえ，本稿では記譜法を図 3 のように定義した． その特徴は以下のようなものである．

- 音価の基本単位は 16 分音符とする．8 分音符以上の単位で 3 連符も許容する．
- 音階はルート音を中心に上下にそれぞれ 2 段階ずつ，計 5 段階の音高で構成される．ルートからの音高差は，半音単位で下から-5, -2, 0, +3, +5 のマイナーペンタトニックスケールとする．
- 歌詞は日本語のみを許容する．1 音符に任意の数のモーラを含むことができる．

また，ラップスタイル歌唱の音楽的表現のために必要になると思われる要素のための記号を以下のように定義した．

- **グリッサンド記号**：下降方向 (A) と上昇方向 (B) の 2 種類．
  始点の音高が指定する高さになるよう記述する．下降・上昇の程度の指定は許容しない．
- **アクセント記号**：音符に対して与える強勢の記号．
- **無声化歌詞記号**：直前のモーラを無声化して発音することを明示する記号．

3. HMM 歌声合成システム

音声合成のための方式は大きくはルールベースのものとコーパスベースのものに分かれる．VOCALOID[3] で採用されているようなルールベースの方式では自然な歌声を合成するための韻律のルールなどを人手により決定しなければならない．現状ではラップスタイルの自然な歌声を合成するためのルールが定かでなく，それを決定すること自体が容易なことではない．またルールを決定したとしても歌唱スタイルは歌唱者に依存するため，そのルールのその他の歌唱者に対する適切さが保証されないという問題もある．一方，コーパスベースの方式では元の歌唱者の自然歌唱でデータベースを作成し，それをもとに合成が行われる．中でも HMM を用いた歌声合成システムは元歌唱者の声質と歌唱スタイルを自動的に再現可能なシステムである[2]．従来 HMM 歌声合成システムでは譜面として五線譜を用いていたため，記譜法の定かでなかったラップスタイルの歌声合成には向かなかったが，適切に定義されたラップスタイル用の譜面を用いれば，システムによるモデルの自動学習による自然なラップスタイルの歌声合成が期待できる．そこで本稿では，2.2 節で定義されたラップスタイル歌唱用記譜法による譜面を用いて HMM 歌声合成システムによる合成を行う．

HMM による歌声合成システムは，学習部と合成部の 2 つのパートで構成される．学習部では，歌声データベース内の音声から抽出されたスペクトル，基本周波数，ビブラートパラメータの系列を，譜面を変換して得られるコンテキスト依存ラベルに基づき，音素単位の HMM でモデル化する．その際，モデルのパターン数の増加に伴いモデルあたりの学習用データ量が少なくなってしまう問題を回避するために，決定木を用いたコンテキストクラスタリングを行い，状態ごとにモデル間でパラメータを共有する．また，HMM の各状態の継続長および譜面上の当該音符開始タイミングからの時間ずれをガウス分布でモデル化し，パラメータを同時に学習する．近年では音高のモデル化には特別に，話者正規化学習の枠組みを歌声合成に応用した譜面上の音高と観測される対数基本周波数の差分のみをモデル化する音高正規化学習の手法[6] が用いられている．合成部では，ユーザから与えられた

Vol.2012-MUS-94 No.7
Vol.2012-SLP-90 No.7
2012/2/3

表 1 学習用ラップスタイル歌声データの仕様
Table 1 Singing voice data used for model training.

収録楽曲	motsu 氏の既発表曲 11 曲（ラップ部分のみ）
総歌唱時間	21 分 6 秒
楽曲テンポ	BPM 92〜130
歌唱者	motsu 氏（男性・日本語話者）
サンプリング周波数/量子化ビット数	48kHz/16bit
メルケプストラム分析手法	49 次 STRAIGHT メルケプストラム
メルケプストラム分析周期	5 ms
基本周波数抽出手法	SWIPE[7)]
基本周波数分析周期	5 ms

譜面データを合成用のコンテキスト依存ラベル系列に変換し，対応する HMM を連結する．その HMM から出力確率最大となるメルケプストラムおよび対数基本周波数の系列を得て，MLSA フィルタを駆動することで音声を合成する．

4. 主観評価実験

2.2 節で定義したラップスタイル歌唱のための記譜法で記述された譜面を用いて，それに対応する歌声をモデル化し合成する実験を行った．

4.1 学習データの準備

HMM 歌声合成システムには，学習用データとして，歌声とそれに対応する譜面のペアデータが必要となる．そこで本稿ではまずはモデル学習に用いるためのラップスタイル歌声データベースの構築と，それに対応する譜面データを得るところから取り組んだ．歌声データベース構築のために使用する楽曲として，motsu 氏の既発表楽曲の中から，2.2 節のラップ記譜法の定義の際に想定した，短調の，スウィングのリズムを含まない楽曲に該当する 13 曲を選定した．

まず譜面を得るために，前述の 13 曲に対し motsu 氏自身がどのようなメロディ構造の意識を持って歌唱を行っているかを前節の記譜法に従い motsu 氏自身により書き起こした．ただし，この 13 曲に限らず motsu 氏の歌唱スタイルは総じて同一の楽曲の中でも図 1 の (1)〜(4) を幅広く含むことがあるため，前節で定義した記譜法のみで全編通して原曲スタイルそのままに採譜することは難しい．そこで，今回の記譜作業においては原曲で (1), (2) の範囲を外れたフレーズに対しては motsu 氏自身がその範囲のスタイルに収まるような歌唱用フレーズを改めて作成した．以上のようにして得られた譜面に基づき，motsu 氏によるラップスタイル歌唱の収録を新たに行い，対応する歌声を得た．その 13 曲中 11 曲をモデル学習に用い，残りの 2 曲をテストセットとして合成に用いる．学習用歌声データの詳細を**表 1** に示す．

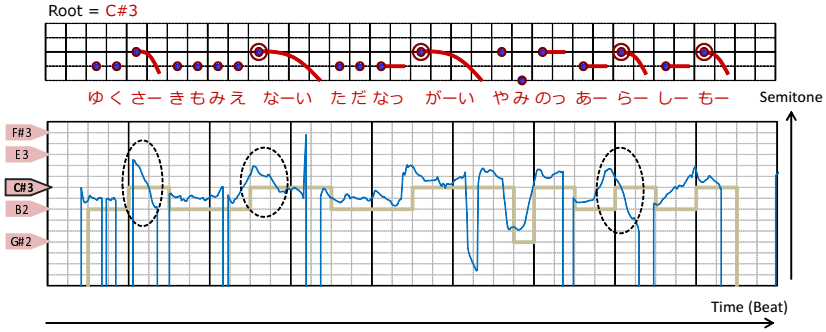

図 4 入力されたラップ譜の一部と合成時にモデルから生成された対数基本周波数系列 (BPM 128)
Fig. 4 A part of input rap score and contour of generated log F_0. (BPM 128)

4.2 HMM によるモデル化

本稿では統計モデルとして，明示的に状態継続長に関するモデルを組み込んだ隠れセミマルコフモデル (Hidden Semi-Markov Models; HSMM)[8)] を用いた．トポロジはスキップなしの left-to-right 型で，状態数は 5 とした．なお，このモデル化においてはグリッサンド記号およびアクセント記号はモデル学習のためのコンテキストとして考慮し，無声化歌詞記号は直前のモーラの母音を無声化母音としてモデル化するための記号として使用した．元となる歌声の歌唱スタイルが HMM によりモデル化されたことを確認するため，ラップスタイル歌声データベースの中から学習に用いなかった楽曲の譜面データを入力としてパラメータ生成を行った．そこで得られた対数基本周波数系列の例を図 4 に示す．図より破線で囲まれた部分で対数基本周波数の下降が確認できる．ここから，グリッサンド技法による歌声の音高下降の現象が HMM によりモデル化されていることが確認できる．

4.3 主観評価実験

本稿におけるラップスタイル歌唱のための記譜法におけるアクセント記号およびグリッサンド記号の導入が，合成音声の自然性にどの程度寄与しているかを確認するため，これらを学習・合成時のコンテキストとして考慮するかどうかについて表 2 の 4 つの手法を用意し，主観評価実験を行った．

歌声データベース内の楽曲で学習に用いなかった 2 曲を合成した音声をそれぞれ 2 小節程

Vol.2012-MUS-94 No.7
Vol.2012-SLP-90 No.7
2012/2/3

表 2 主観評価実験手法
Table 2 Subjective evaluation methods.

	アクセント	グリッサンド
手法 A	考慮しない	考慮しない
手法 B	考慮しない	**考慮する**
手法 C	**考慮する**	考慮しない
手法 D	**考慮する**	**考慮する**

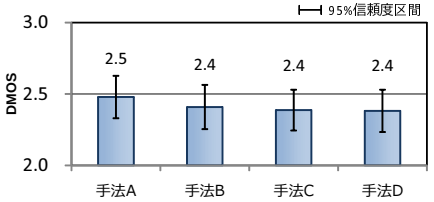

図 5 主観評価実験結果
Fig. 5 Subjective evaluation results.

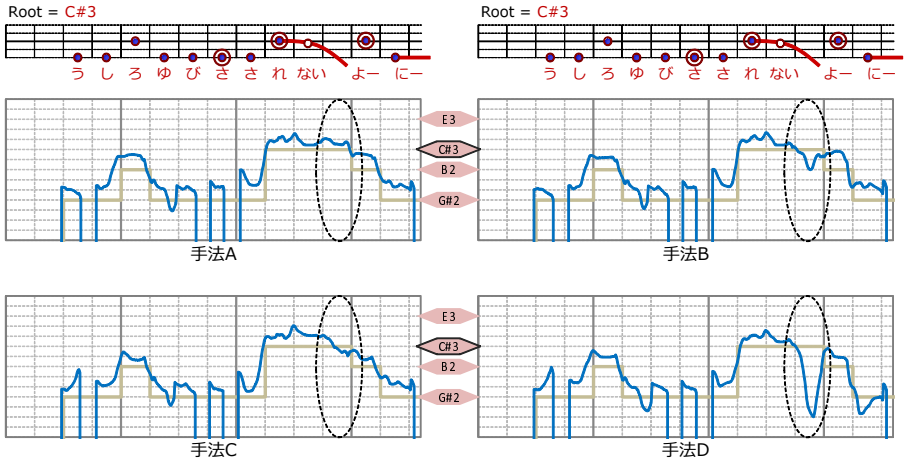

図 6 各条件下での合成音声の対数基本周波数系列 (BPM 100)
Fig. 6 Generated log F_0 contour on each experimental condition (BPM 100).

度の長さを持つ 19 のフレーズに切り分けた．それらのフレーズを被験者に提示し，実際の収録音声の分析合成音をリファレンスとして，それぞれの手法による合成音がどの程度リファレンスに近いと感じられるかを 1（最低）から 5（最高）の 5 段階の評価値 (Degradation Mean Opinion Score; DMOS) により評価する方式で受聴試験を行った．被験者は日本語話者の成人男性 10 名である． 図 5 に主観評価の結果を，図 6 に各手法においてモデルから生成された対数基本周波数系列の例を該当部分の合成用の譜面とともに，それぞれ示す．手法 A〜D の合成音声の対数基本周波数を示す図 6 より破線で囲む部分のように，部分的には各手法において聴感に影響する程度の違いはあるものの，図 5 より手法間における収録音声からの聴感上の違いに有意な差は確認されなかった．

4.4 考　　察

本章における音声合成実験で得られた音声は，ラップスタイルに独特なグリッサンドの技法を再現したものとなっていることが確認された．その一方で，中にはグリッサンド記号を付与していながら対数基本周波数の下降や上昇が発生しないなど入力の期待にそぐわない箇所もまたいくつか見受けられた．その原因としては，例えばコンテキストクラスタリングによる状態共有の結果，グリッサンド記号とは別の要因が音高の下降や上昇の現象の有無を左右する要因となっていた可能性などが考えられる．例えば図 4 の例においてグリッサンド技法による音高の下降現象がよく再現されている，破線で囲まれた部分に対応する音符は歌詞のモーラが持つ母音が/a/であるのに対し，グリッサンド記号が指定されていながらうまく再現されていない一番最後の音符では歌詞のモーラの持つ母音が/o/である．仮に学習の過程で音高の下降現象を表現するような基本周波数のモデルパラメータの推定が「グリッサンド記号が与えられている」という条件より「/a/という音素に対応するモデルである」という条件により強く関連づいて行われていたとすれば，このようなことが発生することが考えられる．そのような場合は学習データの中に偏りがあった可能性が考えられ，もしそうであれば偏りが少なくなるようにデータ量を増加させることで問題の改善が期待される．また，もしこの仮定が正しいとすれば，主観評価実験結果の原因も，今回の学習ではグリッサンドのコンテキストが音高の下降現象を表現するためのパラメータを持つモデルにうまく割り当たらなかった可能性が考えられる．

5. 素片接続型歌声合成器へのパラメータの適用

実験的な試みとして，前節までの HMM による歌声合成で得られたパラメータを用いて，素片接続型の歌声合成システムである VOCALOID[3] による合成を行った．ここでは 4 章

Vol.2012-MUS-94 No.7
Vol.2012-SLP-90 No.7
2012/2/3

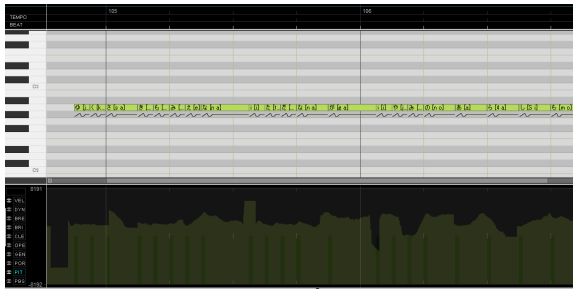

図 7 HMM から出力された音高パラメータを VOCALOID シーケンスとして与えた例
Fig. 7 An example of VOCALOID pitch parameters converted from the parameters generated from the HMMs.

の合成実験で得られた対数基本周波数系列を VOCALOID 用ピッチベンドパラメータに変換し，VOCALOID のシーケンスに与える．VOCALOID エディタ上でそのシーケンスを表示した際の様子を図 7 に示す． ユーザはまず本稿で定義する直観的なラップ用譜面の入力によりこのようなパラメータを得た上で，細かくパラメータを調整したい箇所に対しては任意の修正を施すことができる．また素片接続型の合成器のメリットとして，クリアな合成音声を得ることができる．なお，素片接続型合成器に対して音高を絶対的な物理量として指定することは可能であるが，音量や音素継続長は HMM で学習したモデルと VOCALOID 用の素片の間の基準に不一致があるために，HMM から得られた値を直接適用するだけでは自然な結果が得づらい．こういった問題の解決のための手段としては，リファレンス音声に似せた音声を合成するためのパラメータを反復的に推定する VocaListener[9] のような手法を用いることが考えられる．

6. ま と め

本稿では，ラップスタイルという短時間のうちに音高などの特徴が大きく変動するスタイルのための記譜法を定義した．そしてその記譜法にもとづく譜面とそれに対応するラップスタイル歌声データベースを構築し，そのデータを用いて，HMM 歌声合成システムの枠組みを利用してラップスタイルの歌声合成実験を行った．合成音声において，ラップスタイルの歌唱に独特の音高を急激に変化させるグリッサンド技法による現象が再現されていることが確認された．一方で，譜面における歌唱表現のための付加記号として用意したアクセントとグリッサンドの 2 種類の記号で定義したコンテキスト要因の有効性を確認するための主観評価実験では，それぞれのコンテキスト要因の有効・無効の手法間で，合成音声の自然音声からの聴感上の違いに有意な差が認められなかった．この原因としては学習データに偏りがあった可能性が考えられる．また，実験的な試みとして素片接続型の歌声合成システムである VOCALOID に HMM 歌声合成システムから出力された音高パラメータを与えて合成を行った．

今後の課題としては，偏りの少ない学習データの準備や状態共有決定木の適切なサイズ設定の検討が挙げられる．また本稿で設けた多くの条件を撤廃しさらなる幅広いラップスタイルの表現をするための検討が重要な課題となる．

7. 謝 辞

本研究のために，プロアーティストの視点からラップ歌唱スタイルについてのご意見，建設的なご提案から数多くの貴重なデータまでをご提供くださいました motsu 氏，ならびにその機会を与えてくださいましたエイベックス・マネジメント社様に心より感謝申し上げます．

参 考 文 献

1) H. Kenmochi, “VOCALOID and Hatsune Miku phenomenon in Japan,” Proc.InterSinging 2010, pp.1–4, 2010.
2) K.Oura, A.Mase, T.Yamada, S.Muto, Y.Nankaku, and K.Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System - Sinsy,” Proc.SSW7, pp.211–216, 2010.
3) H.Kenmochi and H.Ohshita, “VOCALOID-Commercial Singing Synthesizer Based on Sample Concatenation,” Proc.INTERSPEECH 2007, pp.4011–4010, 2007.
4) 午前三時のラップ講座 [DVD BOOK], ドレミ楽譜出版社 (2005).
5) M.O.V.E Official Website, http://electropica.com/index.html.
6) 間瀬絢美, 大浦圭一郎, 南角吉彦, 徳田恵一, “音高正規化学習を用いた HMM 歌声合成の検討,” 日本音響学会秋季講論集, vol.I, 1-8-20, pp.283–284, 2010.
7) A.Camacho, “SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music,” Ph.D.Thesis, University of Florida, 2007.
8) H.Zen, T.Masuko, K.Tokuda, T.Kobayashi, and T.Kitamura, “A Hidden Semi-Markov Model-Based Speech Synthesis System,” Proc.IEICE Trans., vol.90-D, no.5, pp.825–834, 2007.
9) T.Nakano, and M.Goto, “VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation,” Proc.SMC 2009. pp.343–348, 2009.