社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (9)
2008－SLP－70 (9)
2008／2／8

歌声合成システム VOCALOID—現状と課題

剣持秀紀・大下隼人
ヤマハ株式会社 サウンドテクノロジー開発センター
hideki_kenmochi@gmx.yamaha.com, hayato_ohshita@gmx.yamaha.com

"VOCALOID"とは、ヤマハが開発した素片連結型の歌声合成技術およびその応用商品の総称である。2007年8月末に発売されたその応用商品「初音ミク」(クリプトン・フューチャー・メディア株式会社)は、音楽制作用のソフトウェアとしては異例の販売本数を記録している。本稿では、VOCALOIDの基本構成、合成アルゴリズムを紹介し、今後の課題と展開について論じる。

Singing synthesis system "VOCALOID" Current situation and todo lists

Hideki Kenmochi, Hayato Ohshita
Center for Advanced Sound Technologies, Yamaha Corporation

"VOCALOID" is a concatenative singing synthesis technology developed by Yamaha Corporation, and also a trademark for its application products. Its application software "Hatsune Miku" released in the end of August 2007 by Crypton Future Media Inc., recorded an extraordinary number of sales as software for musical creation domain. In this paper, we would like to introduce its overview, its synthesis algorithm, and discuss future tasks and prospects.

1. はじめに

"VOCALOID"とはヤマハが開発した素片連結型の歌声合成技術およびその応用商品の総称である。現在のところ、商品はヤマハからではなく、ヤマハとライセンス契約を結んだ各社が独自に制作した歌手ライブラリに、ヤマハが開発したソフトウェアが同梱される形で、ライセンス供与先の製品として発売されている。

最初のバージョンを搭載した製品は2004年から発売されており、5タイトル(英語3タイトル、日本語2タイトル)が発売された。2007年には改良版である Vocaloid2 を発表した。Vocaloid2の応用商品である「初音ミク」(クリプトン・フューチャー・メディア株式会社)は、8月末の発売からわずか3ヶ月で25,000本以上の売上を記録し、音楽制作ソフトウェアとしては異例のヒットとなっている。

また、ユーザが作成したオリジナル曲が画像共有サイトに投稿され、そこでのヒット曲がカラオケや着うた®[1]に配信されたり、「初音ミク」のキャラクターグッズが発売されるなど、通常の音楽ソフトウェアの枠を遥かに超えた現象が起きている。動画共有サイトでは、「初音ミク」を使って制作されたオリジナル楽曲を別のユーザが自分の声で歌ったもの(すなわち、

[1] 「着うた」は株式会社ソニー・ミュージックエンタテインメント (SME) の登録商標である。

合成音声のための作品を人間がカバーした作品)が投稿されるなど、非常に興味深い状況となっている。さらに、2007 年末には第 2 弾の「鏡音リン・レン」が発売され、こちらも好評を得ている。

© Crypton Future Media Inc.

図 1. Vocaloid2 「初音ミク」

このように、Vocaloid は当初の予想を超えて幅広く受け入れられている状況ではあるが、もちろんまだ完全に人間の代わりになるようなものではない。また、当初想定していたプロフェッショナルな(あるいはプロフェッショナルに近い)音楽制作シーンでは、今のところまだ広く利用されるまでには至っていない。本稿では、Vocaloid のシステム構成の概略と合成方式について説明を行い、現在のシステムの課題について議論し、今後の展開について述べる。

2. Vocaloid システム構成

Vocaloid の基本構成を図 2 に示す。

ユーザは(a)スコアエディタを用いて、歌詞と音符、および歌声に必要な表情を入力する。この情報は(c)合成エンジンに送られ、合成エンジンは(b)歌手ライブラリを参照しながら歌声を合成し、出力する。

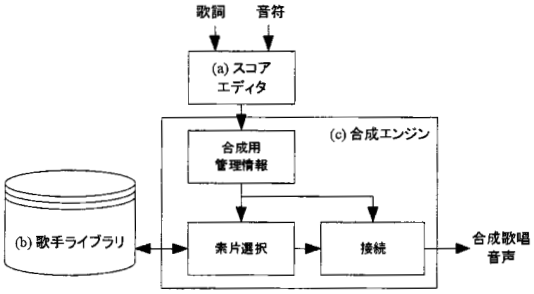

図 2 Vocaloid 基本構成

以下、それぞれの構成要素について詳しく述べる。

(a)スコアエディタ

スコアエディタのスクリーンショットを図 3 に示す。

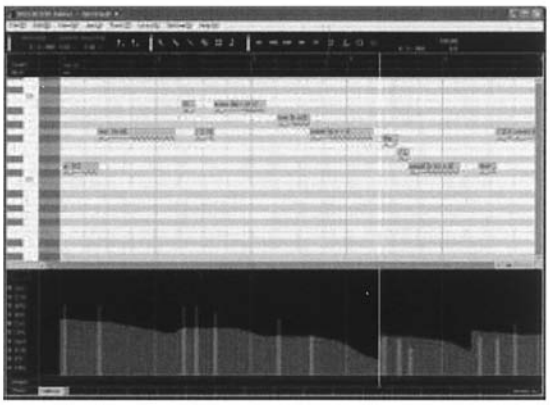

図 3 スコアエディタ

音符はピアノロールで入力する。歌詞は音符をダブルクリックすることで入力可能となる。日本語の場合は仮名あるいはローマ字で、英語の場合は単語そのものを入力する。歌詞は内部の発音辞書により、音声記号に自動的に発音記号に変換される。発音記号を直接編集することも可能である。英語の場合は、複数の音節により構成される単語があるが、この場合は内部の発音辞書に含まれる音節区切りに従い、自動的に複数の音節に分割され、複数の音符に割り当てられる。

アタックやビブラートなどの表情は、音符近

くに表示されるアタックやビブラートのアイコンをマウスで直接操作するか、ダブルクリックして表示されるウィンドウで細かく指定することで与えることができる。スコアエディタの下の部分はコントロールトラックとなっており、合成パラメータを直接間接に操作することができるようになっている。

スコアエディタによる音符・歌詞の入力だけでなく、あらかじめ入力しておいた歌詞をもとにMIDIキーボードで演奏する機能も実装されている。これにより、キーボードで「歌う」という新しい音楽の演奏スタイルが可能となっている。

スコアエディタで入力された情報は、合成時に専用のMIDIメッセージ(VOCALOID MIDI)に変換されて合成エンジンに送られる。一体化されたシステムの場合には、合成エンジンとのやり取りは必ずしもMIDIである必要はないが、合成エンジンを独立して扱うことの将来的な利便性を考慮し、内部的にMIDIの形式でデータをやり取りするようになっている。

合成エンジンに送るMIDIメッセージとして、Note On/Offなどの通常のMIDIメッセージは使用できない。歌声の場合、音符が指定するタイミングで音節に含まれる母音が発音され、子音はそれよりも前に発声されるからである。そこで、VOCALOID MIDIでは合成に必要な全ての情報 (Note On/Off に相当する情報さえも) 事前にディレイ情報付でNRPN(Non Registered Parameter Number)のフォーマットで送っている。すなわち、「今から D[ms]後に、ノート番号 n、duration が d[ms]で歌詞が L であるような音符を鳴らしなさい。」という内容を合成エンジンに送っている。

Vocaloid MIDIの詳しい仕様については各製品に付属するユーザーマニュアルを参照されたい。

Vocaloid2 の各製品には、VOCALOID MIDIを合成エンジンに直接送ることができる VSTプラグイン(Vocaloid Playback VST Instrument)も同梱されているので、VOCALOID MIDIを出力する VST ホストを作成すれば、スコアエディタが無くても Vocaloid 合成エンジンによる合成が可能となる。

(b) 歌手ライブラリ

歌手ライブラリは、実際の歌手の歌唱データを元にした音声素片を含むデータベースである。VOCALOID では音声素片の単位として、diphone と伸ばし音を使用している。歌声ライブラリには、対象とする言語で起こりうる全てのC-V、V-Cの組み合わせおよび全ての母音の伸ばし音が含まれている。

音声素片の元になるデータは複数のピッチで収録される。できるだけ多くのピッチで収録した方が合成音のクオリティが向上すると期待されるが、長時間歌い続けることによる声質の変化や歌手の疲れ(身体面および精神面)、歌手へのギャラなどを考慮し、ある程度のところでの妥協が必要となる。収録したデータは半自動で処理され、人手によるチェック、修正を経て歌声ライブラリが完成する。

(c) 合成エンジン

合成エンジンは、歌詞、音符、表情その他のパラメータに従い、必要な音声素片を歌声データベースから取り出し、連結する。

前述のように、歌声を合成する場合、音符の開始位置(Note On)は、音符を構成する音節の最初の位置ではなく、音節に含まれる母音の位置でなければならない。そこで、Vocaloid では内部に「合成スコア」を持ち、音節の母音の開始部分が音符開始のタイミングに合うように素片の再生位置を調整し、タイミングがちょうど良く聞こえるようにする。

合成スコアには、各時刻でのピッチや各種合

成パラメータの変化も描かれ、合成時に参照される。ピッチに関しては、指定された音符とアタック、ビブラートのパラメータをもとに、ピッチのカーブが内部的に計算され、合成スコアに格納される。素片選択時には、これらのパラメータをもとに、最も適した素片が選択される。

図 4 に“Sing a song”([sIN a sO:N])という歌詞で歌う場合のタイミング調整とピッチカーブの例を示す。

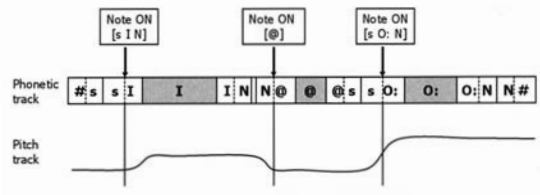

図 4　素片タイミング調整とピッチカーブ

素片を接続する時に、素片のピッチを所望のピッチに変換する必要があるが、たとえピッチを合わせたとしても、単純に接続しただけでは素片間の音色の違いから不自然な合成音やノイズが発生する。これを防ぐために、伸ばし音区間で、隣り合う diphone のスペクトル包絡を補間することで伸ばし音のスペクトル包絡とする。図 5 に“sing”([sIN])という歌詞で合成する場合の例を示す。“sing”([sIN])という歌詞の音符の伸ばし音部分のスペクトル包絡は、[s-I]の最終フレームと[I－N]の最初のフレームのスペクトル包絡を時間的に補間することで求められる。これにより原理的に接続部で音色の突然の変化が発生しないようになっている。diphone 区間は、素片のスペクトル包絡をそのまま使用する。位相に関しても接続部で突然の変化が起こらないように調整される。

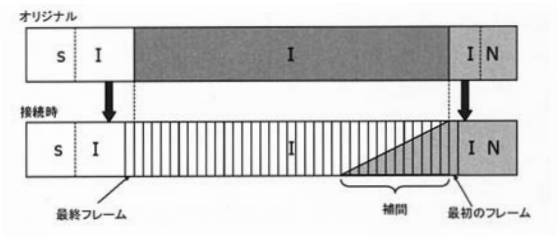

図 5 スペクトル包絡の補間

スペクトル包絡の補間を効率的に行い、かつ合成音の音色をユーザがある程度コントロールできるようにするため、式(1)で表される曲線に、中心周波数 F_i，バンド幅 Bw_i，強度 Amp_i の二次のフィルタがいくつか加算された形でスペクトル包絡を表現する。(このままでは、実際のスペクトル包絡と異なることになるので、誤差分を最終的に加える。)

$$E_s(f) = Gain + SlopeDepth(e^{Slope \cdot f} - 1) \quad (1)$$

式(1)を図示すると図 6 のようになる。

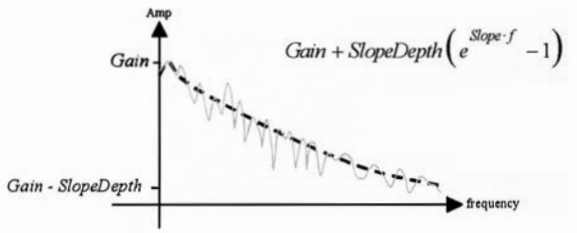

図 6 スペクトル包絡の表現

スペクトル包絡の補間は、式(1)の各パラメータおよびフィルタの係数(F_i, Bw_i, Amp_i)を補間することで行う。

また、これらを変更することで、音色に変化を持たせることが可能になる。例えば、式(1)の *Slope* を変化させることで張りのある声や穏やかな声に変化させることが可能である。

以上のようにして、所望のピッチ、スペクトル包絡が定まったので、選択された素片をそれらに合うように変換する。変換部分の信号処理のブロックダイヤグラムを図 7 に示す。

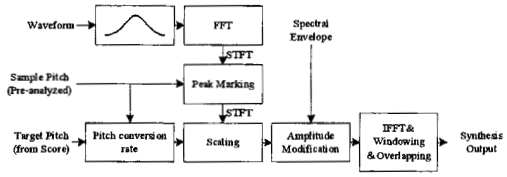

図 7 合成エンジンの信号処理

ピッチ変換は図 8 に示すように、スペクトルを周波数軸上でスケーリングすることで行われる。倍音に相当するピークの近傍のスペクトルの形状はできるだけ元のものを保つように、非線形にスケーリングが行われる。ピッチ変換時には、倍音の周波数が完全に整数倍になっていると仮定し、i 番目の倍音に相当する周波数の位相に対して、以下の式で補償が行われる。

$$\Delta\varphi_i = 2\pi f_0(i+1)(T-1)\Delta t \quad (2)$$

ただし、T は f_{0T} (所望のピッチ)と f_0 (素片のオリジナルのピッチ)の比であり，Δt はフレーム長である。

所望のスペクトル包絡に合うようにピークの強度が調整される。

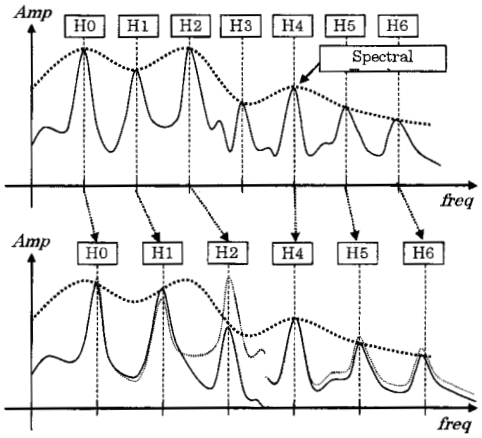

図 8 ピッチ変換と音色の調整

周波数領域でのピッチ変換と音色の調整の後、IFFT と Windowing & Overlapping を行うことで合成音声が得られる。

3. 今後の課題

Vocaloid2「初音ミク」のヒットの要因としては、声質とイメージキャラクタがマッチしたこと、動画共有サイト等、作品の発表の場が存在したこと、適度に漠然としたイメージのみを提供し、ユーザがキャラクタに共感できる余地を与えたこと等、本稿の限られたページ数では書きつくせないほど様々な要因が考えられるが、少なくとも声自体が明確なキャラクタ性を持ち、「このツールだけにしかこの声は出せない」ということが商品として大きな魅力になったであろうということは想像に難くない。

Vocaloid をリリースして以来、「何時間もかけて合成音声を作りこむくらいなら、歌手を呼んできたほうが安いし早い」と良く言われたが、「初音ミク」やそれに続くシリーズのように、声に明確なキャラクタ性があり、「このツールだけにしかこの声は出せない」ということになれば、その論理に勝てる(可能性がある)ということがわかった。

今後ともライセンス先各社と協力し、歌手ライブラリを充実させ、ユーザの選択肢を広げていきたいと考えている。海外でも Vocaloid2 のタイトルとして、PowerFX Systems AB(スウェーデン)が "Sweet Ann"を、Zero-G Limited(イギリス)が "Prima"を発売しており、徐々に広がりを見せている。

一方で、今後、音楽制作のシーンで引き続き使っていただき、かつプロフェッショナルな音楽制作現場にも定番的に浸透するためには、奇をてらわずに、品質や表現力の向上を続けることが最も重要であると考えている。特に、いわゆる「シャウト」や「だみ声」は音楽表現上も非常に重要な要素であるので、それらを自然に再現できるようにしていきたいと考えている。

また、ソロのボーカルではなく、[7]で筆者らが提案したコーラスの音声素片をもとにコ

ーラスを合成するシステムは、プロフェッショナルな音楽制作現場に適していると考えている。今後はこちらの方面でも改良を続けたい。

4. 謝辞

VOCALOID の信号処理部分は、ヤマハと Pompeu Fabra 大学(バルセロナ)の Music Technology Group (MTG) との共同研究によって開発された。MTG のスタッフ、特に Jordi Bonada 氏、Alex Loscos 氏に感謝します。

5. 参考文献

[1] Bonada, Loscos, Kenmochi, "Sample-based Singing-voice Synthesizer by Spectral Concatenation", Proc. of SMAC 03, 439-442, 2003.
[2] Bonada et al., "Spectral Approach to the Modeling of the Singing Voice", Proc. of the 11th AES Convention, 2001.
[3] Bonada et al., "Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models", Proc. of ICMC, 2001.
[4] http://www.zero-g.co.uk/
[5] http://www.crypton.co.jp/
[6] http://www.powerfx.com
[7] 剣持, 大下, Bonada, Loscos, "コーラス音声の合成" 日本音響学会講演論文集 1-Q-23 (2006-4)