Vol.2015-MUS-107 No.59
2015/5/24

歌声合成ソフトウェアVOCALOID^{TM}4における表現力向上への取り組み

橘　誠[1,a)]

概要：歌声合成ソフトウェア VOCALOID^{TM} は 2014 年 12 月におよそ 3 年ぶりのメジャーバージョンアップを行った VOCALOID^{TM}4 が発売された．本稿では，この VOCALOID^{TM}4 において歌声の表現力向上を目的に新たに搭載した，グロウル，クロスシンセシスの機能について，その概要を紹介する．グロウルはスペクトルモーフィングに基づく手法により，少量のグロウルの声質を含んだサンプルを予め用意することで通常発声の歌声に所望の度合いでグロウルの声質を付与することが可能である．またクロスシンセシスでは，合成時にこれまで行っていた音色の補間処理を，異なる歌声ライブラリ間に対しても行えるよう拡張することで，2 つの異なる歌声ライブラリから新たな音色を合成することができるようになった．さらにユーザインタフェースの改良点として，合成されるピッチカーブを可視化するピッチレンダリング機能およびその周辺の機能強化についても述べる．

キーワード：歌唱合成，VOCALOID^{TM}，グロウル，クロスシンセシス，ピッチレンダリング

VOCALOID^{TM}4 : new features for expressive singing voice synthesis

Abstract:
VOCALOID^{TM}4 was released as a major version update of singing synthesis software VOCALOID^{TM}in Dec. 2014. This paper introduces its new functions such as “Growl” and “Cross-Synthesis” for expressive singing voice synthesis. The “Growl” function is based on a spectral morphing technique. By using a small amount of recorded “Growl” samples, an arbitrary amount of “Growl” voice quality can be added to a modal singing voice. The “Cross-Synthesis” technique is an extension which performs timbre interpolation between different singer libraries, allowing synthesis of intermediate voice timbres between them. Furthermore, the “Pitch Rendering” function and its related features are presented as examples of UI improvements.

Keywords: Singing voice synthesis, Growl, Cross-Synthesis, Pitch rendering

1. はじめに

VOCALOID^{TM} はヤマハが開発した歌声合成システムで，2004 年に最初のバージョンを発売し，2007 年に VOCALOID^{TM}2 ， 2011 年に VOCALOID^{TM}3 をそれぞれ発売している．そして 2014 年 12 月におよそ 3 年ぶりとなるメジャーバージョンアップを行い VOCALOID^{TM}4 を発売した [1].

[1] ヤマハ株式会社 事業開発部 VOCALOID プロジェクト
VOCALOID Project, Business Development Division, Yamaha Corporation
[a)] makoto.tachibana@music.yamaha.com

VOCALOID^{TM}4 における主な新機能としては，以下の点が挙げられる．

グロウル　声を激しくふるわせ，うなるような効果を得ることができ，ブルースやロック，スクリーモなど幅広いジャンルで表現力を向上できる．

クロスシンセシス　2 つの異なる歌声ライブラリの音色をブレンドし，新たなオリジナルの音色を作ることができる．

ピッチレンダリング　ピッチの変化やビブラートの掛かり具合をピアノロール上に描画し，視認することができる．

ピッチスナップモード　合成エンジンが自動的に作り出し

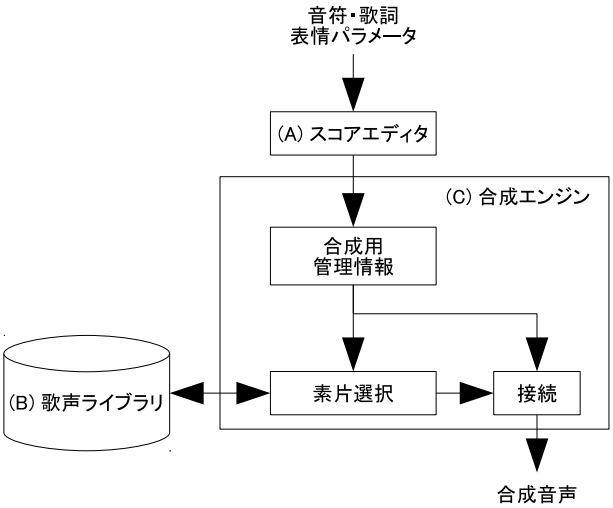

図 1 VOCALOID^{TM} システムの概要
Fig. 1 Outline of VOCALOID^{TM}system.

てきた自然なピッチカーブをオフにすることで，ロボットボイスのような歌い方を簡単に生成できる．

リアルタイムレコーディング MIDI キーボードを演奏しながらノートを入力することができる．

本稿では，これらの特徴のうち歌声の表現力向上に大きく関係しているグロウル，クロスシンセシスについて，その技術概要を紹介する．またユーザインタフェースの改良点として導入したピッチレンダリングおよびその周辺の機能強化についても述べる．

2. VOCALOID^{TM} の概要

2.1 VOCALOID^{TM} システムの概要

まず，簡単に VOCALOID^{TM} システム [2] の概要について紹介する．VOCALOID^{TM} は，主に (A) スコアエディタ，(B) 歌声ライブラリ，(C) 合成エンジンから構成される（図 1）．

(A) のスコアエディタは歌詞と音符などの楽譜情報を入力するためのインタフェースである．歌詞と音符以外にもビブラートの表現や，音符のアタックや前の音符からのポルタメントの掛かり具合，子音の長さなどの音符に付随する情報，さらにはピッチや音量，声の太さや明るさなどの時間変化を操作可能である．

（B）の歌声ライブラリは，声の提供者（歌手や声優）の実際の歌唱データから取り出した音素素片を集めたもので，素片はダイフォンを基本単位とした C-V（子音-母音），V-C（母音-子音）に加えて，V（母音）の伸ばし音を使用している．また，歌声の声域による音色の変化に対応するため，同一の音韻の組合せに対して，複数の異なるピッチで発声されたサンプルを持つことができる．なお，電子楽器のサンプリング音源とは異なり，声の提供者への肉体的，精神的負担を考慮すると，声の提供者の発声可能な音域を半音単位で全て収録するといったことは現実的ではなく，代表的な数個のピッチを収録している．

(C) の合成エンジンでは，入力されたスコアに従って歌声ライブラリから必要な素片を取り出し，発音タイミングを調整した上で，周波数領域でピッチ変換と接続時の音色の合わせこみを行いながら素片同士を接続し，合成音を出力する．

2.2 VOCALOID^{TM}3 での改良点

ここで，2011 年に発表した VOCALOID^{TM}3 [3] での改良点において，本稿と関連している内容について簡単に述べる．

まず (B) 歌声ライブラリにおいて，V-C-V（母音-子音-母音）や C-C-V（子音-子音-母音）などのダイフォンより大きな素片単位のサンプルを持てるようになった．これにより，ダイフォン同士を接続する場合に比べて素片の接続で生じる歪を低減したり，先行・後続の母音に依存した子音の変化を再現できるといった効果が見込まれる．なお，この大きな素片単位のサンプルを歌声ライブラリに収録することは必須ではなく，オプションとして適切な効果が得られたものを任意に追加可能である．そのため歌声ライブラリによって収録されている素片の種類や個数が異なっていることに留意する必要がある．

また (C) 合成エンジンで素片間の音色調整を行う際に，ピッチの異なる複数の素片サンプル間で音色を補間する機能が導入された．これまでは，歌声ライブラリに同一の音素の組合せでピッチの異なる複数のサンプルが収録されていた場合に，合成するピッチによっては突然選択されるサンプルが切り替わり，不自然な印象を与える場合があったが，この音色の補間機能の導入により，合成するピッチに従って素片サンプル間の補間比率が調整されることで，滑らかな音色変化を得ることが可能になった．

3. グロウル

近年，グロウル系統のノンモーダルな発声の歌声の分析，合成に関する検討が盛んに行われている [4–8]．中でも文献 [7,8] で提案されたスペクトルモーフィングによる手法は VOCALOID^{TM} の枠組みとの親和性が高く，今回 VOCALOID^{TM}4 で新たに導入することとした．この手法の利点として以下の点が挙げられる．

(1) 実際のグロウル音声のサンプルを利用することで，モデルベースの手法に比べ，簡単に高品質な合成音が得られる．

(2) 歌声ライブラリの収録とは別に用意したノンパラレルな少量（〜数秒程度）のサンプルを用いて実現可能．また合成音の音色（スペクトル包絡）は，歌声ライブラリのサンプルに合わせ込まれるため，歌声ライブラリの声の提供者とグロウルサンプルの提供者は必ずし

も同一である必要はない．

(3) グロウル成分の付与の度合いを任意にコントロール可能．

図 2 に VOCALOID^{TM}4 におけるグロウル処理の流れを示す．通常の歌声ライブラリに含まれる素片サンプルは FFT 後にピッチ変換・音色調整が行われる．一方，グロウルのサンプルは，まず時間領域で目標の合成ピッチに合うようにリサンプリングを行う．これによりサブハーモニクスを含むスペクトルの構造を一様に伸縮させることができる．しかしながらリサンプリング処理した波形は，サブハーモニクス以外のノイズ成分もシフトされてしまい，そのまま使用すると不自然な合成結果を生じてしまう．そこで，リサンプリングされた波形を FFT し，調波のインデックスを元の調波の位置に近づけるようマッピングを行う．そして歌声ライブラリの素片サンプルと振幅・位相を合わせ込むようフィルタリングを行い，所望の分量をスペクトルモーフィングにより付与する．

グロウルサンプルの保持方法として，例えば以下のような形態が考えられる．

- システム全体で共通のグロウルサンプルを持つ（グローバル）
- 歌声ライブラリ毎に固有のグロウルサンプルを持つ（ライブラリ依存）
- 使用される歌詞・音符等に応じて使い分ける（コンテキスト依存）

VOCALOID^{TM}4 では歌声ライブラリ毎に使用するグロウルサンプルを使い分けるライブラリ依存の方式を採った．これにより，歌声ライブラリ毎に特有の声質でグロウルの効果を生じさせることが可能である．

また，1 つの歌声ライブラリに対して，ピッチの異なる複数のグロウルサンプルを持てることとし，合成する音域によって自動的に使い分けが行われる．なお，複数のグロウルサンプルを持つ場合に，グロウルサンプル同士でサンプルを補間するような処理は行っていない．

グロウルを付加する度合いは，新たに時系列パラメータとして導入した “GWL” によって調整することが可能である．図 3 にその様子を示す．この図では音符の終端に向かって次第にグロウルが付加される度合いが大きくなっている．

4. クロスシンセシス

クロスシンセシスは，異なる歌声ライブラリ間で音色を補間する手法で，2013 年 9 月の SIGMUS 第 100 回記念シンポジウムの講演 [9] において，初めてその概要が紹介された．この背景として，これまでにも VOCALOID^{TM} の歌声ライブラリの中には，“sweet,” “dark,” “power” といった表情のバリエーションを持たせて収録した複数の歌声ライブラリが多数提供されていることが挙げられる．しかし

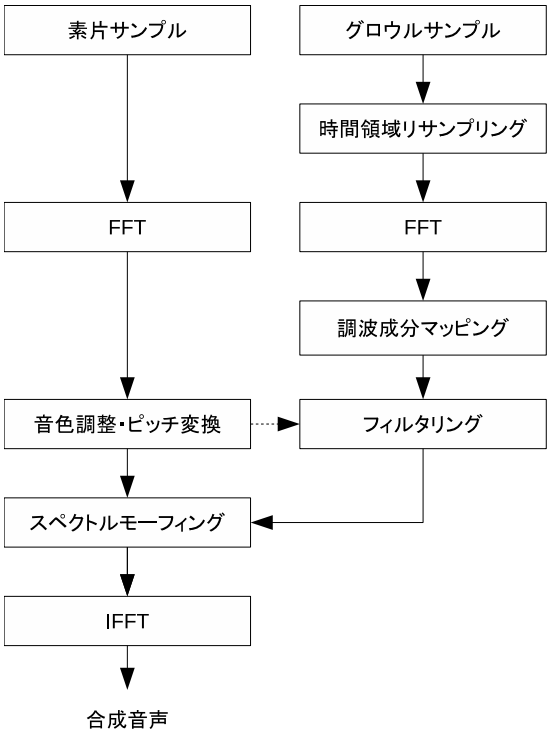

図 2 グロウル処理の流れ
Fig. 2 Flow of synthesizing Growl voice.

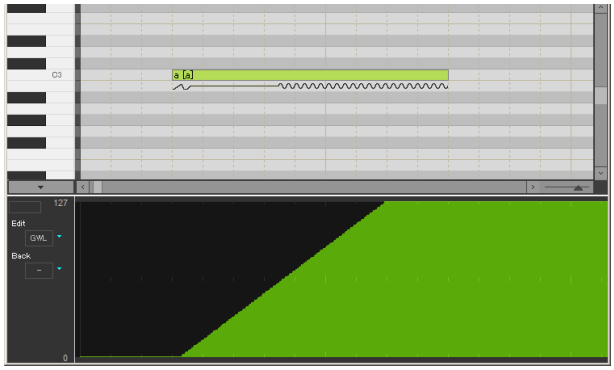

図 3 グロウルパラメータ “GWL”
Fig. 3 New control parameter “GWL.”

ながら，歌声ライブラリによるバリエーション表現では，音符と音符の切れ目などで使用する歌声ライブリをスイッチする必要があり，1 つのノートを発音中に徐々に別の表情の音色に変化させるといったことができなかった．

そこで，2.2 節で示したピッチの異なる素片間の音色を補間する手法を，異なるライブラリ間に拡張することを行った．図 4 にクロスシンセシスの概要を示す．まず歌声ライブラリ 1 において，通常の合成と同様に使用する素片サンプルを選択する．次に別の歌声ライブラリにおいて，選択された素片サンプルに類似したサンプルを選択する．この時の条件としては，素片サンプルの音素の組み合わせが一致すること，また合成するピッチに一番近いサンプルを選択することとした．

なお VOCALOID^{TM} では長い音符を合成する際に，V

（母音）の伸ばし音のサンプルが使用されるが，この V のサンプルはスペクトルの微細構造の変動とピッチの微妙な揺らぎを再現するために使用され，ベースとなるのは V に先行する C-V サンプルの最終フレームの音色となる．そのため従来は C-V の最終フレームを必要な長さ分コピーして使用していたが，補間比率に応じて動的に最終フレームの音色調整も行っている．

なお，V-C-V や C-C-V 等のダイフォンを超える単位のサンプルについては，歌声ライブラリへの登録が必須ではないため，歌声ライブラリによっては同一の音素の組合せを持つサンプルが見つからない場合が想定される．このような場合，V-C-V や C-C-V が見つからなければ，合成に使用するフレーム位置を考慮して V-C，C-V などの細分化した素片サンプルに対しても探索を行う．また，処理量削減のため歌声ライブラリ 2 においては複数の素片同士の音色の補間は行っていない．歌声ライブラリ 1，2 において使用される素片が決定したら，ユーザが指定した所望の割合で音色を補間する．この処理は，VOCALOID^{TM} ではスペクトル包絡を Excitation plus Resonance(EpR) パラメータ [10] で表現しているため，2 つのサンプル間で対応するパラメータ同士を補間することで実現する．

このように，クロスシンセシスではサンプル同士のモーフィングを行うのではなく，単にスペクトル包絡の補間のみを行っている．そのため，後段処理であるピッチ変換に使用されるサンプルは，歌声ライブラリ 1 のものが使用され，スペクトルの微細構造などは補間の比率を変えても変化しない．例えば，歌声ライブラリ 1 が張りのある歌い方，ライブラリ 2 が息成分の多いささやくような歌い方であるといったように声質が大きく異なる場合に，ライブラリ 1，2 の割合をそれぞれ 0%，100% とした時と，ライブラリ 1，2 を入れ替えて 100%，0% とした場合に出力される歌声の声質も異なってしまう．そこで，歌声ライブラリ 1 を**プライマリシンガー**，歌声ライブラリ 2 を**セカンダリシンガー**と明示的に区別している．

音色の補間比率の調整は時系列パラメータ “XSY” によって調整する．図 5 にその一例を示す．図のコントロールパラメータ描画領域の下段に表示されている “VY1V4_ Normal” がプライマリシンガー，上段の “VY1V4_ Power” がセカンダリシンガーのライブラリ名となっており，1 つの音符内で “Normal” → “Power” → “Normal” へと次第に音色が変化する．

5. ピッチレンダリング

5.1 ピッチレンダリング機能

VOCALOID^{TM} では，ピッチベンドやダイナミクスのような時系列パラメータに加えて，ノート単位でピッチの立ち上がり方（ベンドの深さ，長さ）やビブラートの種類などを変更することができる．しかしながら，これらのパラメータは UI 上からは直接確認することができなかったり，抽象化された表現で描画されていたため，パラメータを調整するたびに合成音を聴きながら実際のピッチの変化を確認する必要があった．そこで VOCALOID^{TM}4 ではピアノロール上に合成時のピッチの変化を描画するピッチレンダリング機能を搭載した．

図 6 にノート単位のパラメータの 1 つである「ベンドの

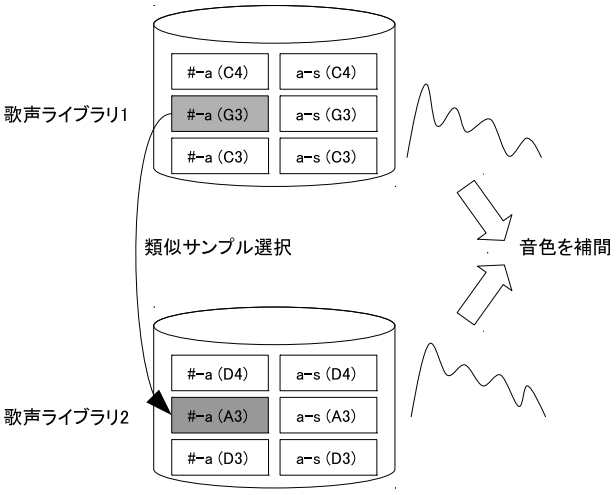

図 4 クロスシンセシスの概要
Fig. 4 Outline of Cross-Synthesis.

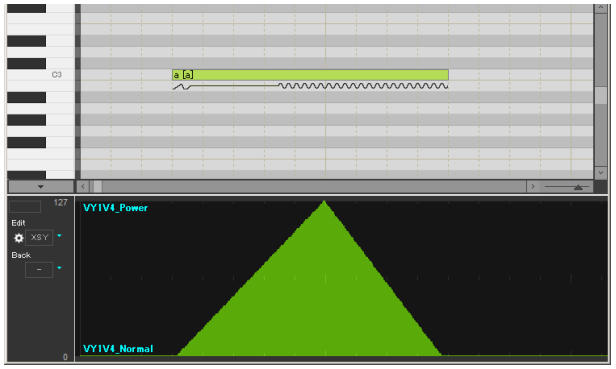

図 5 クロスシンセシスパラメータの例
Fig. 5 Example of controlling Cross-Synthesis.

図 6 「ベンドの長さ」を変化させた時のピッチレンダリング結果
Fig. 6 Pitch rendering result when changing “Bend Length.”

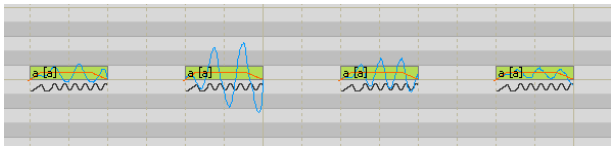

図 7 「ビブラートの種類」を変化させた時のピッチレンダリング結果
Fig. 7 Pitch rendering result when changing “Vibrato Type.”

長さ」を 0, 25, 50, 100 と変化させた時のピッチレンダリング結果を示す．パラメータの変化を視覚的にも把握できるようになっている．また，図 7 には，ビブラートの種類を “Normal,” “Extreme,” “Fast,” “Slight” にそれぞれ設定した場合のピッチレンダリング結果を示す．ビブラートが付加されたピッチ曲線は水色で描画されている．これまでの抽象化されたビブラート曲線（ノートの下の黒線）では把握することができなかった，ビブラートの種類の違いによるピッチ変動の振幅や周期が可視化されていることが確認できる．

5.2 処理の流れ

VOCALOID^{TM} の合成エンジンにおいて，ピッチ情報は以下の順に決定される．

(1) 入力された音符の音高情報に基づき，階段状のピッチ遷移を生成．
(2) 音符の付加情報（ベンドの長さ，深さ）と前後の音符の繋がり方，ポルタメントの情報をもとに，いくつかの数式で表現されるエクスプレッションモデル [11] を適用し，音符間を滑らかに繋ぐ曲線を生成．
(3) 歌声ライブラリから素片選択を行い，サンプルの長さとノートのベロシティの値に応じて子音の長さを決定．これによりノートオンに先行するピッチ曲線の描画範囲が決定される．
(4) ビブラートによる変動およびピッチベンドによる変動を付与．
(5) 伸ばし音の区間では V（母音）の伸ばし音のサンプルが持っているピッチの微細変動を付与．

このうち，ピッチレンダリング機能では (4) の状態を描画している．(3) の処理で歌声ライブラリへアクセスしており，レンダリングの結果の主にピッチ曲線の開始，終了位置などはライブラリによって変化している．またピッチレンダリング時は合成エンジン内での素片サンプルに対する信号処理は行わないため，通常の合成時に比べて高速に結果を取得することが可能である．

5.3 ピッチスナップモード

前節の (2) で適用されるエクスプレッションモデルの処理を行わないことで，ノート間の音程を階段状に繋ぐピッチ遷移を得ることができる．この状態で合成エンジンを動作させることができる機能を新たに “ピッチスナップモード” としてユーザが選択できるようにした．ピッチスナップモードがオンになっている場合には，人間らしい歌い方とは異なるロボットボイスのような表現を簡単に生成できる．図 8 にピッチスナップモードの有無による音高遷移の変化の違いをピッチレンダリング結果として示す．このように通常は異なる音高間を滑らかに遷移しているピッチがピッチスナップモードでは階段状に変化している．

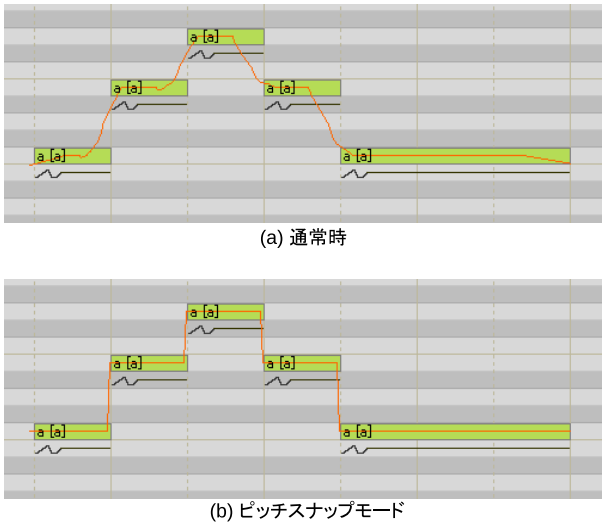

図 8 通常時とピッチスナップモードのピッチ遷移の違い
Fig. 8 Comparing pitch curve with Pitch Snap Mode ON/OFF.

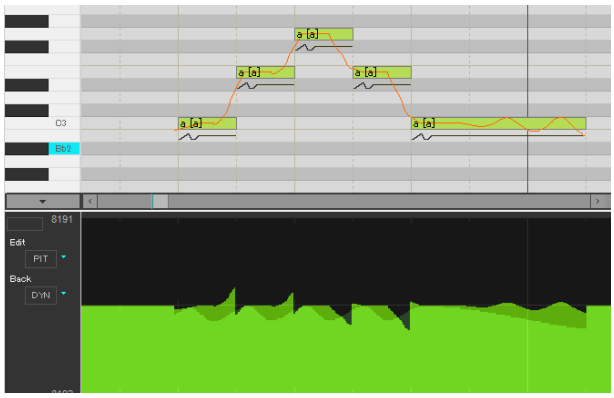

図 9 エクスプレッションモデルとビブラートのピッチ・音量変化のコントロールパラメータへの変換例
Fig. 9 Importing pitch and dynamics change of expression model and vibrato into pitch bend and dynamics parameters.

他にもピッチスナップモードではエクスプレッションモデルによるピッチ，ダイナミクスの変化を考慮する必要が無いため，ピッチ変化を Job Plugin [3] などを利用して外部から与える場合にも，これまでよりも簡単に直接ピッチの値を指定することが可能になると考えられる．またこれとは逆にエクスプレッションモデルによるピッチ，音量の変化をピッチベンドとダイナミクスのコントロールパラメータに変換する機能も搭載し，ピッチスナップモードと併用することでピッチの遷移を可視化されたパラメータとして直接編集することも可能になった．図 9 にコントロールパラメータに変換した例を示す．図 8(a) と同様のピッチの遷移が，ピッチベンドパラメータによって生成されている．

6. おわりに

本稿では，VOCALOID^{TM}4 において新たに搭載したグロウル，クロスシンセシスといった歌声合成の表現力向上への取り組みおよび，ピッチレンダリング機能について紹介した．今後は，グロウル以外の多様なノンモーダル発声への取り組みや，歌唱スタイル [12] の付与によるさらなる表現力の向上などが課題である．

参考文献

[1] http://www.vocaloid.com/: VOCALOID^{TM} 公式サイト.
[2] 剣持秀紀, 大下隼人：歌声合成システム VOCALOID, 情処学音楽情報科学研報，Vol. 2007, No. 102, 2007-MUS-072, pp. 25–28 (2007).
[3] 剣持秀紀：歌声合成ソフトウェア VOCALOID3 と VOCALOID Job Plugin，情処学音楽情報科学研報，Vol. 2012-MUS-94, No. 4, pp. 1–4 (2012).
[4] 溝渕翔平，西村竜一，入野俊夫，河原英紀：声道形状と声帯音源特性を用いたグロウル系歌唱音声への実時間変換の提案，情処学音楽情報科学研報，Vol. 2015-MUS-106, No. 12, pp. 1–6 (2015).
[5] 西脇裕展，坂野秀樹，旭　健作：スクリーム唱法による音声の高品質分析合成を可能とする音声特徴量に関する検討，2014 年春音講論集，Vol. I, 2-Q5-16, pp. 491–492 (2014).
[6] 加藤圭造，伊藤彰則：グロウル・スクリーム歌唱音声の音響的特徴と聴覚印象の考察，信学技報，Vol. 112, No. 422, SP2012-105, pp. 43–48 (2013).
[7] Bonada, J. and Blaauw, M.: Generation of growl-type voice qualities by spectral morphing, *Proc. ICASSP 2013*, pp. 6910–6914 (2013).
[8] Bonada, J.，Blaauw, M.，才野慶二郎，久湊裕司：スペクトルモーフィングによるグロウル系統の歌唱音声合成，情処学音楽情報科学研報，Vol. 2013-MUS-100, No. 24, pp. 1–6 (2013).
[9] 剣持秀紀：歌声合成技術とその未来，情処学音楽情報科学研報，　Vol. 2013-MUS-100, No. 26, p. 1 (2013).
[10] Bonada, J., Celma, Ò., Loscos, A., Ortolà, J., Serra, X., Yoshioka, Y., Kayama, H., Hisaminato, Y. and Kenmochi, H.: Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models, *Proc. of ICMC* (2001).
[11] 剣持秀紀，藤本　健：ボーカロイド技術論，chapter 3, pp. 76–78, ヤマハミュージックメディア (2014).
[12] 橘　誠，才野慶二郎，久湊裕司：HMM 音声合成技術の歌唱スタイル生成 Job Plugin への応用，信学技報，Vol. 113, No. 366, SP2013-94, pp. 123–128 (2013).