Vol.2017-MUS-115 No.17
2017/6/17

高品質音声分析合成を用いた基本周波数の実時間操作インタフェースの実装

渡邊 優介[1,a)]　森勢 将雅[1]　小澤 賢司[1]

概要：本研究では，音声分析合成系 WORLD を用いたリアルタイム音声合成によって基本周波数（F0）変換を行うインタフェースを検討している．Vocaloid のように，歌詞と譜面の情報から歌声を生成する歌声合成ソフトウェアは，近年においても進化し続け，様々な歌唱表現を歌声に付加できるようになった．一方，プリセットとして用意されていないユーザ独自の歌唱表現を付加する場合には，職人芸的な技術や煩雑な作業が必要である．本稿では，スワイプ操作による直感的な F0 のリアルタイム変換機能を有するインタフェースを設計し，iPad で動作するアプリケーションを試作した．本稿では，本インタフェースが有する機能として，リアルタイム合成機能や録音機能，合成操作の記憶機能などについて説明する．最後に，実装した機能が歌唱表現をデザインするための機能として妥当であるか否かについて考察する．

キーワード：音声分析合成，音声加工，基本周波数，実時間操作インタフェース

1. はじめに

音楽コンテンツ制作において，Vocaloid [1] や UTAU[*1] などの歌声合成ソフトウェアによる人工的な歌声を用いた音楽創作が新しい可能性を示している．これら音楽創作に使われるソフトウェアは，楽曲の譜面と歌詞のテキスト情報によって歌声を合成する手法を用いている．このような方式では，ある程度の品質は担保されるが，ユーザがイメージする魅力的な歌唱をデザインすることは難しい．イメージした歌声を容易に作り込むことが可能な歌唱デザイン用インタフェースが実装されれば，計算機による音楽表現の可能性が広がり，歌唱デザインの生産性向上が期待できる．

Vocaloid では，簡単な操作により誰でもビブラートやグロウルなどの歌唱表現を自由に付加することが可能である．Vocaloid をより表情豊かに歌わせる取り組みとして，ユーザの「歌い方」における F0 と音量・音色を対話的に真似るよう歌声合成するシステム VocaListener が提案されている [2], [3]．このシステムは，Vocaloid のインタフェース上で制作した歌唱に対して適用できるプラグイン[*2]として導入されている．VocaListener によって，Vocaloid で制作した歌唱に対してユーザの歌い方の歌唱表現を真似させる加工が自動で行えるようになった．一方，VocaListener を用いて歌唱表現を真似させるためには，ユーザが目的とする歌い方を表現できる必要がある．

ユーザが目的とする歌唱表現をデザインするために，合成音声による歌声や歌声の加工技術が注目されている．一方，ソフトウェアのプリセットによって用意された歌唱表現だけではなく，ユーザ独自が考える歌唱表現を魅力的に歌わせるためには，職人芸的な技術が必要とされる．現在の歌声合成・加工ソフトウェアでは，何度も試行錯誤的にデザインした歌唱表現をレンダリングし，聴覚で確認することを行う煩雑な作業が必要である．煩雑な作業を軽減するために，ユーザのイメージした歌唱表現を作るための手助けとなるインタフェースが求められている．リアルタイムでレンダリングされた音声を聴きながら多くの歌唱表現を試すことができれば，ユーザが持っている歌唱表現のイメージを作り込む一助となることが期待できる．

本稿では，ユーザの持つ歌唱表現のイメージへ向けて音声の高さに相当する基本周波数（F0）を加工できるインタフェース SOUND STONE を提案する．本インタフェースによって，ユーザは簡単に様々な歌声をリアルタイムに加工することができる．

以下，2 章では，本研究に関連するソフトウェアや研究事例について紹介した後，本研究の位置付けについて述べる．3 章では，本インタフェースを実現するために要求される事項について述べる．4 章では，3 章に従って実装した

[1] 山梨大学
University of Yamanashi
[a)] g16tk018@yamanashi.ac.jp
[*1] http://utau2008.web.fc2.com/
[*2] https://www.vocaloid.com/products/show/v4_vocalistener

インタフェースの各種機能について述べる．5 章では，実装した機能が歌唱表現をデザインするための機能として妥当であるか否かについて考察を述べる．最後に 6 章で，本インタフェースの今後の展望について述べる．

2. 関連研究と歌唱デザインの観点からの課題

本章では，本研究で目指すインタフェースに関連するソフトウェアや研究事例について紹介し，歌唱デザインの観点における課題について述べる．

2.1 Vocaloid

Vocaloid は，合成音声による歌声合成ソフトウェアとして世界的に利用されている．本ソフトウェアは，譜面と歌詞の情報を入力するだけで容易に人間らしい歌声で歌わせることができる．歌唱表現の付加については，一般的なビブラートやこぶしなどから，グロウルなどまで様々な表現が可能である．一方，プリセットで用意されていないユーザが独自でイメージした歌唱表現を Vocaloid で表現するためには，ソフトウェアが提供するパラメタを調整しながらレンダリングし，聴覚で確認する煩雑な作業を何度も行わなければならない．

2.2 統計的歌声合成

より人間らしく歌わせるための歌声合成手法として，統計モデルの 1 つである HMM (Hidden Markov Model) による手法が提案されている [4]．本手法は，音声合成ソフトウェアである CeVIO[*3]に導入されている．CeVIO では，Vocaloid と同様に譜面と歌詞の情報を入力するだけで自然な歌声で歌わせることが可能である．また，パラメタによって歌声の感情表現や声質を設定することができる．Sinsy [5] は，同様に HMM による歌声合成手法である．Web 上に譜面情報をアップロードすることで歌声合成機能を用いることができる特徴を持つ．パラメタによって，歌う言語や声質，ビブラート強度などを調整することができる．

深層構造を持つ Deep Neural Network (DNN) に基づく統計的学習法を音声合成に用いる手法が提案されている [6]．本手法を歌声合成に対応させる手法 [7] は，合成された歌声の自然性において高い品質を示している．

これらの手法によって，より自然な歌声・歌い方で歌わせることが可能になる．しかし，自然な歌声であることと，ユーザが持つイメージは必ずしも合致しない．ユーザが持つイメージとは異なる場合，Vocaloid と同様に作り込むための作業は必要となる．

[*3] http://cevio.jp/

2.3 歌声を加工するためのソフトウェア

歌声を編集する F0 制御ソフトウェアとして Auto-Tune[*4]がある．本ソフトウェアの機能であるオートマチックモードでは，ボーカリストの特徴的かつ表現豊かな歌いまわしを残したまま自動で F0 を補正することが可能である．つまり，歌声を入力するだけで自動で F0 制御をするインタフェースが実現されている．本ソフトウェアは，一般的に歌声を自然に F0 制御することで歌としての正しいメロディに調整する技術として使われる．近年では，比較的新しい歌唱表現として，いわゆる「ケロケロボイス」の生成などにも使われている．

歌声編集ソフトウェアの 1 つとして Melodyne[*5]がある．本ソフトウェアは，ユーザの歌声を基に F0 制御した合成音声を生成することで，正しい F0 で歌ったかのように加工することや，しゃくれなどの歌唱表現を付加することが可能である．

Auto-Tune はリアルタイムに再生しながら編集・加工することが可能であるが，Melodyne はオフラインで操作することを前提にしているという違いがある．これらのソフトウェアは，F0 や声色の補正などに力を入れており，ビブラートやこぶしなどの歌唱表現のデザインに特化しているわけではない．

2.4 音声の実時間操作に関する研究

我々の研究グループでは，本システムのように音声や歌唱表現を実時間で変換合成するシステムを提案してきた．その 1 つとして，音声分析合成システム STRAIGHT [8], [9] によって合成された音声の実時間操作を可能にしたインタフェースが提案されている [10]．本インタフェースを用いることで，入力された音声の F0 やフォルマントをリアルタイムで加工することが可能である．

v.morish [11] は，2 人の歌唱者の歌いまわしと声質をユーザの操作に対して実時間でモーフィング率を制御可能なインタフェースである．このインタフェースは，歌いまわしと声質を縦軸と横軸で表しており，ユーザは楽曲の再生中にモーフィング率を制御できる．また，オフラインでモーフィング率の時系列を操作するインタフェースも実装することで，オンライン・オフラインの歌唱デザインの支援について検討している．

これらの研究は，リアルタイムに動作する音声分析合成インタフェースである．本研究で目指すインタフェースは，歌唱表現のデザインを支援するというコンセプトにおいてこれらの研究と異なる．

2.5 本研究の位置づけ

本章では，本稿で提案するインタフェースに関連するソ

[*4] http://www.autotune.mu/products/auto-tune-8/
[*5] http://www.celemony.com/

フトウェアや研究について紹介した．本研究で目指すインタフェースは，直感的な操作性とリアルタイムなレンダリング結果のフィードバックによって，イメージした歌唱表現をデザインすることが目的であり，他研究とは位置付けが異なる．次章では，本研究で目指すインタフェースを実現するために必要な機能要求について述べる．

3. F0 の実時間操作インタフェースの検討

本章では，歌唱デザイン支援を目指した実時間操作インタフェースを実装する場合に要求される機能について述べる．その後，機能を実現するために必要となる音声分析合成手法の検討について述べる．

3.1 歌唱デザイン支援を目指した実時間操作インタフェースに対する要求

歌唱表現のデザインにおける煩雑さを改善するインタフェースを実現するためには，直感的な操作性とユーザの処理を反映した結果がリアルタイムでフィードバックされる機構が不可欠である．例えば，Melodyne の F0 制御システムのリアルタイムフィードバックでは，録音された歌声に対し，F0 を変化させた際の音がリアルタイムでフィードバックされることで聴覚を用いた確認が可能である．しかし，これは歌声の F0 を正しく補正することが目的であり，新たな歌唱表現のデザインを行うことが目的ではないため歌唱表現のデザインには適しているとは言い難い．そこで，歌唱をデザインすることを目的としたインタフェースを実装することを考えると，以下に示す機能が要求される．

- ユーザの操作と合成された音声が直結するリアルタイムな変換音声のレンダリング
- 容易に任意の音声を加工することが可能
- ユーザの操作によってデザインした結果の保存

これらの機能を実装することで，ユーザは操作に直結したフィードバックによって歌唱表現をデザインすることができる．また，容易に任意の音声を加工可能とすることで，様々な発話を数多く試すことができるため，様々な歌唱デザインの実現が期待できる．

3.2 音声分析合成手法

本研究で目指すインタフェースを実装するにあたり，音声をリアルタイムかつ高品質に合成する必要がある．高品質な音声の分析合成が可能な手法として STRAIGHT [8], [9] や TANDEM-STRAIGHT [12]，WORLD [13] などが提案されている．これらは，ボコーダ方式 [14] による音声分析合成手法である．ボコーダ方式における音声分析合成の流れについて図 1 に示す．

基本周波数（F0）は，人間の知覚する音高にほぼ対応している．スペクトル包絡は，音韻や音色に関する情報を有する音響パラメタである．非周期性指標は，音のかすれの

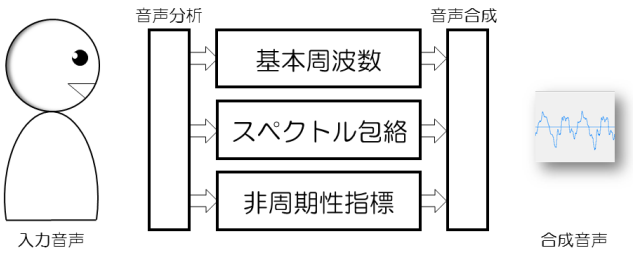

図 1 ボコーダ方式における音声分析合成の流れ

程度を表している．図 1 に示すように，音声を各パラメタに分解した後に再度合成することで音声合成を実現している．この方式による音声合成は，各音響パラメタを独立して操作し合成することができるため，音声加工に有用である特徴をもつ．また WORLD はその方式の中でも，STRAIGHT や TANDEM-STRAIGHT と比較して合成音声の品質が高く処理速度が速いことが示されている [13]．

WORLD (D4C edition [15]) では，高速かつ高性能な F0 推定法である DIO [16] を導入しており，その高速性は後述する録音機能において重要である．このように，WORLD では，高速かつ高精度に各音響特徴パラメタを導出するために独自の方法を提案し推定しているため，実時間操作インタフェースを実装する本研究に適していると考える．さらに，WORLD による実時間音声合成を実現するための拡張と実装例 [17]（以下，実時間 WORLD という）が提案され，Github にてソースコード[*6]が提供されている．以上から本研究では，音声分析合成手法の中でも実時間 WORLD を用いてインタフェースを実装することとした．

3.3 WORLD による実時間音声合成の仕組み

実時間 WORLD では，前節で示した 3 つの音響パラメタから，任意の N サンプル単位で実時間合成を行う機能を用意している．さらに，N サンプル単位で逐次的に合成するために，これらの音響パラメタへのポインタを有するリングバッファが実装されている．リングバッファと音響パラメタとの関係を図 2 に示す．図 2 の下部にある 8 等分された円は，8 つのポインタを持つことができるリングバッファを示している．合成を行う音響パラメタを保持するポインタをリングバッファにリンクすることで逐次追加していき，波形を得るという形で実時間音声合成を行っている．半永久的なリアルタイム音声合成を実現するために，リングバッファにリンクされた音響パラメタは，合成されると自動的にリングバッファから外れるようにしている．本研究で目指すインタフェースでは，実時間 WORLD において合成対象となるフレームを 1 フレーム単位でリングバッファへリンクさせることで，任意のフレームにおける逐次的な音声合成を実現している．

[*6] https://github.com/mmorise/World

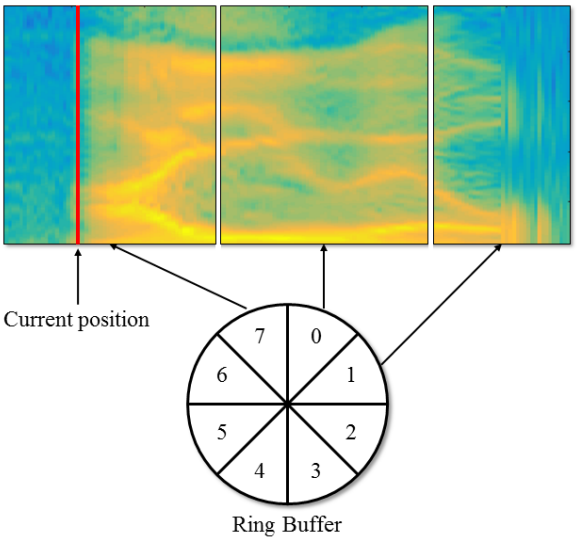

図 2　リングバッファと音響パラメタとの関係．下部の円は，音響パラメタとリンクするためのポインタを持つリングバッファである．この図では，8 つのリングバッファとして実装されている．Current position は，現在までに合成された時刻を示している．Current position に基づいて，合成済みの音声パラメタを持つポインタへのリンクを自動的に破棄している．

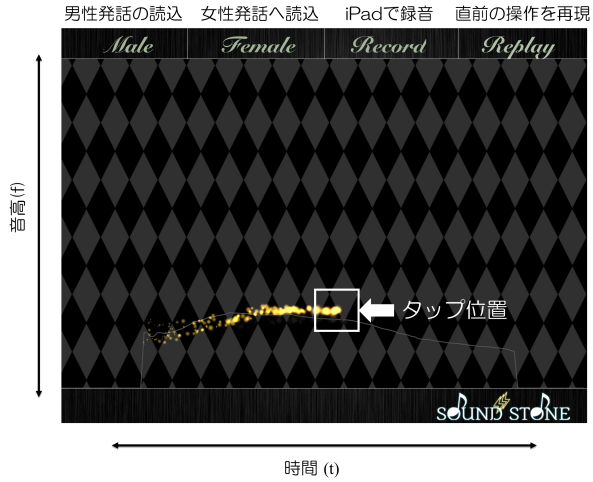

図 3　SOUND STONE の実行画面．横軸は時間，縦軸は F0 である．炎の玉の位置は，現在再生中の時刻と F0 となる．炎の玉は操作する際に指に追従する動作をする．画面下にある薄い線分は，加工対象となる音声の F0 軌跡を表している．

表 1　Recording conditions.

サンプリング周波数	16 kHz
量子化ビット数	16 bit
マイクロホン	iPad が認識した標準入力
チャンネル数	1

4. リアルタイム F0 制御インタフェース SOUND STONE

本章では，前章で示した要望を満たすインタフェース SOUND STONE の実装について説明する．まず，SOUND STONE の実装環境と機能について述べた後，基本機能である「タップした位置に対応した音声合成」とその他 4 つの機能について具体的に説明する．

4.1 SOUND STONE の実装環境と機能

SOUND STONE を開発した OS は，macOS Sierra Ver. 10.12.3 である．統合開発環境は，Xcode Ver. 8.3.2 であり，開発言語は Swift 3.1 である．SOUND STONE におけるリアルタイム出力を実現するために，iOS で動作するソフトウェアフレームワークである AVFoundation を用いて実装した．SOUND STONE は，以下の機能を備えている．

- スワイプ操作による音声の直感的なリアルタイム変換合成機能
- 録音した音声を分析し加工対象とする機能
- スワイプ操作の動きを記憶し，同じ動きを再現する機能

これらの機能により，ユーザの操作を合成音声のレンダリング結果に直結させることや，録音による任意の音声を変換すること等，前章で示した要望を満たしている．実装した SOUND STONE の実行画面を図 3 に示す．

4.2 タップした位置に対応した実時間音声合成

SOUND STONE では，画面上をスワイプ操作することで音声変換し，変換結果が出力される．図 4 では，SOUND STONE をタップした際の音声合成の流れを示している．

GUI で示している操作画面の横軸は時間，縦軸は F0 である．実行デバイスの横幅を音声データのフレーム数で割った値をマージンサイズとしている．つまり，座標の値をマージンサイズで除算した値が WORLD におけるフレーム番号となる．縦軸は周波数であり，60〜800 Hz までを表している．タップ位置を上にずらせば，そのフレームの F0 は高くなり，下にずらせば低くなる．また，適度に上下に揺らせば，擬似的にビブラートの歌唱を表現することも可能である．

操作画面には，加工対象となっている音声の F0 を薄い線分として図示している．これにより，ユーザが音声を操作するうえでの F0 操作の参考にすることができる．

4.3 プリセットとして用意された発話の読込

図 3 に示している「Male」と「Female」ボタンは，それぞれプリセットとして用意された男性音声と女性音声の発話/aiueo/を読み込む機能である．この機能は，後述する録音機能による音声とは異なり，高品質な合成音声に対する歌唱表現のデザインを可能にすることを意図している．

4.4 加工対象音声の録音機能

歌唱デザインを行う際，様々な発話や音色が対象となる．様々な音に対して直感的かつリアルタイムに加工でき

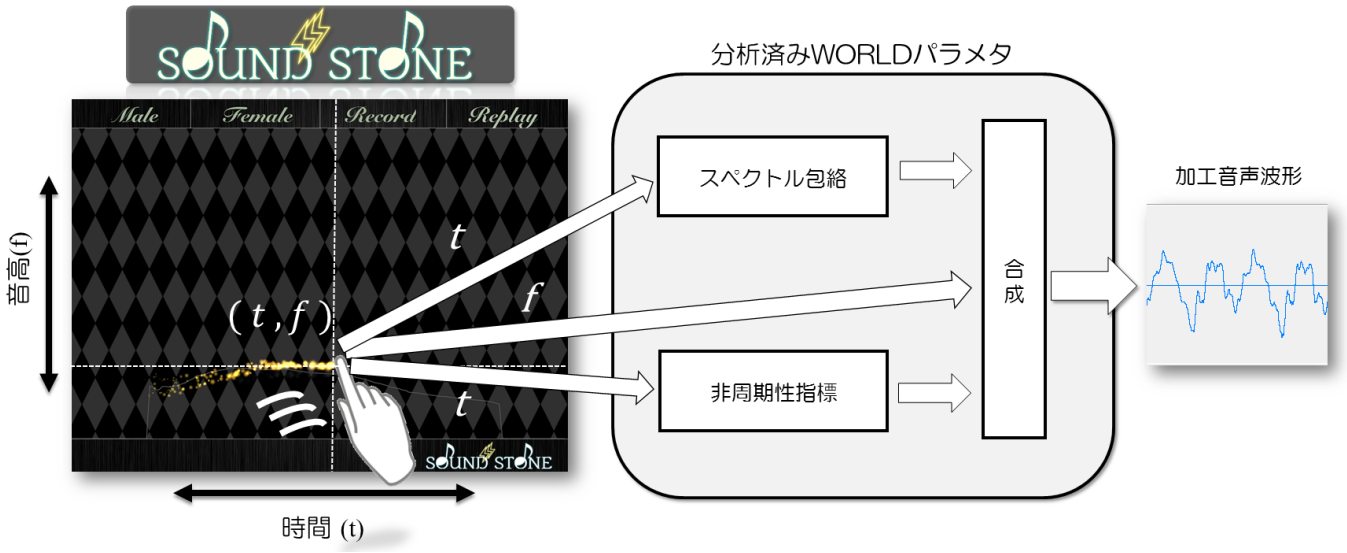

図 4　SOUND STONE における操作を音声合成に適用する流れ

ることは，歌唱表現のイメージを作り込む一助として有用な機能である．本機能は，様々な音声を容易に加工できるようにするために実装した．「Record」ボタンは，任意の発話や歌声を録音することができる．SOUND STONE の Record ボタンでは，押下したタイミングから，離すまでにマイクから得られた音を録音する．録音する音声のパラメタを表 1 に示す．録音後，WORLD による音声分析を行い，録音した音声に対する歌唱デザインが可能となる．

4.5　直前の加工を再現する機能

SOUND STONE は，リアルタイムでレンダリング可能なインタフェースである．SOUND STONE で歌唱デザインを行う際，直前に行った歌唱表現を再度再生できるようにすることで，どのような軌跡やタイミングで加工すればイメージに近づくか確認することができる．「Replay」ボタンは，直前の操作を再現（リプレイ）する機能である．Replay ボタンでは，画面上で加工操作を始めたタイミングから指を離すまでのユーザの操作の動きを記憶し，同じ時系列による歌声合成を可能にする．

5.　機能に関する考察

本章では，SOUND STONE に実装した機能について考察する．特に，SOUND STONE の直感的な操作性と必要な録音環境について述べる．

5.1　直感的な操作性

直感的な操作性という観点において重要なことは，操作に直結したフィードバックを得ることである．SOUND STONE では，ユーザが素早いスワイプ操作を行っても遅延のないレンダリングを行うことが重要である．

リアルタイムの音声合成を実装する際，スワイプ操作中は，タップしているフレームを 1 フレームずつリングバッファへリンクしている．WORLD によって音声波形が合成された際，サンプル数 512 の音声波形を再生する．SOUND STONE では，「フレームをリングバッファへリンクする時間間隔」と「再生するサンプル数」を適切に設定することでシームレスな再生を実現する．再生するサンプル数を増加させると，フレームをリンクする時間間隔を増加させなければ遅延が発生する．一方，サンプル数を減少させると，フレームをリンクする時間間隔を減少させなければ，途切れた音声が出力される．また，サンプル数を減少させたうえで，それに伴いフレームをリンクする時間間隔を減少させる場合，デバイスの処理速度が十分でないと合成が間に合わず途切れた音声が出力される．

現在行われている非公式の主観評価では，サンプル数を 512，リンクする時間間隔を 32 ms とすることで遅延なくシームレスな音声合成の実現に成功している．操作性に関する非公式の主観評価の結果，一般的な歌唱表現の 1 つであるビブラートの表現に対し，32 ms という遅延は操作感に強く影響しないことが示唆されている．

5.2　音声分析を行うために必要な録音環境

SOUND STONE では，様々な音声の加工を可能にするために，録音機能が実装されている．録音機能を使う際，防音室などの録音環境を必要としないことは，利便性において強みである．一般的な話し声を 60 dB とした場合，騒音レベルが 35 dB 程度の環境で録音すると，25 dB 程度の SNR に対応する F0 分析手法であれば，動作する見通しである．

今回，音声分析手法として用いる F0 分析法である DIO は，StoneMask との併用により対雑音性の向上に対する取り組みがなされており，SNR 20 dB 程度までなら十分に

分析可能であると示されている [18]．SOUND STONE の実装にあたり非公式の主観評価として，iPad の内蔵マイクロホンの音質と，防音室ではない環境において十分な分析が可能であるか確認を行った．その結果，騒音レベルが 35 dB 程度の環境で録音機能を利用し，十分に加工ができる音声合成が可能であることを確認した．

録音環境次第では，ファンや電源ノイズなどの低域周波数に対する雑音が分析結果に影響を与えることが確認された．WORLD では，より低 SNR で推定可能な F0 推定法 Harvest [19] が実装されているため，録音環境に応じて適切な F0 推定法を選択可能な機能の実装も検討している．

6. 今後の展望

本稿では，リアルタイムによる F0 の制御を直感的に行うことが可能なインタフェースの機能について紹介した．SOUND STONE の操作性等の具体的な評価は，今後検討していく．現状では SOUND STONE の評価は，直感的操作性の観点から「イメージした歌唱表現（ビブラートやしゃくれなど）をどれくらいイメージ通りに表現することができるか」をコンセプトに行いたいと考える．例えば，ビブラートなどの歌唱表現を行っている歌声を呈示し，SOUND STONE を用いて同様の歌唱表現を再現させることで，どれくらいイメージ通りに表現できたかを評価することができると考えられる．

今後の課題としては，歌唱デザインにおいて重要な要素の 1 つである音色の変換の実装が挙げられる．現状の実装では，変換可能なパラメタは F0 のみとなっており，音色に対する変換はできない．実時間 WORLD では，F0 のみでなくスペクトログラムの変更による合成も実時間で行うことが可能であるため，SOUND STONE の機構を変えずともスペクトログラムの変換を組み込むことが技術的に可能である．しかし，スペクトログラムの変換は F0 の変換とは異なり，時間・周波数・パワーの 3 次元の値の変換となる．3 次元のスペクトログラムの変換を直感的に操作させることは，容易ではない．歌唱デザインのためのスペクトログラムの直感的変換を可能とするインタフェースについては，今後も多くの検討が必要であると考えられる．

謝辞 本研究は，科研費 15H02726, 16H05899, 16K12511, 16K12464 の支援を受けて実施された．

参考文献

[1] H. Kenmochi and H. Ohshita, “VOCALOID – Commercial singing synthesizer based on sample concatenation,” in Proc. INTERSPEECH2007, Special session, pp. 4009–4010, 2007.
[2] T. Nakano and M. Goto, “VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation,” in Proc. SMC, pp. 343–348, 2009.
[3] T. Nakano and M. Goto, “VocaListener2: A singing synthesis system able to mimic a user’s singing in terms of voice timbre changes as well as pitch and dynamics,” in Proc. ICASSP2011, pp. 453–456, 2011.
[4] 大浦圭一郎, “統計モデルに基づいた歌声合成技術の最先端,” 電子情報通信学会誌, Vol. 98, No. 6, pp. 460–466, 2015.
[5] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku and K. Tokuda, “Recent development of the HMM-based singing voice synthesis system – Sinsy,” in Proc. Speech Synthesis Workshop, pp. 211–216, 2010.
[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” CoRR, arXiv preprint arXiv1609.03499, 2016.
[7] B. Merlijn and J. Bonada, “A neural parametric singing synthesizer,” arXiv preprint arXiv1704.03809, 2017.
[8] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneusfrequency-based F0 extraction,” Speech Communication, Vol. 27, pp. 187–207, 1999.
[9] 河原英紀, “Vocoder のもう一つの可能性を探る–音声分析変換合成システム STRAIGHT の背景と展開–,” 日本音響学会誌, Vol. 63, No. 8, pp. 442–449, 2007.
[10] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino and H. Kawahara, “Implementation of realtime STRAIGHT speech manipulation system,” Acoust. Sci. Tech., Vol. 28, No. 3, pp. 140–146, 2007.
[11] M. Morise, M. Onishi, H. Kawahara, H. Katayose, “v.morish’09: A morphing-based singing design interface for vocal melodies, ” Lecture Notes in Computer Science, LNCS 5709 (in Proc. ICEC2009), pp. 185–190, 2009.
[12] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno, “A temporally stable power spectral representation for periodic signals and applications to interfarence-free spectrum, f0, and aperiodicity estimation,” in Proc. ICASSP2008, pp.3933–3936, 2008.
[13] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. Syst., Vol. E99-D, pp. 1877–1884, 2016.
[14] H. Dudley, “Remaking Speech,” J. Acoust. Soc. Am., Vol. 11, No. 2, pp. 169–177, 1939.
[15] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, Vol. 84, pp. 57–65, 2016.
[16] M. Morise, H. Kawahara and H. Katayose, “Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech,” AES 35th International Conference, CD-ROM, pp. 11–13, 2009.
[17] 森勢将雅, “音声分析合成システム WORLD により実時間音声合成を実現するための拡張と実装例,” 情報処理学会音楽情報科学研究会 (夏のシンポジウム), Vol. 2016-MUS-112, No. 20, pp. 1–6, 2016.
[18] M. Morise and H. Kawahara, “TUSK: A framework for overviewing the performance of F0 estimators,” in Proc. INTERSPEECH2016, pp. 1790–1794, 2016.
[19] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in Proc. INTERSPEECH2017, 2017.